Sampling

Prabidhik KC

2022-11-23

We will be workin on sampling.

```
## Loading the necessary libraries

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(gov50data)
```

```
class_years <- read_csv("class_years.csv")
```

```
## Rows: 122 Columns: 1
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): year
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
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
```

```
class_years %>%
  slice_sample(n = 100) %>%
  summarize(fir_prop = mean(year == "First-Year")) %>%
  pull()
```

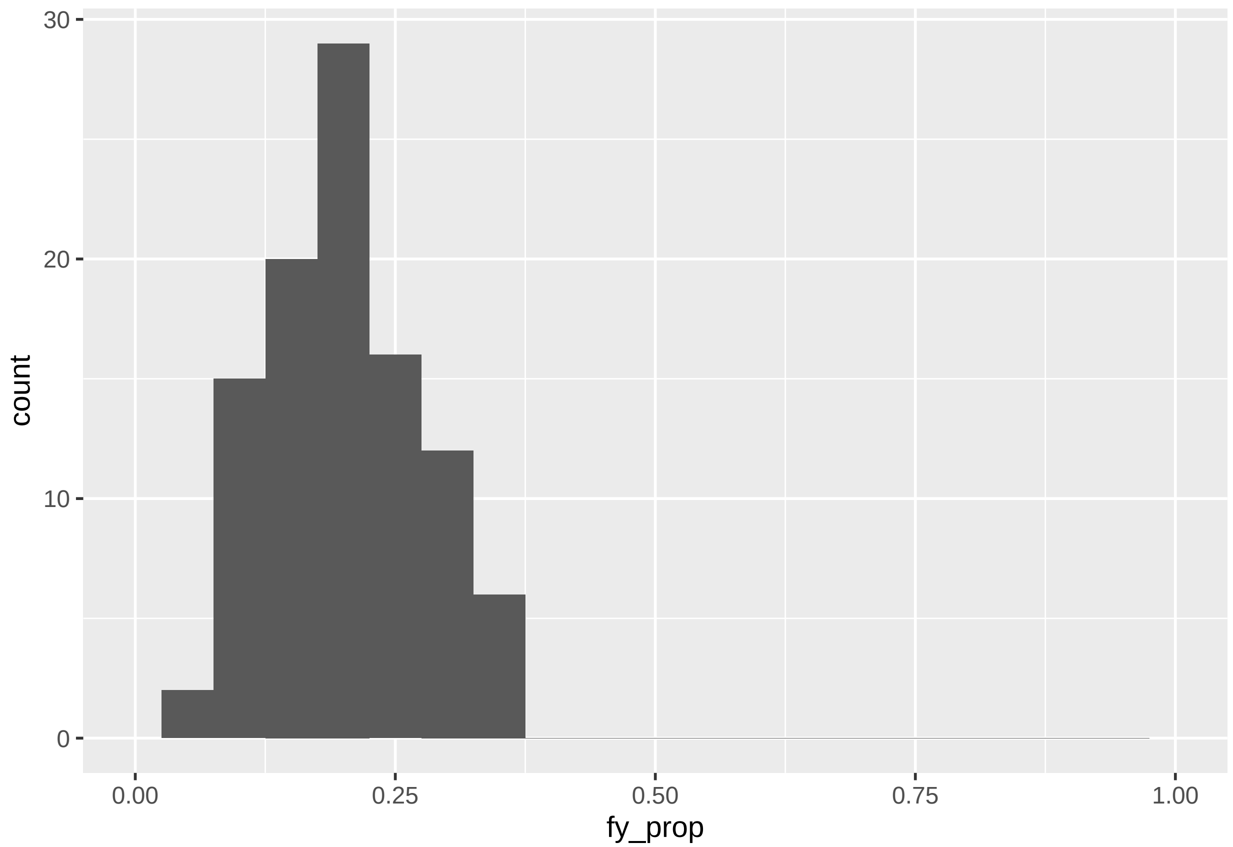
```
## [1] 0.22
```

```
library(infer)
```

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

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

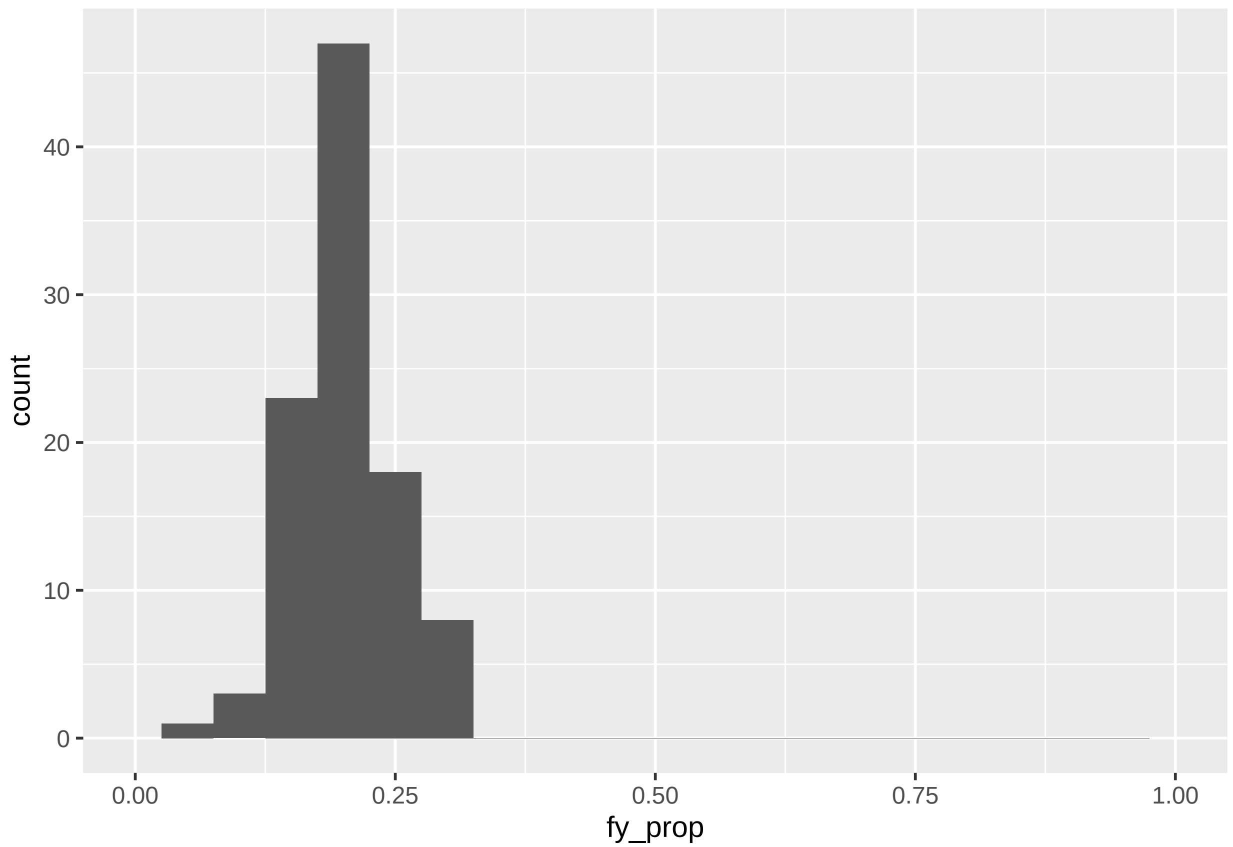
```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



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

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

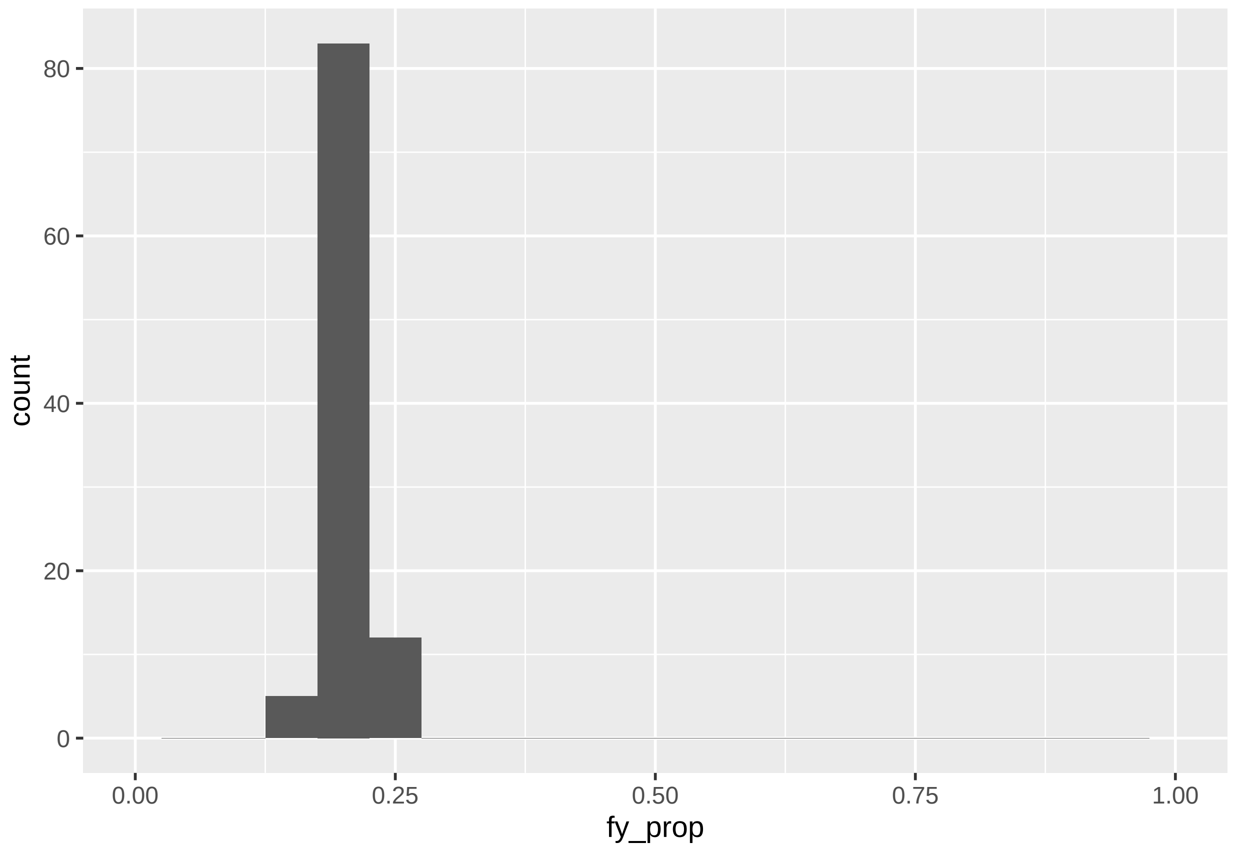
```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



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

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

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



```
## Lets check the standard deviation as we increase the sample size

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

```
## [1] 0.07384806
```

```
samples_n50 %>%
  summarize(sd(fy_prop)) %>%
  pull()
```

```
## [1] 0.04462685
```

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

```
## [1] 0.01866342
```

We observe standard deviation decreasing as we increase the sample size.

```r
samples_n100 <- class_years %>%
  rep_slice_sample(n = 100, reps = 1000) %>%
  group_by(replicate) %>%
  summarize(fy_prop = mean(year == "First-Year")) %>%
  summarize(mean(fy_prop)) %>%
  pull()
```