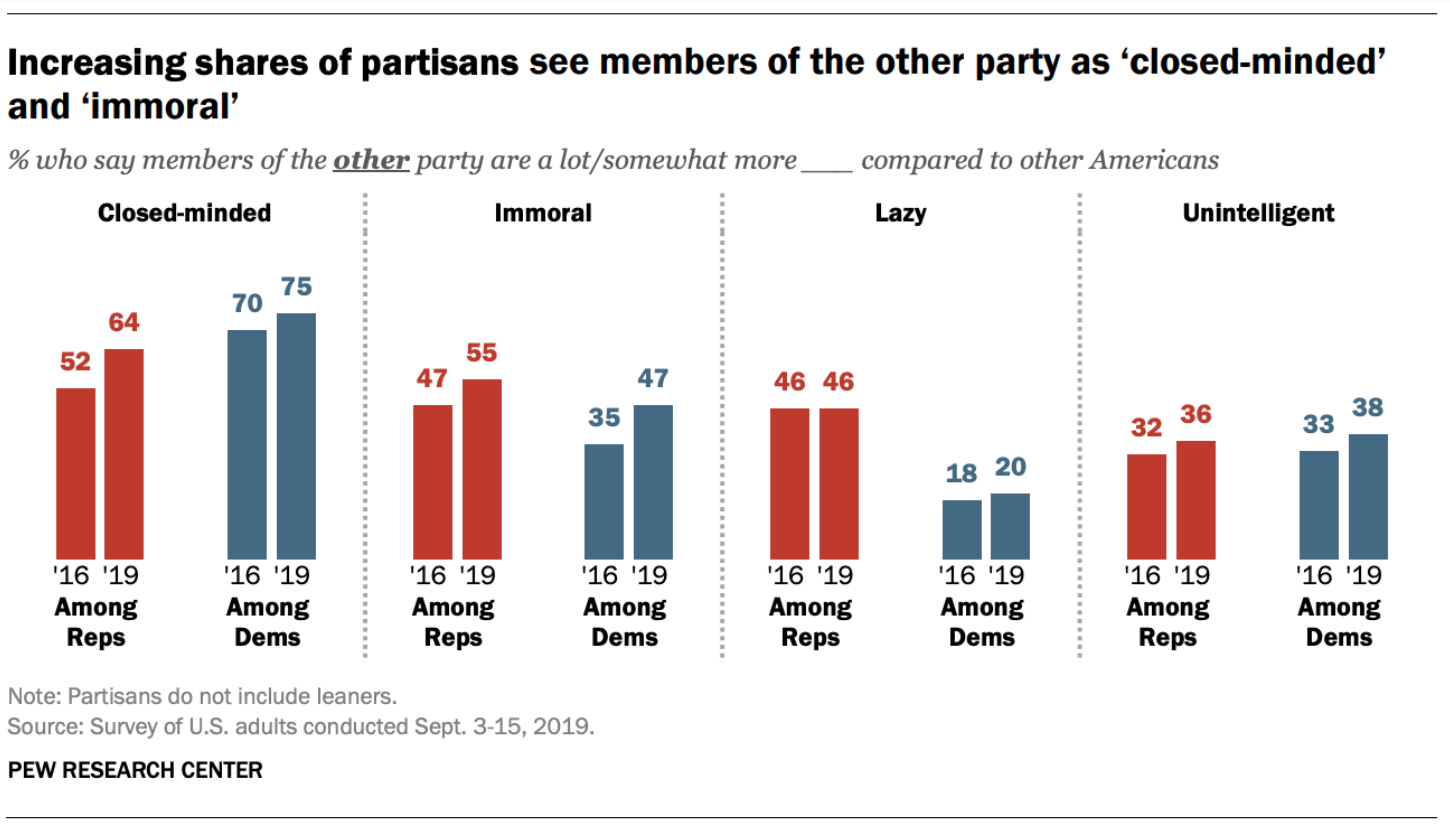
Confidence Intervals for Differences in Proportions

Returning to Pew . . .



Was there really an increase in the proportion of Democrats that view Republicans as lazy or is that just sampling variability?

The Data

```
pew <- data.frame(party = "Democrat",
                  year = rep(c(2016, 2019),
                             c(4947, 4947)),
                  lazy = c(rep(c("yes", "no"),
                               c(890, 4057)),
                           rep(c("yes", "no"),
                               c(989, 3958))))
```

```
slice(pew, 1:5)
```

```
##      party year lazy
## 1 Democrat 2016  yes
## 2 Democrat 2016  yes
## 3 Democrat 2016  yes
## 4 Democrat 2016  yes
## 5 Democrat 2016  yes
```

The Data

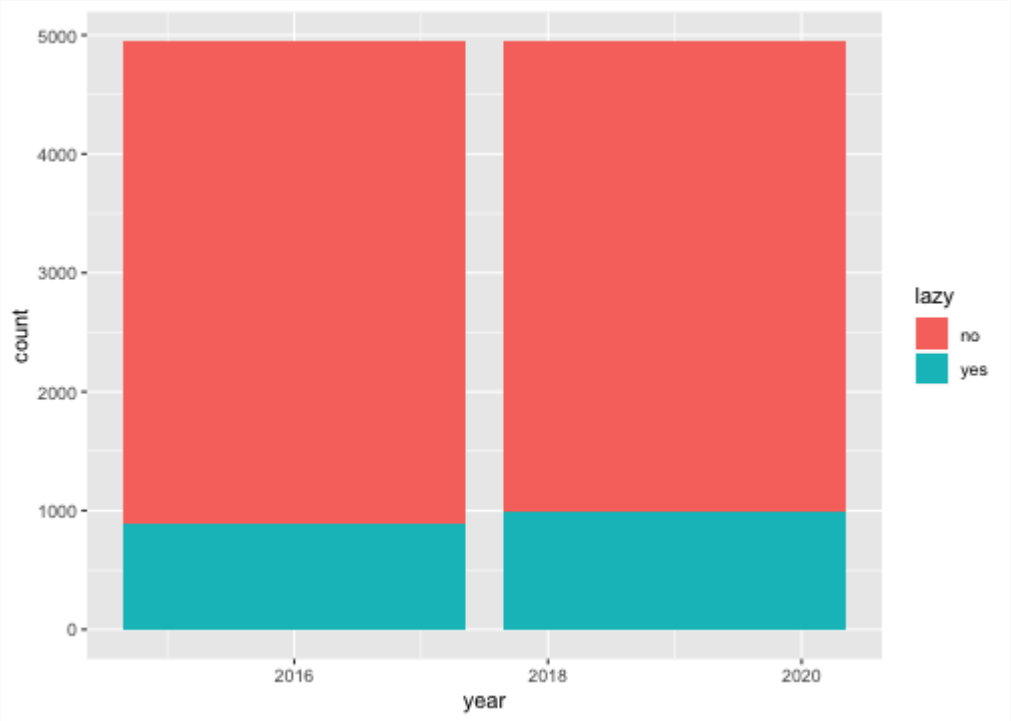
```
pew <- data.frame(party = "Democrat",
                  year = rep(c(2016, 2019),
                             c(4947, 4947)),
                  lazy = c(rep(c("yes", "no"),
                               c(890, 4057)),
                           rep(c("yes", "no"),
                               c(989, 3958))))
```

```
slice(pew, 4946:4950)
```

```
##       party year lazy
## 1 Democrat 2016   no
## 2 Democrat 2016   no
## 3 Democrat 2019  yes
## 4 Democrat 2019  yes
## 5 Democrat 2019  yes
```

Visualization

```
library(tidyverse)
ggplot(pew, aes(x = year, fill = lazy)) +
  geom_bar()
```



Point estimate

```
library(infer)
(point_est <- pew %>%
  specify(response = lazy,
          explanatory = year,
          success = "yes") %>%
  calculate(stat = "diff in props",
            order = c(2019, 2016)) %>%
  pull())
```

```
## Error: The explanatory variable of year is not appropr
## since 'diff in props` is expecting the explanatory var
```

Point estimate

```
library(infer)
(point_est <- pew %>%
  mutate(year = factor(year)) %>%
  specify(response = lazy,
          explanatory = year,
          success = "yes") %>%
  calculate(stat = "diff in props",
            order = c(2019, 2016)) %>%
  pull())
```

```
## [1] 0.02001213
```

```
pew <- mutate(pew, year = factor(year))
```

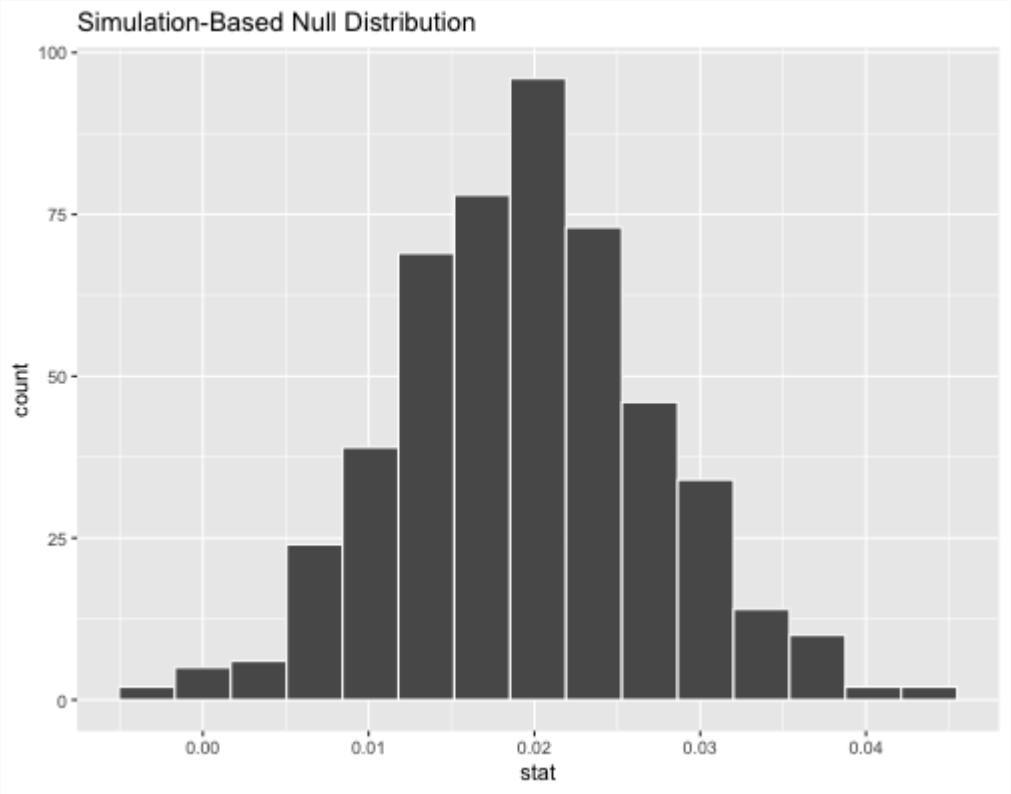
Bootstrapping the SE

```
(boot <- pew %>%
  specify(response = lazy,
          explanatory = year,
          success = "yes") %>%
  generate(reps = 500,
           type = "bootstrap") %>%
  calculate(stat = "diff in props",
            order = c(2019, 2016)))
```

```
## # A tibble: 500 x 2
##    replicate    stat
##        <int>   <dbl>
##  1         1 0.0255
##  2         2 0.00809
##  3         3 0.0135
##  4         4 0.0354
##  5         5 0.0112
##  6         6 0.0110
##  7         7 0.0112
```

The Bootstrap Distribution

```
boot %>%
  visualize()
```



The Bootstrap SE

```
(boot_se <- boot %>%
  summarize(se = sd(stat)) %>%
  pull())
```

```
## [1] 0.007679403
```

Construct the CI

```
c(point_est - 1.96 * boot_se,
  point_est + 1.96 * boot_se)
```

```
## [1] 0.004960499 0.035063759
```

Alternative: Normal Approximation

Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to be normal:

- each proportion separately follows a normal model
- the two samples are independent of one another

The standard error can be estimated with:

$$\widehat{SE} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

3.35 HIV in sub-Saharan Africa. In July 2008 the US National Institutes of Health announced that it was stopping a clinical study early because of unexpected results. The study population consisted of HIV-infected women in sub-Saharan Africa who had been given single dose Nevaripine (a treatment for HIV) while giving birth, to prevent transmission of HIV to the infant. The study was a randomized comparison of continued treatment of a woman (after successful childbirth) with Nevaripine vs. Lopinavir, a second drug used to treat HIV. 240 women participated in the study; 120 were randomized to each of the two treatments. Twenty-four weeks after starting the study treatment, each woman was tested to determine if the HIV infection was becoming worse (an outcome called *virologic failure*). Twenty-six of the 120 women treated with Nevaripine experienced virologic failure, while 10 of the 120 women treated with the other drug experienced virologic failure.[50]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for independence of treatment and virologic failure.

(c) Complete the hypothesis test and state an appropriate conclusion. (Reminder: verify any necessary conditions for the test.)

```
hiv <- data.frame(treatment = rep(c("Nevaripine", "Lopinavir"),
                                  c(120, 120)),
                  outcome = c(rep(c("worse", "not worse"),
                                  c(26, 94)),
                              rep(c("worse", "not worse"),
                                  c(10, 110))))
table(hiv)
```

```
##             outcome
## treatment    not worse worse
##   Lopinavir        110    10
```