Summary of chapters 5–9: Inferential statistics and hypothesis testing

Ernesto F. L. Amaral

September 27, 2022
Advanced Methods of Social Research (SOCI 420)

Source: Healey, Joseph F. 2015. "Statistics: A Tool for Social Research." Stamford: Cengage Learning. 10th edition. Chapters 5 (pp. 122–142), 6 (pp. 144–159), 7 (pp. 160–184), 8 (pp. 185–215), 9 (pp. 216–246).

TEXAS A&M UNIVERSITY

Outline

- Sampling distribution
- Confidence interval and confidence level
- Hypothesis testing
- Two-sample test of means
- Two-sample test of proportions

Sampling distribution

- Sampling distribution is the probabilistic distribution of a statistic for all possible samples of a given size (*n*)
  - It is the distribution of a statistic (e.g., proportion, mean) for all possible outcomes of a certain size
- Central tendency and dispersion
  - Mean is the same as the population mean
  - Standard deviation is referred as standard error
    - It is the population standard deviation divided by the square root of *n*
    - We have to take into account the complex survey design to estimate the standard error (**`svyset`** command in Stata)

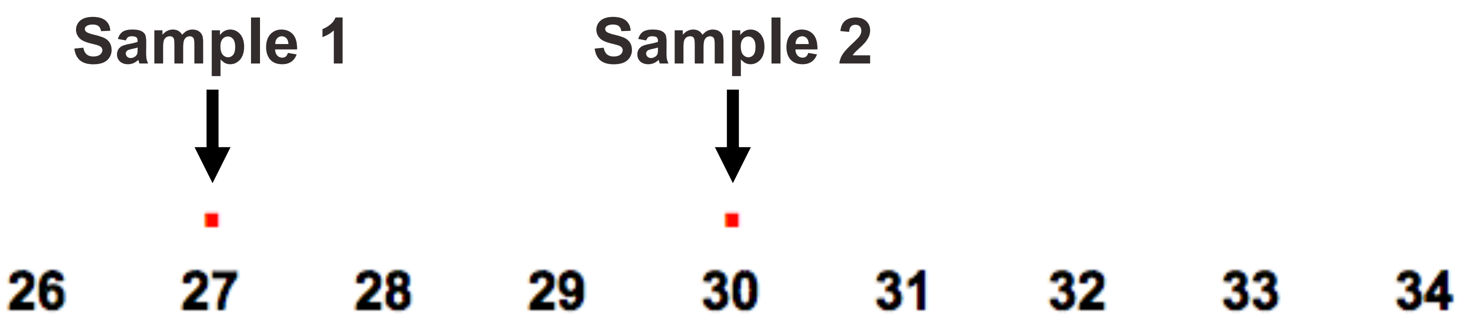
Linking sample and population

- Every application of inferential statistics involves three different distributions
  - Population: empirical; unknown
  - Sampling distribution: theoretical; known
  - Sample: empirical; known

- In inferential statistics, the sample distribution links the sample with the population

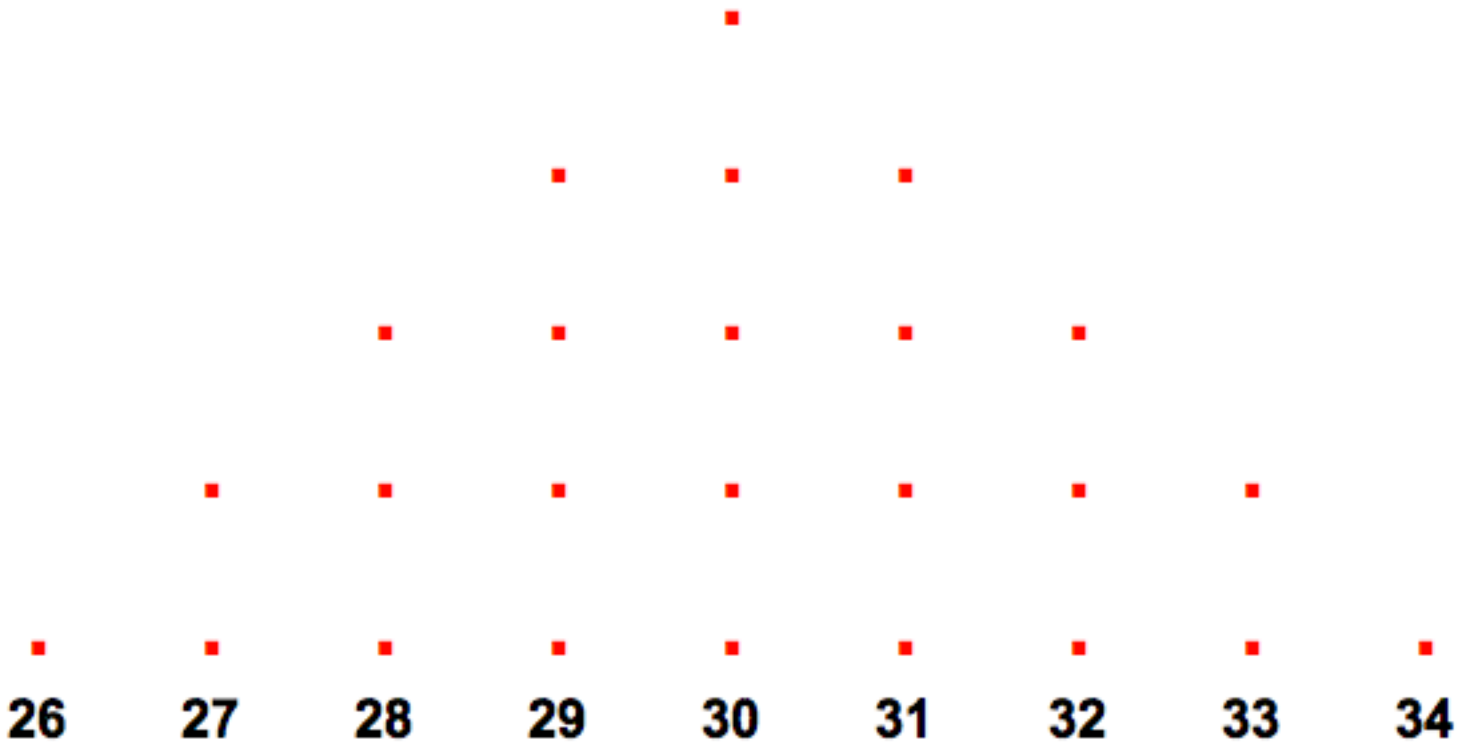
Example

- Suppose we want to gather information on the age of a community of 10,000 individuals
  - Sample 1: *n*=100 people, plot sample's mean of 27
  - Replace people in the sample back to the population
  - Sample 2: *n*=100 people, plot sample's mean of 30
  - Replace people in the sample back to the population



Example

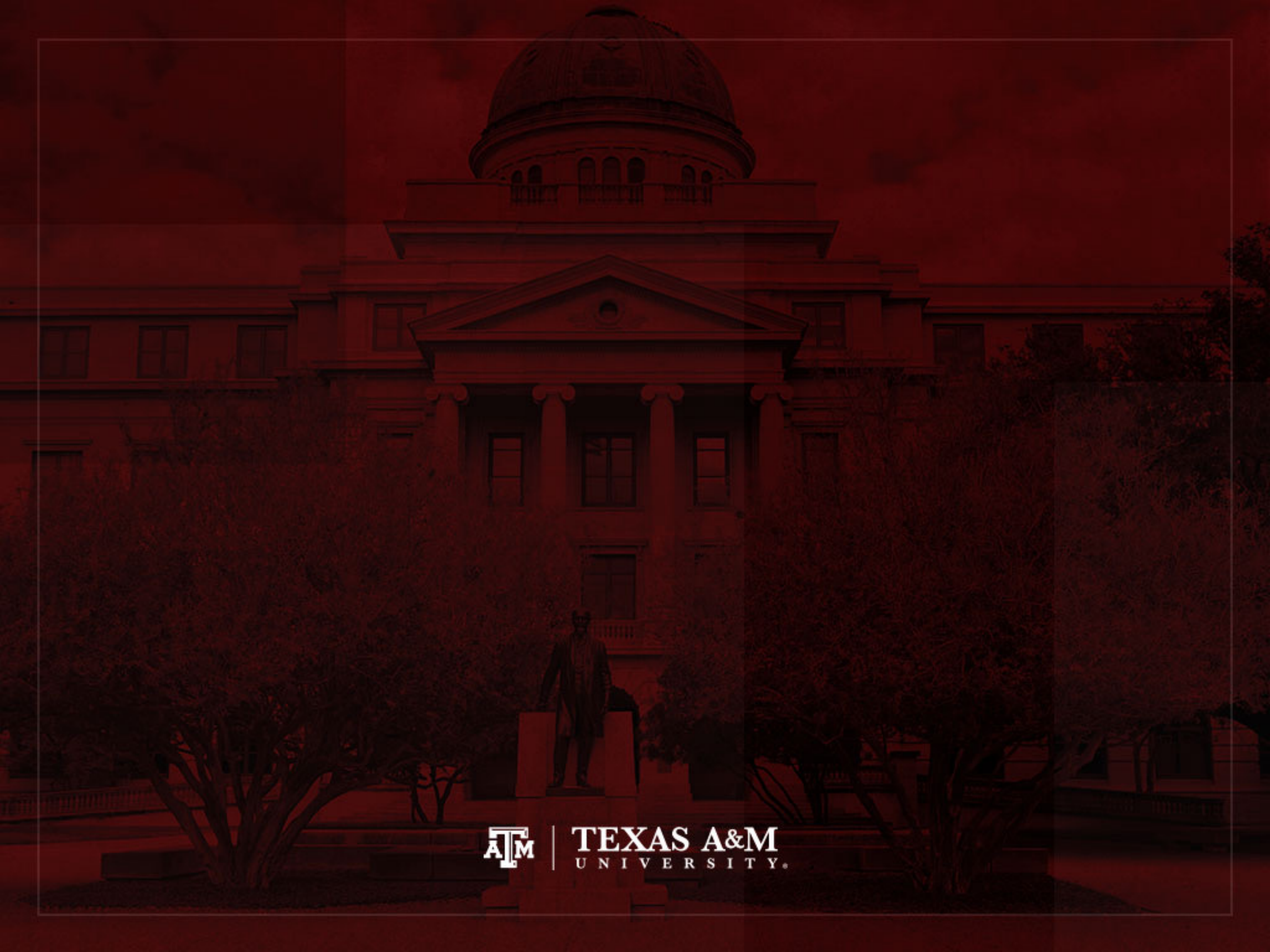
- We repeat this procedure: sampling, replacing
  - Until we have exhausted every possible combination of 100 people from the population of 10,000
  - Sampling distribution has a normal shape



Symbols

Distribution	Shape	Mean	Standard deviation	Proportion
Samples	Varies	$\bar{X}$	s	P_s
Populations	Varies	μ	σ	P_u
Sampling distributions	Normal	$\mu_{\bar{X}}$		
of means		$\mu_{\bar{X}}$	$\sigma_{\bar{X}} = \sigma/\sqrt{n}$	
of proportions		μ_p	σ_p	

Source: Healey 2015, p.157.


TEXAS A&M
UNIVERSITY

Confidence interval & level

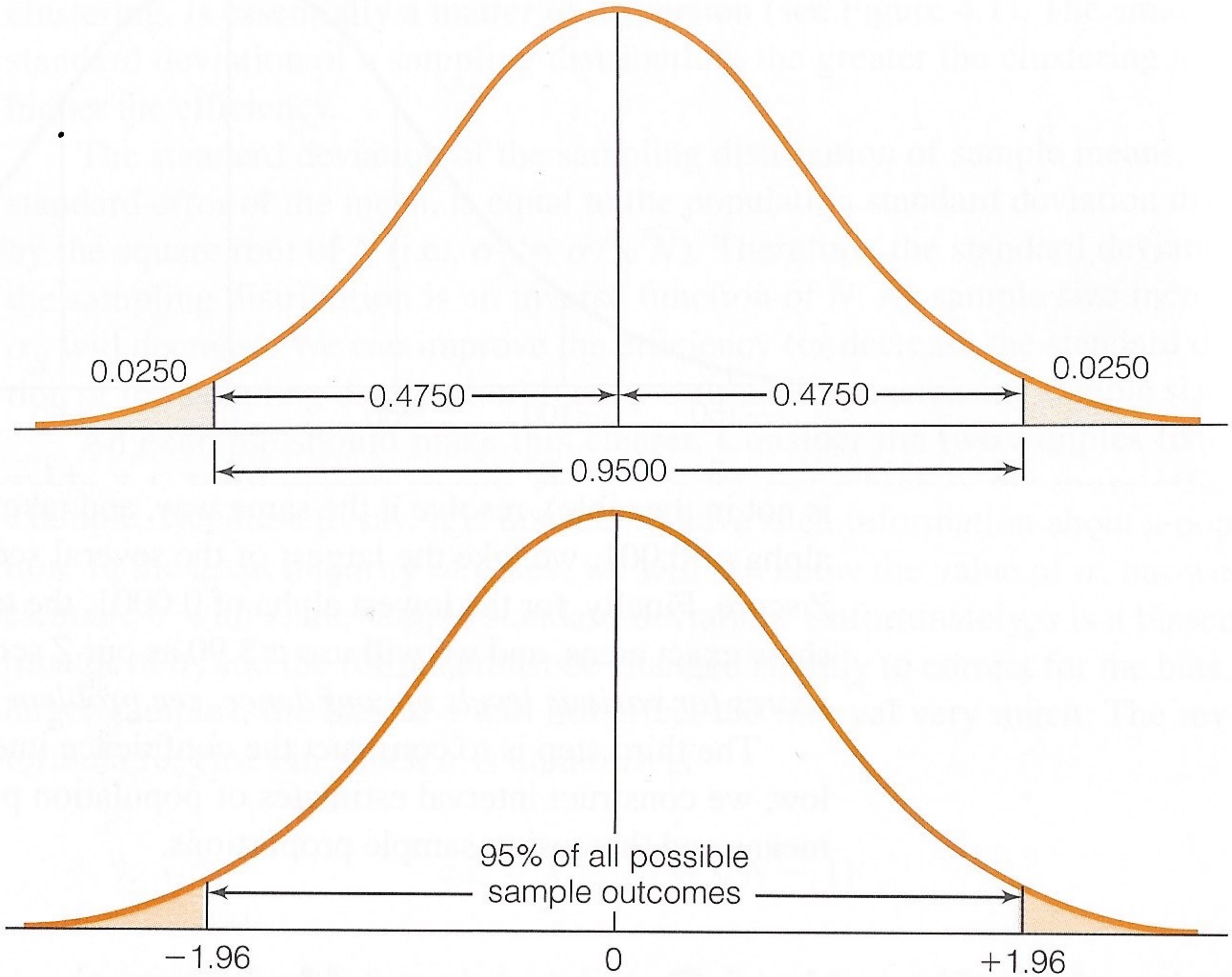
- **Confidence interval** is a range of values used to estimate the true population parameter
  - We associate a confidence level (e.g. 0.95 or 95%) to a confidence interval
- **Confidence level** is the success rate of the procedure to estimate the confidence interval
  - Expressed as probability $(1–\alpha)$ 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)

Confidence level, α, and Z

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

Source: Healey 2015, p.165.

Z score for significance level = α = 0.05



Source: Healey 2015, p.164–165.

Confidence intervals for sample means

- For large samples (*N*≥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

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

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

Example 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
```

	Mean	Linearized 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.

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

$\sqrt{P_u(1-P_u)/n}$ = 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 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
```

	Proportion	Linearized Std. Err.	Logit [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

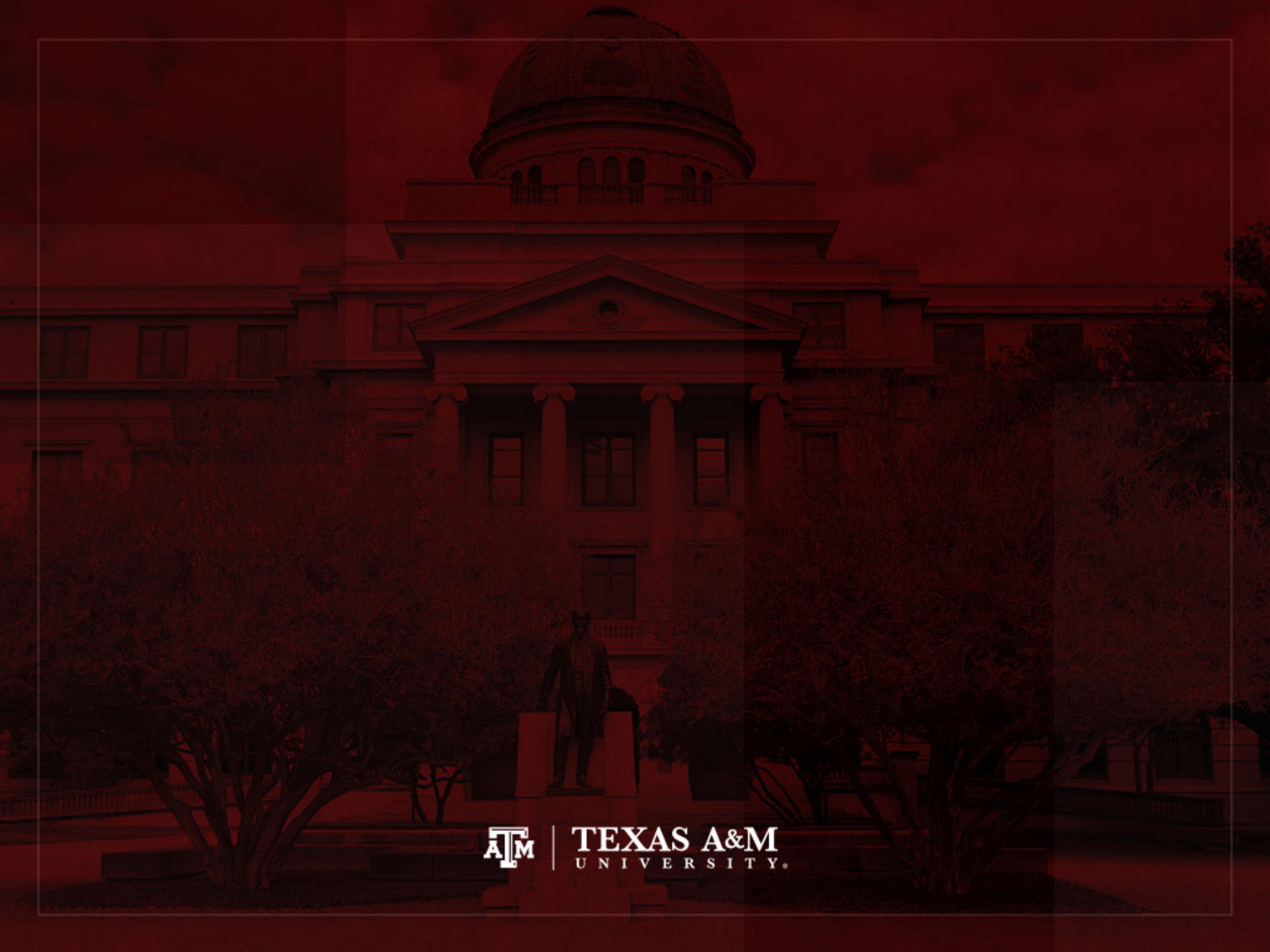
Obs.: Sample size of 3,184,099 individuals.
Source: 2018 American Community Survey.

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.


TEXAS A&M
UNIVERSITY

Hypothesis testing

- We analyze a difference between two sample statistics
  - We compare means or proportions of two samples from specific sub-groups of the population
- This is the question under consideration
  - "Is the difference between the samples large enough to allow us to conclude (with a known probability of error) that the populations represented by the samples are different?"

Source: Healey 2015, p.217.

Null hypothesis (H_0)

- Null hypothesis (H_0) indicates that populations are the same
  - Assuming that the H_0 is true, there is no difference between the parameters of the two populations
  - Equal sign (=) is used in the H_0
- We **reject** the H_0 and say there is a difference between the populations
  - If difference between sample statistics is significant
  - Or if the size of the estimated difference is unlikely

Alternative hypothesis (H_1)

- Alternative hypothesis or research hypothesis (H_1) indicates that populations are different
  - Different sign (≠), greater than sign (>), or less than sign (<) can be used in the H_1
  - Based on theory (previous studies), you should have a H_1 that states the direction of the difference (> or <)
  - If it is an exploratory study, H_1 will state that there is a difference (≠), but you don't know the direction
- We **accept** the H_1 and say there is a difference between the populations
  - If difference between sample statistics is significant

Decisions about hypotheses

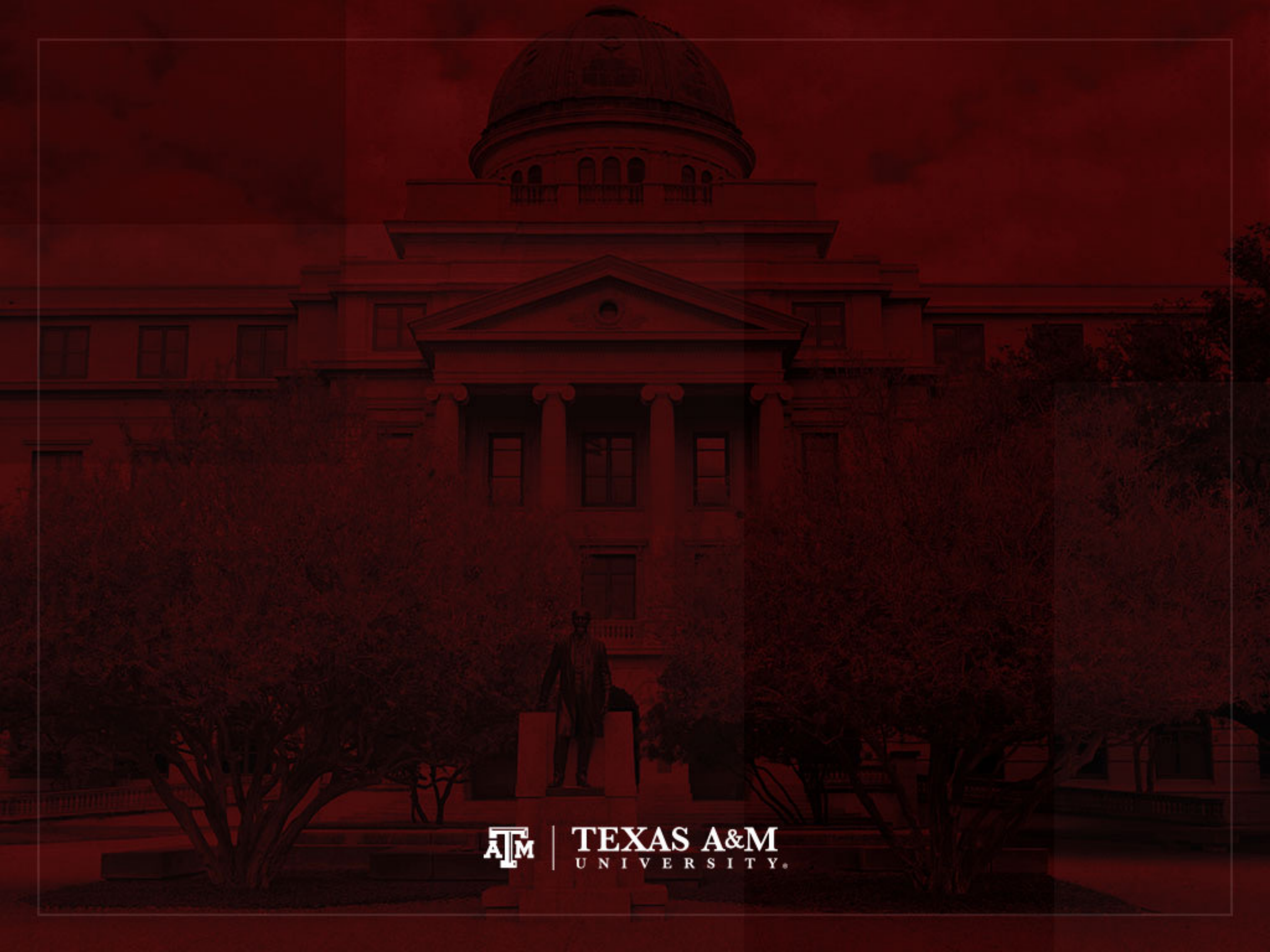
Hypotheses	***p* < *α***	***p* > *α***
Null hypothesis (H_0)	Reject	Do not reject
Alternative hypothesis (H_1)	Accept	Do not accept

- ***p*-value** is the probability of not rejecting the null hypothesis
- If a statistical software gives only the two-tailed *p*-value, divide it by 2 to obtain the one-tailed *p*-value

Significance level (*α*)	**Confidence level**
0.10 (10%)	90%
0.05 (5%)	95%
0.01 (1%)	99%
0.001 (0.1%)	99.9%

Outcomes of hypothesis testing

- Result of a specific analysis could be
  - Statistically significant and
    - Important (large magnitude)
  - Statistically significant, but
    - Unimportant (small magnitude)
  - Not statistically significant, but
    - Important (large magnitude)
  - Not statistically significant and
    - Unimportant (small magnitude)


TEXAS A&M
UNIVERSITY

Two-sample test of means

- Are means of two sub-groups for a specific variable different with statistical significance?

- Obtained *t*

$$t(obtained) = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}-\bar{X}}}$$

- Pooled estimate of the standard error

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

ACS: income by sex

- We know the average income by sex from the 2018 ACS

```
. table sex if income!=0, c(mean income)
```

Sex	mean(income)
Male	61704.38
Female	41238.01

- What causes the difference between male income of \$61,704.38 and female income of \$41,238.01?
- Real difference? Or difference due to random chance?

t-test for income by sex

- Men have an average income that is significantly higher than the female average income
  - The difference between male and female income was large and unlikely to have occurred by random chance (p<0.05) in 2018

```
. ttest income if income!=0, by(sex)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Male	812,666	61704.38	84.78448	76431.5	61538.2	61870.55
Female	761,647	41238.01	56.24931	49090.11	41127.76	41348.26
combined	1574313	51802.82	52.17731	65467.72	51700.56	51905.09
diff		20466.36	103.1275		20264.24	20668.49

```
    diff = mean(Male) - mean(Female)                     t = 198.4570
Ho: diff = 0                            degrees of freedom =  1.6e+06

    Ha: diff < 0              Ha: diff != 0              Ha: diff > 0
 Pr(T < t) = 1.0000      Pr(|T| > |t|) = 0.0000      Pr(T > t) = 0.0000
```

t-test = *t* =

diff. mean / std. error =

20,466.36 / 103.13 =

198.46

Source: 2018 American Community Survey.

Edited table

Table 1. Two-sample *t*-test of individual average wage and salary income for the U.S. population by sex, 2018

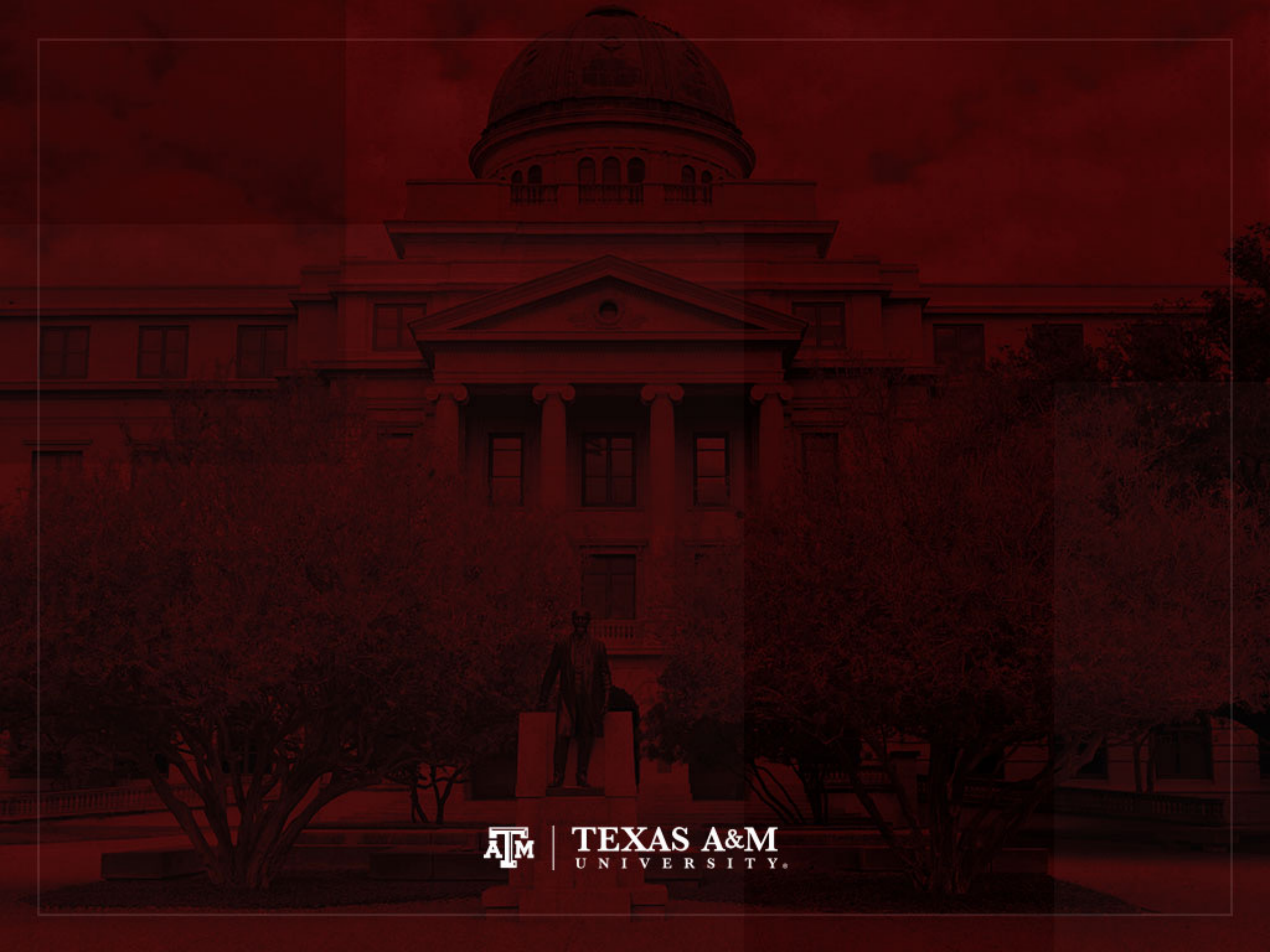
Sex	2018
Male	61,704.38
	(84.79)
Female	41,238.01
	(56.25)
Difference	20,466.36***
	(103.13)
Sample size	1,574,131

Note: Standard errors are reported in parentheses.

*Significant at p<0.10; **Significant at p<0.05; ***Significant at p<0.01.

No sample weight was utilized for this test.

Source: 2018 American Community Survey.


TEXAS A&M
UNIVERSITY

Two-sample test of proportions

- Are proportions of two sub-groups for a specific variable different with statistical significance?
- Obtained Z score

$$Z(obtained) = \frac{P_{s1} - P_{s2}}{\sigma_{p-p}}$$

- Pooled estimate of the standard error

$$\sigma_{p-p} = \sqrt{P_u(1 - P_u)} \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

- Population proportion

$$P_u = \frac{n_1 P_{s1} + n_2 P_{s2}}{n_1 + n_2}$$

ACS: internal migration by sex

- We know the proportion of internal migrants by sex based on the 2018 ACS

```
. tab dommig sex, col nofreq
```

dommig	Sex: Male	Sex: Female	Total
0	94.31	94.95	94.64
1	5.69	5.05	5.36
Total	100.00	100.00	100.00

```
. count if dommig!=. & sex!=.
  3,167,213
```

- What causes the difference between the percentage of men who are internal migrants (5.69%) and the percentage of women who are internal migrants (5.05%)?
  - Real difference? Or difference due to random chance?

Test of proportion for internal migration by sex

- Men are more likely to be internal migrants than women
  - The difference between the percentage of men who are internal migrants and the percentage of women who are internal migrants was large and unlikely to have occurred by random chance ($p<0.05$) in 2018

```
. prtest dommig, by(sex)

Two-sample test of proportions                Male: Number of obs =   1.6e+06
                                            Female: Number of obs =   1.6e+06
------------------------------------------------------------------------------
       Group |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Male |   .0568569    .000186                       .0564924    .0572214
      Female |   .0505412   .0001723                       .0502035    .0508788
-------------+----------------------------------------------------------------
        diff |   .0063157   .0002535                       .0058188    .0068126
             |  under Ho:   .0002532   24.94   0.000
------------------------------------------------------------------------------
        diff = prop(Male) - prop(Female)                          z =  24.9396
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 1.0000         Pr(|Z| > |z|) = 0.0000          Pr(Z > z) = 0.0000
```

prop. test = z =

diff. mean / std. error =

0.0063 / 0.0003 =

24.9396

Source: 2018 American Community Survey.

ACS: international migration by sex

- We know the proportion of international migrants by sex based on the 2018 ACS

```
. tab intmig sex, col nofreq
```

intmig	Sex: Male	Female	Total
0	99.43	99.45	99.44
1	0.57	0.55	0.56
Total	100.00	100.00	100.00

```
. count if intmig!=. & sex!=.
  3,014,232
```

- What causes the difference between the percentage of men who are international migrants (0.57%) and the percentage of women who are internal migrants (0.55%)?
  - Real difference? Or difference due to random chance?

Test of proportion for international migration by sex

- Men are more likely to be international migrants than women
  - The difference between the percentage of men who are international migrants and the percentage of women who are internal migrants was large and unlikely to have occurred by random chance ($p<0.05$) in 2018

```
. prtest intmig, by(sex)

Two-sample test of proportions                  Male: Number of obs =   1.5e+06
                                              Female: Number of obs =   1.5e+06

       Group |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+-------------------------------------------------------------------
        Male |   .0057487   .0000623                       .0056265    .0058709
      Female |   .0054624   .0000593                       .0053461    .0055787
-------------+-------------------------------------------------------------------
        diff |   .0002863   .0000861                       .0001176    .0004549
             |  under Ho:    .000086     3.33   0.001
-------------------------------------------------------------------------------
        diff = prop(Male) - prop(Female)                          z =   3.3286
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.9996         Pr(|Z| > |z|) = 0.0009          Pr(Z > z) = 0.0004
```

prop. test = *z* =

diff. mean / std. error =

0.0003 / 0.0001 =

3.3286

Source: 2018 American Community Survey.

Edited table

Table 2. Test of proportions for internal and international migration status for the U.S. population by sex, 2018

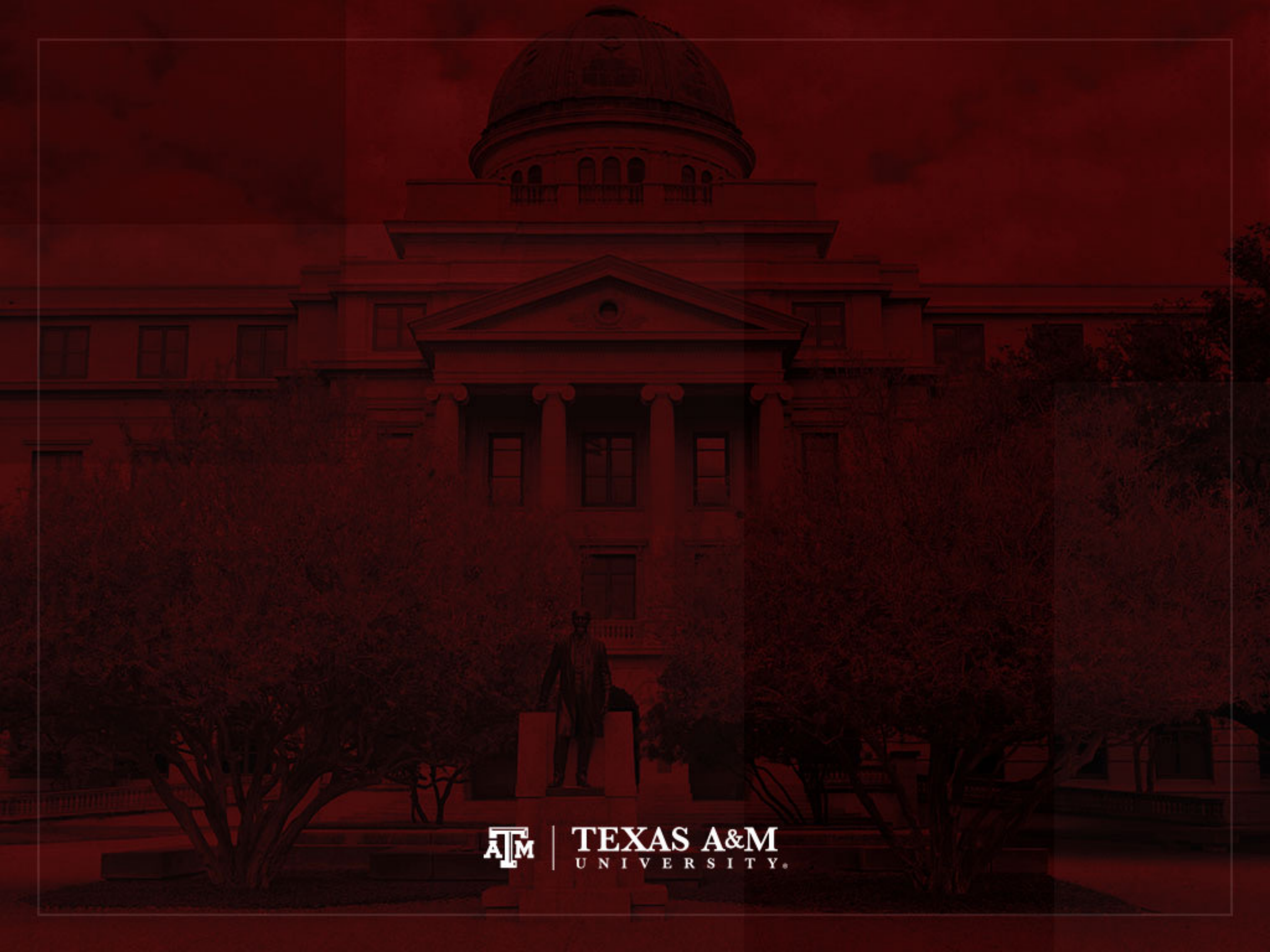
Sex	Internal migration	International migration
Male	0.0569	0.0058
	(0.0002)	(0.0001)
Female	0.0505	0.0055
	(0.0002)	(0.0001)
Difference	0.0063***	0.0003***
	(0.0003)	(0.0001)
Sample size	3,167,213	3,014,232

Note: Standard errors are reported in parentheses.

*Significant at p<0.10; **Significant at p<0.05; ***Significant at p<0.01.

No sample weight was utilized for this test.

Source: 2018 American Community Survey.


TEXAS A&M
UNIVERSITY