Principles of Data Collection

Match the definition with correct term

The target group about which you'd like to make inferences

A. Sample
B. Population
C. Summary statistic
D. Anecdote

Match the definition with correct term

The individual unit on which you make observations

A. Parameter
B. Sample
C. Case
D. Census

Principles of Data Collection for Inference

case/observational unit - the individual unit on which you make observations (a row in a data frame) (In `survey141` data set, a case is a single student)

population - the target group of observational units about which you'd like to make inferences (size of the population: N)

sample - a subset of the population on which you have data (size of the sample: n)

anecdote - very small sample of data collected haphazardly (usually $n = 1$)

census - sample = population (complex, expensive, and sometimes impossible to achieve)

Sampling considerations

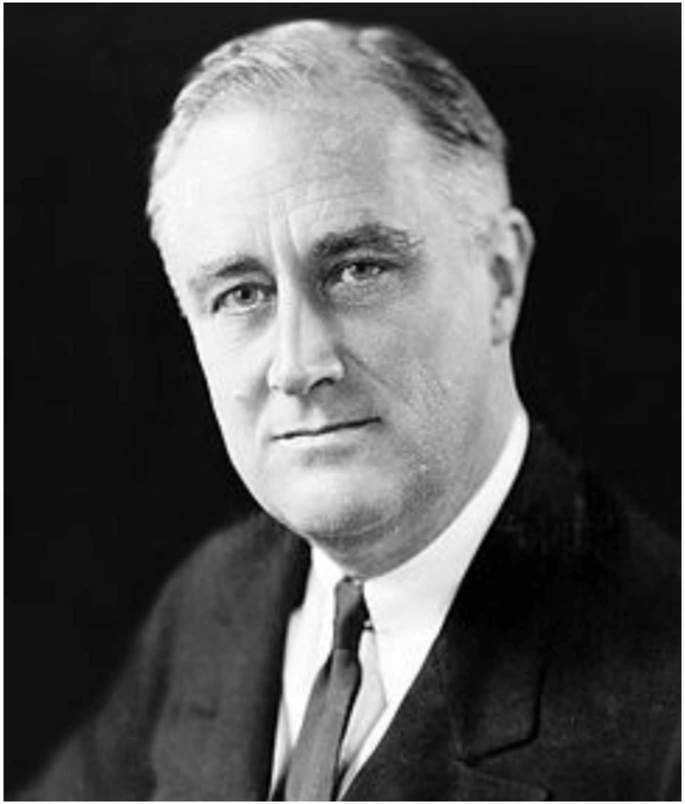
You're a senior Psychology major conducting a study that examines procrastination among Reed students. How should you select a sample?

A. Post a link to your survey on your Facebook page.

$$n \approx 100$$

B. Get a list of Reed student emails from the Registrar, take a simple random sample (SRS), and email that sample.

Initial $n = 100$, Final $n = 34$



Landon v. FDR, 1936

Literary Digest polled 10 million Americans, 2.4 million responded.

N = 128 million, n = 2.4 million

Prediction: 43% for FDR

Result: 62% for FDR

What went wrong?

Literary Digest surveyed

- magazine subscribers
- registered car owners
- registered telephone owners

These groups have a much higher income on average than the typical american. In 1936 the Great Depression is still in full swing, so the typical (poorer) american was more supportive of FDR.

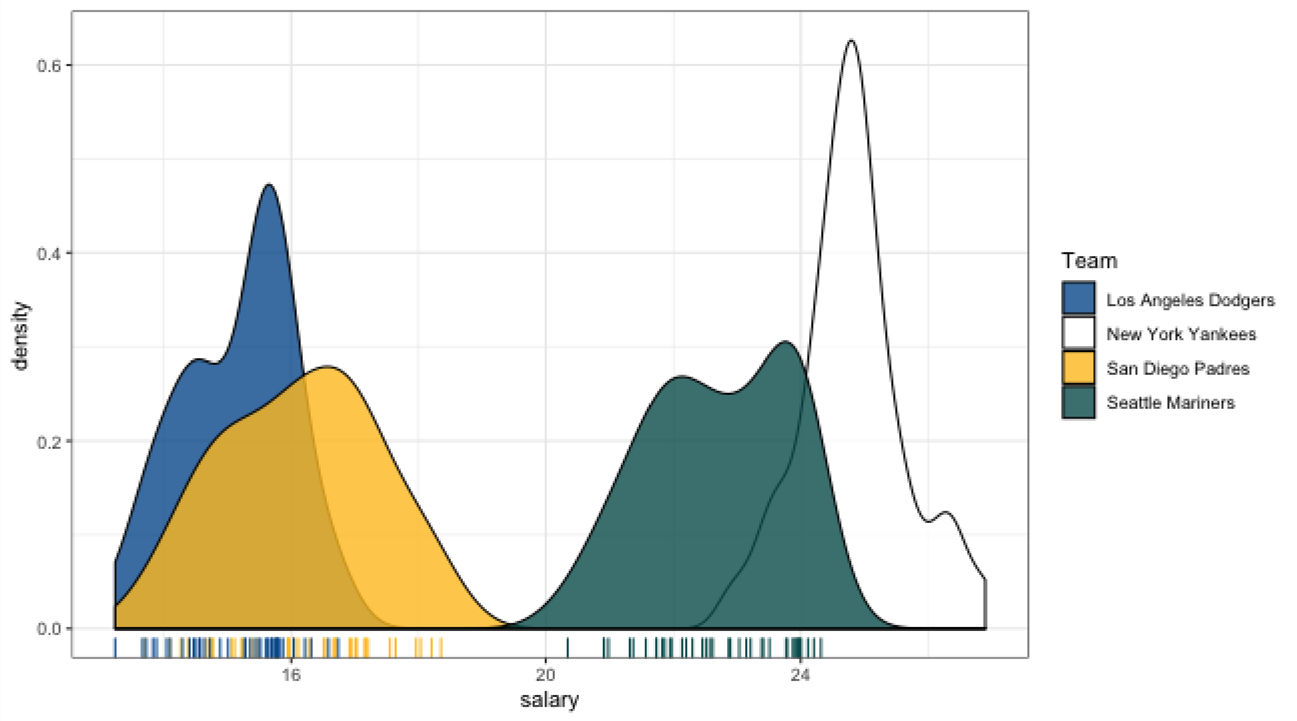
Their sampling method was **biased** (not representative).

Sampling Strategies: stratification

```r
teams <- c(rep("New York Yankees", 40),
           rep("San Diego Padres", 40),
           rep("Seattle Mariners", 40),
           rep("Los Angeles Dodgers", 40))
salary <- c(rnorm(40, mean = 25),
            rnorm(40, 16),
            rnorm(40, 23),
            rnorm(40, 15))
df <- data.frame(teams, salary)
head(df)
```

```
##              teams   salary
## 1 New York Yankees 26.19713
## 2 New York Yankees 24.44472
## 3 New York Yankees 24.79961
## 4 New York Yankees 25.59124
## 5 New York Yankees 25.88720
## 6 New York Yankees 24.48721
```

Population view



Simple Random Sample (SRS)

```
# population mean
mean(df$salary)
```

```
## [1] 19.70155
```

```
# SRS
df %>% sample_n(40) %>% summarize(mean(salary))
```

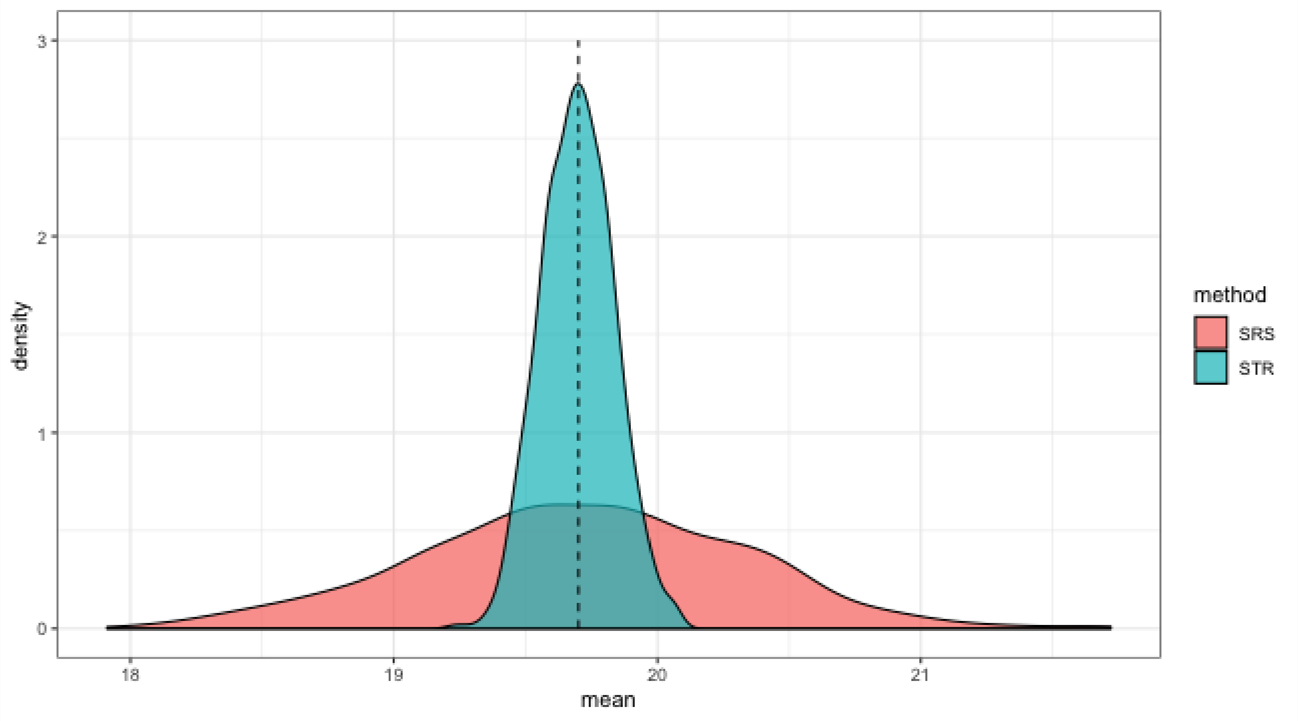
```
##   mean(salary)
## 1     20.30406
```

Stratified Sample

```
# Stratified sample
df %>%
  group_by(teams) %>%
  sample_n(10) %>%
  ungroup() %>%
  summarize(mean(salary))
```

```
## # A tibble: 1 x 1
##   mean(salary)
##            <dbl>
## 1           19.7
```

Long-run performance



Sampling Strategies

SRS: Unbiased, easy, but can be high variance.

Stratified Sampling: Divide population into strata that are *similar* within and *different* between. Draw SRS from within each strata.