Introductory Statistics with Excel

Andrew Parnell

University College Dublin

Class 1 - Basics

Introduction

In this course we will cover:

- Class 1: Graphical and tabular summaries of data
- Class 2: Basics of experimental design and probability distributions
- Class 3: Hypothesis testing
- Class 4: Confidence intervals and t-tests
- Class 5: Statistical modelling and ANOVA
- Class 6: Linear regression and control charts

There are also four Excel practicals.

I assume you have basic mathematical ability, some experience with using Excel, but very few formulae are used in these notes

Data sets

We will use three main data sets throughout this course:

1. The milk yield of 14 cows in kg
2. A genomic breeding value of cows based on a genetic marker expression level
3. A two-way unbalanced experiment on Clenbuterol based on 3 different doses over 3 different runs

As well as a number of small ones for illustration. All are included in the `course_data.xls` file

Data set 1

Questions:

What is the shape of the distribution of milk yield for cows?

Are there any outlying values?

Is the mean milk yield for these cows 120kg?

How certain can we be about the mean milk yield?

Data set 2

Questions:

Is there a relationship between the two variables?

If so, what is the shape of that relationship?

Is one variable causing the other to change?

Data set 3

Questions:

How important is the fortification?

Which is more variable, the runs or the replicates?

Turning data into information

In this class we will look at:

- The general characteristics of data
- Samples and populations
- Different types of data
- Summaries of location and scale
- Outliers and missing data
- Summarising data via graphics

Raw data

Often, we are first presented with **raw data**, e.g.:

	Expression	Breeding value
1	1.58	5.43
2	0.09	3.54
3	0.82	5.14
4	0.58	3.60
5	1.91	8.07
6	1.88	6.91
⋮	⋮	⋮

The raw data is where we always start. We need to turn it into useful information.

Observations and variables

- An **observation** is a set of answers on a single unit, such as a cow or a person. In the previous slide the expression and breeding value of a single cow constitute one observation
- A **variable** is a single characteristic of a set of observations, such as the breeding value

We traditionally arrange the data set so that the observations make up the rows of a table and the variables make up the columns.

The number of observations is referred to as the sample size, often represented by the letter n, e.g. $n = 120$.

Samples and populations

We will usually use **sample** data to make inferences about the larger **population** represented by the data. When data are collected from all the members of a population it is called a **census**. Most commonly, a census is too expensive to conduct.

The distinction between a sample and a population can be unclear. If we are only interested in the cows at one particular farm then we have just conducted a **census**. If we are interested in learning about a larger population (e.g. all cows, or all Irish cows) then we have a **sample**.

Parameters and statistics

A summary measure created from the raw sample data is called a **statistic**. A summary measure from an entire population is called a **parameter**.

Most often, we are interested in estimating the **parameters** from sample data. We try to create **statistics** which are good estimators of the parameters

Types of variable

There are generally three types of variable: **categorical**, **ordinal**, and **quantitative**. Some examples:

Variable	Possible Answers	Variable Type
Height in cm	Measured height in cm	Quantitative
Tenderness rating	1=Poor,..., 10=Excellent	Ordinal
Sex	Male or Female	Categorical
⋮	⋮	⋮

Not all numbers are quantitative variables (e.g. your telephone number)

Summarising categorial and ordinal data

Summarising categorical variables

Categorical and ordinal variables can be summarised in a **frequency** or **relative frequency** table. A frequency table gives just the counts of the data whilst the relative frequencies give the percentages.

Example: seat-belt use by 18 year olds

Response	Frequency (Count)	Relative Frequency (Percentage)
Always	1686	55.4%
Most times	578	19.0%
Sometimes	414	13.6%
Rarely	249	8.2%
Never	115	3.8%
Total	3042	100.0%

Summarising categorical variables 2

When there is more than one categorical/ordinal variable, we can use a cross-tabulation

Example: seat-belt use by 18 year olds

	Always	Most Times	Sometimes	Rarely	Never	Total
Female	915	276	167	84	25	1467
	(62.4%)	(18.8%)	(11.4%)	(5.7%)	(1.7%)	(100%)
Male	771	302	247	165	90	1575
	(49.0%)	(19.2%)	(15.7%)	(10.5%)	(5.7%)	(100%)

Often, the relative frequencies are more informative as they account for the sample size

Categorical/Ordinal variables can be summarised numerically by the **mode**; the category with the highest frequency. For males and females in this example the mode is the category 'Always'.

Summarising quantitative variables: location

The following are all way summarising the **central location** of the data:

Means	Computed by adding together all the values of a variable and dividing by the sample size.
Medians	Computed by placing the data in size order and finding the middle value
Quartiles	Computed by placing the data in size order and finding the values 25% and 75% of the way through

Whilst the mean, median and the mode are all forms of **average**, Excel uses the function `average` to compute only the mean.

Example: finding means and medians

Our 14 cows' milk yields were :
101.7 103.3 111.6 113.2 123.4 130.9 142 143.5 146.1 155.1 155.5 169.3 169.6 170.7
To find the mean we compute:

$$\begin{aligned}\text{mean} &= \frac{169.6 + 142 + 103.3 + 111.6 + 123.4 + 143.5 + 155.1 + 101.7 + 170.7 + 113.2 + 130.9 + 146.1 + 169.3 + 155.5}{14} \\ &= \frac{1935.9}{14} = 138.3 \text{ kg}\end{aligned}$$

For the median we do not have a middle value so we find the mean of the 7th and 8th values giving:

$$\text{median} = \frac{142 + 143.5}{2} = 142.75 \text{ kg}$$

What can we say about the difference between the mean and median here?

Example 2: CEO pay

The top 7 salaries of the highest paid CEOs in the US in 2009 was:
$556m (LJ Ellison; Oracle), $222m (RR Irani; Occidental Petroleum), $154m (JB Hess; Hess), $116m (MD Watford; Ultra Petroleum), $90m (MG Papa; EOG Resources), $87m (WR Berkley; WR Berkley), $68m (MK Rose; Burlington Santa)

The mean is now:

$$\text{mean} = \frac{556 + 222 + 154 + 116 + 90 + 87 + 68}{7} = \$184.7\text{m}$$

and the median:

$$\text{median} = \$116\text{m}$$

Why are the mean and the median different here?

Some notes about means and medians

- The mean is much more sensitive to **outliers**: extreme observations. It is easy for it do be dragged up and down
- The median is much more **robust**, however it does not have as many nice theoretical properties as the mean (see later classes on sampling distributions)
- Often raw data are given as **codes**, e.g. Male $= 0$, Female $= 1$, so that the mean of this variable represents the proportion of Females in the sample. You need to be careful that the codes are set correctly for this to work. For example, if Male $= 1$ and Female $= 2$, the mean is going to be meaningless!
- Modes are generally not calculated for quantitative variables until a **probability distribution** (see later classes) is assumed for the data

Summarising quantitative variables: scale

- The scale or spread of a quantitative variable is usually represented by the **variance/standard deviation**, the **range** or the **inter-quartile range (IQR)**
- The **range** is simply the maximum value minus the minimum
- The **IQR** is the difference between the upper quartile and the lower quartile
- The **variance** is the sum of squared deviations from the mean, divided by the sample size minus 1. The **standard deviation** is the square root of the variance
- As before, the **variance** is more affected by outliers than the **IQR**, but has nicer theoretical properties

Example: milk yields

- Going back to the cow data we have:
  `101.7 103.3 111.6 113.2 123.4 130.9 142 143.5 146.1 155.1 155.5 169.3 169.6 170.7`
- The range is $170.7 - 101.7 = 69$kg
- The quartiles are 115.75 and 155.40 so the IQR $= 155.40 - 115.75 = 39.6$kg
- The mean was 138.3 so the variance can be calculated as:

$$\text{(standard deviation)}^2 = \frac{(101.7 - 138.3)^2 + (103.3 - 138.3)^2 + \ldots + (170.7 - 138.3)^2}{14 - 1} = 604.18 \text{ (kg)}^2$$

- And the standard deviation as:

$$\sqrt{\text{variance}} = \sqrt{604.18} = 24.58 \text{ kg}$$

Some details about measures of scale

- Note that the range, the IQR, and the standard deviation are all in the **original** units (e.g. kilograms), whereas the variance is in **squared** units, so is often not as helpful to report
- In the variance calculation we divide by $n - 1$ so as to get an **unbiased** estimator of the population variance
- It's hard to interpret a standard deviation value by itself. It's often more useful to compare the standard deviations across different groups

Outliers

An **outlier** is an extreme or unusual observation. Common scenarios are:

- **It is a legitimate data value and represents the natural variability for the variable(s) measured.** In this case we should not remove it as it provides important information about the data
- **A mistake was made in taking the measurement.** In this case we should check with the experimenter before discarding or ignoring it
- **The observation actually belongs to a different population.** In this case it should either be discarded or the study re-run to include more samples from the wider population

Statistical graphics

Statistical graphics

Drawing graphs from data is one of the most **important** and **under-rated** parts of statistics

A well-structured and suitably labelled graph can summarise a variable (or set of variables) far more efficiently than any other format

Producing graphs: golden rules

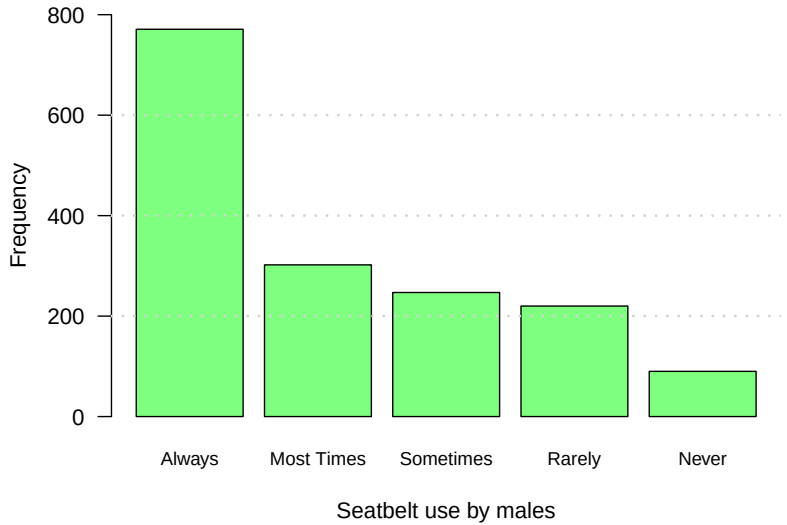
- A graph should be self-explanatory, so that the reader does not need any further information to understand what is going on
- There should always be clear labels on the vertical (y) axis and the horizontal (x) axis
- If necessary, include a caption beneath or above the plot that explains what the graph is showing
- Add extra touches, such as grid lines, colours, different point styles, etc, if they make the plot easier to understand.

Creating a good graph looks easy but requires a lot of thought and is actually very hard!

Bar charts and histograms

- Bar charts and histograms might look alike but they are very different.
- Bar charts are used on the frequencies of categorical or ordinal variables. Histograms are used on quantitative variables only
- Bar charts are very simple to create. They height of the bar represents the frequency or relative frequency of that category.
- Histograms are a bit more complicated:
  1. Decide on a number of cut-off points or bins
  2. Count the number of observations within each bin.
  3. The height of the bar is the count (or relative count) inside each bin.

Example bar chart



Bar charts 2

Some considerations:

- Bar charts can be horizontal as well as vertical, which can sometimes make them easier to read
- If using relative frequencies, it can sometimes be helpful to 'stack' the bars on top of each other so they total 100%
- Use grid-lines if necessary so that the reader can clearly see the differences between different bar heights

Creating a histogram

Some data on driving speeds (in kph):

61 64 66 70 71 71 74 75 76 82 83 86 86 87 89 89 92 92 92 92 92 93 93
94 94 96 97 98 98 99 99 99 100 102 102 104 104 104 105 105 106 107
107 108 108 109 110 111 112 113 113 114 116 116 117 118 119 119 120
120 122 124 125 125 126 128 129 130 130 130 131 132 133 136 136 142
142 143 146 146 148 149 150 150 151 151 152

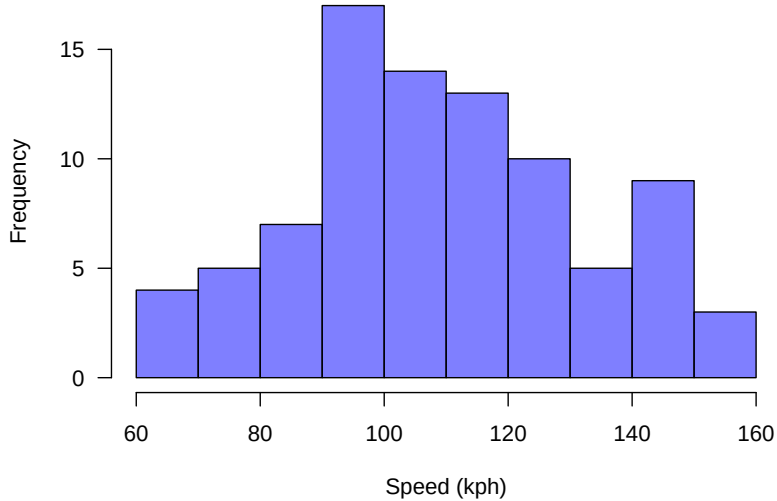
Suppose we set the bins at

60-70, 71-80, ..., 141-150, 151-160.

We would obtain frequencies of:

61-70	71-80	81-90	91-100	101-110	111-120	121-130	131-140	141-150	151-160
4	5	7	17	14	13	10	5	9	3

Example histogram

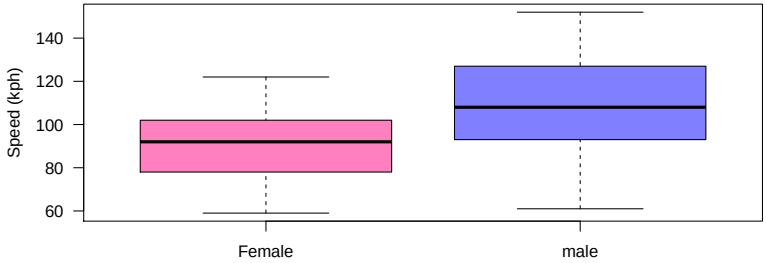


Considerations in drawing histograms

- Choosing the number/size of bins is very important. If you choose **too many narrow** bins the histogram will look very bumpy. If you choose **too few fat** bins then some of the interesting features will be lost
- Multiple histograms on top of each other can be useful in determining the differences between two groups
- From a histogram it can be easy to spot **skew**, whereby values are more spread out on one side of the graph than the other. If a graph is not skewed it is described as **symmetric**

Box plots

Box plots (or Box and whisker plots) are a very neat way of comparing multiple groups side-by-side.



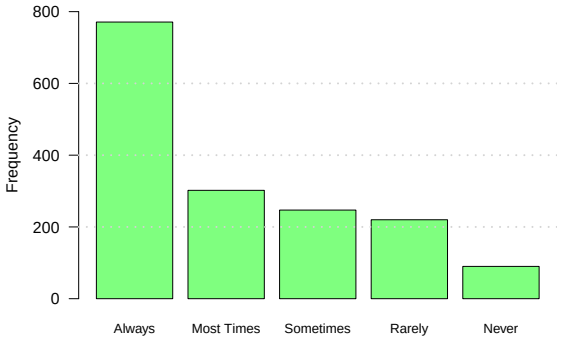
The central line is the median, whilst the edges of the box are the quartiles, and the whiskers are the extremes.

The collection of the median, quartiles and extremes is sometimes called the **five number summary**.

Pie charts

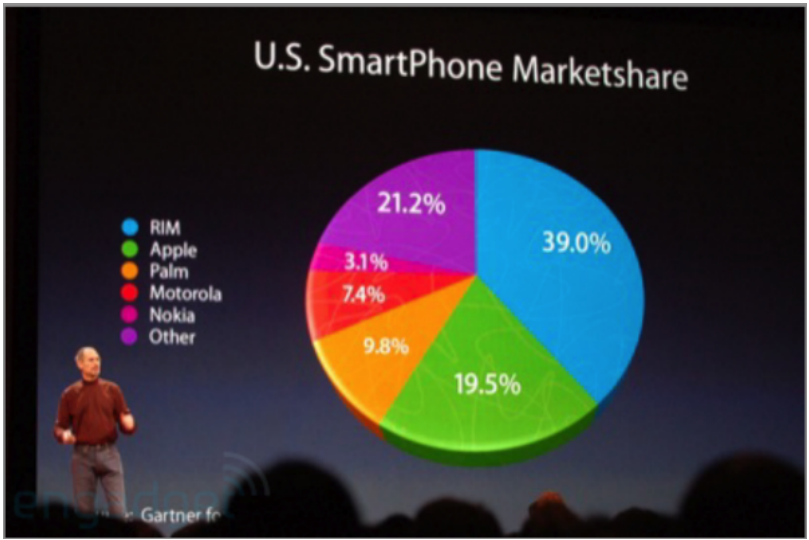
Pie charts a waste of time. Do not create pie charts.

Seatbelt use by males



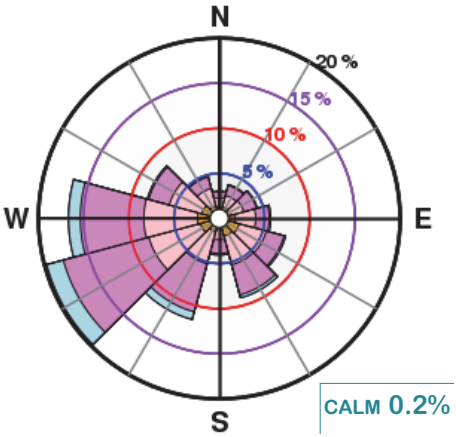
Seatbelt use by males

The classic


U.S. SmartPhone Marketshare
RIM
Apple
Palm
Motorola
Nokia
Other
21.2%
39.0%
3.1%
7.4%
9.8%
19.5%

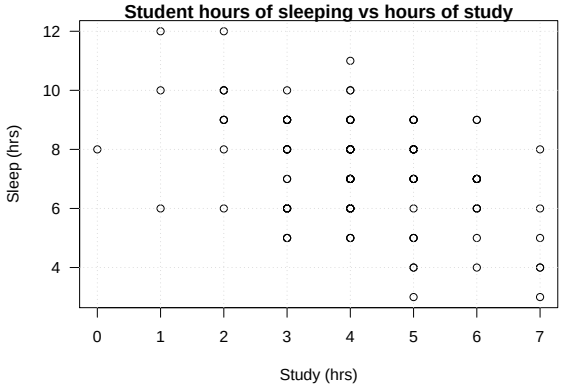
But...

Similar to a pie chart, yet a bit more useful, is a rose plot. The following represents wind speed at Dublin airport:



Scatter plots

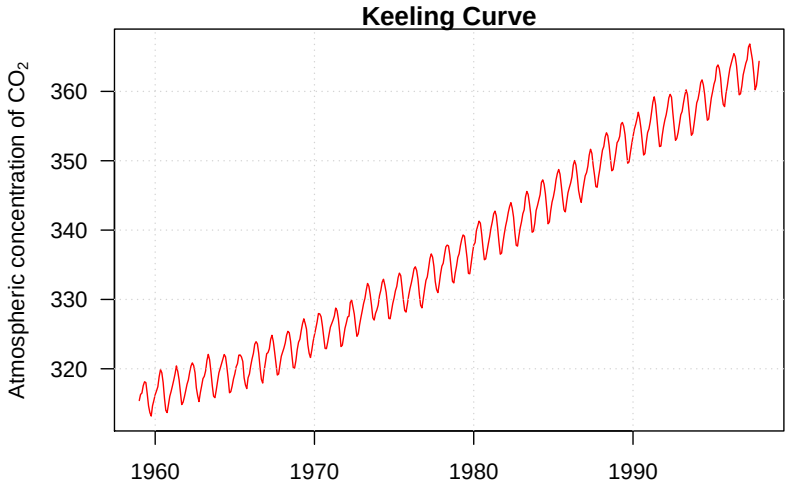
The neatest way to represent two quantitative variables simultaneously is a scatter plot.



Each observation is a point on the graph, e.g. hrs of study $= 6$, hrs of sleep $= 8$.

Line plots

If a scatter plot has a natural ordering in the x-axis, it is usual to join the points to create a line plot:



Class 1 summary

- Always think about **populations/samples**, and **parameters/statistics**
- Summarise data with measures of **location** (mean/median/mode) and **scale** (standard deviation, inter-quartile range)
- Use statistical graphics **carefully** and **thoughtfully** to present your data