Lecture 5b: Estimation procedures

Ernesto F. L. Amaral

September 28–October 03, 2023
Introduction to Sociological Data Analysis (SOCI 600)

www.ernestoamaral.com

Source: Healey, Joseph F. 2015. "Statistics: A Tool for Social Research." Stamford: Cengage Learning. 10th edition. Chapter 7 (pp. 160–184).

TEXAS A&M UNIVERSITY

Outline

- Explain the logic of estimation, role of the sample, sampling distribution, and population
- Define and explain the concepts of bias and efficiency
- Construct and interpret confidence intervals for sample means and sample proportions
- Explain relationships among confidence level, sample size, and width of the confidence interval

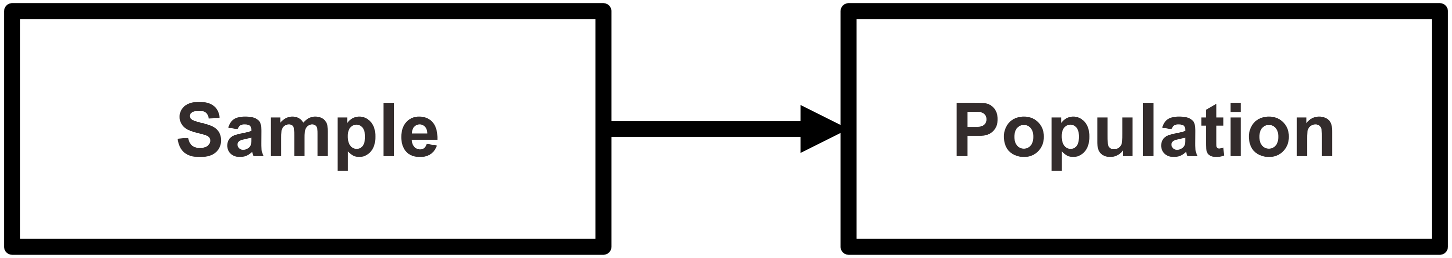
Sample and population

- In estimation procedures, statistics calculated from random samples are used to estimate the value of population parameters

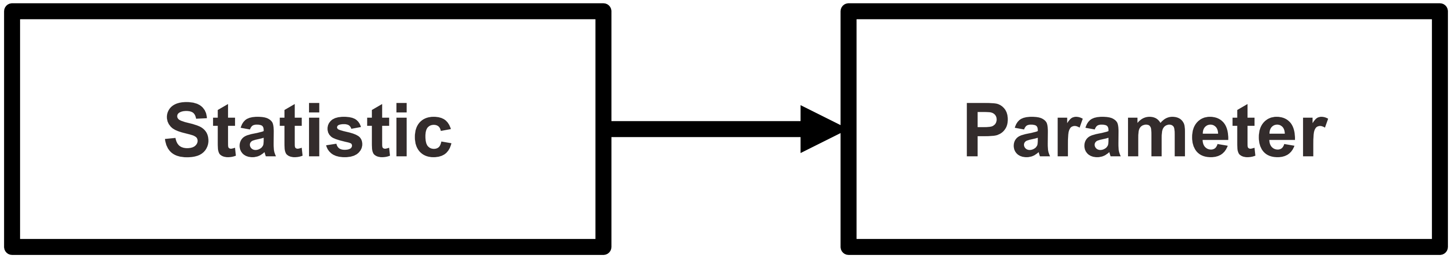
- Example
  - If we know that 42% of a random sample drawn from a city are Republicans, we can estimate the percentage of all city residents who are Republicans

Terminology

- Information from samples is used to estimate information about the population



- Statistics are used to estimate parameters



Basic logic

- Sampling distribution is the link between sample and population

- The values of the parameters are unknown, but the characteristics of the sampling distribution are defined by two theorems (previous chapter)

Two estimation procedures

- **A point estimate** is a sample statistic used to estimate a population value
  - 68% of a sample of randomly selected Americans support capital punishment (GSS 2010)
- **An interval estimate** consists of confidence intervals (range of values)
  - Between 65% and 71% of Americans approve of capital punishment (GSS 2010)
  - Most point estimates are actually interval estimates
  - Margin of error generates confidence intervals
  - Estimators are selected based on two criteria
    - Bias (mean) and efficiency (standard error)

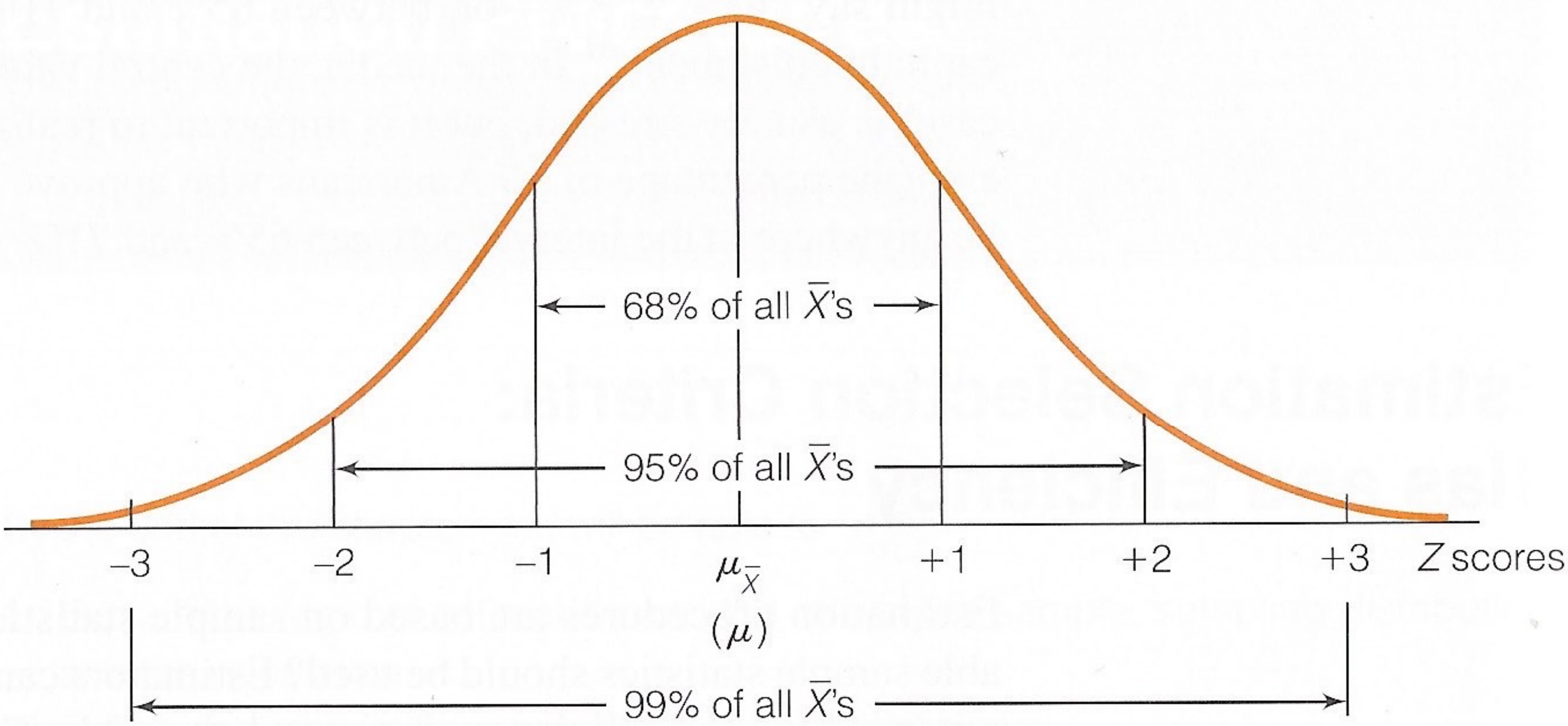
Bias

- An estimator is unbiased if the mean of its sampling distribution is equal to the population value of interest
- The mean of the sampling distribution of sample means ($\mu_{\bar{X}}$) is the same as the population mean (μ)
- Sample proportions (P_s) are also unbiased
  - If we calculate sample proportions from repeated random samples of size n...
  - Then, the sampling distribution of sample proportions will have a mean (μ_p) equal to the population proportion (P_u)
- Sample means and proportions are unbiased estimators
  - We can determine the probability that they are within a certain distance of the population values

Example

- Random sample to get income information
- Sample size (*n*): 500 households
- Sample mean ($\bar{X}$): $45,000
- Population mean (μ): unknown parameter
- Mean of sampling distribution ($\mu_{\bar{X}} = \mu$)
  - If an estimator ($\bar{X}$) is unbiased, it is probably an accurate estimate of the population parameter (μ) and sampling distribution mean ($\mu_{\bar{X}}$)
  - We use the sampling distribution (which has a normal shape) to estimate confidence intervals

Sampling distribution



Source: Healey 2015, p.162.

Efficiency

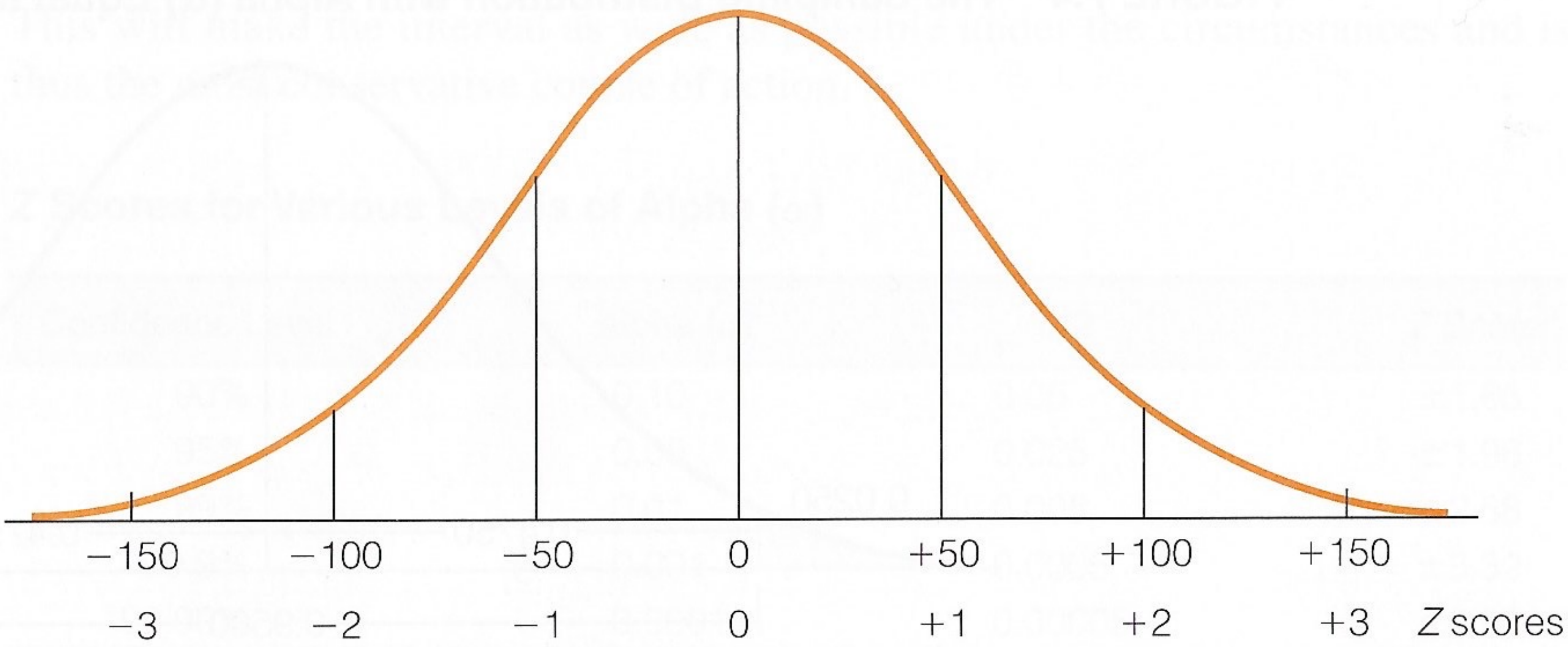
- Efficiency is the extent to which the sampling distribution is clustered around its mean
- Efficiency or clustering is a matter of dispersion
  - The smaller the standard deviation of a sampling distribution, the greater the clustering and the higher the efficiency
  - Larger samples have greater clustering and higher efficiency
  - Standard deviation of sampling distribution: $\sigma_{\bar{X}} = \sigma/\sqrt{n}$

Statistics	**Sample 1**	**Sample 2**
Sample mean	$\bar{X}_1 = \$45{,}000$	$\bar{X}_2 = \$45{,}000$
Sample size	$n_1 = 100$	$n_2 = 1000$
Standard deviation	$\sigma_1 = \$500$	$\sigma_2 = \$500$
Standard error	$\sigma_{\bar{X}} = 500/\sqrt{100} = \$50.00$	$\sigma_{\bar{X}} = 500/\sqrt{1000} = \$15.81$

Source: Healey 2015, p.163.

Sampling distribution

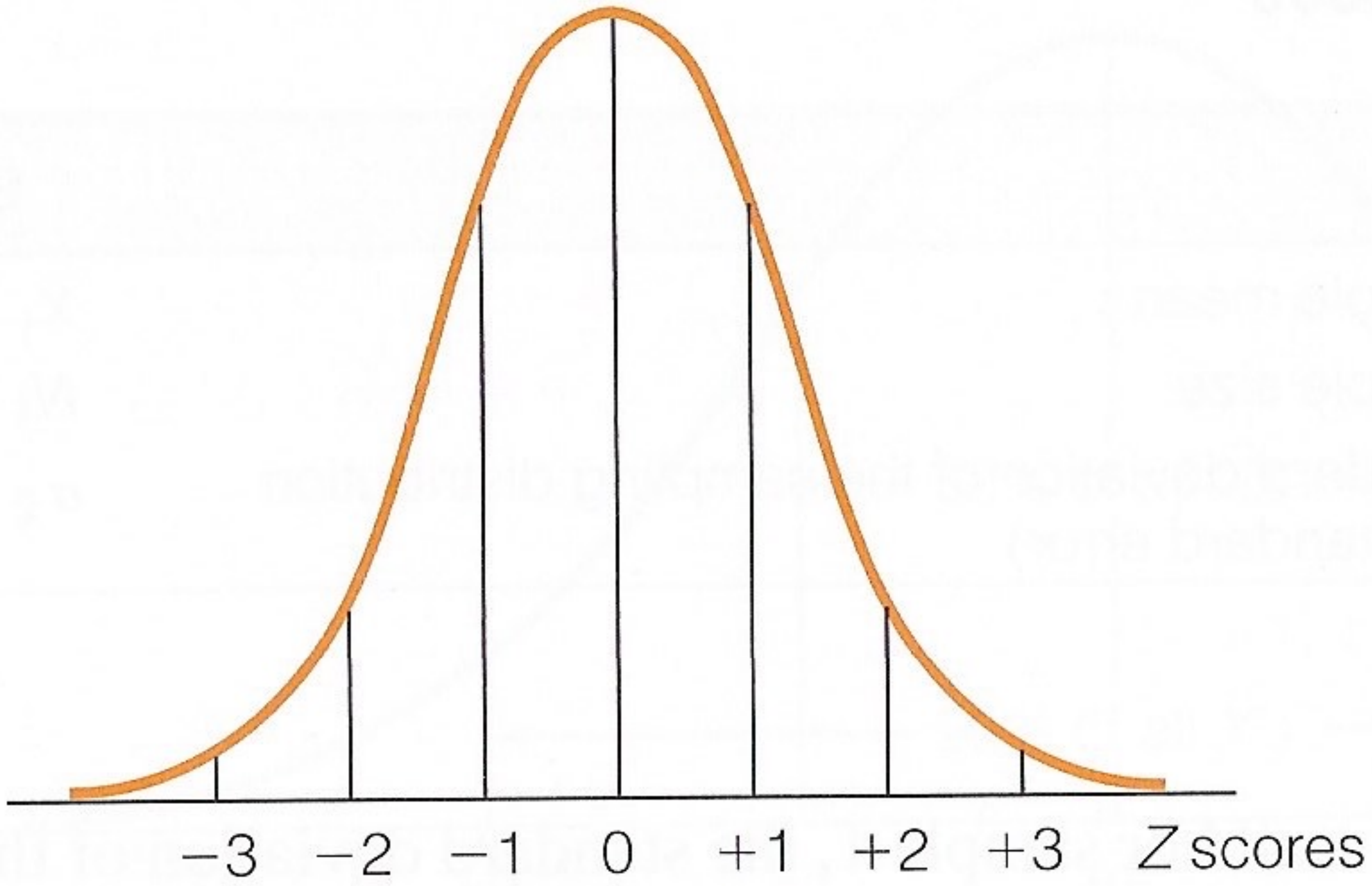
n = 100; $\sigma_{\bar{X}}$ = \$50.00



Source: Healey 2015, p.163.

Sampling distribution

n = 1000; $\sigma_{\bar{X}}$ = $15.81



Source: Healey 2015, p.164.

Confidence interval & level

- **Confidence interval** is a range of values used to estimate the true population parameter
  - We associate a confidence level (e.g. 0.95 or 95%) to a confidence interval
- **Confidence level** is the success rate of the procedure to estimate the confidence interval
  - Expressed as probability $(1–\alpha)$ or percentage $(1–\alpha)*100$
  - α is the complement of the confidence level
  - Larger confidence levels generate larger confidence intervals
- Confidence level of 95% is the most common
  - Good balance between precision (width of confidence interval) and reliability (confidence level)

Interval estimation procedures

- Set the alpha (α)
  - Probability that the interval will be wrong
- Find the *Z* score associated with alpha
  - In column c of Appendix A of textbook
    - If the *Z* score you are seeking is between two other scores, choose the larger of the two *Z* scores
  - In Stata: `display invnormal(α)`
- Substitute values into appropriate equation
- Interpret the interval

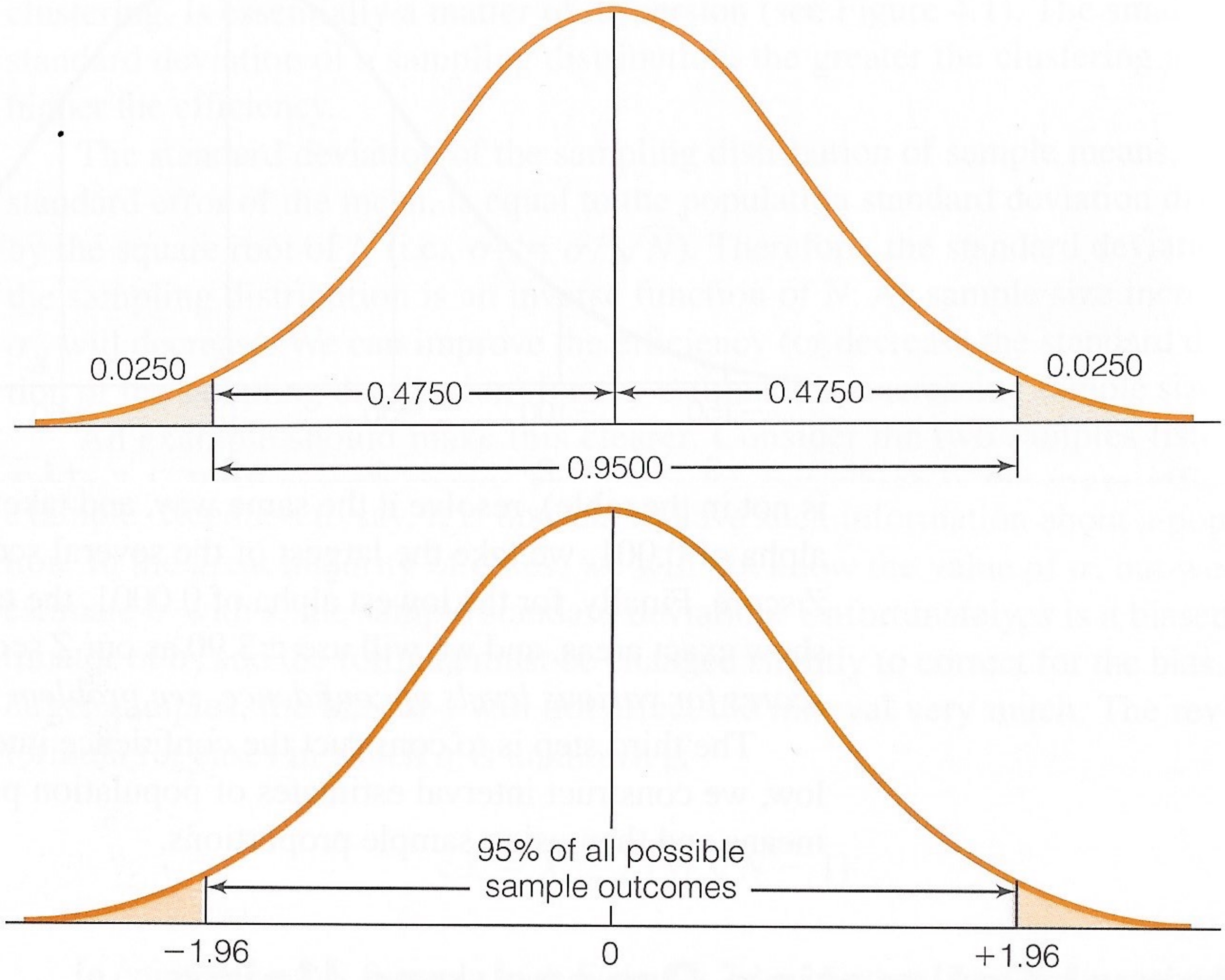
Example to find *Z* score

- Setting alpha (α) equal to 0.05
  - 95% confidence level: $(1-\alpha)*100$
  - We are willing to be wrong 5% of the time
- If alpha is equal to 0.05
  - Half of this probability is in the lower tail ($\alpha/2=0.025$)
  - Half is in the upper tail of the distribution ($\alpha/2=0.025$)
- Looking up this area, we find a Z = 1.96

```
di invnormal(.025)
-1.959964
```

```
di invnormal(1-.025)
di invnormal(.975)
1.959964
```

Finding Z for sampling distribution with $\alpha = 0.05$



Source: Healey 2015, p.164–165.

Confidence level, α, and Z

Confidence level (1 – α) * 100	**Significance level alpha (α)**	**α / 2**	***Z* score**
90%	0.10	0.05	$\pm$1.65
95%	**0.05**	**0.025**	**$\pm$1.96**
99%	0.01	0.005	$\pm$2.58
99.9%	0.001	0.0005	$\pm$3.32
99.99%	0.0001	0.00005	$\pm$3.90

Source: Healey 2015, p.165.

Confidence intervals for sample means

- For large samples (*n*≥100)
- Standard deviation (*σ*) **known** for population

$$c.i. = \bar{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$$

c.i. = confidence interval

$\bar{X}$ = sample mean

Z = score determined by the alpha level (confidence level)

$\sigma/\sqrt{n}$ = sample deviation of the sampling distribution (standard error of the mean)

$\pm Z(\sigma/\sqrt{n})$ = margin of error

Example for means: Large sample, σ known

- Sample of 200 residents
- Sample mean of IQ is 105
- Population standard deviation is 15
- Calculate a confidence interval with a 95% confidence level (α = 0.05)
  - Same as saying: calculate a 95% confidence interval

$$c.i. = \bar{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right) = 105 \pm 1.96\left(\frac{15}{\sqrt{200}}\right) = 105 \pm 2.08$$

  - Average IQ is somewhere between 102.92 (105–2.08) and 107.08 (105+2.08)

Interpreting previous example

$$n = 200;\ 102.92 \leq \mu \leq 107.08$$

- **Correct:** We are 95% certain that the confidence interval contains the true value of μ
  - If we selected several samples of size 200 and estimated their confidence intervals, 95% of them would contain the population mean (μ)
  - The 95% confidence level refers to the success rate to estimate the population mean (μ). It does not refer to the population mean itself
- **Wrong:** Since the value of μ is fixed, it is incorrect to say that there is a chance of 95% that the true value of μ is between the interval

Source: Triola 2008, p.250–303.

Confidence intervals for sample means

- For large samples ($n \geq 100$)
- Standard deviation (σ) **unknown** for population

$$c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{n-1}}\right)$$

c.i. = confidence interval

$\bar{X}$ = sample mean

Z = score determined by the alpha level (confidence level)

$s/\sqrt{n-1}$ = sample deviation of the sampling distribution (standard error of the mean)

$\pm Z(s/\sqrt{n-1})$ = margin of error

Example for means: Large sample, σ unknown

- Sample of 500 residents
- Sample mean income is $45,000
- Sample standard deviation is $200
- Calculate a 95% confidence interval

$$c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{n-1}}\right) = 45{,}000 \pm 1.96\left(\frac{200}{\sqrt{500-1}}\right)$$

$$c.i. = 45{,}000 \pm 17.54$$

  - Average income is between $44,982.46 (45,000–17.54) and $45,017.54 (45,000+17.54)

Example from ACS

- We are 95% certain that the confidence interval from $49,926.89 to $50,161.07 contains the true average wage and salary income for the U.S. population in 2018

```
. ***95% confidence level
. svy, subpop(if income!=. & income!=0): mean income
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =   2,351      Number of obs     =    3,214,539
Number of PSUs   = 1410976      Population size   =  327,167,439
                                Subpop. no. obs   =    1,574,313
                                Subpop. size      =  163,349,075
                                Design df         =    1,408,625

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
      income |   50043.98   59.74195      49926.89    50161.07
--------------------------------------------------------------

.
. ***Standard deviation
. estat sd

-------------------------------------
             |       Mean   Std. Dev.
-------------+-----------------------
      income |   50043.98    61547.67
-------------------------------------
```

Obs.: Only individuals with some wage and salary income are included (exclude those with zero income).

Source: 2018 American Community Survey.

Edited table

Table 1. Summary statistics for individual average wage and salary income of the U.S. population, 2018

Summary statistics	Value
Mean	50,043.98
Standard deviation	61,547.67
Standard error	59.74
95% confidence interval	
Lower bound	49,926.89
Upper bound	50,161.07
Sample size	1,574,313

Obs.: Only individuals with some wage and salary income are included (exclude those with zero income).
Source: 2018 American Community Survey.

Interpreting previous example

n = 1,574,313; 49,926.89 $\leq \mu \leq$ 50,161.07

- **Correct:** We are 95% certain that the confidence interval contains the true value of μ
  - If we selected several samples of size 1,574,313 and estimated their confidence intervals, 95% of them would contain the population mean (μ)
  - The 95% confidence level refers to the success rate to estimate the population mean (μ). It does not refer to the population mean itself
- **Wrong:** Since the value of μ is fixed, it is incorrect to say that there is a chance of 95% that the true value of μ is between the interval

Source: Triola 2008, p.250–303.

Example from GSS

- We are 95% certain that the confidence interval from $35,324.83 to $39,889.96 contains the true average income for the U.S. adult population in 2004

```
. svy: mean conrinc, over(year)
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =     307          Number of obs   =      4,522
Number of PSUs   =     597          Population size = 4,611.7099
                                    Design df       =        290

        2004: year = 2004
        2010: year = 2010
        2016: year = 2016

--------------------------------------------------------------
             |             Linearized
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
conrinc      |
        2004 |   37607.39   1159.734      35324.83    39889.96
        2010 |   31537.11   1216.566      29142.69    33931.53
        2016 |    34649.3   1267.614      32154.41    37144.19
--------------------------------------------------------------
Note: Variance scaled to handle strata with a single sampling
      unit.
```

Source: 2004, 2010, 2016 General Social Surveys.

Edited table

Table 1. Mean, standard error, 95% confidence interval, and sample size of individual average income of the U.S. adult population, 2004, 2010, and 2016

Year	Mean	Standard Error	95% Confidence Interval		Sample Size
			Lower Bound	Upper Bound	
2004	37,607.39	1,159.73	35,324.83	39,889.96	1,688
2010	31,537.11	1,216.57	29,142.69	33,931.53	1,202
2016	34,649.30	1,267.61	32,154.41	37,144.19	1,632

Source: 2004, 2010, 2016 General Social Surveys.

Confidence intervals for sample proportions

$$c.i. = P_s \pm Z\sqrt{\frac{P_u(1-P_u)}{n}}$$

c.i. = confidence interval

P_s = sample proportion

Z = score determined by the alpha level (confidence level)

$\sqrt{P_u(1-P_u)/n}$ = sample deviation of the sampling distribution (standard error of the proportion)

$\pm Z(\sqrt{P_u(1-P_u)/n})$ = margin of error

Note about sample proportions

- The formula for the standard error includes the population value
  - We do not know and are trying to estimate (P_u)
- By convention we set P_u equal to 0.50
  - The numerator [$P_u(1–P_u)$] is at its maximum value
  - $P_u(1–P_u) = (0.50)(1–0.50) = 0.25$
- The calculated confidence interval will be at its maximum width
  - This is considered the most statistically conservative technique

Example for proportions

- Estimate the proportion of students who missed at least one day of classes last semester
  - In a random sample of 200 students, 60 students reported missing one day of class last semester
  - Thus, the sample proportion is 0.30 (60/200)
  - Calculate a 95% (alpha = 0.05) confidence interval

$$c.i. = P_s \pm Z\sqrt{\frac{P_u(1-P_u)}{n}} = 0.3 \pm 1.96\sqrt{\frac{0.5(1-0.5)}{200}}$$

$$c.i. = 0.3 \pm 0.08$$

Example from ACS

- We are 95% certain that the confidence interval from 5.2% to 5.3% contains the true proportion of internal migrants in the U.S. population in 2018

```
. svy: prop migrant
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =   2,351                Number of obs    =     3,184,099
Number of PSUs   = 1410889                Population size =  323,541,502
                                          Design df        =     1,408,538

--------------------------------------------------------------------------
                       |             Linearized          Logit
                       |  Proportion  Std. Err.     [95% Conf. Interval]
-----------------------+--------------------------------------------------
               migrant |
           Non-migrant |   .9418963    .000259     .9413866    .9424019
      Internal migrant |   .0524799   .0002463     .0519993    .0529647
 International migrant |   .0056239   .0000823     .0054649    .0057874
--------------------------------------------------------------------------
```

Source: 2018 American Community Survey.

Edited table

Table 2. Summary statistics for migration status of the U.S. population, 2018

Migration status	Proportion	Standard Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Non-migrant	0.9419	0.0003	0.9414	0.9424
Internal migrant	0.0525	0.0003	0.0520	0.0530
International migrant	0.0056	0.0001	0.0055	0.0058

Obs.: Sample size of 3,184,099 individuals.
Source: 2018 American Community Survey.

Interpreting previous example

n = 3,184,099; $5.2 \leq P_u \leq 5.3$

- **Correct:** We are 95% certain that the confidence interval contains the true value of P_u
  - If we selected several samples of size 3,184,099 and estimated their confidence intervals, 95% of them would contain the population proportion (P_u)
  - The 95% confidence level refers to the success rate to estimate the population proportion (P_u). It does not refer to the population proportion itself
- **Wrong:** Since the value of P_u is fixed, it is incorrect to say that there is a chance of 95% that the true value of P_u is between the interval

Source: Triola 2008, p.250–303.

Example from GSS

- We are 95% certain that the confidence interval from 2.6% to 4.7% contains the true proportion of the U.S. adult population who thinks the number of immigrants to the country should increase a lot in 2004

```
. svy: prop letin1 if year==2004
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      109        Number of obs   =      1,983
Number of PSUs   =      218        Population size = 1,979.3435
                                   Design df       =        109

      _prop_1: letin1 = increased a lot
      _prop_2: letin1 = increased a little
      _prop_3: letin1 = remain the same as it is
      _prop_4: letin1 = reduced a little
      _prop_5: letin1 = reduced a lot
```

	Proportion	Linearized Std. Err.	[95% Conf.	Interval]
letin1				
_prop_1	.0348265	.005221	.0258369	.0467936
_prop_2	.0653852	.0060495	.0543699	.078447
_prop_3	.3517117	.0128957	.3265967	.3776749
_prop_4	.2829629	.0118188	.2601357	.3069621
_prop_5	.2651137	.0127052	.2407073	.2910462

Source: 2004 General Social Survey.

Edited table

Table 2. Proportion, standard error, 95% confidence interval, and sample size of opinion of the U.S. adult population about how should the number of immigrants to the country be nowadays, 2004, 2010, and 2016

Opinion About Number of Immigrants	Proportion	Standard Error	95% Confidence Interval		Sample Size
			Lower Bound	Upper Bound	
2004					1,983
Increase a lot	0.0348	0.0052	0.0258	0.0468	
Increase a little	0.0654	0.0060	0.0544	0.0784	
Remain the same	0.3517	0.0129	0.3266	0.3777	
Reduce a little	0.2830	0.0118	0.2601	0.3070	
Reduce a lot	0.2651	0.0127	0.2407	0.2910	
2010					1,393
Increase a lot	0.0426	0.0061	0.0320	0.0564	
Increase a little	0.0944	0.0096	0.0771	0.1152	
Remain the same	0.3589	0.0166	0.3268	0.3923	
Reduce a little	0.2452	0.0121	0.2220	0.2700	
Reduce a lot	0.2588	0.0146	0.2310	0.2887	
2016					1,845
Increase a lot	0.0586	0.0069	0.0462	0.0740	
Increase a little	0.1163	0.0091	0.0993	0.1358	
Remain the same	0.4028	0.0117	0.3797	0.4264	
Reduce a little	0.2305	0.0097	0.2118	0.2504	
Reduce a lot	0.1918	0.0101	0.1724	0.2128	

Source: 2004, 2010, 2016 General Social Surveys.

Width of confidence interval

- The width of confidence intervals can be controlled by manipulating the confidence level
  - The confidence level increases
  - The alpha decreases
  - The *Z* score increases
  - The confidence interval is wider

Example: $\bar{X}$ = \$45,000; *s* = \$200; *n* = 500

Confidence level	Alpha (α)	*Z* score	Confidence interval	Interval width
90%	0.10	$\pm$1.65	\$45,000 $\pm$ \$14.77	\$29.54
95%	0.05	$\pm$1.96	\$45,000 $\pm$ \$17.54	\$35.08
99%	0.01	$\pm$2.58	\$45,000 $\pm$ \$23.09	\$46.18
99.9%	0.001	$\pm$3.32	\$45,000 $\pm$ \$29.71	\$59.42

Width of confidence interval

- The width of confidence intervals can be controlled by manipulating the sample size
  - The sample size increases
  - The confidence interval is narrower

Example: $\bar{X}$ = $45,000; *s* = $200; *α* = 0.05

n	**Confidence interval**	**Interval width**
100	*c.i.* = $45,000 ± 1.96(200/√99) = $45,000 ± $39.40	$78.80
500	*c.i.* = $45,000 ± 1.96(200/√499) = $45,000 ± $17.55	$35.10
1000	*c.i.* = $45,000 ± 1.96(200/√999) = $45,000 ± $12.40	$24.80
10000	*c.i.* = $45,000 ± 1.96(200/√9999) = $45,000 ± $3.92	$7.84

Summary: Confidence intervals

- Sample means, large samples (*n*>100), population standard deviation known

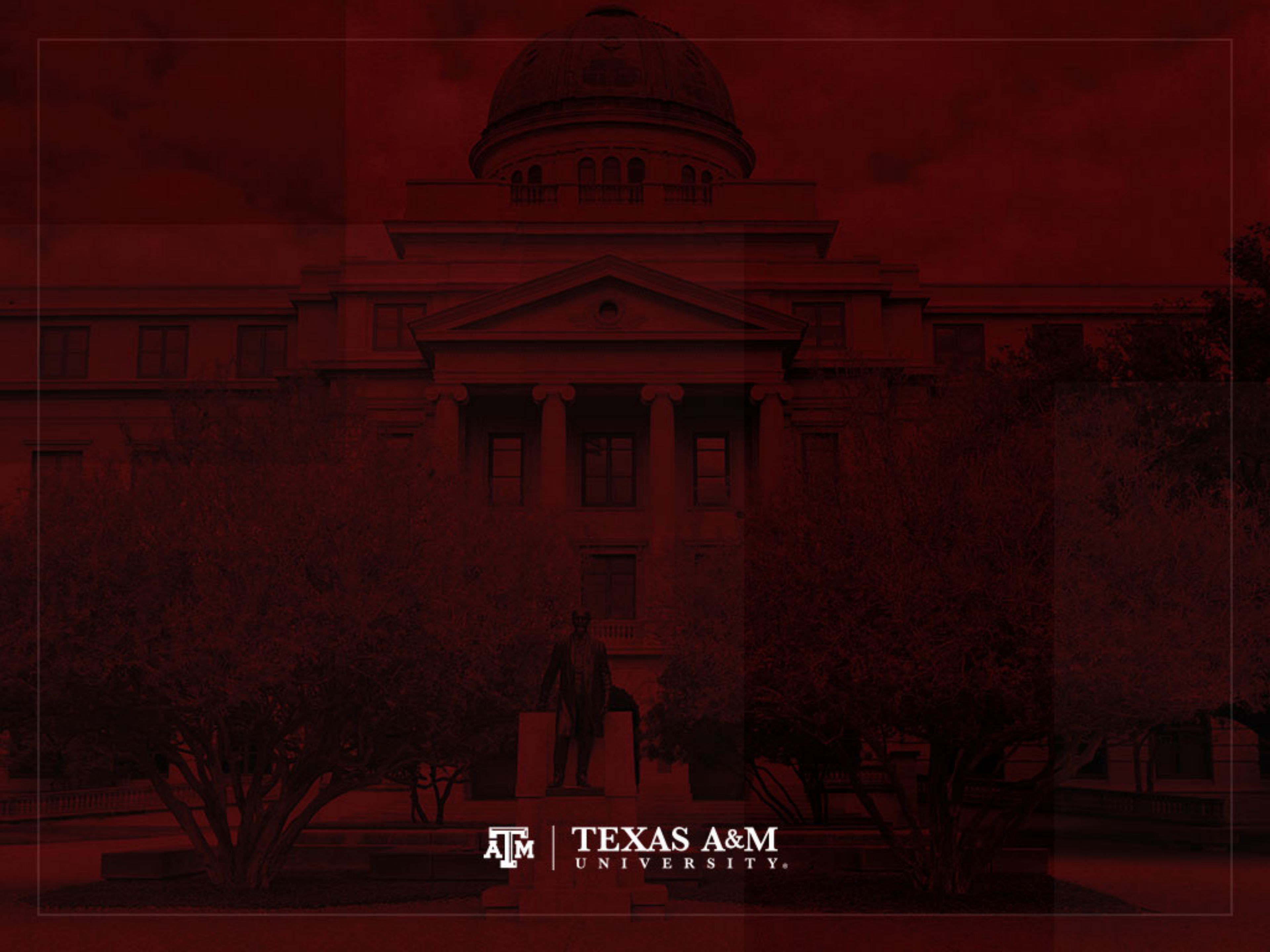
$$c.i. = \bar{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$$

- Sample means, large samples (*n*>100), population standard deviation unknown

$$c.i. = \bar{X} \pm Z\left(\frac{s}{\sqrt{n-1}}\right)$$

- Sample proportions, large samples (*n*>100)

$$c.i. = P_s \pm Z\sqrt{\frac{P_u(1-P_u)}{n}}$$


ATM
TEXAS A&M
UNIVERSITY