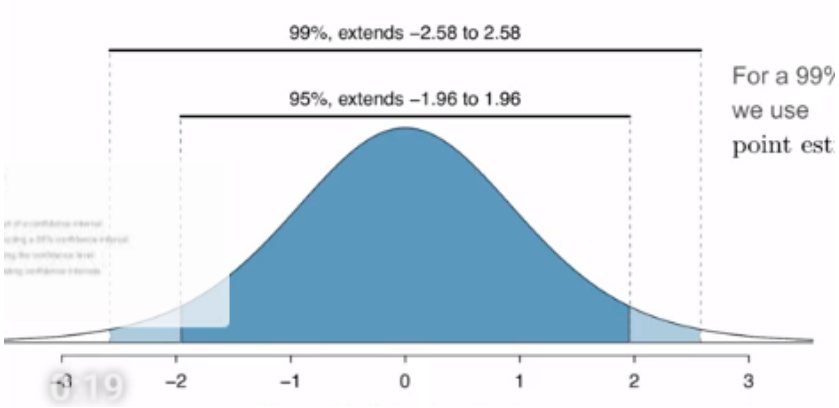
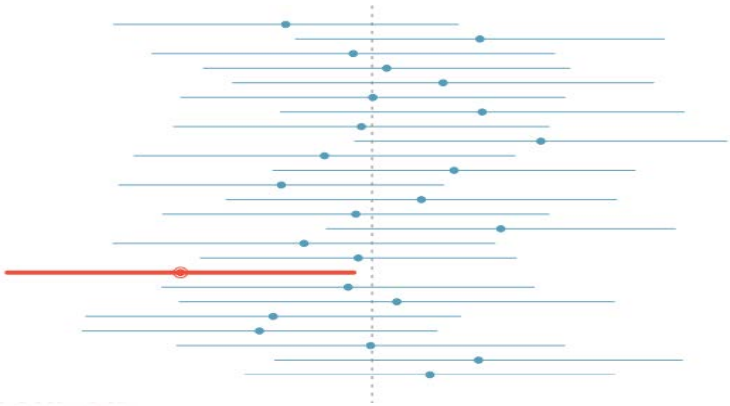
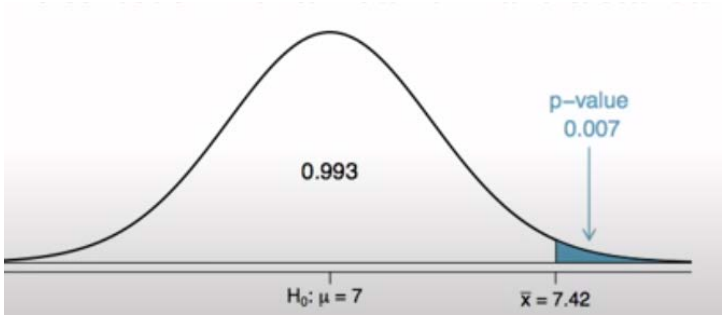
foundations for inference

- sampling distribution
- central limit theorem (CLT)







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.

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.

foundations for inference

sampling variability

central limit theorem

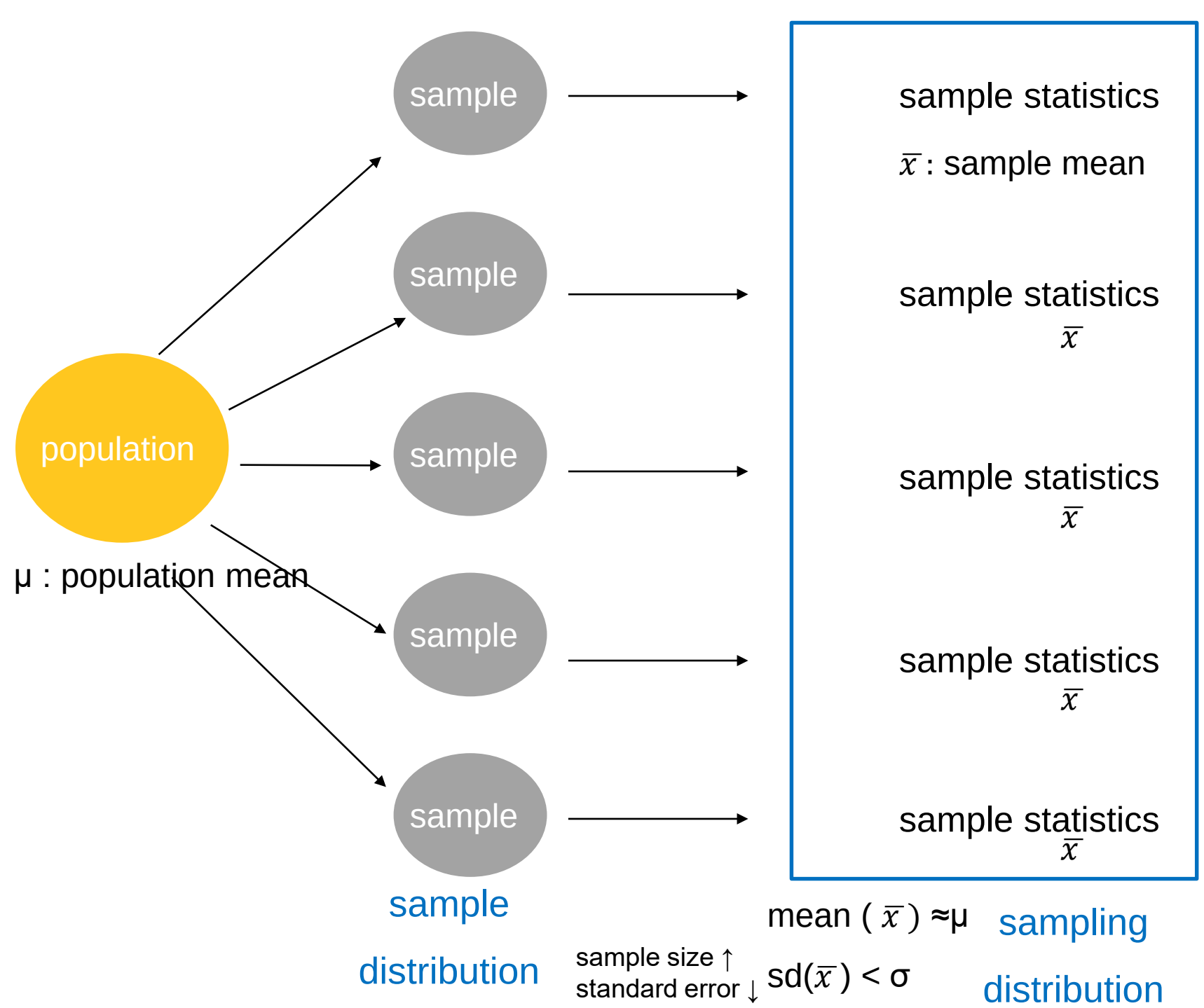
statistical inference

confidence interval

hypothesis testing

significance level

statistical power


sample
sample
sample
sample
sample
population
μ : population mean
sample statistics
$\bar{x}$: sample mean
sample statistics
$\bar{x}$
sample statistics
$\bar{x}$
sample statistics
$\bar{x}$
sample statistics
$\bar{x}$
sample
distribution
sample size ↑
standard error ↓
mean ($\bar{x}$) ≈μ
sd($\bar{x}$) < σ
sampling
distribution

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?

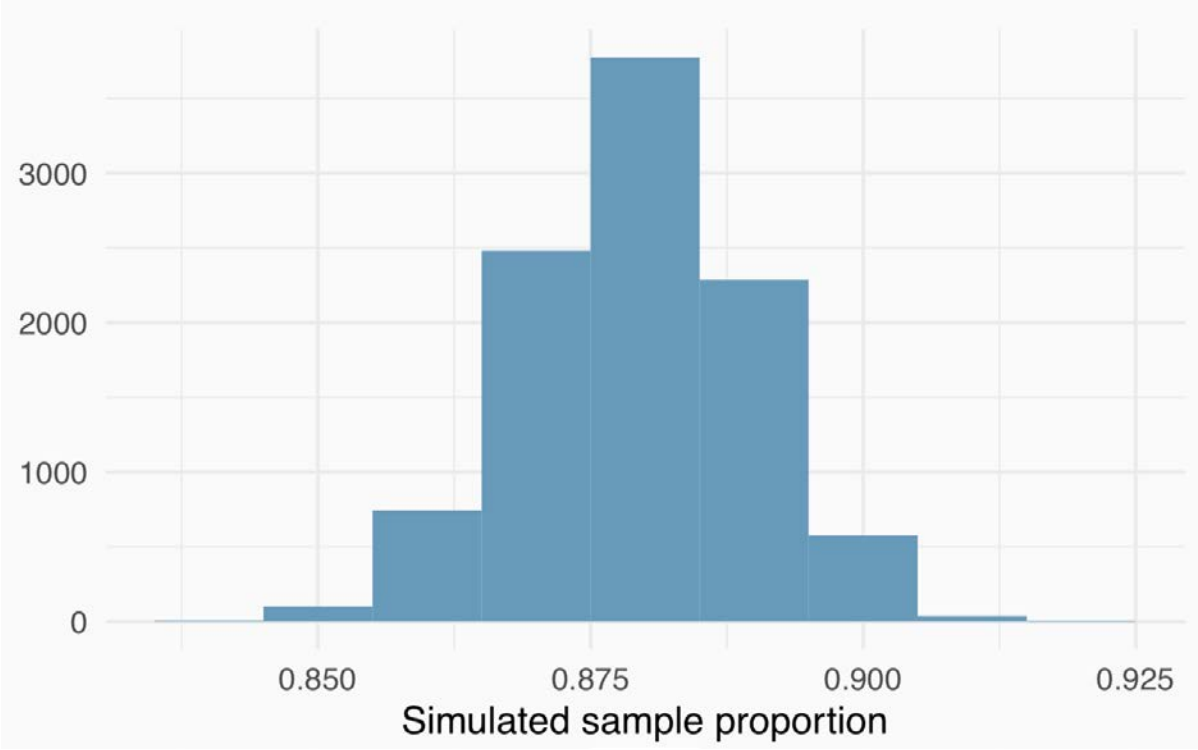
More likely.

Suppose that you don’t have access to the populaion of all Ameri- can 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 proportion 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.

sampling distribution

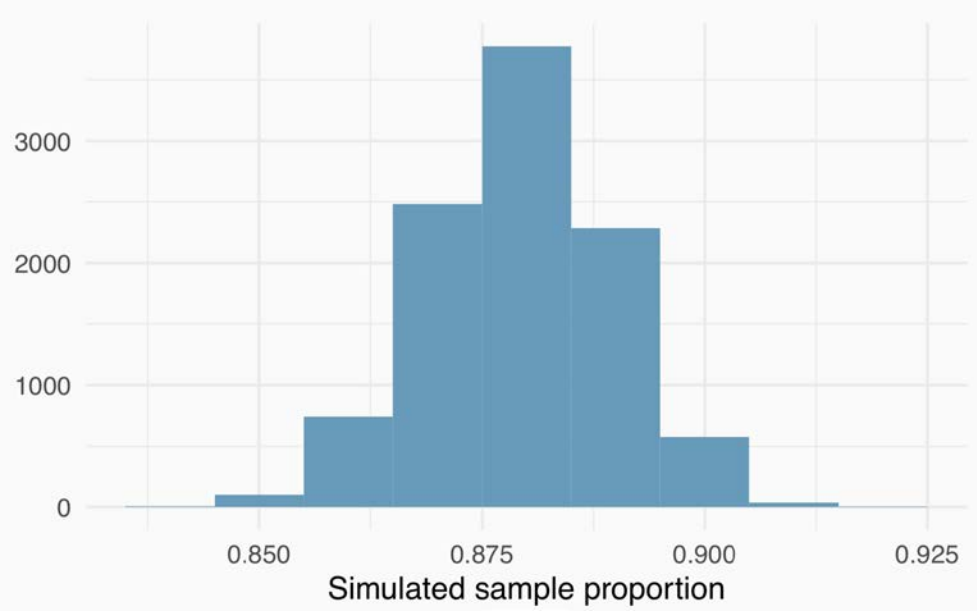
Suppose you were to repeat this process many times and plot the results. What you just constructed is called a sampling distribution.



sampling distribution

What is the shape and center of this distribution?

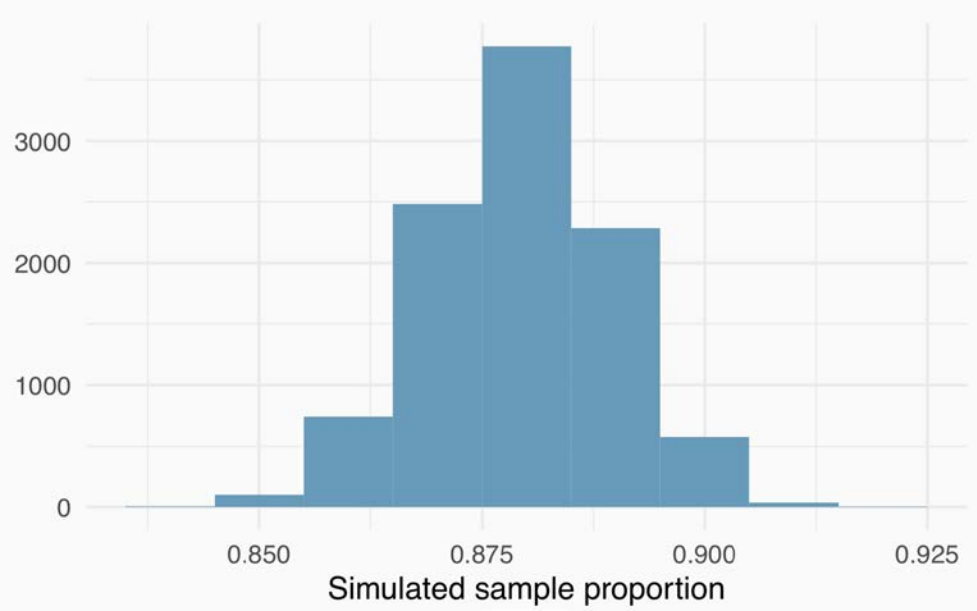
The distribution looks symmetric and somewhat bell-shaped.



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 (CLT): The distribution of sample means is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size.

$$\bar{x} \sim N(mean = \mu, SE = \frac{\sigma}{\sqrt{n}})$$

Conditions for the CLT:

Independence: sampled observations must be independent

- random sample/assignment
- if sampling without replacement, n<10% of population

Sample size/skew: Either the population distribution is normal, or if the population distribution is skewed, the sample sixe is large (*rule of thumb:n>30)*

This distribution of the population is also something very difficult to verify because we often do not know what the population looks like.

Practice

Which of the below visualizations is not appropriate for checking the shape of the distribution of the sample, and hence the population?

a) histogram

b) boxplot

c) normal probability plot

d) barplot

Homework

Problem 1: Suppose you have a slightly right skewed population distribution of annua incomes in a developed nation, with mean $30,000 and standard deviation $20,000. Suppose you take 10,000 random samples of size 625 from this population. Which of the following is most likely to be the distribution of the means of these samples?

a) Right skewed, mean = $30000, SD = $20000

b) Nearly normal, mean = $30000, SD = $20000

b) Nearly normal, mean = $30000, SD = $20000 / $\sqrt{625}$ = $800

c) Nearly normal, mean = $30000, SD = $20000 / $\sqrt{10000}$= $200

d) Left skewed, mean = $30000, SD = $20000 / $\sqrt{625}$ = $800

Homework

Problem 2: Find a news or research report in which a sample proportion is cited.

Discuss how it is used to estimate a population proportion.

Problem 3: Find a news or research report in which a sample mean is cited.

Discuss how it is used to estimate a population mean.