DATA 8

Fall 2016

Lecture 27, October 28

Variability of the Sample Mean

Slides created by Ani Adhikari and John DeNero

Announcements

- Work on your project!
- A short homework will be released today.
- The server will be down (intentionally) during lecture, for work needed to prevent Dirty COW problems. Google it!

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

CLT for Sample Mean

- Large random sample drawn with replacement from any population
- Probability histogram for the sample mean is
  - Roughly normal
  - Centered at the population mean
  - SD = ?????

(Demo)

Variability of the Sample Mean

- Fix a sample size
- Draw all possible random samples of that size
- Compute the mean of each sample
- You'll end up with a lot of means
- The distribution of those is the *probability distribution of the sample mean*
- It's roughly normal, centered at the population mean
- SD = (population SD)/$\sqrt{\text{(sample size)}}$

Accuracy of the Sample Mean

- Greater if SD of sample mean is smaller
- Doesn’t depend on population size
- Increases as sample size increases, because SD of sample mean decreases
- For 3 times the accuracy, you have to multiply the sample size by a factor of $3^2 = 9$
- **Square Root Law:** If you multiply sample size by a factor, accuracy goes up by the square root of the factor

Designing Your Sample

- Suppose you want to estimate a population %, e.g. what percent will vote for Candidate A
- Need to construct a confidence interval, and want a narrow one
- How large should the sample be?

Width of 95% Confidence Interval

- CLT says distribution of sample proportion is roughly normal, centered at population proportion
- 95% confidence interval:
  - Center **± 2 SD** of sample proportion
- That's 2 SDs of the sample proportion on both sides of the center
- Total width: 4 SDs of the sample proportion

$$= 4 \times (\text{SD of 0-1 population}) / \sqrt{(\text{sample size})}$$

Control the Width

- Suppose you're OK with the total width being 3% but no more
- $4 \times (\text{SD of 0-1 population}) / \sqrt{(\text{sample size})} \leq 0.03$
- $\sqrt{(\text{sample size})} \geq 4 \times (\text{SD of 0-1 population}) / 0.03$

(Demo)

Bound the 0-1 Population SD

- $\sqrt{(\text{sample size})} \geq 4 \times (\text{SD of 0-1 population})/0.03$
- **SD of 0-1 population $\leq$ 0.5**
- $\sqrt{(\text{sample size})} \geq 4 \times 0.5/0.03 = 66.6666\ldots$
- sample size $\geq (66.6666\ldots)^2 = 4444.44\ldots$
- sample size: 4445