Comparing proportions & odds

Dr. Maria Tackett

10.21.19

[Click for PDF of slides]()

Announcements

- Lab 06 due **Tuesday, Oct 22 at 11:59p**
- Reading 07 for Wednesday
- Project Proposal **due Wed, Oct 30 at 11:59p**
- Datathon **November 2 - 3**

Packages

```r
library(tidyverse)
library(knitr)
library(broom)
library(fivethirtyeight)
```

Modeling Binary Outcomes

FiveThirtyEight March Madness

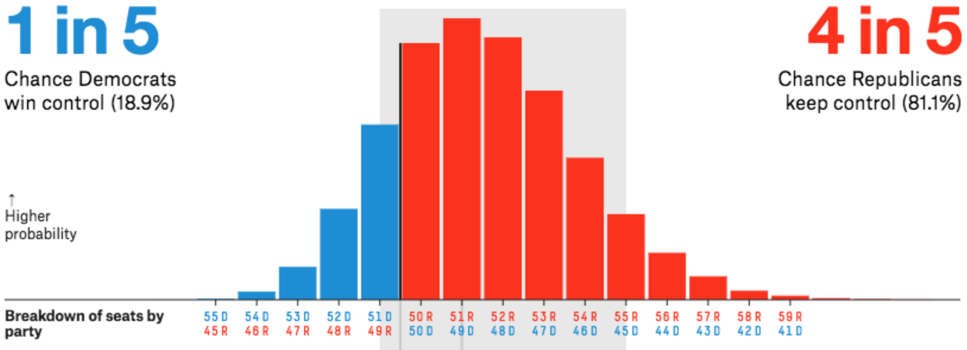
2018 March Madness Predictions

*Live Win Probabilities are "derived using **logistic regression analysis**, which lets us plug the current state of a game into a model to produce the probability that either team will win the game."*

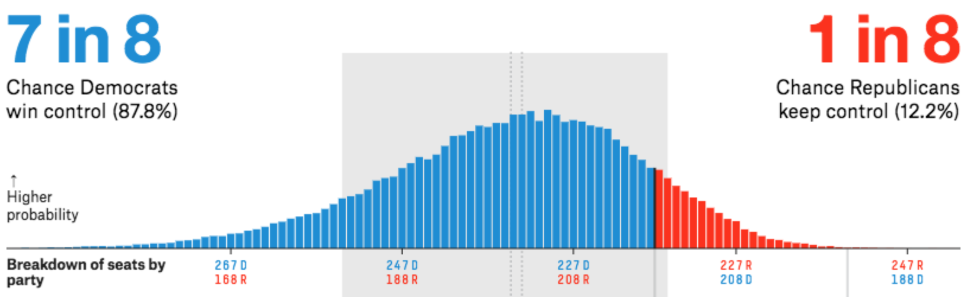
-"How Our March Madness Predictions Work"

STA 210

2018 Election Forecasts



[FiveThirtyEight.com Senate forecast]



[FiveThirtyEight.com House forecast]

Our models are probabilistic in nature; we do a lot of thinking about these probabilities, and the goal is to develop ***probabilistic estimates*** *that hold up well under real-world conditions.*

[-"How FiveThirtyEight's House, Senate, and Governor Models Work"]

STA 210

Is it rude to recline your seat on a plane?

```
flying <- flying %>%
  filter(!is.na(recline_rude)) %>%
  mutate(rude = if_else(recline_rude %in%
                        c("Somewhat", "Very"), 1, 0))
```

Source: *41 Percent of Fliers Think You're Rude If You Recline Your Seat*

Response Variable, Y

- Y is a binary response variable
    - 1: yes
    - 0: no
- $Mean(Y) = p$
    - p is the proportion of "yes" responses in the population
- $Variance(Y) = p(1 - p)$

STA 210

Sampling Distribution of Sample Proportion

- $\hat{p}$: average of binary responses in the sample
  - Called the sample proportion
  - This is the statistic, i.e. the estimate of p
- Given $\hat{p}$ is the sample proportion based on a sample of size n from a population with population proportion p:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

...assuming n is "large" (more than 5 "yes" and 5 "no")

STA 210

Confidence Interval for a Single Proportion

- Approximate C% confidence interval for p is

$$
\begin{aligned}
&\hat{p} \pm z^* SE(\hat{p}) \\
=&\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\end{aligned}
$$

where z^* is the critical value calculated from the $N(0, 1)$ distribution

```
# Critical value for 90% CI
qnorm(0.95)
```

```
## [1] 1.644854
```

STA 210

Opinions about reclining: 90% CI

```
crit.val <- qnorm(0.95)
```

```
flying %>%
  summarise(n = n(),
            p_hat = sum(rude)/n,
            se = sqrt(p_hat*(1-p_hat)/n),
            lb = p_hat - crit.val*se,
            ub = p_hat + crit.val*se)
```

```
## # A tibble: 1 x 5
##       n p_hat     se    lb    ub
##   <int> <dbl>  <dbl> <dbl> <dbl>
## 1   854 0.412 0.0168 0.384 0.440
```

We are 90% confident that the interval 0.384 to 0.44 contains the true proportion of fliers who think reclining your seat on a plane is rude.

Sampling Distribution for Difference in Two Proportions

- Let $\hat{p}_1$ and $\hat{p}_2$ be sample proportions from independent random samples of size n_1 and n_2, respectively:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

... assuming n_1 and n_2 are "large" (at least 5 "yes" and "no" in each sample)

Confidence Interval for Difference in Proportions

- Approximate C% confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE(\hat{p}_1 - \hat{p}_2)$$

$$= \; (\hat{p}_1 - \hat{p}_2) \pm z^* \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where z^* is the critical value calculated from the $N(0, 1)$ distribution

STA 210

Opinions about reclining by age

```
flying %>%
  filter(age %in% c("18-29", "30-44")) %>%
  group_by(age, rude) %>%
  summarise(n = n()) %>%
  spread(rude, n) %>%
  kable(format="markdown")
```

age	0	1
18-29	78	94
30-44	143	79

Is there a significant difference in the proportion of 18-29 year olds versus 30-44 year olds who think reclining a seat on a plane is rude?

Opinions about reclining by age: 90% CI

```
flying %>%
  filter(age %in% c("18-29", "30-44")) %>%
  group_by(age) %>%
  summarise(n = n(),
            p_hat = round(sum(rude)/n,3)) %>% kable(format="markdo
```

age	n	p_hat
18-29	172	0.547
30-44	222	0.356

1. Calculate a 90% confidence interval for the difference in proportion of 18-29 year olds and 30-44 year olds who think reclining a seat on a plane is rude. Interpret the interval.

2. Based on the interval, is there evidence of a significant difference in proportions between the two groups?

STA 210

What are some potential difficulties with reporting results using the difference in proportions? or proportions/percentages in general?

Odds

- Given p, the population proportion of "yes" responses, the corresponding **odds** of a "yes" response is

$$\omega = \frac{p}{1-p}$$

- The *sample odds* are $\hat{\omega} = \frac{\hat{p}}{1-\hat{p}}$

- **Ex.**
  - proportion of fliers who think reclining is rude: 0.412.
  - odds a flier thinking reclining is rude: 0.701 to 1

Properties of the odds

- odds ≥ 0
- If $\hat{p} = 0.5$, then odds $= 1$
- If odds of "yes" $= \omega$, then the odds of "no" $= \frac{1}{\omega}$
- If odds of "yes" $= \omega$, then $\hat{p} = \frac{\omega}{(1+\omega)}$

STA 210

Odds ratio

- Suppose we have two populations with proportions p_1 and p_2 and odds ω_1 and ω_2
- The odds ratio is $\phi = \frac{\omega_1}{\omega_2}$
  - *Estimate*: $\hat{\phi} = \frac{\hat{\omega}_1}{\hat{\omega}_2}$
- Good alternative to the difference in proportions
- **Intepretation:** The odds of "yes" in group 1 is ϕ times the odds of "yes" in group 2

Why use Odds Ratio?

- In practice, the odds ratio is more consistent across levels of confounding variables
- The odds ratio is more easily interpreted / understood
- The odds ratio can be easily extended to regression analysis

Sampling distribution of log(odds ratio)

- Let $\hat{\omega}_1$ and $\hat{\omega}_2$ be sample odds from independent random samples of size n_1 and n_2, respectively:

$$\log(\hat{\phi}) = \log\left(\frac{\hat{\omega}_1}{\hat{\omega}_2}\right) \approx N\left(\log\left(\frac{\omega_1}{\omega_2}\right), \frac{1}{n_1 p_1(1-p_1)} + \frac{1}{n_2 p_2(1-p_2)}\right)$$

... assuming n_1 and n_2 are "large" based on the thresholds for difference in proportions

STA 210

Confidence Interval for Log Odds Ratio

- Approximate C% confidence interval for $\log(\phi)$ is

$$\log(\hat{\phi}) \pm z^* \times SE[\log(\hat{\phi})]$$

$$= \log(\hat{\phi}) \pm z^* \times \sqrt{\frac{1}{n_1\hat{p}_1(1-\hat{p}_1)} + \frac{1}{n_2\hat{p}_2(1-\hat{p}_2)}}$$

where z^* is the critical value calculated from the $N(0,1)$ distribution

STA 210

Confidence Interval for Odds Ratio

Suppose LB and UB are the lower and upper bounds of the C% confidence interval for the log odds ratio, $\log(\phi)$

The C% confidence interval for the odds ratio, ϕ is

$$\exp\{LB\} \text{ to } \exp\{UB\}$$

Opinions about reclining seat

```
flying %>%
  filter(age %in% c("18-29", "30-44")) %>%
  group_by(age) %>%
  summarise(n = n(),
            p_hat = round(sum(rude)/n,3),
            odds = round(p_hat/(1-p_hat),3))
```

```
## # A tibble: 2 x 4
##   age        n p_hat  odds
##   <ord> <int> <dbl> <dbl>
## 1 18-29   172 0.547 1.21
## 2 30-44   222 0.356 0.553
```

1. Calculate a 90% confidence interval for the odds ratio of 18-29 versus 30-44 year olds who think reclining a seat on a plane is rude. Interpret the interval.

2. Based on the interval, is there evidence of a significant difference in the odds between the two groups?

Hypothesis Test for Odds Ratio

- We want to test whether two groups have equal odds, i.e. $\phi = \frac{\omega_1}{\omega_2} = 1$
- Null Hypothesis: $H_0 : \log(\phi) = \log\left(\frac{\omega_1}{\omega_2}\right) = 0$
- Test Statistic:

$$z = \frac{\log(\hat{\phi}) - 0}{SE_0[\log(\hat{\phi})]} = \frac{\log(\hat{\phi}) - 0}{\sqrt{\frac{1}{n_1\hat{p}_c(1-\hat{p}_c)} + \frac{1}{n_2\hat{p}_c(1-\hat{p}_c)}}}$$

- p-value: proportion of $N(0, 1)$ distribution as extreme or more extreme than the test statistic

STA 210

Standard error $SE_0[\log(\hat{\phi})]$

- The null hypothesis is that odds ratio is 1, i.e. the proportions are equal

- To calculate standard error, we estimate $\hat{\pi}_c$, the sample proportion from the combined data

$$SE_0[\log(\hat{\phi})] = SE_0\left[\log\left(\frac{\hat{\omega}_1}{\hat{\omega}_2}\right)\right] = \sqrt{\frac{1}{n_1\hat{p}_c(1-\hat{p}_c)} + \frac{1}{n_2\hat{p}_c(1-\hat{p}_c)}}$$

STA 210

Opinions about reclining seat

Do the odds of thinking it's rude to recline a seat on a plane differ between 18-29 and 30-44 year olds?

$$H_0 : \log(\phi) = 0$$
$$H_a : \log(\phi) \neq 0$$

- $\hat{p}_c$ = 0.439
- **18 - 29**: $n = 172$, $\hat{\omega} = 1.208$
- **30 - 44**: $n = 222$, $\hat{\omega} = 0.553$

1. Calculate the test statistic.
2. Calculate p-value and make a conclusion.