DSC 10, Spring 2018
Lecture 20

Designing Experiments

sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

Last Time

The Normal Distribution

Every bell-shaped curve is called "the normal distribution"

- The average (center) could be different
- The standard deviation (spread) could be different
- These two numbers alone determine the whole shape

Percent in Range	**All Distributions**	**Normal Distribution**
average $\pm$ 1 SD	at least 0%	about 68%
average $\pm$ 2 SDs	at least 75%	about 95%
average $\pm$ 3 SDs	at least 88.888...%	about 99.73%

Central Limit Theorem

If the sample is

- large, and
- drawn at random with replacement,

Then, *regardless of the distribution of the population,*

the probability distribution of the sample average (or sample sum or sample proportion) is roughly bell-shaped

Variability of the Sample Mean

- Fix a large sample size
- Draw *all* possible random samples of that size
- Compute the mean of each sample (lots of them)
- The distribution of those is the *probability distribution of the sample mean*
- It's a normal distribution, centered at the population mean

sample mean's average = population average

sample mean's SD = (population SD) / $\sqrt{\text{sample size}}$

Discussion Question 1

Population: Incomes with mean \$10,000 and SD \$20,000

Sample: 100 chosen uniformly at random with replacement

What's the chance that the sample average is above \$14,000?

A. 2.5%
B. 37%
C. 75%
D. I need a hint

sample mean's average = population average

sample mean's SD = (population SD) / $\sqrt{\text{sample size}}$

Percent in Range	All Distributions	Normal Distribution
average $\pm$ 1 SD	at least 0%	about 68%
average $\pm$ 2 SDs	at least 75%	about 95%
average $\pm$ 3 SDs	at least 88.888...%	about 99.73%

Discussion Question 1: Solution

Population: Incomes with mean \$10,000 and SD \$20,000

Sample: 100 chosen uniformly at random with replacement

Question: What's the chance that the sample average is above \$14,000?

- SD of sample mean = population SD / $\sqrt{\text{sample size}}$
  = \$20,000 / 10
  = \$2,000
- \$14,000 is 2 SD above the population mean
- About 95% are within 2 SD of the population mean
- About **2.5% are above**; about 2.5% are below

Discussion Question 2

Population: A perfect bell shape. Mean 10; SD 20

Sample: 100 chosen uniformly at random with replacement

What's the chance that *all* are below 50?

A. 2.5%
B. 95%
C. 97.5%
D. None of the above
E. I need a hint

sample mean's average = population average

sample mean's SD = (population SD) / $\sqrt{\text{sample size}}$

Percent in Range	All Distributions	Normal Distribution
average $\pm$ 1 SD	at least 0%	about 68%
average $\pm$ 2 SDs	at least 75%	about 95%
average $\pm$ 3 SDs	at least 88.888...%	about 99.73%

Discussion Question 2: Solution

Population: A perfect bell shape. Mean 10; SD 20
Sample: 100 chosen uniformly at random with replacement
Question: What's the chance that *all* are below 50?

- 50 is 2 population SD above the population mean
- The chance of drawing one value below 50 is 97.5%
- The chance of drawing 100 below 50 is **0.975 ** 100**

Discussion Question 3

You want to estimate the height of the tallest person on campus. You sample 100 people at random and compute a 99.9999% confidence interval using the bootstrap. Its upper bound is 6'4".

A 6'5" person walks by! What might have gone wrong?

A. Standard deviation of the population is too large to estimate
B. Sample size is too small for 99.9999% confidence interval
C. Height of tallest person is difficult to estimate with bootstrap
D. Empirical distribution of height of tallest person is not bell-shaped
E. More than one of the above

Discussion Question 4

You want to estimate the average compensation for SF workers by randomly sampling workers.

How many workers should you sample at random in order to get a 95% confidence interval with a width of $10,000 or less?

(Demo)

Choosing a Sample Size

Designing your sample

- You want to estimate what proportion of voters will vote for Candidate A in an upcoming election.

- How many people should you sample at random in order to get a 95% confidence interval with a width of 0.03 or less?

Width of 95% Confidence Interval

- A sample proportion is a sample mean, so CLT applies
- CLT says the distribution of a sample proportion is roughly normal, centered at population proportion
- **95%** confidence interval:
  - Center **± 2 SDs** of the sample proportion
- Total width = 4 SDs of the sample proportion

  = 4 x (population SD)/$\sqrt{\text{(sample size)}}$

Control the Width

- Suppose you want a width of no more than 0.03
- Total width = 4 SDs of the sample proportion

  = 4 x (population SD)/$\sqrt{\text{(sample size)}}$ ≤ 0.03

- Solve for sample size

  sample size ≥ (4 x (population SD) / 0.03)^2

Problems

- We don’t know the population SD.
- We have to take a sample, compute width of confidence interval, and adjust sample size.
    - Not practical to take a sample when trying to figure out how big of a sample to take...
- We aren’t guaranteed that our interval will be as narrow as we want.
- Can we address these issues?

(Demo)

Bound the Population SD

Fact: SD of population of 0's and 1's is always ≤ 0.5

sample size $\geq$ $(4 \times (0.5) / 0.03)^2$

$\geq$ $(4 \times (\text{population SD}) / 0.03)^2$

Choose a sample size of at least $(4 \times (0.5) / 0.03)^2 = 4445$.