# Statistics M2 Final Project

**Proposal deadline:** November 13, 2017

**Project deadline:** December 15, 2017: a clear report written in English or French, with appropriate data visualizations, prepared in RMarkdown

In your final project, you need to describe a data analysis problem, give some example data, and demonstrate a dependent measure and an analysis that demonstrably works to answer the question posed in the problem.

A **data analysis problem** has two parts: data—some kind of measurements or observations that are subject to variability even when they come from the same source—and a problem, some kind of information or answer that you'd like to extract from this data.

*Artificial example.* The data is a set of test results from computer chips in my computer chip factory. I would like to find out why some of them are defective. I am going to evaluate whether chips with at least one anomalous test result tend to come from any one particular computer chip machine.

*Less artificial example.* The data is from an infant conditioned head-turn study, in which 10-month-olds were trained to turn their heads toward a speaker when they heard a change in stimulus, and experimenters coded whether and when they turned their heads. The study played sequences of syllables and every so often switched one of the phonemes. For some babies, the phonemic change chosen was a contrast that exists in their native language, and for others it was not. I am going to evaluate whether 10-month-olds react more consistently to switches in native-language phonemes than to switches in non-native phonemes.

You are neither expected nor allowed to make use of the complete data set that you would need in order to answer the question. Rather, you are to make use of **example data.** This could come from several possible sources. A good source would be an existing study done on the same type of measurement but for some entirely different study (for example, computer chip diagnostics from a different factory, or infant conditioned head-turn data from a different experiment with different stimuli and a different question). Another source would be a pilot study or, failing this, a small subset of data already collected for a full study. Or you could simply invent your data by sampling from the built-in random sampling functions in R, but I would prefer you not do this. (The fact that you created your data from scratch will make it hard for you to pretend that you're not sure what the best way to analyse it is.)

You will then justify a choice of **dependent measure.** This might simply be a more thorough explanation of how you arrive at a single, pre-selected dependent measure (such as your binary variable "at least one anomalous test result," or whatever measure of "consistency" you wish to apply to infants' head-turning behaviour after switches in the stream of stimuli). You would then give an a priori justification of why you think this measure is better than several possible alternatives, or, at least, why it is not worse, along with some possible concerns you have about it. Critically, you must explain why you believe that this measure provides information about the question you'd like to answer. You may instead investigate, and try and choose between, more than one dependent measure, which you should do empirically. You'll need to find a way of using your sample data to assess which of the measures is best at resolving questions of the kind you intend to pose.

You will need to describe the proposed study in detail, and then use your example data to develop **an analysis that demonstrably works.** It's the demonstration more than the analysis that is critical here. Obviously, you need to choose an analysis that makes sense. So, for a simple comparison of two groups, you might propose some reasonable test statistic and say you'll do a permutation test. Your task is then to construct various possible outcomes by manipulating your example data, covering a range of possible answers to your question, from the positive to the negative to the downright weird. You should also try and manipulate parameters of the study, such as overall size, balance between groups, and sources of non-independence between observations. Obviously, you do not need to cover every imaginable possibility, but you need to cover enough to allow you to conclude that your analysis will tend to give correct answers. Be sure to include in your description of the analysis anything you have to say about how and whether you will exclude certain data points, and what your criteria will be—these criteria need to be scripted into your analysis and evaluated on all the artificial data sets.

**For your proposal, due on November 13th.**

You need to write a short (perhaps one-page) plan for what you are going to do, that specifies the question, the data, and any ideas you have about what the analysis might be. So far, you know how to do some simple analyses for comparing two groups and (after Assignment 3) for testing non-independence across groups. Between now and the end of the class, we will develop a few more analyses for standard problems, including the analysis of factorial designs (two or more crossed group variables), and of cases where you have continuous or ordered predictors. There may, thus, be things that you feel you don't know how to do yet. Please try and be clear about what you want to do, so that, if you have doubts, we have time to talk about the appropriate techniques in class or one on one as soon as possible. Do not feel obliged to do a complicated statistical analysis. I suggest you propose the simplest thing that you think could work.

We will also talk about some simple techniques for manipulating data to simulate alternate outcomes. You will see one that works in very simple cases in Assignment 3, and you will get a chance to start practising more of them as of Assignment 4, which comes out on November 20th. In your proposal, suggest various types of strategies you would like to try, and we can talk about how you might execute them in the meantime. This will also give me a better idea of what you need to see in class.

**START NOW.**

There are a lot of details you will need to take care of. Notably, do not overlook the problem of how to read your example data into R. You have probably never done this, and we have not talked about it, and we are not going to talk about it because we don't have time. But, unless you are using completely artificial example data, you will need to somehow load your data from a file or files into a table in R, and you will need to structure that table in a reasonable way. It is very likely that whatever data you or someone else may have is saved in some kind of a file that will be hard to read in R, and in some kind of an arrangemenet that is not at all like the kind of table you're going to wnat to use. So this As soon as you decide on your data, think about the way in which it's currently stored, and the arrangement of the table in which you'd like to have it in order to be able to analyze it. You are strongly encouraged to give details about this in your proposal.