Tutorial: Non-response bias in surveys

Non-response is a huge issue in many surveys (Groves and Peytcheva, 2008). Survey non-response leads to significant bias if response is correlated with the survey indicators of interest. We use a simple example from the Framingham study to illustrate this concept.

Source: Groves, R.M. and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias. *Public opinion quarterly*, 72(2): 167-89.
(I found a free draft via Google.)

Example:

- Suppose blood samples from the participants at baseline got lost; rather than measure everyone in the population again, the study investigators decided to try to estimate the baseline prevalence of high cholesterol (cholesterol $>$ 240 mg/dL). They randomly sampled 400 individuals and asked them to return to the study center to have their cholesterol measured again, knowing that not all 400 would return for the re-test.
- The willingness of a participant to revisit the lab was correlated with the frailty of the individual, sex, and prior knowledge of high cholesterol. With a lot of missing data, we would expect to obtain biased of high cholesterol prevalence.
- Prevalence of high cholesterol at baseline was 43.1% in the Framingham cohort.

We consider three different scenarios:

A. Low response rate

B. Moderate reponse rate

C. High response rate

Exercise: Calculate the prevalence of high cholesterol for each of the three response rate settings, as well as for the complete sample of 400 individuals.

```
proportion highchol
proportion highcholA highcholB highcholC
proportion highcholA
proportion highcholB
proportion highcholC
```

As suspected, bias increases with the amount of missingness.

We have baseline covariate data from the Framingham study. We can estimate the probability that a sampled individual returns to have his/her cholesterol tested again as a function of these covariates.

If we knew these probabilities *exactly*, we could obtain an unbiased estimate of high cholesterol prevalence at baseline. In this example, we do have these probabilities (`pA`, `pB`, and `pC`

in the dataset).

Exercise: Calculate the prevalence of high cholesterol for each of the three response rate settings using the survey weights, and compare to the complete-case data.

```
gen wA = 1/pA
gen wB = 1/pB
gen wC = 1/pC

proportion highchol
proportion highcholA [pweight=wA]
proportion highcholB [pweight=wB]
proportion highcholC [pweight=wC]
```

Here we recovered unbiased estimates. However in practice, we will never exactly know `pA`, `pB`, and `pC`. Many methods have been developed to address survey non-response, including multiple imputation and weighting for non-response. Maximizing the response rate is always the best policy.