Agenda

1. DataFest information session Feb 7th, 7p-8p!
2. Smithies in SDS (SSDS) information session Feb 8th, 7p-8p!
3. Laptop loan program
4. Recap of Samplng Techniques
5. Experimental Design

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.

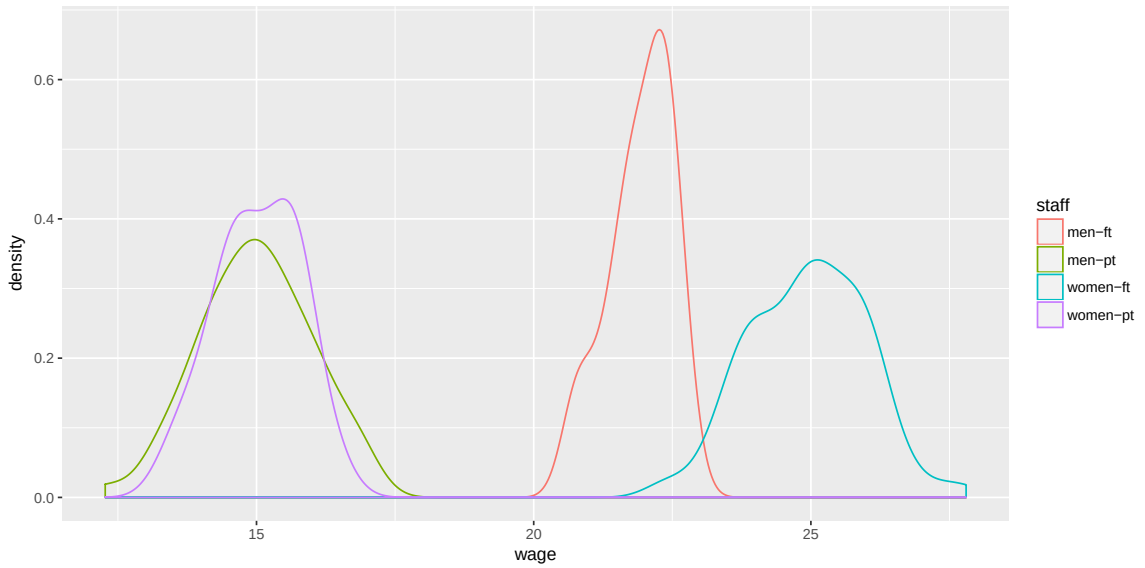
```
staff <- c(rep("women-ft", 90),
           rep("women-pt", 18),
           rep("men-ft", 9),
           rep("men-pt", 63))

wage <- c(rnorm(n = 90, mean = 25), rnorm(18, 15), rnorm(9, 22), rnorm(63, 15))

ds <- data.frame(staff, wage)
```

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.

```
library(mosaic)
library(ggplot2)
qplot(data = ds, x = wage, color = staff, geom = "density")
```



Our goals is to know 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 directly.

```
mean(~wage, data = ds)

## [1] 20.31593
```

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. This mean is now an *estimate* of the mean wage of all 180 workers.

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

## [1] 20.4122
```

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] 20.32956
```

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!

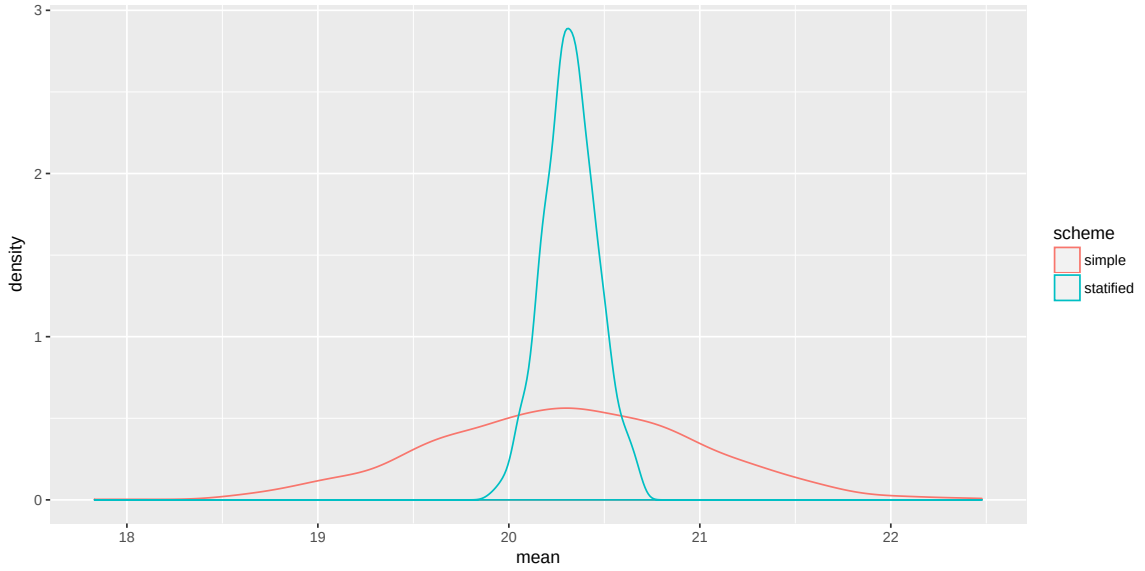
```
# Comparison
SRS <- do(1000) * mean(~wage, data = sample(ds, 40))
STR <- do(1000) * mean(~wage, data = bind_rows(
                                sample(filter(ds, staff == "women-ft"), 20),
                                sample(filter(ds, staff == "women-pt"), 4),
                                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))

head(sim)

##       mean scheme
## 1 20.31110 simple
## 2 20.28608 simple
## 3 21.01234 simple
## 4 20.38969 simple
## 5 20.08515 simple
## 6 20.44397 simple
```

```
qplot(data = sim, x = mean, color = scheme, geom = "density")
```



Experimental Design One experimental study tested for gender bias in hiring decisions within the academic sciences.

> Moss-Racusin, C. A. et al. (2012). Science faculty's subtle gender biases favor male students. PNAS, 109(41).
>
> **Abstract:** In a randomized double-blind study (n = 127), science faculty from research-intensive universities rated the application materials of a student who was randomly assigned either a male or female name for a laboratory manager position. Faculty participants rated the male applicant as significanly more competent and hireable than the (identical) female applicant. These participants also selected a higher starting salary and offered more career mentoring to the male applicant. (See materials here)

1. Controlling

2. Randomization

3. Replication

4. Blocking

1. What is the best way to answer each of the questions below: an experiment or an observational study? Explain your choices.

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

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

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

   (d) Does seeing someone else yawn make people more likely to yawn?

   (e) Is having 1 versus 2 children better in terms of marital satisfaction?

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. Students sign up to be subjects in a psychology experiment. When they arrive, they are told that interviews are running late and are taken to a waiting room. The experimenters then stage a theft of a valuable object left in the waiting room. Some subjects are alone with the thief, and others are in pairs – these are the treatments being compared. Will the subject report the theft? Afterwards, a consent form is given and the true nature of the experiment is explained to them. Do you think this study is ethically OK?

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

1. The amount of ice cream sold in New England and the number of deaths by drowning

2. The salary of U.S. ministers and the price of vodka

3. The number of doctors in a region and the number of crimes committed in that region

4. The amount of coffee consumed and the prevalence of lung cancer