Gov 50: 16. Sampling

Matthew Blackwell

Harvard University

Roadmap

1. Sampling exercise
2. Sampling framework
3. Polls

1/ Sampling exercise

Data on class years enrolled in Gov 50

```r
library(gov50data)
class_years
```

```
## # A tibble: 122 x 1
##    year
##    <chr>
##  1 Senior
##  2 Junior
##  3 Sophomore
##  4 Junior
##  5 Graduate Year 2
##  6 Sophomore
##  7 Professional Year 2
##  8 First-Year
##  9 Sophomore
## 10 Junior
## # ... with 112 more rows
```

What proportion of the class is first years?

```
class_years |>
  count(year) |>
  mutate(prop = n / nrow(class_years))
```

```
## # A tibble: 9 x 3
##   year                      n    prop
##   <chr>                 <int>   <dbl>
## 1 First-Year               25 0.205
## 2 Graduate Year 1           2 0.0164
## 3 Graduate Year 2           1 0.00820
## 4 Junior                   31 0.254
## 5 Not Set                   3 0.0246
## 6 Professional Year 2       2 0.0164
## 7 Senior                   14 0.115
## 8 Sophomore                43 0.352
## 9 Year 1, Semester 1        1 0.00820
```

Let's take some samples!

We can use the `slice_sample()` function to take a random sample of rows of a tibble:

```
class_years |>
  slice_sample(n = 5)
```

```
## # A tibble: 5 x 1
##   year
##   <chr>
## 1 Sophomore
## 2 Junior
## 3 Junior
## 4 Sophomore
## 5 Sophomore
```

Another sample

```r
class_years |>
  slice_sample(n = 5)
```

```
## # A tibble: 5 x 1
##   year
##   <chr>
## 1 Junior
## 2 Not Set
## 3 First-Year
## 4 First-Year
## 5 Sophomore
```

Sample proportion of first-years

```
class_years |>
  slice_sample(n = 20) |>
  summarize(fy_prop = mean(year == "First-Year"))
```

```
## # A tibble: 1 x 1
##   fy_prop
##     <dbl>
## 1    0.15
```

Repeated sampling

We sometimes want to draw multiple samples from a tibble. For this we can use `rep_slice_sample()` from the `infer` package:

```
library(infer)
class_years |>
  rep_slice_sample(n = 5, reps = 2)
```

```
## # A tibble: 10 x 2
## # Groups:   replicate [2]
##    replicate year
##        <int> <chr>
##  1         1 First-Year
##  2         1 Sophomore
##  3         1 First-Year
##  4         1 Sophomore
##  5         1 First-Year
##  6         2 Junior
##  7         2 First-Year
##  8         2 Sophomore
##  9         2 First-Year
## 10         2 Sophomore
```

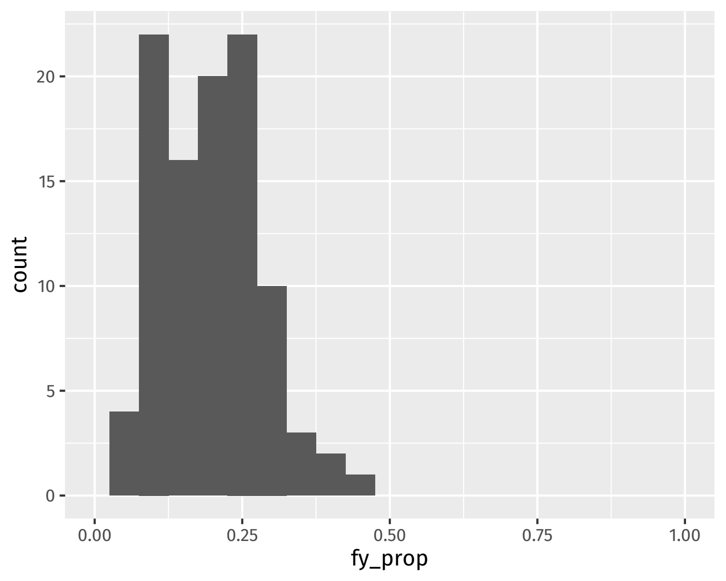
Simulate many separate studies being done

```
samples_n20 <- class_years |>
  rep_slice_sample(n = 20, reps = 100) |>
  group_by(replicate) |>
  summarize(fy_prop = mean(year == "First-Year"))
samples_n20
```

```
## # A tibble: 100 x 2
##    replicate fy_prop
##        <int>   <dbl>
##  1         1    0.25
##  2         2    0.4
##  3         3    0.3
##  4         4    0.4
##  5         5    0.2
##  6         6    0.25
##  7         7    0.1
##  8         8    0.25
##  9         9    0.35
## 10        10    0.1
## # ... with 90 more rows
```

Distribution of these proportions

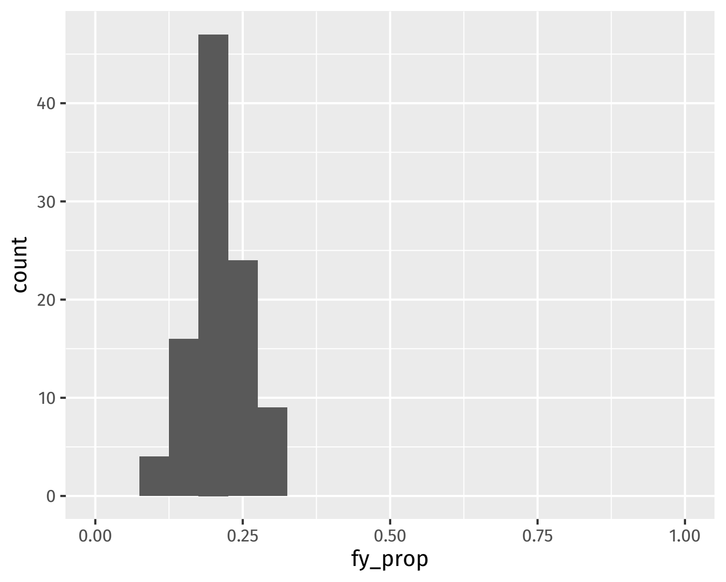
```r
samples_n20 |>
  ggplot(mapping = aes(x = fy_prop)) +
  geom_histogram(binwidth=0.05) +
  lims(x = c(0, 1))
```


count
20
15
10
5
0
0.00
0.25
0.50
0.75
1.00
fy_prop

What if the sample sizes are bigger?

```
samples_n50 <- class_years |>
  rep_slice_sample(n = 50, reps = 100) |>
  group_by(replicate) |>
  summarize(fy_prop = mean(year == "First-Year"))

samples_n50 |>
  ggplot(mapping = aes(x = fy_prop)) +
  geom_histogram(binwidth=0.05)  +
  lims(x = c(0, 1))
```

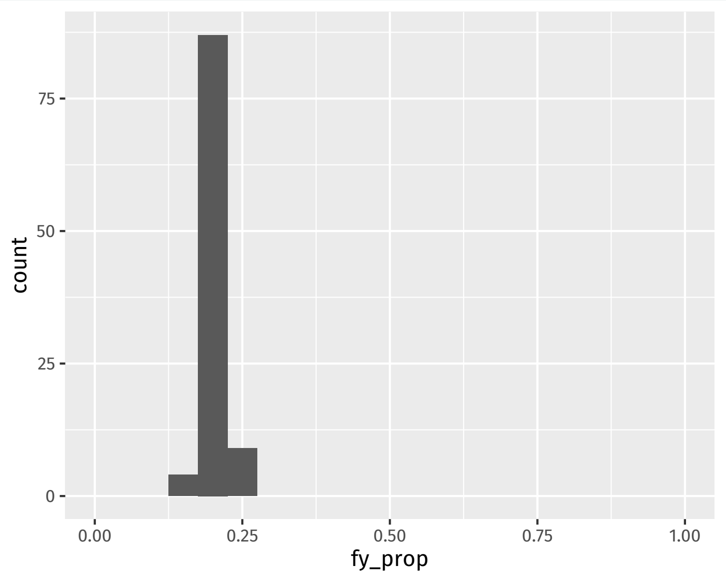


count
40
30
20
10
0
0.00
0.25
0.50
0.75
1.00
fy_prop

What if the sample sizes are bigger?

```
samples_n100 <- class_years |>
  rep_slice_sample(n = 100, reps = 100) |>
  group_by(replicate) |>
  summarize(fy_prop = mean(year == "First-Year"))

samples_n100 |>
  ggplot(mapping = aes(x = fy_prop)) +
  geom_histogram(binwidth=0.05)  +
  lims(x = c(0, 1))
```


count
75
50
25
0
0.00
0.25
0.50
0.75
1.00
fy_prop

Sample size and variability across samples

```
samples_n20 |>
  summarize(sd(fy_prop)) |> pull()
```

```
## [1] 0.0849
```

```
samples_n50 |>
  summarize(prop_sd = sd(fy_prop)) |> pull()
```

```
## [1] 0.0427
```

```
samples_n100 |>
  summarize(prop_sd = sd(fy_prop)) |> pull()
```

```
## [1] 0.0147
```

2/ Sampling framework

Populations

Population: group of units/people we want to learn about.

Population parameter: some numerical summary of the population we would like to know. - population mean/proportion, population standard deviation.

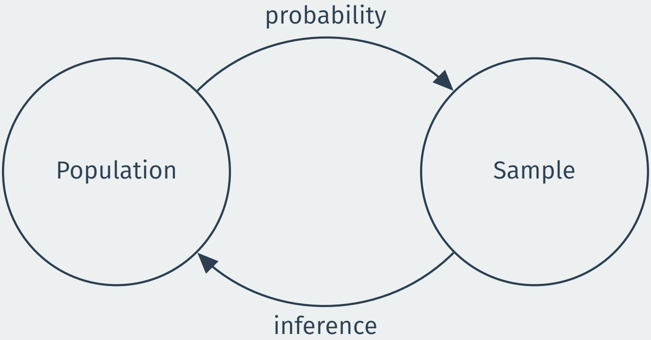
Census: complete recording of data on the entire population.

Samples

Sample: subset of the population taken in some way (hopefully randomly).

Estimator or sample statistic: numerical summary of the sample that is our "best guess" for the unknown population parameter.

Sampling framework



Sampling at random

Random sample: units selected into sample from population with a non-zero probability.

Simple random sample: all units have the same probability of being selected into the sample.

Our sampling exercise

- **Population**: all students enrolled in Gov 50.
- **Population parameter**: population proportion of first-years enrolled in Gov 50
  - Population proportions often denoted p
- **Sample**: simple random sample of different sizes.
- **Sample statistic/estimator**: sample proportion of first-years
  - Estimators often denoted with a hat: $\hat{p}$
  - We saw the $\hat{p}$ varies with the random sample taken.

Expected value

The **expected value** of a sample statistic, $\mathbb{E}[\hat{p}]$, is the average value of the statistic across repeated samples.

```
samples_n100 |>
  summarize(mean(fy_prop)) |> pull()
```

```
## [1] 0.205
```

The **expected value** of a sample proportion from a simple random sample is equal to the population proportion, $\mathbb{E}[\hat{p}] = p$

Standard error

The **standard error** is the standard deviation of the sample statistic across repeated samples.

```
samples_n100 |>
  summarize(sd(fy_prop)) |> pull()
```

```
## [1] 0.0147
```

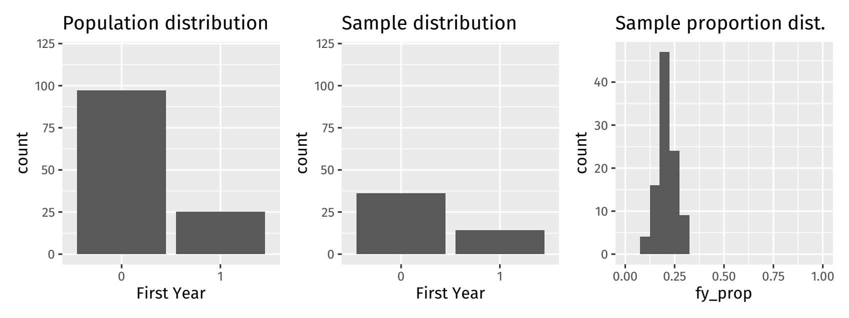
Tells us how far away, on average, the sample proportion will be from the population proportion.

Standard error vs population standard deviation

The **standard error** is the SD of the statistic across repeated samples.

Should not be confused with the population standard deviation or sample standard deviation, both of which measure how far **units** are away from a mean.

The three distributions



3/ Polls

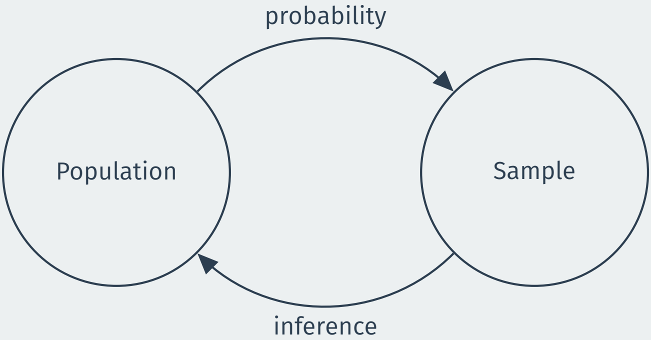
How popular is Joe Biden?

- What proportion of the public approves of Biden's job as president?
- Latest Gallup poll:
    - Sept 1st-16th
    - 812 adult Americans
    - Telephone interviews
    - Approve (42%), Disapprove (56%)

Poll in our framework

- **Population**: adults 18+ living in 50 US states and DC.
- **Population parameter**: population proportion of all US adults that approve of Biden.
  - Census: not possible.
- **Sample:** random digit dialing phone numbers (cell and landline).
- **Point estimate**: sample proportion that approve of Biden

Where are we going?



We only get 1 sample. Can we learn about the population from that sample?