Stratified Sampling and Cluster Sampling: Takeaways

Syntax

- Sampling randomly from three strata using and then combining into a single dataframe:

```
stratum_1 <- df %>%
    filter(condition) %>%
    sample_n(1)
stratum_2 <- df %>%
    filter(condition) %>%
    sample_n(2)
stratum_3 <- df %>%
    filter(condition) %>%
    sample_n(7)
combined <- bind_rows(stratum_1, stratum_2, stratum_3)
```

- Using the split-apply-combine workflow to stratify, randomly sample, and estimate:

```
df %>%
  # Split: stratify
  group_by(strata_col) %>%
  # Apply: sample n observations for each stratum
  sample_n(n) %>%
  # Apply & combine: calculate mean value for each stratum, combine results
  summarize(mean = mean(col))
```

- Sampling randomly 25% of the units within each stratum:

```
df %>%
  group_by(stratum_column) %>%
  sample_frac(.25)
```

Concepts

- To make our samples representative we can try different sampling methods:
    - **Simple random sampling**
    - **Stratified sampling**
    - **Cluster sampling**
- Choosing strata:
    - Minimize variability within each stratum
    - Maximize variabiity between strata
    - Stratification criterion should be strongly correlated with the property you're trying to measure

- When we describe a sample or a population, we do **descriptive statistics**. When we try to use a sample to draw conclusions about a population, we do **inferential statistics** (we *infer* information from the sample about the population).

Resources

- The Wikipedia entry on stratified sampling.
- The Wikipedia entry on cluster sampling.