STAT011 Statistical Methods I

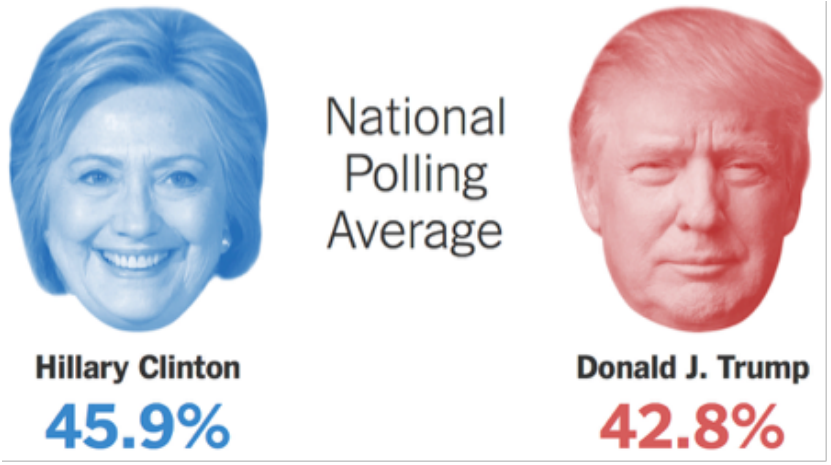
Lecture 19 Inference for Two Proportions

Lu Chen
Swarthmore College
4/4/2019

Review

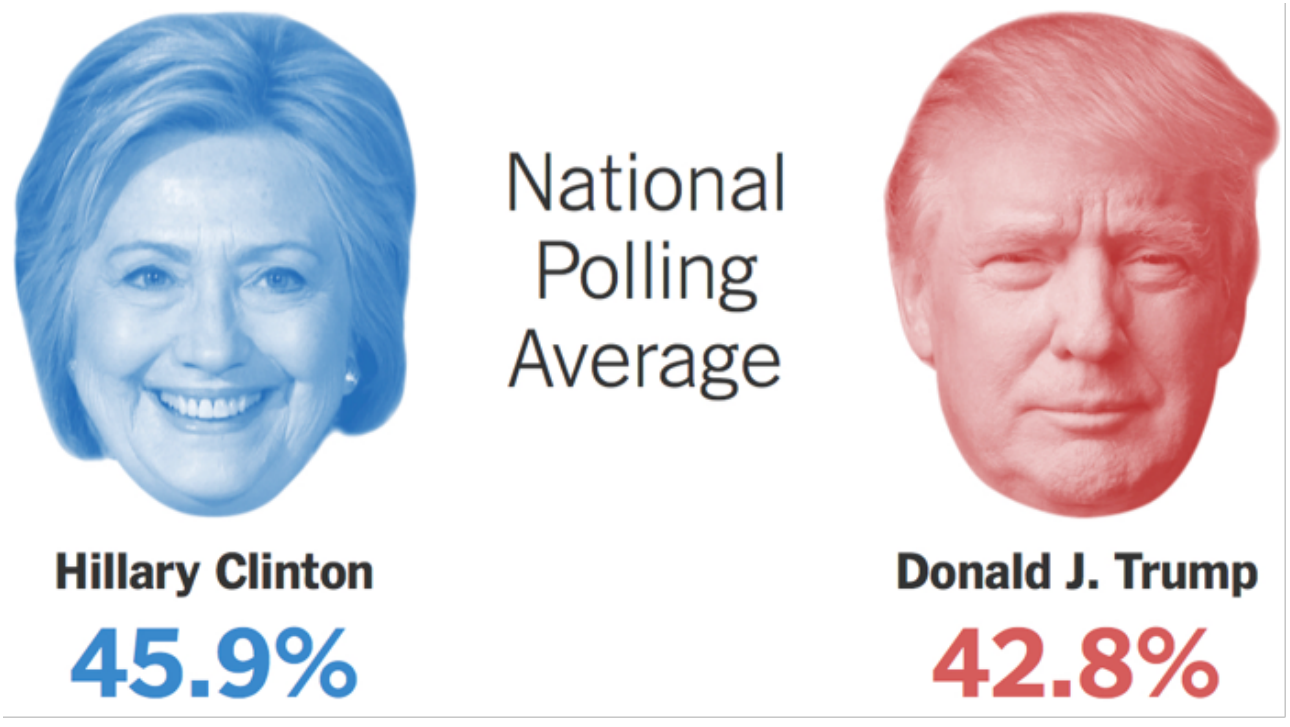
- Motivation example: an analysis of many statistical analyses
- Data
- Bernoulli distribution $X \sim Bernoulli(p)$
- Sampling distribution of a sample proportion $\hat{p} \overset{approx.}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$
- Inference for a population proportion
  - p is unknown…
  - Large sample C.I. for a population proportion $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - Large sample z test for a population proportion $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{approx.}{\sim} N(0, 1)$
- Examples

Outline



- Motivation example
- Two-proportion problem
- Large sample confidence interval for a difference in proportions
- Large sample significance test for a difference in proportions
- Examples

Election polls



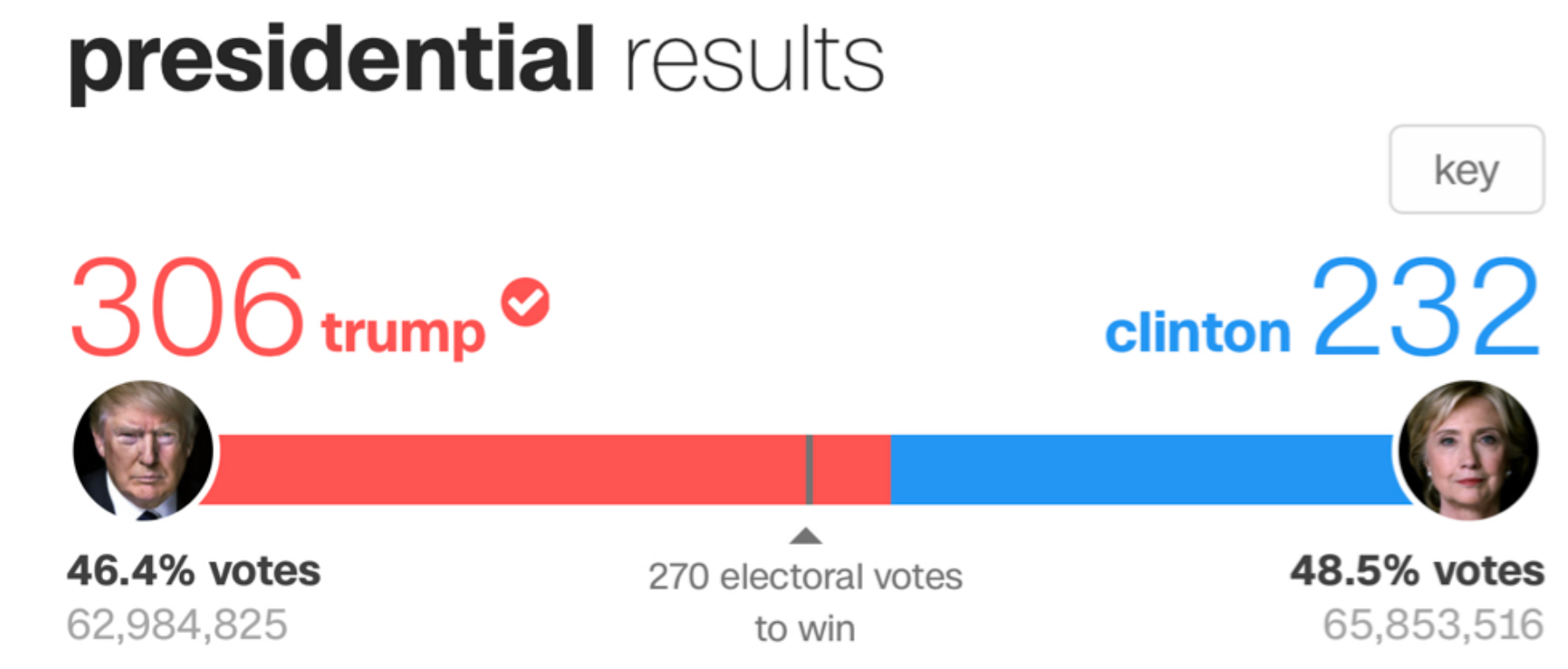
- 11/8/2016 NY Times
- http://www.nytimes.com/interactive/2016/us/elections/polls.html

Election polls

Polls	Dates	Type, Respondents	Clinton	Trump	Margin
YouGov/Economist	11/4 - 11/7	Online3,669	45	41	Clinton +4
IBD/TIPP NEW	11/4 - 11/7	Live Phone1,107	41	43	Trump +2
Insights West NEW	11/4 - 11/7	Online940	45	41	Clinton +4
Bloomberg/Selzer	11/4 - 11/6	Live Phone799	46	43	Clinton +3
Lucid/The Times-Picayune	11/4 - 11/6	Online931	45	40	Clinton +5

- 376 total polls from January 4th to November 7th
- A total of 1,001,325 subjects
  - 471,126 (47.1%) for Clinton
  - 414,797 (41.4%) for Trump
- "Polls conducted more recently and polls with a larger sample size are given greater weight in computing the averages" - NY Times

Popular vote



2/16/2017 CNN

Popular vote versus 376 polls

	Total	Clinton	Trump
Popular vote	128,838,341	65,853,516 (48.5%)	62,984,825 (46.4%)
376 polls	1,001,325	471,126 (47.1%)	414,797 (41.4%)

The Atlantic

> Clinton is ahead in the popular vote totals, meaning that polls showing her ahead by a few points in head-to-head matchups with Trump were wrong in magnitude, but not directionality.

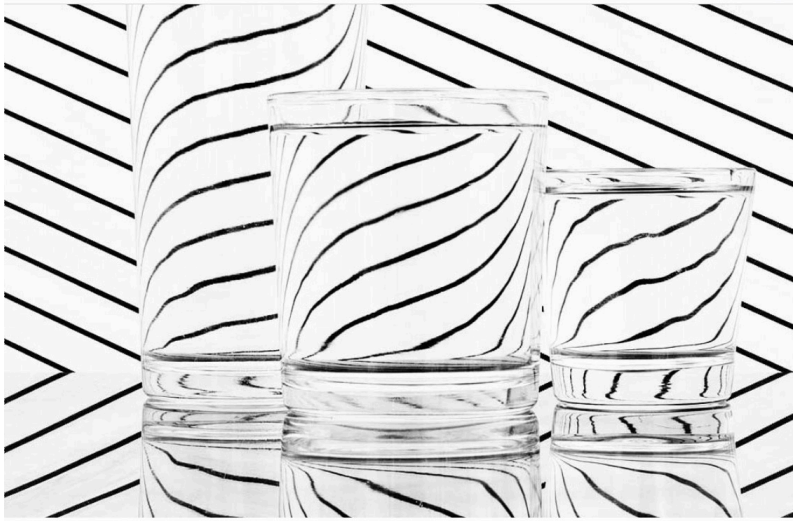
Polls?

The Atlantic AA

Rick Rycroft / AP

What Went Wrong With the 2016 Polls?

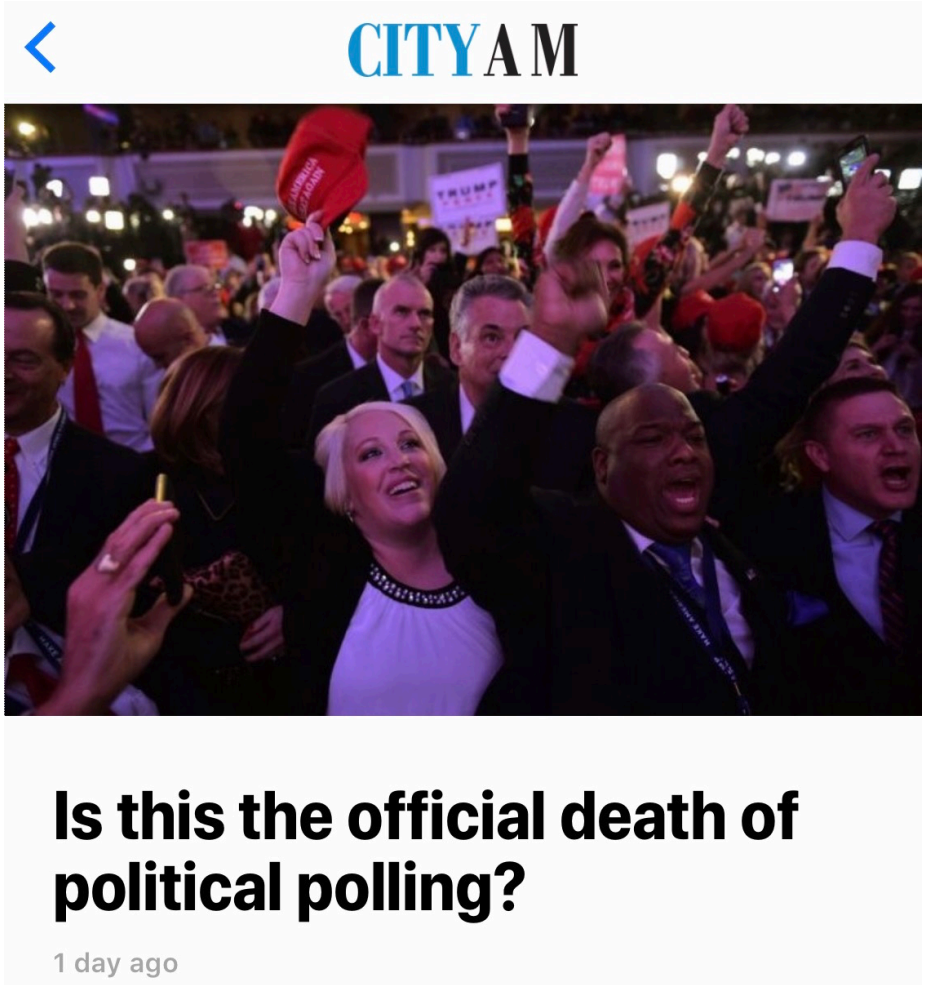
HARVARD BUSINESS REVIEW



Why Pollsters Were Completely and Utterly Wrong

1 day ago

Polls?


CITY A M
Is this the official death of political polling?
1 day ago

Popular vote versus 376 polls

	Total	Clinton	Trump
Popular vote	128,838,341	65,853,516 (48.5%)	62,984,825 (46.4%)
376 polls	1,001,325	471,126 (47.1%)	414,797 (41.4%)

- Polls from conservative media had similar results.
- Polls from Trump Camp had similar results.
- Why did ALL these polls over-estiamte the difference between the two?

Possible reasons

- Weighting: weights were chosen by pollers and usually not reported
- "Shy Trump" effect
- Low-response rate (< 15%): introduces more errors and self-selection bias
- Data-collection methods - sampling bias

  online

  live telephone

  interactive voice response

Did we all believe Clinton would win because of bad data, or did we ignore bad data because we believed Clinton would win?

How different polls work

NY Times polling notes

- **Online Polls**
  - Most online polls are based on panels of self-selected respondents. Internet access is not yet evenly distributed across socioeconomic and demographic groups.
- **Live Telephone Polls**
  - An interviewer asks questions of a respondent by telephone. … about half of U.S. households do not have a landline.
- **Interactive Voice Response Polls (I.V.R.)**
  - I.V.R. employ an automated, recorded voice to call respondents … Most I.V.R. polls call only landlines.

376 polls

Number of polls by polling methods and results

	Online	Live Phone	I.V.R.
Clinton led	188	126	27
Trump led or even	8	14	13
Total	**196**	**140**	**40**

- Is there a relationship between results and polling method?
  Do the three polling methods have different results?
- Conditional distribution
  Proportion of Clinton led conditional on polling method
  Proportion of Trump led or even conditional on polling method

376 polls

```
head(polls, 3) # look at the first 3 rows of the polls data
```

```
##        Result Method
## 1 Clinton led Online
## 2 Clinton led Online
## 3 Clinton led Online
```

```
table.polls <- table(polls); table.polls
```

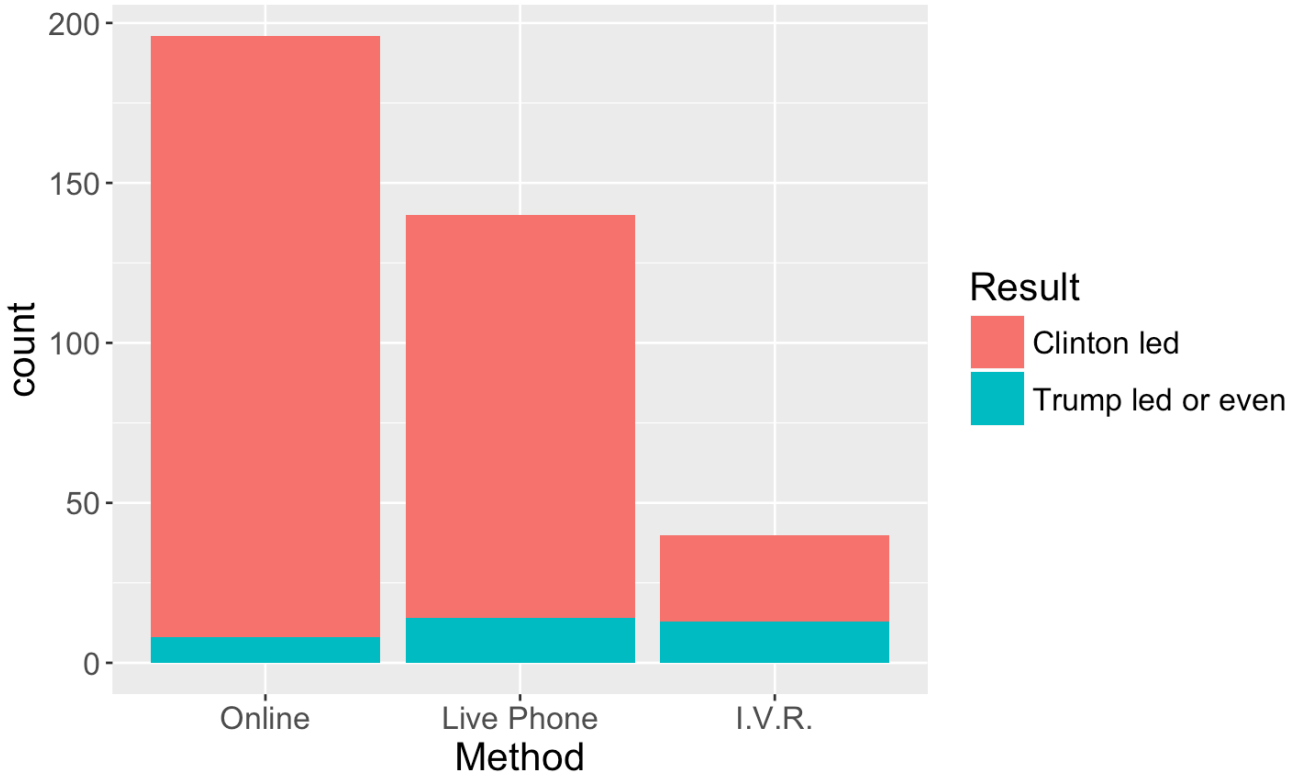
```
##                    Method
## Result              Online Live Phone I.V.R.
##   Clinton led          188        126     27
##   Trump led or even      8         14     13
```

```
prop.table(table.polls, margin = 2) ## Conditional distribution by columns
```

```
##                    Method
## Result                  Online Live Phone     I.V.R.
##   Clinton led       0.95918367 0.90000000 0.67500000
##   Trump led or even 0.04081633 0.10000000 0.32500000
```

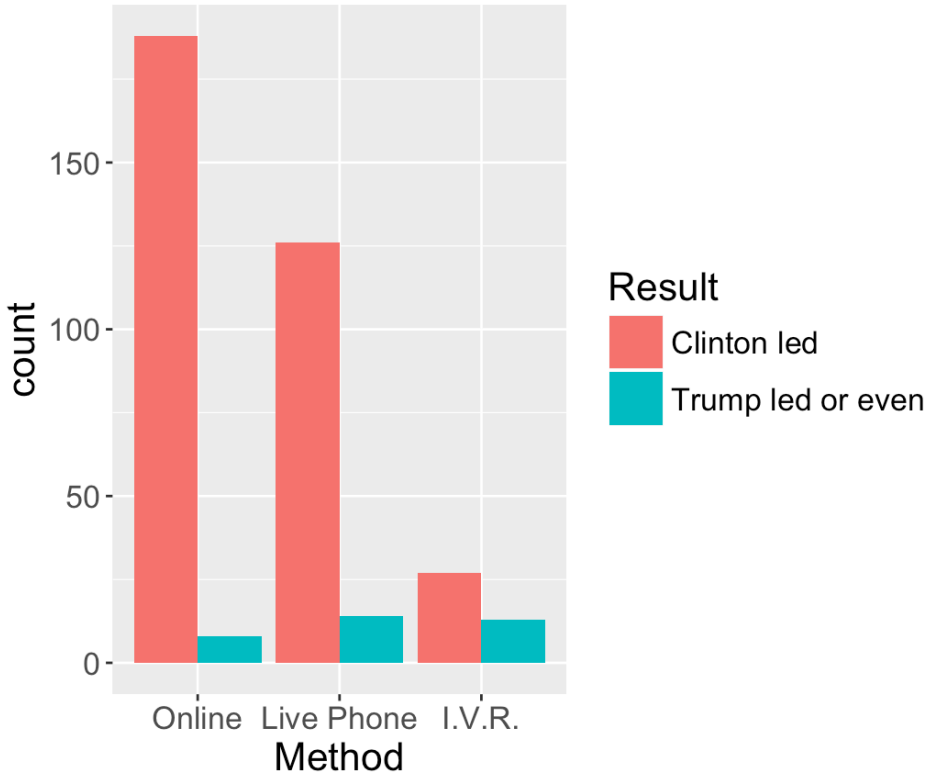
376 polls

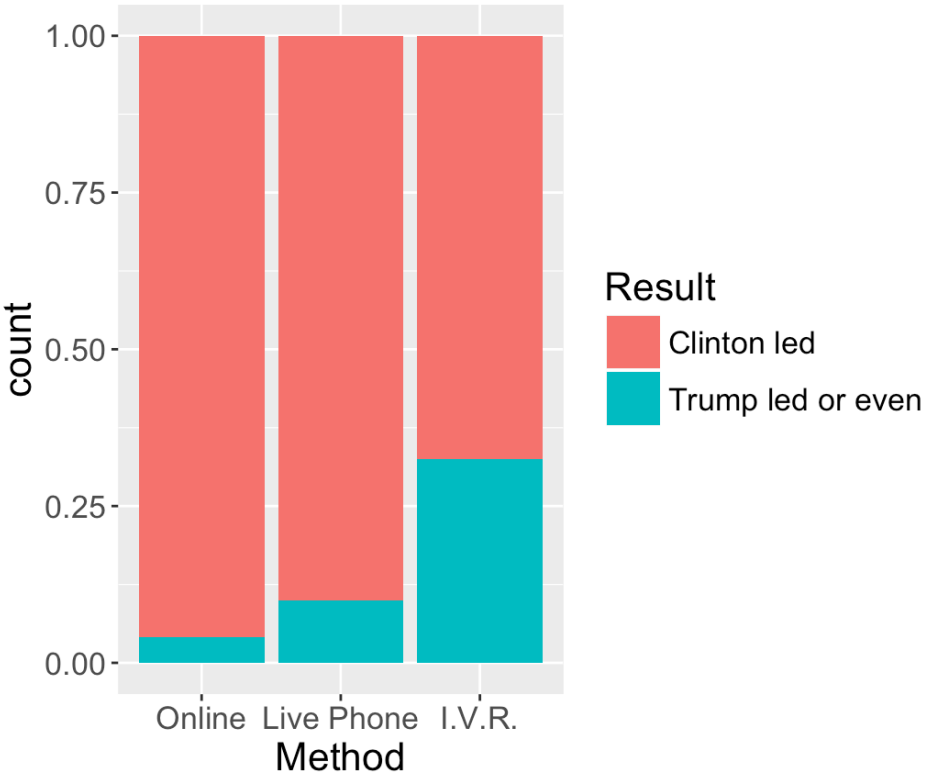
```
library(ggplot2)
theme_update(text=element_text(size=16))
ggplot(polls,aes(Method, fill=Result))+geom_bar()
```



376 polls

```
gp <- ggplot(polls,aes(Method, fill=Result))
gp+geom_bar(position="dodge") # side by side
gp+geom_bar(position="fill") # conditional proportions
```





376 polls

	Online	Live Phone	I.V.R.
Clinton led	188	126	27
Trump led or even	8	14	13
Total	**196**	**140**	**40**
Conditional proportions of Clinton led	0.959	0.900	0.675

- Are the proportions of Clinton-led the same for different polling methods?
  - Let's do pairwise comparisons today. In the next lecture, we will learn the method for testing whether three or more proportions are the same.
- Inference for two proportions p_1 and p_2:
  - Confidence interval for $p_1 - p_2$
  - Significance test $H_0 : p_1 = p_2$

Inference for two proportions

$$\hat{p}_1 \overset{approx.}{\sim} N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right) \text{ and } \hat{p}_2 \overset{approx.}{\sim} N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right)$$

- Estimate for $p_1 - p_2$: $\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$
- Mean of $\hat{p}_1 - \hat{p}_2$: $p_1 - p_2$
- SD of $\hat{p}_1 - \hat{p}_2$: $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
- Distribution of $\hat{p}_1 - \hat{p}_2$

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

Two-proportion vs. one-proportion problems

	One-proportion	Two-proportion
Data	Binary	Binary vs. binary
Parameter of interest	p	p_1, p_2
Estimate	$\hat{p} = \frac{X}{n}$	$\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$
Mean	p	$p_1 - p_2$
SD	$\sqrt{\frac{p(1-p)}{n}}$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
Distribution	$\hat{p} \overset{approx.}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$	$(\hat{p}_1 - \hat{p}_2) \overset{approx.}{\sim} N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$

Inference for two proportions

Since p_1 and p_2 are unknown,

- For confidence interval,

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

- For testing $H_0 : p_1 = p_2$, denote $\hat{p}$ as the pooled estimator for $p = p_1 = p_2$.

$$SE_{\hat{p}_1-\hat{p}_2} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$

Large sample CI for difference in proportions

Choose two independent SRSs of size n_1 and n_2 from a large population having proportions p_1 and p_2 of successes. The estimate of the difference in the population proportions is $D = \hat{p}_1 - \hat{p}_2$. The **standard error of D** is

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

and the **margin of error** for confidence level C is $m = z^* \text{SE}_D$, where the critical value z^* is the value for the standard Normal density curve with area C between $-z^*$ and z^*. An **approximate level C confidence interval for $p_1 - p_2$** is

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

for number of successes/failures in each sample being at least 10.

Large sample test for difference in proportions

To test the hypothesis $H_0 : p_1 = p_2$, compute the z **statistic**

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where the **pooled estimator** of p_1 and p_2 is $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$. In terms of a standard Normal random variable Z, the approximate P-value for a test of H_0 against

$$H_a : p_1 > p_2 \text{ is } P(Z \geq z)$$
$$H_a : p_1 < p_2 \text{ is } P(Z \leq z)$$
$$H_a : p_1 \neq p_2 \text{ is } 2P(Z \geq |z|)$$

As a general rule, we will use the z test when the number of successes and the number of failures in each of the samples are at least 5.

376 polls

	Online	Live Phone	I.V.R.
Clinton led	188	126	27
Trump led or even	8	14	13
Total	**196**	**140**	**40**
Conditional proportions of Clinton led	0.959	0.900	0.675

- $\hat{p}_1 = 0.959, n_1 = 196, \hat{p}_2 = 0.900, n_2 = 140$
- **95% confidence interval** for $p_1 - p_2$:

$$(0.959 - 0.900) \pm 1.96\sqrt{\frac{0.959\times(1-0.959)}{196} + \frac{0.9\times(1-0.9)}{140}} = 0.059 \pm 0.057$$

- We are 95% confident that the true difference in proportion of Clintion led between the two methods (online vs live phone) is between 0.002 and 0.116.

376 polls

	Online	Live Phone	I.V.R.
Clinton led	188	126	27
Trump led or even	8	14	13
Total	**196**	**140**	**40**
Conditional proportions of Clinton led	0.959	0.900	0.675

- $\hat{p}_1 = 0.959, n_1 = 196, \hat{p}_2 = 0.900, n_2 = 140$
- **Level 0.05 test** for $H_0 : p_1 = p_2$ vs. $H_a : p_1 \neq p_2$:

$$\hat{p} = \frac{188+126}{196+140} = 0.935 \text{ and } z = \frac{(0.959-0.900)-0}{\sqrt{0.935\times(1-0.935)\times(\frac{1}{196}+\frac{1}{140})}} = 2.16$$

- $P = 0.031 < 0.05$, `2*(1-pnorm(2.16))`. We reject H_0 at level 0.05. There is significant difference in proportion of Clintion led between the two methods (online vs live phone).

R function

```r
prop.test(x = c(188, 126), n = c(196, 140), correct = FALSE) # Online vs. Live Phone
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(188, 126) out of c(196, 140)
## X-squared = 4.6749, df = 1, p-value = 0.03061
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.002290546 0.116076801
## sample estimates:
##    prop 1    prop 2
## 0.9591837 0.9000000
```

- 95% CI: $[0.002, 0.116]$
- 0.05 test: $z = \sqrt{4.67} = 2.16$ and $P = 0.031 < 0.05$
- Note here the "X-squared" value is the square of the z statistic; the former is always positive while z could be positive or negative.

R function

```r
prop.test(x = c(188, 27), n = c(196, 40), correct = FALSE) # Online vs. I.V.R.
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(188, 27) out of c(196, 40)
## X-squared = 33.095, df = 1, p-value = 8.774e-09
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1364159 0.4319515
## sample estimates:
##    prop 1    prop 2
## 0.9591837 0.6750000
```

- 95% CI: $[0.136, 0.432]$ does not contain 0.
- 0.05 test: $z = \sqrt{33.1} = 5.75$ and $P = 8.8 \times 10^{-9} < 0.05$

R function

```r
prop.test(x = c(126, 27), n = c(140, 40), correct = FALSE) # Live Phone vs. I.V.R.
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(126, 27) out of c(140, 40)
## X-squared = 12.353, df = 1, p-value = 0.0004403
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.07158061 0.37841939
## sample estimates:
## prop 1 prop 2
##  0.900  0.675
```

- 95% CI: $[0.072, 0.378]$
- 0.05 test: $z = \sqrt{12.35} = 3.51$ and $P = 0.0004 < 0.05$

376 polls

Three comparisons

	Online vs LivePhone	Online vs I.V.R.	LivePhone vs I.V.R.
Difference	0.059	0.284	0.225
Confidence interval	[0.002, 0.116]	[0.136, 0.432]	[0.072, 0.378]
Test *P*-value	0.031	8.8×10^{-9}	0.0004

- The three polling methods are significantly different from each other. Online and I.V.R. have the largest difference.

Summary

- Motivation example
- Two-proportion problem
- Large sample confidence interval for a difference in proportions

  $\hat{p}_1 - \hat{p}_2 \pm z^* \text{SE}_D$ where $\text{SE}_D = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
- Large sample significance test for a difference in proportions

  $z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\text{SE}_{Dp}}$ where $\text{SE}_{Dp} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ and $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$
- Examples

Midterm 2

- **Tuesday 4/16 during class time**
  - Review class on Thursday 4/11
  - Practice problems available on Friday 4/12
- Homework 9
  - Covers Lecture 18, 19 and 20
  - Due on Friday 4/12 11:59 pm
  - Solutions available on Saturday 4/13
- There will be office hours on Thursday 4/11 2:40 pm - 3:40 pm