Lecture 14: Large and small sample inference for proportions

Statistics 101

Mine Çetinkaya-Rundel

March 13, 2012

Midterm evaluation, n = 43

- Lectures: Videotaping - probably too late for this semester but I'll look into it, 81% think pace is about right, loving the clicker questions, slides can be hard to see - printing them out or using a computer seems to help some, class can get loud at times, solutions to clicker questions - one incentive to come to class, and feel free to as your team mates/email/come to office hours if you miss any
- HW: Answer keys are posted, avg time spent on HW 3 hrs, with a standard deviation of 1hr, you should not be losing points on the entire question on a HW if all you missed was a small calculation error
- Labs: Most think labs relate well to lectures, about 1/3 don't, saving your code in an R script - see Lab 1, last question on lab assignments - intended to get you think about how various components of the course tie into each other, 70% have been collaborating with their team members
- 40 out of 41 who responded think stats is useful

Review question

Which of the following is a data set?

"Scientists predict that global warming may have big effects on the polar regions within the next 100 years. One of the possible effects is that the northern ice cap may completely melt. Would this bother you a great deal, some, a little, or not at all if it actually happened?

(I)

	attitude	group
1	A great deal	Duke
2	A great deal	Duke
⋮		
85	Not at all	Duke
86	Some	US
87	A great deal	US
⋮		
764	Not at all	US
765	A great deal	US

(II)

	Duke	US
A great deal	58	454
Some	15	124
A little	9	52
Not at all	3	50

(a) I and II
(b) Only I
(c) Only II
(d) Neither

Melting ice cap

- We are interested in finding out if there is a significant difference between the proportions of Duke students and US public who would be bothered a great deal by the melting of the northern ice cap.

H_0: $p_{Duke} = p_{US}$

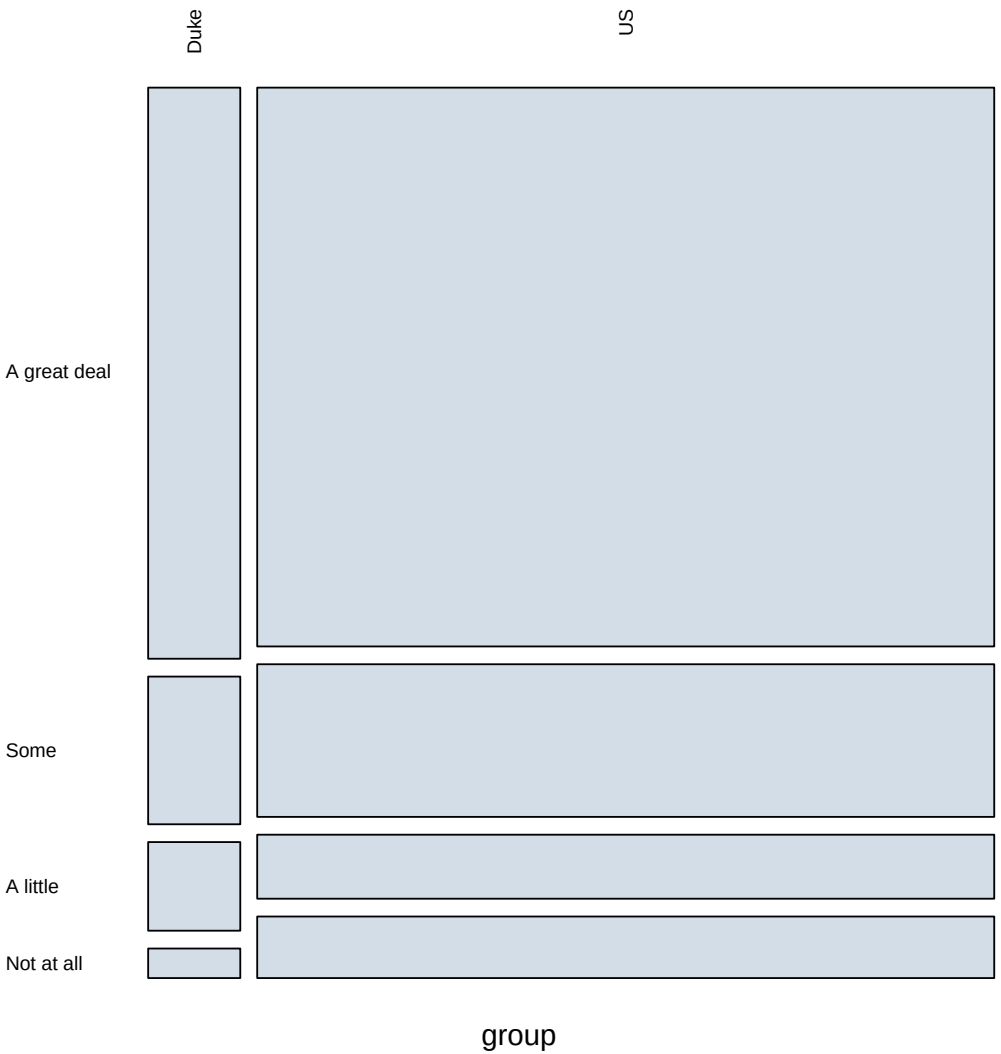
H_A: $p_{Duke} \neq p_{US}$

H_0 : $p_{Duke} - p_{US} = 0$

H_A : $p_{Duke} - p_{US} \neq 0$

- Parameter: Difference between population proportions, $p_{Duke} - p_{US}$
- Point estimate: Difference between sample proportions, $\hat{p}_{Duke} - \hat{p}_{US}$

Exploratory analysis



	Duke	US
# of successes	58	454
n	85	680
$\hat{p}$	0.682	0.668

Checking assumptions & conditions

1. *Independence within groups:*
   - The US group is sampled randomly and we're assuming that the Duke group represents a random sample as well.
   - 85 $<$ 10% of all college Duke students and 680 $<$ 10% of all Americans.

   We can assume that the attitudes of Duke students in the sample are independent of each other, and attitudes of US residents in the sample are independent of each other as well.
2. *Independence between groups:* The sampled Duke students and the US residents are independent of each other.
3. *Normality:*
   We need at least 10 *expected* successes and 10 *expected* failures in the two groups.

Flashback to working with one proportion

- When constructing a confidence interval for a population proportion, we check if the *observed* number of successes and failures are at least 10.

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

- When conducting a hypothesis test for a population proportion, we check if the *expected* number of successes and failures are at least 10.

$$np \geq 10 \qquad n(1-p) \geq 10$$

In the above formula p comes from the null hypothesis.

Finding expected number of successes when comparing two population proportions

- Similar to the one sample case, when constructing a confidence interval for the difference between two population proportions, we check if the *observed* number of successes in each group and failures are at least 10.

$$Group1 : n_1\hat{p}_1 \geq 10 \qquad n_1(1-\hat{p}_1) \geq 10$$

$$Group2 : n_2\hat{p}_2 \geq 10 \qquad n_2(1-\hat{p}_2) \geq 10$$

- But in this case the null hypothesis simply states the population proportions are equal to each other, and doesn't set them equal to a given value.

$$H_0 : p_1 = p_2$$

Then, we need to first find a common proportion for the two groups, and use that in our analysis.

Pooled estimate of a proportion

- Since H_0 implies that both samples come from the same population, we pool the two samples to calculate a *pooled* estimate of the sample proportion.
- This simply means finding the proportion of total successes among the total number of observations.

Pooled estimate of a proportion

$$\hat{p} = \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2}$$

Pooled estimate of a proportion - in context

	Duke	US
# of successes	58	454
n	85	680
$\hat{p}$	0.682	0.668

$$\hat{p} = \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2}$$
$$= \frac{58 + 454}{85 + 680}$$
$$= \frac{512}{765}$$
$$= 0.669$$

Why is the pooled estimate closer to $\hat{p}_{US}$ than $\hat{p}_{Duke}$?

Number of expected successes and failures

Duke

$$85 \times 0.669 = 56.865$$
$$85 \times (1 - 0.669) = 28.135$$

US

$$680 \times 0.669 = 454.92$$
$$680 \times (1 - 0.669) = 225.08$$

There are at least 10 expected successes and 10 expected failures in both groups. So we can proceed with the test.

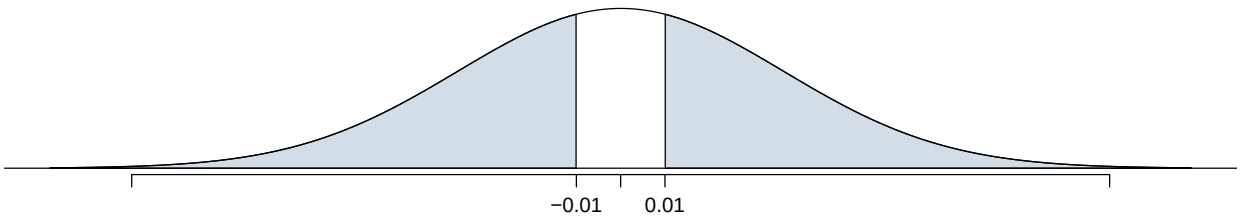
What if there weren't at least 10 expected successes and failures? What method can we use then?

The hypothesis test

- Hypotheses: $H_0 : p_{Duke} = p_{US}$; $H_A : p_{Duke} \neq p_{US}$
- Assumptions and conditions are satisfied
- Test statistic:

$$Z = \frac{(0.682 - 0.668) - 0}{\sqrt{\frac{0.669 \times (1-0.669)}{85} + \frac{0.669 \times (1-0.669)}{680}}}$$
$$= \frac{0.014}{0.054}$$
$$= 0.26$$

- p-value: $2 \times P(Z > 0.26) = 2 \times (1 - 0.6026) = 0.7948$



What's the conclusion?

The 95% confidence interval

$$
\begin{aligned}
\textit{point estimate} &\pm \textit{ME} \\
(0.682 - 0.668) &\pm 1.96 \times \sqrt{\frac{0.682 \times (1 - 0.682)}{85} + \frac{0.668 \times (1 - 0.668)}{680}} \\
0.014 &\pm 1.96 \times 0.0536 \\
0.014 &\pm 0.105 \\
(-0.091 &, \quad 0.119)
\end{aligned}
$$

Clicker question

Which of the below is the correct interpretation of this 95% confidence interval?

$$p_{Duke} - p_{US} = (-0.091, 0.119)$$

We are 95% confident that Duke students who would be bothered a great deal about the melting of the northern ice cap are

(a) 9.1% to 11.9% lower

(b) 9.1% to 11.9% higher

(c) 9.1% lower to 11.9% higher

(d) 9.1% higher to 11.9% lower

than those in the US population.

Recap - comparing two proportions

- Population parameter: $p_1 - p_2$, point estimate: $\hat{p}_1 - \hat{p}_2$
- Assumptions and conditions:
  - independence within groups (random sample and 10% condition met for both groups)
  - independence between groups
  - at least 10 successes and failures
- $SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
- *Only* when conducting a hypothesis test where $H_0 : p_1 = p_2$
  - Pooled proportion: $\hat{p} = \frac{\#\ suc_1 + \#suc_2}{n_1 + n_2}$
  - Use the pooled proportion for calculating expected number of successes and failures, and in the calculation of the standard error

Reference - standard error calculations

	one sample	two samples
mean	$SE = \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
proportion	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

- When working with means, it's very rare that σ is known, so we usually use s.
- When working with proportions,
  - if doing a hypothesis test, p comes from the null hypothesis
  - if constructing a confidence interval, use $\hat{p}$ instead

When to retreat

- The inference tools that we have learned that rely on the CLT and the normal distribution require the following two assumptions:
  1. The individual observations must be independent.
  2. Sample size and skew should not prevent the sampling distribution from being nearly normal.
     - means: $n > 50$, population distribution not extremely skewed
     - proportions: at least 10 successes and 10 failures
- If conditions for a statistical technique are not satisfied:
  1. learn new methods that are appropriate for the data
  2. consult a statistician
  3. ignore the failure of conditions → this option effectively invalidates any analysis and may discredit novel and interesting findings
- Next, we'll jump ahead to Chapter 6 and learn how to analyze smaller samples *(tools you might need for your project)*.
- Next week we'll come back to Chapter 5 to work on chi-square testing *(tools you won't need for your project)*.

Gender discrimination - another look

- In 1972, as a part of a study on gender discrimination, 48 male bank supervisors were each given the same personnel file and asked to judge whether the person should be promoted to a branch manager job that was described as “routine”.
- The files were identical except that half of the supervisors had files showing the person was male while the other half had files showing the person was female.
- It was randomly determined which supervisors got “male” applications and which got “female” applications.
- Of the 48 files reviewed, 35 were promoted.
- The study is testing whether females are unfairly discriminated against.

B.Rosen and T. Jerdee (1974), “Influence of sex role stereotypes on personnel decisions”, J.Applied Psychology, 59:9-14.

Gender discrimination - another look

The table below shows the gender distribution of the promoted files.

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

Parameter and point estimate

Do these data provide convincing evidence that females are unfairly discriminated against?

- *Parameter of interest:* Difference between the proportions of *all* males and females who are equally qualified who get promoted.

$$p_m - p_f$$

- *Point estimate:* Difference between the proportions of equally qualified males and females *in the sample* who get promoted.

$$\hat{p}_m - \hat{p}_f$$

Hypotheses

Clicker question

If the study is testing whether females are unfairly discriminated against, what are the appropriate hypotheses for this study?

$$p = \text{proportion who get promoted}$$

(a) $H_0 : p_m - p_f = 0$
$H_A : p_m - p_f \neq 0$

(b) $H_0 : p_m - p_f = 0$
$H_A : p_m - p_f < 0$

(c) $H_0 : p_m - p_f = 0$
$H_A : p_m - p_f > 0$

(d) $H_0 : \hat{p}_m - \hat{p}_f = 0$
$H_A : \hat{p}_m - \hat{p}_f \neq 0$

Assumptions and conditions

1. *Independence:* Since it was randomly determined which supervisors got “male” applications and which got “female” applications, we can assume that the decisions are independent.
2. *Normality:* There are only 3 male files that did not get promoted.

So what do we do?

Since the sample size isn’t large enough to use CLT based methods, we use a simulation method instead.

Set up

1. We'll let a face card represent *not promoted* and a non-face card represent a *promoted*. Consider aces as face cards.
   - Set aside the jokers (if any).
   - Take out 3 aces → there are exactly 13 face cards left in the deck.
   - Take out a number card → there are exactly 35 non-face cards left in the deck.
2. Shuffle the cards and deal them intro two groups of size 24, representing males and females.
3. Count and record how many files in each group are promoted (non-face cards).
4. Calculate the proportion of promoted files in each group and take the difference (male - female).
5. We repeat this many many times and plot the randomization distribution of $\hat{p}_{sim,m} - \hat{p}_{sim,f}$.

Simulation results - shuffling

Clicker question

What is the difference you calculated?

(a) More than 0.2

(b) Between 0.1 and 0.2

(c) Between -0.1 and 0.1

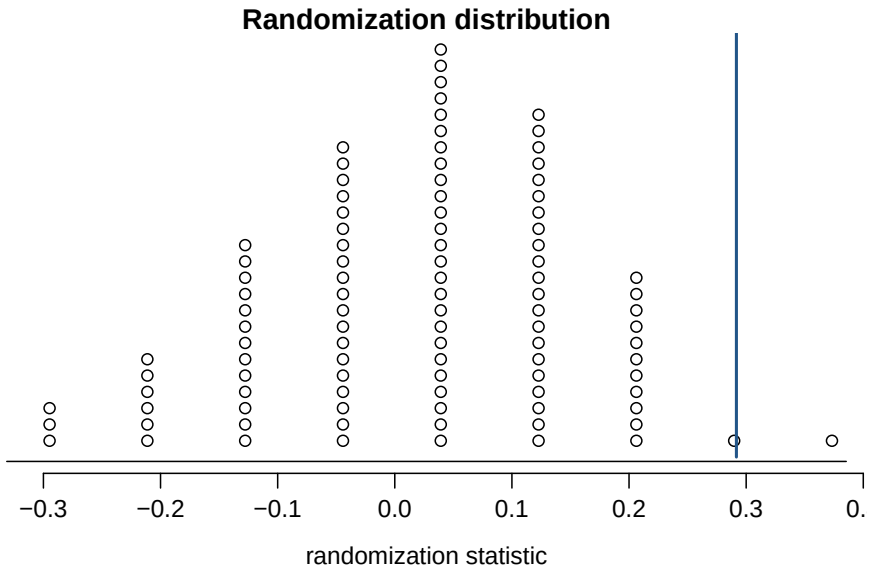
(d) Between -0.1 and -0.2

(e) Less than -0.2

Randomization test

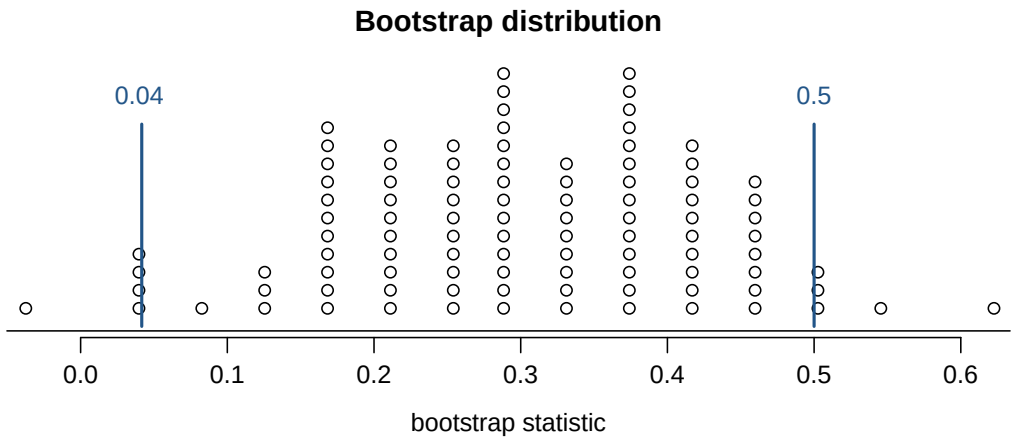
```
# load data
discr = read.csv("http://stat.duke.edu/courses/Spring12/sta101.1/lec/discr.csv", h = T)

# randomization test
inference(discr$promotion, discr$gender, est = "proportion", type = "ht", method = "simulation",
      order = c("male","female"), outcome = "promoted", null = 0, alternative = "greater", nsim =
      100)
# Two categorical variables
# Difference between two proportions
# n_male = 24 ; n_female = 24
# Observed difference between proportions = 0.2917
# H0: p_male - p_female = 0
# HA: p_male - p_female > 0
# Randomizing, please wait...
# p-value:  0.01
```



Bootstrap confidence interval

```
inference(discr$promotion, discr$gender, est = "proportion", type = "ci", method = "simulation",
      order = c("male","female"), outcome = "promoted", nsim = 100)
# Two categorical variables
# Difference between two proportions
# n_male = 24 ; n_female = 24
# Observed difference between proportions = 0.2917
# Bootstrapping, please wait...
# 95 % Bootstrap interval = ( 0.04 , 0.5 )
```



How do we interpret this interval?

Paul the Octopus

http://www.youtube.com/watch?v=3ESGpRUMj9E

Paul the Octopus - psychic?

- Paul the Octopus predicted 8 World Cup games, and predicted them all correctly
- Does this provide convincing evidence that Paul actually has psychic powers?
- How unusual would this be if he was just randomly guessing (with a 50% chance of guessing correctly)?
- Hypotheses:

$H_0 : \ p = 0.5$
$H_A : \ p > 0.5$

Assumptions and conditions

1. *Independence:* We can assume that each guess is independent of another.
2. *Normality:* The number of expected successes is *smaller than 10*.

$$8 \times 0.5 = 0.4$$

So what do we do?

Since the sample size isn't large enough to use CLT based methods, we use a simulation method instead.

Simulate

Clicker question

Flip a coin 8 times. Did you get all heads?

(a) Yes

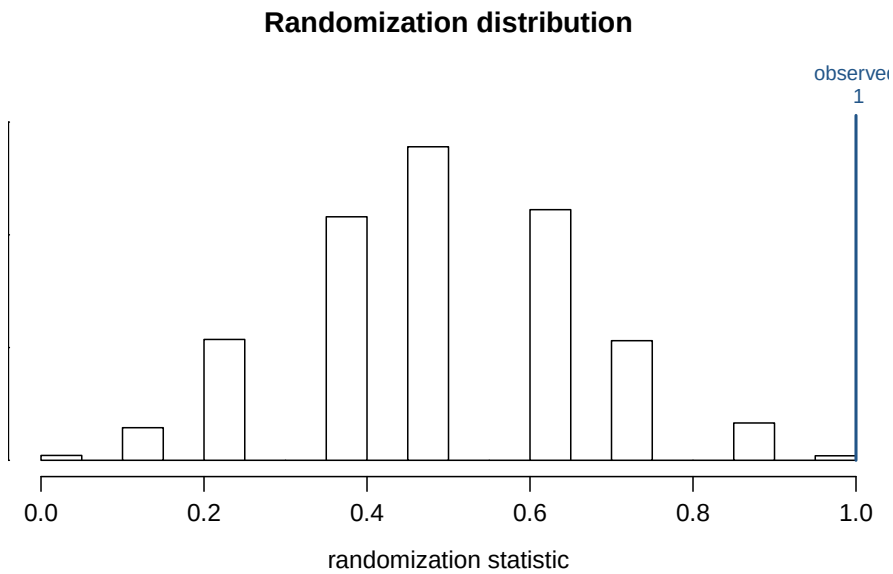
(b) No

What proportion got all heads? What's the p-value? What's the conclusion of the hypothesis test?

Randomization test

```
# data
paul = factor(rep("yes", 8), levels = c("yes","no"))

# test
inference(paul, est = "proportion", type = "ht", method = "simulation", outcome = "yes", null = 0.5,
      alternative = "greater")
# One categorical variable
# Single proportion
# Observed proportion = 1
# Randomizing, please wait...
# H0: p = 0.5
# HA: p > 0.5
# p-value:  0.0029
```



Conclusions

Clicker question

Which of the following is false?

(a) If in fact Paul was randomly guessing, the probability that he would get the result of all 8 games correct is 0.0029.

(b) Reject H_0, the data provide convincing evidence that Paul did better than randomly guessing.

(c) We may have made a Type I error.

(d) The probability that Paul is psychic is 0.0029.