MHA Statistics Course - Day 2

Pacific University

Dr. Chester Ismay
Email: chester@pacificu.edu

2017/01/27

Slides available at http://bit.ly/mha17-day2

Recall Hans Rosling's 4 minute presentation

- What are the most important things you remember from the video from yesterday?
- How did Hans use data viz to tell the story he wanted to tell?
- What did you learn from the workshop yesterday?

These slides available at http://bit.ly/mha17-day2

Workshop Agenda - Saturday

- Part 4: Simulating Randomness
  - Random sampling
  - Common terms of inference
  - Simulation
  - Sampling distribution
- Part 5: Inferential Statistics
  - Hypothesis testing
  - Confidence intervals
- Part 6: Workshop Review

These slides available at http://bit.ly/mha17-day2

Sampling seen as tasting soup

Key terms

- population/sample
- sampling
- representative sample
- bias
- generalizability
- parameter/statistic

These slides available at http://bit.ly/mha17-day2

Sampling seen as tasting soup

These slides available at http://bit.ly/mha17-day2

Visualizing sampling

- Let's explore data on profiles of San Francisco OKCupid users.
- This data contains information about/questions asked of 59,946 OkCupid users who were
    - living within 25 miles of San Francisco,
    - had active profiles on June 26, 2012,
    - were online in the previous year, and
    - had at least one picture in their profile.

These slides available at http://bit.ly/mha17-day2

Visualizing sampling

- Let's think of these 59,946 users as our population

These slides available at http://bit.ly/mha17-day2

Visualizing sampling

- Let's think of these 59,946 users as our population
  - Visualize their self-reported heights in inches

These slides available at http://bit.ly/mha17-day2

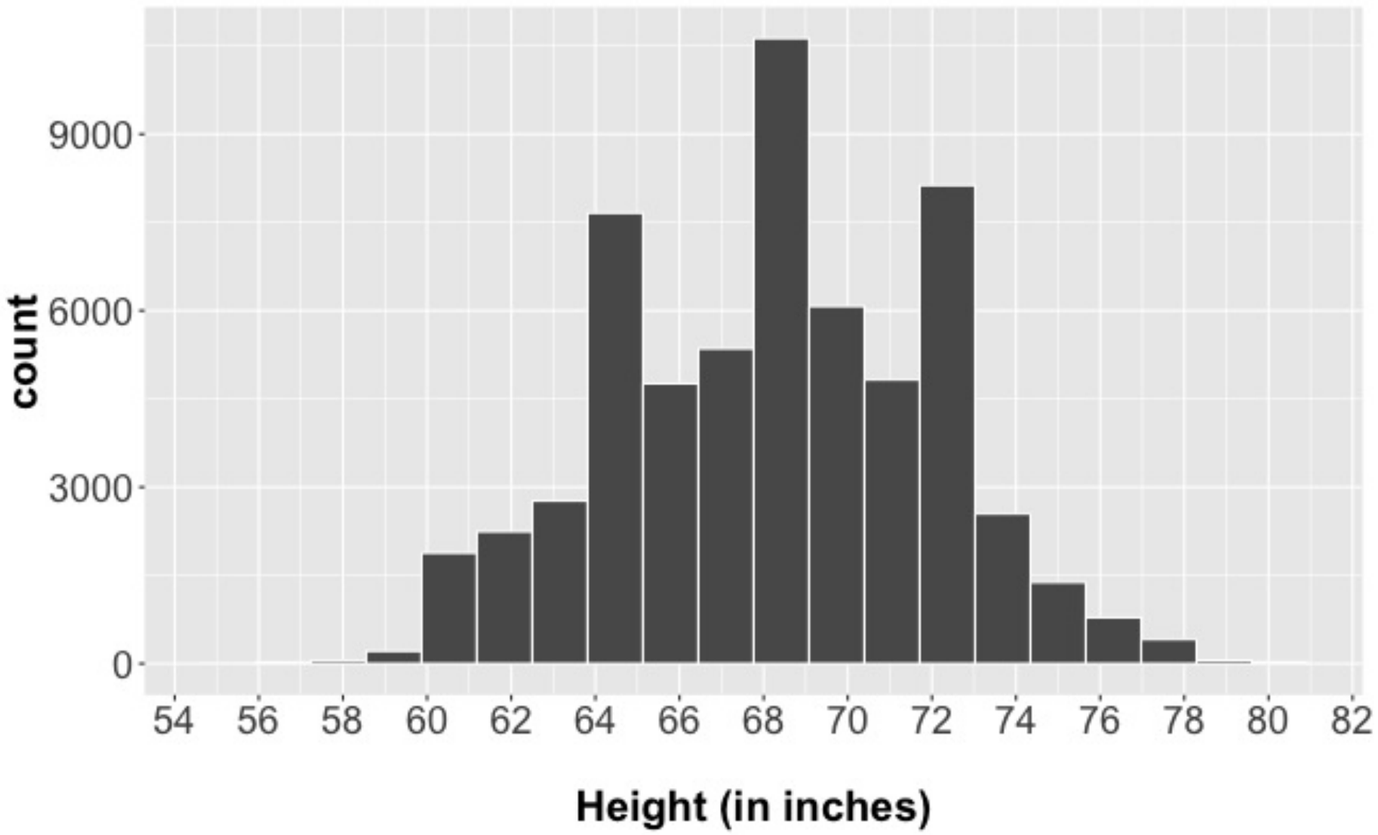
Visualizing sampling

- Let's think of these 59,946 users as our population
  - Visualize their self-reported heights in inches
  - Then take samples from the population of heights and visualize those

These slides available at http://bit.ly/mha17-day2

The population

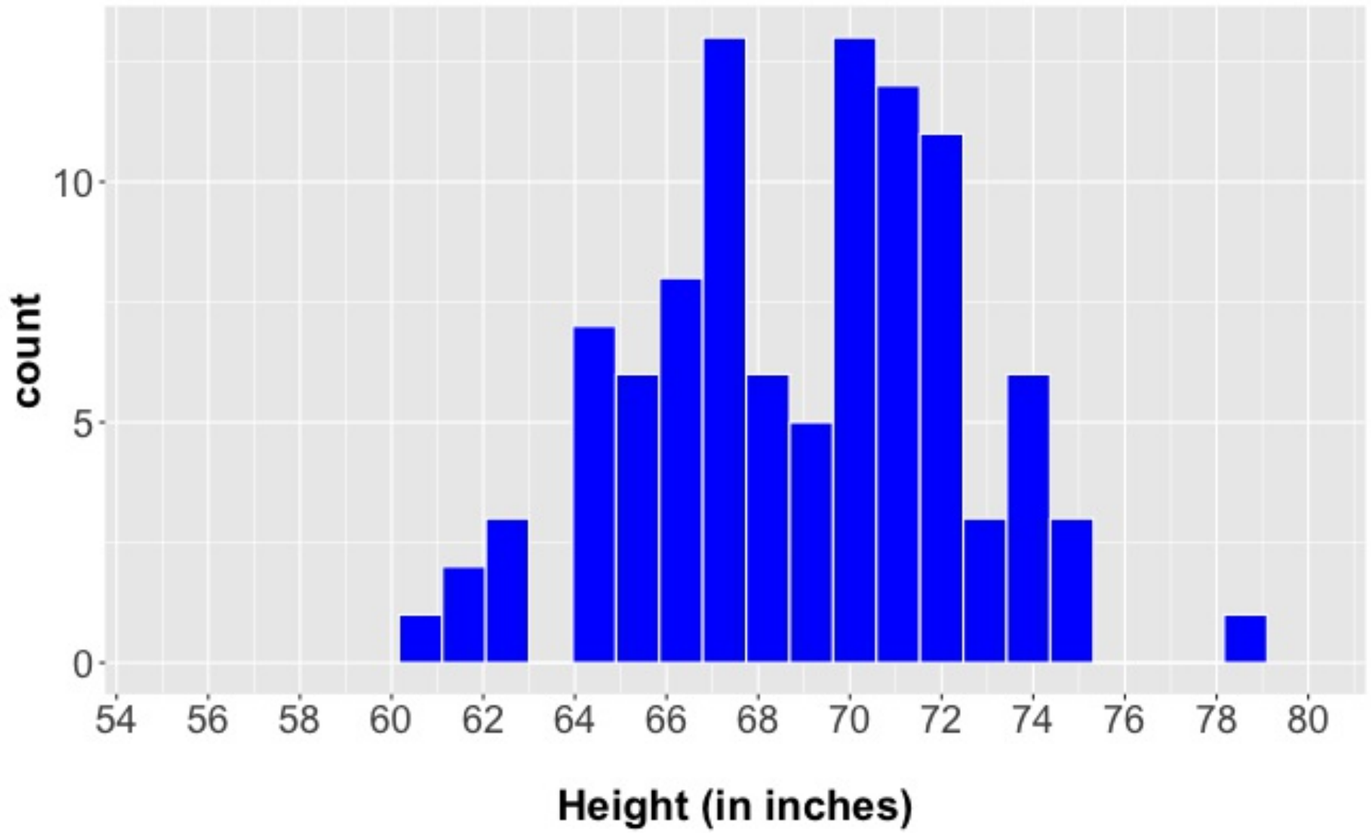
Self-reported heights of San Francisco OKCupid users in 2012



These slides available at http://bit.ly/mha17-day2

One random sample

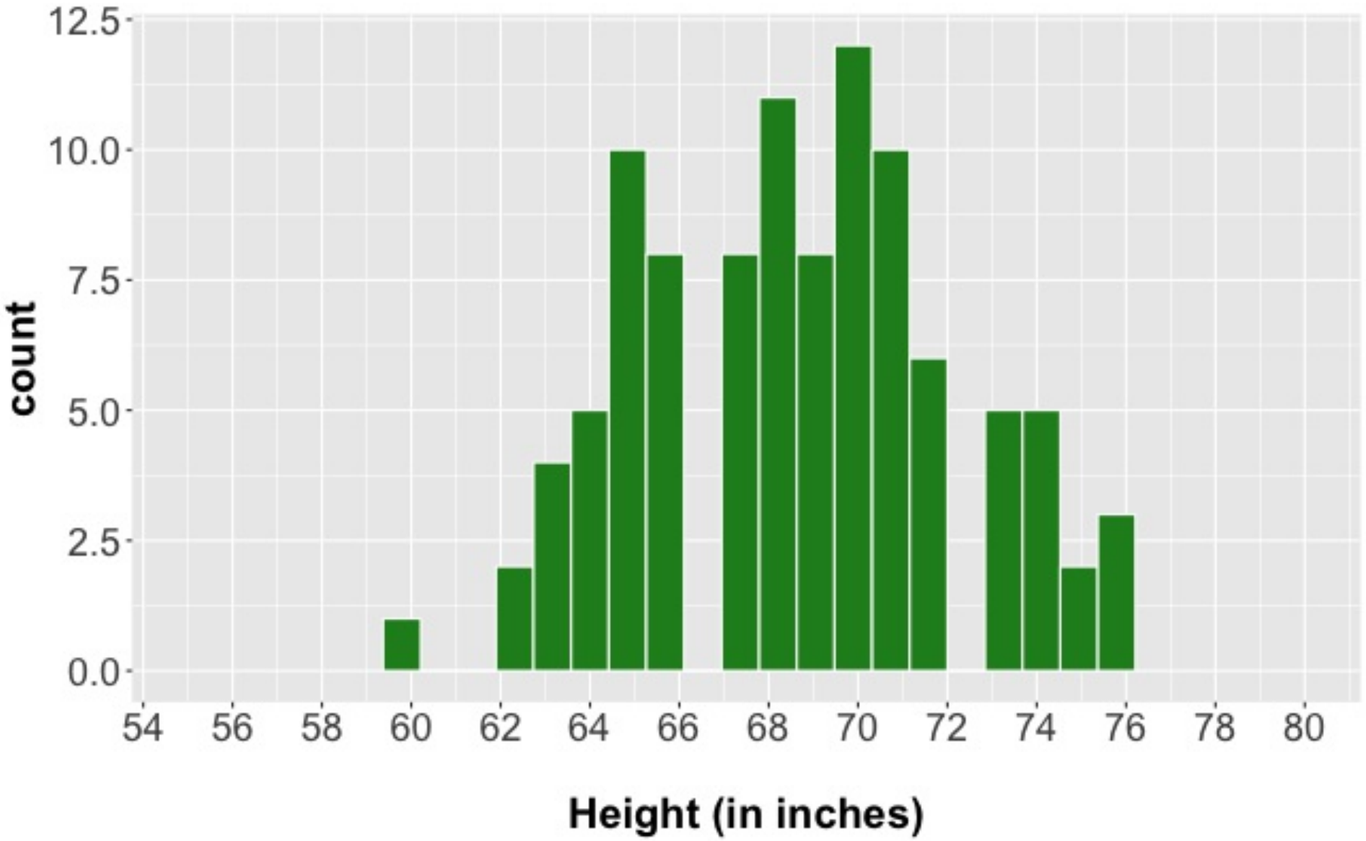
Self-reported heights of 100 San Francisco OKCupid users in 2012



These slides available at http://bit.ly/mha17-day2

Another random sample

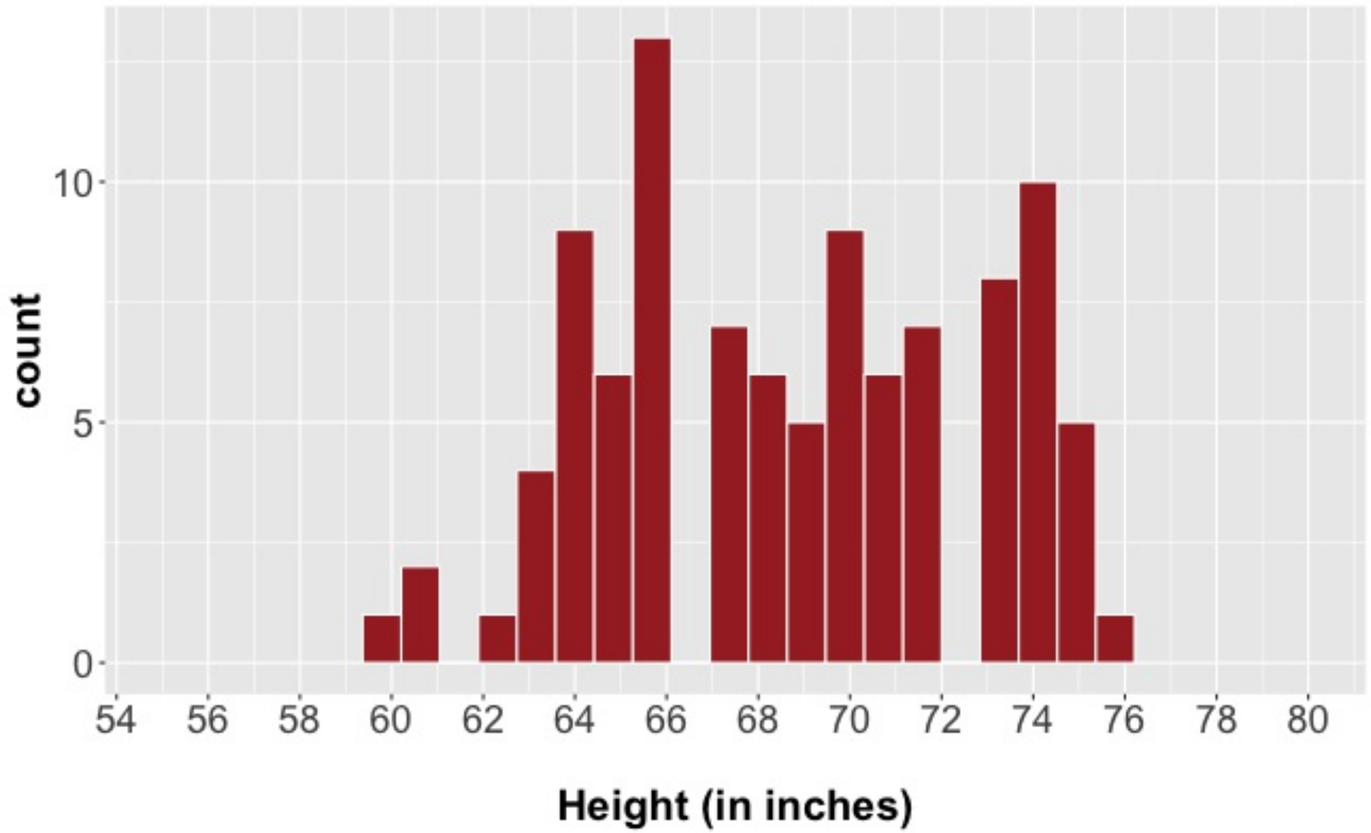
Self-reported heights of 100 San Francisco OKCupid users in 2012



These slides available at http://bit.ly/mha17-day2

One more random sample

Self-reported heights of 100 San Francisco OKCupid users in 2012



These slides available at http://bit.ly/mha17-day2

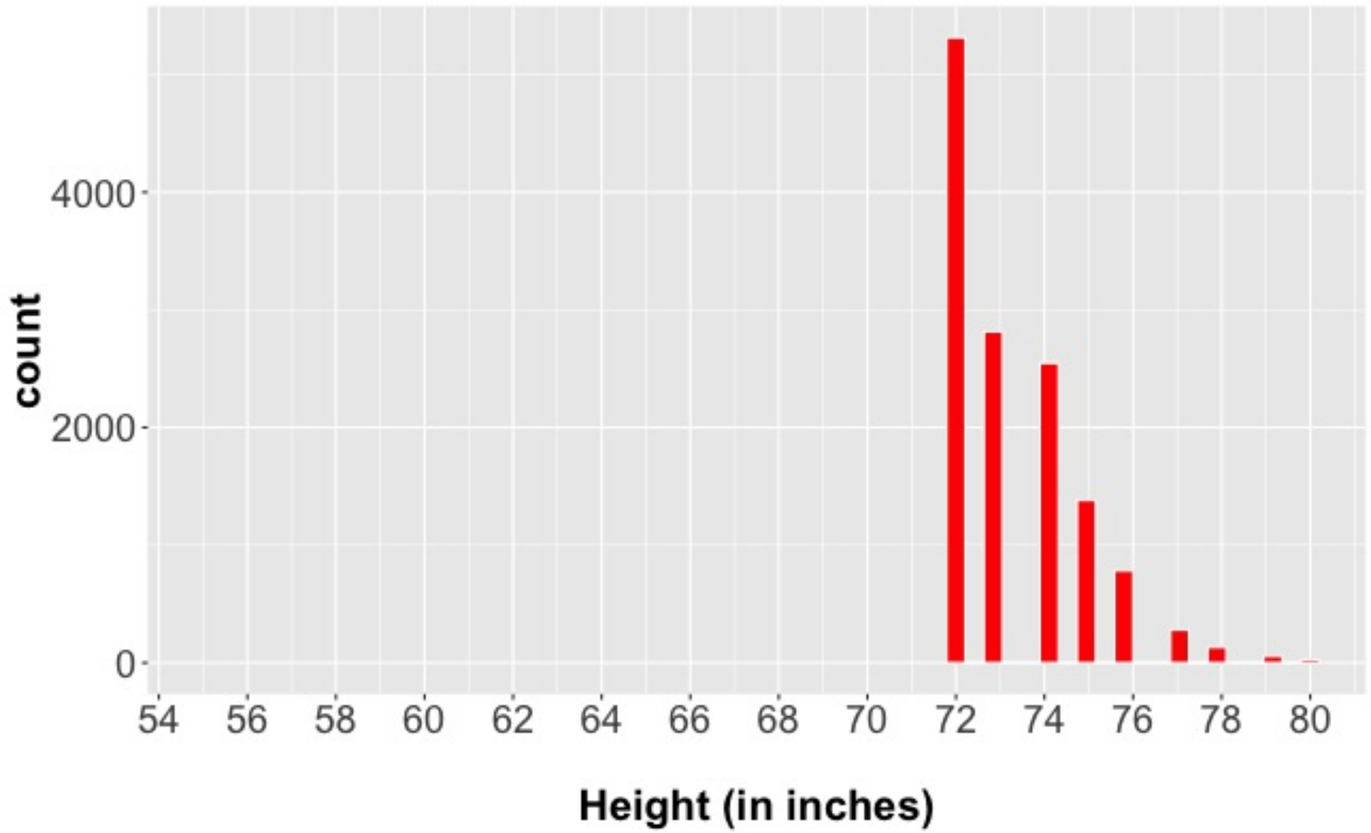
A convenience sample

Self-reported heights of 100 San Francisco OKCupid users in 2012

These slides available at http://bit.ly/mha17-day2

A convenience sample

Self-reported heights of 100 San Francisco OKCupid users in 2012



These slides available at http://bit.ly/mha17-day2

The sampling distribution of the mean

- Let's suppose we were interested in estimating the mean height of all OKCupid San Francisco profiles. To do so, we could look at the mean of the height variable in our first sample:

- Mean of sample 1: 68.85 inches

The sampling distribution of the mean

- Or we could have just used the mean of sample 2
- Mean of sample 2: 68.65 inches

The sampling distribution of the mean

- Or the mean of sample 3
- Mean of sample 3: 68.77 inches

The sampling distribution of the mean

- Or the mean of sample 4?
- Mean of sample 4: 73.349 inches

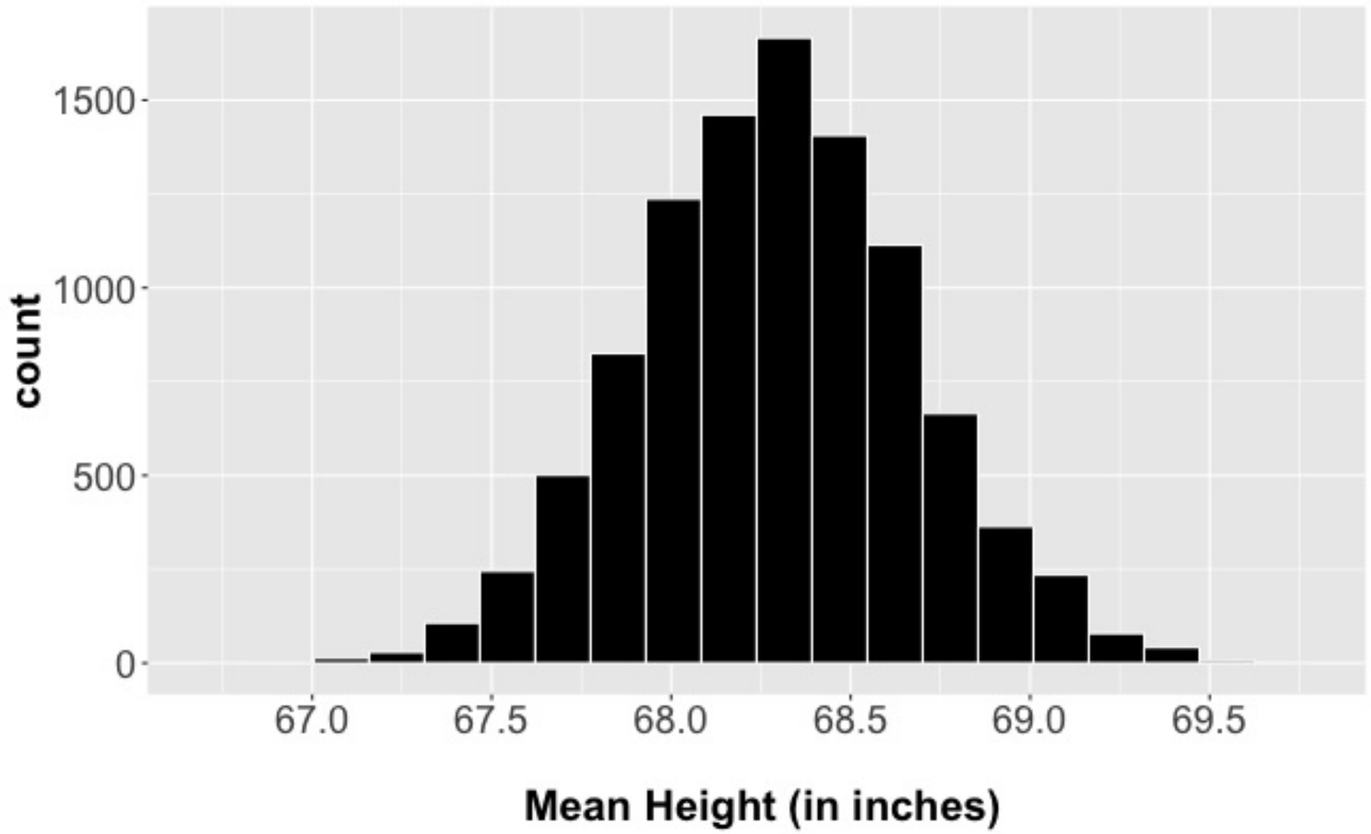
The sampling distribution of the mean

- The fact that the estimate of our population mean changes based on the sample selected is called sampling variability.
- You can view the sampling variability by looking at sample means from many different samples in what is known as the sampling distribution of the mean.

These slides available at http://bit.ly/mha17-day2

The sampling distribution of the mean

Mean height (inches) from random samples of OKCupid profile heights



These slides available at http://bit.ly/mha17-day2

Gapminder data

We will be working with data derived from the data Hans Rosling gave in his presentation from yesterday.

- You can view the full data set in interactive form here
- Please download the CSV of this data to your computer here

These slides available at http://bit.ly/mha17-day2

Comparisons to make based on inference

1. Was the proportion of Autocratic countries lower in 2007 than in 1957?
2. Did European countries have a statistically different mean GDP per capita in 2002 than Asian countries?
3. Was the mean life expectancy for countries in South America in 2007 greater than 75 years?
4. Were the majority of African countries in 1962 Democratic?

These slides available at http://bit.ly/mha17-day2

BREAK TIME

Think about two other statistical questions you could answer based on this data and write them down.

These slides available at http://bit.ly/mha17-day2

The Lady Tasting Tea

- Read over the exerpt from the book The Lady Tasting Tea by David Salsburg here

- Read up until the Learning check

Salsburg, David. 2001. The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century. First Edition. New York, NY: W.H. Freeman.

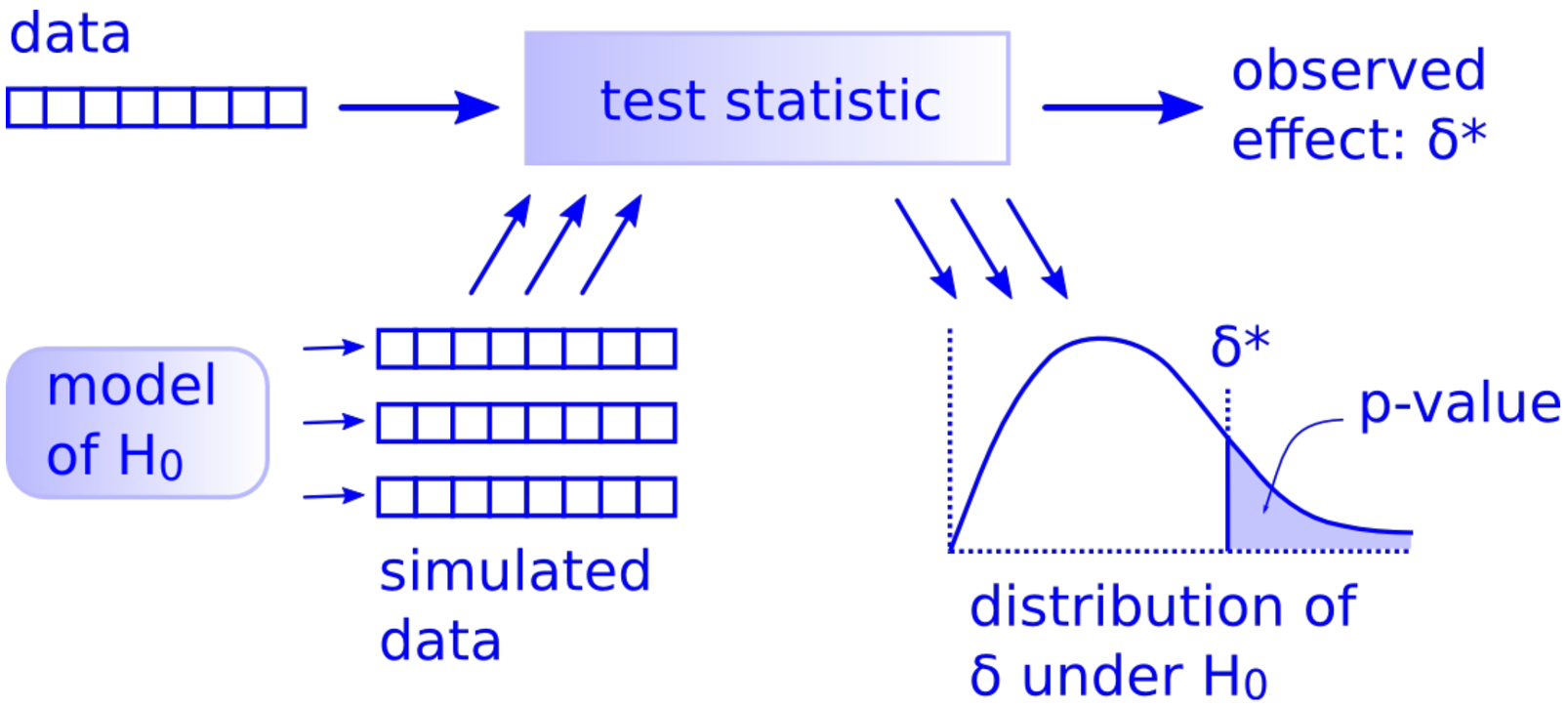
These slides available at http://bit.ly/mha17-day2

Using StatKey

DEMO

These slides available at http://bit.ly/mha17-day2

There is Only One (Hypothesis) Test



These slides available at http://bit.ly/mha17-day2

BREAK TIME

What do the different pieces of the diagram look like for each of our questions?

1. Was the proportion of Autocratic countries lower in 2007 than in 1957?
2. Did European countries have a statistically different mean GDP per capita in 2002 than Asian countries?
3. Was the mean life expectancy for countries in South America in 2007 greater than 75 years?
4. Were the majority of African countries in 1962 Democratic?

These slides available at http://bit.ly/mha17-day2

Four common types of problems in inference

- Test for Single Mean
- Test for Single Proportion
- Test for Difference in Means
- Test for Difference in Proportions

These slides available at http://bit.ly/mha17-day2

Four common types of problems in inference

- Test for Single Mean
- Test for Single Proportion
- Test for Difference in Means
- Test for Difference in Proportions

Match each of these tests with the problems on the previous slide

- To assist think about
  - how many variables are in the problem
  - what type of variable each is

These slides available at http://bit.ly/mha17-day2

Answers

1. Was the proportion of Autocratic countries lower in 2007 than in 1957? - Test for Difference in Proportions

These slides available at http://bit.ly/mha17-day2

Answers

1. Was the proportion of Autocratic countries lower in 2007 than in 1957? - Test for Difference in Proportions
2. Did European countries have a statistically different mean GDP per capita in 2002 than Asian countries? - Test for Difference in Means

These slides available at http://bit.ly/mha17-day2

Answers

1. Was the proportion of Autocratic countries lower in 2007 than in 1957? - Test for Difference in Proportions
2. Did European countries have a statistically different mean GDP per capita in 2002 than Asian countries? - Test for Difference in Means
3. Was the mean life expectancy for countries in South America in 2007 greater than 75 years? - Test for Single Mean

These slides available at http://bit.ly/mha17-day2

Answers

1. Was the proportion of Autocratic countries lower in 2007 than in 1957? - Test for Difference in Proportions
2. Did European countries have a statistically different mean GDP per capita in 2002 than Asian countries? - Test for Difference in Means
3. Was the mean life expectancy for countries in South America in 2007 greater than 75 years? - Test for Single Mean
4. Were the majority of African countries in 1962 Democratic? - Test for Single Proportion

These slides available at http://bit.ly/mha17-day2

Was the proportion of Autocratic countries lower in 2007 than in 1957?

Test for Difference in Proportions

- Data for this problem can be viewed interactively here
- The summarized data to be entered into StatKey is at the top of the interactive data
- Back into StatKey

These slides available at http://bit.ly/mha17-day2

Did European countries have a statistically different mean GDP per capita in 2002 than Asian countries?

Test for Difference in Means

- Data for this problem can be viewed interactively here
- Download a CSV of this data to upload in StatKey here
- Back into StatKey

These slides available at http://bit.ly/mha17-day2

Was the mean life expectancy for countries in South America in 2007 greater than 75 years?

Test for Single Mean

- Data for this problem can be viewed interactively here
- Download a CSV of this data to upload in StatKey here
- Back into StatKey

These slides available at http://bit.ly/mha17-day2

Were the majority of African countries in 1962 Democratic?

Test for Single Proportion

- Data for this problem can be viewed interactively here
- The summarized data to be entered into StatKey is at the top of the interactive data
- Back into StatKey

These slides available at http://bit.ly/mha17-day2

Resources

- ModernDive: An Introduction to Statistical and Data Sciences via R
- Free DataCamp Course (in development)
  Effective Data Storytelling using the tidyverse
- Inference Mind Map
- Filled-in Examples of "There is Only One Test"

These slides available at http://bit.ly/mha17-day2

Reflection

- What are your biggest take-aways from this workshop?
- What advice would you give to future workshop attendees about the workshop?
- What did you like most about the workshop?
- What could be improved for future workshop participants?

These slides available at http://bit.ly/mha17-day2

Thanks!

- Slides created via the R package `xaringan` by Yihui Xie.
- Email me at chester@pacificu.edu
- Source code for these slides is on GitHub

These slides available at http://bit.ly/mha17-day2