Comparing proportions across group and Error bars

Te Rutherford

February 3, 2015

Today

- Interpretation of confidence interval
- Inference with confidence interval
- Comparing rates/proportions from multiple groups
- Error bars

Problem 7.54 in the handout

- The proportion of people satisfy MSDS $= 11/100$
- 150 people have been surveyed.
- First step: we want to create a data vector using `rep` command. There are 150 * 11/100 successes and 150 * 89/100 failures.

```
msds_data = c(rep(0, 134), rep(1, 16))
```

Find 95% confidence interval of the proportion

```
library(bootstrap)
b= bcanon(msds_data, nboot= 20000, theta=mean,
          alpha = c(0.025, 0.975))
b$confpoints
```

```
##       alpha bca point
## [1,] 0.025      0.06
## [2,] 0.975      0.16
```

- By putting in `alpha = c(0.025, 0.975)`, we only get the percentiles that we want.
- If you don't put this in, you might not get the percentiles you want.

Interpretation of Confidence Interval

- You should report the estimated proportion along with the confidence interval.
  - The estimated proportion of people who satisfy MSDS is 0.10 (0.06, 0.16).
  - The estimated proportion of people who satisfy MSDS is 0.10 ($\pm$0.06).
- Theoretical interpretation
  - If we repeat this study many times and draw many 95% confidence intervals, 95% of the intervals will cover the true proportion.
- Practical interpretation
  - We are 95% confident that the true proportion is in the interval of (0.06, 0.16).

Effects of sample size

- In general, as we increase sample size, the confidence interval becomes smaller.

```
library(bootstrap)
msds_data = c(rep(0, 134), rep(1, 16))
b= bcanon(msds_data, nboot= 20000, theta=mean,
          alpha = c(0.025, 0.975))
b$confpoints
```

```
##       alpha bca point
## [1,] 0.025    0.0600
## [2,] 0.975    0.1533
```

```
library(bootstrap)
msds_data = c(rep(0, 1340), rep(1, 160))
b= bcanon(msds_data, nboot= 20000, theta=mean,
          alpha = c(0.025, 0.975))
b$confpoints
```

Inference with confidence interval

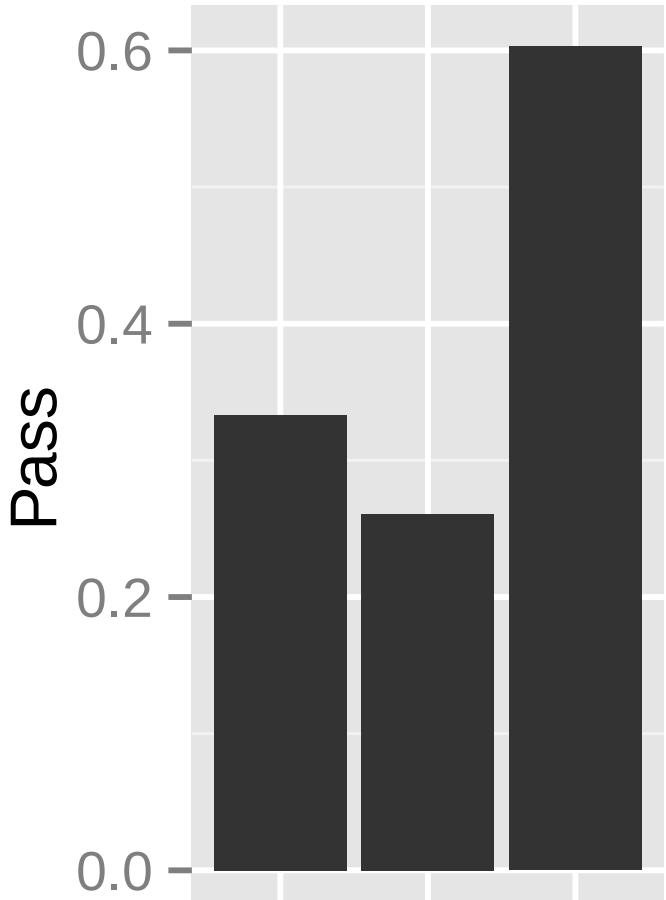
- We are 95% confident that the true proportion is not anything outside the interval of (0.06, 0.16)
- The true proportion is *significantly different* from 0.05.

What if we want to compare the proportions from two groups?

- New command for plotting the summary statistics (without `ddply`)

Plot

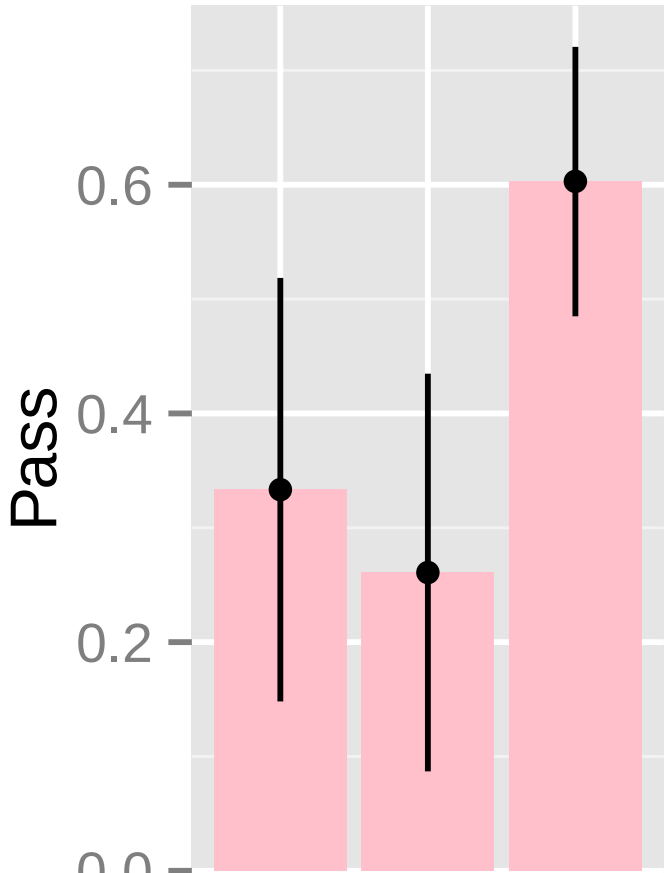
```
library(ggplot2)
data = read.csv(
  'http://www.typ-stats.com/datasets/RicciData.csv')
ggplot(data=data,mapping=aes(x=Race, y=Pass)) +
  stat_summary(fun.y=mean, geom='bar')
```



Error bar

Draw the bars first and then draw the confidence intervals (error bars).

```
ggplot(data=data,mapping=aes(x=Race, y=Pass)) +
  stat_summary(fun.y=mean, geom='bar',fill='pink')+
  stat_summary(fun.data=mean_cl_boot, geom='pointrange')
```



Error bar (2)

This kind is a bit ugly.

```
ggplot(data=data,mapping=aes(x=Race, y=Pass)) +
  stat_summary(fun.y=mean, geom='bar') +
  stat_summary(fun.data=mean_cl_boot, geom='errorbar')
```