The effect of statistical model selection on subsequent infrerence
Timur Hassan, Department of Mathematics & Statistics, La Trobe University

My vacation scholarship project was concerned with the effect of statistical model selection on subsequent inference. This topic addresses the issue of naive confidence intervals in factorial design experiments having a significantly lower minimum coverage than the claimed nominal coverage.

I explored the effect of preliminary hypothesis tests, namely, what happens when we set up hypotheses as follows:

$$H_0: \theta=0 \quad vs \quad H_a: \theta\neq0.$$

It is a common practise that the statistician will test these hypotheses and accept the result as truth. Suppose that we accept H_0 and set θ to 0. We then set up a confidence interval for θ. This confidence interval will have random endpoints and will have width such that our estimate (a fixed value) will fall within this confidence interval with a prescribed probability, say 0.95. If the true θ is different from 0, then the probability that the confidence interval covers θ drops from 0.95, sometimes to extremely low values, depending on a number of factors.

I used Monte Carlo simulation techniques implemented in R to prepare plots of estimated coverage probability against θ. These plots showed that the coverage probability in my 2^3 factorial design model dropped to about 0.89 where the nominal coverage was 0.95.

I thoroughly enjoyed this topic and would be interested in pursuing it if resources were available. The Big Day In event in Sydney was a blast and I would recommend it to anybody that is given the opportunity to participate in the following years. The CSIRO speakers were very interesting and at the same time hilarious. The environment was very relaxed and every seminar that I sat in was always interesting and often entertaining.

I am very grateful to my supervisor Paul Kabaila, CSIRO and AMSI.