Unit 5: Inference for categorical data
1. Inference for a single proportion

Sta 101 – Spring 2019

Duke University, Department of Statistical Science

Dr. Ellison

Slides posted at
https://www2.stat.duke.edu/courses/Spring19/sta101.001/index.html

Outline

Organizing Analyses We Know

Tip: Break it down by types of variables involved in the research question.

https://www.facebook.com/konmarimethod/

Organizing Analyses We Know

Tip: Break it down by types of variables involved in the research question.

- Any **numerical** variables? How many?
- Any **categorical** variables? How many?

Organizing Analyses We Know

Tip: Break it down by types of variables involved in the research question.

- Any **numerical** variables? How many?
- Any **categorical** variables? How many?
  - How many **levels** does each categorical variable have?

Organizing Analyses We Know

Tip: Break it down by types of variables involved in the research question.

- Any **numerical** variables? How many?
- Any **categorical** variables? How many?
  - How many **levels** does each categorical variable have?
- If two variables, which variable(s) are **explanatory** and which variable is **response**?

Organizing

What analyses do we know so far?

https://www.facebook.com/konmarimethod/

Organizing

What analyses do we know so far?

Is there more than one way to conduct them?

Organizing

What analyses do we know so far?

Is there more than one way to conduct them?

If so, when should we use each way (ie. what conditions must be met)?

Organizing

Types of analyses that **HAVE one specified Population Parameter of Interest** and

Organizing

Types of analyses that **HAVE one specified Population Parameter of Interest** and

- We can create a **confidence interval** for the population parameter.

Organizing

Types of analyses that **HAVE one specified Population Parameter of Interest** and

- We can create a **confidence interval** for the population parameter.
- We can conduct a **hypothesis test** for the population parameter using:

Ho: $Pop.Param = \#$
Ha: $Pop.Param\ (\neq\ or\ <\ or\ >)\ \#$

Organizing (read over later)

What's new in Unit 5?

		Analyses and Ways to Conduct Them (that we know so far)	
Types of Variable(s) Involved	**Population Parameter**	**Confidence Interval for the Population Parameter**	**Hypothesis Test for the Population Parameter** *Ho: Pop. Param = # Ha: Pop. Param (≠ or < or >) #*
Single Numerical Variable	μ	CLT Confidence Interval (Unit 3+4) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 3+4) Bootstrap Hypothesis Test (Unit 4)
	μdiff	CLT Confidence Interval (Unit 4) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 4) Bootstrap Hypothesis Test (Unit 4)
	Median	Bootstrap Confidence Interval (Unit 4)	Bootstrap Hypothesis Test (Unit 4)
Single Categorical Variable (2 levels)	**p**	CLT Confidence Interval (Unit 5) Bootstrap Confidence Interval (Unit 4)	CLT Hypothesis Test (Unit 5) Bootstrap Hypothesis Test (Unit 4) Randomization Testing (Unit 5-Selecting ball/chips out of bag, rolling dice)
Numerical Response Variable Categorical Explanatory Variable (2 levels)	μ1-μ2	CLT Confidence Interval (Unit 4) Bootstrap Confidence Interval (Unit 5)	CLT Hypothesis Test (Unit 4) Randomization Testing (Unit 1-Shuffling Cards)
	Median1-Median2	Bootstrap Confidence Interval (Unit 5)	Randomization Testing (Unit 1-Shuffling Cards)
Categorical Response Variable Categorical Explanatory Variable (both have 2 levels)	**p1-p2**	CLT Confidence Interval (Unit 5) Bootstrap Confidence Interval (Unit 5)	CLT Hypothesis Test (Unit 5) Randomization Testing (Unit 1-Shuffling Cards)

Organizing

Types of analyses that **DON'T HAVE one specified Population Parameter of Interest** and

- No confidence interval framework
- Hypothesis tests set up are different

https://www.facebook.com/konmarimethod/

Organizing (read over later)

What's new in Unit 5?

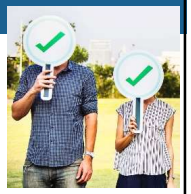
Types of Variables	Analysis	Hypotheses
Numerical Response Variable Categorical Explanatory Variable (>2 levels)	**ANOVA**	Ho: μ1=μ2=…=μk Ha: at least one pair of groups has means that are different
Single Categorical Variable (>2 levels)	**Chi-Squared Goodness of Fit Test**	Ho: The data follows the specified distribution. Ha: The data does not follow the specified distribution.
Categorical Response Variable Categorical Explanatory Variable (at least one has >2 levels)	**Chi-Squared Independence Test**	Ho: The two variables are independent/not associated. Ha: The two variables are dependent/associated.

Focus today:

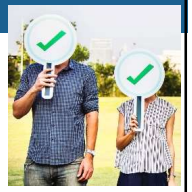
Categorical Variable (2 levels)

Single Population Proportion, *p*

p = proportion of one of the levels

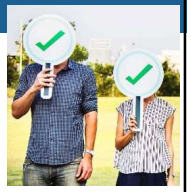


Focus today:
Categorical Variable (2 levels)
Single Population Proportion, *p*
p = proportion of one of the levels



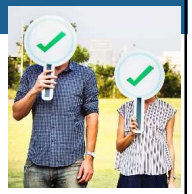
- Construct confidence intervals and hypothesis tests using **Central Limit Theorem methods**.

Focus today:
Categorical Variable (2 levels)
Single Population Proportion, *p*
p = proportion of one of the levels



- Construct confidence intervals and hypothesis tests using **Central Limit Theorem methods**.
- Conduct a hypothesis test for *p* with **randomization testing**.
- Create a confidence interval for *p* with **bootstrapping**.

Focus today:
Categorical Variable (2 levels)
Single Population Proportion, *p*
p = proportion of one of the levels



When to use each method

- Construct confidence intervals and hypothesis tests using **Central Limit Theorem methods**.
- Conduct a hypothesis test for *p* with **randomization testing**.
- Create a confidence interval for *p* with **bootstrapping**.

Outline

1. Housekeeping
2. Main ideas

Hypothesis Testing and Confidence Intervals with a Single Categorical Variable

1. Theoretical Hypothesis Testing and Confidence Intervals: The CLT also describes the distribution of $\hat{p}$
2. Interpretations: CI vs. HT determines observed vs. expected counts / proportions
3. CLT Conditions: Only use CLT based methods if the sample size is large enough for a nearly normal sampling distribution

3. Applications
   1. Single population proportion, large sample
   2. Single population proportion, small sample
4. Recap
5. Summary

Organizing CLT Results for Different Sample Statistics

When other certain conditions are met...

$\bar{x} \sim N(mean = \mu, standard\ dev./error = \frac{\sigma}{\sqrt{n}})$

When other certain conditions are met...

$\bar{x}_1 - \bar{x}_2 \sim N(mean = \mu_1 - \mu_2, standard\ dev./error = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$

When other certain conditions are met...

$\hat{p} \sim N(mean = p, standard\ dev./error = \sqrt{\frac{p(1-p)}{n}})$

https://www.facebook.com/konmarimethod/

Distribution of $\hat{p}$

Central Limit Theorem for Proportions

When certain conditions are met...

$\hat{p} \sim N(mean = p, standard\ dev./error = \sqrt{\frac{p(1-p)}{n}})$

Conditions:

- **Independence**:
  a. Random sample/assignment AND
  b. n<10% of population
- **Shape/Skew**:
  a. "Success/Failure Conditions:"
     a. np ≥ 10 AND
     b. n(1-p) ≥ 10

Distribution of $\hat{p}$

Why is are the following called the **success**/**failure** conditions?

np ≥ 10
n(1-p) ≥ 10

Distribution of $\hat{p}$

Why is are the following called the **success**/**failure** conditions?

p=proportion of "successes"
1-p=proportion of "failures"

np ≥ 10
n(1-p) ≥ 10

Why is are the following called the **success/failure** conditions?

p=proportion of "successes"

$1-p$=proportion of "failures"

Number of successes ← $np \geq 10$

Number of failures ← $n(1-p) \geq 10$

Central Limit Theorem for Proportions

When certain conditions are met...

$$\hat{p} \sim N\left(mean = p, standard\ dev./error = \sqrt{\frac{p(1-p)}{n}}\right)$$

Conditions:

- **Independence**:
  a. Random sample/assignment AND
  b. n<10% of population
- **Shape/Skew**:
  a. "Success/Failure Conditions:"
     a. np ≥ 10 (ie: at least 10 successes) AND
     b. n(1-p) ≥ 10 (ie: at least 10 failures)

Clicker question

Suppose p = 0.05. What shape does the distribution of $\hat{p}$ have in random samples of n = 100.

(a) unimodal and symmetric (nearly normal)
(b) bimodal and symmetric
(c) right skewed
(d) left skewed

Clicker question

Suppose p = 0.05. What shape does the distribution of $\hat{p}$ have in random samples of n = 100.

~~(a) unimodal and symmetric (nearly normal)~~ **p and n don't satisfy SF Conditions.**

$\boldsymbol{np = 5 < 10}$

$\boldsymbol{n(1-p) = 95 \geq 10}$

~~(b) bimodal and symmetric~~

(c) right skewed

(d) left skewed

Clicker question

Suppose p = 0.05. What shape does the distribution of $\hat{p}$ have in random samples of n = 100.

~~(a) unimodal and symmetric (nearly normal)~~ **p and n don't satisfy SF Conditions.**

~~(b) bimodal and symmetric~~

(c) right skewed

(d) left skewed

$$np = 5 < 10$$
$$n(1-p) = 95 \geq 10$$

If p and n don't satisfy SF conditions:

- Not unimodal and symmetric
- Sampling distribution for $\hat{p}$ has one peak at p
- think about boundaries for $\hat{p}$, ie [0,1].

Clicker question

Suppose p = 0.5. What shape does the distribution of $\hat{p}$ have in random samples of n = 100.

(a) unimodal and symmetric (nearly normal)

(b) bimodal and symmetric

(c) right skewed

(d) left skewed

Clicker question

Suppose p = 0.5. What shape does the distribution of $\hat{p}$ have in random samples of n = 100.

(a) unimodal and symmetric (nearly normal) **p and n satisfy SF Conditions.**

(b) bimodal and symmetric

(c) right skewed

(d) left skewed

$$np = 50 \geq 10$$
$$n(1-p) = 50 \geq 10$$

Outline

1. Housekeeping
2. Main ideas

Hypothesis Testing and Confidence Intervals with a Single Categorical Variable

1. Theoretical Hypothesis Testing and Confidence Intervals: The CLT also describes the distribution of $\hat{p}$
2. Interpretations: CI vs. HT determines observed vs. expected counts / proportions
3. CLT Conditions: Only use CLT based methods if the sample size is large enough for a nearly normal sampling distribution

3. Applications
   1. Single population proportion, large sample
   2. Single population proportion, small sample
4. Recap
5. Summary

CI vs. HT determines observed vs. expected counts / proportions

Central Limit Theorem for Proportions
When certain conditions are met…

$\hat{p} \sim \boldsymbol{N}(mean = p, standard\ dev./error = \sqrt{\frac{p(1-p)}{n}})$

For instance…

$P(\hat{p} < obs) = P(Z < \frac{obs - mean}{stand.dev})$=probability

Use z-tables

CI vs. HT determines observed vs. expected counts / proportions

Central Limit Theorem for Proportions
When certain conditions are met…

$\hat{p} \sim \boldsymbol{N}(mean = p, standard\ dev./error = \sqrt{\frac{p(1-p)}{n}})$

For instance…

$P(\hat{p} < obs) = P(Z < \frac{obs - mean}{stand.dev})$=probability

Use z-tables

CI vs. HT determines observed vs. expected counts / proportions

Central Limit Theorem for Proportions
When certain conditions are met…

$\hat{p} \sim \boldsymbol{N}(mean = p, standard\ dev/error. = \sqrt{\frac{p(1-p)}{n}})$

(1-α)% *Confidence Interval*

$(point\ estimate) \pm z^*_{\alpha/2} SE$

Test Statistic for Hypothesis Test

$z = \frac{(point\ estimate) - (null\ value)}{SE}$

CI vs. HT determines observed vs. expected counts / proportions

Central Limit Theorem for Proportions
When certain conditions are met…

$\hat{p} \sim \boldsymbol{N}(mean = p, standard\ dev/error. = \sqrt{\frac{p(1-p)}{n}})$

(1-α)% *Confidence Interval*

$(point\ estimate) \pm z^*_{\alpha/2} SE$

Test Statistic for Hypothesis Test

$z = \frac{(point\ estimate) - (null\ value)}{SE}$

CI vs. HT determines observed vs. expected counts / proportions

Central Limit Theorem for Proportions

When certain conditions are met…

$\hat{p} \sim N(mean = p, standard\ dev/error. = \sqrt{\frac{p(1-p)}{n}})$

(1-α)% *Confidence Interval*

$(point\ estimate) \pm z^*_{\alpha/2} SE$

Problem: However, in Confidence intervals and Hypothesis Testing, we don't know what ***p*** is….

Test Statistic for Hypothesis Test

$$z = \frac{(point\ estimate) - (null\ value)}{SE}$$

CI vs. HT determines observed vs. expected counts / proportions

Problem: In Confidence intervals and Hypothesis Testing, we don't know what ***p*** is but we need to plug it in in two parts of the analyses.

What part of the analyses do we need to plug in *p*?	Confidence Interval for *p*	Hypothesis Test for *p*

CI vs. HT determines observed vs. expected counts / proportions

Problem: In Confidence intervals and Hypothesis Testing, we don't know what ***p*** is but we need to plug it in in two parts of the analyses.

What part of the analyses do we need to plug in *p*?	Confidence Interval for *p*	Hypothesis Test for *p*
Checking SF Conditions: np≥10 n(1-p)≥10		
Calculating Standard Error: $SE = \sqrt{\frac{p(1-p)}{n}}$		

CI vs. HT determines observed vs. expected counts / proportions

Problem: In Confidence intervals and Hypothesis Testing, we don't know what ***p*** is but we need to plug it in in two parts of the analyses.

What part of the analyses do we need to plug in *p*?	Confidence Interval for *p*	Hypothesis Test for *p*
Checking SF Conditions: np≥10 n(1-p)≥10	Solution: Plug in the **observed** $\hat{p}$ for *p*.	
Calculating Standard Error: $SE = \sqrt{\frac{p(1-p)}{n}}$	Solution: Plug in the **observed** $\hat{p}$ for *p*.	

CI vs. HT determines observed vs. expected counts / proportions

Problem: In Confidence intervals and Hypothesis Testing, we don't know what ***p*** is but we need to plug it in in two parts of the analyses.

What part of the analyses do we need to plug in *p*?	Confidence Interval for *p*	Hypothesis Test for *p*
Checking SF Conditions: np≥10 n(1-p)≥10	Solution: Plug in the **observed** $\hat{p}$ for *p*.	Solution: Plug in the **null value p0** for *p*.
Calculating Standard Error: $SE = \sqrt{\frac{p(1-p)}{n}}$	Solution: Plug in the **observed** $\hat{p}$ for *p*.	Solution: Plug in the **null value p0** for *p*.

CI vs. HT determines observed vs. expected counts / proportions

Answer:

Remember, when doing a HT always assume H_0 is true!

- **S-F:** Number of successes and failures for checking the success-failure condition for the nearly normal distribution of $\hat{p}$:
  - CI: use observed proportion → $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$
  - HT: use null value of the proportion → $np_0 \geq 10$ and $n(1-p_0) \geq 10$
- **SE:** Proportion of success for calculating the standard error of $\hat{p}$:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

  - CI: use observed proportion → $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - HT: use null value of the proportion → $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

NEW CI vs. HT determines observed vs. expected counts / proportions

Interpreting…

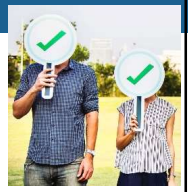
np = expected number of *successes*
n(1-p) = expected number of *failures*

n$\hat{p}$ = observed number of *successes*
n(1-$\hat{p}$) = observed number *failures*

np_0 = expected number of *successes (assuming Ho is true)*
$n(1-p_0)$ = expected number of *failures (assuming Ho is true)*

Outline

Single Population Proportion, *p*



When to use each method

- Construct confidence intervals and hypothesis tests using **Central Limit Theorem methods**.
- Conduct a hypothesis test for *p* with **randomization testing**.
- Create a confidence interval for *p* with **bootstrapping**.

Simulation vs. theoretical inference

Conditions:

Confidence Interval

- **Independence**:
  a. Random sample/assignment
  b. 10% rule
- **Shape/Skew**:
  a. "Success/Failure Conditions:" At least 10 successes and failures (ie: $n\hat{p} \geq 10$, $n(1-\hat{p}) \geq 10$)

Condition met:

(1-α)% *Confidence Interval With CLT Methods*

$$\hat{p} \pm z^*_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Condition NOT met:

Make a **bootstrap** confidence interval.

Simulation vs. theoretical inference

Conditions:

Hypothesis Test

- **Independence**:
  a. Random sample/assignment
  b. 10% rule
- **Shape/Skew**:
  a. "Success/Failure Conditions:" At least 10 successes and failures (ie: $np_0 \geq 10$, $n(1-p_0) \geq 10$)

Condition met:

CLT Methods

Test Statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Condition NOT met:

Conduct hypothesis test for p with **randomization testing** – balls/chips in a bag, dice, etc.

Central Limit Theorem for Proportions

When certain conditions are met…

$$\hat{p} \sim N\left(mean = p, standard\ dev/error. = \sqrt{\frac{p(1-p)}{n}}\right)$$

(1-α)% *Margin of Error*

$$ME = z^*_{\alpha/2} SE$$

Problem: In Margin of Error calculation:

- We don't know what ***p*** is.
- We may also not know what $\hat{p}$ is.

- **Calculating the necessary sample size for a CI with a given margin of error:**
  - Option 1: If there is a previous study, use $\hat{p}$ from that study

Recap on CLT based methods

- **Calculating the necessary sample size for a CI with a given margin of error:**
  - Option 1: If there is a previous study, use $\hat{p}$ from that study
  - Option 2: If not, use $\hat{p} = 0.5$:
    - if you don't know any better, 50-50 is a good guess
    - $\hat{p} = 0.5$ gives the most conservative estimate of the standard error, which gives the highest possible sample size

Outline

1. Housekeeping
2. Main ideas

Hypothesis Testing and Confidence Intervals with a Single Categorical Variable

1. Theoretical Hypothesis Testing and Confidence Intervals: The CLT also describes the distribution of $\hat{p}$
2. Interpretations: CI vs. HT determines observed vs. expected counts / proportions
3. CLT Conditions: Only use CLT based methods if the sample size is large enough for a nearly normal sampling distribution

3. Applications
   1. Single population proportion, large sample
   2. Single population proportion, small sample
4. Recap
5. Summary

Application exercise: App Ex 5.1

See course website for details.

Outline

1. Housekeeping
2. Main ideas

Hypothesis Testing and Confidence Intervals with a Single Categorical Variable

1. Theoretical Hypothesis Testing and Confidence Intervals: The CLT also describes the distribution of $\hat{p}$
2. Interpretations: CI vs. HT determines observed vs. expected counts / proportions
3. CLT Conditions: Only use CLT based methods if the sample size is large enough for a nearly normal sampling distribution

3. Applications
   1. Single population proportion, large sample
   2. Single population proportion, small sample
4. Recap
5. Summary

Outline

Let's conduct a hypothesis test on *p* (the proportion of Duke students that are vegetarian/vegan) with **randomization testing**.

Clicker question

Are you vegetarian or vegan?

(a) Yes, I am vegetarian or vegan
(b) No, I am neither vegetarian nor vegan

Clicker question

A variety of studies suggest that 8% of college students are vegetarian or vegan. Assuming that this class is a representative sample of Duke students, which of the following are the correct set of hypotheses for testing if the proportion of Duke students who are vegetarian is different than the proportion of vegetarian college students at large.

(a) $H_0 : p = 0.08;\ H_A : p \neq 0.08$
(b) $H_0 : p = 0.08;\ H_A : p < 0.08$
(c) $H_0 : \hat{p} = 0.08;\ H_A : \hat{p} \neq 0.08$
(d) $H_0 : \hat{p}_{Duke} = \hat{p}_{all\ college};\ H_A : \hat{p}_{Duke} \neq \hat{p}_{all\ college}$
(e) $H_0 : p_{Duke} = p_{all\ college};\ H_A : p_{Duke} \neq p_{all\ college}$

Clicker question

A variety of studies suggest that 8% of college students are vegetarian or vegan. Assuming that this class is a representative sample of Duke students, which of the following are the correct set of hypotheses for testing if the proportion of Duke students who are vegetarian is different than the proportion of vegetarian college students at large.

(a) $H_0 : p = 0.08; H_A : p \neq 0.08$

(b) $H_0 : p = 0.08; H_A : p < 0.08$

(c) $H_0 : \hat{p} = 0.08; H_A : \hat{p} \neq 0.08$

(d) $H_0 : \hat{p}_{Duke} = \hat{p}_{all\ college}; H_A : \hat{p}_{Duke} \neq \hat{p}_{all\ college}$

(e) $H_0 : p_{Duke} = p_{all\ college}; H_A : p_{Duke} \neq p_{all\ college}$

"A variety of studies…" usually suggests we can assume that we "know" the pop. parameter it's referring to (0.08).

Simulate by hand

Goal in Randomization Testing:

1. Set up hypotheses.

Ho: p=0.08 (=8/100)

Ha: p≠0.08

Example 1:

Simulate by hand

Goal in Randomization Testing:

1. Set up hypotheses.
2. Create a **Randomization Distribution:** an approximation for the sampling distribution that assumes Ho is true (or is centered at the null value)

Ho: p=0.08 (=8/100)

Ha: p≠0.08

Example 1:

- *100 chips in a bag: 8 green (veg), 92 white (non veg).*
- *Sample randomly n times from the bag, **with replacement** (n = observed sample size=144)*
- *Calculate $\hat{p}$, the proportion of greens (successes) in the random sample of size $n=144$, record this value.*
- *Repeat many times.*

Simulate by hand

Goal in Randomization Testing:

1. Set up hypotheses.
2. Create a **Randomization Distribution:** an approximation for the sampling distribution that assumes Ho is true (or is centered at the null value)
3. Calculate p-value with this randomization distribution: **P-value**=% of simulated points in the randomization distribution that are at least as extreme as the sample statistic observed.

Ho: p=0.08 (=8/100)

Ha: p≠0.08

Example 1:

- *100 chips in a bag: 8 green (veg), 92 white (non veg).*
- *Sample randomly n times from the bag, **with replacement** (n = observed sample size=144)*
- *Calculate $\hat{p}$, the proportion of greens (successes) in the random sample of size $n=144$, record this value.*
- *Repeat many times.*
- Calculate the proportion of simulations where $\hat{p}$ is at least as different from 0.08 (null value) as the observed sample proportion.

Clicker question

What would we expect this randomization distribution to be centered at?

a) 0
b) 0.08
c) the proportion of vegans/vegetarians in this class
d) Some number higher than 0.08.

Clicker question

What would we expect this randomization distribution to be centered at?

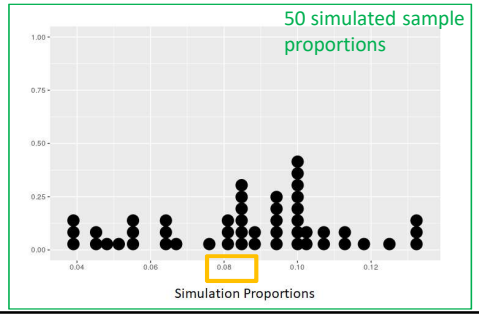
a) 0
b) 0.08 – The null value.
c) the proportion of vegans/vegetarians in this class
d) Some number higher than 0.08.

Simulate by hand

Ho: p=0.08
Ha: p≠0.08

Example:

P-value=



HT in R

```
n_veg = # of veg's in class
n_nonveg = # of nonveg's in class

sta101 = data.frame(veg = c(rep("yes", n_veg), rep("no", n_nonveg)))

inference(y = veg, data = sta101, success = "yes",
          statistic = "proportion", type = "ht",
          null = 0.08, alternative = "twosided",
          method = "simulation")
```

Outline

Let's create a bootstrap confidence interval for p (the proportion of Duke students that are vegetarian/vegan).

NEW Bootstrap interval for a single proportion

How would the simulation scheme change for a bootstrap interval for the proportion of Duke students who are vegetarians?

NEW Bootstrap interval for a single proportion

How would the simulation scheme change for a bootstrap confidence interval for the proportion of Duke students who are vegetarians?

Step 1: Create a Bootstrap Distribution

- ~~100 chips in a bag: 8 green (veg), 92 white (non veg).~~
- Sample randomly n times from the ~~bag~~ original sample, with replacement (n = observed sample size)
- Calculate $\hat{p}$, the proportion of vegetarians (successes) in the random sample of size n, record this value.
- Repeat many times.

NEW Bootstrap interval for a single proportion

How would the simulation scheme change for a bootstrap confidence interval for the proportion of Duke students who are vegetarians?

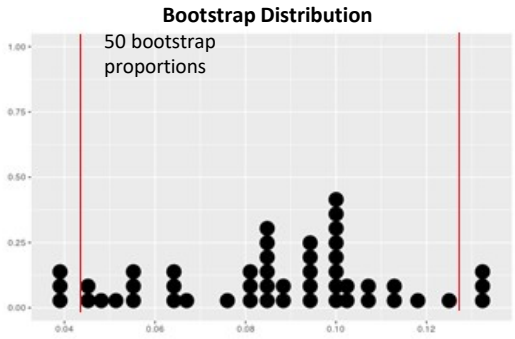
Step 1: Create a Bootstrap Distribution

- ~~100 chips in a bag: 8 green (veg), 92 white (non veg).~~
- Sample randomly n times from the ~~bag~~ original sample, with replacement (n = observed sample size)
- Calculate $\hat{p}$, the proportion of vegetarians (successes) in the random sample of size n, record this value.
- Repeat many times.

Step 2: Using the Bootstrap distribution, find a confidence interval using the Percentile Method or the SE Method

Simulate by hand

- Find a confidence interval using the Percentile Method or the SE Method



Percentile Method: Middle 44=(0.88)(50) points*

88% Confidence Interval: (4.2%,12.4%)

CI in R

```
inference(y = veg, data = sta101, success = "yes",
          statistic = "proportion", type = "ci",
          method = "simulation", boot_method = "se")
```

Summary of main ideas

1. The CLT also describes the distribution of $\hat{p}$
2. CI vs. HT determines observed vs. expected counts / proportions
3. Only use CLT based methods if the sample size is large enough for a nearly normal sampling distribution