Session 3

Sample Size Estimation

Matt Denwood

2023-07-13

An introduction to day 2

Building on day 1

Yesterday we looked at simulating data - we will continue that theme today!

BUT: my code looks a bit different to Giles's code, e.g.:

```
# If necessary:
## install.packages(c("tidyverse","pbapply"))
library("tidyverse")
library("pbapply")
```

REMEMBER: the coding style is not important as long as the output is the same

Background to sample size calculations

Power calculation

Power is defined as the proportion of experiments that can be expected to give p-values of $<= 0.05$ (or whatever alpha is chosen), conditional on the specified parameters. Power calculations can be done using:

- Approximation methods, e.g. power.t.test:

```
power.t.test(n = 150, delta = 0.25, sd = 1)
##
##      Two-sample t test power calculation
##
##               n = 150
##           delta = 0.25
##              sd = 1
##       sig.level = 0.05
##           power = 0.5785239
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Sample size estimation

The goal is typically to find the minimum sample size that corresponds to >= 80% power, for a specified set of parameters. This can be done in one of two ways:

- Using approximation methods directly i.e.:

```
power.t.test(n = NULL, delta = 0.25, sd = 1, power = 0.8)
##
##      Two-sample t test power calculation
##
##               n = 252.1281
##           delta = 0.25
##              sd = 1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

- By trying different sample sizes (using either approximation methods or simulation):

Sample size calculation for LCM

Determining the objective

Let's take a simple 2-test, 2-population Hui-Walter model as an example.

Group discussion:

- What parameters do we need to simulate a dataset? Which of these are experimental/controllable parameters, and which are nuisance parameters?
- What might we be interested in estimating from the model?
- How can we maximise the efficiency of fitting the model to each dataset we simulate?

Exercise

Write a function to:

1. Take input parameters in two arguments: controllable (2xN), and nuisance (2xSe, 2xSp, 2xPrev)

Obtaining an answer

The answer depends on the objective ... and there are many things that might be the objective, including:

- Width of 95% CI for sensitivity for one or both tests
- Width of 95% CI for specificity for one or both tests
- Width of 95% CI for prevalence in one or both populations
- Something more complex, like proving one test has a higher Se/Sp than the other (maybe using Bayesian p-values)
- Several / all of the above

Group discussion:

- How would we expect these things to vary depending on:

Additional considerations

Discussion:

- How should we deal with uncertainty in parameter values? Integrate over them!
- How best to deal with multiple dimensions of N (i.e. total samples and distribution of samples)?
- What about more complex scenarios e.g. 3 tests, including covariance?

Further reading

If you are interested in making this more complicated (!), you can read through some related work here:
https://www.costmodds.org/projects/covetlabLCM/sample_size_calculation.html

Remember the bonus session 4!