**Presentation Briefing**

The aim of these presentations is to give you a chance to explore an area of probability or statistics that is not formally covered in the module. The presentation enables other students to learn from your research. Being able to clearly explain statistical concepts is an important skill that you will need later in your degree or in the workplace.

Each presentation will be 20 minutes in length plus extra time for questions at the end. Your target audience is the rest of the class. The presentation should probably cover the theory, simple example(s) which are easily understood, its implementation in R, and discuss a real world application (perhaps from your other subject area). You may use Powerpoint or other software to help illustrate the concepts. Each student will research a different topic. Topics will be allocated on a first-come first-served basis. Presentations will take place in Week 13.

**Possible Topics**

**Survey sampling.** An important way of collecting data is through surveys. These can be conducted face-to-face, by telephone, online etc. For statistics it is important that those surveyed are representative of the wider population of interest. For this topic you should research methods for constructing a representative survey, by considering who to survey, how to survey, and how to correct for any inherent biases in the sampling process. To get you started, this webpage lists some sampling methods: <https://explorable.com/methods-of-survey-sampling>

**Missing data.** In textbook examples we always have complete data sets - i.e. none of the observations are missing. In real life, however, there are nearly always missing observations. How we handle missing data depends on why the data is missing and what we are trying to find out. For this topic you should research ways to handle the missing data and recommend methods for different situations. You could start your research by reading this: <https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4>

**Density estimation.** We assume that our sample data comes from an unobserved underlying probability density function, which we assume is the distribution of the population. Using the sample mean to estimate the population mean is good but it only gets us one number, whereas density estimation is using the sample to estimate the whole underlying pdf. There are lots of different methods, but the non-parametric methods mentioned here are a good starting point: <https://machinelearningmastery.com/probability-density-estimation/>

**Maximum likelihood.** There are different ways of estimating the parameters of a distribution. For example, the method of moments suggests that the sample mean is a good estimator of the population mean. A more general approach is to consider the likelihood function and find the parameter values that maximise the likelihood. For this topic you should explain what the likelihood function is, how you maximise it, and compare the benefits of ML to other methods. See, for a starting point, <http://mathworld.wolfram.com/MaximumLikelihood.html>

**Non-parametric testing.** Non-parametric statistics involves analysing data without making any assumption that the data comes from the usual parameterised distributions. Often these methods work by considering only each observation's rank, or by permuting the labels to see how this affects a particular statistic. For this topic you should consider a selection of non-parametric hypothesis tests, give an intuitive explanation and illustrate their use. Some common tests are mentioned here: <https://www.statisticshowto.datasciencecentral.com/parametric-and-non-parametric-data/>

**Bootstrap.** Most of the time we make some assumption regarding the distribution when calculating confidence intervals or performing hypothesis tests. To avoid making this assumption we can keep resampling the same observations to see how much a particular statistic, such as the sample mean, would change if the original sample had been slightly different. For this topic you should explain the bootstrapping approach and how it can be applied to confidence intervals and hypothesis tests.

See: <https://towardsdatascience.com/an-introduction-to-the-bootstrap-method-58bcb51b4d60>

**Bayesian statistics.** This is a huge topic area but one that students should be aware of. In classical statistics we let the data alone do the talking. In Bayesian statistics we take our prior knowledge and then use data to update our knowledge. Some argue this is what humans naturally do when evaluating the world around us, while others argue that we should remove our prior beliefs from the scientific approach. For this topic you should explain the Bayesian approach and illustrate its use. See <https://towardsdatascience.com/bayesian-statistics-for-data-science-45397ec79c94>

**Cluster analysis.** Sometimes data comes from two or more groups. For example, we may have two or more species and we wish to analyse how the species differ. For clustering, however, we only have measurements and don't know which group an observation comes from. The question is, do the data naturally fall into different clusters, which may or may not correspond to groups of interest. For this topic you should research different methods for conducting cluster analysis. Some common methods are mentioned here: <http://www.statsoft.com/Textbook/Cluster-Analysis>

**Statistical learning.** Suppose we have data on specimens from two different species. If we then collect more observations, can we work out which group this new observation came from. Statistical learning is about establishing rules which optimally classify new observations correctly. For this topic you should research several statistical learning approaches, give an intuitive justification and illustrate their use. See the discriminant and tree based methods mentioned in the introductory book found here: <http://faculty.marshall.usc.edu/gareth-james/>

**Time series.** Many statistical methods assume the observations are independent of each other. However, observations collected through time are not independent. For example, today's share price depends on yesterday's share price. A family of statistical models has been developed to handle time-dependent data. For this topic you should research such models, discuss how to estimate the parameters and outline the uncertainties of future predictions from past data. The early sections of this webpage should be useful: <http://www.statsoft.com/textbook/time-series-analysis>