Basics of Data Analysis

School of Engineering

August 18, 2017

Introduction

- This is a 2 hour workshop which is intended to serve as an introduction to data analysis.
- The workshop does not provide an exhaustive exploration of all facets of data analytics.
- In the interests of brevity, the workshop will cover some key points on the following topics:
  1. Collecting Data
  2. Storing and Manipulating Data
  3. Creating Effective Graphs
  4. Inference and Statistical Testing
  5. Elementary Ordinary Least Squares Regresssion (OLS)

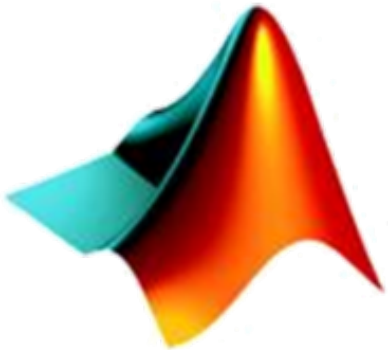
Stucture of the Workshop

Section	Time
Collecting Data	5 min
Storing & Manipulating Data	5 min
Creating Effective Pictures	5 min
Activity 1	15 min
Statistical Analysis	15 min
Activity 2	10 min
Ordinary Least Squares Regression	15 min
Activity 3	10 min

Software for Examples

There are a number of different analysis software packages that you can use. A few of the more popular packages are:

We will be using:



MATLAB

Measurement: Precision and Accuracy

When collecting data in a scientific context we take measurements. There are 2 important aspects when looking at measurements: **precision** and **accuracy**

Definition: **Accuracy**

Accuracy refers to the closeness of a measured value to a standard or known value.

Example

In lab you obtain a weight measurement of 3.2 kg for a given substance, but the actual or known weight is 10 kg, then your measurement is not accurate. In this case, your measurement is not close to the known value.

Measurement: Precision and Accuracy

Definition: Precision

Precision refers to the closeness of two or more measurements to each other.

Example

If in a lab you weigh a given substance five times, and get 3.2 kg each time, then your measurement is very precise.

Precision is independent of accuracy. You can be very precise but inaccurate, as described above. You can also be accurate but imprecise.

Measurement: Precision and Accuracy

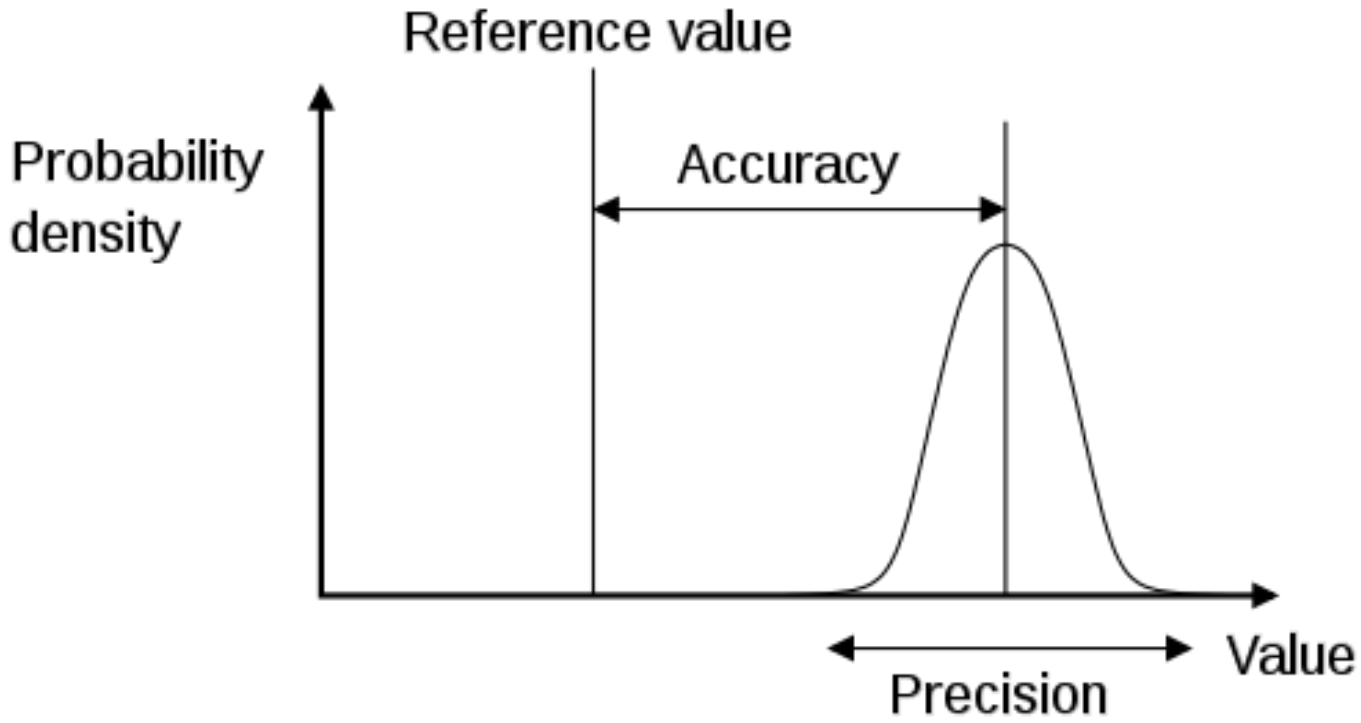
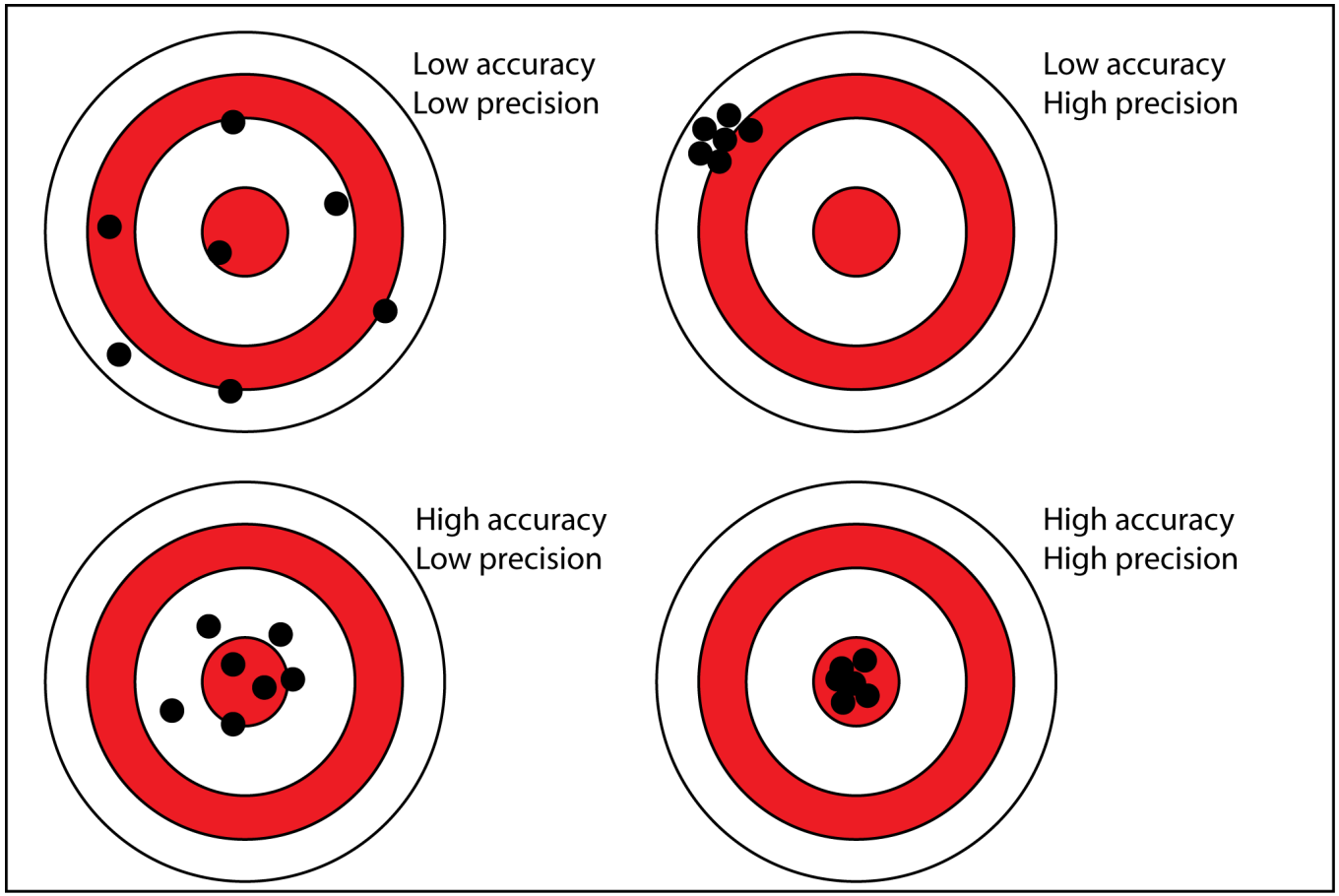


Figure : Accuracy is how far away from the true value a measurement is, and precision is the degree to which repeated measurements under unchanged conditions show the same results

Measurement: Precision and Accuracy



Measurement: Precision and Accuracy

- Accuracy and precision are dependent on the instrument used to take measurements
- Good science will report the **accuracy** and the **precision** of the instruments used to take measurements.

Finding the Absolute Accuracy of an Instrument

1. Find a known value of the phenomena you are measuring, V_A
2. Take a measurement of the phenomena using the instrument, V_O
3. Use the following formula to determine the percentage error:

$$\epsilon = \frac{V_O - V_A}{V_A} \times 100\%$$

Measurement: Precision and Accuracy

Estimating the Precision of an Instrument

1.
2.
3.

It is important to note that the **precision** and **accuracy** are normally given in the data sheet that accompanies your instrument.

Some examples of sentence starters for reporting your precision and accuracy are:

-
-

The Importance of Collecting Data on an Experimental Control

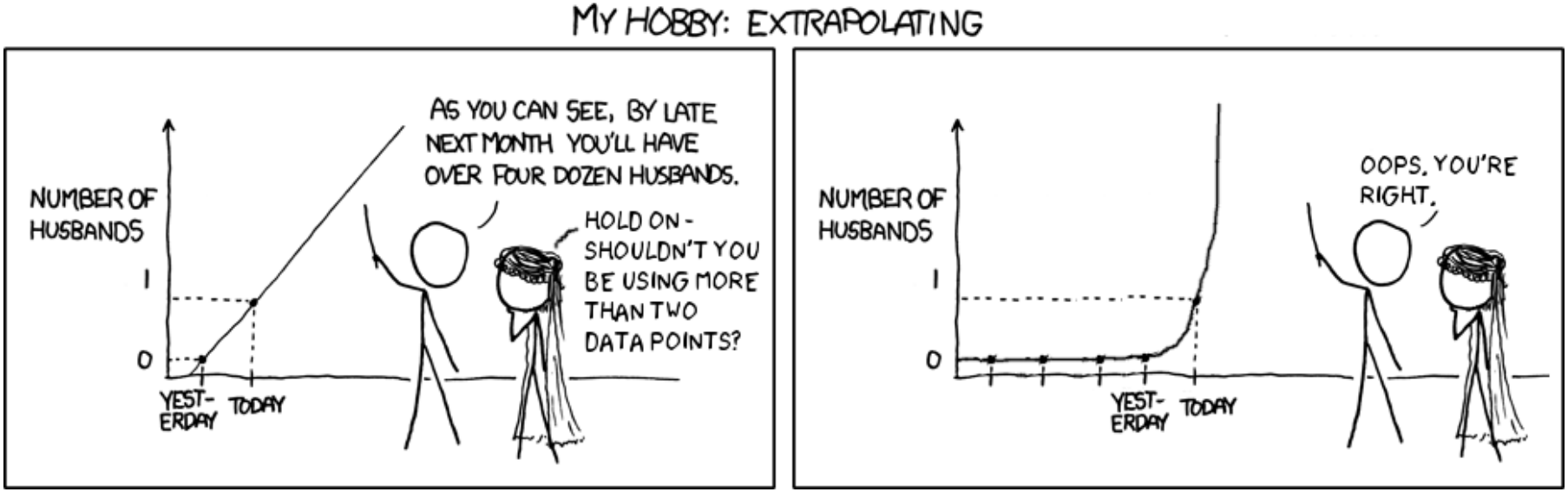
A controlled experiment often compares the results obtained from experimental samples against **control samples**.

Control Samples are practically identical to the experimental sample except for the one aspect whose effect is being tested (the independent variable).

Collection of Data

Understanding the Domain of your Dependant Variables

- Try to have some understanding of the domain you want to investigate for your dependent variables.
- Perhaps even collect data for some margin beyond the edges of these domains.
- This foresight will help you to avoid extrapolation in your analysis - extrapolation is bad science.



Keep Data Storage and Data Analysis Separate

Definition: Data Storage

Data storage simply means recording each observation made.

Definition: Data Analysis

Data analysis means creating graphics, tables or using statistics to uncover patterns or relationships in the data.

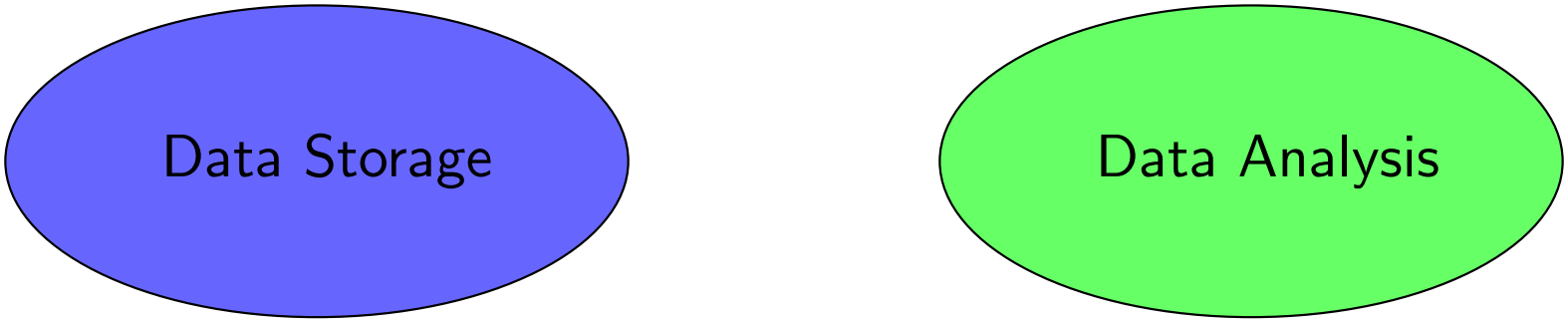


Figure : Think of these two tasks as separate non-overlapping domains - combining the two can make your analysis inflexible, or end up creating extra work for yourself.

Keep Data Storage and Data Analysis Separate

	A	B	C	D	E	F	G	H	I	J	K	L
1						Canopy Cover					Area not occupied (%)	Canopy Cover (%)
2				Rep#	AV Grass Height	Total number of unoccupied Dots						
3	Date	Time	Species			North	East	South	West	AV		
4	13/05/2017	7am-11am	Native	1	40	67	71	65	68	68	70	
5	13/05/2017	7am-11am	Native	2	45	51	71	84	32	60	62	
6	13/05/2017	7am-11am	Native	3	30	48	85	44	62	60	62	
7	13/05/2017	7am-11am	Native	4	60	53	67	44	20	46	48	
8	13/05/2017	7am-11am	Native	5	40	31	6	5	32	19	19	
9	**AV**				43.00					50.30	52.31	47.69
10	**SE**				4.90					8.68	9.03	9.03
11	22/05/2017	10:30am-1:00pm	Native	1	11	14	13	6	6	10	10	
12	22/05/2017	10:30am-1:00pm	Native	2	55	24	36	43	60	41	42	
13	22/05/2017	10:30am-1:00pm	Native	3	24	16	9	2	21	12	12	
14	22/05/2017	10:30am-1:00pm	Native	4	33	55	50	37	50	48	50	
15	22/05/2017	10:30am-1:00pm	Native	5	36	51	56	65	27	50	52	
16	**AV**				31.80					32.05	33.33	66.67
17	**SE**				7.25					8.78	9.13	9.13

Figure : An example of mixing these two domains

How Should I Structure My Data?

Definition: Observation

Observation refers to any data collected during the scientific activity.

Definition: Variable

A **variable** is any factor, trait, or condition that can exist in differing amounts or types.

When collecting your data, one good practice is to record your data in a matrix array with each **row** representing an observation, and each **column** representing a variable.

How Should I Structure My Data?

variables →

observations ↓

	A	B	C	D	E	F	G	H	I
1	pclass	survived	name	sex	age	sibsp	parch	ticket	fare
2	1	1	Allen, Miss. Elisabeth Walton	female	29	0	0	24160	211.3375
3	1	1	Allison, Master. Hudson Trevor	male	0.917	1	2	113781	151.5500
4	1	0	Allison, Miss. Helen Loraine	female	2	1	2	113781	151.5500
5	1	0	Allison, Mr. Hudson Joshua Creighton	male	30	1	2	113781	151.5500
6	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25	1	2	113781	151.5500
7	1	1	Anderson, Mr. Harry	male	48	0	0	19952	26.5500
8	1	1	Andrews, Miss. Kornelia Theodosia	female	63	1	0	13502	77.9583
9	1	0	Andrews, Mr. Thomas Jr	male	39	0	0	112050	0.0000
10	1	1	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53	2	0	11769	51.4792
11	1	0	Artagaveytia, Mr. Ramon	male	71	0	0	PC 17609	49.5042
12	1	0	Astor, Col. John Jacob	male	47	1	0	PC 17757	227.5250
13	1	1	Astor, Mrs. John Jacob (Madeleine Talmadge Force)	female	18	1	0	PC 17757	227.5250
14	1	1	Aubart, Mme. Leontine Pauline	female	24	0	0	PC 17477	69.3000
15	1	1	Barber, Miss. Ellen "Nellie"	female	26	0	0	19877	78.8500
16	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80	0	0	27042	30.0000
17	1	0	Baumann, Mr. John D	male		0	0	PC 17318	25.9250
18	1	0	Baxter, Mr. Quigg Edmond	male	24	0	1	PC 17558	247.5208
19	1	1	Baxter, Mrs. James (Helene DeLaudeniere Chaput)	female	50	0	1	PC 17558	247.5208
20	1	1	Bazzani, Miss. Albina	female	32	0	0	11813	76.2917
21	1	0	Beattie, Mr. Thomson	male	36	0	0	13050	75.2417
22	1	1	Beckwith, Mr. Richard Leonard	male	37	1	1	11751	52.5542
23	1	1	Beckwith, Mrs. Richard Leonard (Sallie Monypeny)	female	47	1	1	11751	52.5542

Figure : In this matrix array, each **observation** is recorded in a **row**, and each **variable** is listed over a **column**.

How Should I Structure My Data?

	A	B	C	D	E	F	G	H	I	J	K	L
1						Canopy Cover					Area not occupied (%)	Canopy Cover (%)
2				Rep#	AV Grass Height	Total number of unoccupied Dots						
3	Date	Time	Species			North	East	South	West	AV		
4	13/05/2017	7am-11am	Native	1	40	67	71	65	68	68	70	
5	13/05/2017	7am-11am	Native	2	45	51	71	84	32	60	62	
6	13/05/2017	7am-11am	Native	3	30	48	85	44	62	60	62	
7	13/05/2017	7am-11am	Native	4	60	53	67	44	20	46	48	
8	13/05/2017	7am-11am	Native	5	40	31	6	5	32	19	19	
9	**AV**				43.00					50.30	52.31	47.69
10	**SE**				4.90					8.68	9.03	9.03
11	22/05/2017	10:30am-1:00pm	Native	1	11	14	13	6	6	10	10	
12	22/05/2017	10:30am-1:00pm	Native	2	55	24	36	43	60	41	42	
13	22/05/2017	10:30am-1:00pm	Native	3	24	16	9	2	21	12	12	
14	22/05/2017	10:30am-1:00pm	Native	4	33	55	50	37	50	48	50	
15	22/05/2017	10:30am-1:00pm	Native	5	36	51	56	65	27	50	52	
16	**AV**				31.80					32.05	33.33	66.67
17	**SE**				7.25					8.78	9.13	9.13

Figure : This is a bad way to structure your data set

How Should I Structure My Data?

	A	B	C	D	E	F	G	H	I	J	K
1	start_time	stop_time	date	type	species	rep_no	av_grass_height	cc_north	cc_east	cc_south	cc_west
2	7:00:00 AM	11:00:00 AM	13/05/2017	AVN	native	1	40	67	71	65	68
3	7:00:00 AM	11:00:00 AM	13/05/2017	AVN	native	2	45	51	71	84	32
4	7:00:00 AM	11:00:00 AM	13/05/2017	AVN	native	3	30	48	85	44	62
5	7:00:00 AM	11:00:00 AM	13/05/2017	AVN	native	4	60	53	67	44	20
6	7:00:00 AM	11:00:00 AM	13/05/2017	AVN	native	5	40	31	6	5	32
7	10:30:00 AM	1:00:00 PM	22/05/2017	AVN	native	1	11	14	13	6	6
8	10:30:00 AM	1:00:00 PM	22/05/2017	AVN	native	2	55	24	36	43	60
9	10:30:00 AM	1:00:00 PM	22/05/2017	AVN	native	3	24	16	9	2	21
10	10:30:00 AM	1:00:00 PM	22/05/2017	AVN	native	4	33	55	50	37	50
11	10:30:00 AM	1:00:00 PM	22/05/2017	AVN	native	5	36	51	56	65	27
12	9:30:00 AM	12:00:00 AM	5/06/2017	AVN	native	1	40	12	15	0	28
13	9:30:00 AM	12:00:00 AM	5/06/2017	AVN	native	2	30	56	51	21	77
14	9:30:00 AM	12:00:00 AM	5/06/2017	AVN	native	3	25	4	1	10	2

Figure : This is the data set on the previous slide reformatted to make it more readily usable for exploration and analysis

Manipulating Data: Large Datasets

- Large datasets will need to be manipulated in a software package like MATLAB
- Large datasets typically arise from scientific instruments that sample at high frequencies
- Do not try to manage large datasets in a spreadsheet application - this will be unwieldy, and provide little insight
- Take some time to understand how to work with your dataset in your chosen software package - each package will have its own quirks

Manipulating Data: Medium Datasets

- You can manipulate a medium size data set directly in a software package like MATLAB, however, this is often clunky and can make basic visualisation of your data a pain.
- An easier way to manipulate data is with a spreadsheet software package, like Excel or Google Sheets.
- The main advantage of using a spreadsheet software is that it will make visualising your data set easier, and it is very simple to save as a **.csv** file which you can then import to an analysis software package.
- **NOTE: this slide is not suggesting that you should undertake exploration or analysis of your dataset in a spreadsheet. Often these applications are ill-equipped to perform the exploration or analysis that you require**

Tips for Storing Data: Large Datasets

- Combining large datasets from two or more different experiments will most likely not be helpful
- When you save your data sets make sure you use an appropriate naming convention, for example:

  `experiment_<number of experiment>_<yyyymmdd>`

- Using a naming convention like the one above will automatically keep you files in order

Tips for Storing Data: Medium Datasets

- Don't be afraid to add redundant information to your master file - it may help with your analysis later.
- Try to have only one data file called your **master**, this may mean adding extra variables like *date*. **Note: never delete any of your data files - even after combining them into a single data file**
- When you update your master file, you should not overwrite the old one - this will help to provide some version control. A good naming convention would be:

`master_<yyyymmdd>`

Why Use Graphs?

- Data analysis is a process, which can be split up into three main operations:
  1. **Collection:** collecting data, obviously
  2. **Exploration:** uncovering patterns, trends, and gathering corroborative evidence to support arguments
  3. **Analysis:** using statistics to make inferential claims

- One of the key tools to help explore the data is to use graphs

- Graphs can help you find corroborative evidence for your hypotheses, and can be used to help strengthen arguments

- Graphs can also be used to help you uncover patterns or relationships which may not be initially evident

To Plot or Not to Plot?

- The purpose of plotting scientific data is to visualise variation or show relationships between variables
- Not all data sets require a plot
- If there are only one or two points, it is easy to examine the numbers directly - little or nothing is gained by putting them on a graph
- Similarly, if there is no variation in the data, it is easy enough to see or state the fact without using a graph of any sort.

Graphing: The Fundamentals

- Ensure that you are **labelling your axes** and including the **units of measurement** on these axes - many thesis submissions neglect this aspect
- Every figure has a caption placed beneath it that describes the content in a few lines
- The caption usually starts with a sequential figure number that is used for reference elsewhere in the paper.

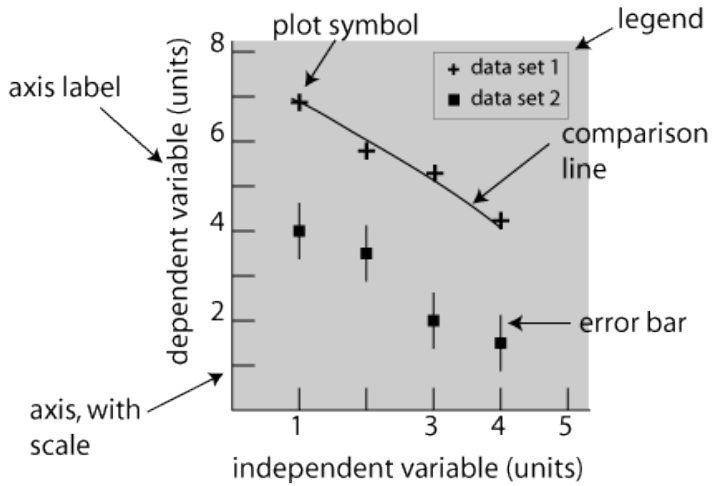


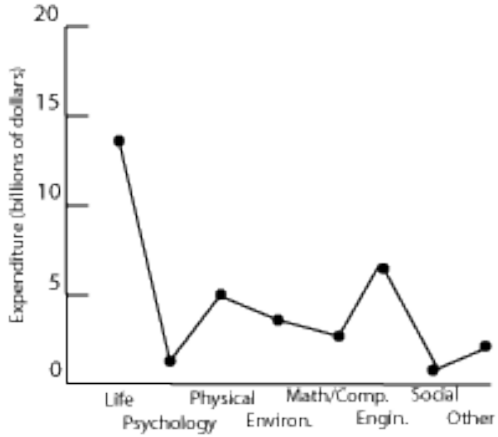
Figure : A caption will help to explain what the figure is about

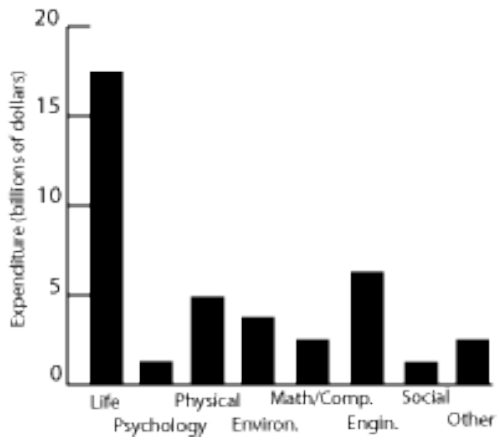
Graphing: The Fundamentals

- A legend in the form of a text box in the plot area is sometimes used to identify the symbols associated with each data set.
- If present, the legend must be placed in the plot area so that it does not detract from the display of data.
- In formal technical publications legend information is often placed in the caption, but a legend may be useful for other presentations such as posters or talks.

Graphing: The Fundamentals

When a graph is appropriate, it must be of an appropriate type to avoid misleading the reader.





- Both plots above show US research expenditures by discipline in 2000.
- The scatter plot on the left is incorrect - it suggests a relationship between the variables on the two axes implied by the connecting lines.
- The horizontal axis is just a list of disciplines with no inherent ordering, no relationship can exist.

Choosing the Right Graph (Small Dataset)

- There is a limited amount of choice when it comes to plotting small data sets
- In fact, some plots of small data sets can be misleading and promote erroneous conclusions

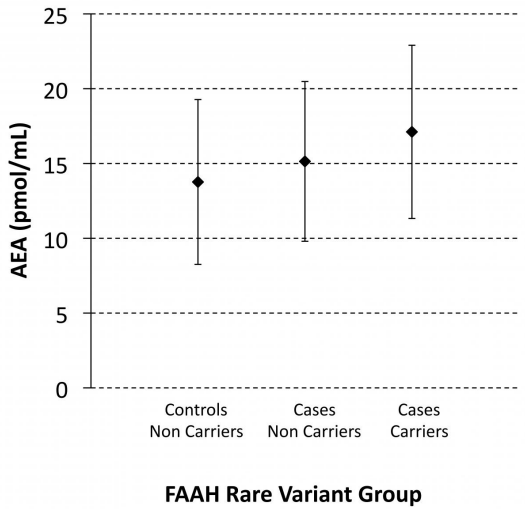


Figure : A plot which can work for small data sets is plotting the mean, maximum, and minimum value, however, be wary of drawing any strong conclusions from these graphs

Choosing the Right Graph (Med. & Large Dataset): Distributions

A picture can significantly aid the process of describing the distribution of your data for a variable.

Type of Graph	Appropriate Use
Histogram	Used with continuous and some discrete variables
Box Plot	Used with continuous and discrete variables
Ordered Bar Charts	Used when an ordinal metric can be placed on a categorical variable

Choosing the Right Graph (Med. & Large Dataset): Distributions

- A histogram works well for a continuous variable
- A histogram can also work for some discrete variables
- Histograms should be avoided for categorical variables
- Be judicious when choosing your bin size - the wrong choice can reduce the effectiveness of your histogram

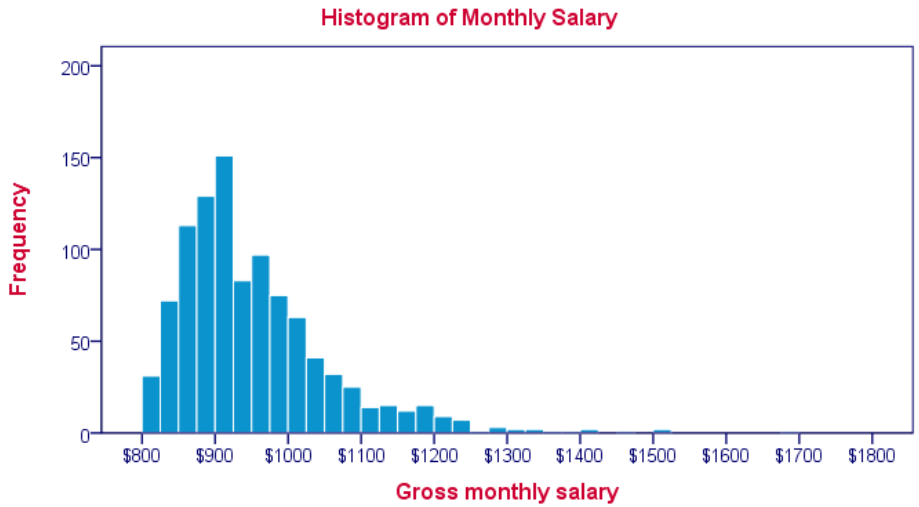


Figure : An example of a histogram

Choosing the Right Graph (Med. & Large Dataset): Distributions

- If your bin size is too big you risk all of your data falling into only one or two categories
- If you bin size is too small, you risk only one observation falling into each bin
- Both of these situations will yield bad histograms

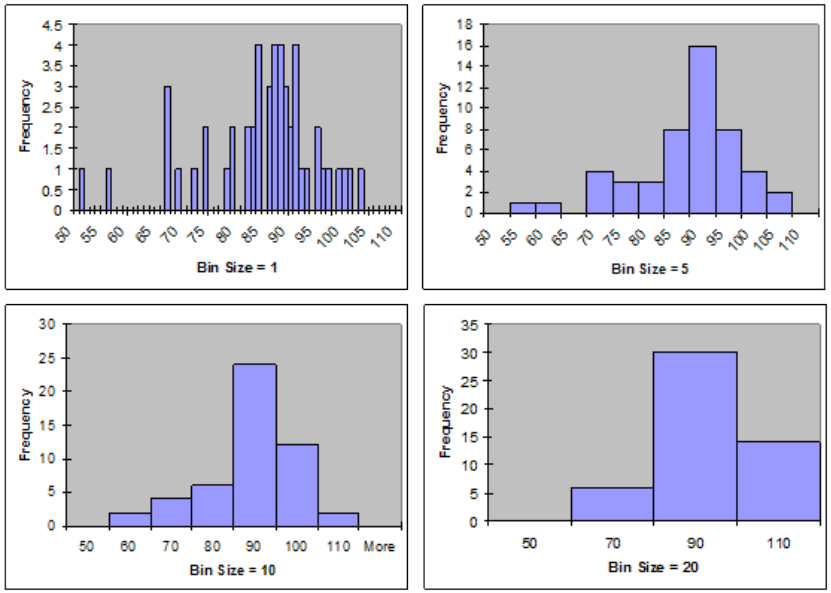


Figure : Most software packages will choose an optimal bin size for you

Choosing the Right Graph (Med. & Large Dataset): Distributions

- A **box plot** is a concise way to show the distribution of data
- These plots are generally very effective when comparing the distributions of two or more data sets for a variable
- **Box plots** are very useful for showing how a distribution changes over time

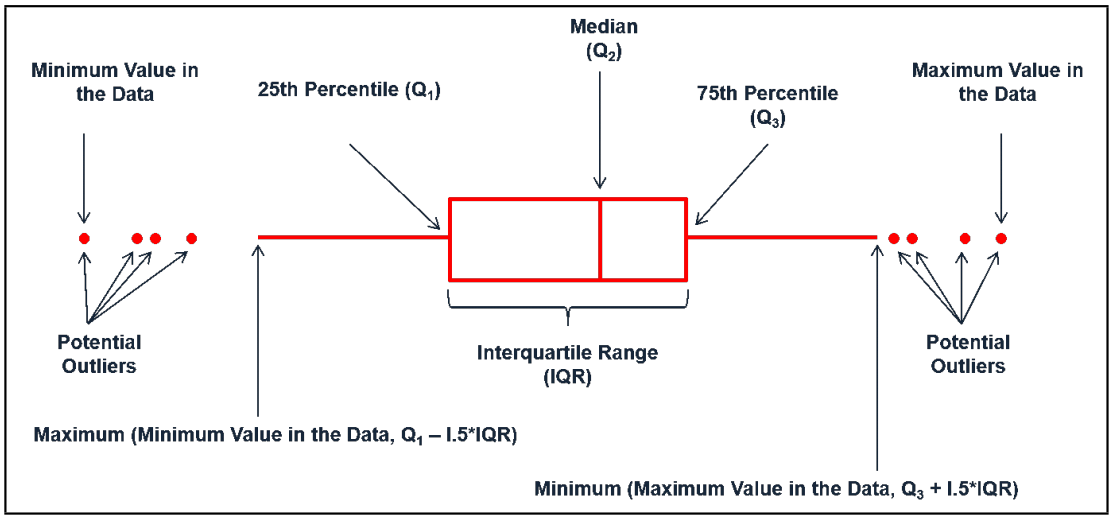


Figure : The key features of a **box plot**

Choosing the Right Graph (Med. & Large Dataset): Distributions

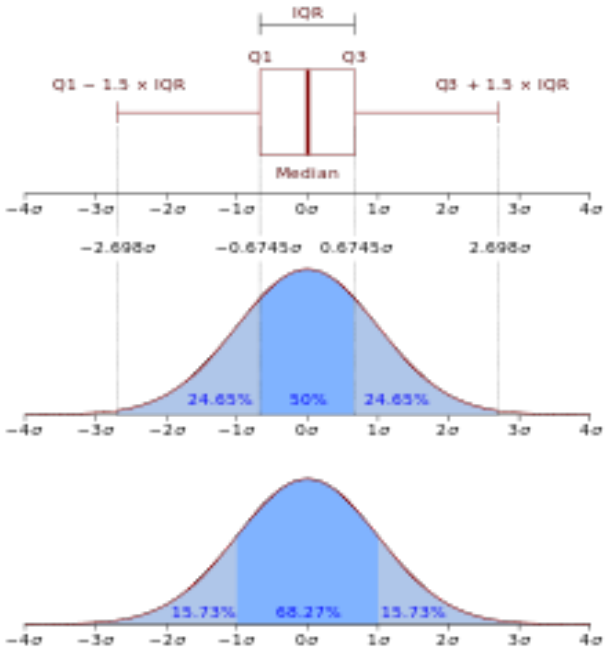


Figure : Mapping the **box plot** to a distribution

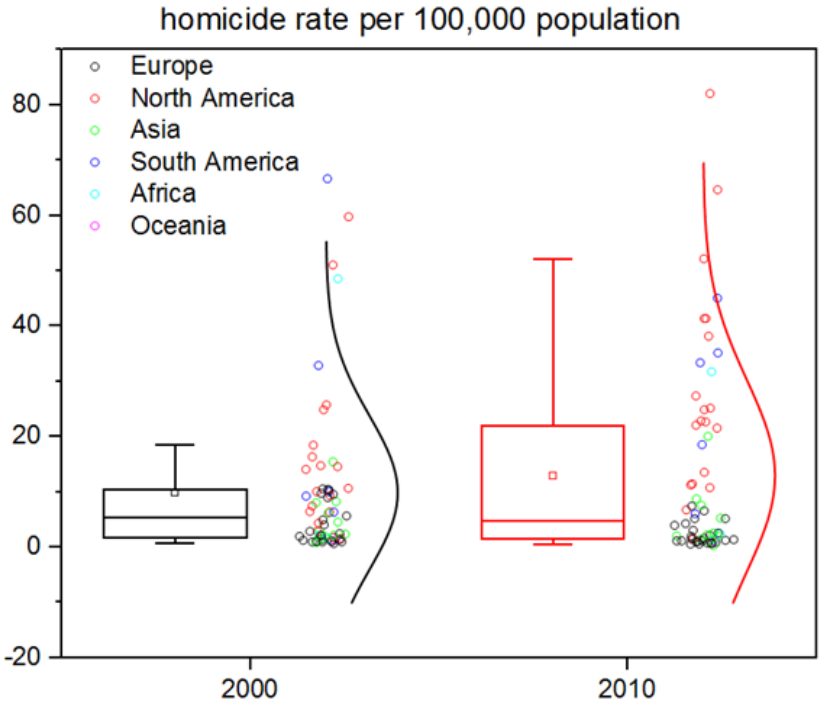


Figure : How a **box plot** maps to a skewed distribution

Choosing the Right Graph (Med. & Large Dataset): Distributions

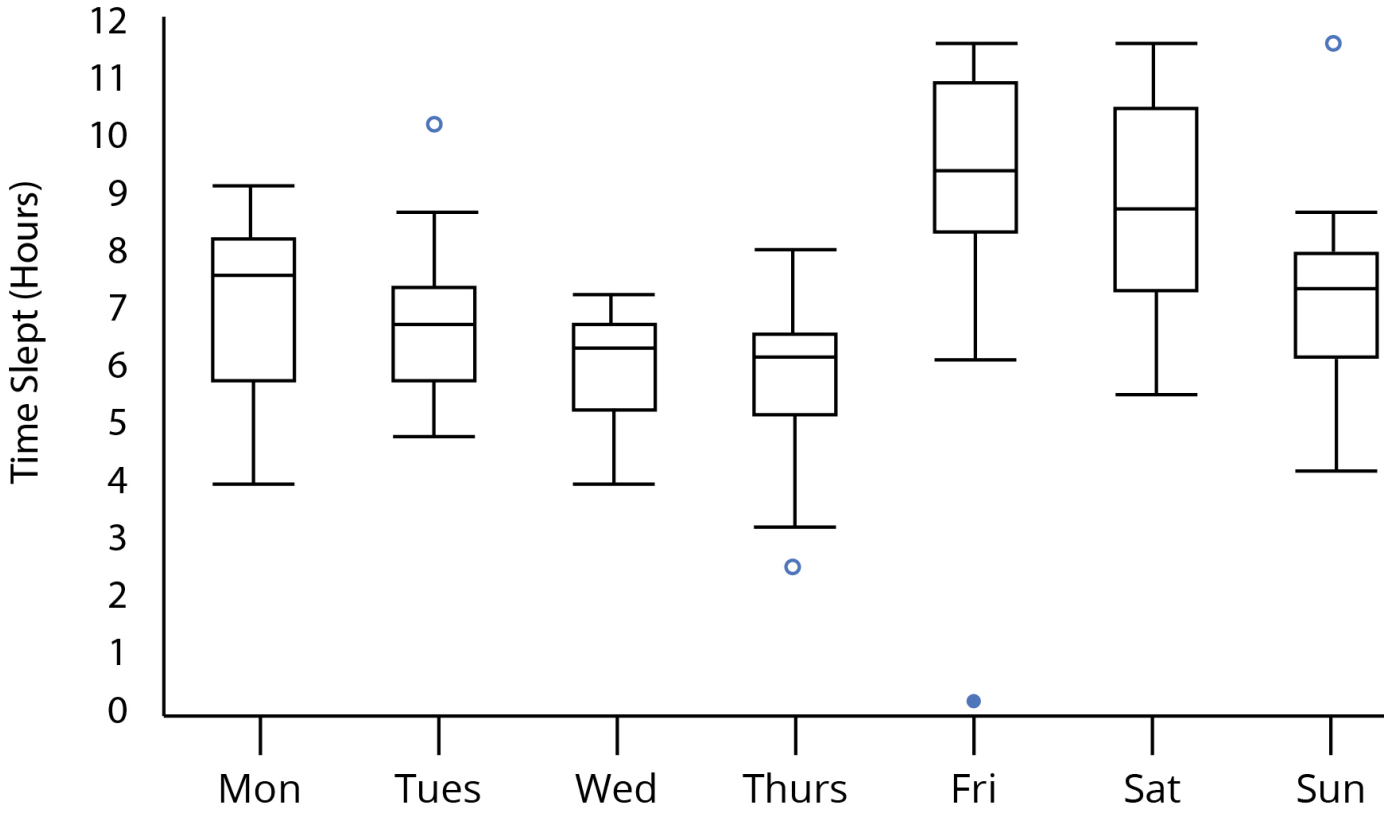


Figure : An example showing how **box plots** can be used to explain a changing distribution over time

Choosing the Right Graph (Med. & Large Dataset): Distributions

If your variable is categorical, but has an ordinal metric which can be applied to it, then a **bar chart** will effectively describe the distribution

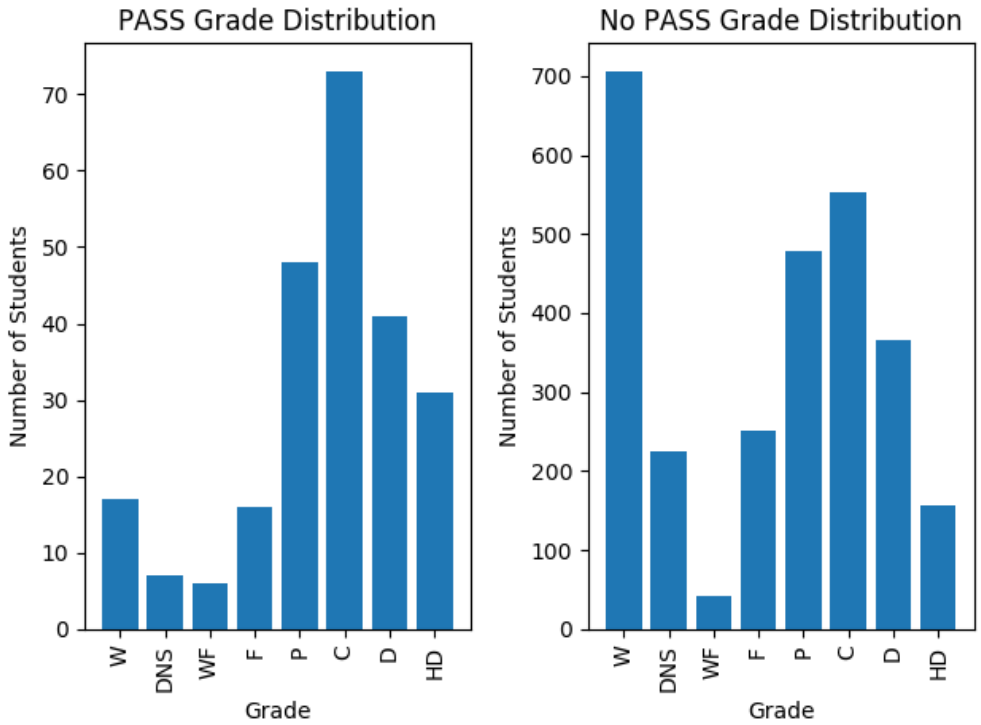


Figure : Grade data is categorical, but has an underlying order to it. The side by side graphs show a shift in the grade distribution

Choosing the Right Graph (Med. & Large Dataset): Scatter Plots

- Scatter plots can give you an idea of the relationship between two variables - a great precursor to a regression analysis
- Scatter plots work really well for continuous variables
- One variable is on the x-axis, another variable is on the y-axis, and each point in the Cartesian plane is an observation

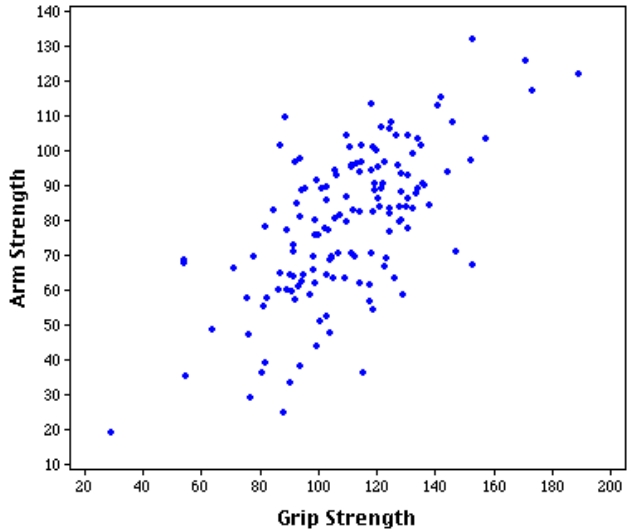


Figure : A scatter plot of two continuous variables

Choosing the Right Graph (Med. & Large Dataset): Scatter Plots

Using scatter plots for **discrete data** can still be effective, but can be a bit tricky to get an effective picture because observations can overlap each other

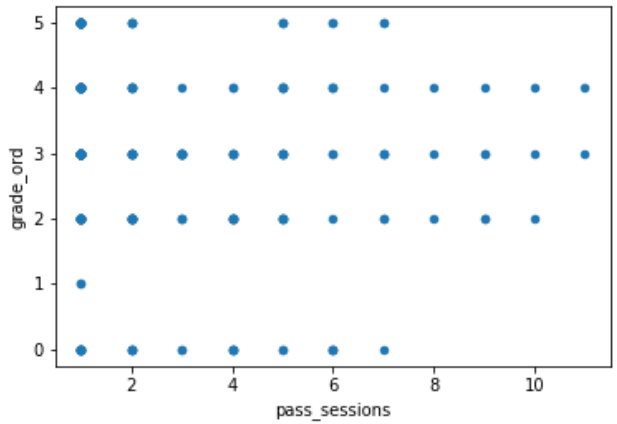


Figure : A scatter plot of discrete data

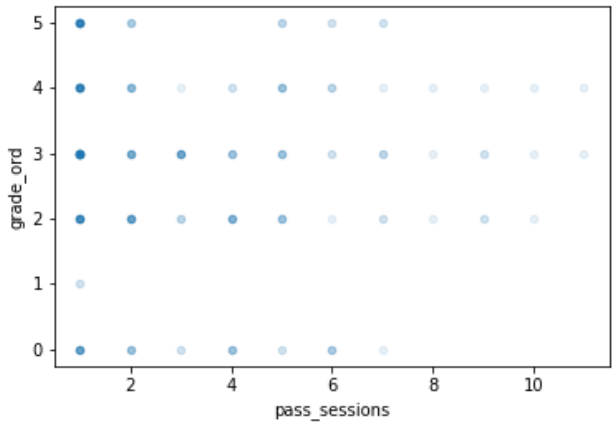


Figure : A better scatter plot of discrete data - each observation carries a 0.1 visual weight, helping readers to better understand where observation locations

Choosing the Right Graph (Med. & Large Dataset): Two Datasets on One Graph?

- There are very few situations where it is appropriate to use two different scales on the same plot
- It is very easy to mislead the viewer of the graphic - this is bad science

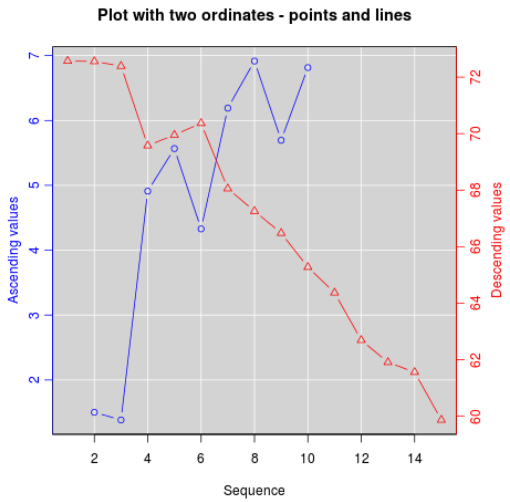


Figure : The intersecting plots seem to imply something

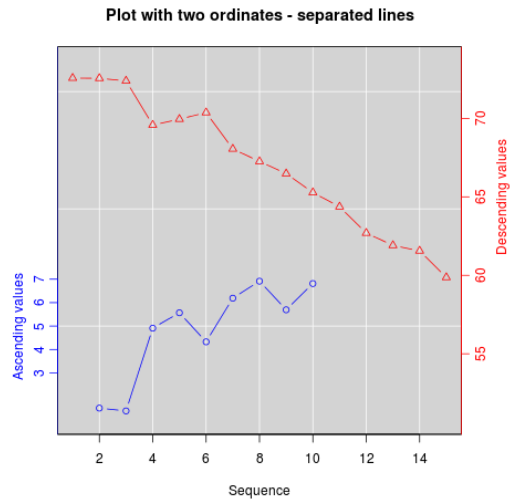


Figure : A simple change of scale reveals the original plot is misleading

It's now time to attempt activity 1

What is this section about?

This section of the workshop is really only about three things:

- Understanding **population parameters**, **sample statistics**, and the **sampling distribution of means**;
- How you can **estimate population parameters** using **confidence intervals**
- Understanding how to correctly set up a **hypothesis test**, why you might like to do this, and when this statistical analysis is appropriate;

Populations and Samples

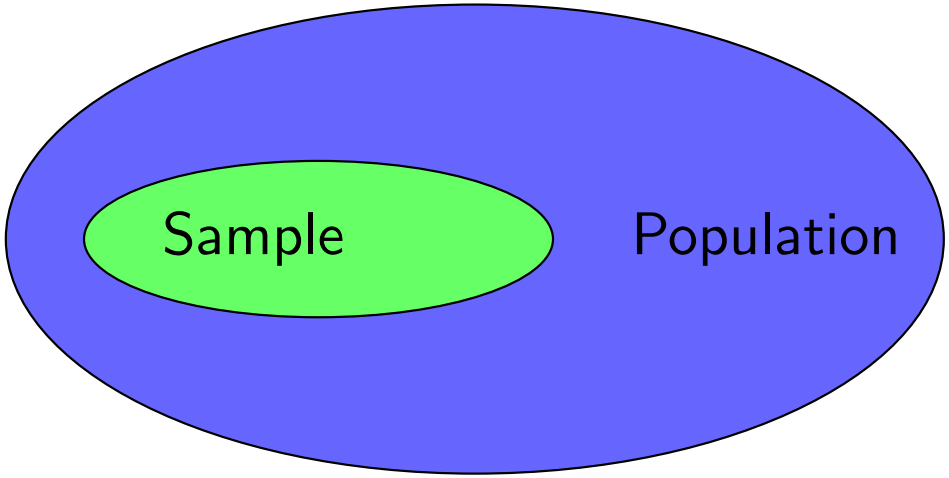


Figure : A sample is a small set of a population

Definition: Population

A **population** is a set of similar items or event which is of some interest for some question or experiment

Definition: Sample

A **sample** is a set of data collected from a statistical population

Parameters and Statistics

Definition: Parameter

A **parameter** is any summary number, like an average or percentage, that describes the entire **population**.

Definition: Statistic

A **statistic** is any summary number, like an average or percentage, that describes a **sample**.

Table : Shows the correct notation used to represent parameters and statistics

	Population Parameter	Sample Statistic
Mean	μ	$\bar{x}$
Standard Deviation	σ	s

The Sampling Distribution of Means

Definition: Continuous Random Variable

A **continuous random variable**, usually denoted as X, is a variable whose value is assigned from a set of possible outcomes according to some probability distribution.

- Suppose we start taking samples from some population defined by a random variable X
- For each of these samples we calculate the mean, $\bar{x}$
- Each $\bar{x}$ will be different due to sampling variability

The Sampling Distribution of Means

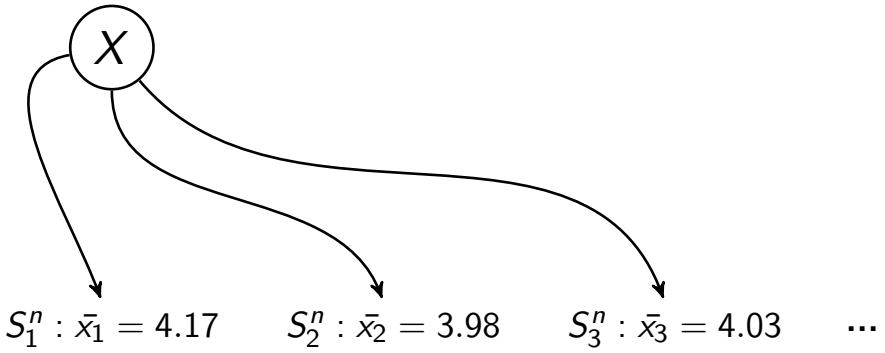


Figure : Samples of size n, denoted by S_i^n, are taken from the population represented by random variable X. The distribution of X is unknown, but it has some true mean μ, and some true standard deviation σ. For each sample the mean, $\bar{x}_i$, is calculated.

- The sample means form a distribution of their own, $\overline{X}$
- According to the central limit theorem, the sample means are normally distributed, that is $\overline{X} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}})$

The Sampling Distribution of Means

If the sample size, n, is big enough then,

$$\overline{X} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}})$$

This is true **irrespective** of whether or not X is normally distributed.

- The mean of the sample means is the true population mean, that is:

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

- The standard deviation of the sample means, also called the *standard error*, is given by:

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

The t-distribution

- Statistical test should be done using z-scores if the population parameter, σ , is known
- The problem with this is that we often don't know the true population parameter σ
- Additionally if our underlying distribution is not normal and our sample is too small, then we can get erronous results from inference
- What do we do instead?
- **Use the t-distribution**

The t-distribution

If the population parameter, σ, is unknown, or the sample size is less than 30, then:

- We use the sample statistic s to approximate σ; and
- The normal distribution is abandoned in favour of the student's t-distribution

The key conditions to check prior to using the t-distribution:

- When **sample size is less than 15**, use t-distribution only when population is very close to normal
- When **sample size is between 15 and 30**, the t-distribution can be used if the variable is not far from normal
- When **sample size is large**, we can always use t-distribution provided there are no extreme outliers that cannot be removed

Inference

Now that the justification for using the t-distribution has been covered, we turn our focus to inference which relies on t-scores.

NOTE: these inference techniques also work for z-scores, if they are appropriate to use

Definition: **Inference**

Statistical **inference** is the theory, methods, and practice of forming judgements about the parameters of a population.

We will look at two forms of inference:

- Estimating population means using confidence intervals
- Hypothesis testing

Estimating the Population Mean Using Confidence Intervals

- Suppose we want to estimate an actual population mean μ - we may only be able to obtain $\bar{x}$, the mean of a sample randomly selected from the population of interest
- We can use $\bar{x}$ to find a range of values:

$$\text{Lower value} < \mu < \text{Upper value}$$

- This range of values provides us with some confidence that it contains the population mean
- The range of values is called a **confidence interval**

Defintion: Confidence Interval

A **confidence interval** is a range of values so defined that there is a specified probability that the value of a parameter lies within it.

Estimating the Population Mean Using Confidence Intervals

- We are interested in finding a (1-α)100% confidence interval for the population mean μ, where $0 < |\alpha| < 1$
- The formula for the confidence interval in words is:

$$\text{Sample mean} \pm (\text{t-multiplier} \times \text{standard error})$$

The formula for the **confidence interval** in notation is:

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

The *t-multiplier*, which we denote as $t_{\alpha/2,n-1}$, depends on the sample size through $n-1$ (called the *degrees of freedom*), and the confidence level $(1-\alpha) \times 100$ through $\alpha/2$.

Estimating the Population Mean Using Confidence Intervals

Example

Example of confidence interval calculation using t distribution

Hypothesis Testing

The general idea of **hypothesis testing** involves:

1. Making an initial assumption
2. Collecting evidence (data)
3. Based on the available evidence (data), deciding whether to reject or not reject the initial assumption

Every hypothesis test regardless of the population parameter involved requires the above three steps.

Hypothesis Testing

Example

Is normal body temperature really 36.67°C?

Consider the population of many adults. A researcher wants an answer to the question: "Is the average adult body temperature 36.67°C?" To answer his research question, the researcher starts by assuming that the average adult body temperature was 36.67°C.

Then, the researcher goes out and tries to find evidence that refutes his initial assumption. In doing so, he selects a random sample of 130 adults. The average body temperature of the 130 sampled adults is 37.81°C.

Then, the researcher uses the data he collected to make a decision about his initial assumption. It is either likely or unlikely that the researcher would collect the evidence he did given his initial assumption that the average adult body temperature is 36.67°C?

Hypothesis Testing

Example

(Continued) Is normal body temperature really 36.67°C?

If it is likely, then the researcher does not reject his initial assumption that the average adult body temperature is 36.67°C. There is not enough evidence to do otherwise.

If it is unlikely, then:

- either the researcher's initial assumption is correct and he experienced a very unusual event;
- or the researcher's initial assumption is incorrect.

In the practice of statistics, if the evidence (data) we collected is unlikely in light of the initial assumption, then we reject our initial assumption.

Hypothesis Testing

- In statistics, we always assume the null hypothesis is true. That is, the null hypothesis is always our initial assumption.
- In statistics, the data are the evidence.
- In statistics, we always make one of two decisions:
  - We either **reject the null hypothesis**; or
  - We **fail to reject the null hypothesis**

Steps for Hypothesis Testing

STEP 1 - Setting up two competing hypotheses

- Each hypothesis test includes two hypothesis about the population
- One is the null hypothesis, notated as H_0, which is a statement of a particular parameter value
- This hypothesis is assumed to be true until there is evidence to suggest otherwise
- The second hypothesis is called the alternative, or research, hypothesis, written as H_a
- The alternative hypothesis is a statement of a range of alternative values in which the parameter may fall

Steps for Hypothesis Testing

STEP 2 - Set some level of significance called alpha

- This value is used as a probability cutoff for making decisions about the null hypothesis
- The most common alpha value is 0.05 or 5%. Other popular choices are 0.01 (1%)

STEP 3 - Calculate a test statistic

- Gather sample data and calculate a test statistic where the sample statistic is compared to the parameter value
- The test statistic is calculated under the assumption the null hypothesis is true

Steps for Hypothesis Testing

STEP 4 - Calculate probability value (p-value)

- A p-value is found by using the test statistic to calculate the probability of the sample data producing such a test statistic or one more extreme
- A small p-value means that we reject the null hypothesis
- A *small p-value* means that the *probability of getting this sample is low, assuming the null hypothesis is true*

STEP 5 - Make a test decision about the null hypothesis

- In this step we decide to either reject the null hypothesis or decide to fail to reject the null hypothesis
- Notice we do not make a decision where we will accept the null hypothesis

Steps for Hypothesis Testing

STEP 6 - State an overall conclusion

- Once we have found the p-value, and made a statistical decision about the null hypothesis, we want to summarise our results into an overall conclusion for our test
- Include a sentence starter on what you can say for rejection of null `text`
- Include a sentence starter on what you can say for failure to reject null `text`

Type I and Type II Errors

When doing hypothesis testing, two types of mistakes may be made and we call them Type I error and Type II error.

Decision	*Reality*	
	H_0 is true	H_0 is false
Reject H_0	Type I error	Correct
Fail to Reject H_0	Correct	Type II error

Figure : Table shows the errors that can be made when performing hypothesis testing

- If we reject H_0 when H_0 is true, we commit a **Type I** error. The probability of Type I error is denoted by: α.
- If we fail to reject H_0 when H_0 is false, we commit a **Type II** error. The probability of Type II error is denoted by: β.

One Sample t testing of a mean

Two Sample t testing of a mean

It's now time to attempt activity 2

Ordinary Least Squares (OLS) Regression

What is Ordinary Least Squares Regression

- Ordinary least squares (OLS) is a method for estimating the unknown parameters in a linear regression model
- It is one of the methods which we can assign a *line of best fit* to a data set

The linear regression model that would be fit to the data shown on the left, would be:

$$GDP(\%\Delta) = \beta_0 + \beta_1 \cdot UE(\%\Delta)$$

OLS would determine the β_0, and β_1 coefficients.

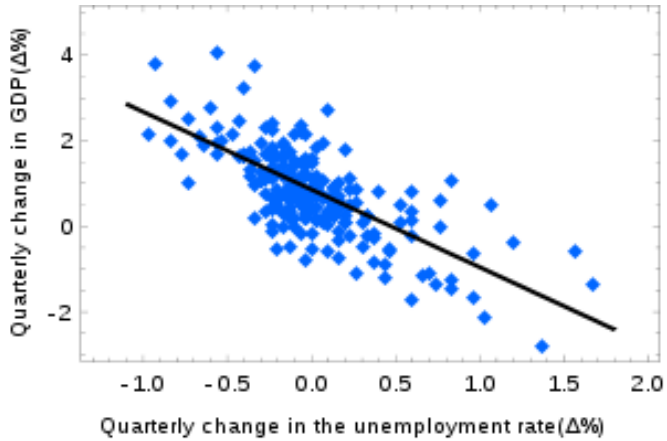


Figure : A scatter plot showing unemployment vs. GDP

What is Ordinary Least Squares Regression

- OLS is more than simply fitting a line to the data - it tests whether that relationship is statistically significant
- Consider the model $y \sim \beta_0 + \beta_1 \cdot x$
- When an OLS is run, it also performs the following statistical tests:

$$H_0 : \beta_0 = 0 \qquad H_0 : \beta_1 = 0$$
$$H_a : \beta_0 \neq 0 \qquad H_a : \beta_1 \neq 0$$

- p-values are reported for each of these tests

What is Ordinary Least Squares Regression

- **If the p-value is low** - it means the x variable is statistically significant in explaining the y variable
- **If the p-value is high** - it means that the x variable is not statistically significant in explaining the y variable

Ordinary Least Squares (OLS) Regression

Important Assumptions of OLS Regression

- Many are familiar with using OLS to fit straight lines (or non-linear transformations) to data sets in their high school, undergraduate, or postgraduate careers
- There are, however, some important assumptions that need to be satisfied if you want to use OLS to show the statistical significance of a relationship in your data
- This workshop will consider the two most important assumptions:
  - **Homoskedasticity assumption**
  - **No autocorrelation assumption**

Important Assumptions: Homoskedasticity

- The homoskedasticity assumption of the fitted model is that the standard deviations of the error terms are constant
- If this assumption is **violated** it is **not appropriate to use OLS** to provide statistical evidence of a relationship
- To test for homoskedasticity we look at the residual plot - we expect to see **no pattern in the residual plot**
- If there is no pattern in the residual plot, then we can assume that the homoskedasticity assumption has not been violated

Ordinary Least Squares (OLS) Regression

Important Assumptions: Homoskedasticity

Figure : The residual plot on the right looks like a random scatter of the residuals - there is no apparent pattern in the residuals

Important Assumptions: Autocorrelation

Definition: Autocorrelation

Autocorrelation, also known as serial correlation, is the correlation of a signal with a delayed copy of itself.

- In OLS there is an assumption that there will be **no autocorrelation present in the residuals**
- We can test for autocorrelation in the residuals once the model has been run
- One type of data where autocorrelation is almost ALWAYS present is time series data
- As a general rule, it is **not appropriate to use OLS regression with time series data**

How to perform OLS Regression Using MATLAB

1. Determine your model prior to fitting it to the data - it is bad science to look at the data first and then fit the model
2. Load your data into MATLAB
3. Fit your model using OLS
4. Check the residual plot to ensure that the homoskedasticty assumption is not violated
5. Perform a statistical test to ensure that there is no autocorrelation in the residuals
6. Read the summary report to determine if the β coefficients are statistically significant to your model

It's now time to attempt activity 3