Lecture 2b: Survey weights

Ernesto F. L. Amaral

February 05–07, 2024
Advanced Methods of Social Research (SOCI 420)

www.ernestoamaral.com

Source: Treiman, Donald J. 2009. Quantitative Data Analysis: Doing Social Research to Test Ideas. San Francisco: Jossey-Bass. Chapter 9 (pp. 195–224).

TEXAS A&M UNIVERSITY

Outline

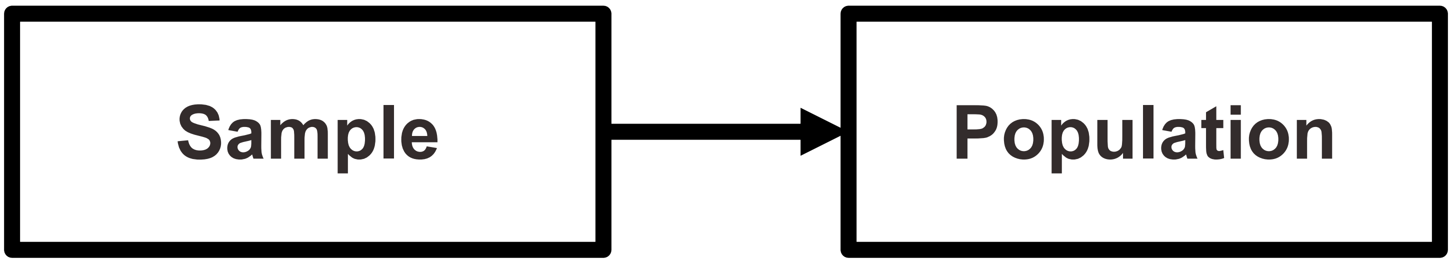
- Inferential statistics
- Survey weights
- Weight options in Stata
- Complex sample cluster design
- Weights in the General Social Survey (GSS)
- Examples of descriptive statistics

Inferential statistics

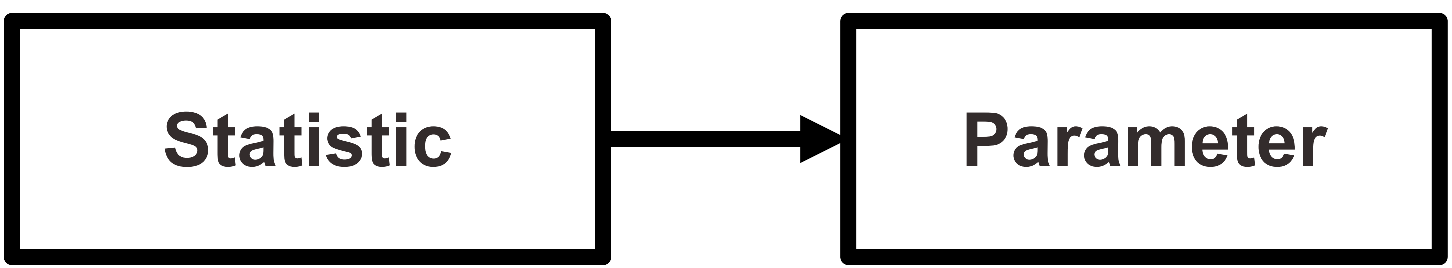
- Social scientists need inferential statistics
  - They almost never have the resources or time to collect data from every case in a population
- Inferential statistics uses data from samples to make generalizations about populations
  - **Population** is the total collection of all cases in which the researcher is interested
  - **Samples** are carefully chosen subsets of the population
- With proper techniques, generalizations based on samples can represent populations

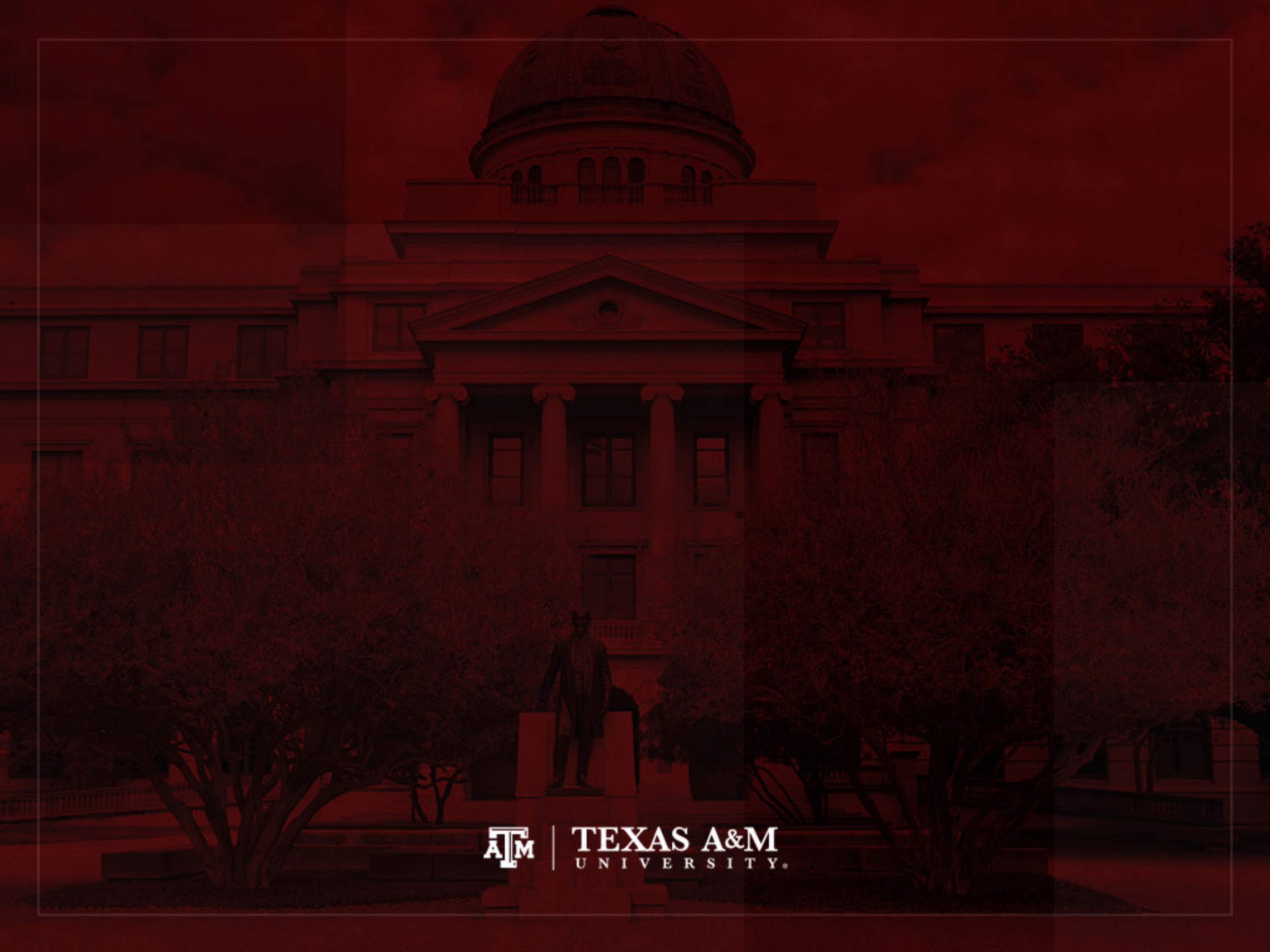
Basic logic and terminology

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



  - Statistics: characteristics of samples
  - Parameters: characteristics of populations
- Statistics are used to estimate parameters




TEXAS A&M
UNIVERSITY

Survey weights

Name	Number of observations collected in the survey	Weight to expand to population size	Weight to maintain sample size
José	1	4	0.8
Maria	1	6	1.2
Total	2	10	2

Survey weight =
Population weight * (Sum of survey weights / Sum of population weights)

Weights for tables

- When we use a sample to estimate the absolute number of people
  - For an area
  - For a specific sub-group
  - We use weights to expand to population size
- If we use a sample to estimate the proportion of people in a specific sub-group
  - And we are not concerned with the absolute value
  - We use weights to maintain the sample size (we focus on percentages)

Weights for regressions

- In a simple linear regression, the test of statistical significance for a β coefficient (t-test) is estimated as

$$t = \frac{\hat{\beta}}{SE_{\hat{\beta}}} = \frac{\hat{\beta}}{\sqrt{\frac{MSE}{S_{xx}}}} = \frac{\hat{\beta}}{\sqrt{\frac{RSS}{df * S_{xx}}}} = \frac{\hat{\beta}}{\sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{(n-2)\sum_i (x_i - \bar{x})^2}}}$$

  - SE_β: standard error of β
  - MSE: mean squared error = RSS / df
  - RSS: residual sum of squares = $\sum_i (y_i - \hat{y}_i)^2$ = $\sum_i \hat{e}_i{}^2$
  - df: degrees of freedom = n–2 for simple linear regression
    - 2 statistics (slope and intercept) are estimated to calculate sum of squares
  - S_{xx}: corrected sum of squares for x (total sum of squares)

Weights for regressions

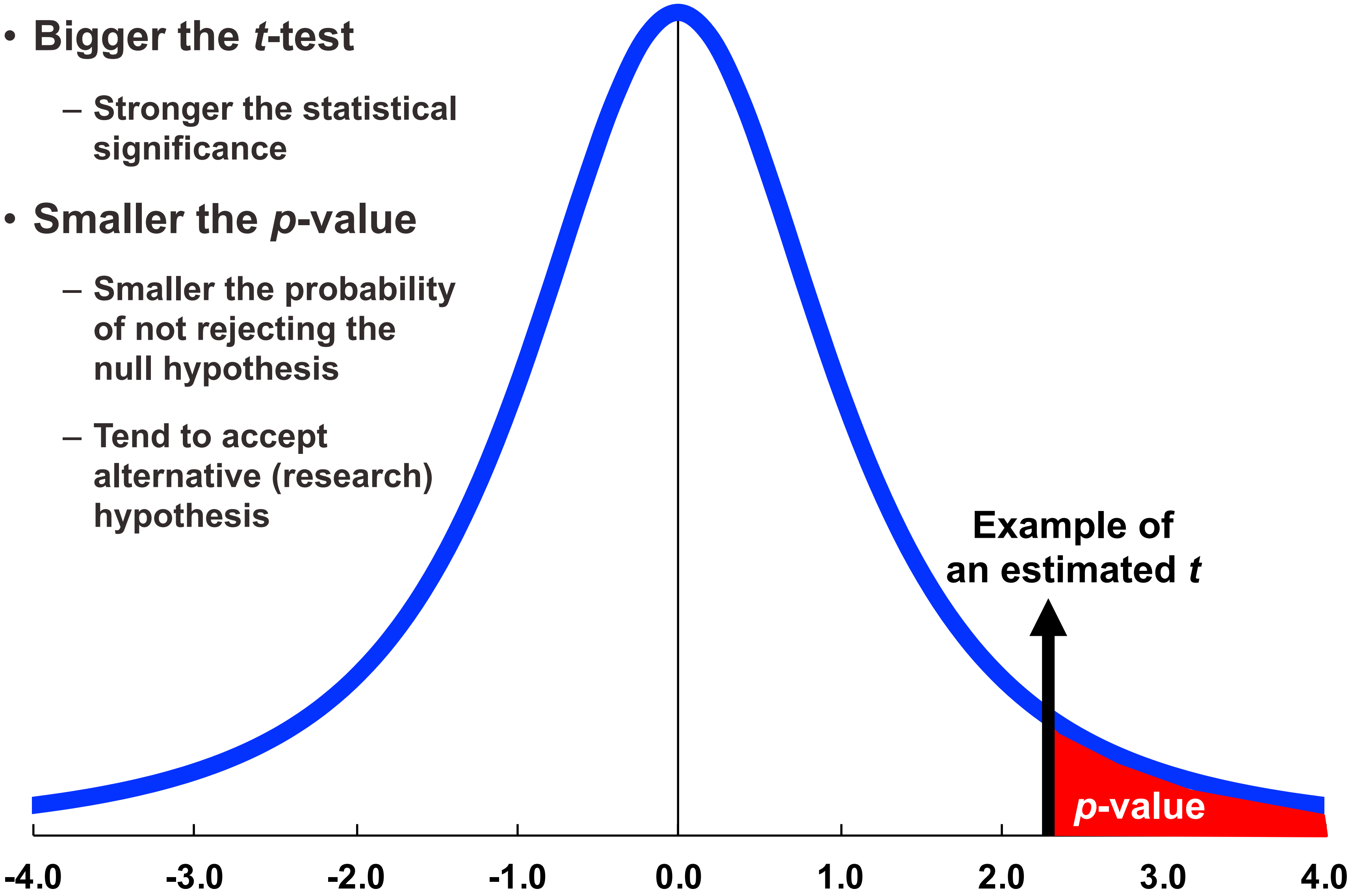
- If we use a weight that expands to the population size (*N*) on regressions
  - We would be incorrectly informing the statistical software that we have a sample with enormous size
  - This would artificially increase the test of statistical significance for the coefficient

$$\uparrow t = \frac{\hat{\beta}}{SE_{\hat{\beta}}} = \frac{\hat{\beta}}{\sqrt{\frac{MSE}{S_{xx}}}} = \frac{\hat{\beta}}{\sqrt{\frac{MSE}{S_{xx}}}} = \frac{\hat{\beta}}{\sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{\uparrow(n-2)\sum_i (x_i - \bar{x})^2}}} \downarrow$$

  - We have to inform the weight related to the sample design, but we should maintain the sample size (*n*)

t distribution (*df* = 2)

- **Bigger the *t*-test**
  - **Stronger the statistical significance**
- **Smaller the *p*-value**
  - **Smaller the probability of not rejecting the null hypothesis**
  - **Tend to accept alternative (research) hypothesis**

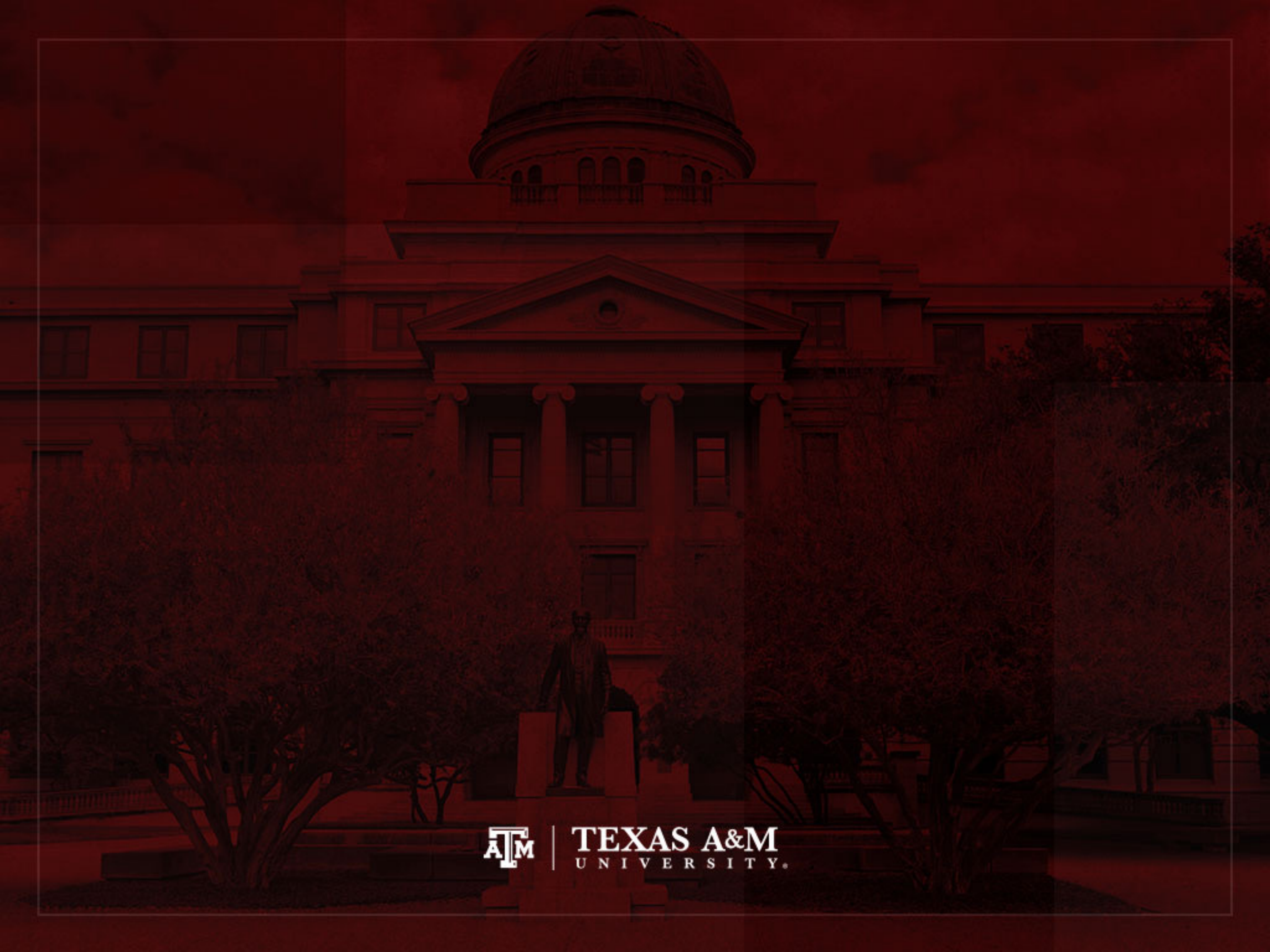


Decisions about hypotheses

Hypotheses	$p < \alpha$	$p > \alpha$
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 (success rate)
0.10 (10%)	90%
0.05 (5%)	95%
0.01 (1%)	99%
0.001 (0.1%)	99.9%


TEXAS A&M
UNIVERSITY

Weight options in Stata

- Frequency weight (fweight)
- "Importance" weight (iweight)
- Analytic weight (aweight)
- Sampling weight (pweight)

Extract of 2018 ACS microdata

	year	strata	cluster	perwt	hhwt	sex	age	income
1	2018	360248	2.018012e+12	56.00	56.00	Male	46	28000
2	2018	360248	2.018012e+12	51.00	51.00	Male	20	5000
3	2018	360248	2.018012e+12	76.00	76.00	Female	84	0
4	2018	360248	2.018012e+12	55.00	55.00	Female	18	1200
5	2018	360248	2.018012e+12	143.00	143.00	Female	56	1500
6	2018	360248	2.018012e+12	198.00	198.00	Male	31	10000
7	2018	360248	2.018012e+12	48.00	48.00	Female	19	2000
8	2018	360248	2.018012e+12	48.00	48.00	Male	25	7000
9	2018	360248	2.018012e+12	65.00	65.00	Female	18	0
10	2018	360248	2.018012e+12	53.00	53.00	Female	18	15000
11	2018	360248	2.018012e+12	17.00	17.00	Male	63	0
12	2018	360248	2.018012e+12	39.00	39.00	Female	18	4000
13	2018	360248	2.018012e+12	104.00	104.00	Male	21	1000
14	2018	360248	2.018012e+12	200.00	200.00	Male	40	80000
15	2018	360248	2.018012e+12	20.00	20.00	Male	33	0
16	2018	360248	2.018012e+12	59.00	59.00	Male	19	2900
17	2018	360248	2.018012e+12	56.00	56.00	Male	55	0
18	2018	360248	2.018012e+12	77.00	77.00	Male	18	9000
19	2018	360248	2.018012e+12	16.00	16.00	Female	41	1100
20	2018	360248	2.018012e+12	46.00	46.00	Male	33	0

Frequency weight

- **FWEIGHT**
  - Expands survey size to the population size
  - Indicates the number of duplicated observations
  - Used on tables to generate frequencies
  - Can be used in frequency distributions only when weight variable is discrete (no fractional numbers)

```
tab x [fweight = weight]
```

"Importance" weight

- **IWEIGHT**
  - Indicates the "importance" of the observation in some vague sense
  - Has no formal statistical definition
  - Any command that supports iweights will define exactly how they are treated
  - Intended for use by programmers who want to produce a certain computation
  - Can be used in frequency distributions even when weight variable is continuous (fractional numbers)

```
tab x [iweight = weight]
```

Analytic weight

- **AWEIGHT**
  - Inversely proportional to the variance of an observation
  - Variance of the *j*th observation is assumed to be σ^2/w_j, where w_j are the weights
  - For most Stata commands, the recorded scale of aweights is irrelevant
  - Stata internally rescales frequencies, so sum of weights equals sample size

```
tab x [aweight = weight]
regress y x1 x2 [aweight = weight]
```

More about analytic weight

- Observations represent averages and weights are the number of elements that gave rise to the average

group	x	y	n
1	3.5	26.0	2
2	5.0	20.0	3

- Instead of

group	x	y
1	3	22
1	4	30
2	8	25
2	2	19
2	5	16

- Usually, survey data is collected from individuals and households (not as averages)
  – Thus, aweights are not appropriate for most cases

Sampling weight

- **PWEIGHT**
  - Denote the inverse of the probability that the observation is included due to the sampling design
  - Variances, standard errors, and confidence intervals are estimated with a more precise procedure
  - Indicated for statistical regressions to estimate robust standard errors
    - Obtain unbiased standard errors of OLS coefficients under heteroscedasticity (i.e., residuals not randomly distributed)
    - Robust standard errors are usually larger than conventional ones

```
regress y x1 x2 [pweight = weight]
```

Summary of Stata weights

WEIGHTS IN FREQUENCY DISTRIBUTIONS		
Weight unit of measurement	**Expand to population size**	**Maintain sample size**
Discrete	**fweight**	**aweight**
Continuous	**iweight**	

WEIGHTS IN STATISTICAL REGRESSIONS should maintain sample size	
Robust standard error	**Adjusted R^2, TSS, ESS, RSS**
pweight	**aweight**
reg *y x*, vce(robust) **reg *y x*, vce(cluster *area*)**	**outreg2**

Example of 2018 ACS weight

```
. sum perwt, d

                        Person weight
-------------------------------------------------------------
      Percentiles      Smallest
 1%           10              1
 5%           19              1
10%           29              1       Obs           3,214,539
25%           52              1       Sum of wgt.   3,214,539

50%           80                      Mean           101.7774
                        Largest       Std. dev.      83.93534
75%          124           1916
90%          195           1990       Variance        7045.14
95%          263           2097       Skewness       2.845116
99%          427           2313       Kurtosis       17.99265
```

Example of 2018 ACS weight

```
. tab sex
```

Sex	Freq.	Percent	Cum.
Male	1,574,618	48.98	48.98
Female	1,639,921	51.02	100.00
Total	3,214,539	100.00	

```
. tab sex [fweight=perwt]
```

Sex	Freq.	Percent	Cum.
Male	161,072,404	49.23	49.23
Female	166,095,035	50.77	100.00
Total	327,167,439	100.00	

```
. tab sex [iweight=perwt]
```

Sex	Freq.	Percent	Cum.
Male	161,072,404	49.23	49.23
Female	166,095,035	50.77	100.00
Total	327,167,439	100.00	

```
. tab sex [aweight=perwt]
```

Sex	Freq.	Percent	Cum.
Male	1,582,595	49.23	49.23
Female	1,631,944	50.77	100.00
Total	3,214,539	100.00	

Example of 2021 GSS weight

```
. sum wtssnrps, d

      person post-stratification weight, nonrespondents
                           adjusted
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .243687        .1723802
 5%      .30024        .1738938
10%    .4057674        .1926333       Obs               4,032
25%    .5423563        .2104285       Sum of wgt.       4,032

50%    .8183308                       Mean                  1
                        Largest       Std. dev.      .7260472
75%    1.212269         6.51434
90%    1.798724        6.903664       Variance       .5271445
95%     2.27083        7.218392       Skewness       2.825826
99%    3.986099        7.557038       Kurtosis       15.89999
```

Example of 2021 GSS weight

```
. tab sex, m
```

respondents sex	Freq.	Percent	Cum.
male	1,736	43.06	43.06
female	2,204	54.66	97.72
.i	19	0.47	98.19
.n	71	1.76	99.95
.s	2	0.05	100.00
Total	4,032	100.00	

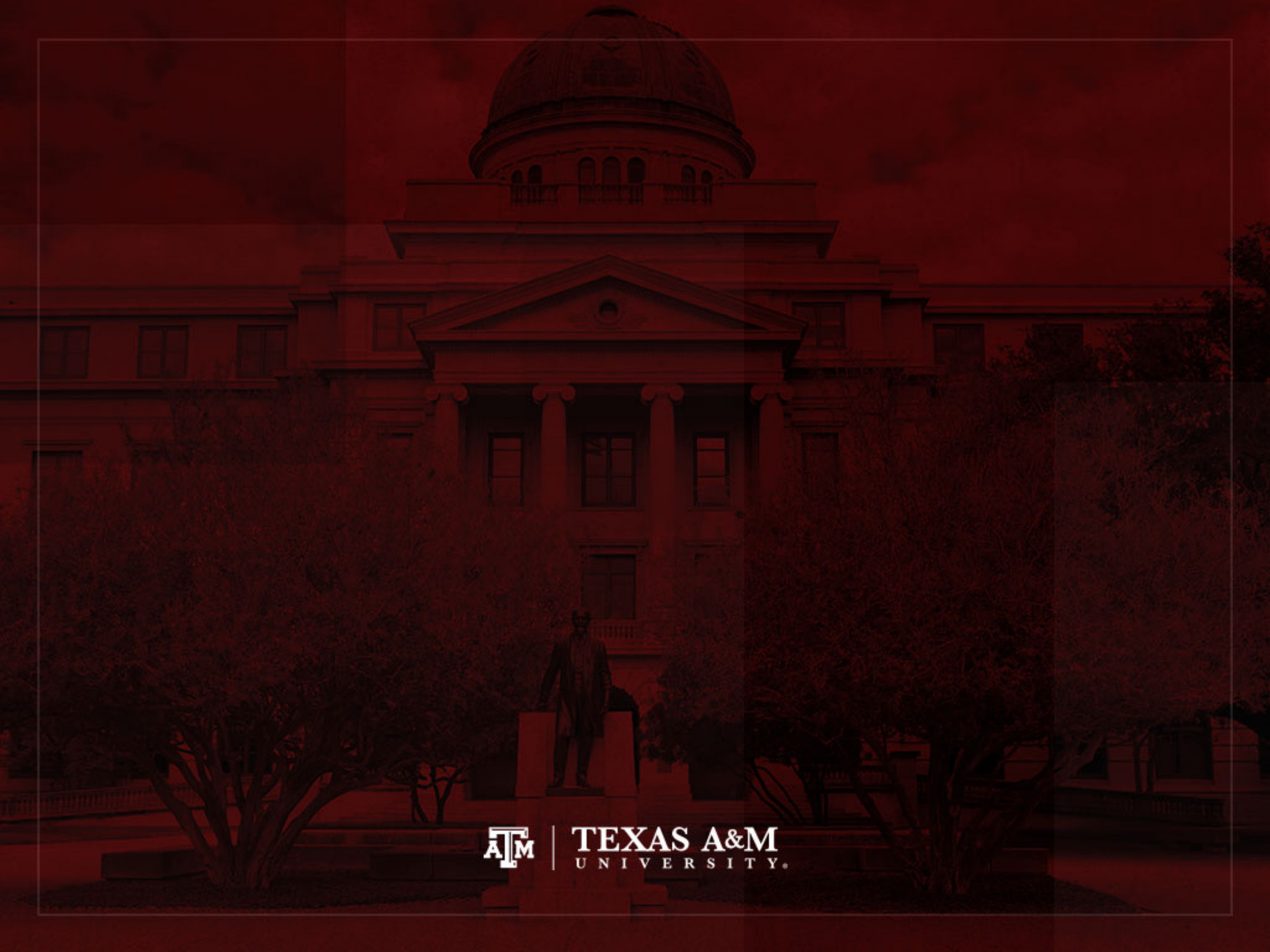
```
. tab sex [fweight=wtssnrps], m
may not use noninteger frequency weights
r(401);
```

```
. tab sex [iweight=wtssnrps], m
```

respondents sex	Freq.	Percent	Cum.
male	1,904.2566	47.23	47.23
female	1,993.21543	49.43	96.66
.i	18.1122752	0.45	97.11
.n	113.299832	2.81	99.92
.s	3.11586052	0.08	100.00
Total	4,032	100.00	

```
. tab sex [aweight=wtssnrps], m
```

respondents sex	Freq.	Percent	Cum.
male	1,904.2566	47.23	47.23
female	1,993.21543	49.43	96.66
.i	18.1122752	0.45	97.11
.n	113.299832	2.81	99.92
.s	3.11586052	0.08	100.00
Total	4,032	100.00	


TEXAS A&M
UNIVERSITY

Complex sample cluster design

- To calculate standard errors correctly, variables for sample cluster design must be used
  - Without design variables, Stata will assume a simple random sample and underestimate standard errors

- Strata are created based on the lowest level of geography available in each sample
  - We use additional statistical techniques that account for the complex sample design to produce correct standard errors and statistical tests

Cluster design for tables

- If we want to estimate a confidence interval for a sample statistic (mean or proportion), we need to inform the complex survey design
- **Confidence interval** is a range of values used to estimate the true population parameter
- **Confidence level** is the success rate of the procedure to estimate the confidence interval
- Larger confidence levels generate larger confidence intervals

Confidence level, *α*, and *Z*

Confidence level (1 – *α*) * 100	**Significance level alpha (*α*)**	***α* / 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.

Confidence intervals from samples

$$c.i. = sample\ estimate \pm margin\ of\ error$$
$$c.i. = sample\ estimate \pm score\ of\ confidence\ level * standard\ \ error$$

- Sample mean ($\bar{x}$), standard deviation (s), *n*<30

$$c.i. = \bar{x} \pm t\left(\frac{s}{\sqrt{n}}\right) \qquad df = n - 1$$

- Sample mean ($\bar{x}$), standard deviation (s), *n*≥30

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

- Sam. proportion (P_s), pop. proportion (P_u), *n*≥30

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

Cluster design for regressions

- We also need to inform cluster design for regressions, because the *t*-test utilizes standard errors

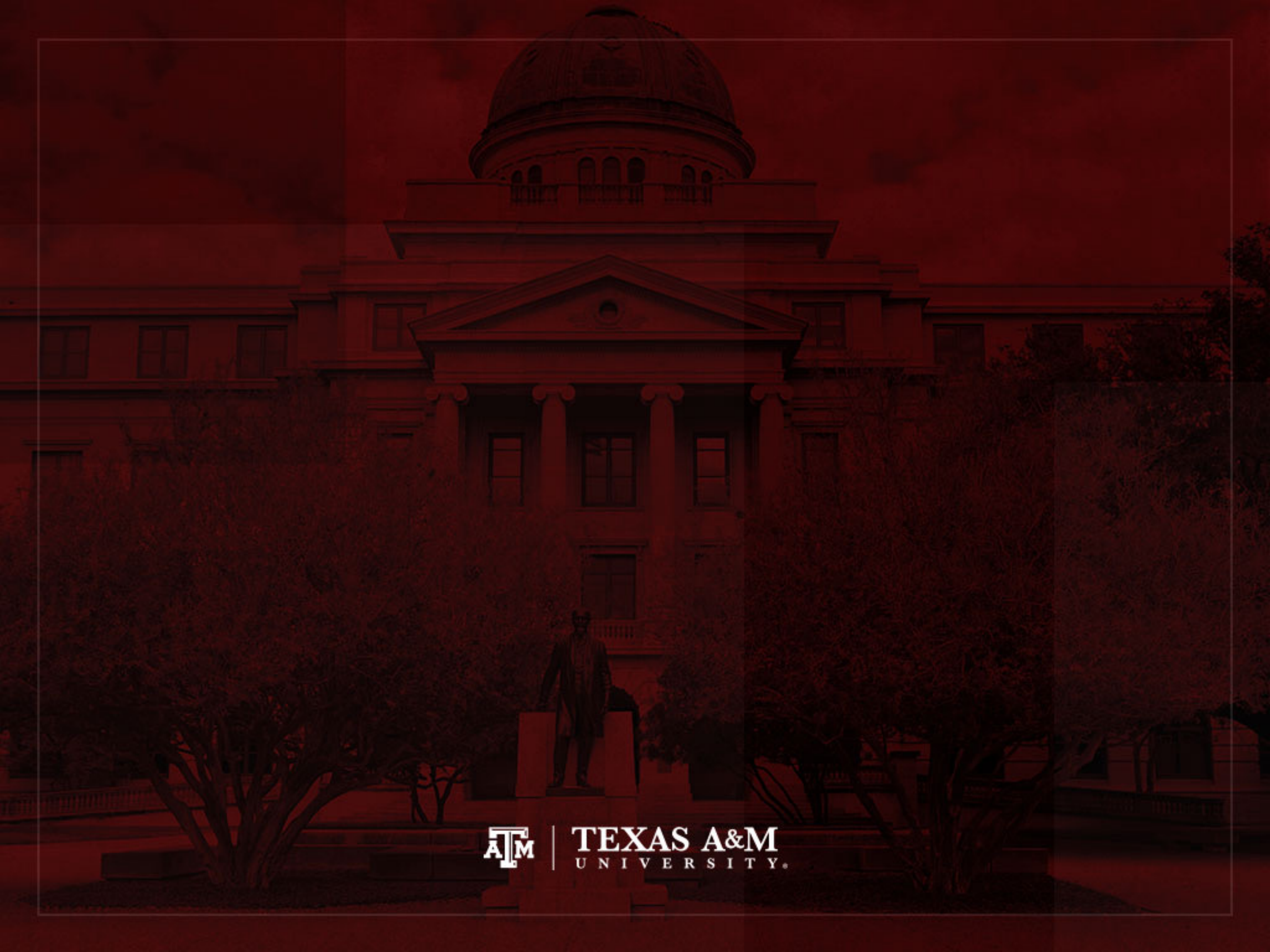
$$t = \frac{\hat{\beta}}{SE_{\hat{\beta}}} = \frac{\hat{\beta}}{\sqrt{\frac{MSE}{S_{xx}}}} = \frac{\hat{\beta}}{\sqrt{\frac{RSS}{df * S_{xx}}}} = \frac{\hat{\beta}}{\sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{(n-2)\sum_i (x_i - \bar{x})^2}}}$$

  - SE_β: standard error of β
  - *MSE*: mean squared error = *RSS* / *df*
  - *RSS*: residual sum of squares = $\sum_i (y_i - \hat{y}_i)^2$ = $\sum_i \hat{e}_i{}^2$
  - *df*: degrees of freedom = *n*–2 for simple linear regression
  - S_{xx}: corrected sum of squares for x (total sum of squares)

Cluster design & standard error

- Sample cluster designs underestimate standard errors, because they tend to select individuals with more similar characteristics from the same clusters
  - Simple random samples would provide more variation (higher standard errors), because they give the same chance of selection for all individuals in the population
- When we inform the cluster design, the standard error tends to increase and statistical significance decreases

$$\downarrow t = \frac{\hat{\beta}}{\uparrow SE_{\hat{\beta}}} = \frac{\hat{\beta}}{\sqrt{\frac{MSE}{S_{xx}}}} = \frac{\hat{\beta}}{\sqrt{\frac{RSS}{df * S_{xx}}}} = \frac{\hat{\beta}}{\sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{(n-2)\sum_i (x_i - \bar{x})^2}}}$$


TEXAS A&M
UNIVERSITY

Weights in GSS

- The General Social Survey (GSS) targets the adult population (18+) living in U.S. households
- Due to the adoption of the sub-sampling design of non-respondents, a weight must be employed when using the GSS 2004 and after
- There are three continuous weight variables
  - WTSS
  - WTSSNR
  - WTSSALL
- They all maintain the original sample size, even in frequency distributions with "iweight"

WTSS

- WTSS variable takes into consideration
  - Sub-sampling of non-respondents
  - Number of adults in the household

- In years prior to 2004, a value of one is assigned to all cases, so they are effectively unweighted
  - Number of adults can be utilized to make this adjustment for years prior to 2004

WTSSNR

- WTSSNR variable takes into consideration
  - Sub-sampling of non-respondents
  - Number of adults in the household
  - Differential non-response across areas
- In years prior to 2004, a value of one is assigned to all cases, so they are effectively unweighted
  - Number of adults can be utilized to make this adjustment for years prior to 2004
  - Area non-response adjustment is not possible

WTSSALL

- WTSSALL takes WTSS and applies an adult weight to years before 2004
- The weight value of WTSSALL is the same as WTSS for 2004 and after
- Researchers who use the GSS data before or after 2004 may consider using the WTSSALL weight variable

```
tab x [aweight = wtssall]
sum x [aweight = wtssall]
```

GSS has a cluster sample

(https://gssdataexplorer.norc.org/gss_stdError)

- First- and second-stage units are selected with probabilities proportional to size
  - Size is defined by number of housing units
- Third-stage units (housing units) are selected to be an equal-probability sample
  - This results in roughly the same number of housing units selected per second-stage sampling unit

GSS variables for cluster design

(https://gssdataexplorer.norc.org/gss_stdError)

- There are two design variables
  - VSTRAT
  - VPSU

- First-stage unit
  - VSTRAT: Variance Stratum
  - National Frame Areas (NFAs): one or more counties

- Second-stage unit
  - VPSU: Variance Primary Sampling Unit
  - Segments: block, group of blocks, or census tract

GSS complex sample design

(https://gssdataexplorer.norc.org/gss_stdError)

- Account for GSS sample design in Stata

```
svyset [weight=wtssall], strata(vstrat) psu(vpsu) singleunit(scaled)
```

- After ”svyset,” you should indicate survey design with the option ”svy” for commands that estimate standard errors

```
svy: mean y
svy: reg y x1 x2
```

Strata with single sampling unit

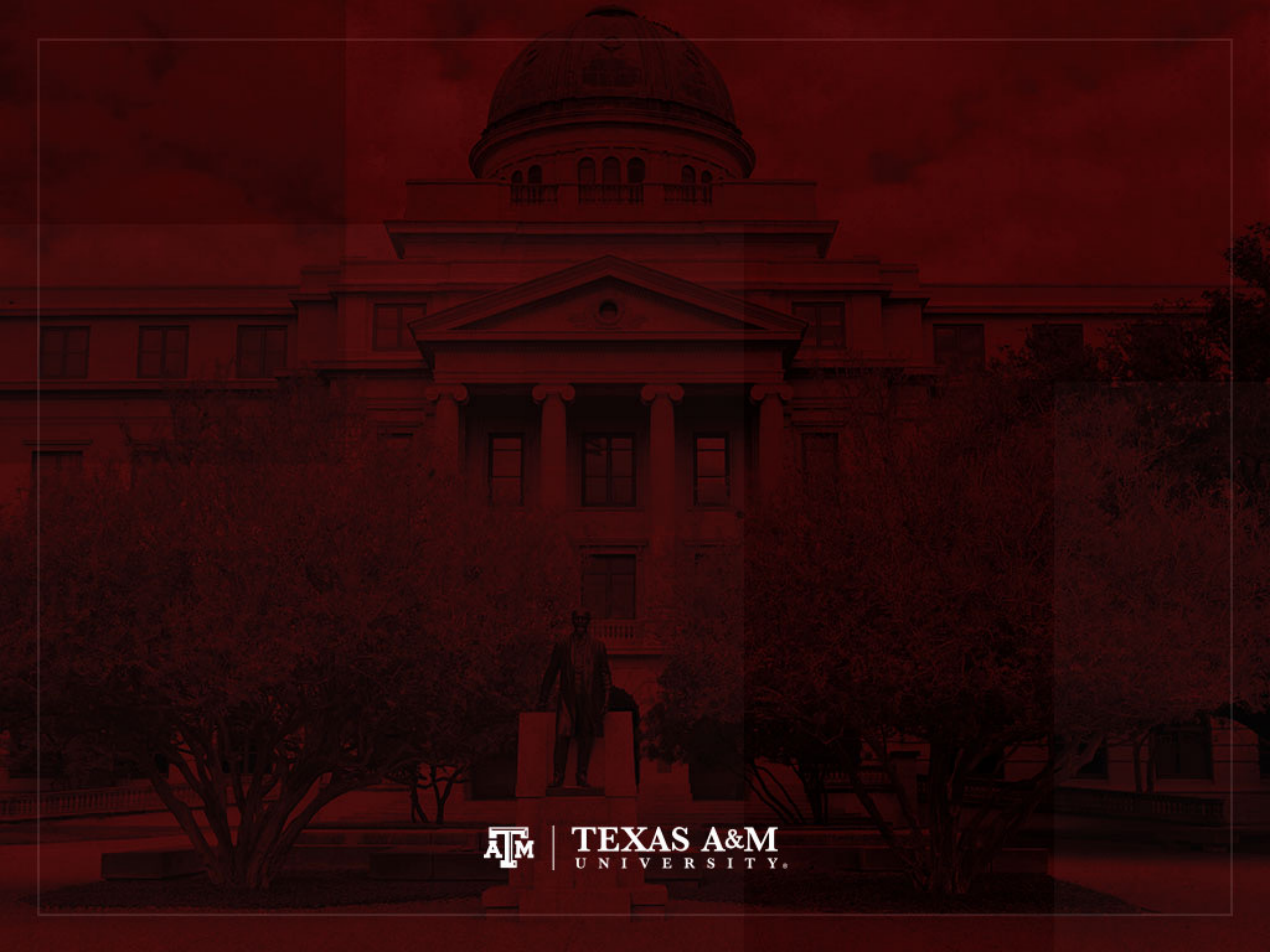
(https://gssdataexplorer.norc.org/gss_stdError)

- VSTRAT and VPSU were created with a minimum of three respondents within a cell
  - If all cases are missing on a variable, you get an error message in Stata
  - ”Missing standard error because of stratum with single sampling unit”
- It is recommended to utilize the ”subpop” option for any subdomain analyses (e.g., for males)

```
svy, subpop(if sex==1): tab x
```

- You can also specify that strata with one sampling unit are ”centered” at grand mean instead of stratum mean

```
svyset [weight=wtssall], strata(vstrat) psu(vpsu) singleunit(centered)
```


TEXAS A&M
UNIVERSITY

Example: 2021 GSS in Stata (nominal-level variable)

```
. tab sex
```

respondents sex	Freq.	Percent	Cum.
male	1,736	44.06	44.06
female	2,204	55.94	100.00
Total	3,940	100.00	

```
. tab sex [iweight=wtssnrps]
```

respondents sex	Freq.	Percent	Cum.
male	1,904.2566	48.86	48.86
female	1,993.21543	51.14	100.00
Total	3,897.472	100.00	

```
. svyset [weight=wtssnrps], strata(vstrat) psu(vpsu) singleunit(scaled)
(sampling weights assumed)

. svy: tab sex
(running tabulate on estimation sample)

Number of strata =      9
Number of PSUs   = 3,492
Number of obs    =     3,940
Population size  = 3,897.472
Design df        =     3,483
```

respondents sex	proportion
male	.4886
female	.5114
Total	1

Key: proportion = **Cell proportion**

Example: 2021 GSS in Stata (ordinal-level variable)

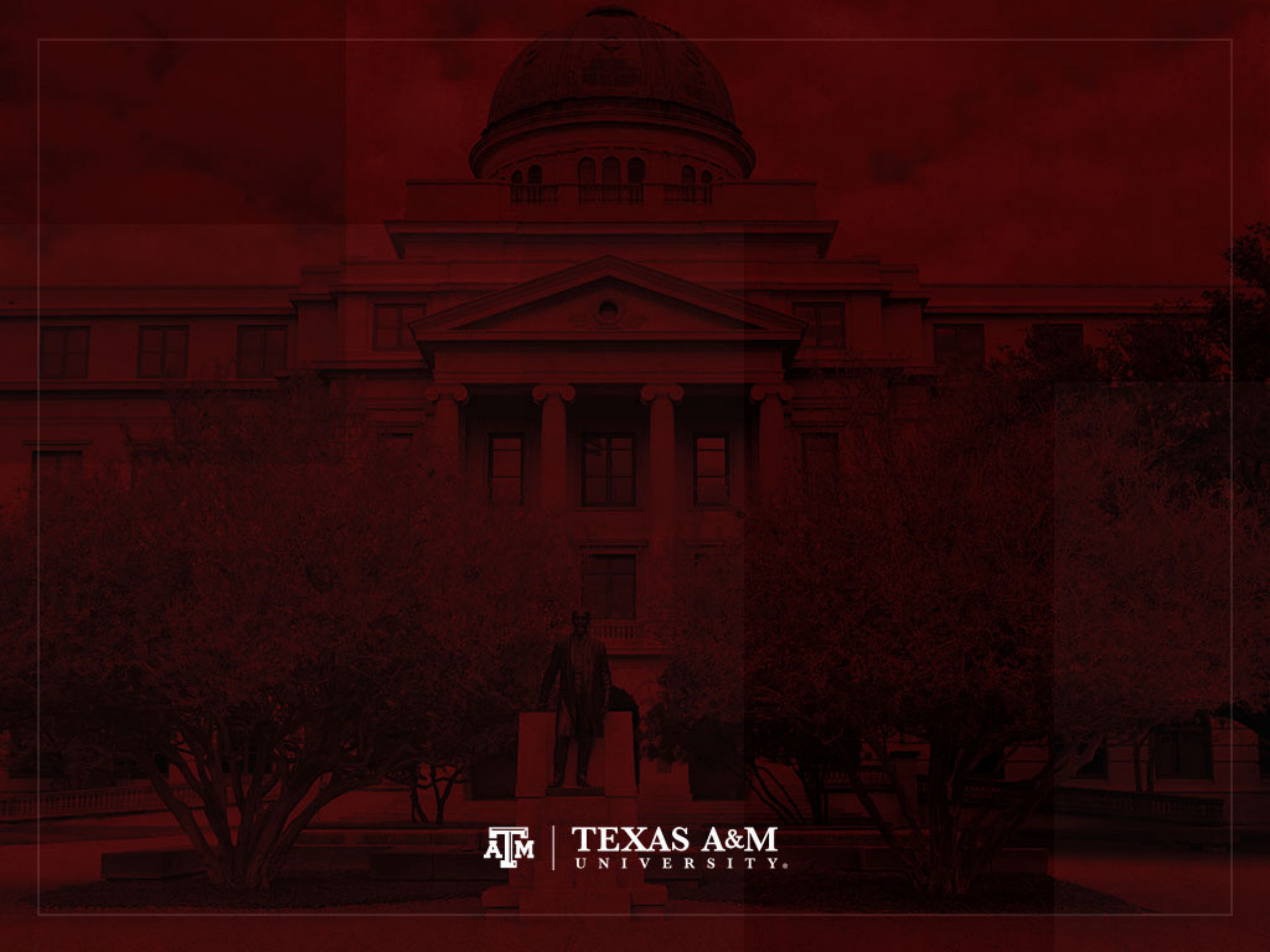
```
. tab degree

       r's highest degree |      Freq.     Percent        Cum.
--------------------------+-----------------------------------
   less than high school |        246        6.14        6.14
             high school |      1,597       39.84       45.97
associate/junior college |        370        9.23       55.20
              bachelor's |      1,036       25.84       81.04
                graduate |        760       18.96      100.00
--------------------------+-----------------------------------
                   Total |      4,009      100.00
```

```
. tab degree [iweight=wtssnrps]

       r's highest degree |      Freq.     Percent        Cum.
--------------------------+-----------------------------------
   less than high school | 480.972702       11.99       11.99
             high school | 1,891.6334       47.15       59.13
associate/junior college | 452.656901       11.28       70.42
              bachelor's |681.8664156       16.99       87.41
                graduate | 505.084448       12.59      100.00
--------------------------+-----------------------------------
                   Total | 4,012.2139      100.00
```

```
. svyset [weight=wtssnrps], strata(vstrat) psu(vpsu) singleunit(scaled)
(sampling weights assumed)

. svy: tab degree
(running tabulate on estimation sample)

Number of strata =      9
Number of PSUs   =  3,543
Number of obs    =      4,009
Population size  = 4,012.2139
Design df        =      3,534

----------------------
r's       |
highest   |
degree    | proportion
----------+-----------
 less tha |      .1199
 high sch |      .4715
 associat |      .1128
 bachelor |      .1699
 graduate |      .1259
          |
    Total |          1
----------------------
Key: proportion = Cell proportion
```

Example: 2021 GSS in Stata (interval-ratio-level variable)

```
. sum conrinc

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
     conrinc |      2,456    41722.79    39243.69        336   170912.6

. sum conrinc [iweight=wtssnrps]

    Variable |     Obs      Weight        Mean   Std. dev.       Min        Max
-------------+-----------------------------------------------------------------
     conrinc |   2,456  2453.15509    37647.74   37376.88        336   170912.6

. svy: mean conrinc
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =     9          Number of obs   =      2,456
Number of PSUs   = 2,241          Population size = 2,453.1551
                                  Design df       =      2,232

--------------------------------------------------------------
             |             Linearized
             |       Mean   std. err.     [95% conf. interval]
-------------+------------------------------------------------
     conrinc |   37647.74   850.3902      35980.1    39315.38
--------------------------------------------------------------

. estat sd

-------------------------------------
             |       Mean   Std. dev.
-------------+-----------------------
     conrinc |   37647.74    37376.87
-------------------------------------
```


TEXAS A&M
UNIVERSITY