Lecture 2a: Basic descriptive statistics

Ernesto F. L. Amaral

September 1, 2022
Introduction to Sociological Data Analysis (SOCI 600)

Source: Healey, Joseph F. 2015. ”Statistics: A Tool for Social Research.” Stamford: Cengage Learning. 10th edition. Chapter 2 (pp. 24–65).

TEXAS A&M UNIVERSITY

Outline

- Frequency distributions
- Proportions, percentages
- Person-years
- Rates, probabilities, ratios
- Percentage change
- Using graphs to present data

Frequency distributions

- Frequency distributions are tables that report the number of cases in each category of a variable
- Frequency distributions summarize distribution of a variable by reporting the number of times each score of a variable occurred
- General rule for categories of frequency distribution
  - Exhaustive
  - Mutually exclusive
    - Each case counted in one and only one category

Frequency distributions

- Useful way to examine variables
- Report the number of cases in each category
- Used with variables at any level of measurement
- For nominal-level variables
  - Count the number of times each category occurs and display the frequencies in table format

Population by gender (fictitious data)

Gender	Frequency
Males	53
Females	60
Total	113

Source: Healey 2015, p.25.

Number of categories

- Greater detail: more categories
- More clarity: fewer categories

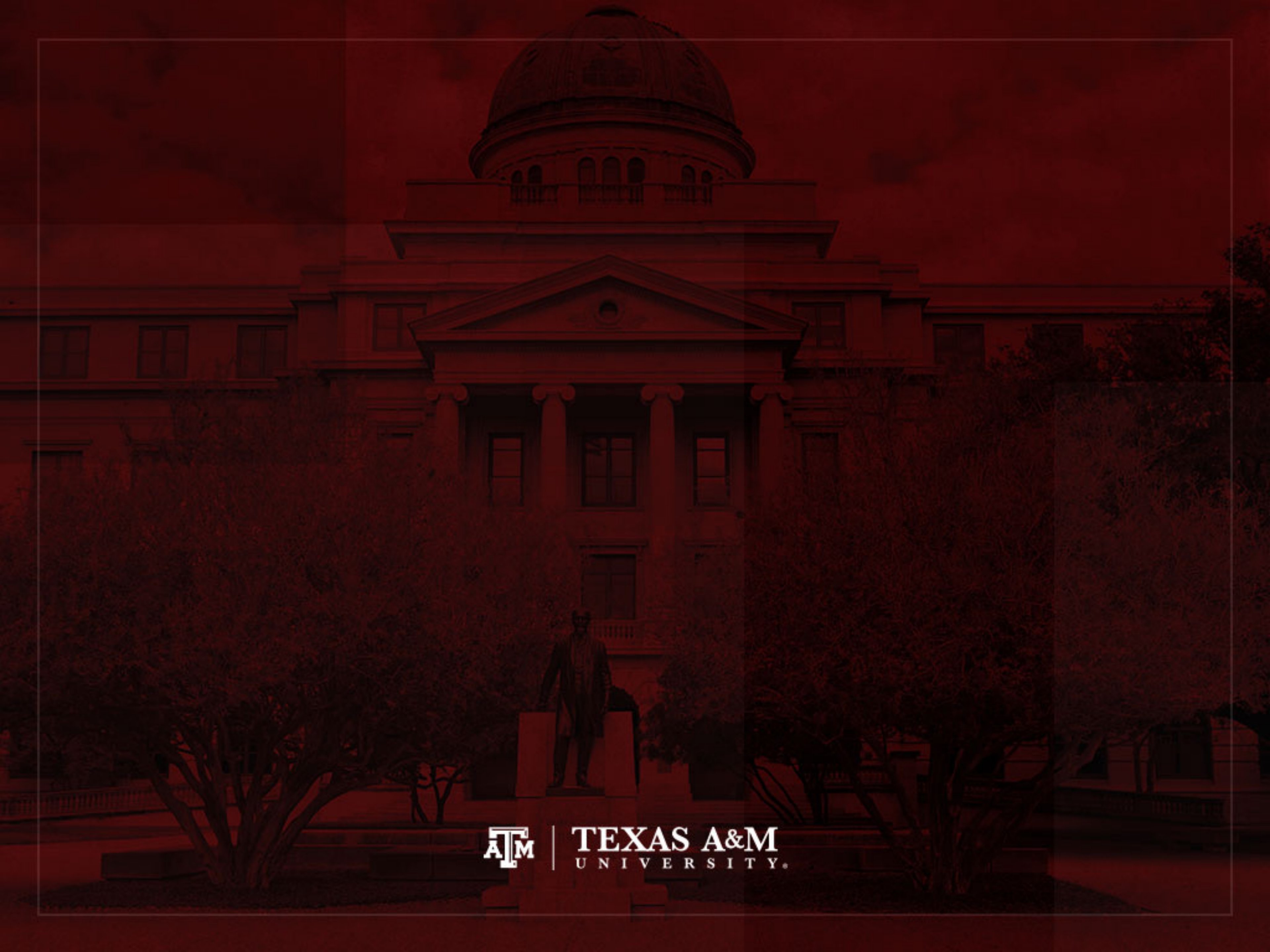
Self-described religious identifications of adult Americans, 2008

Religious group	Frequency
Protestant	116,203,000
Catholic	57,199,000
Jewish	2,680,000
Muslim	1,349,000
Buddhist	1,189,000
Unitarian	586,000
Other	2,992,000
None	34,169,000
Total	216,367,000

Source: Healey 2015, p.26.

Religious group	Frequency
Protestant	116,203,000
Catholic	57,199,000
Jewish	2,680,000
Other	6,116,000
None	34,169,000
Total	216,367,000

Source: Healey 2015, p.26.


TEXAS A&M
UNIVERSITY

Proportions and percentages

- Report relative size
  - Compare the number of cases in a specific category to the number of cases in all categories
  - The part (specific category) is the numerator (*f*)
  - The whole (all categories) is the denominator (*N*)
- What percentage of a group of people is female?
  - The whole is the number of people in the group
  - The part is the number of females

Formulas

$$Proportion = \left(\frac{f}{N}\right)$$

$$Percentage\text{: } \% = \left(\frac{f}{N}\right) \times 100$$

where *f* = frequency or the number of cases in any category

N = the number of cases in all categories

Guidelines

- With small number of cases (less than 20), report actual frequencies
- Always report number of observations along with proportions and percentages
- We can calculate percentages and proportions for variables at all levels of measurement

Nominal-level: Religion

Self-described religious identifications of adult Americans, 2008

Religious group	Frequency	Percentage
Protestant	116,203,000	53.71%
Catholic	57,199,000	26.44%
Jewish	2,680,000	1.24%
Muslim	1,349,000	0.62%
Buddhist	1,189,000	0.55%
Unitarian	586,000	0.27%
Other	2,992,000	1.38%
None	34,169,000	15.79%
Total	216,367,000	100.00%

Source: Healey 2015, p.27.

Nominal-level: College major

Declared major fields on two college campuses (fictitious data)

Major	College A	College B
Business	103	3,120
Natural sciences	82	2,799
Social sciences	137	1,884
Humanities	93	2,176
Total	415	9,979

Declared major fields on two college campuses (fictitious data)

Major	College A	College B
Business	24.82%	31.27%
Natural sciences	19.76%	28.05%
Social sciences	33.01%	18.88%
Humanities	22.41%	21.81%
Total	100.00% (415)	100.01% (9,979)

Source: Healey 2015, p.27.

Ordinal-level: Birth control

Do you strongly agree, agree, disagree, or strongly disagree that the University Health Center should provide condoms and other ”safe sex” items on demand and at no additional cost to students?

Response	Frequency	Percentage
Strongly agree	350	25.55%
Agree	462	33.72%
Disagree	348	25.40%
Strongly disagree	210	15.33%
Total	1,370	100.00%

Aggregating categories...

Response	Frequency	Percentage
Strongly agree or Agree	812	59.27%
Disagree or Strongly disagree	558	40.73%
Total	1,370	100.00%

Source: Healey 2015, p.30–31.

Interval-ratio-level variables

- Frequency distributions for interval-ratio-level variables is more complex than for nominal and ordinal variables
- Large number of scores
- Requires collapsing or grouping of categories
- Decide the number of categories and the width of those categories
- **Class intervals** refer to the categories used in the frequency distribution

Interval-ratio-level: Age

Age of students in a college class (fictitious data)

Interval width = 1 year of age	
Ages	**Frequency**
18	5
19	6
20	3
21	2
22	1
23	1
24	1
25	0
26	1
Total	20

Source: Healey 2015, p.32.

Interval-ratio-level: Stated limits

- **Stated class limits** are separated by a distance of one unit

Age of students in a college class (fictitious data)

Interval width = 2 years of age		
Ages	**Frequency**	**Percentage**
18–19	11	55.0%
20–21	5	25.0%
22–23	2	10.0%
24–25	1	5.0%
26–27	1	5.0%
Total	20	100.0%

Source: Healey 2015, p.32.

Interval-ratio-level: Midpoints

- **Midpoints** are exactly halfway between the upper and lower limits of a class interval and can be found by dividing the sum of the upper and lower limits by 2

Class interval width = 3	
Class interval	**Midpoint**
0–2	1.0
3–5	4.0
6–8	7.0
9–11	10.0

Class interval width = 6	
Class interval	**Midpoint**
100–105	102.5
106–111	108.5
112–117	114.5
118–123	120.5

Source: Healey 2015, p.33.

Interval-ratio-level: Real limits

- **Real class limits** treat the variable as continuous

Stated limits	Real limits
18–19	17.5|–19.5
20–21	19.5|–21.5
22–23	21.5|–23.5
24–25	23.5|–25.5
26–27	25.5|–27.5

Source: Healey 2015, p.34.

Class intervals (stated limits)	Real class limits
3–5	3.0–5.9
6–8	6.0–8.9
9–11	9.0–11.9

Class intervals (stated limits)	Real class limits
100–105	99.5–105.5
106–111	105.5–111.5
112–117	111.5–117.5
118–123	117.5–123.5

Source: Healey 2015, p.35.

Cumulative frequency and cumulative percentage

- These columns inform how many cases fall below a given score or class interval

Age of students in a college class (fictitious data)

Age	Frequency	Cumulative frequency	Percentage	Cumulative percentage
18–19	11	11	55.0%	55.0%
20–21	5	16	25.0%	80.0%
22–23	2	18	10.0%	90.0%
24–25	1	19	5.0%	95.0%
26–27	1	20	5.0%	100.0%
Total	20		100.0%	

Source: Healey 2015, p.36.

Unequal class intervals

- **Open-ended interval** is an alternative to handle a few very high (or low) scores

Age of students in a college class (fictitious data)

Age	Frequency	Cumulative frequency
18–19	11	11
20–21	5	16
22–23	2	18
24–25	1	19
26–27	1	20
28 and older	1	21
Total	21	

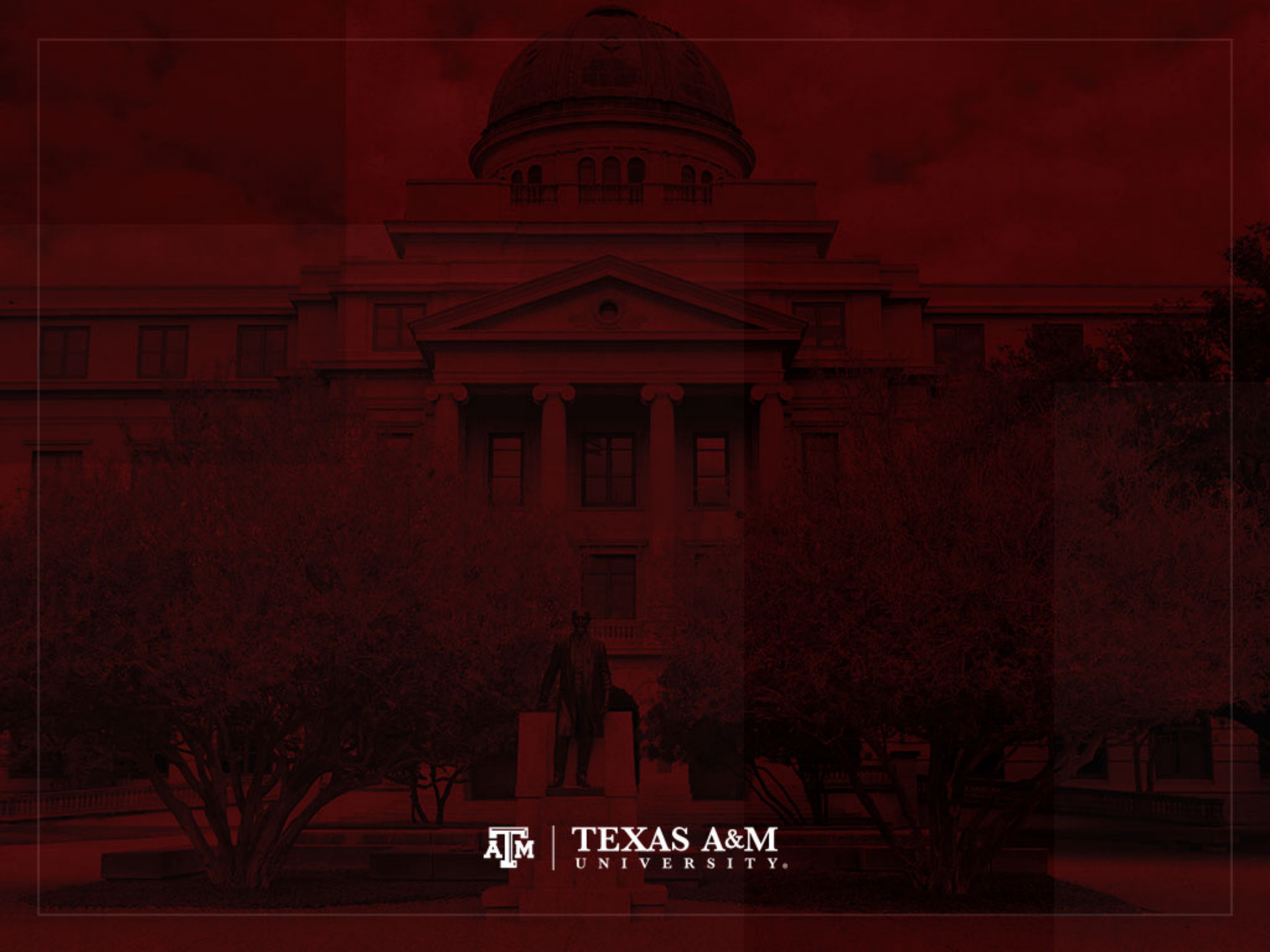
Source: Healey 2015, p.36.

Intervals of unequal size

Distribution of income by household, United States, 2011

Income	Percentage of households	Cumulative percentage
Less than $10,000	7.8%	7.8%
$10,000 to $14,999	5.8%	13.6%
$15,000 to $24,999	11.4%	25.0%
$25,000 to $34,999	10.6%	35.6%
$35,000 to $49,999	13.9%	49.5%
$50,000 to $74,999	18.0%	67.5%
$75,000 to $99,999	11.7%	79.2%
$100,000 to $149,999	12.1%	91.3%
$150,000 to $ 199,999	4.4%	95.7%
$200,000 and above	4.3%	100.0%
Total	100.0% (*N* = 114,991,720)	

Source: Healey 2015, p.37.


TEXAS A&M
UNIVERSITY

Person-years

- **Person-years** is the sum of each individual's time at risk of experiencing an event (e.g. birth, death, migration)
  - For those who do not experience event, person-years is the sum of time until end of period
  - For those who experience event, it is the time until the event

- **Period person-years lived** (PPYL) take into account that people are present during part of the period (fraction of years)
  - Each full year that a person is present in a period, he/she contributes one "person-year" to the total of PPYL
  - Each month a person is present in the population, he/she contributes 1 person-month, or 1/12 person-year, to PPYL

Example of person-years

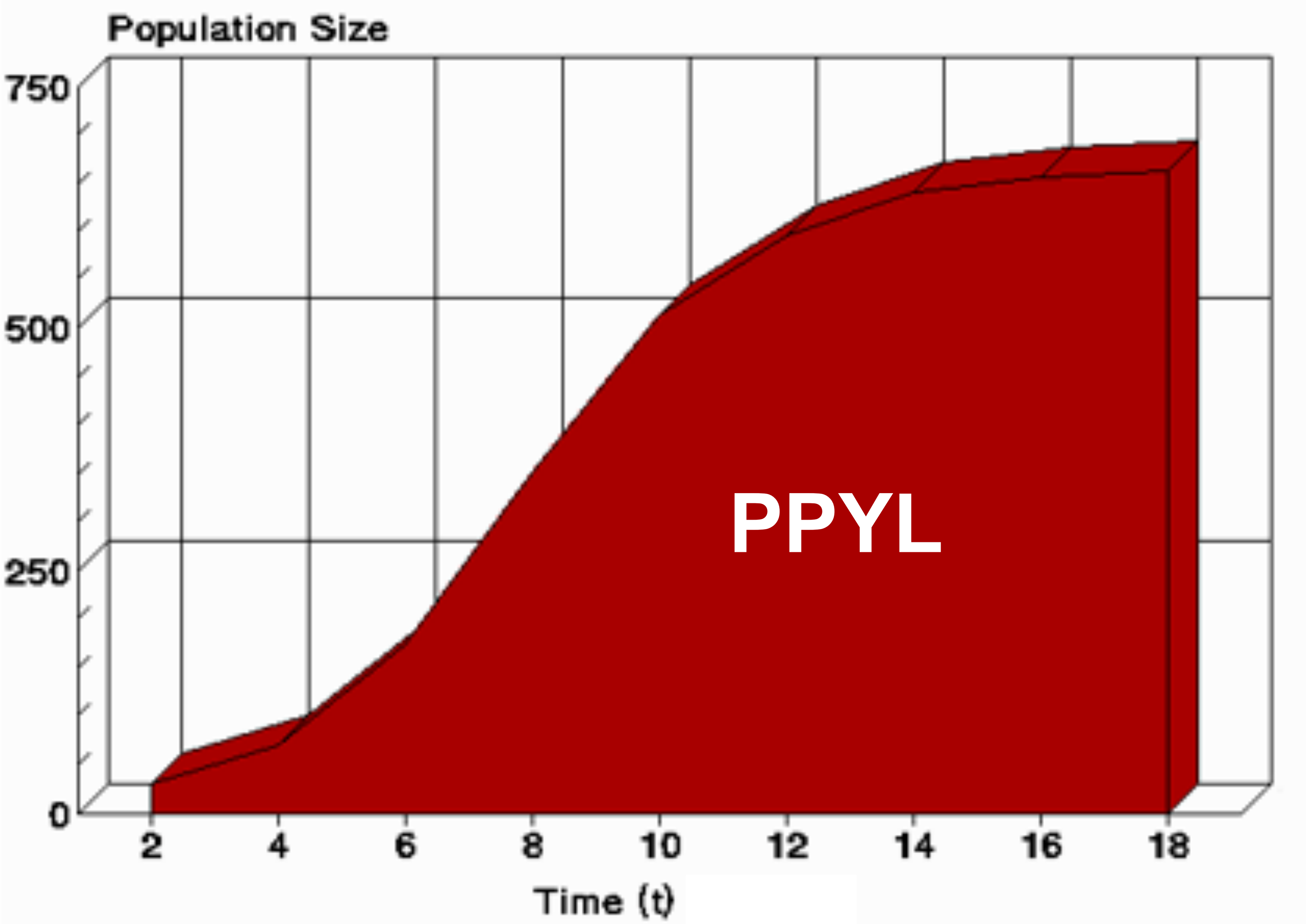
Hypothetical population increasing at the rate of 0.001 per month

Month	Population	Person-years (population / 12)	Approximation for person-years: Mid-period	Approximation for person-years: Average of start and end
January	200.00	16.67		200.00
February	200.20	16.68		
March	200.40	16.70		
April	200.60	16.72		
May	200.80	16.73		
June	201.00	16.75		
July	201.20	16.77	201.20	
August	201.40	16.78		
September	201.61	16.80		
October	201.81	16.82		
November	202.01	16.83		
December	202.21	16.85		202.21
Period person-years lived (PPYL)		**201.10**	**201.20**	**201.11**

Calculating person-years

- Whenever we know the population sizes on each month over the period of a year
  - We can add up the person-years month by month
  - Take the number of people present on each month and divide by 12
  - Add up all monthly contributions
- When our subintervals are small enough
  - Our sum is virtually equal to the area under the curve of population as a function of time during the period...

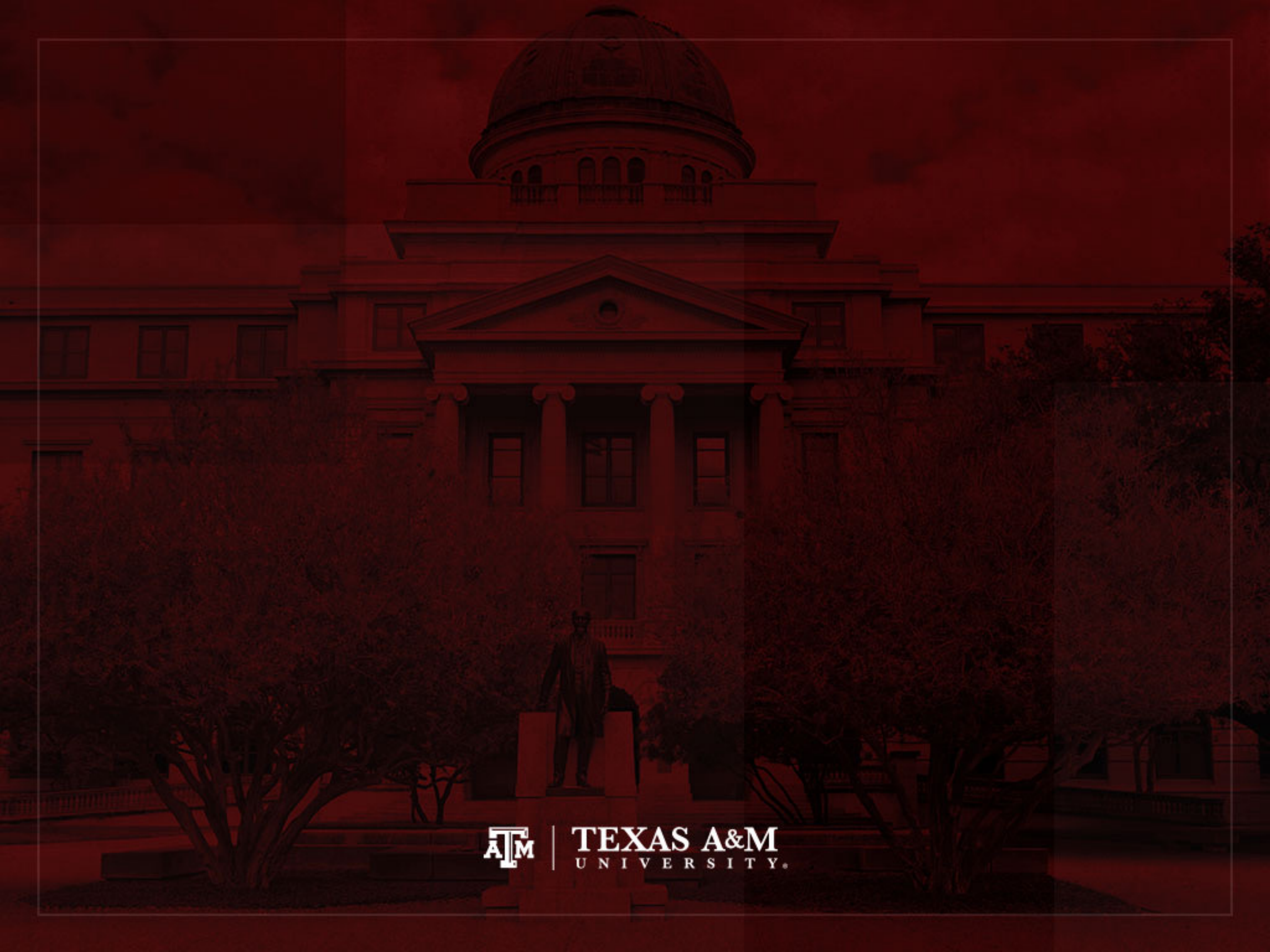
Person-years and areas



Source: https://www2.palomar.edu/users/warmstrong/lmexer9.htm.

Approximation for PPYL

- When sequences of population sizes throughout a period are unknown
  - Take the population in the middle of the period and multiply by the length of the period
    - E.g., for 2005–2015, we take the mid-period count of 308,745,000 people in the U.S. from the 2010 Census and multiply by 10 years to obtain 3,087,450,000 person-years in the period
  - Or take the average of the starting and ending populations and multiply by the length of the period


TEXAS A&M
UNIVERSITY

Rates, probabilities, ratios

- Rates
  - Describe the number of occurrences of an event for a given number of individuals who had the chance to experience that event per unit of time
- Probabilities
  - Divides the number of events by the total number of people at risk in the relevant time frame
- Ratios
  - Compare the size of one group to the size of another group

Rates

(Fleurence, Hollenbeak 2007)

- Rates are an instantaneous measure that range from zero to infinity
  - Rates describe the number of occurrences of an event for a given number of individuals per unit of time
  - Rates consider the time spent at risk
- Numerator
  - Number of events (e.g. births, deaths, migrations)
- Denominator includes time
  - Sum of each individual's time at risk of experiencing an event for a specific population during a certain time period (person-years)
  - We can use approximations for the denominator
    - Population in the middle of the period or
    - Average of starting and ending populations for that period

Examples of rates

- Express the number of actual occurrences of an event (e.g. births, deaths, homicides) vs. number of possible occurrences per some unit of time
- Examples

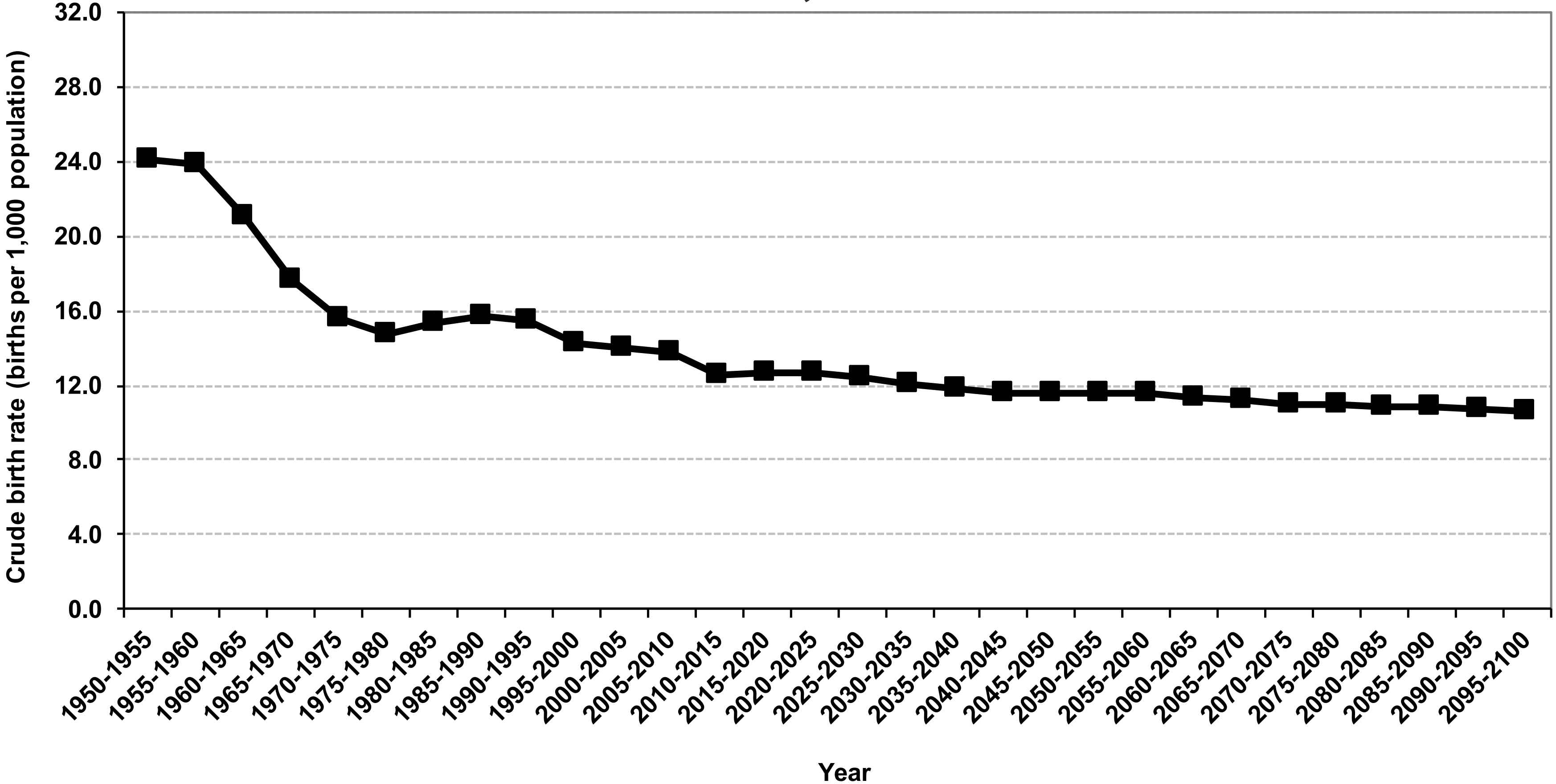
$$Crude\ birth\ rate = \frac{Number\ of\ births}{Total\ population} \times 1000$$

$$Crude\ death\ rate = \frac{Number\ of\ deaths}{Total\ population} \times 1000$$

CBR and CDR

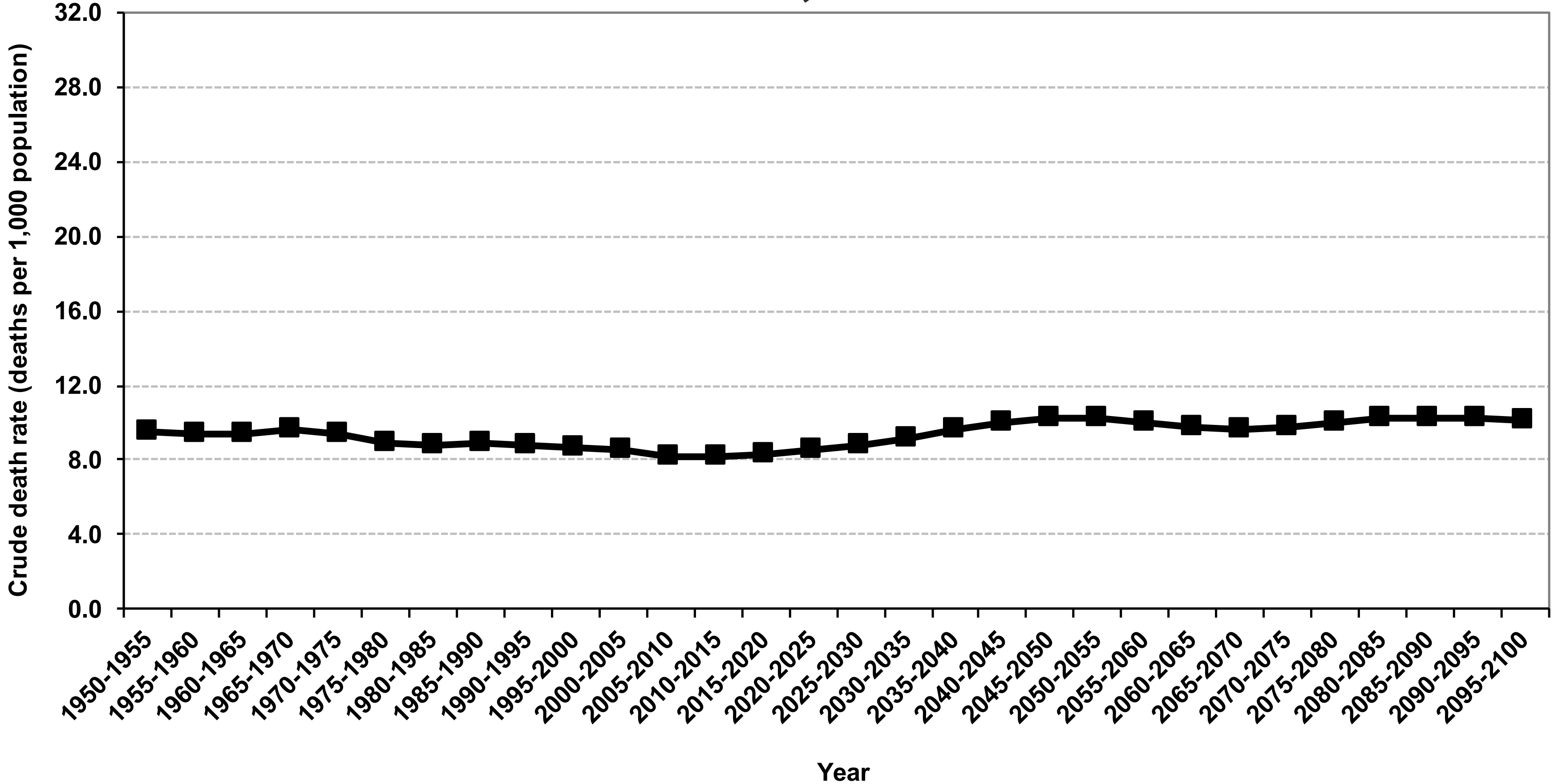
- Crude Birth Rate (CBR or *b*)
  - Number of births to members of the population in the period divided by the total period person-years lived
- Crude Death Rate (CDR or *d*)
  - Number of deaths to members of the population in the period divided by the total period person-years lived

Crude birth rates, United States, 1950–2100



Source: United Nations, World Population Prospects 2017
https://esa.un.org/unpd/wpp/Download/Standard/Population/
(medium variant).

Crude death rates, United States, 1950–2100



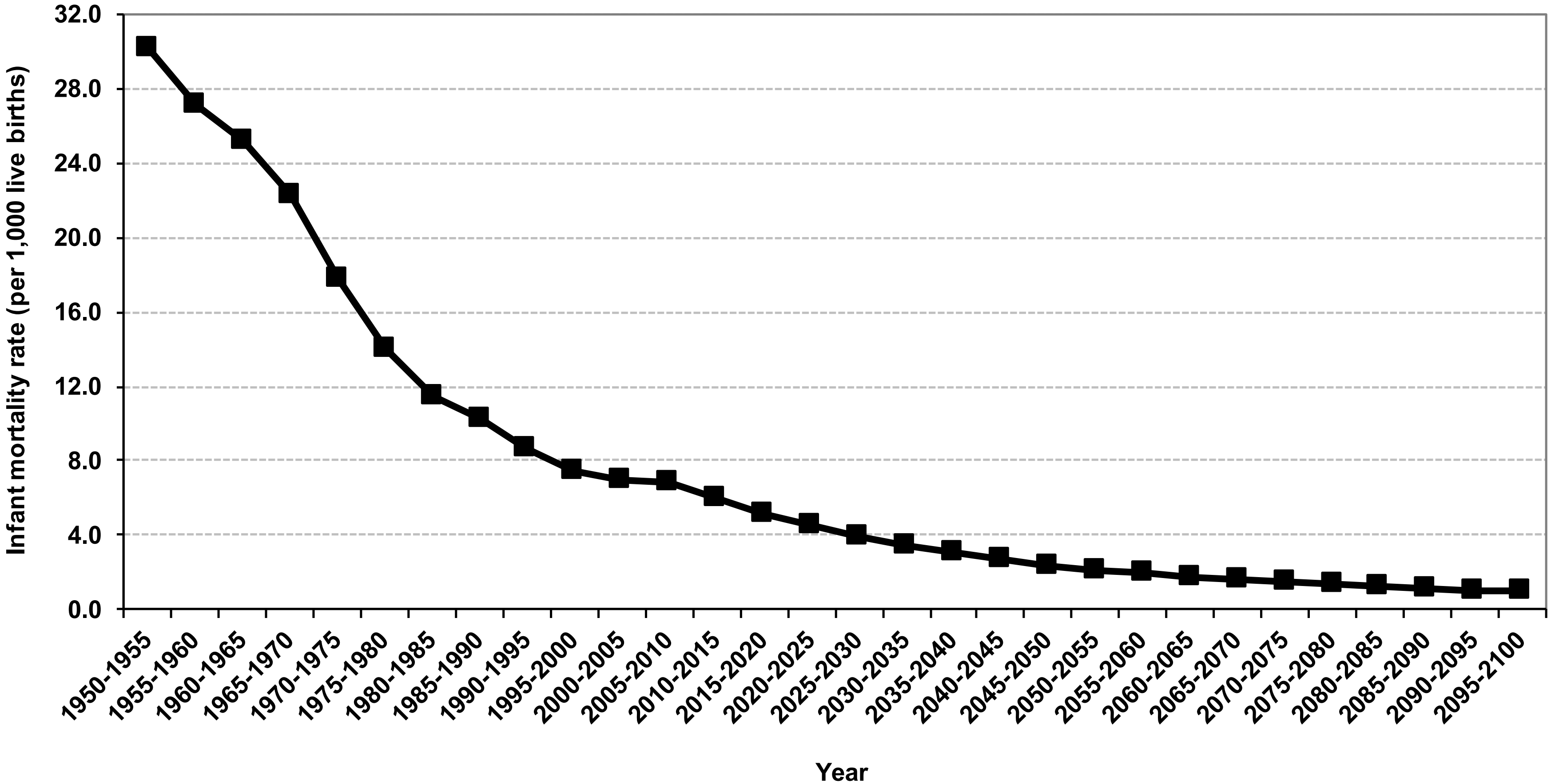
Source: United Nations, World Population Prospects 2017
https://esa.un.org/unpd/wpp/Download/Standard/Population/
(medium variant).

Infant mortality rate (IMR)

$$IMR = \frac{the\ number\ of\ deaths\ under\ age\ 1\ in\ the\ period}{the\ number\ of\ live\ births\ in\ the\ period}$$

- IMR is a period measure
- It uses current information from vital registration
- It can be computed for countries without reliable census or other source for a count of the population at risk by age
- Infants born by teenagers and by older mothers are at higher risk

Infant mortality rates, United States, 1950–2100



Source: United Nations, World Population Prospects 2017
https://esa.un.org/unpd/wpp/Download/Standard/Population/
(medium variant).

Probabilities

(Fleurence, Hollenbeak 2007)

- Probabilities describe the likelihood that an event will occur for a single individual in a given time period and range from 0 to 1
  - Does not include time in the denominator
  - Divides the number of events by the total number of people at risk in the relevant time frame
- Conversion between rates and probabilities:

$$\text{probability: } p = 1 - e^{-rt}$$

$$\text{rate: } r = -1/t * \ln(1-p)$$

- An approximation for the denominator is the population at the beginning of the period

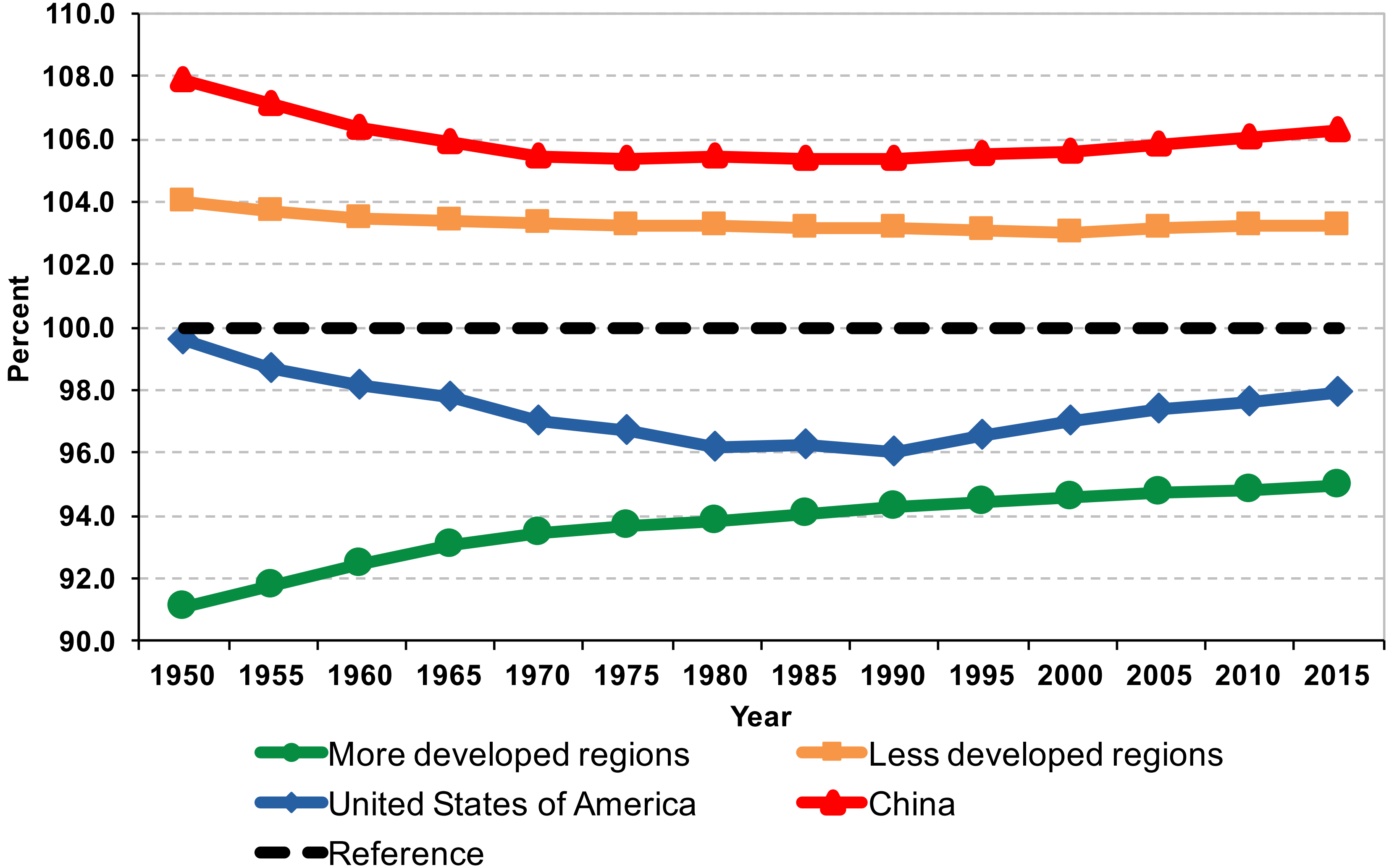
Ratios

- Describe a relationship between two numbers
  - Compare the size of one group to the size of another group
  - Compare the relative sizes of categories
  - Indicate how many times the first number contains the second
  - Denominator is not at "risk" of moving to numerator
  - Optional: multiply by 100 to get percentage

$$Sex\ ratio = \frac{Population\ of\ males}{Population\ of\ females}$$

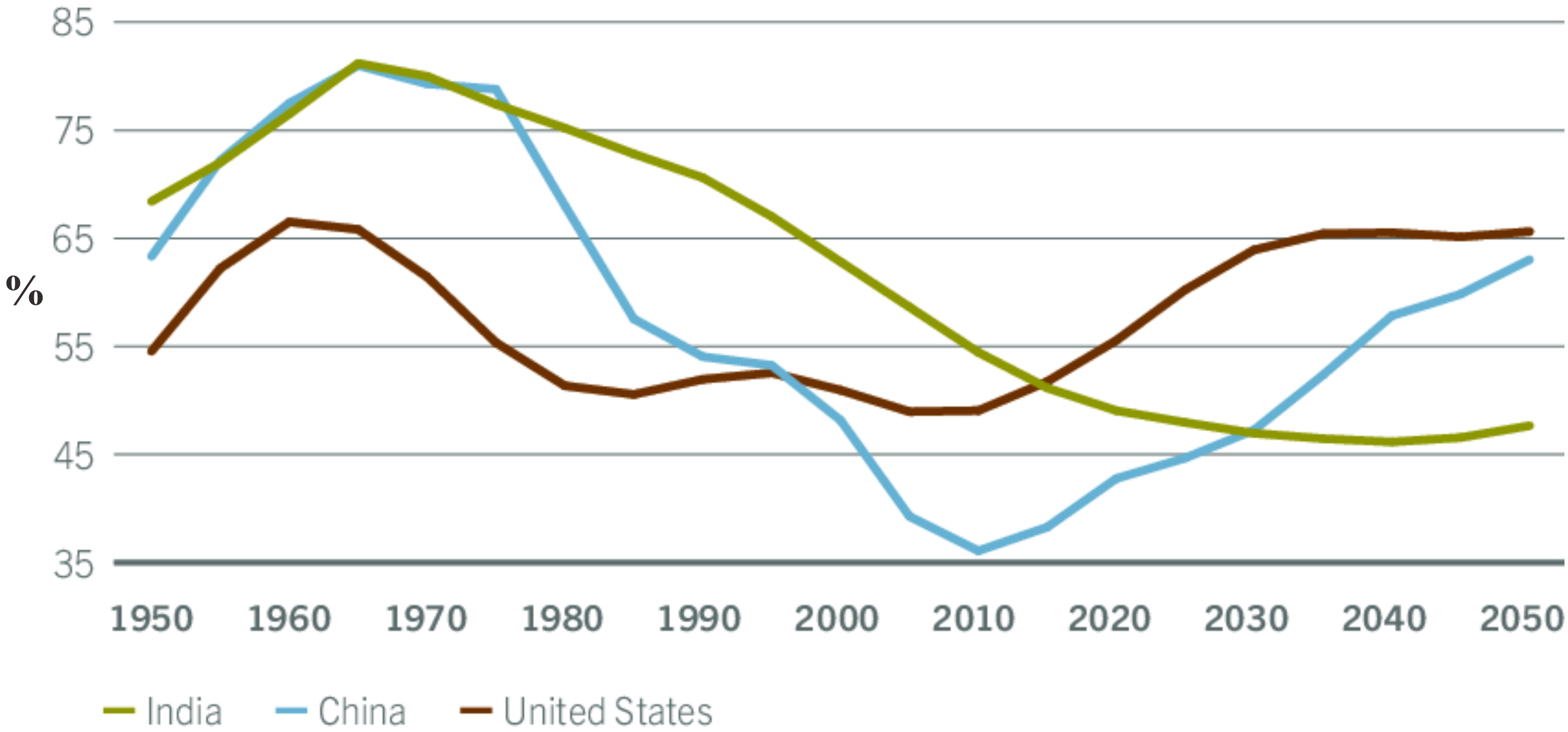
$$Total\ dependency\ ratio = \frac{Pop.\ children\ (0\ to\ 14)\ +\ Elderly\ pop.(65+)}{Working\ age\ population\ (15\ to\ 64)}$$

Sex ratios, 1950–2015



Source: United Nations, World Population Prospects 2017
https://esa.un.org/unpd/wpp/Download/Standard/Population/

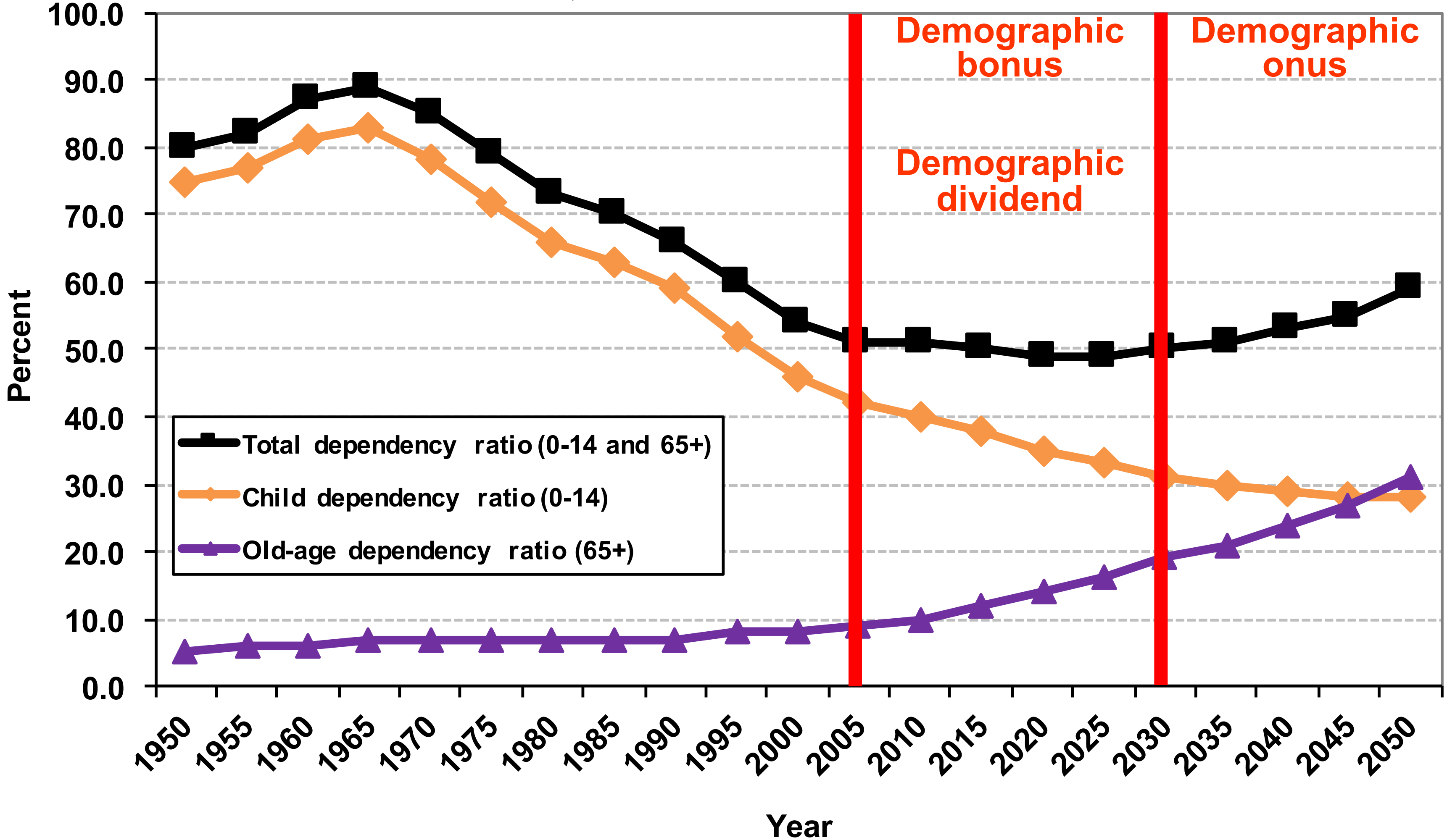
Total dependency ratios, India, China, United States



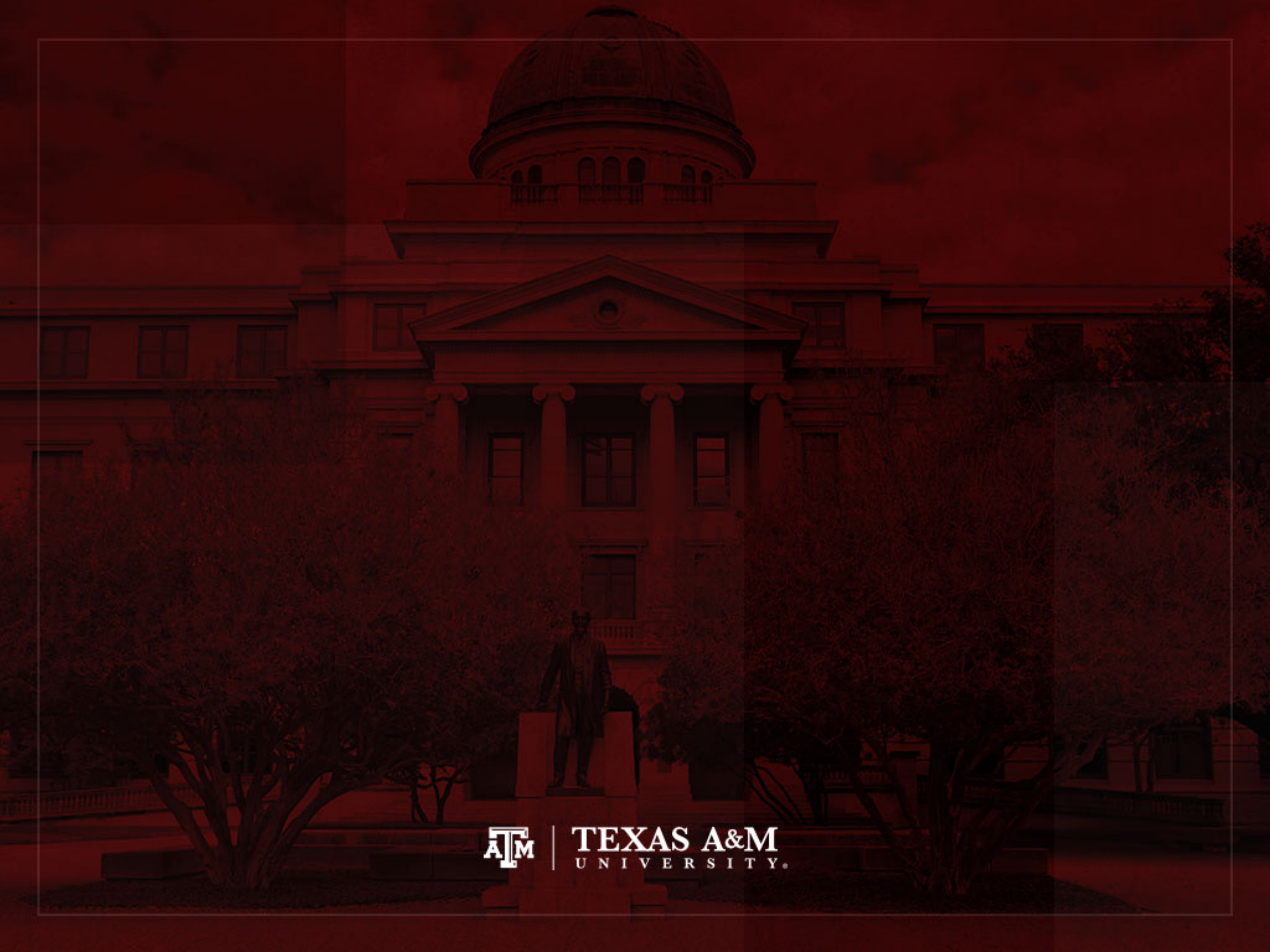
Source: United Nations Population Division

Source: https://www.researchgate.net/figure/Total-dependency-ratios-of-China-India-and-the-US-In_fig3_275071394.

Dependency ratios, Brazil, 1950–2050



Source: United Nations - http://esa.un.org/unpp (medium variant).


TEXAS A&M
UNIVERSITY

Percentage change

- Measures the relative increase or decrease in a variable over time

$$Percent\ change = \left(\frac{f_2 - f_1}{f_1}\right) \times 100$$

  - f_1 is the first (or earlier) frequency
  - f_2 is the second (or later) frequency

- Percentage change can be calculated with percentages, rates, or other values
  - If positive, it indicates an increase from time 1 to 2
  - If negative, it indicates a decrease from time 1 to 2

Example of percentage change

- In a country, the population of college graduates rose from 8% in 2000 to 13% in 2010
- By how much is the population of college graduates higher in 2010, relative to 2000?
- **Percentage point**: the population of college graduates experienced a 5 percentage point increase (13 – 8) in the period
- **Percentage change**: the population of college graduates is 62.5% higher in 2010 than in 2000

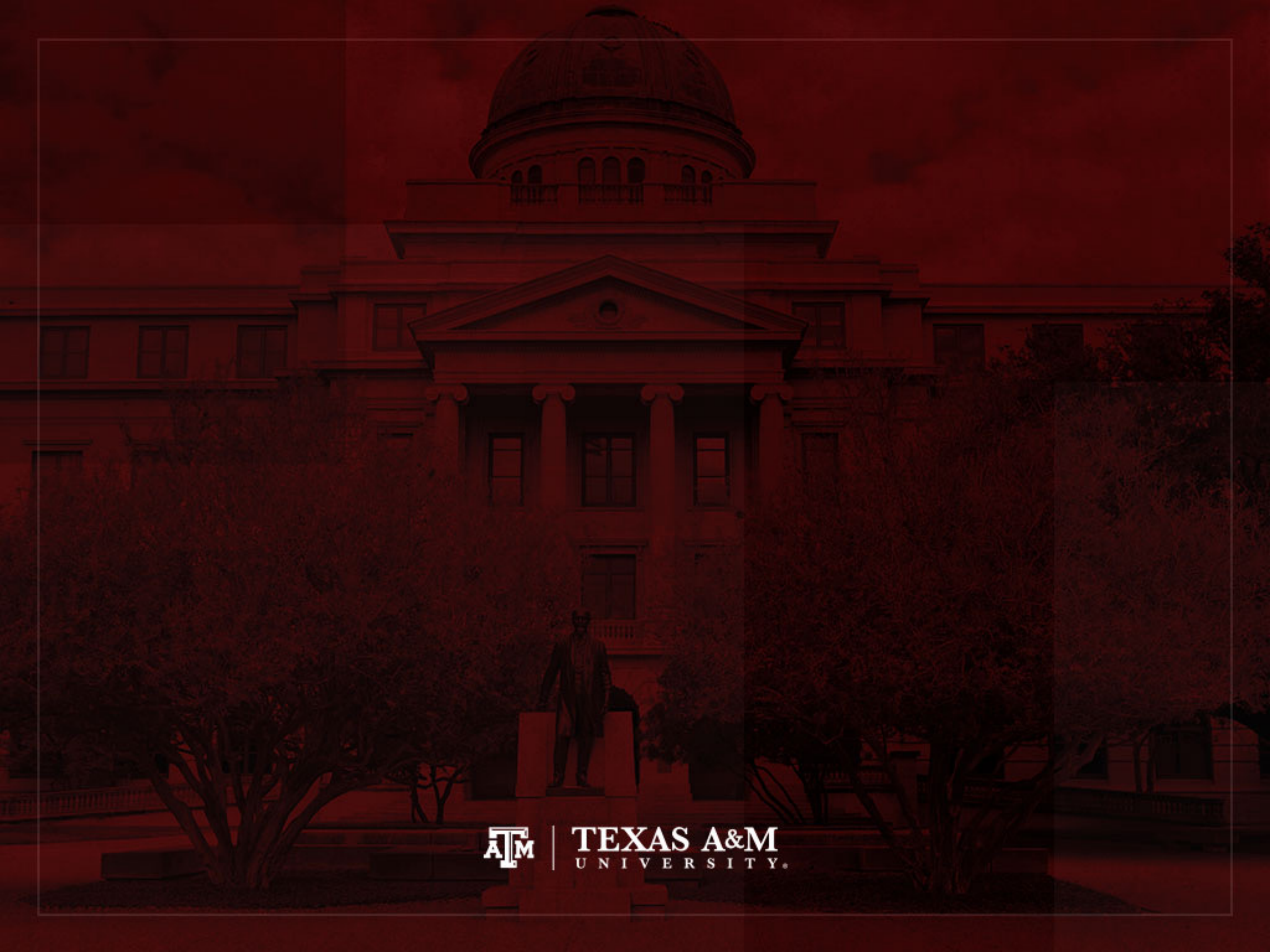
$$Percent\ change = \left(\frac{13-8}{8}\right)\times 100 = \left(\frac{5}{8}\right)\times 100 = (0.625)\times 100 = 62.5\%$$

Example of percentage change

Projected population growth for six nations, 2012–2050

Nation	Population, 2012 (f_1)	Population, 2050 (f_2)	Increase or decrease ($f_2 - f_1$)	Percentage change $(f_2 - f_1)/(f_1)*100$
China	1,350,400,000	1,310,700,000	–39,700,000	–2.96
United States	313,900,000	422,600,000	108,700,000	34.63
Nigeria	170,100,000	402,400,000	232,300,000	136.57
Mexico	116,100,000	143,900,000	27,800,000	23.94
United Kingdom	63,200,000	79,600,000	16,400,000	25.95
Canada	34,900,000	48,600,000	13,700,000	39.26

Source: Healey 2015, p.44.


TEXAS A&M
UNIVERSITY

Using graphs to present data

- Pie charts, bar graphs, histograms, and line charts present frequency distributions graphically

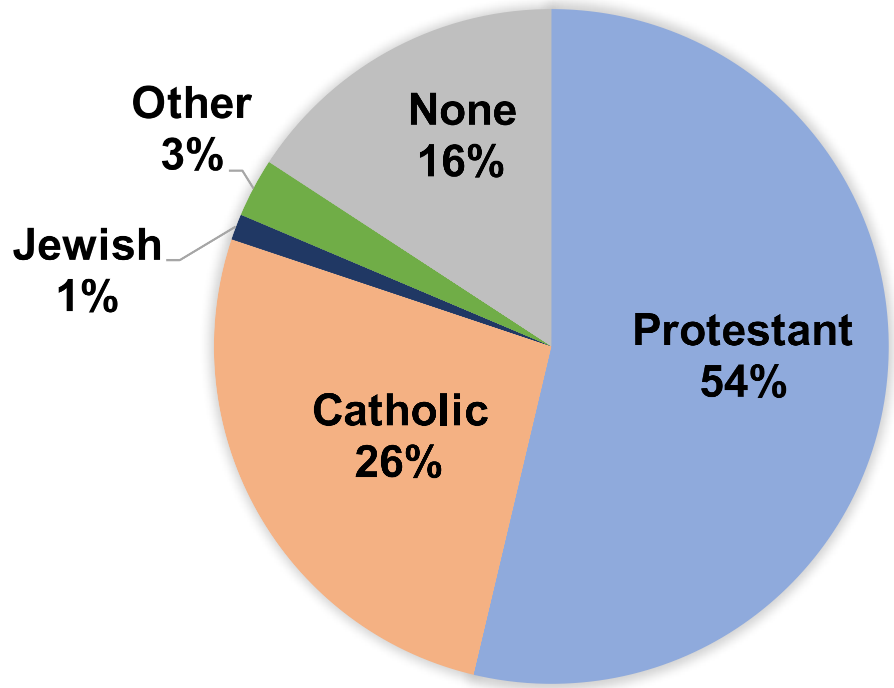
- Graphs and charts are commonly used ways of presenting “pictures” of research results

- Graphs and charts are very useful ways to display the overall shape of a distribution

Pie charts

- Pie charts are useful for discrete variables with only a few categories
- The pie is divided into segments, which are proportional in size to the percentage of cases in each category

Religious Identifications, United States, 2008

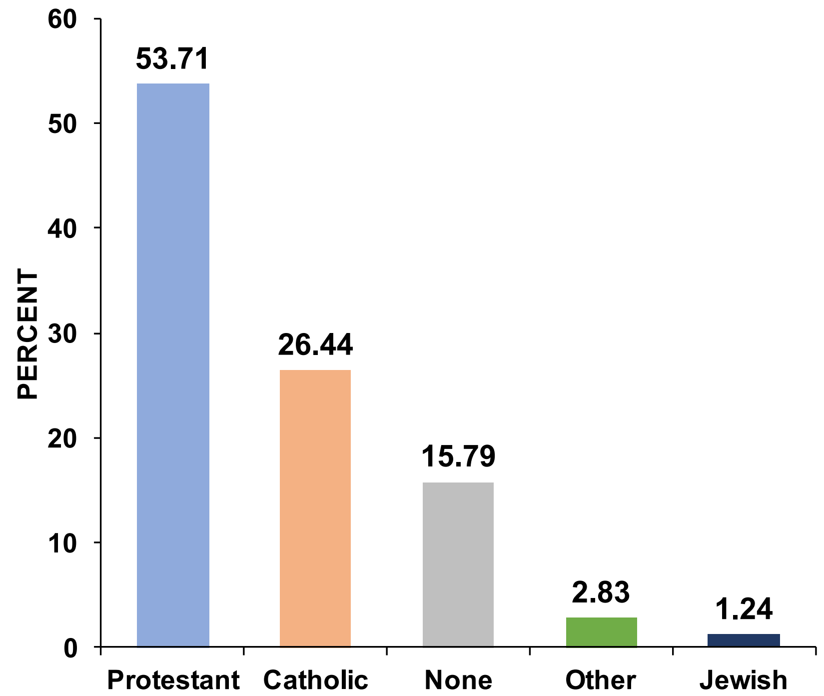


Source: Healey 2015, p.47.

Column charts

- Column charts are useful for discrete variables
- The categories are represented by columns
- The height of these columns corresponds to the number or percentage of cases in each category

Religious Identifications, United States, 2008

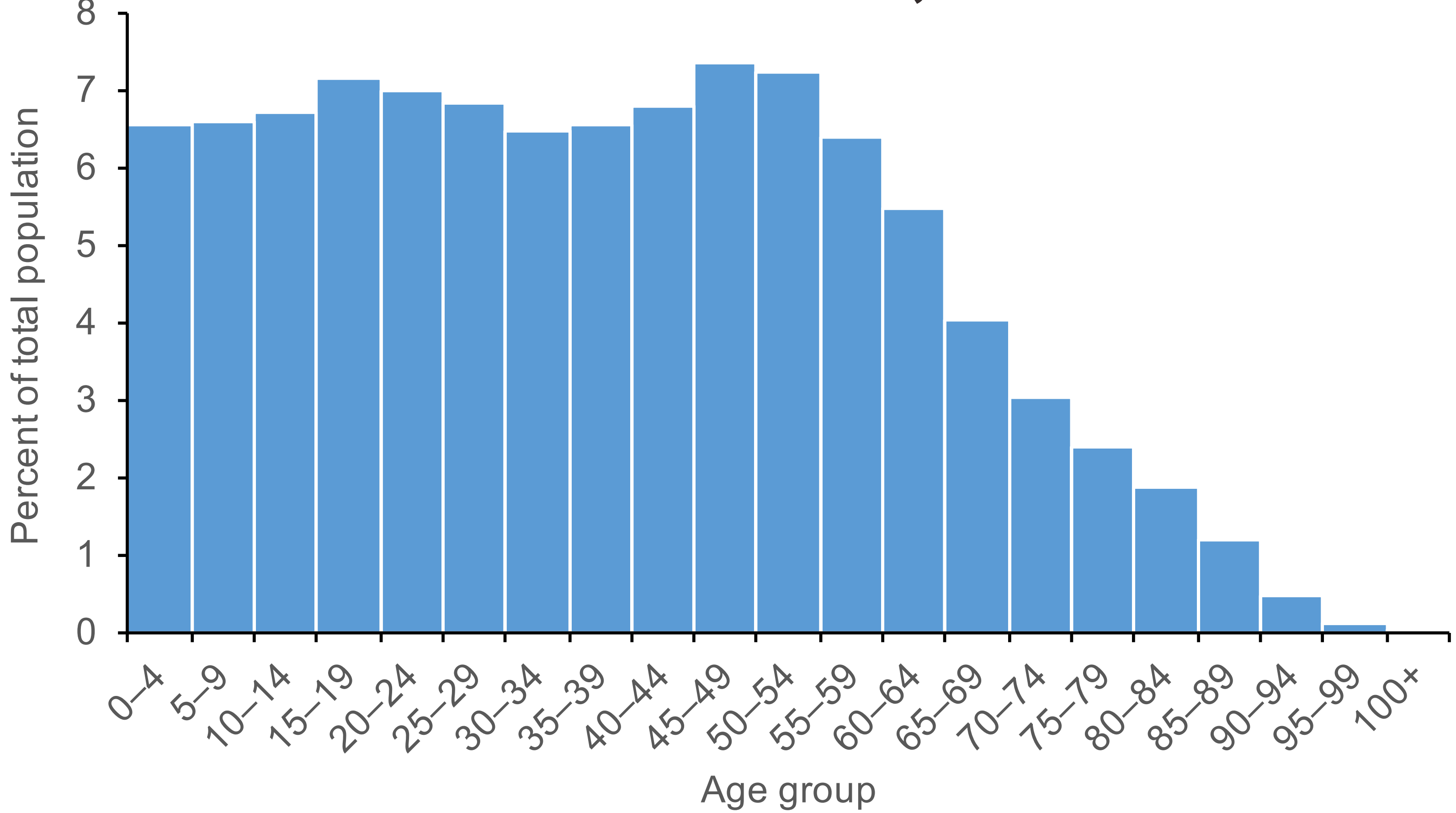


Source: Healey 2015, p.48.

Histograms

- Most appropriate for continuous interval-ratio level variables
- It can be used for discrete interval-ratio level variables
- Look like column charts
- Use real limits instead of stated limits
- Categories (or scores) of the variable border each other (the sides of the columns touch)

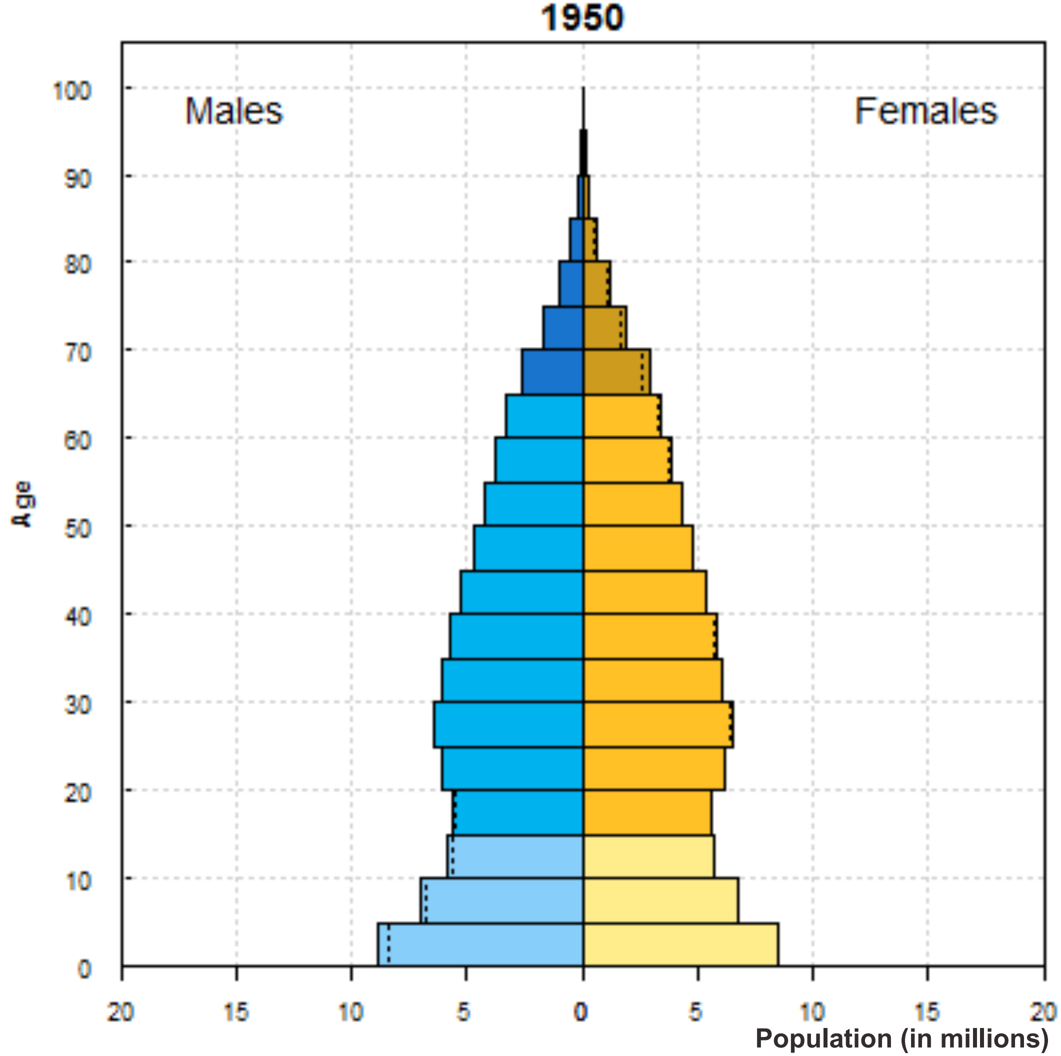
Age distribution, United States, 2010



Source: https://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf.

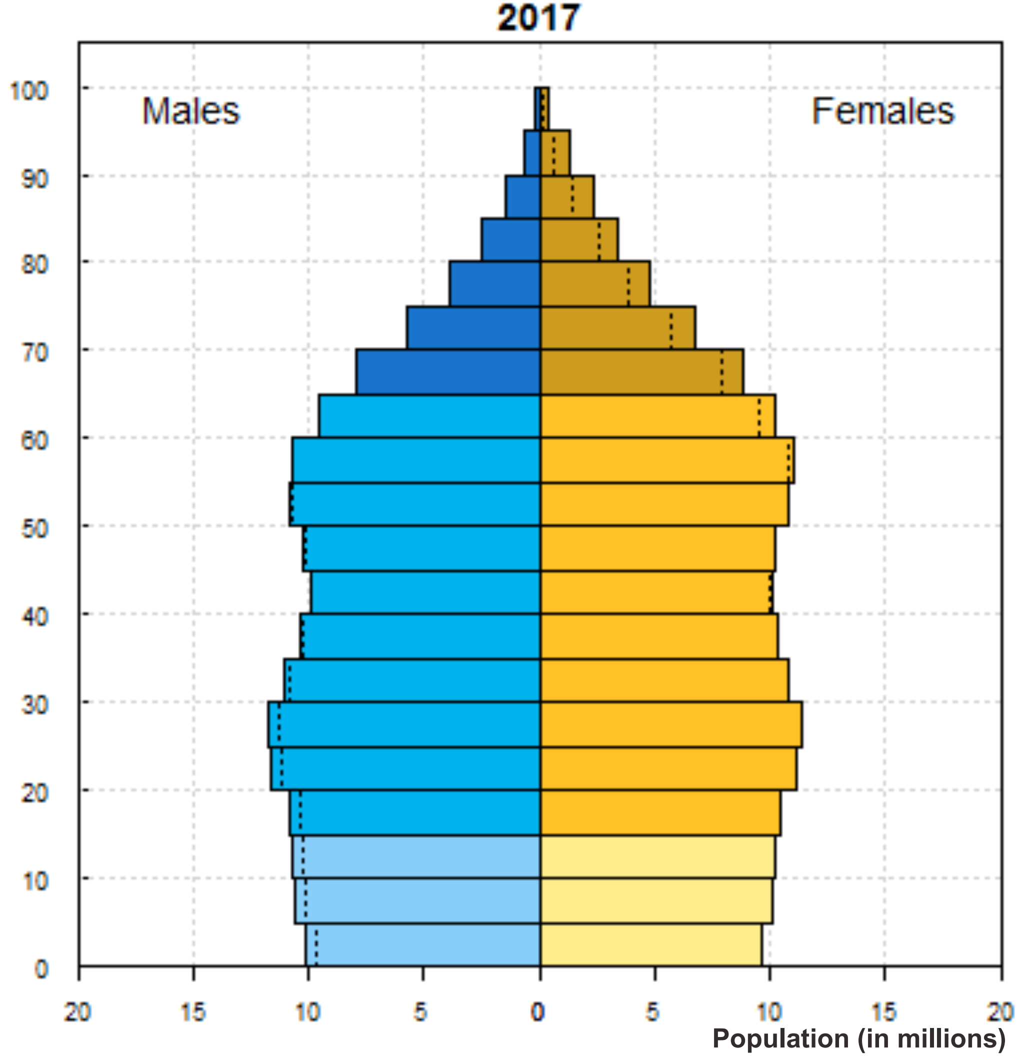
Age-sex structure, United States

Bar chart



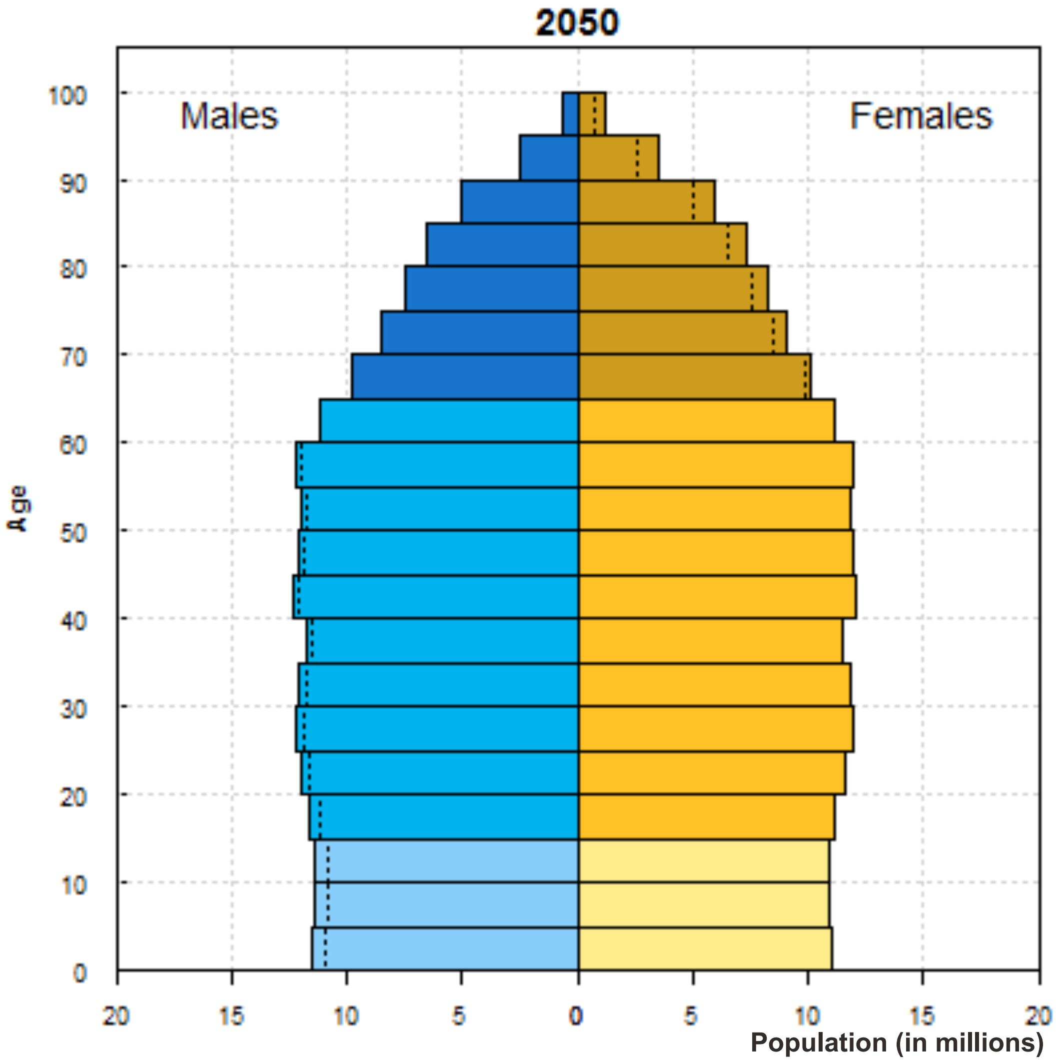
The dotted line indicates the excess male or female population in certain age groups.
Source: United Nations, World Population Prospects 2017
https://esa.un.org/unpd/wpp/Download/Standard/Population/ (medium variant).

Age-sex structure, United States



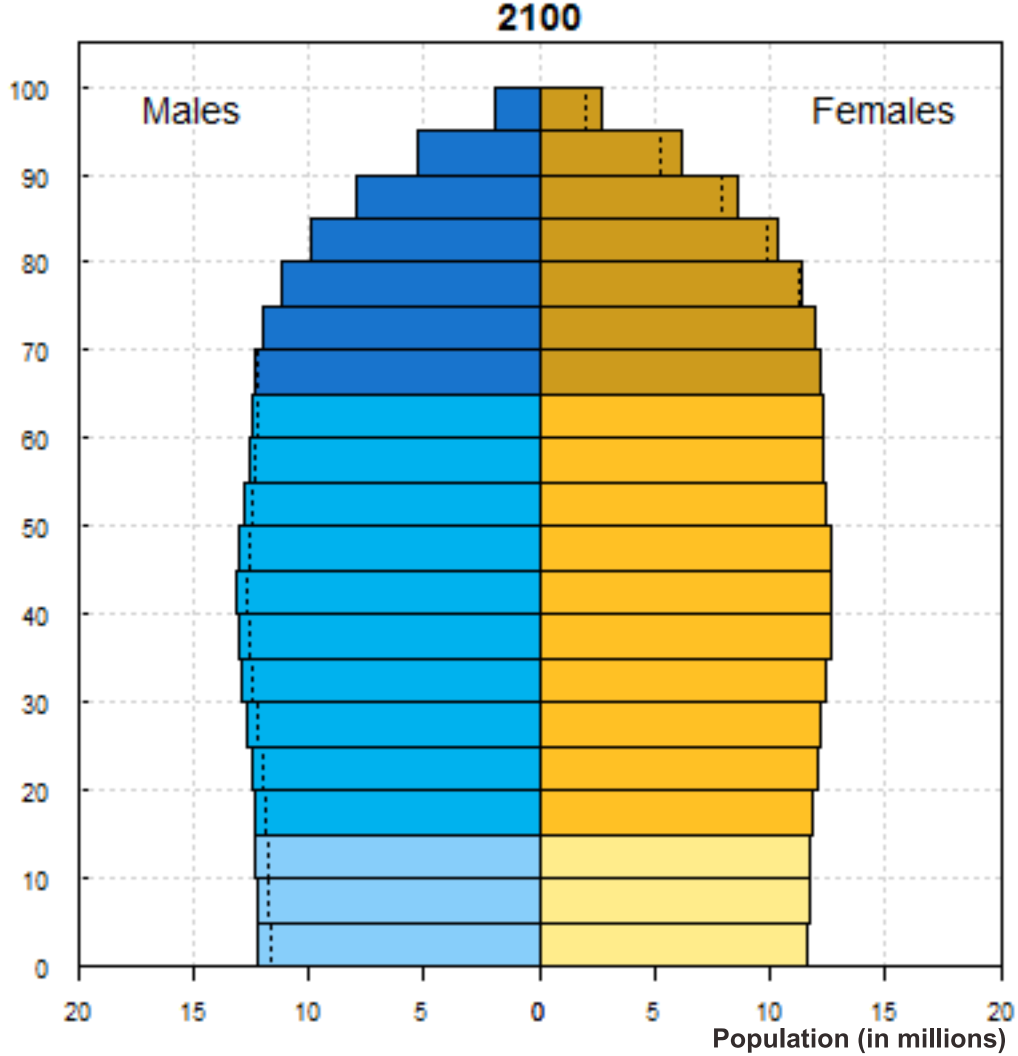
The dotted line indicates the excess male or female population in certain age groups.
Source: United Nations, World Population Prospects 2017
https://esa.un.org/unpd/wpp/Download/Standard/Population/ (medium variant).

Age-sex structure, United States



The dotted line indicates the excess male or female population in certain age groups.
Source: United Nations, World Population Prospects 2017
https://esa.un.org/unpd/wpp/Download/Standard/Population/ (medium variant).

Age-sex structure, United States

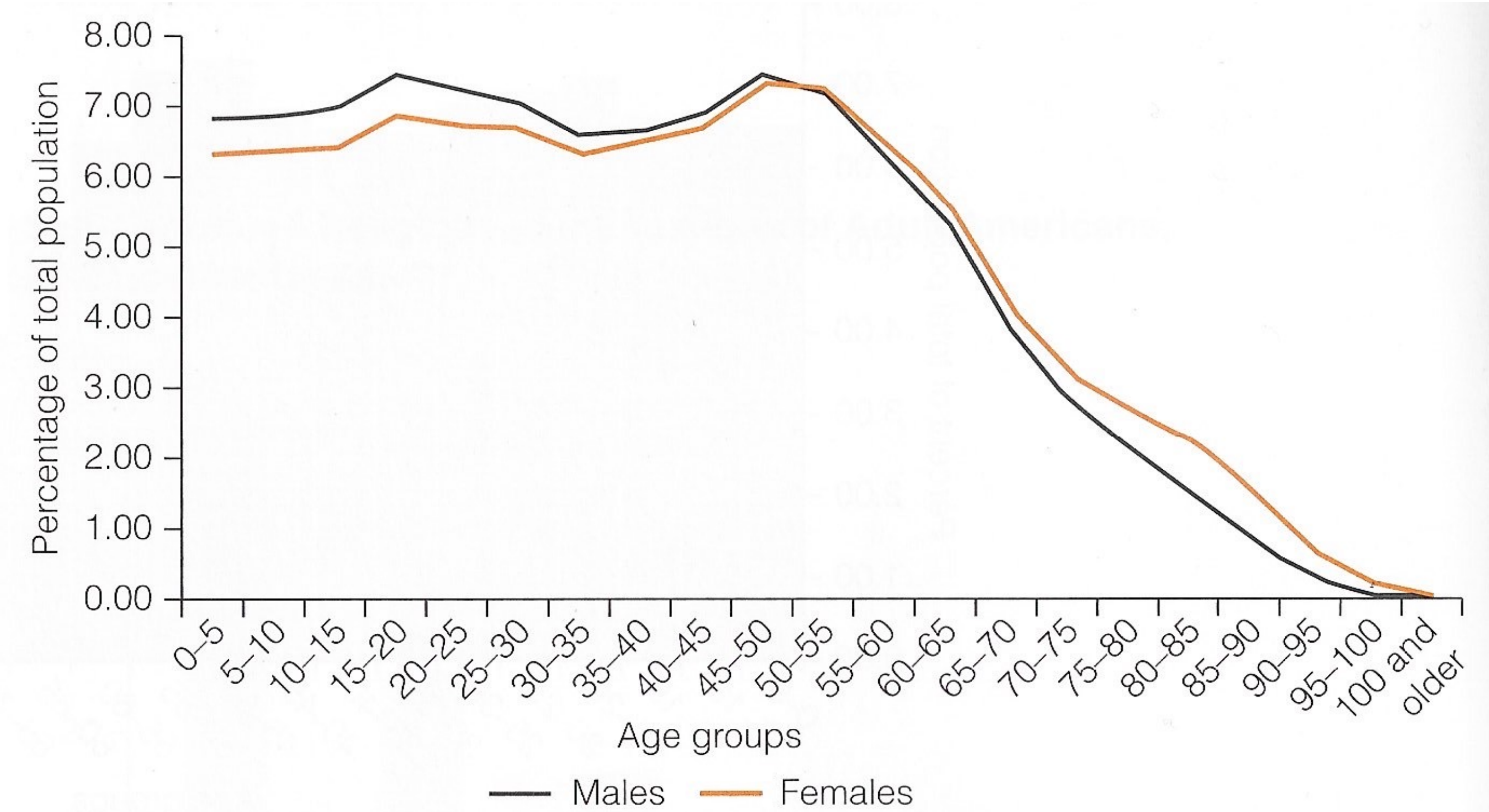


The dotted line indicates the excess male or female population in certain age groups.
Source: United Nations, World Population Prospects 2017
https://esa.un.org/unpd/wpp/Download/Standard/Population/ (medium variant).

Line charts

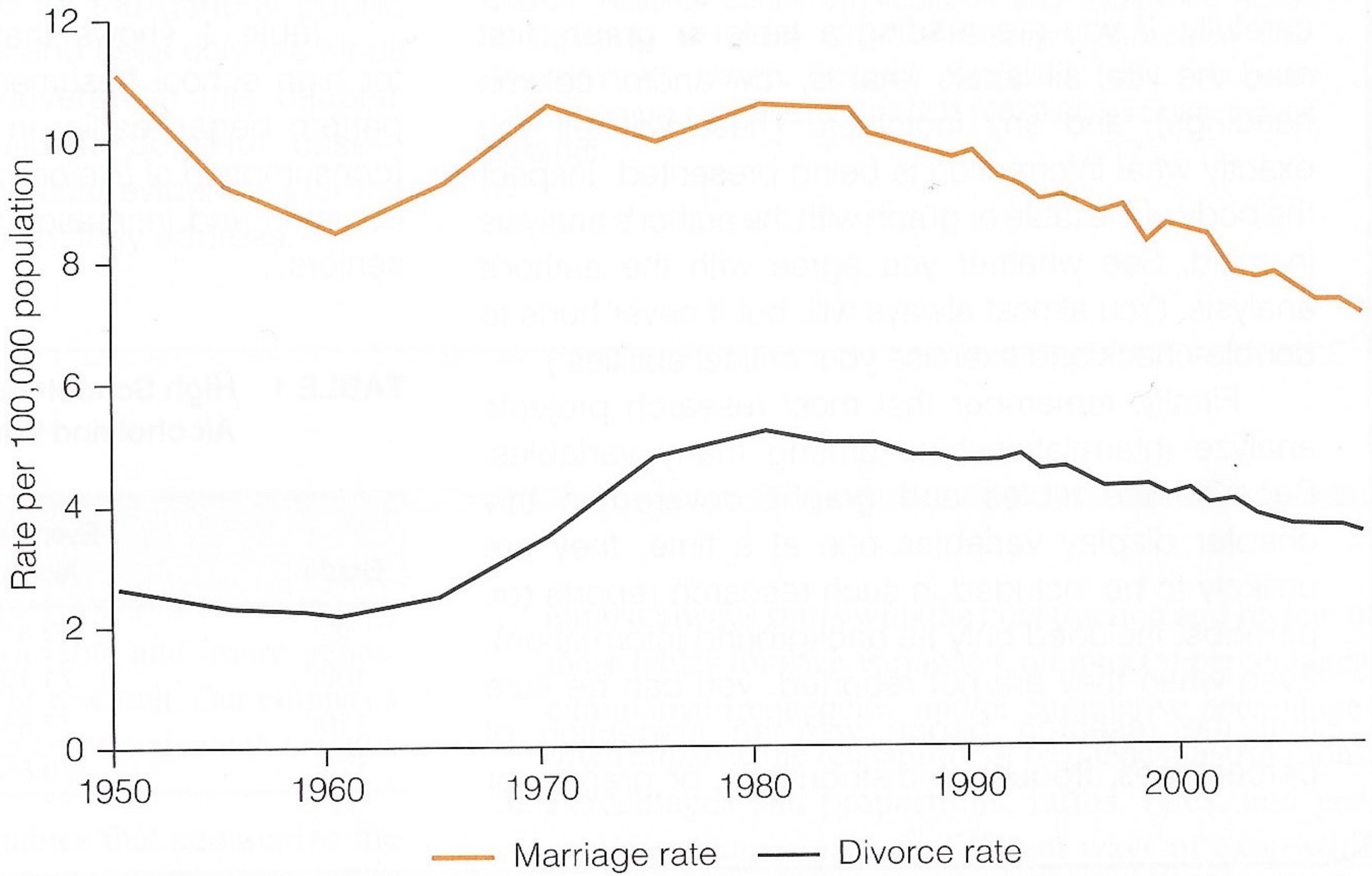
- Sometimes called **frequency polygons**
- Constructed similarly to a histogram, except graph a dot at each category's midpoint and then connect the dots
- Especially appropriate for continuous interval-ratio level variables
- It can be used for discrete interval-ratio level variables

Age distribution by gender, United States, 2010



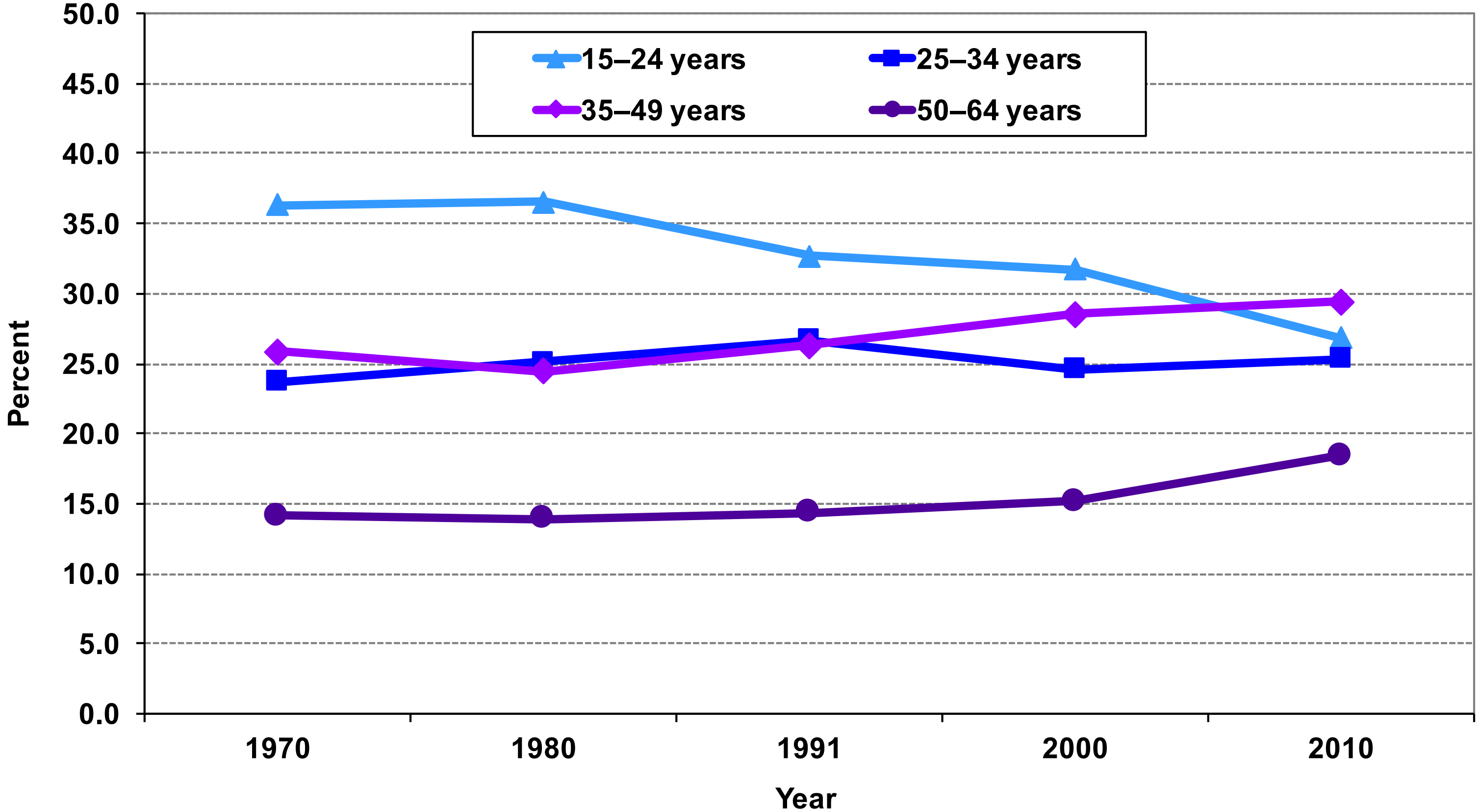
Source: Healey 2015, p.50.

Marriage and divorce rates, United States, 1950–2008



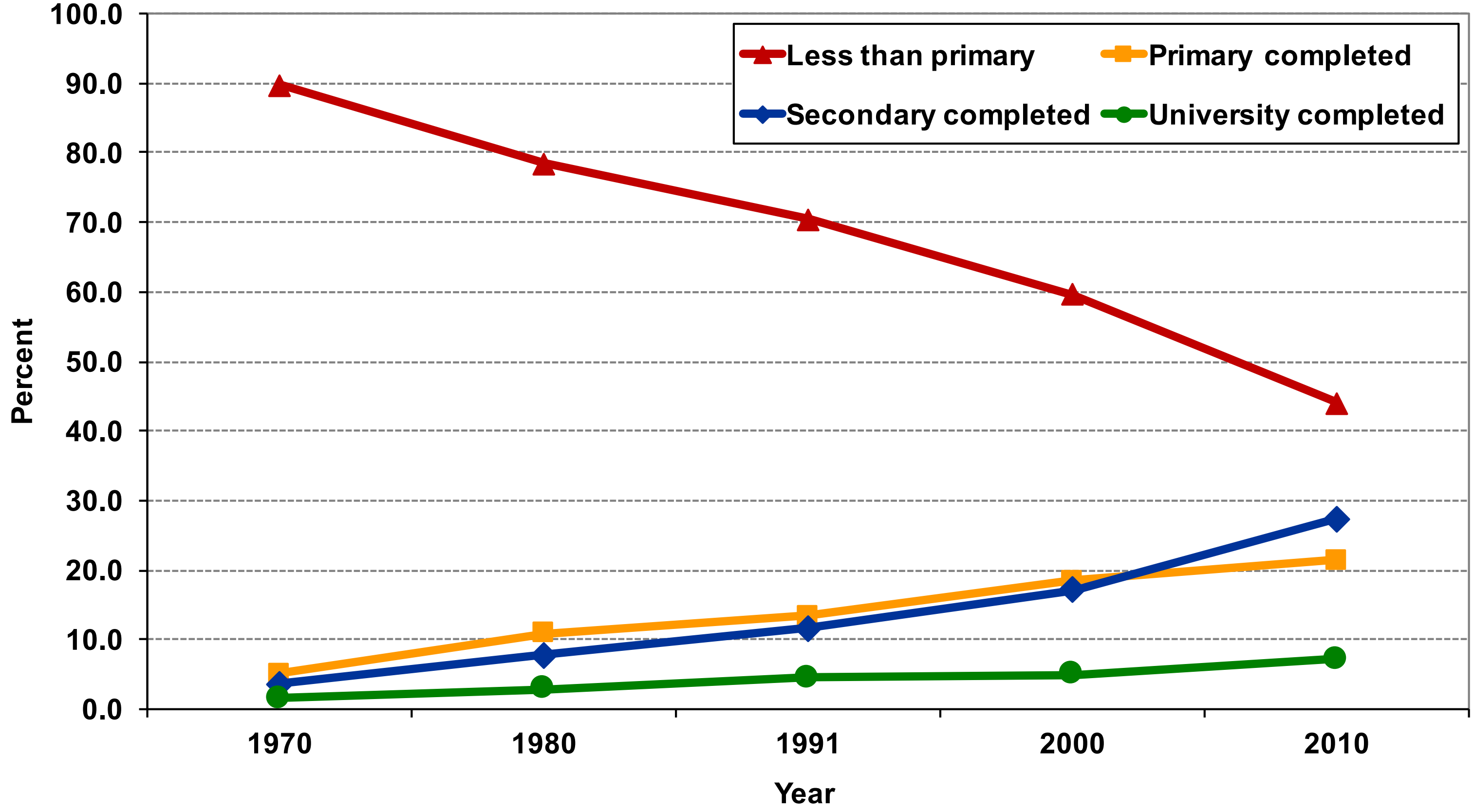
Source: Healey 2015, p.55.

Age distribution, Brazil, 1970–2010



Source: 1970–2010 Brazilian Demographic Censuses.

Education distribution, Brazil, 1970–2010



Source: 1970–2010 Brazilian Demographic Censuses.

Example: 2016 GSS in Stata (nominal-level variable)

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

      pweight: wtssall
          VCE: linearized
  Single unit: scaled
     Strata 1: vstrat
         SU 1: vpsu
        FPC 1: <zero>

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

Number of strata   =        65                 Number of obs     =      2,849
Number of PSUs     =       130                 Population size   = 2,844.2159
                                               Design df         =         65

----------------------
Religious |
group     | proportion
----------+-----------
 Protesta |      .4744
 Catholic |      .2348
   Jewish |      .0199
    Other |      .0534
     None |      .2174
          |
    Total |          1
----------------------
  Key:  proportion  =  cell proportion
```

Example: 2016 GSS in Stata (number of missing cases)

```
. tab religion, m
```

Religious group	Freq.	Percent	Cum.
Protestant	1,371	47.82	47.82
Catholic	649	22.64	70.46
Jewish	51	1.78	72.24
Other	159	5.55	77.78
None	619	21.59	99.37
.	18	0.63	100.00
Total	2,867	100.00	

Edited table (nominal-level variable)

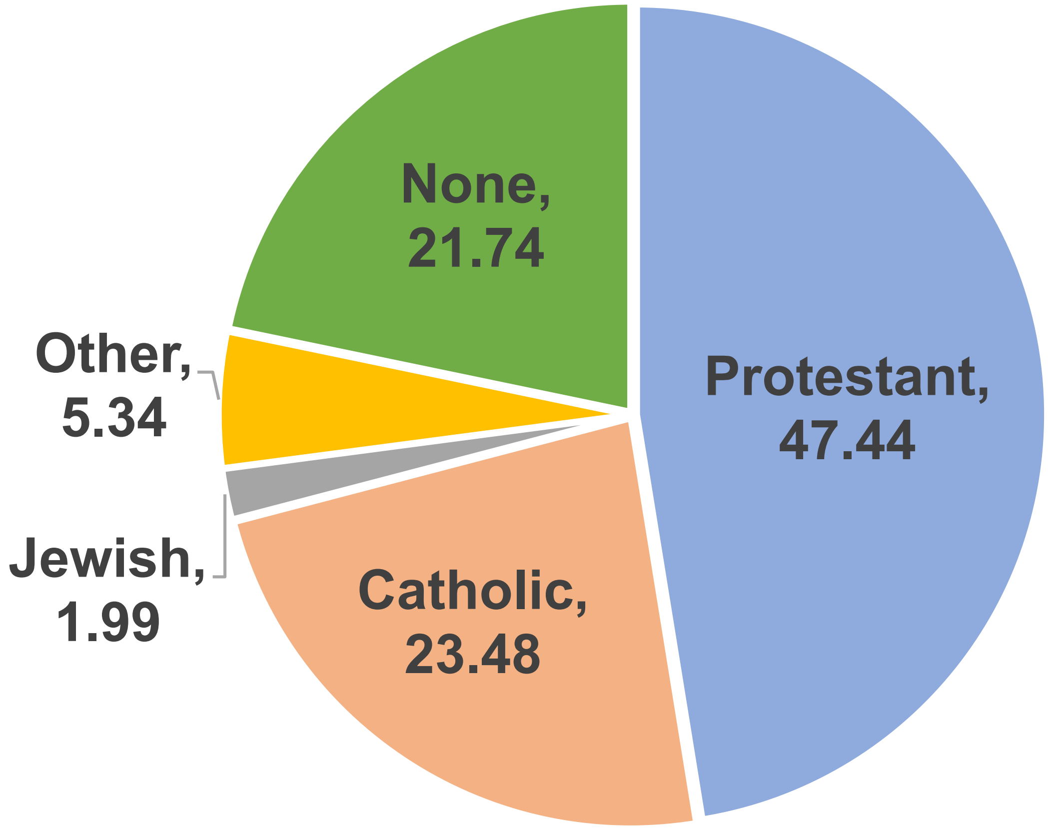
Table 1. Distribution of U.S. adult population by religious preference, 2016

Religion	Percentage
Protestant	47.44
Catholic	23.48
Jewish	1.99
Other	5.34
None	21.74
Total (sample size)	**100.00 (2,849)**
Missing cases	18

Source: 2016 General Social Survey.

Edited figure (nominal-level variable)

Figure 1. Percentage distribution of U.S. adult population by religious preference, 2016



Note: Number of cases is equal to 2,849. Missing cases are equal to 18.
Source: 2016 General Social Survey.

Example: 2016 GSS in Stata (ordinal-level variable)

```
. svy: tab rincome
(running tabulate on estimation sample)

Number of strata   =         65                  Number of obs     =      1,581
Number of PSUs     =        130                  Population size   = 1,641.5236
                                                 Design df         =         65
```

respondents income	proportion
lt $1000	.0169
$1000 to	.0354
$3000 to	.0216
$4000 to	.018
$5000 to	.0179
$6000 to	.0196
$7000 to	.017
$8000 to	.0175
$10000 -	.0674
$15000 -	.072
$20000 -	.0958
$25000 o	.6008
Total	1

Key: proportion = **cell proportion**

Example: 2016 GSS in Stata (number of missing cases)

```
. tab rincome, m
```

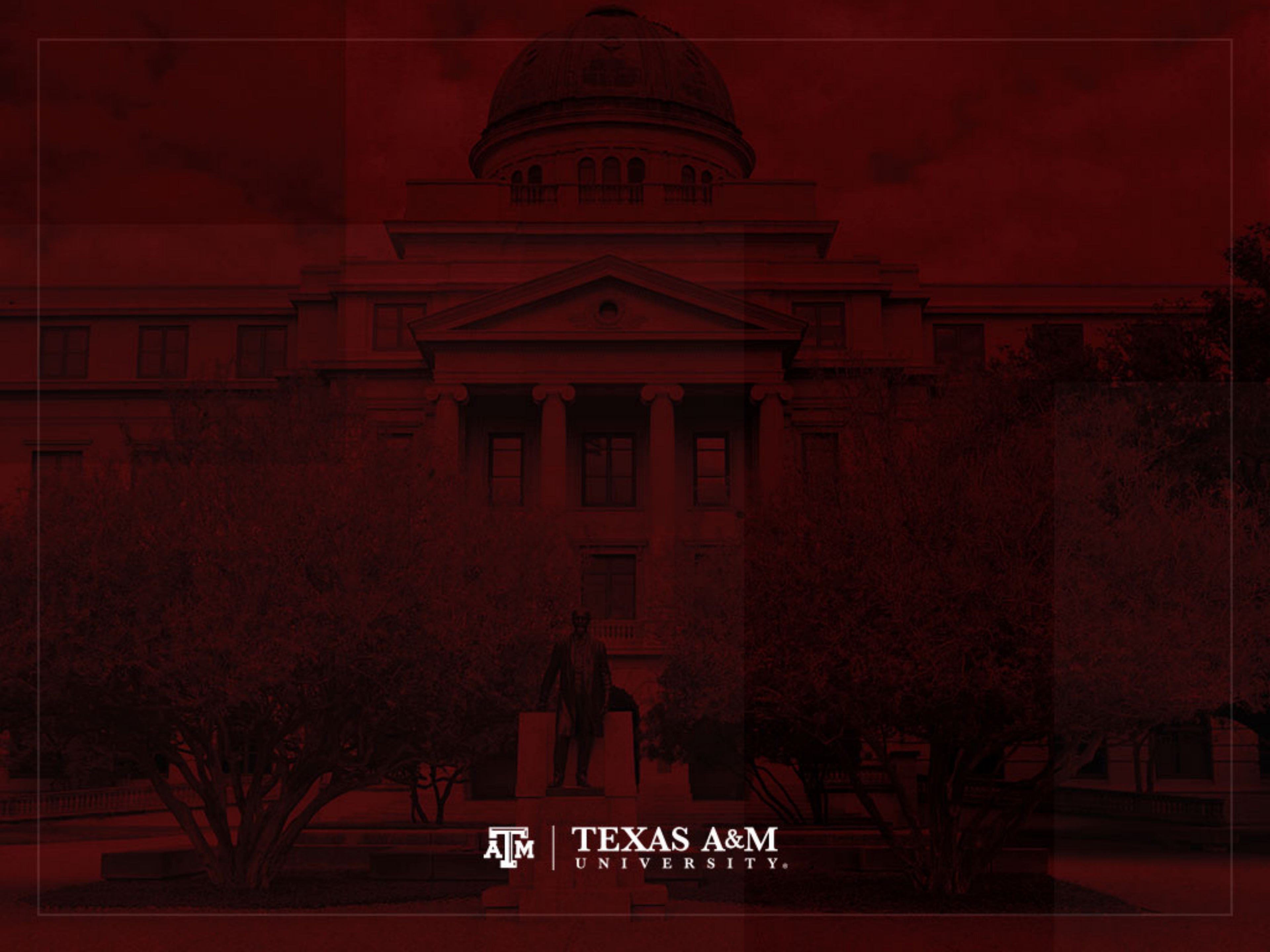
respondents income	Freq.	Percent	Cum.
lt $1000	25	0.87	0.87
$1000 to 2999	51	1.78	2.65
$3000 to 3999	32	1.12	3.77
$4000 to 4999	30	1.05	4.81
$5000 to 5999	31	1.08	5.89
$6000 to 6999	31	1.08	6.98
$7000 to 7999	24	0.84	7.81
$8000 to 9999	34	1.19	9.00
$10000 - 14999	96	3.35	12.35
$15000 - 19999	112	3.91	16.25
$20000 - 24999	138	4.81	21.07
$25000 or more	977	34.08	55.14
.a	150	5.23	60.38
.i	1,136	39.62	100.00
Total	2,867	100.00	

Edited table (ordinal-level variable)

Table 2. Distribution of U.S. adult population by income, 2016

Respondents' income	Percentage	Cumulative percentage
Less than $1,000	1.69	1.69
$1,000 to 2,999	3.54	5.23
$3,000 to 3,999	2.16	7.39
$4,000 to 4,999	1.80	9.19
$5,000 to 5,999	1.79	10.98
$6,000 to 6,999	1.96	12.94
$7,000 to 7,999	1.70	14.64
$8,000 to 9,999	1.75	16.39
$10,000 to 14,999	6.74	23.13
$15,000 to 19,999	7.20	30.33
$20,000 to 24,999	9.58	39.91
$25,000 or more	60.08	100.00
Total (n)	**100.00 (1,581)**	
Refused / Don't know (n)	(150)	
Not applicable (n)	(1,136)	

Source: 2016 General Social Survey.


TEXAS A&M
UNIVERSITY