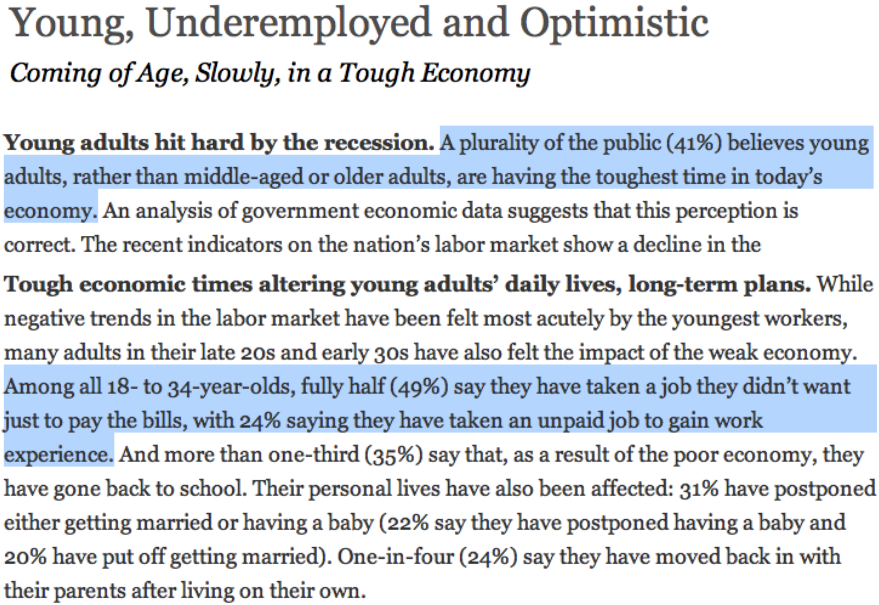
STATISTICS

Marco Caserta
marco.caserta@ie.edu

IE University

1. Variability in estimates
2. Examining the Central Limit Theorem
3. Sampling Distribution of Sample Statistics

1. Variability in estimates
   - Central Limit Theorem
   - Sampling distributions - via CLT
2. Examining the Central Limit Theorem
3. Sampling Distribution of Sample Statistics



Young, Underemployed and Optimistic

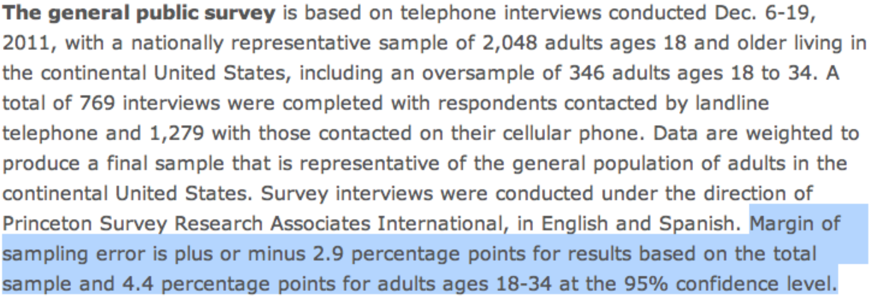
Coming of Age, Slowly, in a Tough Economy

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

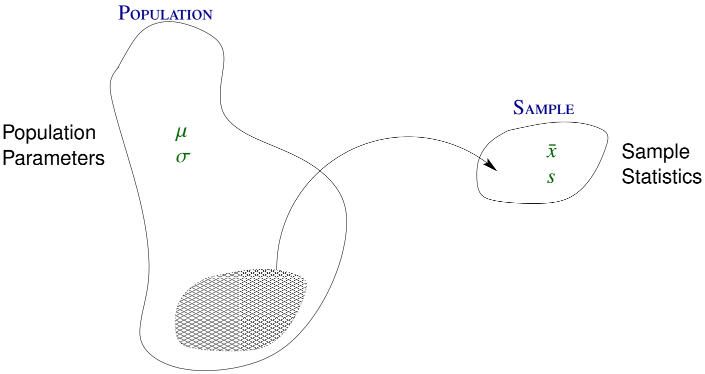
Margin of error



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.

From Population to Sample



Population

Population Parameters

μ

σ

Sample

$\bar{x}$

s

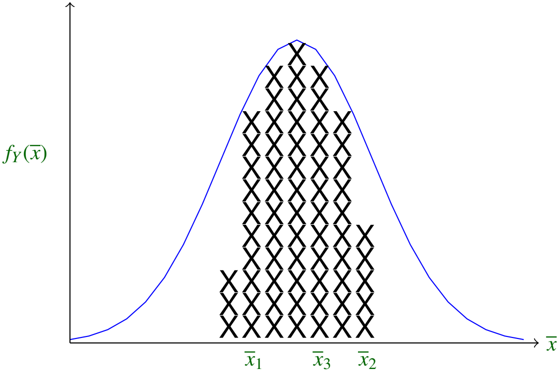
Sample Statistics

Sample Statistics

- Sample statistics are *random variables* used to estimate population parameters. Since they are random variables, we cannot conclude anything at all from a single observation of any of such statistics.
- Given a population parameter, *e.g.*, μ, we could design different sample statistics, *e.g.*, sample mean $\overline{x}$, sample median m, etc., to estimate the parameter. Which estimator is better?

Central Limit Theorem

- CLT to draw useful conclusions about how to estimate the (unknown) population mean μ

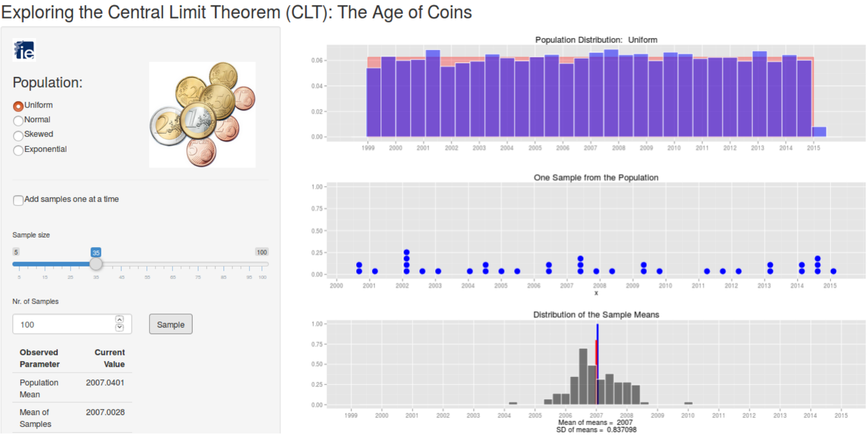


- Sample # 1 → $\overline{x}_1$
- Sample # 2 → $\overline{x}_2$
- Sample # 3 → $\overline{x}_3$
- …

Properties of the distribution function of sample mean $\overline{x}$:

I. mean of distribution *of the samples means* $E(\overline{x}) = \mu$ (UNBIASED ESTIMATOR)

II. st. dev. of distribution *of the samples means* $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$ (MINIMUM VARIANCE ESTIMATOR)



Exploring the Central Limit Theorem (CLT): The Age of Coins

Population:

- ● Uniform
- ○ Normal
- ○ Skewed
- ○ Exponential

□ Add samples one at a time

Sample size: 35 (5–100)

Nr. of Samples: 100

Sample

Observed Parameter	Current Value
Population Mean	2007.0401
Mean of Samples	2007.0028

Parameter estimation

The Idea Behind the Sampling Process

- 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.

Central Limit Theorem

Central Limit Theorem

The distribution of the sample mean is well approximated by a normal model:

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

where SE is represents STANDARD ERROR, which is defined as the standard deviation of the sampling distribution. If σ is unknown, use s.

- It wasn't a coincidence that the sampling distribution we saw earlier was symmetric, and centered at the true population mean.
- Note: Since $SE = \frac{\sigma}{\sqrt{n}}$, as n increases SE decreases.
  - As the sample size increases we would expect samples to yield more consistent sample means, hence the variability among the sample means would be lower.
- Why $E(\bar{x}) = \mu$ and $SE = \frac{\sigma}{\sqrt{n}}$? *Prove it!*

CLT - CONDITIONS

Certain conditions must be met for the CLT to apply:

1. INDEPENDENCE: Sampled observations must be independent. This is difficult to verify, but is more likely if
   - random sampling/assignment is used, and
   - if sampling without replacement, $n <$ 10% of the population.

CLT - conditions

Certain conditions must be met for the CLT to apply:

1. Independence: Sampled observations must be independent. This is difficult to verify, but is more likely if
   - random sampling/assignment is used, and
   - if sampling without replacement, $n <$ 10% of the population.
2. Sample size/skew: Either the population distribution is normal, or if the population distribution is skewed, the sample size is large.
   - the more skewed the population distribution, the larger sample size we need for the CLT to apply
   - for moderately skewed distributions $n > 30$ is a widely used rule of thumb

   This is also difficult to verify for the population, but we can check it using the sample data, and assume that the sample mirrors the population.

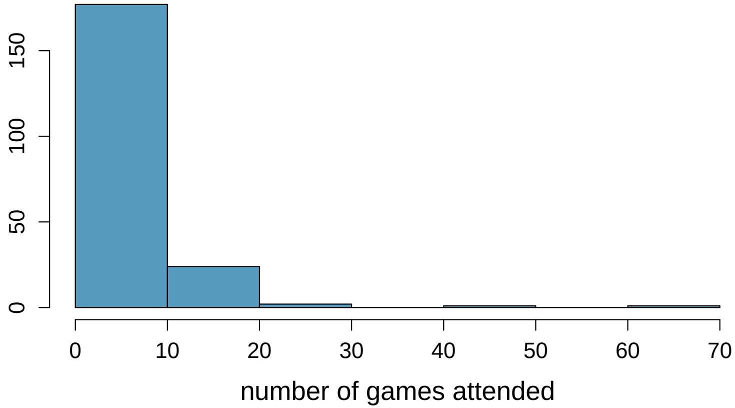
1. Variability in estimates
   - Central Limit Theorem
   - Sampling distributions - via CLT

2. Examining the Central Limit Theorem

3. Sampling Distribution of Sample Statistics

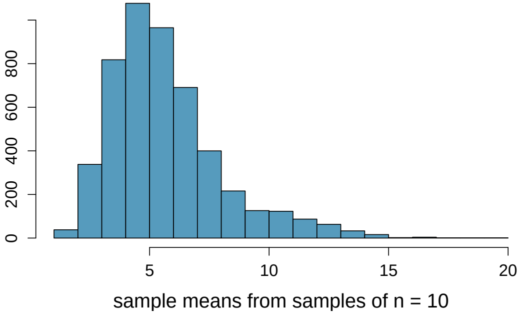
Average number of basketball games attended

Next let's look at the population data for the number of basketball games attended:



Average number of basketball games attended (cont.)

Sampling distribution, n = 10:

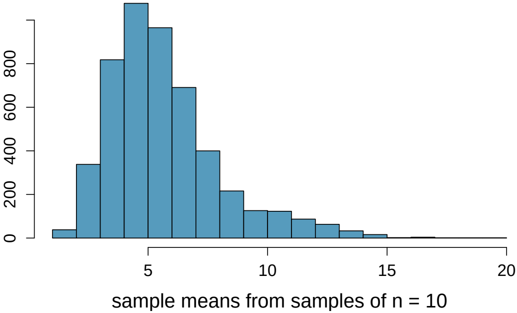


What does each observation in this distribution represent?

Is the variability of the sampling distribution smaller or larger than the variability of the population distribution? Why?

Average number of basketball games attended (cont.)

Sampling distribution, n = 10:



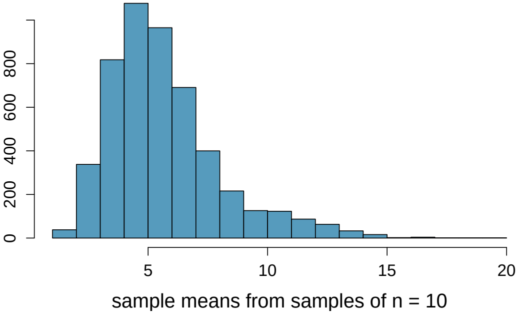
What does each observation in this distribution represent?

Sample mean ($\bar{x}$) of samples of size $n = 10$.

Is the variability of the sampling distribution smaller or larger than the variability of the population distribution? Why?

Average number of basketball games attended (cont.)

Sampling distribution, n = 10:



What does each observation in this distribution represent?

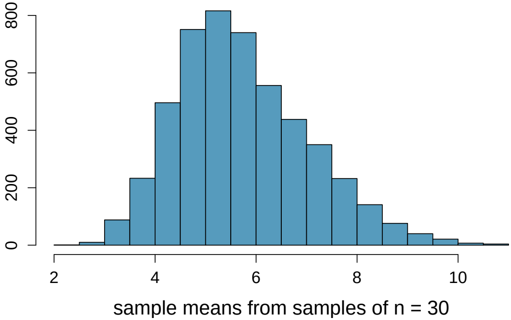
Sample mean ($\bar{x}$) of samples of size $n = 10$.

Is the variability of the sampling distribution smaller or larger than the variability of the population distribution? Why?

Smaller, sample means will vary less than individual observations.

Average number of basketball games attended (cont.)

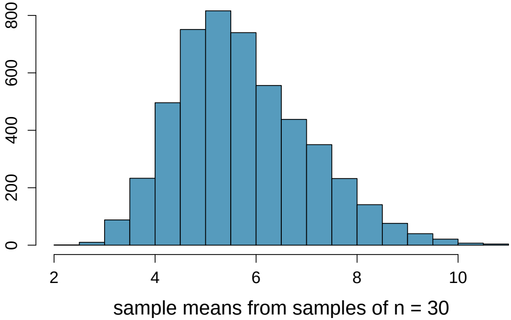
Sampling distribution, n = 30:



How did the shape, center, and spread of the sampling distribution change going from $n = 10$ to $n = 30$?

Average number of basketball games attended (cont.)

Sampling distribution, n = 30:

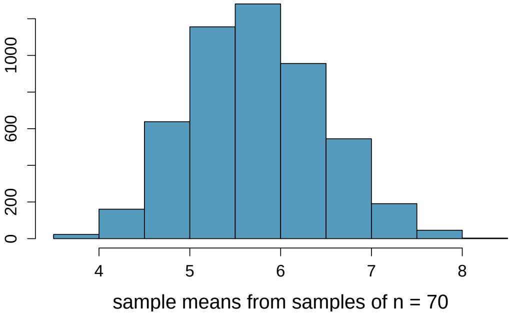


How did the shape, center, and spread of the sampling distribution change going from $n = 10$ to $n = 30$?

Shape is more symmetric, center is about the same, spread is smaller.

Average number of basketball games attended (cont.)

Sampling distribution, n = 70:



Average number of basketball games attended (cont.)

The mean of the sampling distribution is 5.75, and the standard deviation of the sampling distribution (also called the standard error) is 0.75. Which of the following is the most reasonable guess for the 95% confidence interval for the true average number of basketball games attended by students?

(a) 5.75 ± 0.75

(b) $5.75 \pm 2 \times 0.75$

(c) $5.75 \pm 3 \times 0.75$

(d) cannot tell from the information given

Average number of basketball games attended (cont.)

The mean of the sampling distribution is 5.75, and the standard deviation of the sampling distribution (also called the standard error) is 0.75. Which of the following is the most reasonable guess for the 95% confidence interval for the true average number of basketball games attended by students?

(a) 5.75 ± 0.75

(b) $5.75 \pm 2 \times 0.75 \;\rightarrow (4.25, 7.25)$

(c) $5.75 \pm 3 \times 0.75$

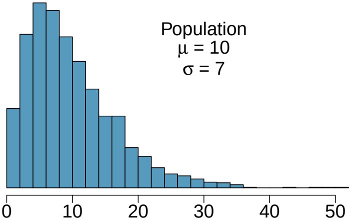
(d) cannot tell from the information given

Four plots: Determine which plot (A, B, or C) is which. We know: distribution for a population ($\mu = 10, \sigma = 7$).

(2) a single random sample of 100 observations from this population,

(3) a distribution of 100 sample means from random samples with size 7, and

(4) a distribution of 100 sample means from random samples with size 49.

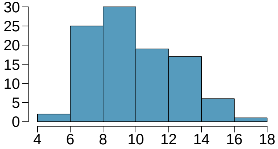


(A) A - (3); B - (2); C - (4)

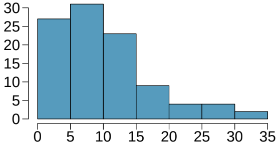
(B) A - (2); B - (3); C - (4)

(C) A - (3); B - (4); C - (2)

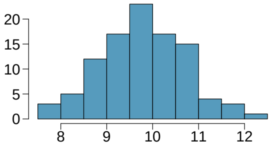
(D) A - (4); B - (2); C - (3)



Plot A



Plot B



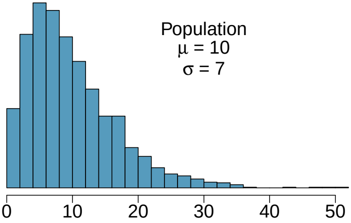
Plot C

Four plots: Determine which plot (A, B, or C) is which. We know: distribution for a population ($\mu = 10, \sigma = 7$).

(2) a single random sample of 100 observations from this population,

(3) a distribution of 100 sample means from random samples with size 7, and

(4) a distribution of 100 sample means from random samples with size 49.

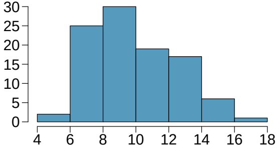


(A) *A - (3); B - (2); C - (4)*

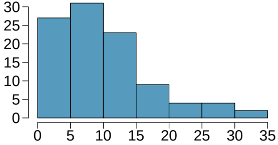
(B) A - (2); B - (3); C - (4)

(C) A - (3); B - (4); C - (2)

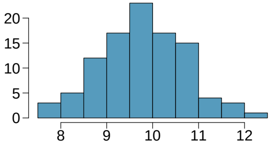
(D) A - (4); B - (2); C - (3)



Plot A



Plot B



Plot C

1. Variability in estimates
   - Central Limit Theorem
   - Sampling distributions - via CLT

2. Examining the Central Limit Theorem

3. Sampling Distribution of Sample Statistics

Sampling Distributions

- Sample Mean ⇒ $\bar{x} = \sum_i x_i/n \sim N(\mu, \sigma/\sqrt{n})$

- Sample Proportion

- Sample Variance

Sampling Distribution of Sample Proportions

- p is the proportion of the population having a certain characteristic
- The sample proportion $\hat{p}$ provides an estimate of p:

$$\hat{p} = \frac{X}{n}, \quad 0 \leq \hat{p} \leq 1$$

- $\hat{p}$ has a binomial distribution, but can be approximated by a normal distribution when $np(1p) > 5$
- $\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$
- $z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$

If the true proportion of voters who support Proposition A is P = .4, what is the probability that a sample of size 200 yields a sample proportion between .40 and .45?

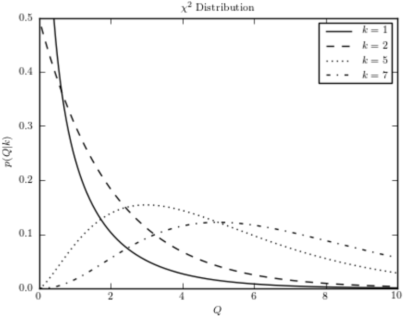
Sampling Distribution of Sample Variance

- The sampling distribution of s^2 has mean σ^2, *i.e.*: $E(s^2) = \sigma^2$
- If the population is normally distributed, then:

$$Var(s^2) = \frac{2\sigma^4}{n-1}$$

and $\frac{(n-1)s^2}{\sigma^2}$ follows a χ^2 distribution with $n-1$ degrees of freedom

- Degrees of freedom?


χ^2 Distribution
$k = 1$
$k = 2$
$k = 5$
$k = 7$
$p(Q|k)$
0.5
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
10
Q