Review

• Our objective: to make confident statements about a parameter (aspect) in the population based on a given sample.

• Most statistical methods are based on the assumption that the sample is only a very tiny proportion of the population. Indeed, it is generally assumed that the population is ‘infinite’ in size (this means if you were to count every individual in the population you would never be able stop).

• In certain cases, a sample may be a large proportion of the population. For example, we may want to understand the voting behaviour of residents in Hearn. To do this, 500 people were interviewed. 500 is a large proportion of the population of Hearn which is about 3000 people. In such cases, the statistical methods we will discuss in this class, are not valid.
Typically, for a statistical analysis to be valid, the sample size should be at most 5% of the population.

- All statistical inference that we do is based on the size of the sample only (not the size of the population).

- The size of a sample (the number in a sample) can influence the quality/reliability of the estimator.

- As the sample size increases the average becomes a ‘better’ estimator of the population mean. By better we do not mean it will be closer, simply that the chance of it being close to the true mean is larger. This is because the sample mean is random and can vary from sample to sample. This idea is an important notion, that will be used frequently in this class.
A representative sample

- When making a confidence statement (inference) about a population based on a sample we need to ensure that the sample is somehow representative of the data.

- For example, if we want to make a confidence statement about the mean height of students at A&M (the population is all students at A&M) based on a sample containing only women. It is likely that this sample will be biased.

- This sample is NOT a representative sample of students at A&M.

- Female students form a subpopulation of the population of all students. The sample is representative sample of female students, rather than the population of all students.
• A ‘representative sample’ has nothing to do with sample size.

• A simple random sample (SRS) is an example of a representative sample. This is where every individual in the population has an equal chance of being selected. No subpopulation is excluded.

  In a SRS there is always a chance that an individual will be selected more than once. But in practice, implementing an SRS is usually not feasible; people do not want to be interviewed twice. This is why the Hearn example given in slide 1, where 500 people out of 3000 are sampled, is not a SRS. Each person that is interviewed is “removed” from the population when selecting the next person to interview.

• If the population size is sufficient large as compared to the sample size, the chance an individual is sampled twice is extremely small.

• This is why we require that the sample size is at most 5% of the
population size.

- Designing an experiment in a good way is extremely important, but something we shall not cover in this course.

- In this course we will mainly assume that the sample is simple random sample.
Samples, Populations and Variables

- The population and sample are made up of *individuals* (these are not necessarily human), these can be people, companies, animals, a chemical etc.

- A variable is a characteristic in the individual that we are interested in. For example, for people it could be height, blood pressure, ethnicity or mother tongue.

  The characteristic of interest varies from individual to individual it is natural to call it a **variable**. We will learn later that since it is variable it is ‘random’.
Different variables in an M&M bag

- In a bag of M&Ms we may be interested in the main colour, number of M&Ms, weight of bag, type of M&M (chocolate or peanut) etc.

<table>
<thead>
<tr>
<th>bag no.</th>
<th>majority colour</th>
<th>number of M&amp;Ms</th>
<th>weight of bag</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>blue</td>
<td>18</td>
<td>2.2 ounces</td>
<td>chocolate</td>
</tr>
<tr>
<td>2</td>
<td>brown</td>
<td>19</td>
<td>2.3 ounces</td>
<td>chocolate</td>
</tr>
<tr>
<td>3</td>
<td>red</td>
<td>12</td>
<td>2.1 ounces</td>
<td>peanut</td>
</tr>
</tbody>
</table>

- **Types of Variables** From the above we can see that variables come in several different types:
  - Numerical continuous: eg. weight (2.2 ounces)
  - Numerical discrete: eg. the number of M&Ms in a bag (18)
- **Binary**: eg. Type (chocolate/peanut)
- **Categorical**: eg. Majority colour (blue/brown/red/green)

- Numerical variables always have a meaningful ordering. Beware of categorical variables disguised as a numerical variable. For example, the number of a bus is not a numerical variable but a categorical variable.

- In statistics we treat different types of variables in different ways.

- There are two types of Numerical variables, numerical discrete and numerical continuous. There is an interesting connection between these two variables. Numerical discrete variables “become” numerical continuous variables when averaged. For example, the number of children in a family is discrete but the average number of children in a family is continuous.
– Therefore in this course we will treat numerical continuous and numerical discrete variables in the much same way (with a few exceptions).
– In more advanced courses (such as STAT652), where more sophisticated models are used. Numerical discrete and numerical continuous variables will be treated differently.

• During the course we will consider different methods for treating different types of variables.
Examples of variables
What type of variables are the following:

• The gender of a randomly chosen person (we can use M/F or 0/1)?
• The number of a randomly selected bus?
• The make of bicycle of a randomly chosen person?
• The number of bicycles owned by a randomly chosen person?
• The height of person?
• Whether a random selected person responds to a drug?
• The predictions of Paul the octopus (win or lose).
**Statistical Analysis comes in three stages**

(1) Data description. When starting a data analysis first use a graphical method to represent the data (Chapter 3, Ott and Longnecker). I.e. histograms, pie charts, line graphs, line and whisker plots etc.

(2) Summary statistics, average (mean), median, variance, quantiles etc. This describes the data set (which can be large) in a few numbers, it also gives us an idea about the spread of the data.

(3) Quantitative techniques (this will be the main focus of the course, Chapter 3-11, Ott and Longnecker). We can evaluate an average, but what does this average tell us about of the true population average (usually called population mean)? How close is the sample average to the population average? We will be finding out a few weeks from now.
Visual Tools

- One picture is more effective than a thousand words.

Reasons to reuse and recycle plastics:

https://www.stat.tamu.edu/~suhasini/plastic.html

- When analyzing data find the plotting tool which best describes the data.

- http://www.stat.tamu.edu/~suhasini/teaching651/temp_faraday.dat
  This is the minimum and maximum monthly temperatures in the Antarctic Peninsula (from 1951 - 2004).

- To understand how the temperatures change over time, below is a time series plot of both the min and max temperatures.
What features do you observe?

- Is there seasonality?
- Is there a slow increase?
- What are the differences between the two plots?
- There are interesting plotting tools, such as time series plots, pie charts etc (see Ott and Longnecker, Chapter 3).
Lecture 2 (MWF)

Opening data in JMP

- Open JMP

- JMP > Preferences > Text Data Files > Import Setting check the **Use best guess** box. This displays the data in the correct way (recognizes spaces or commas as a new column etc.,)

- **Data on hard drive** Go to File > Open. Then you will see a Finder or File Manager. Select file and and press open.

- **Data on internet** Go to File > New (an empty spread sheet will pop up) > File > Internet Open...A window will pop up.

- Paste desired url in the pop-up window.

- You should see the data in a JMP spreadsheet.
The data in JMP

- The symbol on the left indicates how JMP reads each variable.

- The blue right angle triangle mean JMP reads the variables are continuous numerical.
• You can change the “type” of variable by clicking on the symbol/triangle. It can be changed to
  – ordinal (numerical discrete; data with an ordering such as ratings)
  – nominal variable (which is another name for categorical).

• Ensuring the variable is correctly is specified in JMP is important for using the appropriate statistical procedure.
Histograms: A tool for plotting the distribution of numerical variables

- **Range** The smallest interval which contains all the data. For example, the range of 22, 23, 39, 37, 31, 24, 24, 26, 27, 41 is the interval [22, 41].

- **Bin** Partitioning of the Range (usually into equal parts). **Frequency** Number/or percentage of data lying in each bin.

- The histogram is a plot device for checking the frequency of observations in a certain interval. It does this by partitioning the range and plotting frequency of each bin as a height on the graph. The size of each bin is called the **binwidth**.
• Relative frequency histogram is exactly the same as the histogram. The only difference is that height is not the number in the bin but the proportion in the bin. It is usually more useful than a standard histogram since proportions convey more information than the raw numbers.
Lecture 2 (MWF)

Plotting a Histogram

• Example using binwidth 4.

<table>
<thead>
<tr>
<th>interval</th>
<th>[20-24]</th>
<th>[25-29]</th>
<th>[30-34]</th>
<th>[35-39]</th>
<th>[40-44]</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>percent/relative frequency</td>
<td>40%</td>
<td>20%</td>
<td>10%</td>
<td>20%</td>
<td>10%</td>
</tr>
</tbody>
</table>

• The Histogram (in Statcrunch and JMP):
Plotting a Histogram in JMP

- Once data is loaded into JMP.

- **Analyze > Distribution.** A window will pop up will variable. Highlight and (double) click on variable you want to plot. Press OK.

- You can adjust the histogram by selecting red arrow next to the variable and going to Histogram option.

- **To get counts on the y-axis choose Count Axis.** To get proportions/relative frequency (percentage of data on the y-axis) choose Prob Axis. You can change the bin width by selecting Set Bin Width.

- If you click on a block in the histogram, it will be highlighted as a stripped block. The data which contributes to that block will be highlighted on the corresponding spreadsheet.
• To make comparisons between *different* subgroups, highlight the factor variable click on By.
Features in a histogram?

• Outcomes that are most frequent.

• If the sample is a composition of multiple populations (more of this later), these can be seen with multiple modes in the histogram.

• The spread of the data, the main concentration of the data.

• Warning: The histogram and your perception of how the data is distributed depends heavily on the bin width you use. In practice, it is useful to plot several histograms with different bin widths, and compare the plots.
The distribution of M&Ms in a bag

Observe how different bin widths can change your perception of the data.
Using a histogram to compare populations

- A histogram is a very useful tool for comparing samples and seeing whether they come from the same or from different populations. We will learn more quantitative methods of comparison later in the course. What we do now is simply a visual comparison.

- Example

We could expect the temperatures in January in the Antarctic to be more than those in May in the Antarctic (recall that in the Antarctic, January is summer and May is winter) not that all temperatures are in Celsius.

Below are the histograms of data taken in January and a sample taken from May (maximum temp). What do you think?
Comparing temperatures in the Antarctic
• The top plot the summer temperatures and the lower plot are the winter temperatures in the Antarctic between 1951-2005. What do you notice?

• We see that the histograms appear to be a shift of each other.

• How to quantify the main features and the differences?

• There are several ways to do this. One way is to consider a numerical value which describes a feature in the data, and to compare the numerical values from each sample.
  
  – From the point of view of statistical inference, it is much easier to compare numerical values than graphs.
  
  – One measure of center is the average (sample means) of the sample.
Summary of the Histogram

- For discrete variables the relative frequency histogram is an appropriate way to represent the frequency/distribution of a sample or population.
  - The y-axis gives the proportion of the population who live in 2-person households, 3-person households etc. The relatively frequency histogram of the population contains all the information on the distribution of the population.

- However, making a relatively frequency histogram for continuous data is problematic. No relative frequency histogram will contain all the information about the population.

- By construction, the relatively frequency histogram is restricted by the chosen binwidth.
• If the binwidth of the plot is two, you cannot obtain the proportions for bins less than two.
  
  – Consider the heights of all people in the world. Suppose you make a relatively frequency histogram using the bin-width of 0.01m (1cm). This will convey a lot of information about the data - but not all. It will not tell you the proportion of people between 1.60-1.605 meters tall (this binwidth is narrower than 0.01m).
  
  – To get over this problem (and other mathematical issues), we consider a closely related cousin of the relative frequency histogram called the density plot. For numerical continuous variables no relative frequency histogram will convey all the information about the frequency of the variable. However, the density plot will.
The density plot

- The density plot is a little different to the histogram, in the sense that the area under the graph represents the frequency of an event and not the height.

- Look at [https://www.stat.tamu.edu/~suhasini/teaching651/density_functions.pdf](https://www.stat.tamu.edu/~suhasini/teaching651/density_functions.pdf), which gives information about the density.

- To plot the distribution of the population of numerical continuous variables we will always use the density plots.

- The area under the curve is used to calculate probabilities. But the height of the plot will help understand which outcomes are most likely to occur.
The Shape of a distribution/density

The shape of a density gives important information about the population.

• Variables whose distributions tend to be close to symmetric:
  – Heights of a certain gender.
  – Length of bird bills and other biological lengths.

• Variables whose distributions tend to be skewed:
  – Price of houses.
  – Gestation period of a baby.

• Variables whose distributions tend to be multi-modal (have several distinct peaks).
  – The height of adult humans (both sexes),
– Number of M&Ms in a bag (all types, Peanut/Milk chocolate/Peanut butter).

Multi-modal densities suggest that it is a mix of subpopulations.

• Variables whose distribution tends to be flat (uniformly distributed):
  – The numbers in a lottery.

• Each numerical continuous variable will have its own density plot, with its own features.

• In general, the distribution will not be bell shaped. Skewed distributions very common.