Simple Random Sampling

A central problem in the social sciences is that we can't collect measurements about all units in a population we want to study, because typically it is too expensive to do so. One common research strategy for estimating a characteristic of all units in a population is to conduct simple random sampling, measure the characteristic among that sample, and rely on a statistic from those measurements as an estimate of characteristics in the population. In this example, we explore how to estimate the mean of a variable in a population using the sample mean of that variable in a simple random sample of the population.

- **M**odel:

  When we calculate a sample mean, we have an implicit model of the world in mind that justifies our inferences. The model has three components, which map exactly on to the three model elements we described in Chapter 1: a signature, functional equations, and root distributions. In this case the signature is just the conceptualization of a well-defined variable Y with a well defined size and range. For example, if we seek to estimate the average political ideology of residents of Princeton, New Jersey, on a left-right scale, we might presuppose (a) a population size, N, for Princeton and (b) that there is such a thing as ideology (Y) measurable on some scale. The functional equations seem absent here but in fact even in this simple case implicitly there is a measurement node in our model and we assume that measurement itself does not alter Y and that Y is measured without error. The distribution is a speculation on a possible population level distribution for Y. Note that you do not need a distribution to calculate a sample statistic, but if you define a distribution then you can simulation and thus diagnose the design; for example you can assess how well your procedure fares under different distributions.

- **I**nquiry:

  We ask: what is the population mean of Y, i.e. $\frac{1}{N}\sum_1^N Y_i = \bar{Y}$?

- **D**ata strategy:

  In simple random sampling, we draw a random sample without replacement of size n, where every member of the population has an equal probability of inclusion in the sample, $\frac{n}{N}$.

  We measure Y for all n units (every unit in the sample, but *not* every unit in the population).

- **A**nswer strategy:

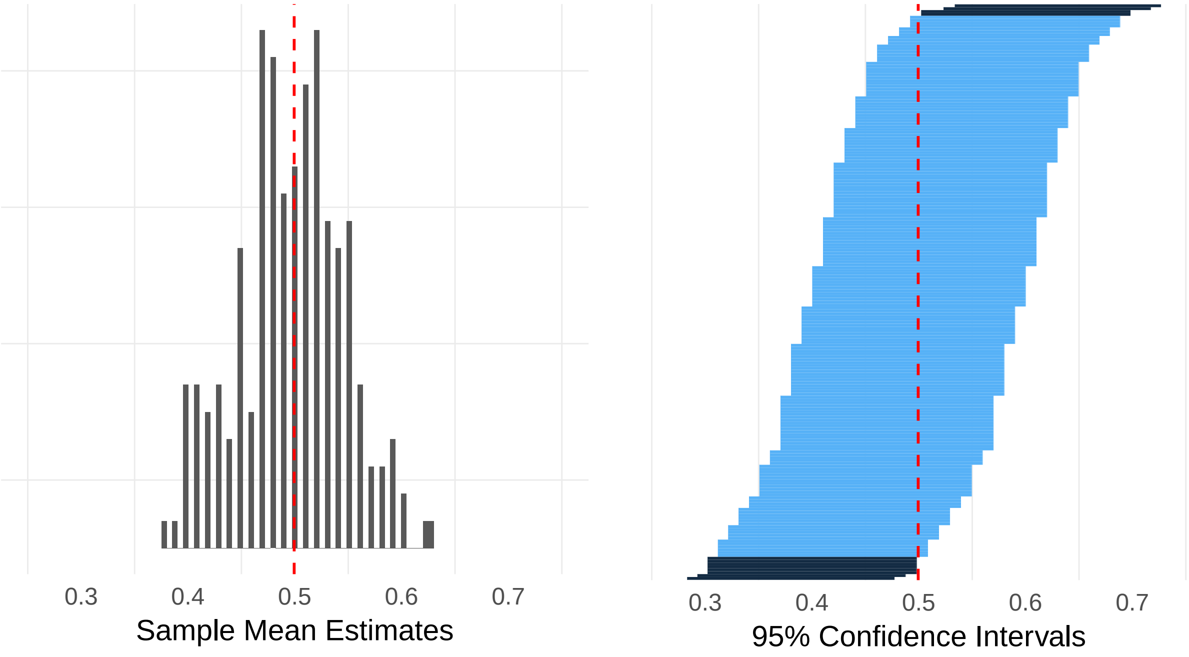
  We estimate the population mean with the sample mean, i.e. $\frac{1}{n}\sum_{i=1}^n Y_i$. Though our inquiry implies our answer should be a single number, typically an answer strategy also provides statistics that help us assess the uncertainty around that single number. Here, the we calculate a confidence interval around the estimate. In order to do so, we calculate the standard error of the sample mean, and then approximate the sampling distribution of the sample mean in order to construct the confidence interval. Our approximation is the t distribution with degrees of freedom of n minus 1, representing the single parameter we estimate. In the code for our answer strategy, we spell out each step in turn.

```r
# Model --------------------------------------------------------------------
population <- declare_population(
  N = 28572, Y = sample(1:7, N, replace = TRUE))

fixed_population <- population()

# Inquiry ------------------------------------------------------------------
estimand <- declare_estimand(Ybar = mean(Y))

# Data Strategy ------------------------------------------------------------
sampling <- declare_sampling(n = 100)

# Answer Strategy ----------------------------------------------------------
estimator <- declare_estimator(
  estimator_function = function(data) {
      est <- mean(data$Y)
      se <- sd(data$Y) / sqrt(nrow(data))
      critical_value <- qt(0.975, df = nrow(data) - 1)
      ci_lower <- est - critical_value * se
      ci_upper <- est + critical_value * se
      data.frame(est, ci_lower, ci_upper)
    },
  estimand = estimand,
  label = "Sample Mean Estimator"
)

# Design ------------------------------------------------------------------
design <- declare_design(fixed_population, estimand, sampling, estimator)
diagnosands <-
  declare_diagnosands(
    bias = mean(est - estimand),
    mean_est = mean(est),
    coverage = mean(ci_lower <= estimand &
                      estimand <= ci_upper)
  )
```

Takeaways

```
diagnosis <- diagnose_design(
  design, sims = 5000, bootstrap = FALSE, diagnosands = diagnosands)
```

estimand_label	estimator_label	bias	mean_est	coverage
Ybar	Sample Mean Estimator	0	3.99	0.95



- Under simple random sampling, the sample mean estimator of the population mean is unbiased. The graph on the left shows the sampling distribution of the estimator: it's centered directly on the true value of the inquiry.
- Confidence intervals **also** have a sampling distribution - they change depending on the idiosyncracies of the sample we happen to draw. We built a 95% confidence interval so that 95% of the time, they will cover the true value.
- As sample size grows, the sampling distribution of the estimator gets tighter, but the coverage of the confidence intervals stays at 95% – just the properties we would want out of our answer strategy.

Exercises

1. We sampled a small number of units relative to the size of the population. We defined our standard error as usual, $\hat{\sigma} \equiv \frac{\hat{\sigma}}{\sqrt{n}}$

   When this ratio is higher, i.e. when the sample represents a substantial proportion of the population, a finite population correction is needed, so we redefine our standard error as:

   $$\hat{\sigma} \equiv \frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

   We show in the diagnosis above that our coverage is correct when sampling 100 units from a 10000-unit population. Now show how coverage changes with and without the finite population correction as a function of N, and identify the sample size at which the correction starts to matter.

2. Modify the declaration to change the distribution of Y from a 1 to 7 ideology scale to something else. Try a binary variable or a variable with a normal distribution. Is the sample mean estimator still unbiased? Interpret your answer.

3. Change the sampling procedure to favor units with higher values of ideology. Is the sample mean estimator still unbiased?