Lecture 2b: Survey weights

Ernesto F. L. Amaral

September 05, 2023
Introduction to Sociological Data Analysis (SOCI 600)

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

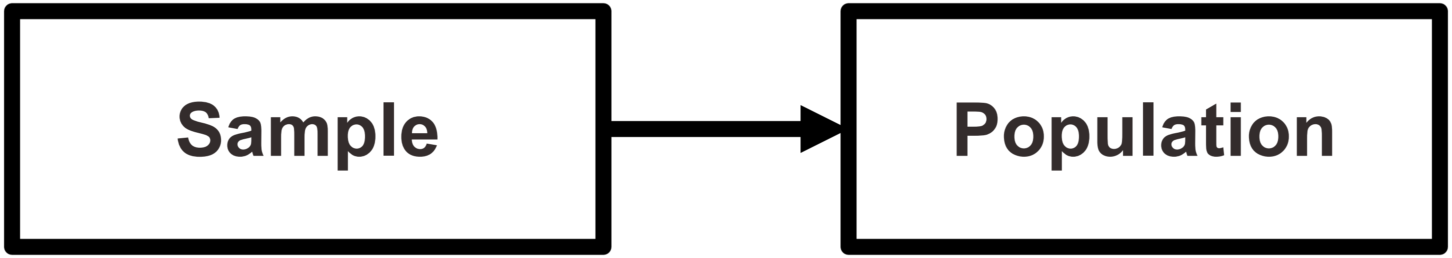
- Inferential statistics
- Survey weights
- Weight options in Stata
- Complex sample cluster design
- Weights in surveys
  - American Community Survey (ACS)
  - General Social Survey (GSS)
- Examples of descriptive statistics

Inferential statistics

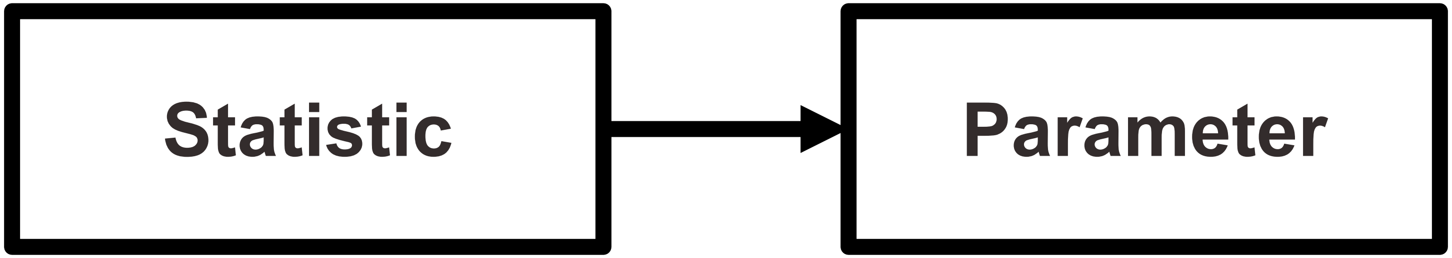
- Social scientists need inferential statistics
  - They almost never have the resources or time to collect data from every case in a population
- Inferential statistics uses data from samples to make generalizations about populations
  - **Population** is the total collection of all cases in which the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

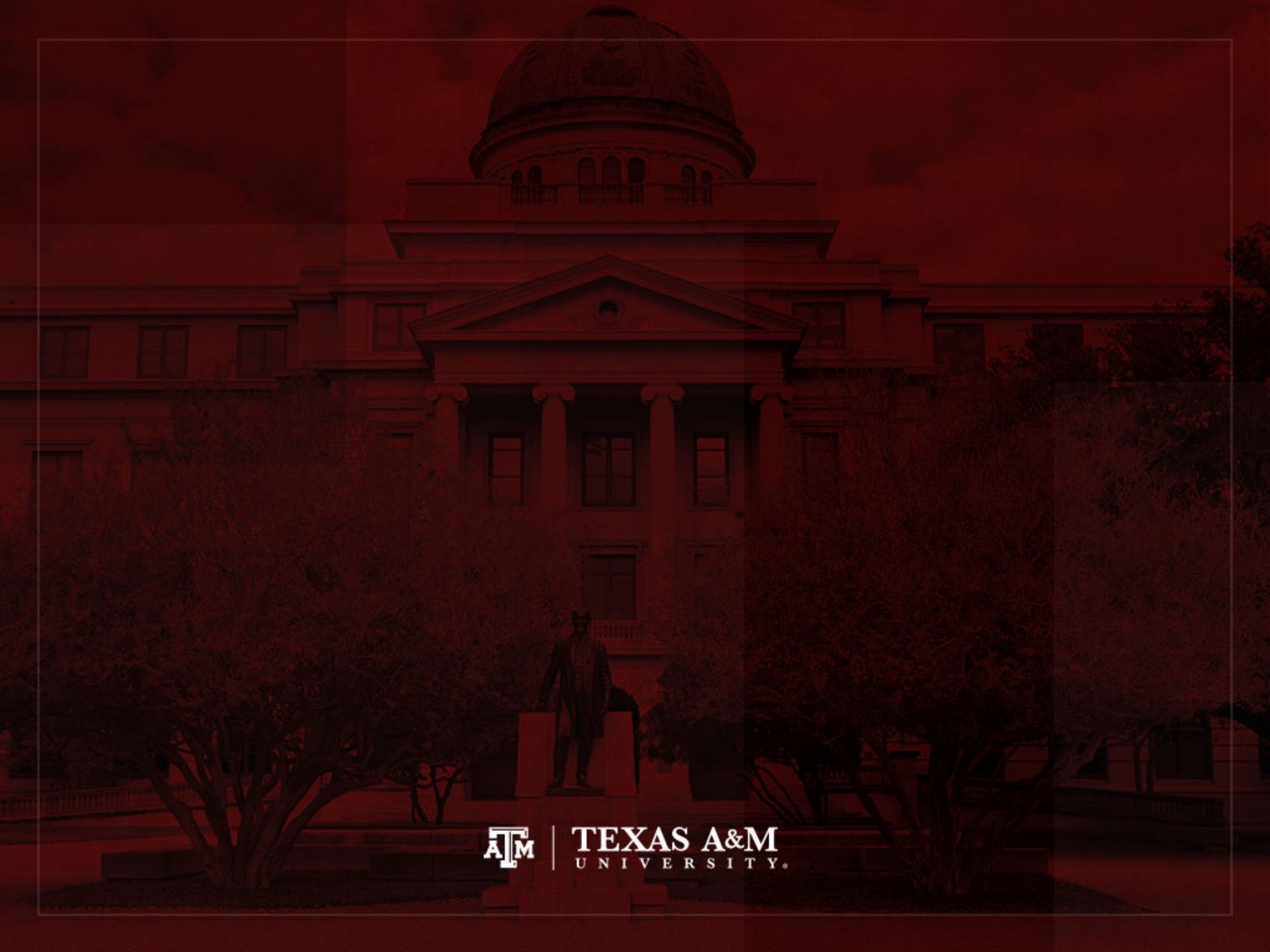
Basic logic and terminology

- 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

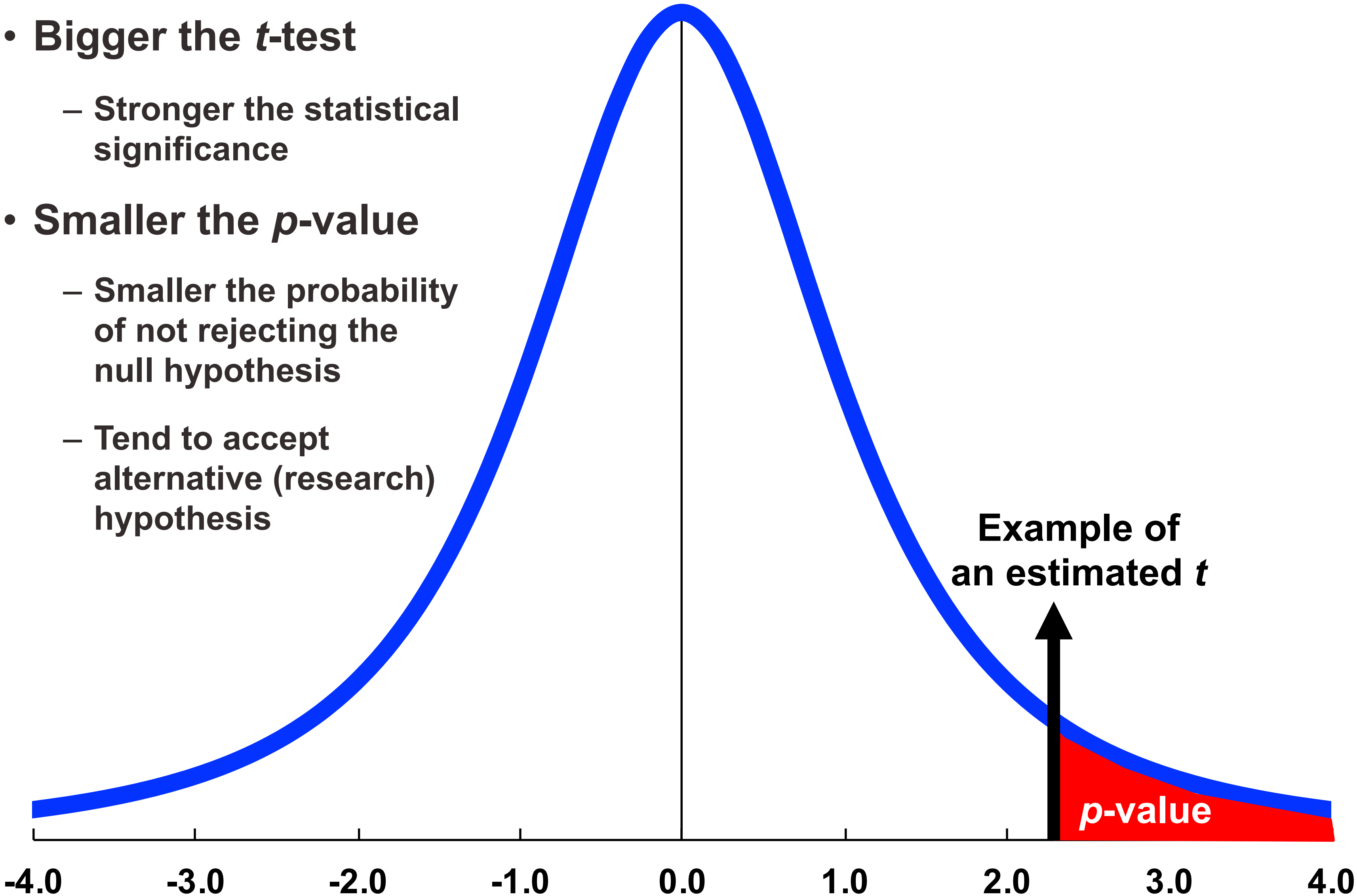
- If we use a weight that expands to the population size (*N*) on regressions
  - We would be incorrectly informing the statistical software that we have a sample with enormous size
  - This would artificially increase the test of statistical significance for the coefficient

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

  - 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**

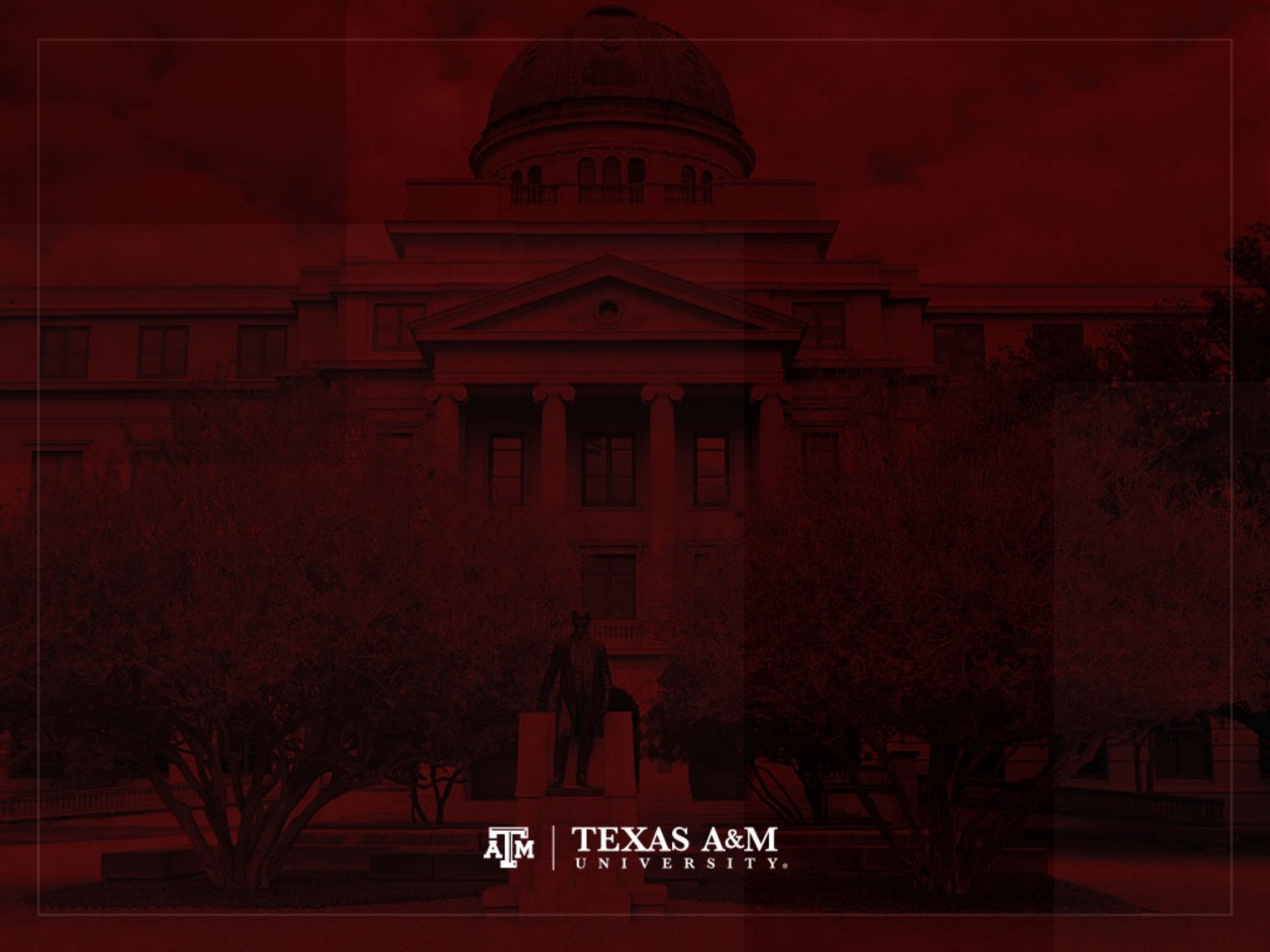


Decisions about hypotheses

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	

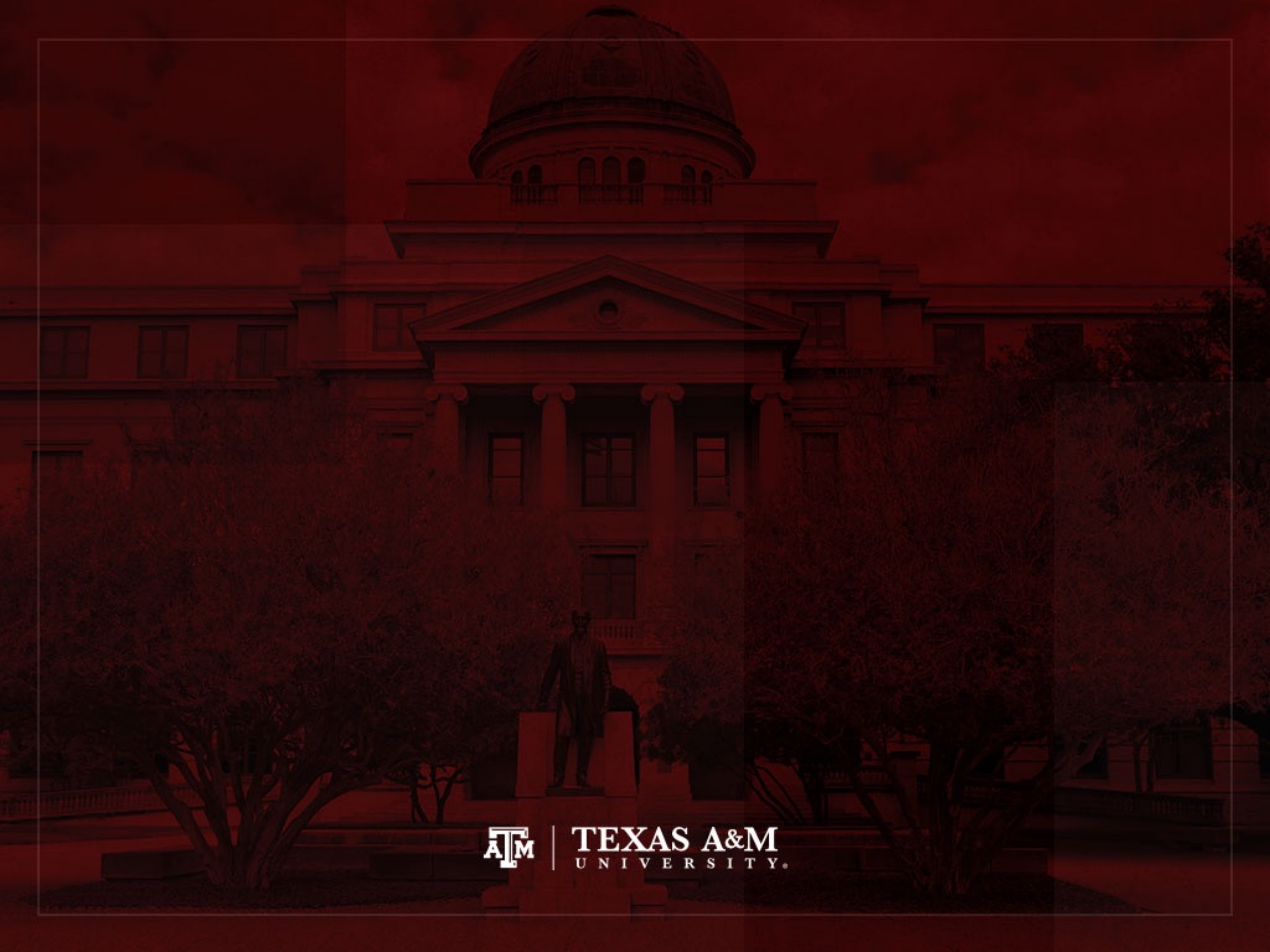
```
. tab sex [fweight=wtssnrps], m
may not use noninteger frequency weights
r(401);
```

```
. 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	$\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 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), $n \geq 30$

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

Cluster design for regressions

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

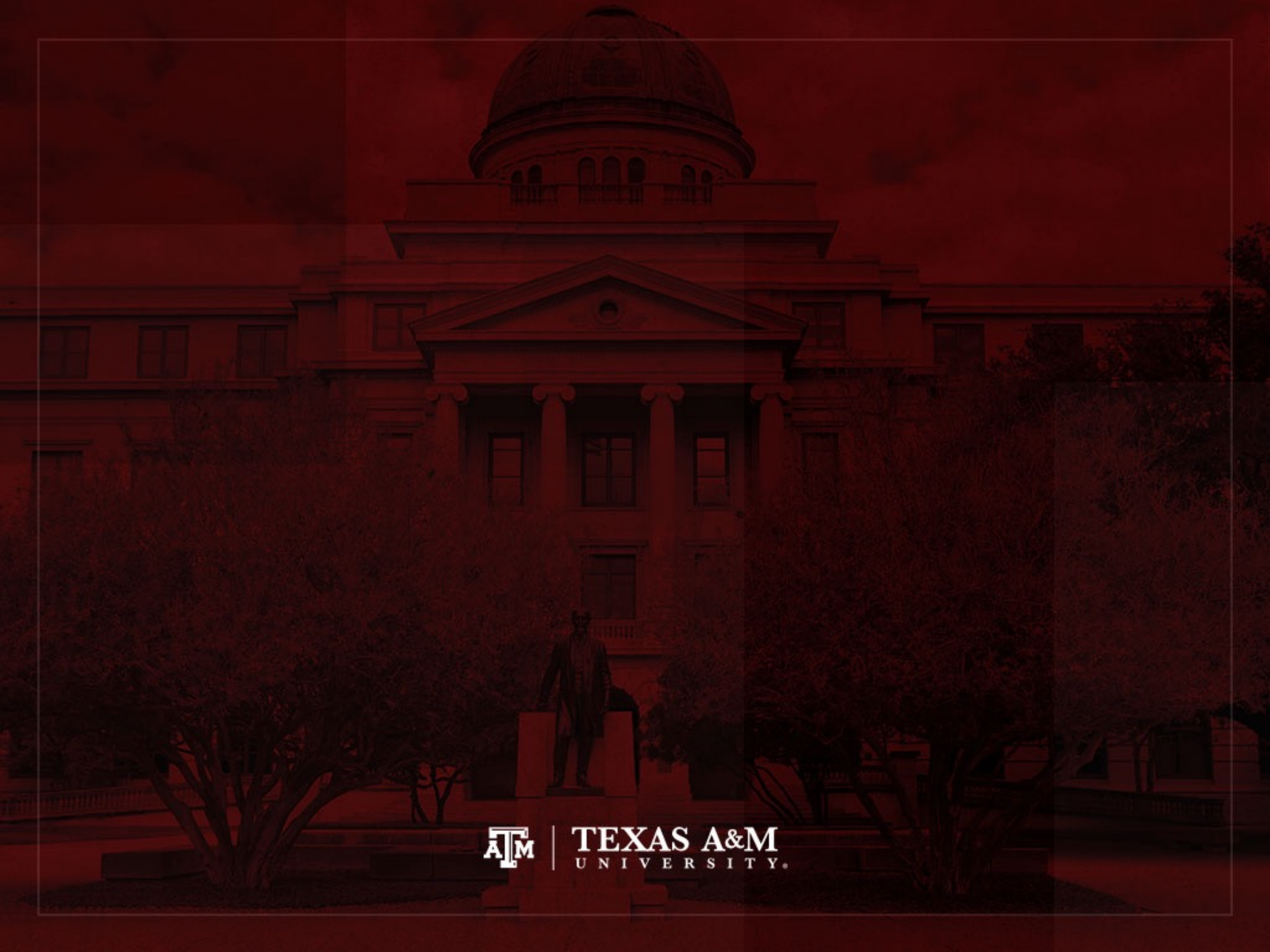
$$t = \frac{\hat{\beta}}{SE_{\hat{\beta}}} = \frac{\hat{\beta}}{\sqrt{\frac{MSE}{S_{xx}}}} = \frac{\hat{\beta}}{\sqrt{\frac{RSS}{df * S_{xx}}}} = \frac{\hat{\beta}}{\sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{(n-2)\sum_i (x_i - \bar{x})^2}}}$$

  - 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 ACS

- In the American Community Survey (ACS) PERWT indicates how many persons in the U.S. population are represented by a given person in an IPUMS sample

  https://usa.ipums.org/usa-action/variables/PERWT#description_section

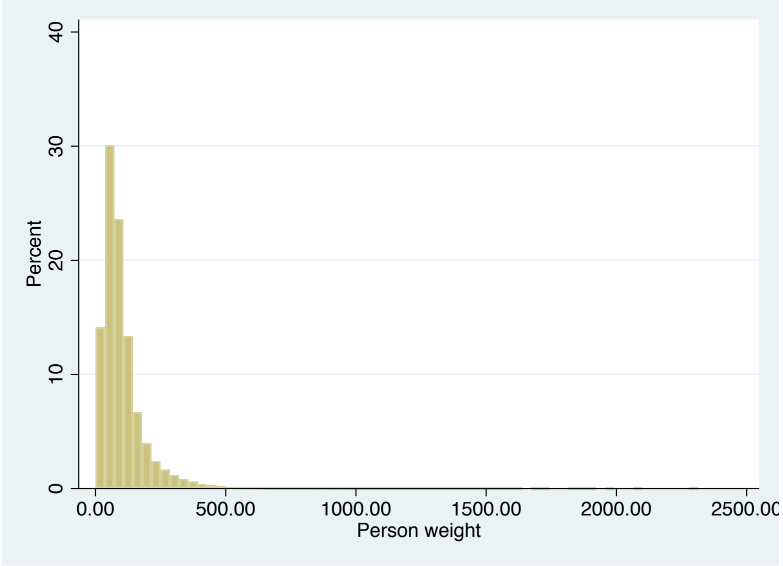
- HHWT indicates how many households in the U.S. population are represented by a given household in an IPUMS sample
  - Users should also be sure to select one person (e.g., PERNUM = 1) to represent the entire household

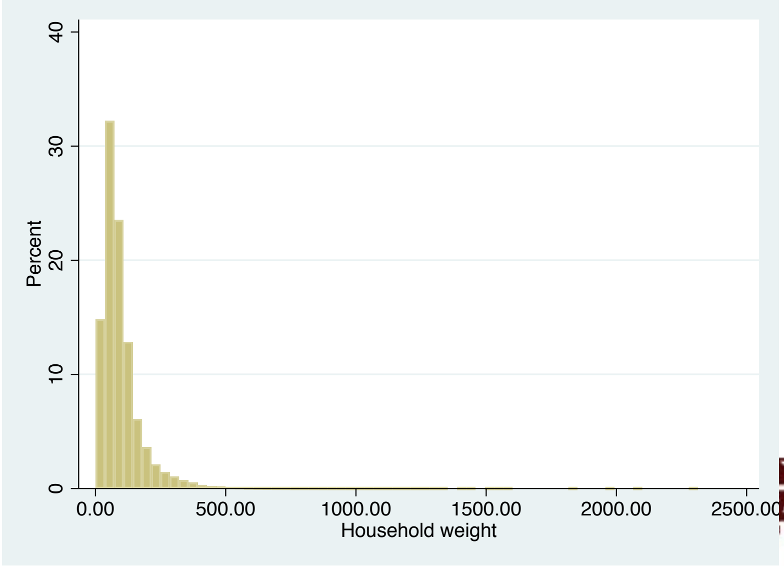
  https://usa.ipums.org/usa-action/variables/HHWT#description_section

Summary of 2018 ACS weights

```
. 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
```

```
. sum hhwt if pernum==1, d

                       Household weight
-------------------------------------------------------------
      Percentiles      Smallest
 1%            8              1
 5%           16              1
10%           25              1       Obs           1,410,976
25%           48              1       Sum of Wgt.   1,410,976

50%           73                      Mean           91.85967
                        Largest       Std. Dev.      75.18581
75%          112           1837
90%          173           1990       Variance       5652.906
95%          234           2097       Skewness        2.88203
99%          386           2313       Kurtosis       19.09996
```





ACS has a cluster sample

- All IPUMS samples are cluster samples
  - Samples are not individual-level samples
  - They are samples of households or dwellings
  - Individuals are sampled as parts of households
    - Information about all individuals within the same household
- Samples are also stratified to some degree
  - U.S. Census Bureau divides population into strata based on key characteristics
  - Sample separately from each stratum
  - Each stratum is proportionately represented in the final sample

Source: https://usa.ipums.org/usa/chapter2/chapter2.shtml

ACS variables for cluster design

- Sampling weight (PERWT or HHWT)
  - It is chosen based on type of research question
- Household strata (STRATA)
  - Integrated variable that represents the impact of the sample design stratification on the estimates of variance and standard errors
  - In the 2005 onward ACS samples, strata are defined as unique Public Use Micro-data Areas (PUMA)
- Household cluster (CLUSTER)
  - Integrated variable which uniquely identifies each household record in a given sample

Source: https://usa.ipums.org/usa/complex_survey_vars/userNotes_variance.shtml

ACS complex sample design

- Account for ACS sample design in Stata

`svyset cluster [pweight=perwt], strata(strata)`

```
. svyset cluster [pweight=perwt], strata(strata)

      pweight: perwt
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: cluster
        FPC 1: <zero>
```

- 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`

Source: https://usa.ipums.org/usa/complex_survey_vars/userNotes_variance.shtml

Mean income

```
. mean income [pweight=perwt]

Mean estimation                   Number of obs   =  2,642,681
```

	Mean	Std. Err.	[95% Conf.	Interval]
income	31175.11	39.98542	31096.74	31253.48

```
. svy: mean income
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =   2,351      Number of obs   =   2,642,681
Number of PSUs   = 1408111      Population size = 262,216,823
                                Design df       =   1,405,760
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
income	31175.11	40.99966	31094.75	31255.47

Source: 2018 American Community Survey.

For subpopulations

- We use the following approach to conduct subpopulation analysis without compromising the data design structure
  - We produce estimates for the population of interest, while incorporating the full sample design information for variance estimation

- Example: only people with 15–64 years of age

```
svyset cluster [pweight=perwt], strata(strata)
svy, subpop(if age>=15 & age<=64): mean var1
```

Source: https://usa.ipums.org/usa/complex_survey_vars/userNotes_variance.shtml

Mean income

```
. svy: mean income
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =     2,351      Number of obs   =     2,642,681
Number of PSUs   =   1408111      Population size =   262,216,823
                                  Design df       =     1,405,760
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
income	31175.11	40.99966	31094.75	31255.47

```
. svy, subpop(if income!=.): 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 =     2,642,681
                                  Subpop. size    =   262,216,823
                                  Design df       =     1,408,625
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
income	31175.11	41.00232	31094.74	31255.47

If we consider that missing cases are part of the population, we need to inform that subpopulation is only non-missing cases

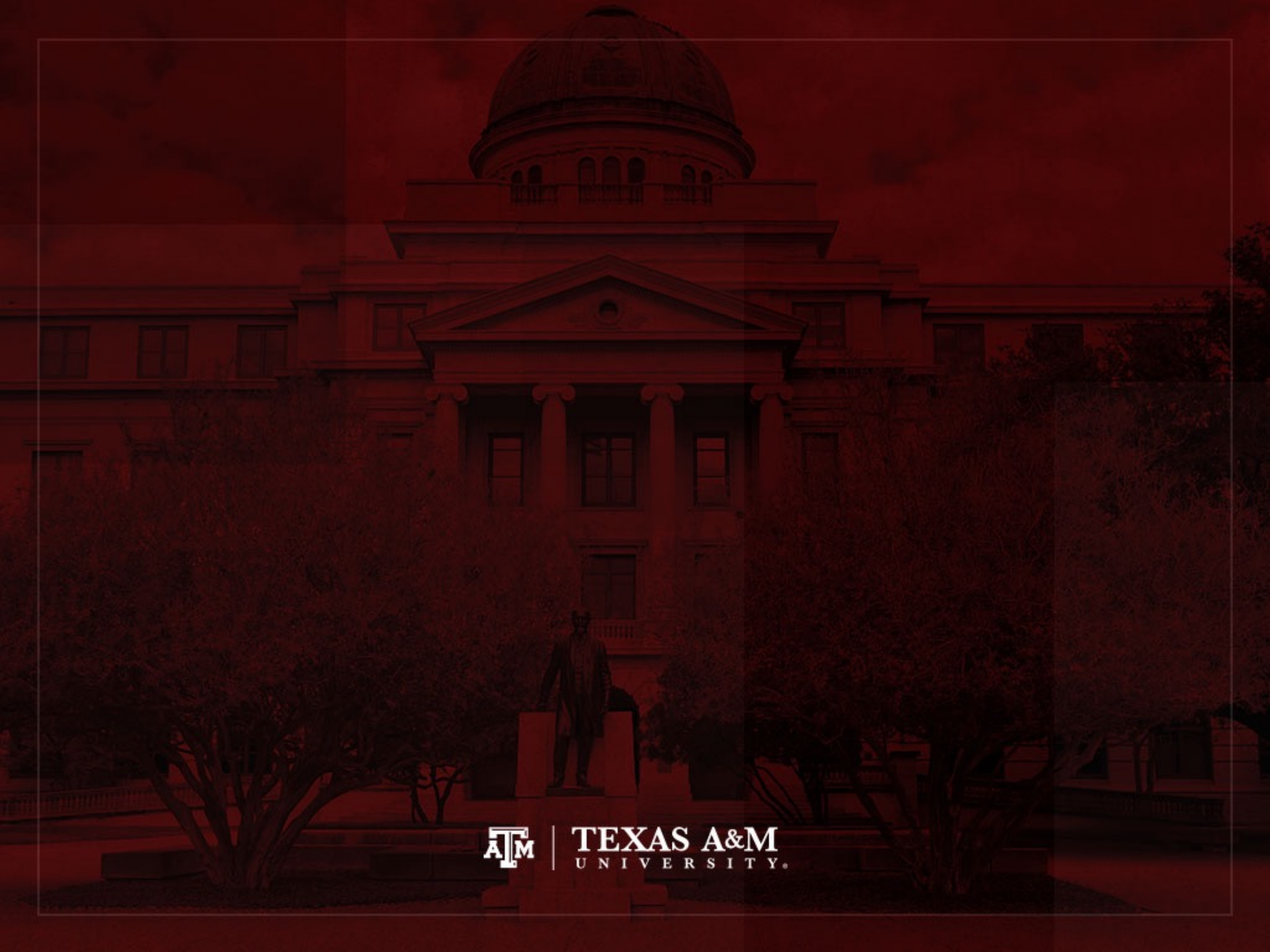
Source: 2018 American Community Survey.

Mean income (15–64)

```
. svy, subpop(if age>=15 & age<=64): mean income
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =    2,351      Number of obs     =    3,175,157
Number of PSUs   = 1410150       Population size   =  323,036,047
                                 Subpop. no. obs   =    2,004,091
                                 Subpop. size      =  209,809,274
                                 Design df         =    1,407,799

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
      income |   36736.34   48.39971      36641.48     36831.2
--------------------------------------------------------------

. svy, subpop(if age>=15 & age<=64 & income!=.): 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   =    2,004,091
                                 Subpop. size      =  209,809,274
                                 Design df         =    1,408,625

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
      income |   36736.34   48.40061      36641.48    36831.21
--------------------------------------------------------------
```

If we consider that missing cases are part of the population, we need to inform that subpopulation is only non-missing cases

Source: 2018 American Community Survey.


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

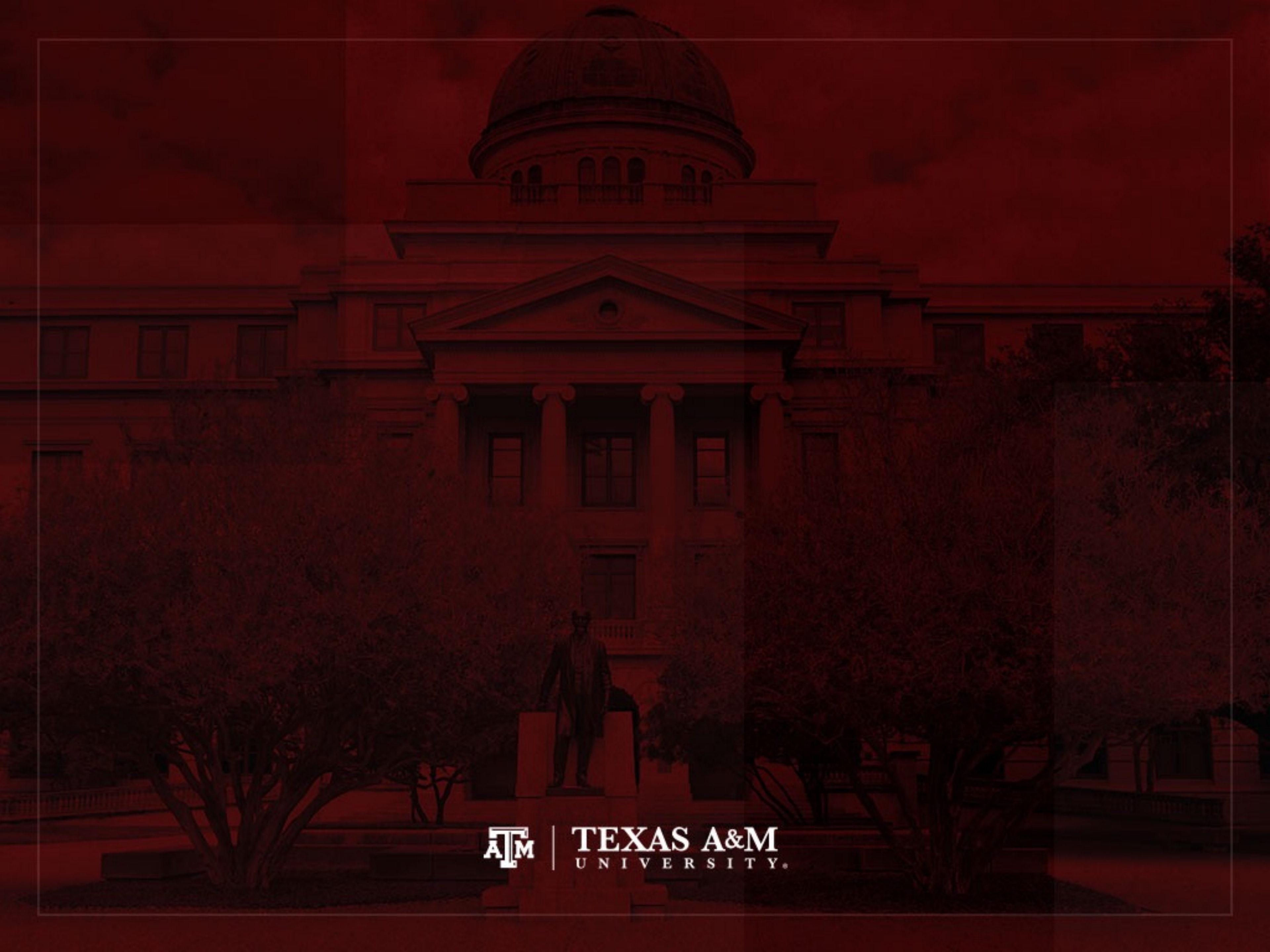
(https://gssdataexplorer.norc.org/gss_stdError)

- 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: 2019 ACS, Texas (nominal-level variable)

```
. tab sex
```

sex	Freq.	Percent	Cum.
male	134,479	49.30	49.30
female	138,297	50.70	100.00
Total	272,776	100.00	

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

sex	Freq.	Percent	Cum.
male	14,389,011	49.62	49.62
female	14,606,870	50.38	100.00
Total	28,995,881	100.00	

```
. svyset cluster [pweight=perwt], strata(strata)

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

Number of strata =       212
Number of PSUs   =   114,016
Number of obs     =   272,776
Population size   = 28,995,881
Design df         =   113,804
```

sex	proportion
male	.4962
female	.5038
Total	1

Key: proportion = **Cell proportion**

Example: 2019 ACS, Texas (ordinal-level variable)

```
. tab educ
```

educational attainment [general version]	Freq.	Percent	Cum.
n/a or no schooling	18,672	6.85	6.85
nursery school to grade 4	23,056	8.45	15.30
grade 5, 6, 7, or 8	21,619	7.93	23.22
grade 9	7,263	2.66	25.89
grade 10	6,783	2.49	28.37
grade 11	7,319	2.68	31.06
grade 12	74,662	27.37	58.43
1 year of college	33,207	12.17	70.60
2 years of college	15,505	5.68	76.28
4 years of college	41,586	15.25	91.53
5+ years of college	23,104	8.47	100.00
Total	272,776	100.00	

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

educational attainment [general version]	Freq.	Percent	Cum.
n/a or no schooling	2,338,799	8.07	8.07
nursery school to grade 4	2,791,197	9.63	17.69
grade 5, 6, 7, or 8	2,627,585	9.06	26.75
grade 9	876,753	3.02	29.78
grade 10	758,921	2.62	32.40
grade 11	825,208	2.85	35.24
grade 12	7,564,180	26.09	61.33
1 year of college	3,606,553	12.44	73.77
2 years of college	1,561,001	5.38	79.15
4 years of college	3,996,149	13.78	92.93
5+ years of college	2,049,535	7.07	100.00
Total	28,995,881	100.00	

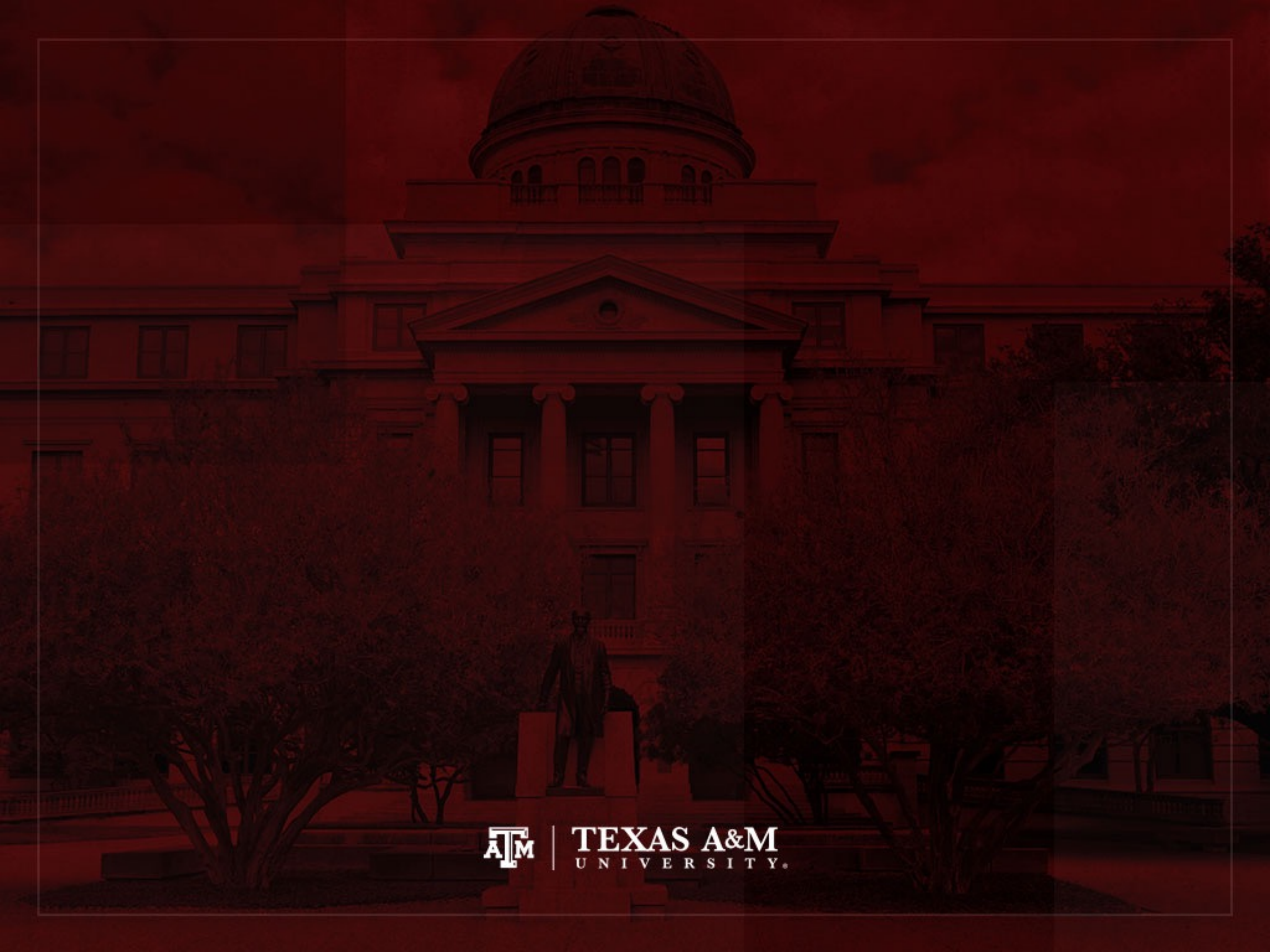
```
. svyset cluster [pweight=perwt], strata(strata)
. svy: tab educ
(running tabulate on estimation sample)

Number of strata =      212
Number of PSUs   =  114,016
Number of obs     =   272,776
Population size   = 28,995,881
Design df         =   113,804
```

educational attainment [general version]	proportion
n/a or n	.0807
nursery	.0963
grade 5,	.0906
grade 9	.0302
grade 10	.0262
grade 11	.0285
grade 12	.2609
1 year o	.1244
2 years	.0538
4 years	.1378
5+ years	.0707
Total	1

Key: proportion = **Cell proportion**

Example: 2019 ACS, Texas (interval-ratio-level variable)

```
. sum income

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      income |    219,299    32291.87    58306.42          0     483000

. sum income [iweight=perwt]

    Variable |     Obs      Weight        Mean   Std. dev.       Min        Max
-------------+-----------------------------------------------------------------
      income | 219,299    22421711    31745.27   53892.93          0     483000

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

Survey: Mean estimation

Number of strata =     212        Number of obs   =    219,299
Number of PSUs   = 113,830        Population size = 22,421,711
                                  Design df       =    113,618

--------------------------------------------------------------
             |             Linearized
             |       Mean   std. err.     [95% conf. interval]
-------------+------------------------------------------------
      income |   31745.27   142.2892      31466.39    32024.16
--------------------------------------------------------------

. estat sd

-------------------------------------
             |       Mean   Std. dev.
-------------+-----------------------
      income |   31745.27   53893.05
-------------------------------------
```


TEXAS A&M
UNIVERSITY

Examples of descriptive statistics

- Nominal-level variable
- Ordinal-level variable
- Interval-ratio-level variable
- Boxplots
- Age-sex structure

Nominal-level variable (Example: 2018 ACS in Stata)

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

raceth	Freq.	Percent	Cum.
White	197,034,851	60.22	60.22
African American	40,373,281	12.34	72.56
Hispanic	59,740,273	18.26	90.82
Asian	18,662,293	5.70	96.53
Native American	2,170,486	0.66	97.19
Ohter races	9,186,255	2.81	100.00
Total	327,167,439	100.00	

```
. count if raceth!=.
  3,214,539
```

Edited table

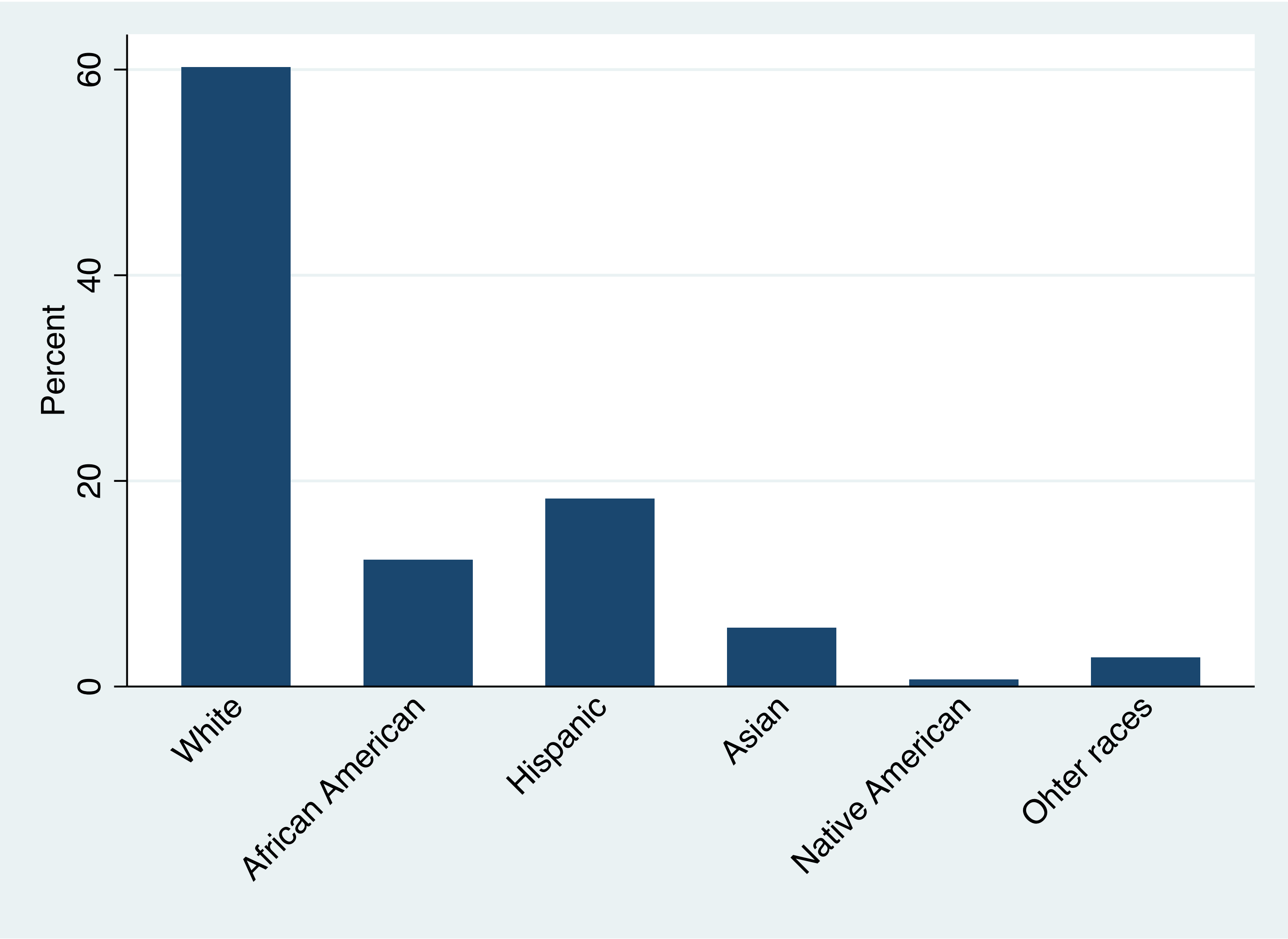
Table 1. Distribution of U.S. population by race/ethnicity, 2018

Race/ethnicity	Percentage
Non-Hispanic White	60.22
Non-Hispanic African American	12.34
Hispanic	18.26
Non-Hispanic Asian	5.70
Non-Hispanic Native American	0.66
Other races	2.81
Total	**99.99**
Population size (N)	327,167,439
Sample size (n)	3,214,539

Source: 2018 American Community Survey.

Column graph for race/ethnicity, 2018

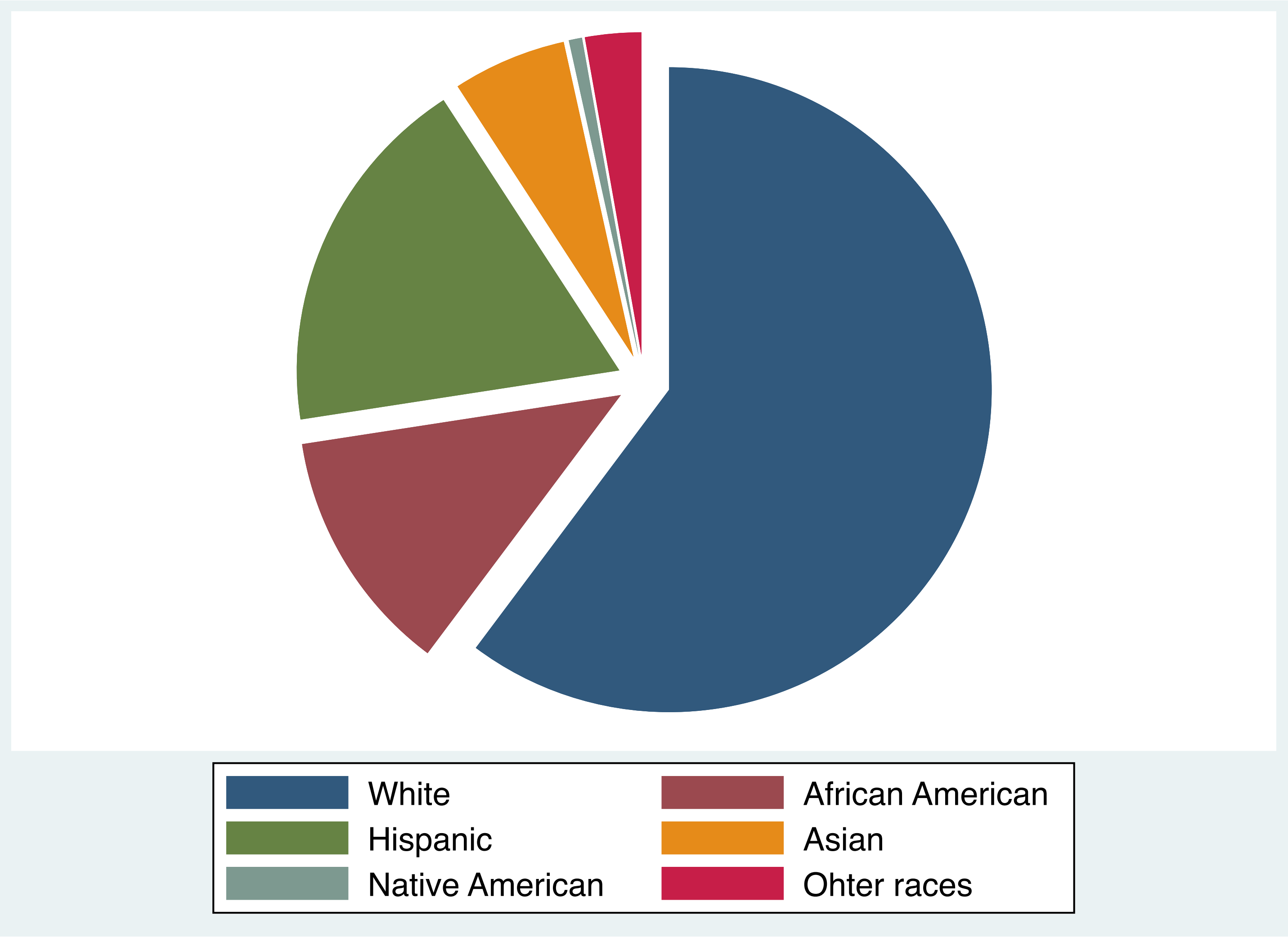
```
graph bar [fweight=perwt], over(raceth,
label(angle(45))) ytitle("Percent")
```



Source: 2018 American Community Survey.

Pie graph for race/ethnicity, 2018

```
graph pie [fweight=perwt], over(raceth) pie(_all, explode)
```



Source: 2018 American Community Survey.

Ordinal-level variable (Example: 2018 ACS in Stata)

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

educgr	Freq.	Percent	Cum.
Less than high school	97,758,814	29.88	29.88
High school	92,183,547	28.18	58.06
Some college	60,822,461	18.59	76.65
College	47,865,798	14.63	91.28
Graduate school	28,536,819	8.72	100.00
Total	327,167,439	100.00	

```
. count if educgr!=.
  3,214,539
```

Edited table

Table 1. Distribution of U.S. population by educational attainment, 2018

Educational attainment	**Percentage**
Less than high school	29.88
High school	28.18
Some college	18.59
College	14.63
Graduate school	8.72
Total	**100.00**
Population size (N)	327,167,439
Sample size (n)	3,214,539

Source: 2018 American Community Survey.

Interval-ratio-level variable (Example: 2018 ACS in Stata)

```
. table year [fweight=perwt] if income!=0, c(min income p25 income p50 income p75 income max income)
```

Census year	min(income)	p25(income)	med(income)	p75(income)	max(income)
2018	4	16400	35000	61000	718000

```
. table year [fweight=perwt] if income!=0, c(iqr income sd income mean income)
```

Census year	iqr(income)	sd(income)	mean(income)
2018	44600	62143.93	50043.98

```
. count if income==. | income==0
  1,640,226
```

Survey design for income

```
. ***Complex survey design
. 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

```
.
. ***Estimate standard deviation
. estat sd
```

	Mean	Std. Dev.
income	50043.98	61547.67

Edited table

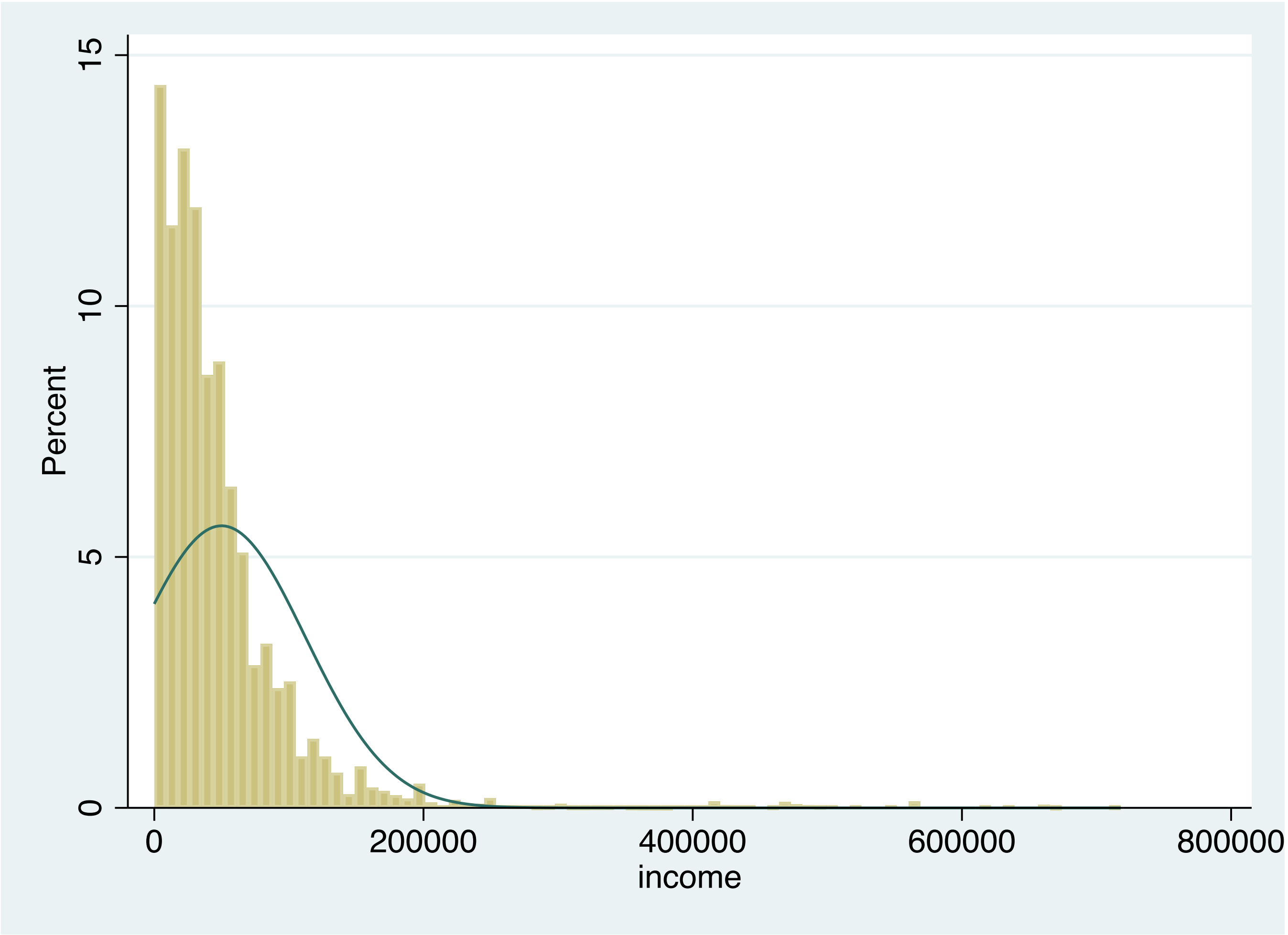
Table 1. Descriptive statistics of respondents' wage and salary income, U.S. population, 2018

Statistics	Income
Mean	50,043.98
Minimum	4.00
25th percentile	16,400.00
Median	35,000.00
75th percentile	61,000.00
Maximum	718,000.00
Range	717,996.00
Interquartile range	44,600.00
Standard deviation	61,547.67
Population size (N)	**163,349,075**
Sample size (n)	**1,574,313**
Missing cases	1,640,226

Source: 2018 American Community Survey.

Histogram of wage and salary income, U.S. population, 2018

```
hist income [fweight=perwt] if income!=0, percent normal
```



Source: 2018 American Community Survey.

Obs.: Only people with some wage and salary income are included (different than zero).

Wage and salary income by sex, 2018 ACS

```
. ***Income
. table year [fweight=perwt] if income!=0, c(mean income p50 income)
```

Census year	mean(income)	med(income)
2018	50043.98	35000

```
. ***Income by sex
. table female [fweight=perwt] if income!=0, c(mean income p50 income)
```

female	mean(income)	med(income)
Male	59014.14	40000
Female	40294.34	30000

Wage and salary income by race/ethnicity, 2018 ACS

```
. ***Income by race/ethnicity
. table raceth [fweight=perwt] if income!=0, c(mean income p50 income)
```

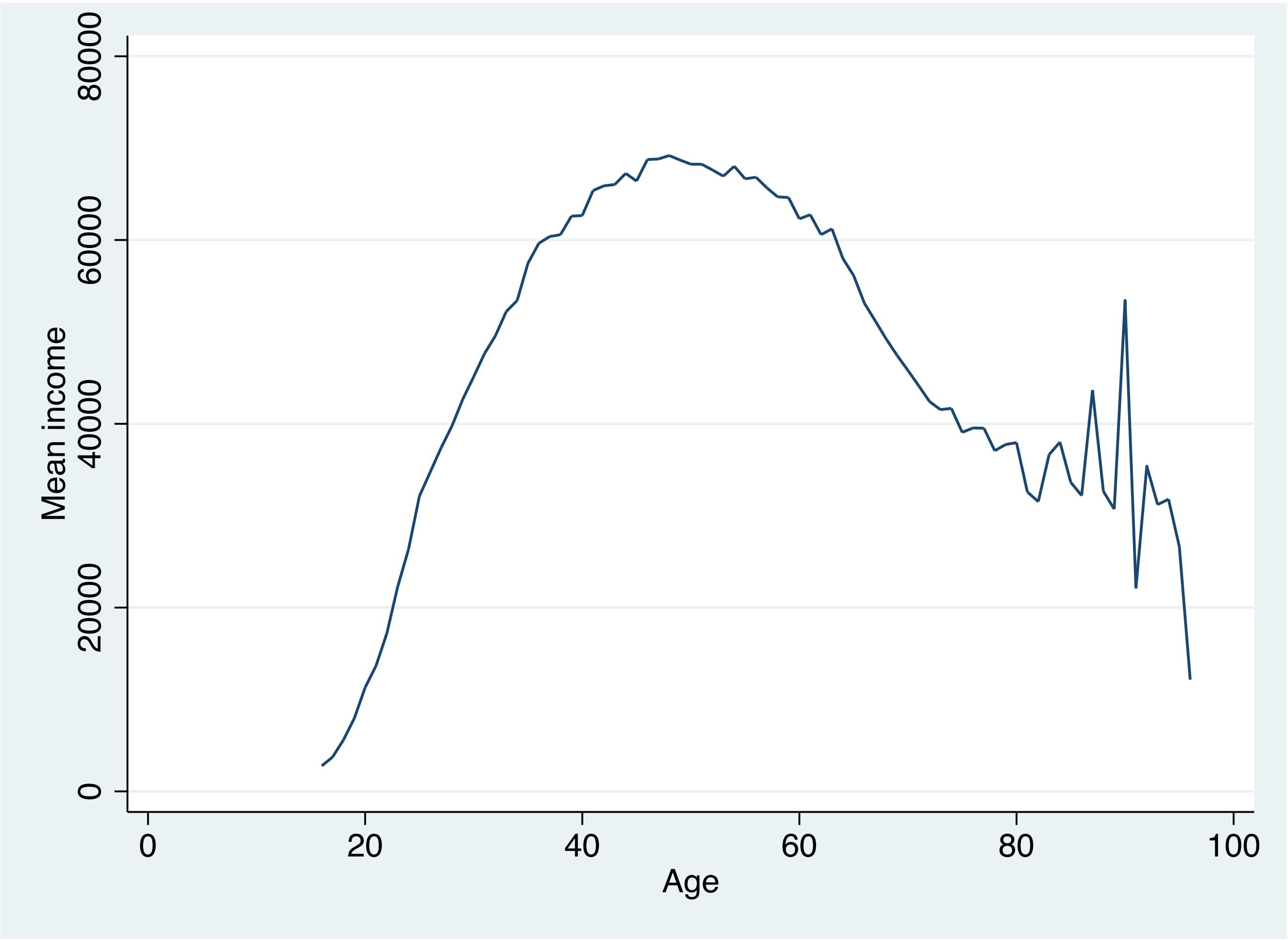
raceth	mean(income)	med(income)
White	55289.18	40000
African American	37183.63	29000
Hispanic	36236.16	27500
Asian	64154.23	43000
Native American	34851.55	27000
Ohter races	44162.79	30000

Wage and salary income by education, 2018 ACS

```
. ***Income by educational attainment
. table educgr [fweight=perwt] if income!=0, c(mean income p50 income)
```

educgr	mean(income)	med(income)
Less than high school	22750.89	18000
High school	34055.76	27000
Some college	39607.05	30300
College	67654.84	50000
Graduate school	98541.49	72000

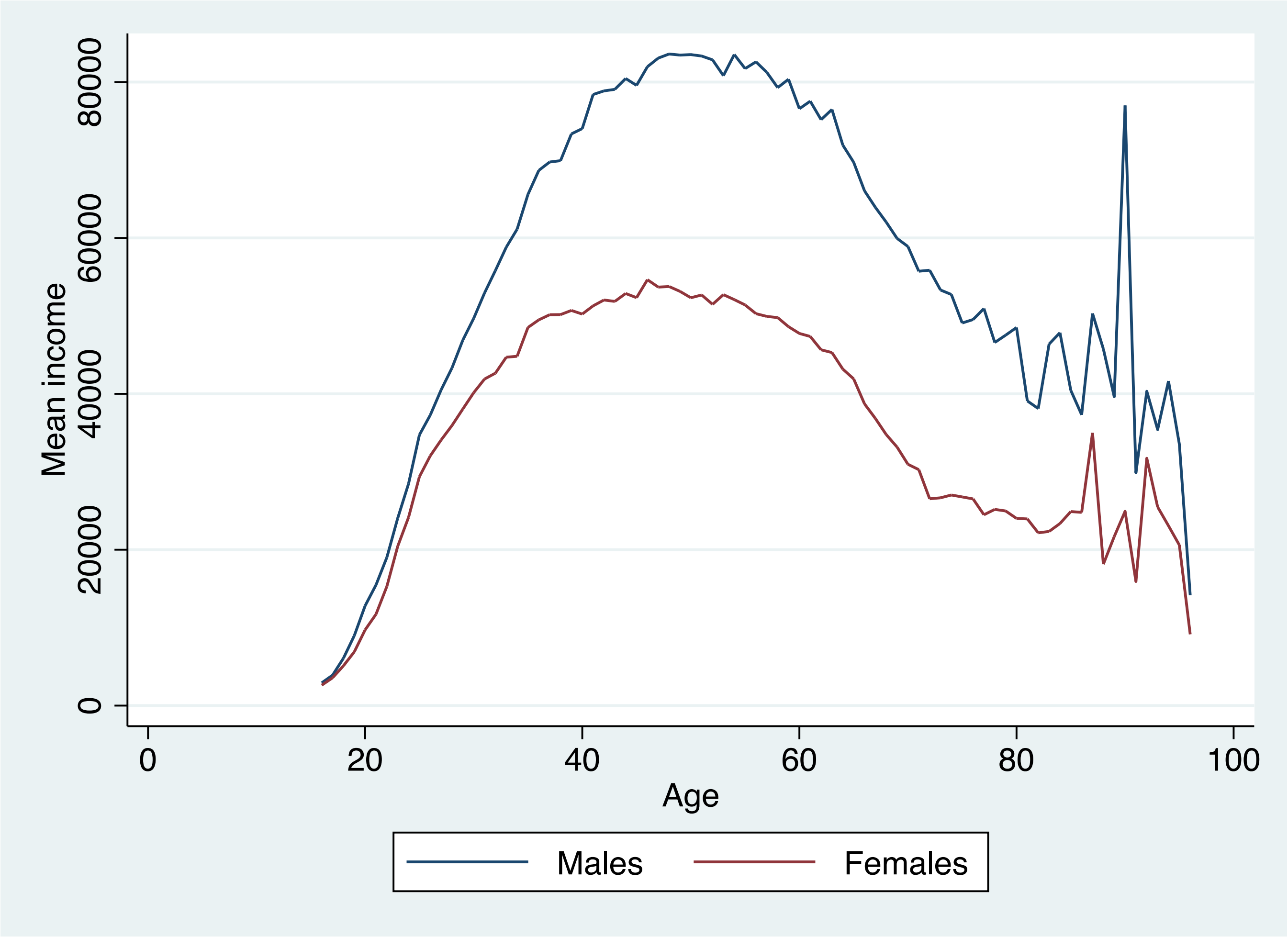
Mean income by age, U.S. population, 2018



Source: 2018 American Community Survey.

Obs.: Only people with some wage and salary income are included (different than zero).

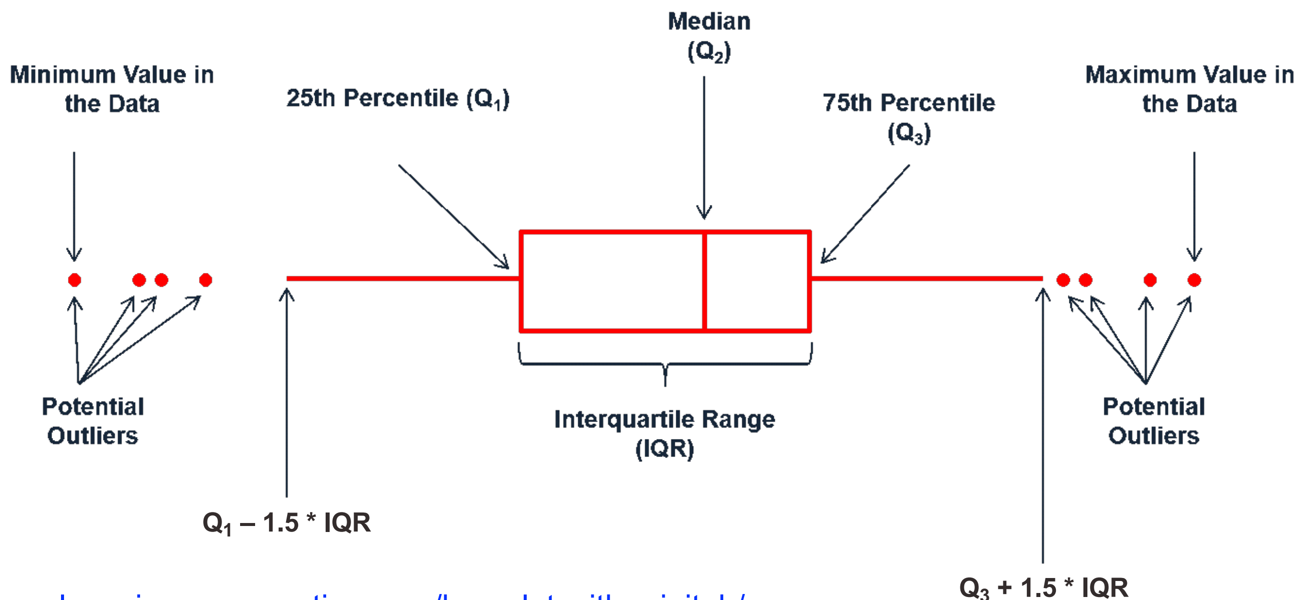
Mean income by age and sex, U.S. population, 2018



Source: 2018 American Community Survey.
Obs.: Only people with some wage and salary income are included (different than zero).

Boxplots

- Boxplot is also known as ”box and whiskers plot”
  - It provides a way to visualize and analyze dispersion
  - Useful when comparing distributions
  - It uses median, range, interquartile range, outliers
  - Easier to read all this information than in tables



Source: https://www.leansigmacorporation.com/box-plot-with-minitab/

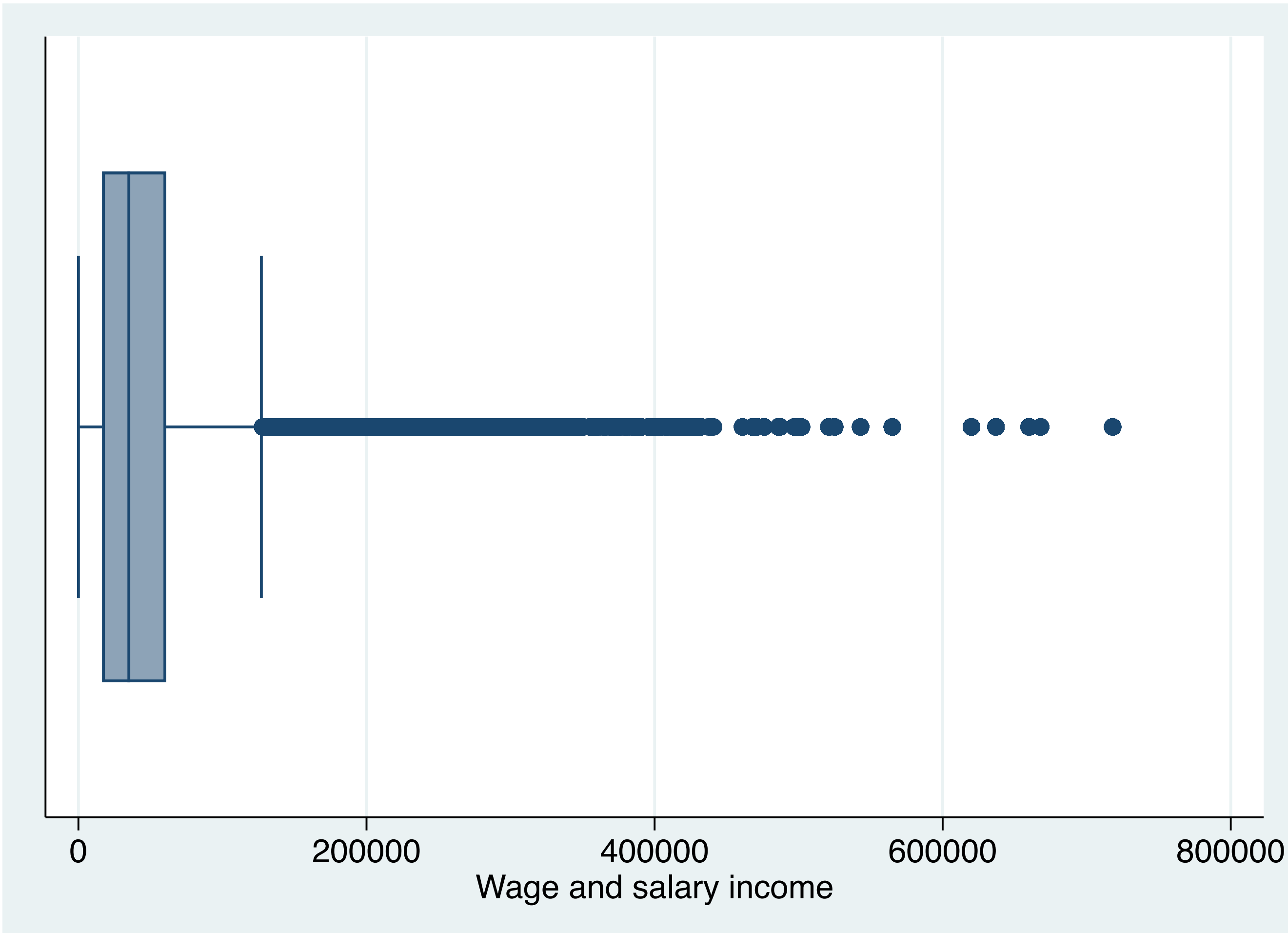
Example: 2018 ACS in Stata

- Generate box plot for respondents’ wage and salary income

```
graph hbox income if income!=0 [fweight=perwt],
        ytitle(Wage and salary income)
```

Edited figure

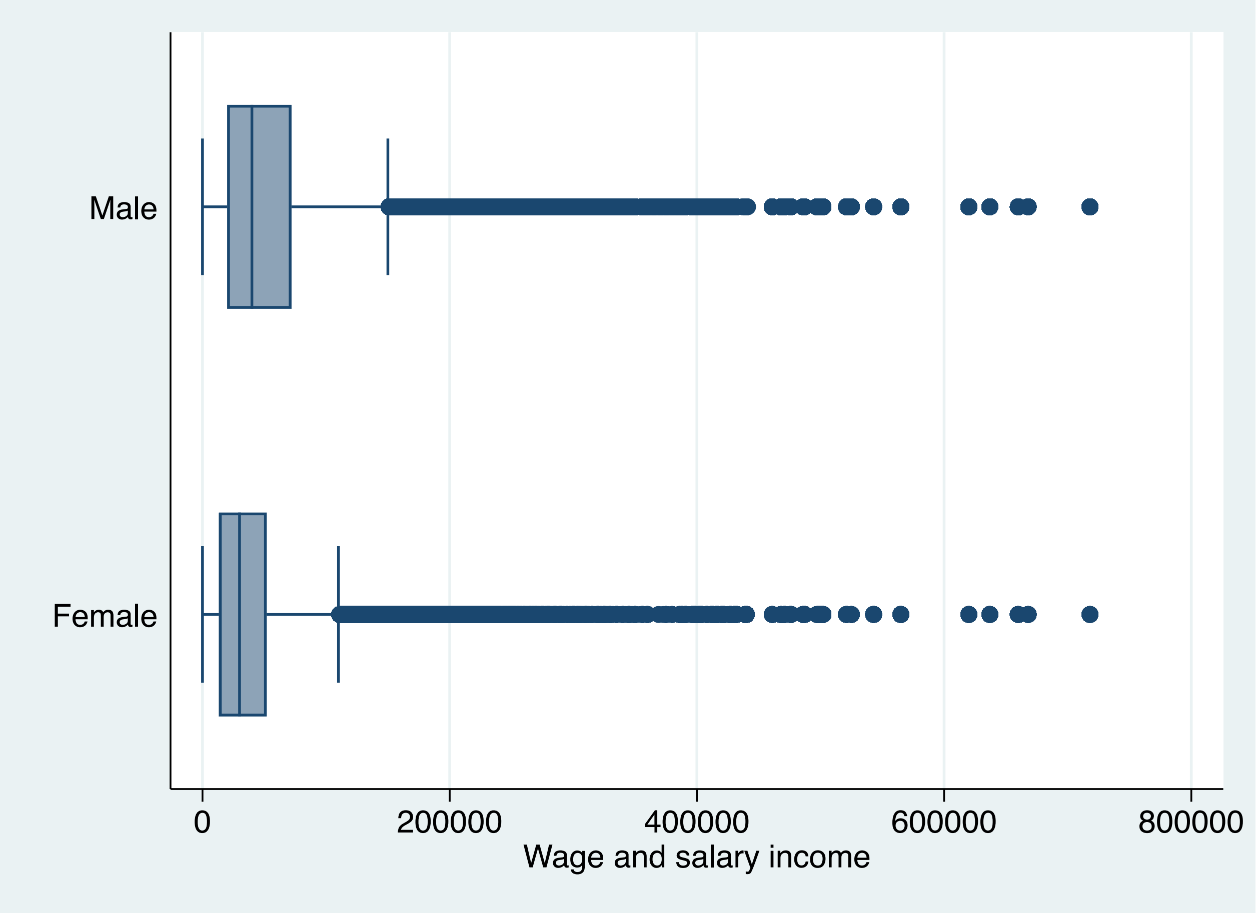
Figure 1. Distribution of respondents' wage and salary income, U.S. population, 2018



Source: 2018 American Community Survey.

Income by sex, 2018

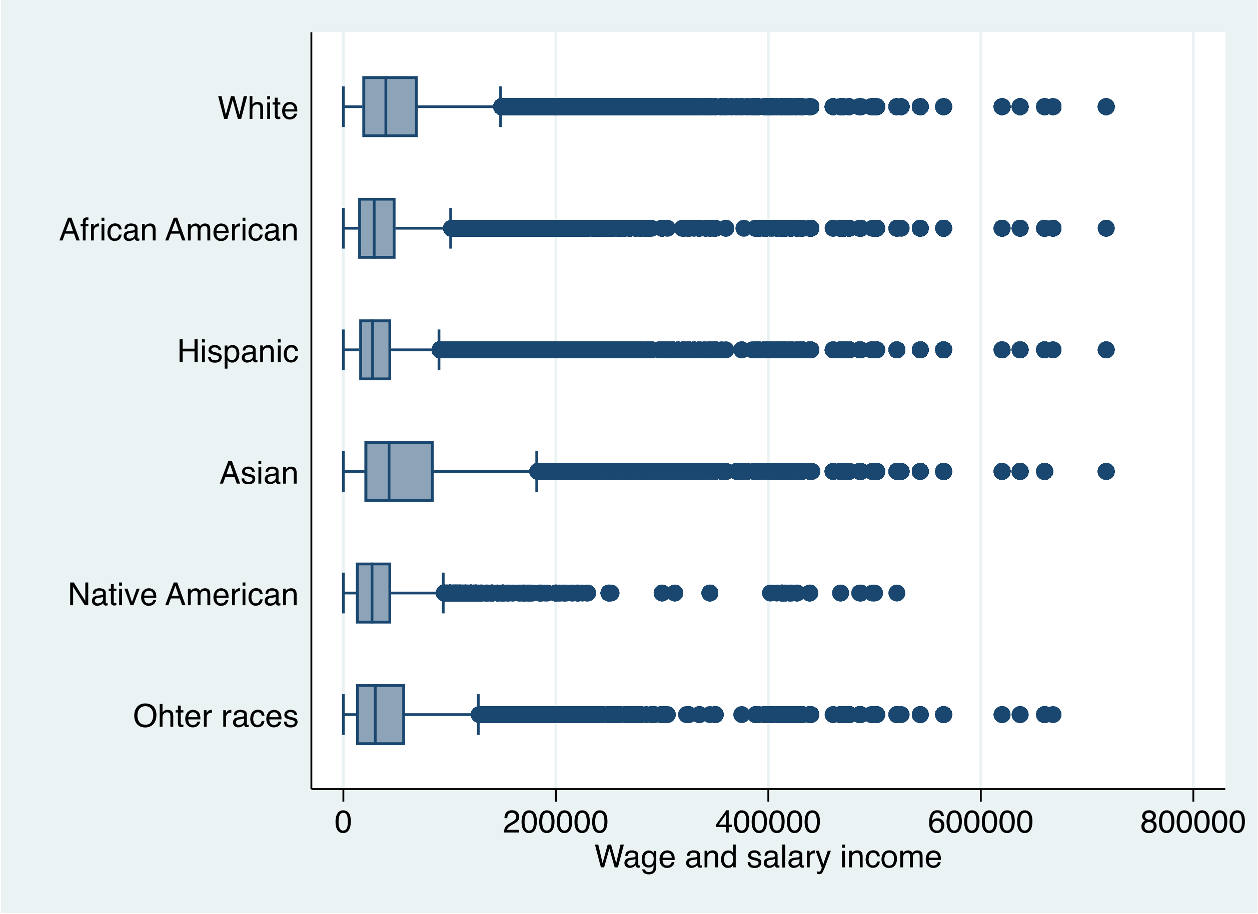
```
graph box income if income!=0 [fweight=perwt],
over(female) ytitle(Wage and salary income)
```



Source: 2018 American Community Survey.

Income by race/ethnicity, 2018

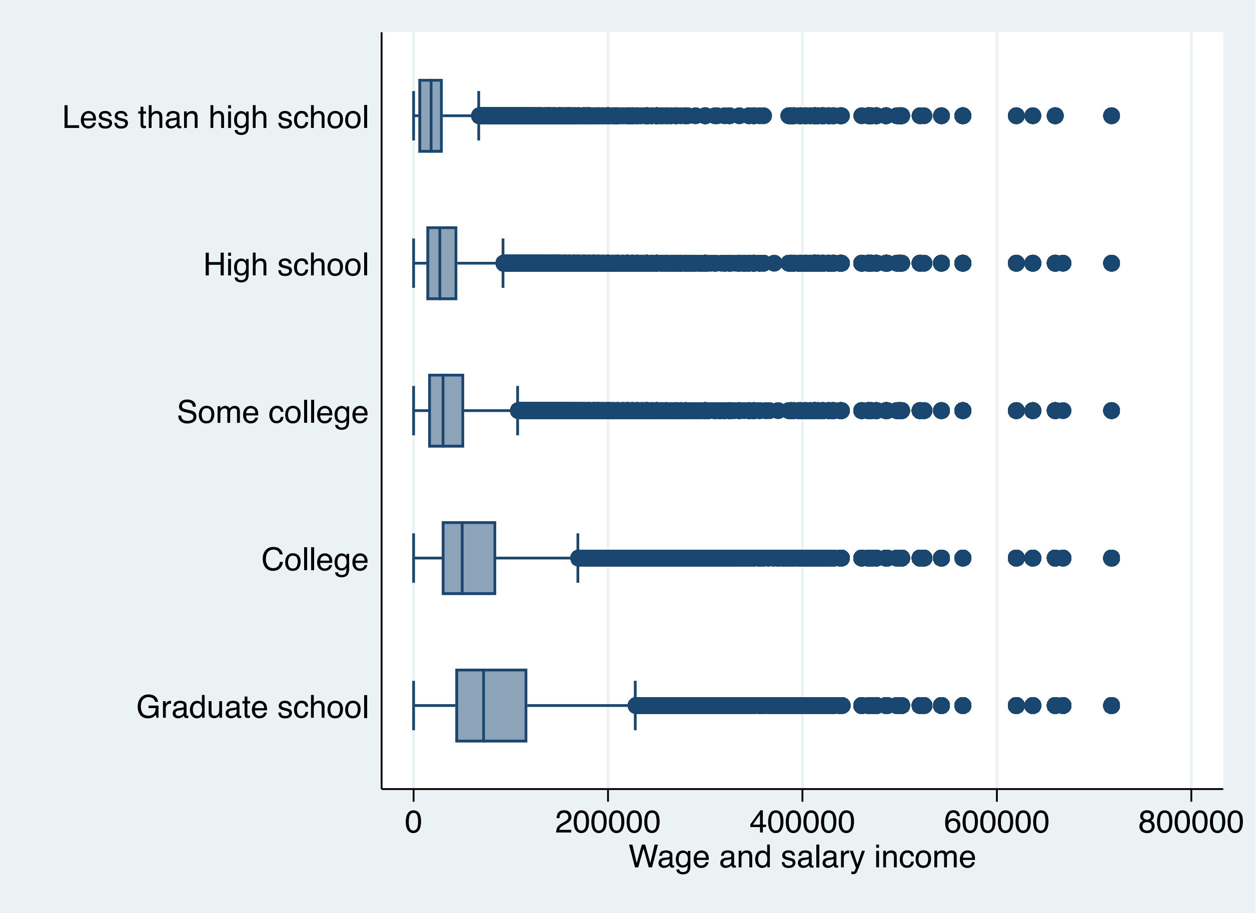
```
graph box income if income!=0 [fweight=perwt],
over(raceth) ytitle(Wage and salary income)
```



Source: 2018 American Community Survey.

Income by education, 2018

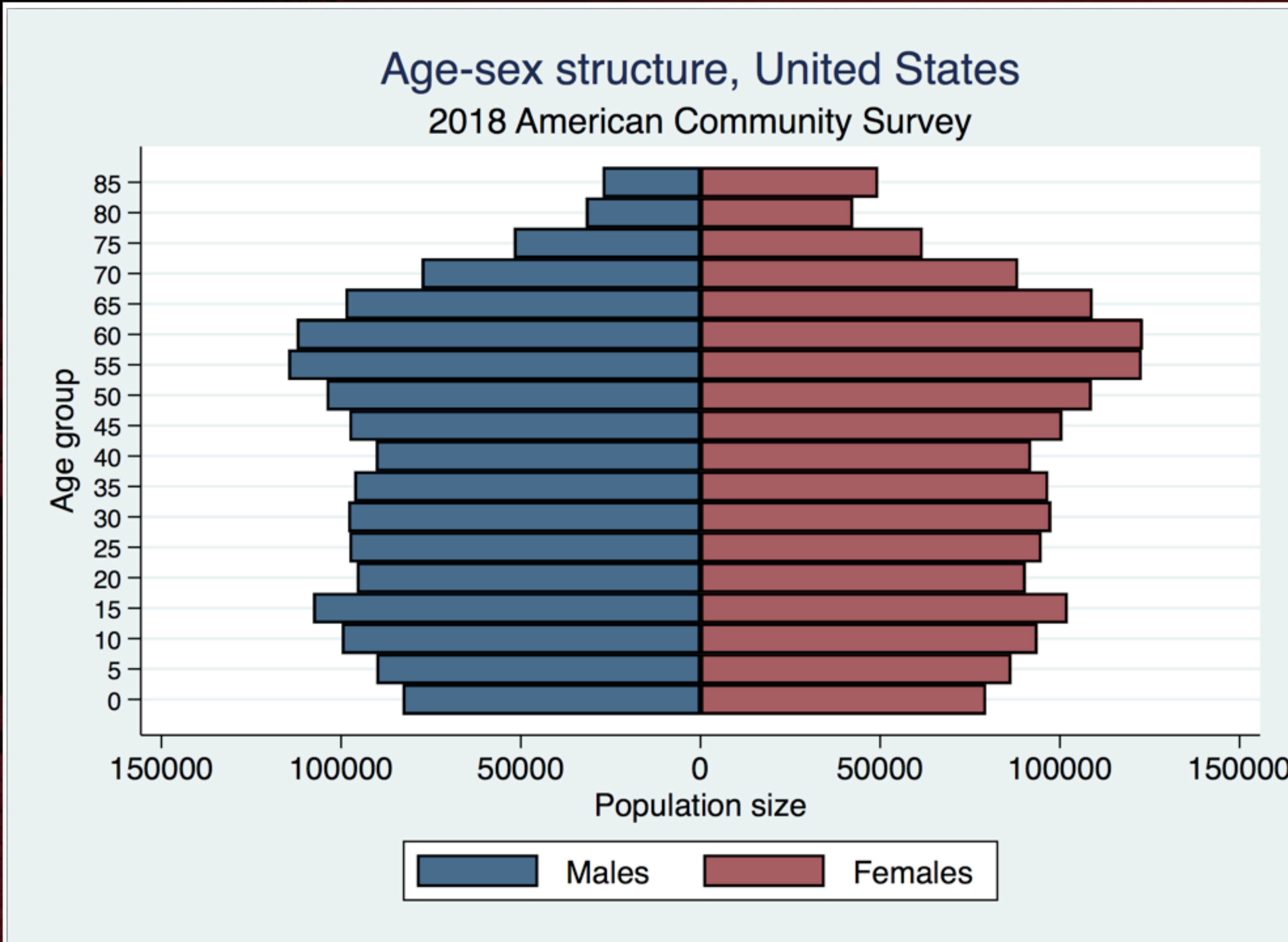
```
graph box income if income!=0 [fweight=perwt],
over(educgr) ytitle(Wage and salary income)
```

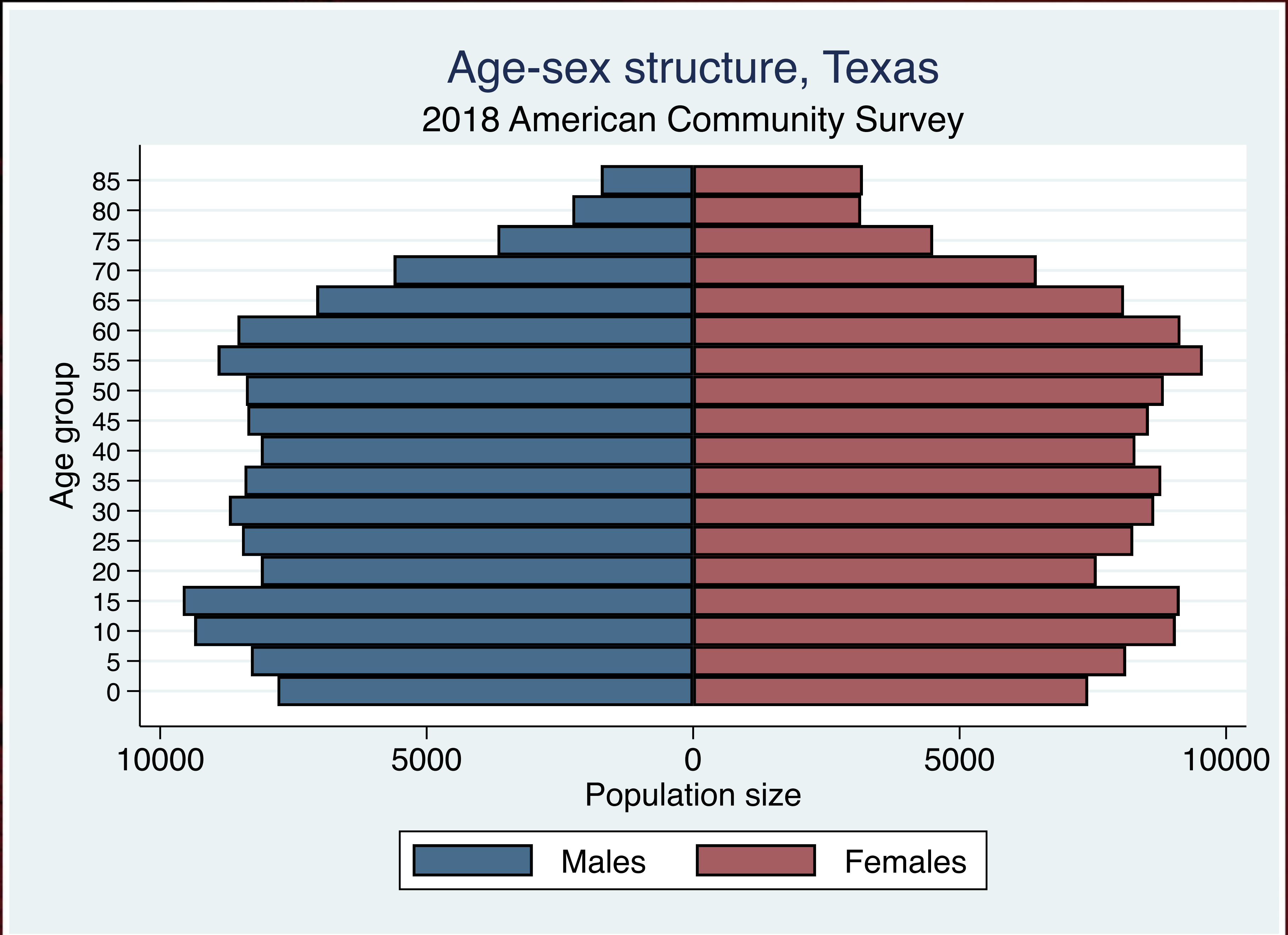


Source: 2018 American Community Survey.

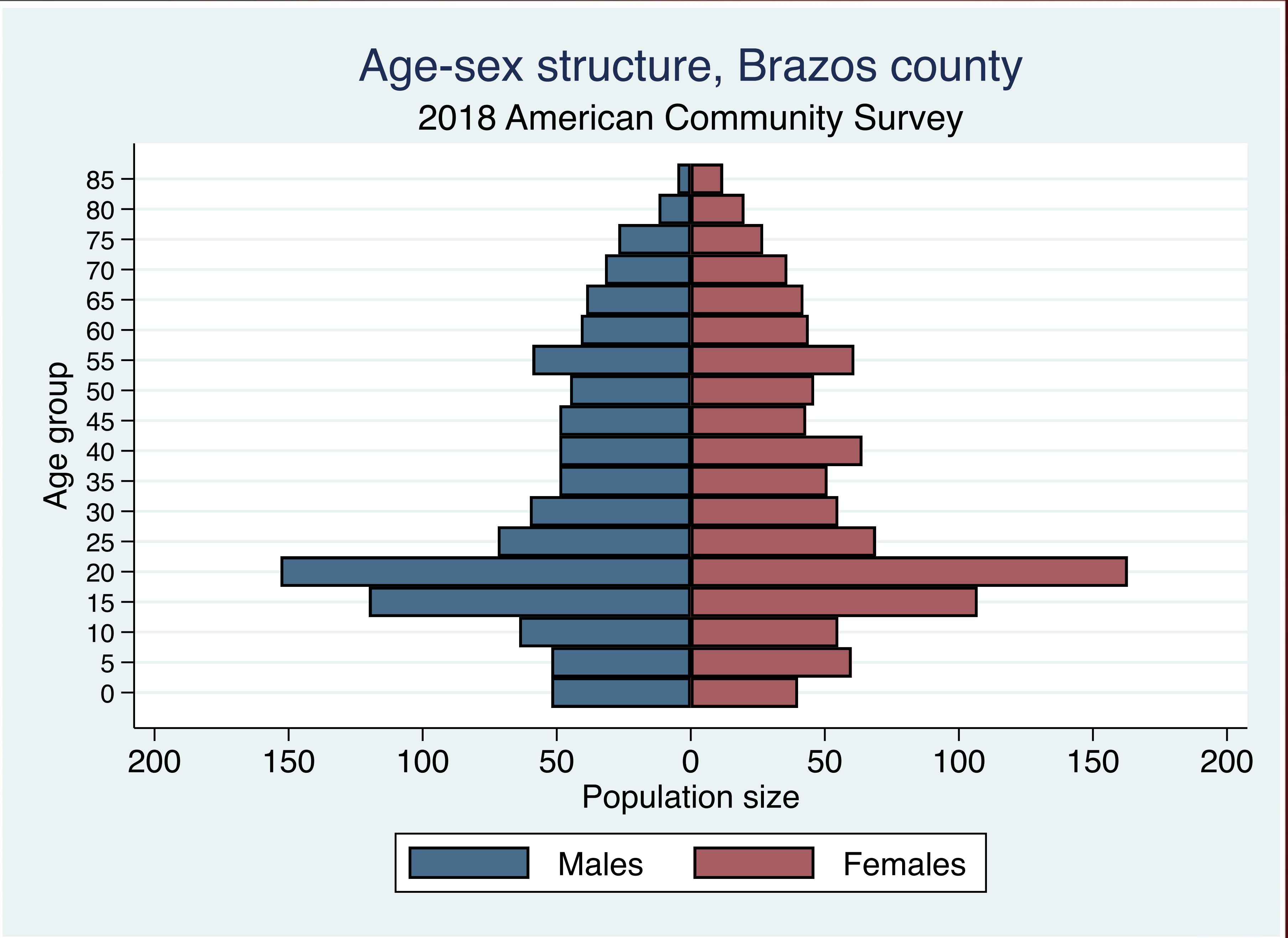
Age-sex structure

```
***Generate five-year age groups variable - automatically
egen age5y = cut(age), at(0,5,10,15,20,25,30,35,40,45,50,55,60,65,70,75,80,85,100)
table age5y, contents(min age max age count age)

***Generate male variable (opposite of female variable)
gen male=!female
tab male female, m nolabel

***Generate variables with male and female totals by five-year age groups
sort age5y
by age5y: egen maletotal=total(male)
by age5y: egen femaletotal=total(female)

***Replace male total by negative value
replace maletotal=-maletotal

***Age-sex structure
twoway bar maletotal age5y [fweight=perwt], horizontal barwidth(5) fcolor(navy) lcolor(black) lwidth(medium) || ///
       bar femaletotal age5y [fweight=perwt], horizontal barwidth(5) fcolor(maroon) lcolor(black) lwidth(medium) ///
       legend(label(1 Males) label(2 Females)) ///
       ylabel(0(5)85, angle(horizontal) valuelabel labsize(*.8)) ///
       ytitle("Age group") ///
       xlabel(-150000 "150000" -100000 "100000" -50000 "50000" 0 50000 100000 150000) ///
       xtitle("Population size") ///
       title("Age-sex structure, United States") ///
       subtitle("2018 American Community Survey")
```

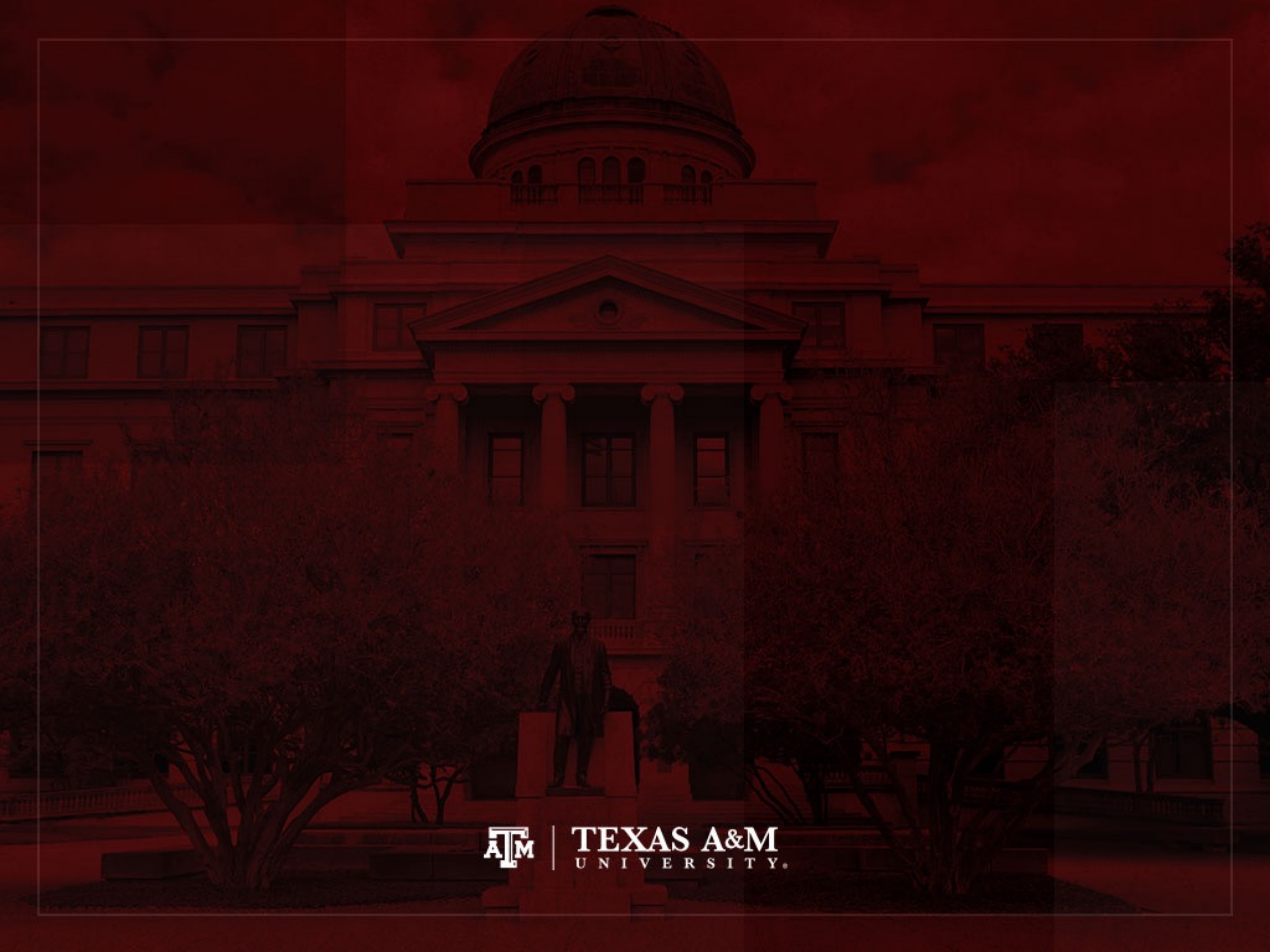

Age-sex structure, United States
2018 American Community Survey
Age group
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
150000
100000
50000
0
50000
100000
150000
Population size
Males
Females



Age-sex structure, Texas
2018 American Community Survey
Age group
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
10000
5000
0
5000
10000
Population size
Males
Females



Age-sex structure, Brazos county
2018 American Community Survey
Age group
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
200
150
100
50
0
50
100
150
200
Population size
Males
Females


TEXAS A&M
UNIVERSITY