POLSCI 9590: Methods I

Sampling and Generalization

Dave Armstrong

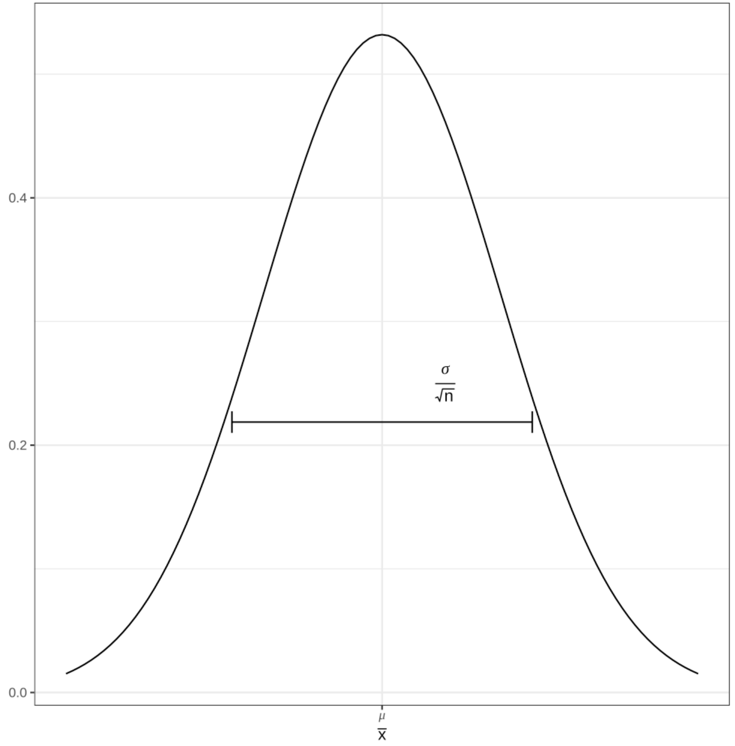
What is a p-value?

A p-value is: the probability that we observe sample statistic at least as extreme as the one we observed if the null hypothesis is true.

- p-values are increasingly controversial, even though almost everyone uses them more or less uncritically.
- p-values can be made arbitrarily small by collecting more data (though this is time/resource-intensive and often impractical or impossible).

The Idea

Theoretically, we know

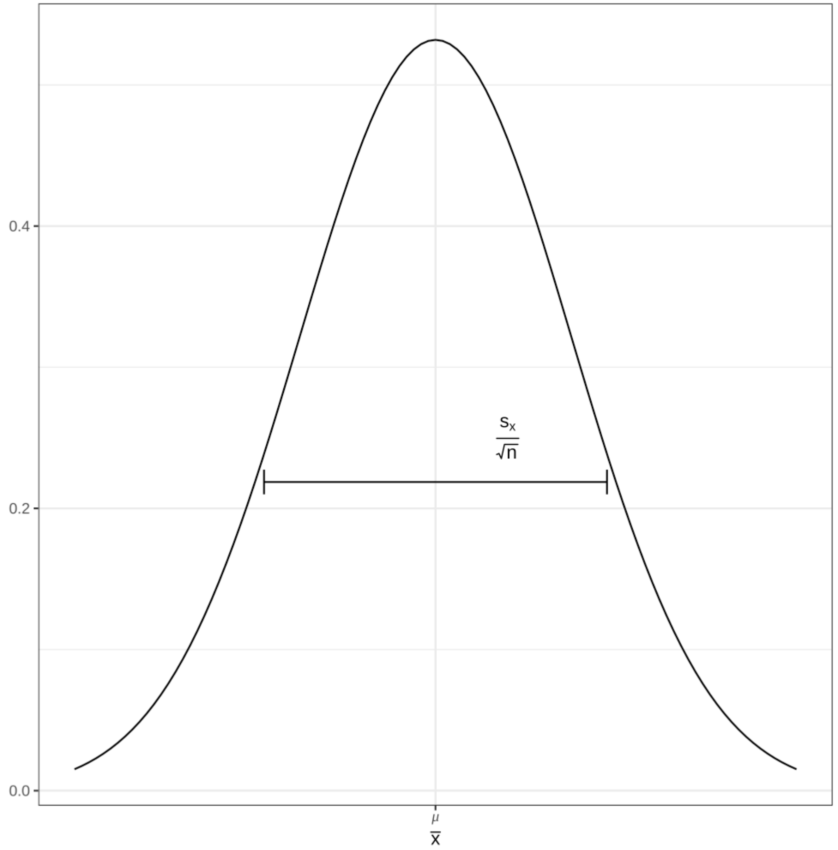


The Idea

We know two pieces of information to help us out:

- $\bar{x}$ - our sample statistic value.
- s_x - the standard deviation of x.

But, we still don't know μ.



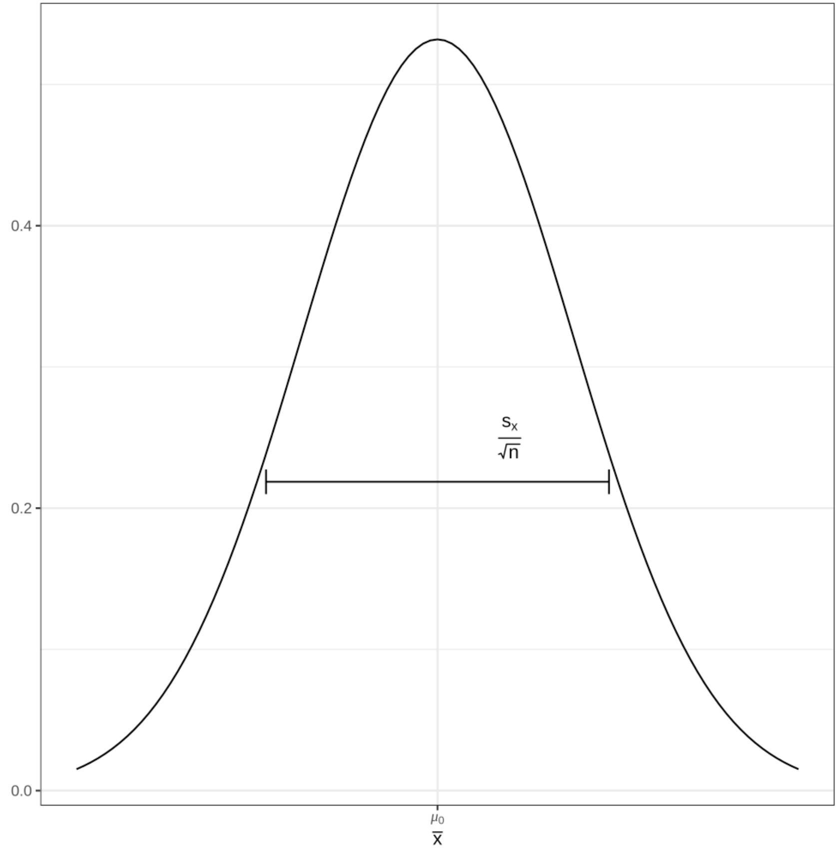
The Idea

We know two pieces of information to help us out:

- $\bar{x}$ - our sample statistic value.
- s_x - the standard deviation of x.

But, we still don't know μ.

- This is where our hypothesis comes in: μ_0



P-values

Now, we know all of the relevant pieces of this distribution. Under the null hypothesis (i.e., if the null hypothesis is true), we know that (approximately):

$$\bar{x} \sim N\left(\mu_0, \frac{s_x}{\sqrt{n}}\right) \text{ or } \bar{x} \sim t_{n-1}\left(\mu_0, \frac{s_x}{\sqrt{n}}\right)$$

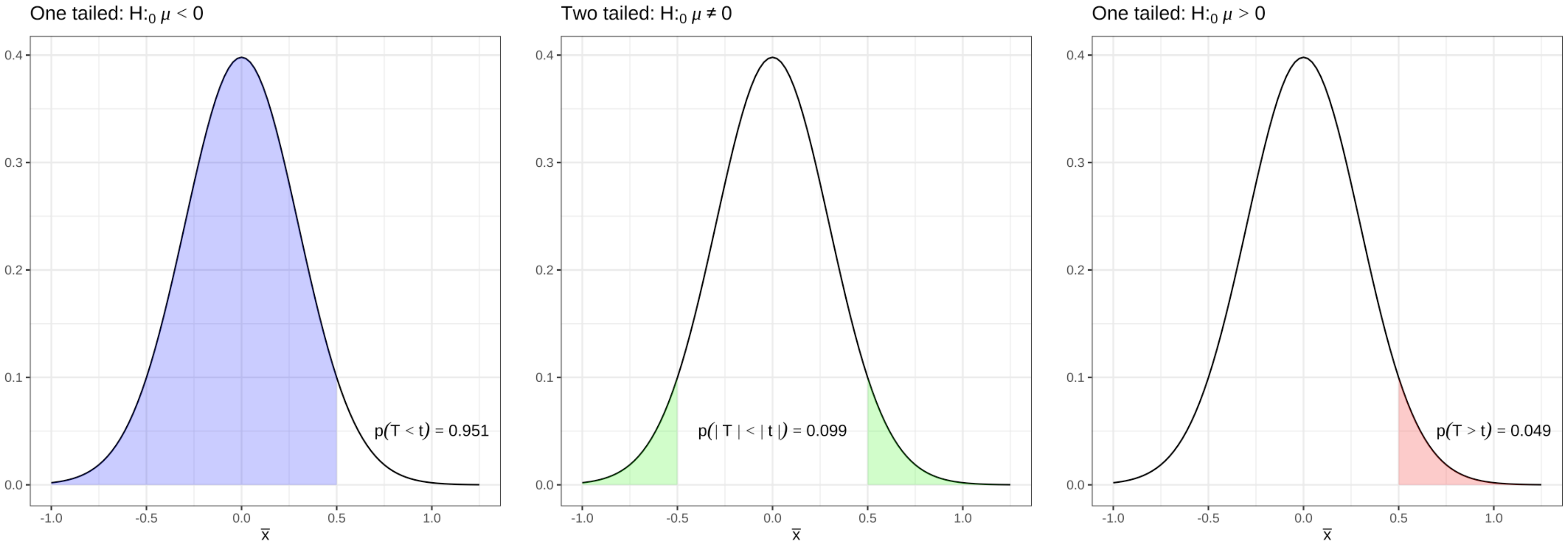
So, we can turn our sample statistic into a z-score.

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

We can then use the normal probability table to figure out what the probability is.

One and Two Tailed Tests

Let's assume $\bar{x} = 0.5$, $\mu_0 = 0$, $s_x = 3$ and $n = 100$



t-test

R Python Stata

```
set.seed(519)
x <- scale(rnorm(100, 0, 1))
x <- x*3 + .5
t.test(x, mu = 0, alternative="two")
```

```
##
##     One Sample t-test
##
## data:  x
## t = 1.6667, df = 99, p-value = 0.09874
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.09526509  1.09526509
## sample estimates:
## mean of x
##       0.5
```

One-sided

R | Python | Stata

```
t.test(x, mu = 0, alternative="less")
```

```
##
##      One Sample t-test
##
## data:  x
## t = 1.6667, df = 99, p-value = 0.9506
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##       -Inf 0.9981173
## sample estimates:
## mean of x
##       0.5
```

```
t.test(x, mu = 0, alternative="greater")
```

```
##
##      One Sample t-test
##
## data:  x
## t = 1.6667, df = 99, p-value = 0.04937
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  0.001882653         Inf
## sample estimates:
## mean of x
##       0.5
```

Hypothesis Test for Proportions

Same as a test for the mean, but

- We use the z distribution.
- We can calculate the standard error based under the null hypothesis directly (rather than estimating it) because regardless of the individual values, the standard deviation of a binary variable is $s = \sqrt{\frac{p(1-p)}{n}}$.

For the normal approximation to work, we need:

- $np \geq 5$
- $n(1-p) \geq 5$

where n is the number of observations in the sample and p is the hypothesized population proportion. If this isn't true, we need a different test.

Proportion Test Example

Let's say that we had a 250 observations on gender and that 110 were males. If we wenated to test $H_0 : p = .5$ against the two-sided alternative, we would do:

R Python Stata

```
prop.test(x=110, n=250, p=.5)
```

```
##
##      1-sample proportions test with continuity correction
##
## data:  110 out of 250, null probability 0.5
## X-squared = 3.364, df = 1, p-value = 0.06664
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3778970 0.5039775
## sample estimates:
##    p
## 0.44
```

Difference of Means

We want to make an inference about the difference between two population parameters, where generally $H_0 : \mu_1 = \mu_2$,

- $H_A : \mu_1 \neq \mu_2$; $H_A \mu_1 - \mu_2 = 0$
- $H_A : \mu_1 < \mu_2$; $H_A \mu_1 - \mu_2 < 0$
- $H_A : \mu_1 > \mu_2$; $H_A \mu_1 - \mu_2 > 0$

Just like any test, we need to make a z- or t-statistic:

$$\frac{\text{Estimate} - H_0 \text{Value}}{\text{SE}}$$

In this case:

$$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_{\bar{x}_1 - \bar{x}_2}}$$

SE of Difference

- Assume different population variances of two groups

$$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \text{ with df: } \frac{\left[\left(\frac{s_1^2}{n_1}\right)+\left(\frac{s_2^2}{n_2}\right)\right]^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

- Assume same population variance of two groups

$$s_p=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

$$s_{\bar{x}_1-\bar{x}_2}=s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \text{ with df: } n_1+n_2-2$$

Which to Choose?

If you choose equal variances and you're wrong...

- Your inferences will be wrong and potential anti-conservative.

If you choose unequal variances and you're wrong ...

- Your inferences may have higher variance than they would have otherwise, so your inferences will be a bit conservative, but this is probably better.

The default in `t.test()` and `tTest()` is to **not** assume that the variances are equal.

CES Example

R Python Stata

```r
library(rio)
library(dplyr)
library(DAMisc)
ces <- import("ces19.dta")
ces <- ces %>% mutate(
  vote_con = case_when(vote == 2 ~ 1,
                       vote %in% c(1,3,4) ~ 0,
                       TRUE ~ NA_real_))
```

```r
tTest("vote_con", "market", data=ces, var.equal=FALSE)
```

```
## Summary:
##             mean       n    se
## 0           -0.3929244 1679 0.3791567
## 1           -0.0149498 664  0.3720461
## Difference -0.3779746 2343 0.01714886
## p-value < 0.001
## --------------------------------
##
##     Welch Two Sample t-test
##
## data:  market by vote_con
## t = -22.041, df = 1237, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is
## 95 percent confidence interval:
##  -0.4116186 -0.3443305
## sample estimates:
## mean in group 0 mean in group 1
##      -0.3929244      -0.0149498
```

Proportion Test

R Python Stata

```
ces <- ces %>% mutate(
  coll_grad = case_when(educ == 3 ~ 1,
                        educ %in% 1:2 ~ 0,
                        TRUE ~ NA_real_))
s <- ces %>%
  # group by the independent variable
  group_by(coll_grad) %>%
  filter(!is.na(coll_grad) & !is.na(vote_con)) %>%
  # summarise the dependent variable
  summarise(n_con = sum(vote_con, na.rm=TRUE),
            n = n())
```

```
s
```

```
## # A tibble: 2 × 3
##   coll_grad n_con     n
##       <dbl> <dbl> <int>
## 1         0   423  1254
## 2         1   256  1117
```

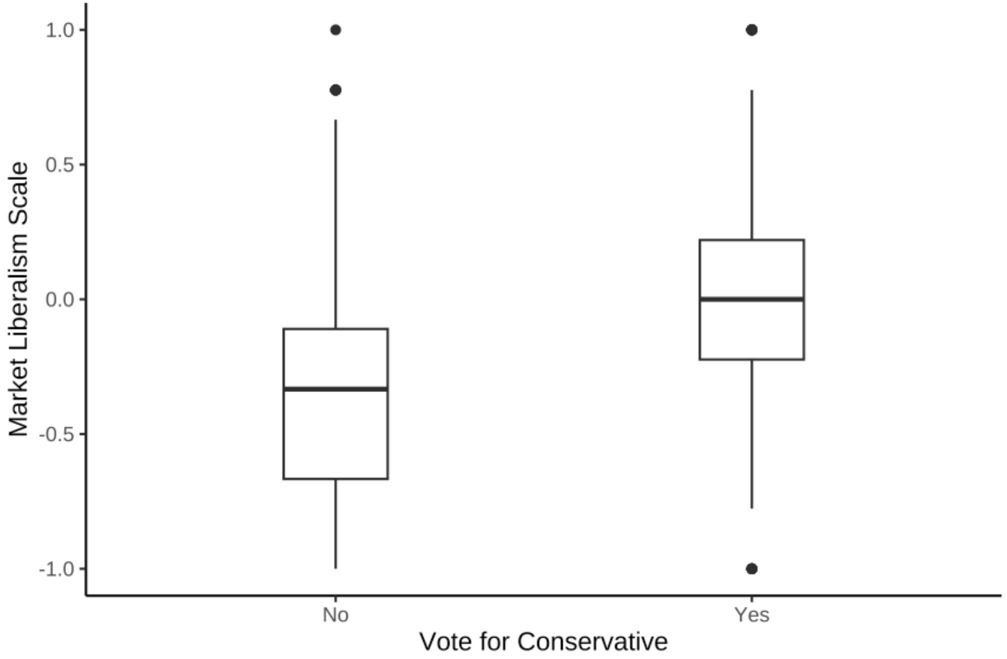
```
prop.test(s$n_con, s$n)
```

```
##
##     2-sample test for equality of proportions with continuity correctio
##
## data:  s$n_con out of s$n
## X-squared = 33.275, df = 1, p-value = 7.999e-09
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.07134009 0.14493042
## sample estimates:
##    prop 1    prop 2
## 0.3373206 0.2291853
```

Visualizing Differences: Box Plot

R Python Stata

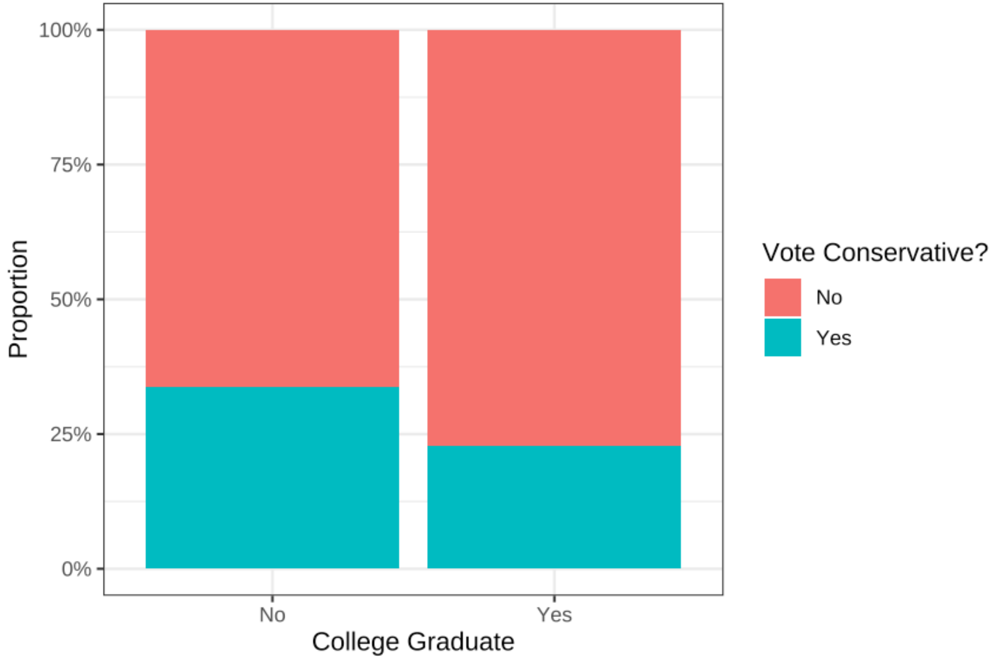
```
ces %>%
  mutate(vc = factor(vote_con,
                     levels=c(0,1),
                     labels=c("No", "Yes"))) %>%
  filter(!is.na(vc)) %>%
ggplot(aes(x=vc, y=market)) +
  geom_boxplot(width=.25) +
  theme_classic() +
  labs(x="Vote for Conservative",
       y="Market Liberalism Scale")
```



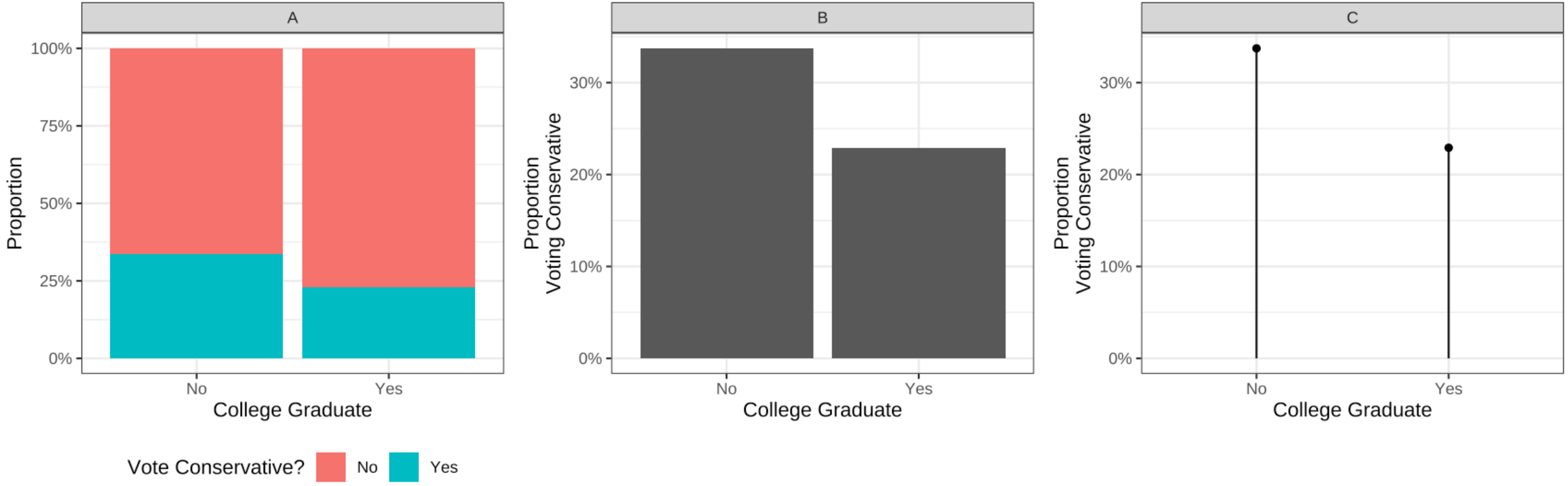
Visualizing Proportions

R   Python   Stata

```r
ces %>% select(coll_grad, vote_con) %>%
  na.omit %>%
  group_by(coll_grad, vote_con) %>%
  summarise(n = n()) %>%
  ungroup() %>%
  group_by(coll_grad) %>%
  mutate(prop = n/sum(n)) %>%
ggplot(aes(x=factor(coll_grad, labels=c("No", "Yes")),
           y=prop,
           fill=factor(vote_con, labels=c("No", "Yes")))) +
  geom_bar(stat="identity", position="stack") +
  theme_bw() +
  labs(x="College Graduate",
       y="Proportion",
       fill="Vote Conservative?") +
  scale_y_continuous(labels=scales::label_percent())
```



What's Best?


A
B
C
100%
75%
50%
25%
0%
Proportion
30%
20%
10%
0%
Proportion
Voting Conservative
No
Yes
College Graduate
Vote Conservative?
No
Yes

Exercises

Using the **ces** data, answer the following questions.

1. Do people who identify with a religion higher feeling thermometer scores for the conservative candidate?
   - Is there a difference between Catholics and Non-Catholic Christians?
   - For each of the results above, make the appropriate graph.
2. Are middle-aged people more likely to turn out to vote than older and younger people?
   - What if you just look at the difference between the 18-34 and 35-54 groups?
   - For each of the results above, make the appropriate graph.