# Unit 5: Inference for categorical data

## 1. Inference for a single proportion

Sta 101 - Spring 2016

Duke University, Department of Statistical Science

Dr. Çetinkaya-Rundel

Slides posted at *http://bit.ly/sta101_s16*

1. Housekeeping

2. 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

3. Applications
    1. Single population proportion, large sample
    2. Single population proportion, small sample

4. Recap

5. Summary

- PS 5 due April 1 (after the midterm) but I strongly recommend that you work on it and ask questions before.
- PA 5 is due Mar 29 (so you can review it before the midterm). It will be available on Mar 23 but will contain material on chi-square testing which we will cover on Mar 28, so you can decide when you want to take it.
- Sunday, Mar 27, 2-3pm at Old Chem 116.

1. Housekeeping

2. 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

3. Applications
    1. Single population proportion, large sample
    2. Single population proportion, small sample

4. Recap

5. Summary

*Central limit theorem for proportions:* Sample proportions will be nearly normally distributed with mean equal to the population mean, $p$, and standard error equal to $\sqrt{\frac{p\,(1-p)}{n}}$.

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

Conditions:

- Independence: Random sample/assignment + 10% rule
- At least 10 successes and 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)

(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)

(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)*
(b) bimodal and symmetric
(c) right skewed
(d) left skewed

1. Housekeeping

2. 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

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

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

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 $\hat{p}$:

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 $\rightarrow n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$

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

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 $\rightarrow n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$
  - HT: use null value of the proportion $\rightarrow np_0 \geq 10$ and $n(1-p_0) \geq 10$

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 $\rightarrow n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$
  - HT: use null value of the proportion $\rightarrow np_0 \geq 10$ 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 vs. HT determines observed vs. expected counts / proportions

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 $\rightarrow n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$
  - HT: use null value of the proportion $\rightarrow np_0 \geq 10$ 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 $\rightarrow SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

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

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 $\rightarrow n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$
  - HT: use null value of the proportion $\rightarrow np_0 \geq 10$ 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 $\rightarrow SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - HT: use null value of the proportion $\rightarrow SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

1. Housekeeping

2. 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

3. Applications
    1. Single population proportion, large sample
    2. Single population proportion, small sample

4. Recap

5. Summary

- If the S-F condition is met, can do theoretical inference: Z test, Z interval
- If the S-F condition is not met, must use simulation based methods: randomization test, bootstrap interval

1. Housekeeping

2. 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

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.

1. Housekeeping

2. 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

3. Applications
    1. Single population proportion, large sample
    2. Single population proportion, small sample

4. Recap

5. Summary

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

Describe a simulation scheme for this hypothesis test.

Describe a simulation scheme for this hypothesis test.

- 100 chips in a bag: 8 green (veg), 92 white (non veg).

**Describe a simulation scheme for this hypothesis test.**

- 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)

Describe a simulation scheme for this hypothesis test.

- 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)
- Calculate $\hat{p}$, the proportion of greens (successes) in the random sample of size $n$, record this value.

**Describe a simulation scheme for this hypothesis test.**

- 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)
- Calculate $\hat{p}$, the proportion of greens (successes) in the random sample of size $n$, record this value.
- Repeat many times.

**Describe a simulation scheme for this hypothesis test.**

- 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)
- Calculate $\hat{p}$, the proportion of greens (successes) in the random sample of size $n$, record this value.
- Repeat many times.
- Calculate the proportion of simulations where $\hat{p}$ is at least as different from 0.08 as the observed sample proportion.

```
n_veg = [fill in based on class data]
n_nonveg = [fill in based on class data]

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")
```

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

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

1. Housekeeping

2. 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

3. Applications
    1. Single population proportion, large sample
    2. Single population proportion, small sample

4. Recap

5. Summary

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

- Calculating the necessary sample size for a CI with a given margin of error:
  - If there is a previous study, use $\hat{p}$ from that study
  - 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 – highest possible sample size
- HT vs. CI for a proportion
  - Success-failure condition:
    - CI: At least 10 *observed* successes and failures
    - HT: At least 10 *expected* successes and failures, calculated using the null value
  - Standard error:
    - CI: calculate using observed sample proportion: $SE = \sqrt{\frac{p(1-p)}{n}}$
    - HT: calculate using the null value: $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

If the S-F condition is not met

- HT: Randomization test – simulate under the assumption that $H_0$ is true, then find the p-value as proportion of simulations where the simulated $\hat{p}$ is at least as extreme as the one observed.
- CI: Bootstrap interval – resample with replacement from the original sample, and construct interval using percentile or standard error method.

1. Housekeeping

2. 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

3. Applications
    1. Single population proportion, large sample
    2. Single population proportion, small sample

4. Recap

5. Summary

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