Agenda

1. Experimental Design

Recap of Lab Session

- Please email me if you still can't log in to the Cloud RStudio Server!
- Remember that your R Markdown document renders in a clean workspace!
- Procedure for submitting HTML files to blackboard (see Resources tab)
- Tricks: the UP key for cycling through commands, the TAB key for auto-complete
- Recommendation: re-type the code from the labs – don't just copy-and-paste

Stratified sampling simulation Recall the stratifield sampling exercise from last time, and suppose that hourly wages were normally distributed with means \$25, \$15, \$22, and \$15, among the 90 women working full-time, 18 women working part-time, 9 men working full-time, and 63 men working part-time, respectively. The following R code builds a data frame that represents one possible reality.

```
> staff <- c(rep("women-ft", 90), rep("women-pt", 18), rep("men-ft", 9), rep("men-pt", 63))
> wage <- c(rnorm(90, mean = 25), rnorm(18, 15), rnorm(9, 22), rnorm(63, 15))
> ds <- data.frame(staff, wage)
> head(ds)
```

```
     staff     wage
1 women-ft 25.55224
2 women-ft 24.43177
3 women-ft 24.96609
4 women-ft 24.81856
5 women-ft 25.30242
6 women-ft 24.71514
```

```
> nrow(ds)
```

```
[1] 180
```

Note that wages are similar *within* the four groups, but dissimilar *among* the groups. This command will draw separate densities for the four groups on the same plot.

```
> require(mosaic)
> qplot(data = ds, x = wage, color = staff, geom = "density")
```

We want to estimate the mean wage among all 180 workers. In this case, since we know the wage of all of the workers, we can just compute it.

«»= mean(wage, data = ds)

But recall that for the purposes of this exercise, we don't actually know all 180 wages, and we are asked to sample 40 of them. We can take a *simple random sample* and compute the mean wage within that sample.

```
> # simple random sampling
> mean(~wage, data = sample(ds, 40))
```

```
[1] 20.95426
```

Note that this is close to the actual mean wage, but not the same. Note also that each time we take a different random sample, we get a different mean wage in that sample.

Now let's implement the *stratified sampling* scheme.

```
> # Stratified sampling
> strat_samp <- bind_rows(
+   sample(filter(ds, staff == "women-ft"), 20),
+   sample(filter(ds, staff == "women-pt"), 4),
+   sample(filter(ds, staff == "men-ft"), 2),
+   sample(filter(ds, staff == "men-pt"), 14)
+ )
> mean(~wage, data = strat_samp)

[1] 20.20743
```

Again, the stratified sample mean is close to the actual value, but not the same. It will also differ each time we take a different random sample. So why might we prefer stratified sampling over simple random sampling?

Let's compare the *distribution* of sample means if we do this many, many times!

```
> # Comparison
> SRS <- do(1000) * mean(~wage, data = sample(ds, 40))
> STR <- do(1000) * mean(~wage, data = bind_rows(
+                                 sample(filter(ds, staff == "women-ft"), 20),
+                                 sample(filter(ds, staff == "women-pt"), 4),
+                                 sample(filter(ds, staff == "men-ft"), 2),
+                                 sample(filter(ds, staff == "men-pt"), 14)
+ ))
> sim <- bind_rows(SRS, STR) %>%
+   mutate(scheme = rep(c("simple", "statified"), each = 1000))
> qplot(data = sim, x = mean, color = scheme, geom = "density")
```

Experimental Design

1. What is the best way to answer each of the questions below: an experiment, a sample survey, or an observational study that is not a sample survey? Explain your choices.

   (a) Are people generally satisfied with how things are going in the country right now?
   Survay

   (b) Do college students learn basic accounting better in a classroom or using an online course?
   Experimental

   (c) How long do your teachers wait on average after they ask their class a question?

2. A study showed that women who work in the production of computer chips have abnormally high numbers of miscarriages. The union claimed that exposure to chemicals used in production caused the miscarriages. Another possible explanation is that these workers spend most of their work time standing up. Illustrate these relationships in a diagram.

3. A study shows that there is a positive correlation between the size of a hospital (measured by its number of beds x) and the median number of days y that patients remain in the hospital. Does this mean that you can shorten a hospital stay by choosing a small hospital? Use a diagram to explain the association.

4. Someone says, "There is a strong positive correlation between the number of firefighters at a fire and the amount of damage the fire does. So sending lots of firefighters just causes more damage." Why is this reasoning wrong?

Activity For each of the following pairs of variables, a statistically signficant positive relationship has been observed. Identify a potential lurking variable that might cause the spurious correlation.

Vivek notices that students in his class with larger shoe sizes tend to have higher grade point averages. Based on this observation, what is the best description of the relationship between shoe size and grade point average?