# BCA Research Project, Semester 1 2018: Supervision meeting 1

1 March 2018

Attendees: Elasma Milanzi, Enes Makalic, Katrina Scurrah and Carl Higgs (supervised, minutes)

Enes and Carl met in late January, however this was the first full supervision meeting for this project. Elasma was also in attendance, and may be able to provide statistical supervision support as well as advice on project scope.

## Project scope

Katrina clarified the broad intended scope of the project which may consider more than just Pearson correlation, but also:

* Other kinds of correlation coefficients (eg. Spearman, others?)
  + So review will aim to cover ‘what kinds of correlation may be of interest in the context of a classical twin study?’ For example, the Spearman or Kendall non-parametric rank correlation statistics
* correlation in the context of mixed models
* role of Mz/Dz ratio: this has important implications in variance component modelling, but what impacts can imbalance in sample size or other aspects have on power to detect a difference in correlations between groups?
* Assumptions and robustness to violation of these

## Meetings

It was clarified that Thursday is probably the best day for all, and probably in the afternoon. These will be held weekly, in general, commencing 15 March.

As such a proposed schedule (saved in accompanying Excel file in case you want to copy dates) is:



This probably already needs refinement – literature review may require more time!

## Background context

Carl met with Enes on 22 January to discuss the project and subsequently sent Enes and Katrina an e-mail outlining his initial understanding of the project; pasted here for the record (you’ve read this – feel free to ignore!):

**Background**

This is a one semester research project for the Master of Biostatistics degree, with the brief

The “classical twin design” aims to estimate components of variation due to shared genetic effects, shared environmental effects, and unshared effects using data from identical and non-identical twins.  A recent publication described a method of calculating the power to detect each of these variance components under certain assumptions.  However, this method did not address power to detect differences in correlations between identical and non-identical twins, which is an important first step in fitting variance components models. 

**Research question(s):**This project will develop methods for estimating power for this first step, using both theory and simulations.

**Current understanding**

Discussing this with Enes today, my current understanding is that the problem of estimating power to detect a difference between two correlations isn't necessarily new, or even specific to twin studies (although it will be great to here your thoughts on this Katrina as I have only shallow knowledge of this area!).  However, evaluation of existing methods in addition to highlighting relative merits may open opportunities for improvements in methodology.

* A common approach uses Fisher's *z*-transformation (ie. inverse hyperbolic tangent) to map the two correlation coefficients from (-1,1) to (-inf,inf) with approximately normal distribution to allow for asymptoptically valid comparisons.  The Stata command 'power twocorrelations' uses this approach
* Alternate approaches to approximating the sampling distribution of rho apparently exist
  + an interesting one might involve bootstrapping.  For example (links not necessarily dealing with differences yet, but may be generalisable; this also may be useful when we move to the simulation study)
    - [Bootstrap power of the generalized correlation coefficient](https://link.springer.com/article/10.1007/BF00162525)
    - [Size and Power Properties of Some Test Statistics for Testing the Population Correlation Coefficient ρ](https://www.omicsonline.org/open-access/size-and-power-properties-of-some-test-statistics-for-testing-thepopulation-correlation-coefficient-2155-6180-1000353.php?aid=91187)
    - Efron and Tibshirani (1994) discuss parametric bootstrapping of corr(x) and comparison w/ Fisher's *z* transformation (p55) -- i haven't looked into this in any detail yet as to whether this might be useful for our purposes
    - an article re calculating confidence intervals for correlation coefficients in Stata has a good review of history of Fisher's transformation and its geometric interpretation, as well as application of bootstrapping for this purpose; might be some parallels with the task of calculating power -- Cox NJ. Speaking Stata: Correlation with confidence, or Fisher's z revisited. Stata J. 2008; 8 (3) : 413-39.
* Enes has an idea - I haven't asked for details about this yet though, so I can approach lit review from blank slate!  Although I am keen to hear more once I am more familiar with area.

I haven't commenced the literature review yet of course; the above are just initial impressions.

**Tasks include**

* Literature review
* Comparison of methods, possibly including novel approach(es)
* Simulation study to compare efficency / accuracy / performance of these
* Possible extras:
  + web app for online power calculator
  + collaboration on paper relating performance of proposed novel approach (later)
* Assessment:
  + Presentation around May
  + Submit report around June

**Plan**:  At this stage, perhaps

* if you could please send me any literature (e.g the 'recent paper' mentioned in project brief) you would recommend as a starting point for my review
* Meet up in about a month w/ Katrina & Enes to:
  1. discuss particularities of this question to variance component models in twin studies
  2. share my proposed timeline for the semester
  3. share any progress with my literature review, thus far

**Availabilities for meetings**

* **Would 3pm on Friday 2 March in Enes' office work for our next meeting?**I would suggest Monday of that week, but I'm arriving back from Tasmania that day.
* In general, I work Tues - Thursday, so would ideally meet on a Monday or Friday.

**My background fyi**

* I have an MPH (epi/biostats focus; finished 2015 at UoM)
* I have completed subjects in genetic epidemiology and bioinformatics; however my understanding in this very large area of statistics is more broad than deep
* I work in a kind of 'data scientist' like role in the spatial team in the [Healthy Liveable Cities group](http://cur.org.au/research-programs/healthy-liveable-cities-group/)in Centre for Urban Research at RMIT; work is in general more computational (data management, GIS) than statistical.  The broad area of research is relationship between built environment and health/wellbeing through social determinants of health lens and with policy-relevant focus.  We use multilevel models and make maps.
* I work with Stata, R,  Python and SQL (and have some familiarity with Javascript and Java)
* I also work with Koen Simons, a colleague of yours in CEB; incidentally, he was my supervisor on my research project last year.
* Enes floated the possibility of doing the computation for this project in MATLAB; the opportunity to formally learn to use this is interesting to me.  I'll keep this option open while I conduct the literature review.  I have used it briefly previously, so I might reinstall and see if anything comes back!