# FMS1203S: Randomness in scientific thinking

Week 2

Problem with sampling

# Introduction

- Problems of sampling arise when you'd like to know a characteristic of some population of interest but it isn't feasible to measure every individual.
- Example: estimate the average height of NUS students.
  - One approach - measure the height of every student (expensive, time consuming).
  - Another approach - sample a small subset of students.
- It's important that the sample be *random*.
- For a *random* sample, provided the sample size is not too small, the population characteristic can be pretty well estimated.
- One of the most difficult parts of sampling is to obtain a truly random sample representative of the population of interest.

# An Example of biased sampling

In a survey a random sample of people is taken and asked for a list of friends, the average number of friends of sample is computed.

Another random sample is taken from the friend lists of the first sample and also asked for a list of friends and the average number of friends in this sample is also computed.

The people sampled at the second stage have, on average, many more friends than do the people in the original sample.

This suggests that, on average, your friends are more popular than you are. Is this correct?

# Sampling techniques

Sampling is usually done in the following steps:

- Define the population of interest. In the example above of student heights, this is straightforward - it is the collection of all NUS students.
- Define a *sampling frame*, the list to use to reach the members of the population. In our example, suppose we had a list of contact phone numbers for students. If we sample students by choosing randomly from the list of phone numbers then the phone number list defines the sampling frame.
- Specify a sampling method.
- What sample size should be used?
- Implement the sampling program and collect the data.
- Identify and address any problems that arose (for example, what is the effect of refusal by some of the people you sampled to participate in your survey or study?)

# Homework Exercise: estimating your vocabulary

- You are given a dictionary that is sufficiently large that it contains every word that you know. One way to estimate your vocabulary is as follows:
  - Suppose there are $N$ words in the dictionary. Randomly sample $n$ of them and see how many of these words that you know. Call this number $x$.
  - An estimate of the proportion of words you know in the dictionary is $x/n$.
  - Therefore an estimate of the number of words you know is $N \times x/n$.
  - Of course, this estimate will vary from sample to sample ...

# Estimating your vocabulary

Example: suppose you sample 100 words at random from a dictionary with 100,000 words.

You know 50 of the 100 words (one half).

So we estimate that you know about half the 100,000 words in the dictionary, so your vocabulary size is 50,000.

# Some things to think about

- How would you work out the number of words in the dictionary if you didn’t want to count every word?
- What happens if the dictionary doesn’t contain every word that you know?
- We need to choose words in our dictionary so that each word has the same chance of being chosen. How do you do this?
- How would you decide on how many words to use in your sample?

# Readings for next week

Group one: Milner, A. and Calel, R. (2012). Are first-borns more likely to attend Harvard? *Significance*, 9, pp. 37–39.

Group two: Moon, N. (2010). Curtains at Number Ten: predicting the general election. *Significance*, 7, pp. 24–26.

Group three: Goldacre, B. (2007). When the facts get in the way of a good story. *Significance*, 4(2), 84–85.

Group four: Milgram, S., Mann, L. and Harter, S. (1965). The Lost-letter Technique: A Tool of Social Research. *The Public Opinion Quarterly*, 29(3), 437–438.

Group five: Baker, Stephen (2009). They've Got Your Number: Data, Digits and Destiny - how the Numerati are changing our Lives. (Chap 7, Lover)