Advanced Statistics
DS2003 (BDS-4A)
Lecture 04

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

24 February, 2022

Previous Lecture

- The Poisson Distribution
- Negative Binomial Distribution
- Normal Distribution

The Normal Distribution

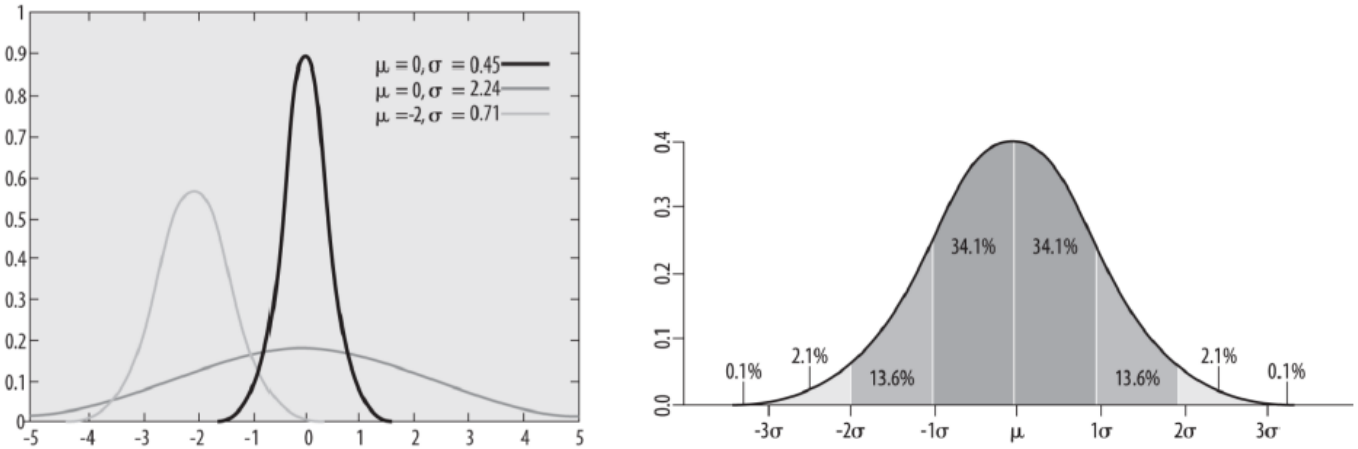


Figure 3.1: (left) All normal distributions have the same shape but differ to their μ and σ: they are shifted by μ and stretched by σ. (right) Percent of data failing into specified ranges of the normal distribution.

The Normal Distribution

- The amount of data that falls into the different regions:
  - $\mu \pm \sigma \rightarrow$ *68% data*
  - $\mu \pm 2\sigma \rightarrow$ *95% data*
  - $\mu \pm 3\sigma \rightarrow$ *99.7% data*

Point Estimates and Sampling Variability

Parameter estimation

- We are often interested in *population parameters*.
- Since complete populations are difficult (or impossible) to collect data on, we use *sample statistics* as *point estimates* for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the *margin of error* associated with our point estimate.
- But before we get to quantifying the variability among samples, let's try to understand how and why point estimates vary from sample to sample.

Suppose we randomly sample 1,000 adults from each state in the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Suppose we randomly sample 1,000 adults from each state in the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Suppose we randomly sample 1,000 adults from each state in the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Not the same, but only somewhat different.

Young, Underemployed and Optimistic

Coming of Age, Slowly, in a Tough Economy

Young adults hit hard by the recession. A plurality of the public (41%) believes young adults, rather than middle-aged or older adults, are having the toughest time in today's economy. An analysis of government economic data suggests that this perception is correct. The recent indicators on the nation's labor market show a decline in the

Tough economic times altering young adults' daily lives, long-term plans. While negative trends in the labor market have been felt most acutely by the youngest workers, many adults in their late 20s and early 30s have also felt the impact of the weak economy. Among all 18- to 34-year-olds, fully half (49%) say they have taken a job they didn't want just to pay the bills, with 24% saying they have taken an unpaid job to gain work experience. And more than one-third (35%) say that, as a result of the poor economy, they have gone back to school. Their personal lives have also been affected: 31% have postponed either getting married or having a baby (22% say they have postponed having a baby and 20% have put off getting married). One-in-four (24%) say they have moved back in with their parents after living on their own.

http://pewresearch.org/pubs/2191/young-adults-workers-labor-market-pay-careers-advancement-recession

Margin of error

The general public survey is based on telephone interviews conducted Dec. 6-19, 2011, with a nationally representative sample of 2,048 adults ages 18 and older living in the continental United States, including an oversample of 346 adults ages 18 to 34. A total of 769 interviews were completed with respondents contacted by landline telephone and 1,279 with those contacted on their cellular phone. Data are weighted to produce a final sample that is representative of the general population of adults in the continental United States. Survey interviews were conducted under the direction of Princeton Survey Research Associates International, in English and Spanish. Margin of sampling error is plus or minus 2.9 percentage points for results based on the total sample and 4.4 percentage points for adults ages 18-34 at the 95% confidence level.

41% ± 2.9%: We are 95% confident that 38.1% to 43.9% of the public believe young adults, rather than middle-aged or older adults, are having the toughest time in today's economy.

49% ± 4.4%: We are 95% confident that 44.6% to 53.4% of 18-34 years olds have taken a job they didn't want just to pay the bills.

Suppose the proportion of American adults who support the expansion of solar energy is $p = 0.88$, which is our parameter of interest. Is a randomly selected American adult more or less likely to support the expansion of solar energy?

Suppose the proportion of American adults who support the expansion of solar energy is $p = 0.88$, which is our parameter of interest. Is a randomly selected American adult more or less likely to support the expansion of solar energy?

More likely.

Suppose that you don’t have access to the populaion of all American adults, which is a quite likely scenario. In order to estimate the proportion of American adults who support solar power expansion, you might sample from the population and use your sample propor- tion as the best guess for the unknown population proportion.

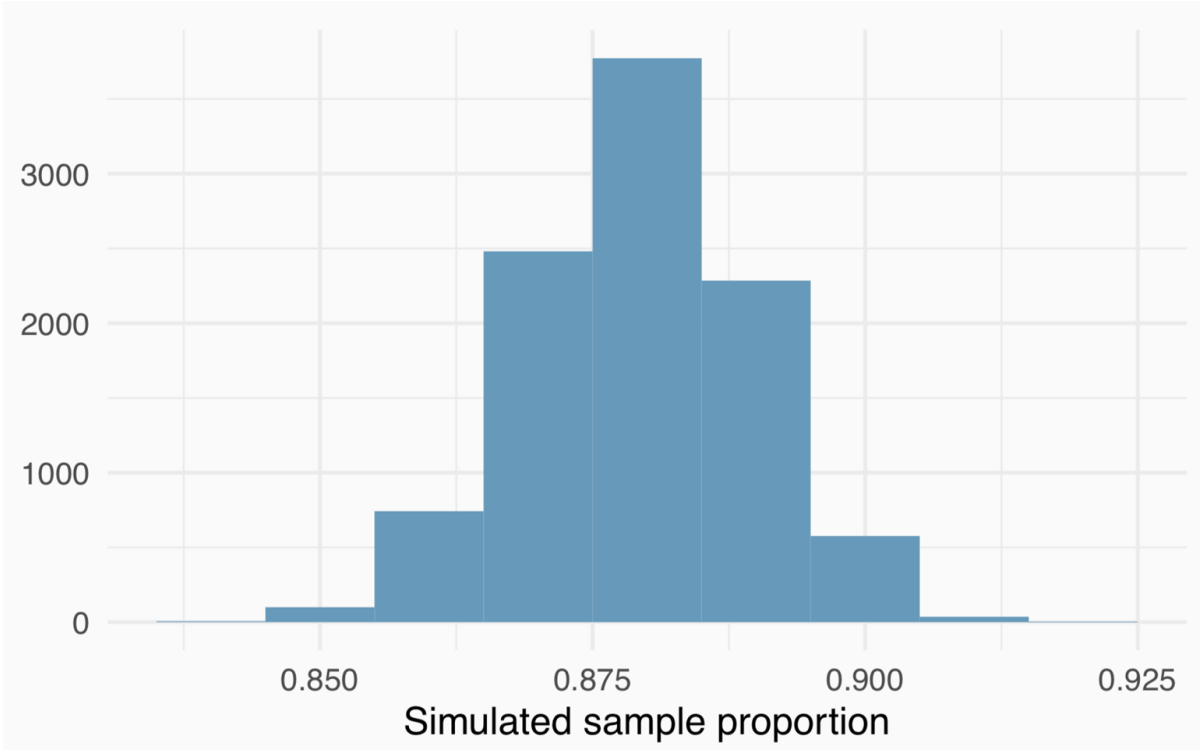
- Sample, with replacement, 1000 American adults from the population, and record whether they support solar power or not expansion.
- Find the sample proportion.
- Plot the distribution of the sample proportions obtained by members of the lass.

```
# 1. Create a set of 250 million entries, where 88% of
# them are "support" and 12% are "not".

pop_size <- 250000000
possible_entries <- c(rep("support", 0.88 * pop_size),
                      rep("not", 0.12 * pop_size))

# 2. Sample 1000 entries without replacement.

sampled_entries <- sample(possible_entries, size = 1000)

# 3. Compute p-hat: count the number that are "support",
# then divide by # the sample size.

sum(sampled sampled_entries == "support") / 1000
```

Sampling distribution

Suppose you were to repeat this process many times and plot the results.

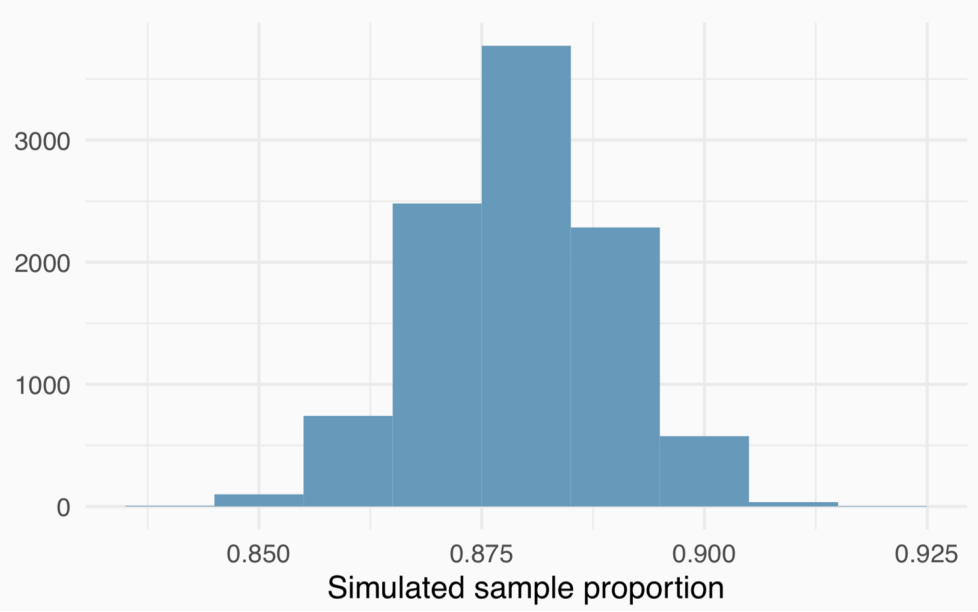
What you just constructed is called a sampling distribution.



Sampling distribution

What is the shape and center of this distribution?

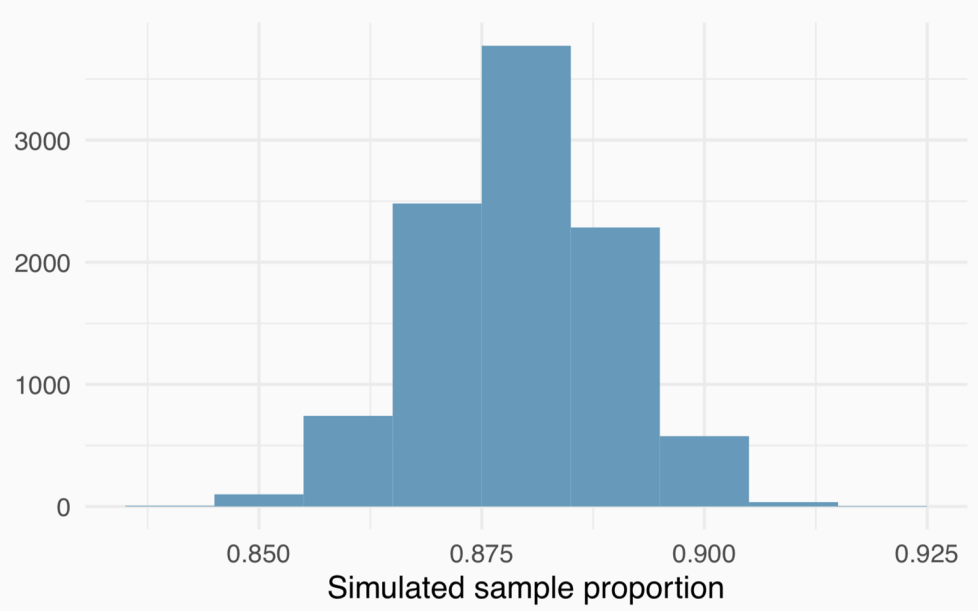
The distribution looks symmetric and somewhat bell-shaped.



Sampling distribution

Based on this distribution, what do you think is the true population proportion?

The center of the distribution: about 0.88.



Sampling distributions are never observed

- In real-world applications, we never actually observe the sampling distribution, yet it is useful to always think of a point estimate as coming from such a hypothetical distribution.

- Understanding the sampling distribution will help us characterize and make sense of the point estimates that we do observe.

Central Limit Theorem

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

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

It wasn't a coincidence that the sampling distribution we saw earlier was symmetric, and centered at the true population population.

We won't go through a detailed proof of why $SE = \sqrt{\frac{p\ (1-p)}{n}}$

but note that as n increases SE decreases.

- As n increases samples will yield more consistent $\hat{p}$s, i.e. variability among $\hat{p}$s will be lower.

Sources

- openintro.org/os (Chapter 5.1)