Quantitative Methods – I (Statistics)

A. Y. 2024-25

Prof. Marco Stefanucci

Chapter 2
Organizing and Graphing Data

Organizing and Graphing data: Road Map

1. Distributions
2. Frequency Distributions: Absolute, Relative, Percentage
3. Cumulative Distributions: Absolute, Relative, Percentage
4. Appropriate Graphs: Pie charts, Bar charts, Pareto charts, Histogram
5. Shapes of the distributions

Univariate and multivariate distributions

Raw statistical information takes the form of a **unit distribution** **simple** (univariate) o **multiple** (multivariate), depending on the number of variables that are considered.

Univariate distribution: a single variable is considered.

Units	Values of X
u_1	x_1
u_2	x_2
$\vdots$	$\vdots$
u_i	x_i
$\vdots$	$\vdots$
u_n	x_n

We wish to describe and summarize the main facts about X.

Univariate and multivariate distributions

Multivariate distribution:

Units	Values of X	Values of Y	...	Values of Z
u_1	x_1	y_1	...	z_1
u_2	x_2	y_2	...	z_1
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
u_i	x_i	y_i	...	z_i
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
u_n	x_n	y_n	...	z_n

We can also discuss dependence and association among the variables.

Frequency Distributions

Raw data (i.e. in their original form): typically so large to look meaningless ➔ immense importance of Descriptive Statistics.

Ex: age of 50 students…

21	19	24	25	29	34	26	27	37	33
18	20	19	22	19	19	25	22	25	23
25	19	31	19	23	18	23	19	23	26
22	28	21	20	22	22	21	20	19	21
25	23	18	37	27	23	21	25	21	24

Frequency Distributions

A frequency distribution is a tabular way of summarizing the distribution of a character.

Collection of Raw Data: ex. Age of 50 students

21	19	24	25	29	34	26	27	37	33
18	20	19	22	19	19	25	22	25	23
25	19	31	19	23	18	23	19	23	26
22	28	21	20	22	22	21	20	19	21
25	23	18	37	27	23	21	25	21	24

From raw data to frequency distribution

Frequency Distribution

Age	Nr. Of Students
18	3
19	8
20	3
21	6
22	5
23	6
24	2
25	6
26	2
27	2
28	1
29	1
31	1
33	1
34	1
37	2
Total	50

Frequency Distribution – Absolute (f_i)

Lists all values/categories (x_i) and associated number of elements (f_i)

Ex: age of 50 students…

Variable

Values (x_i) (Categories/ Classes)

Age	Nr. Of Students
18	3
19	8
20	3
21	6
22	5
23	6
24	2
25	6
26	2
27	2
28	1
29	1
31	1
33	1
34	1
37	2
Total	50

Absolute Frequencies of *i*-thValue (or *i*–th Category / Class) → (f_i)

$N=\sum f_i$ = total nr of elements

Frequency Distribution – Absolute (f_i)

The frequency distribution is used also for qualitative and quantitative variables.

Lists all values/categories (x_i) and associated number of elements (f_i)

Ex: worries about reaching the end of the month…

Variable → | **Response** | **Number of Adults** | ← Frequency column

Response	Number of Adults
Very worried	162
Moderately worried	203
Not too worried	305
Not worried at all	325
Others	20
	Sum = 1015

Category → Not too worried

305 ← Frequency

Organizing data

1.6 The following table lists the number of billionaires in eight countries as of February 2011, as reported in The New York Times of July 27, 2011.

Country	Number of Billionaires
United States	413
China	115
Russia	101
India	55
Germany	52
Britain	32
Brazil	30
Japan	26

Variable

Observation

Data set

Character (or category)

Source: Forbes, International Monetary Fund.

Briefly explain the meaning of a member, a variable, a measurement, and a data set with reference to this table.

a. What is the variable for this data set?
b. How many observations are in this data set?
c. How many characters does this data set contain?

a. Number of Billionaires by Country
b. n=413+115+101+55+52+32+30+26=824
c. 8 *(United States, China, etc.)*

Frequency distributions for quantitative variables

In general, a frequency table provides a useful summary if the number of values x_j is small.
For discrete variable with a large number of values and for continuous variables we have to group the values into classes to achieve an adequate level of synthesis. However, the grouping is by and large arbitrary and the mapping of the unit distribution (u_i, x_i) into a frequency distribution carries with it an information loss.
The construction of the frequency distribution entails:

i subdividing the range of values that X can take into nonoverlapping intervals (aka classes)

ii counting the number of observations falling within each class

Frequency distributions for quantitative variables

For continuous variables there are several alternative ways of defining the classes:

Left-open, right-closed classes:

$x_j \dashv x_{j+1}, \quad (x_j, x_{j+1}], \quad x_j < x \leq x_{j+1}$

Left-closed, right-open classes:

$x_j \vdash x_{j+1}, \quad [x_j, x_{j+1}), \quad x_j \leq x < x_{j+1}$

The *size* of the class is defined by the difference between the upper and the lower bounds $(x_{j+1} - x_j)$.

Distribution in classes

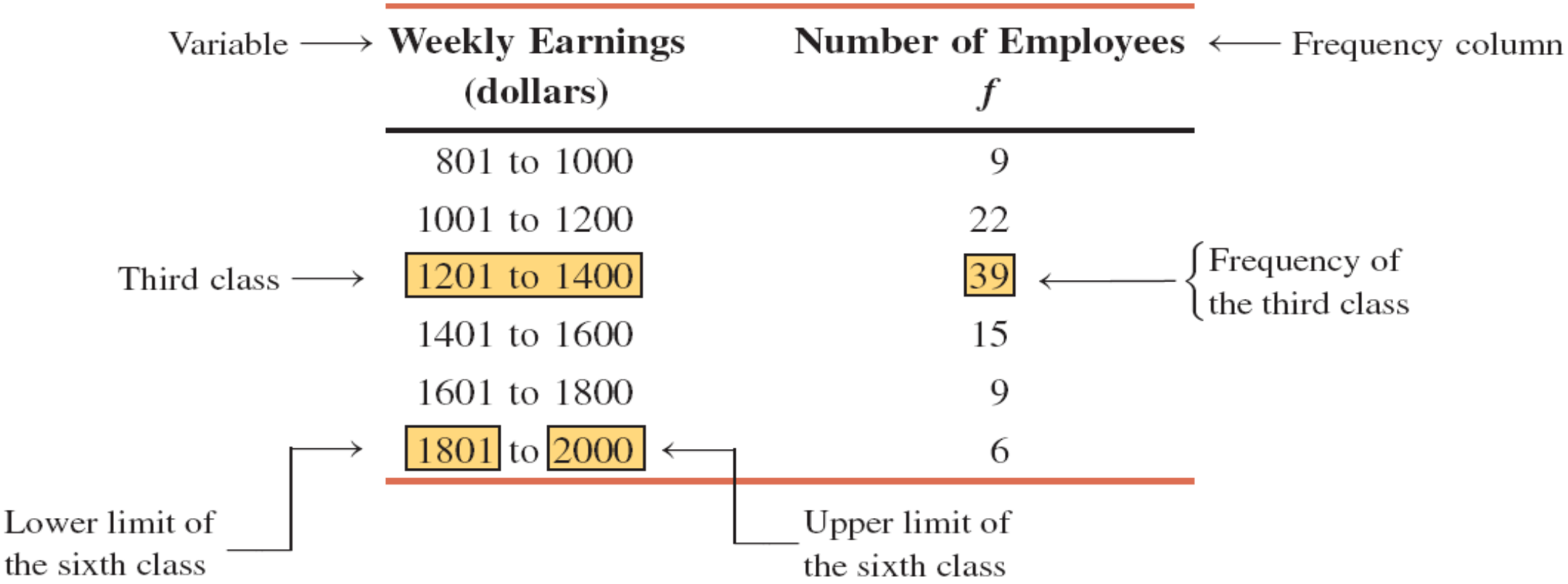
More generally, for quantitative variables is useful to subdivide the range of values that X can take into mutually exclusive and exhaustive intervals or classes

Classes (x_i) and associated number of elements (f_i or m_i)

Distribution in classes

Classes (x_i) and associated number of elements (f_i or m_i)

Ex: weekly earnings (grouped into classes)



Frequency Distribution – Excercise

The following data give the total number of iPads® sold by an internet store in 30 days.

8	25	11	15	29	22	10	5	17	21	22	13	26
16	18	12	9	26	20	16	23	14	19	23	20	16
27	16	21	14									

Construct a frequency distribution table using the following classes:

5-9, 10-14, 15-19, 20-24, 25-29

Frequency Distribution – Excercise

The following data give the total number of iPads® sold by an internet store in 30 days.

8	25	11	15	29	22	10	5	17	21	22	13	26
16	18	12	9	26	20	16	23	14	19	23	20	16
27	16	21	14									

Classes	m_i
5-9	
10-14	
15-19	
20-24	
25-29	

Frequency Distribution – Exercise

The following data give the total number of iPads® sold by an internet store in 30 days.

8	25	11	15	29	22	10	5	17	21	22	13	26
16	18	12	9	26	20	16	23	14	19	23	20	16
27	16	21	14									

Classes	m_i
5-9	3
10-14	
15-19	
20-24	
25-29	

Frequency Distribution – Excercise

The following data give the total number of iPads® sold by an internet store in 30 days.

8	25	11	15	29	22	10	5	17	21	22	13	26
16	18	12	9	26	20	16	23	14	19	23	20	16
27	16	21	14									

Classes	m_i
5-9	3
10-14	6
15-19	8
20-24	8
25-29	5

Frequency Distributions – Relative and Percentage

Relative Frequencies

$$rf_i = \frac{f_i}{n} \qquad \text{with} \sum rf_i = 1$$

Percentage Distribution

$$p_i = 100 \times rf_i \quad \text{with} \sum p_i = 100$$

Frequency Distributions – Relative and Percentage

The relative frequency f_j by which x_j occurs si the ratio of n_j to the total frequency, n:

$$f_j = \frac{n_j}{n}, \quad j = 1, \ldots, K.$$

The percentage frequency is obtained by multiplying f_j by 100:

$$p_j = 100 \cdot \frac{n_j}{n} = 100 \cdot f_j, \quad j = 1, \ldots, K.$$

The following obviously hold:

$$0 \leq f_j \leq 1, \quad \sum_{j=1}^{K} f_j = 1; \quad 0 \leq p_j \leq 100, \quad \sum_{j=1}^{K} p_j = 100.$$

Relative frequency

Relative frequency distribution:

Values of X	Frequency
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_j	f_j
$\vdots$	$\vdots$
x_K	f_K
Total	1

Percentage frequency

Percentage frequency distribution:

Values of X	Relative frequency		Percentage freq.
x_1	f_1	x 100	p_1
x_2	f_2	x 100	p_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_j	f_j	x 100	p_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_K	f_K	x 100	p_K
Total	1		100

Cumulative frequency

For variables that are measured on an ordinal or a quantitative scale we can count the number of cases which have a value not greater than x_j.
This is known as a cumulative frequency.
More formally, the absolute cumulative frequency is defined as

$$N_j = \sum_{k=1}^{j} n_k, \quad j = 1, \ldots, K.$$

Cumulative frequency

Values of X	Frequency	Cumulative freq.
x_1	n_1	$N_1 = n_1$
x_2	n_2	$N_2 = n_1 + n_2$
$\vdots$	$\vdots$	$\vdots$
x_j	n_j	$N_j = n_1 + n_2 + \cdots + n_j$
$\vdots$	$\vdots$	$\vdots$
x_K	n_K	n
Totale	n	

Cumulative relative and percentage frequency

Simularly, we define

- Cumulative relative frequency:

$$F_j = \sum_{k=1}^{j} f_k, j = 1, \ldots, K$$

(note that $F_K = 1$)

- Cumulative percentage frequency:

$$P_j = \sum_{k=1}^{j} p_k, j = 1, \ldots, K$$

(note that $P_K = 100$)

Cumulative relative frequency

Values of X	Rel. freq.	Cumulative rel. freq.
x_1	f_1	$F_1 = f_1$
x_2	f_2	$F_2 = f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$
x_j	f_j	$F_j = f_1 + f_2 + \cdots + f_j$
$\vdots$	$\vdots$	$\vdots$
x_K	f_K	1
Total	1	

Frequency Distributions – Relative and Percentage

Ex: Federal Tax (in classes)

Federal and State Tax (in cents)	Frequency	Relative Frequency	Percentage
27 to less than 36	5	.10	10
36 to less than 45	21	.42	42
45 to less than 54	16	.32	32
54 to less than 63	6	.12	12
63 to less than 72	2	.04	4
	Sum = 50	Sum = 1.00	Sum = 100

Cumulative Distribution – Absolute

For each value (category/class) gives the total number of observations taking that value or lower (or falling below the upper boundary of each class)

Federal and State Tax (in cents)	Frequency f_i	Cumulative Frequency
27|-36	5	5
36|-45	21	26
45|-54	16	42
54|-63	6	48
63|-72	2	50
Total	50	

Cumulative Distribution: Relative and Percentage

Federal and State Tax (in cents)	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency	Percentage Distribution	Cumulative Percentage Distribution
27|-36	5	5	0.1	0.1	10.0	10.0
36|-45	21	26	0.42	0.52	42.0	52.0
45|-54	16	42	0.32	0.84	32.0	84.0
54|-63	6	48	0.12	0.96	12.0	96.0
63|-72	2	50	0.04	1	4.0	100.0
Total	50		1.00		100.0	

Graphic Presentation

It is important to choose the appropriate graphs to make statistical information coherent.

- **The Pie Chart**
- **The Bar Graph**
- **The Statistical Map**
- **The Histogram**
- **Times Series Charts**
- **Distortions in Graphs**

Graphing data

Appropriate Graphs:

- Pie charts, Bar charts, Pareto charts,

  ➔ for Qualitative and Quantitative Discrete

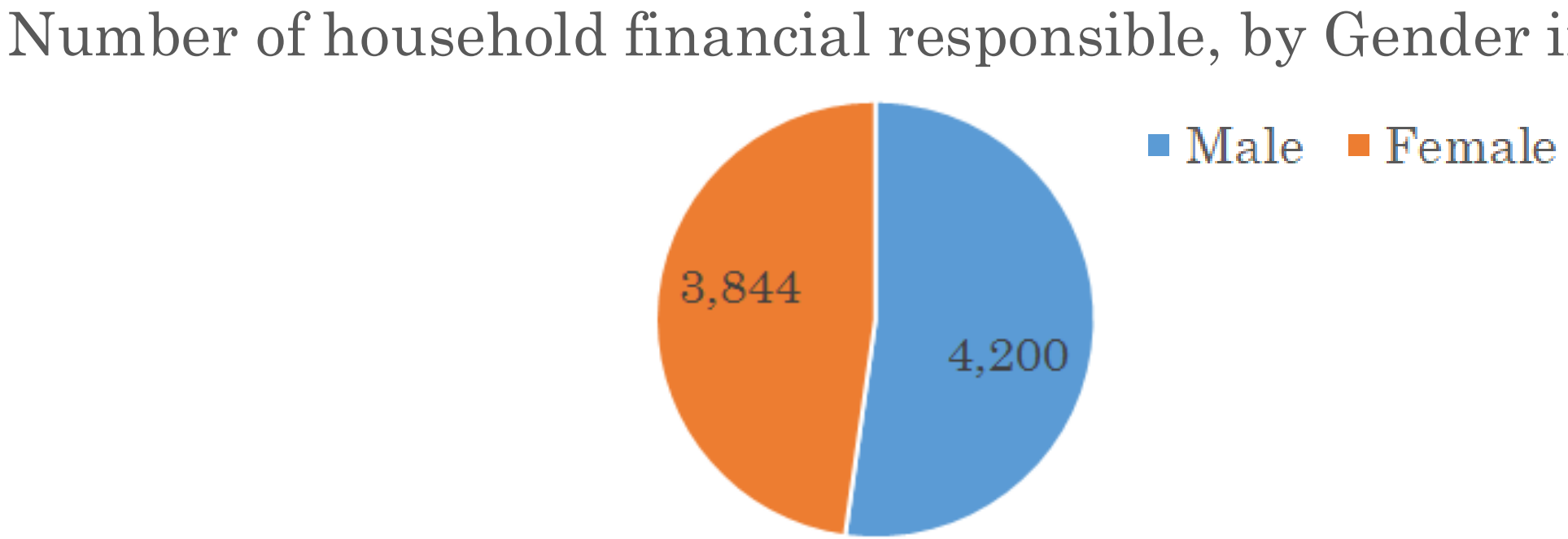
- Histogram

  ➔ for Quantitative Continuous (in classes)

Graphs – Pie Chart

A circle divided into portions that represent the relative frequencies or percentages of each category/value.

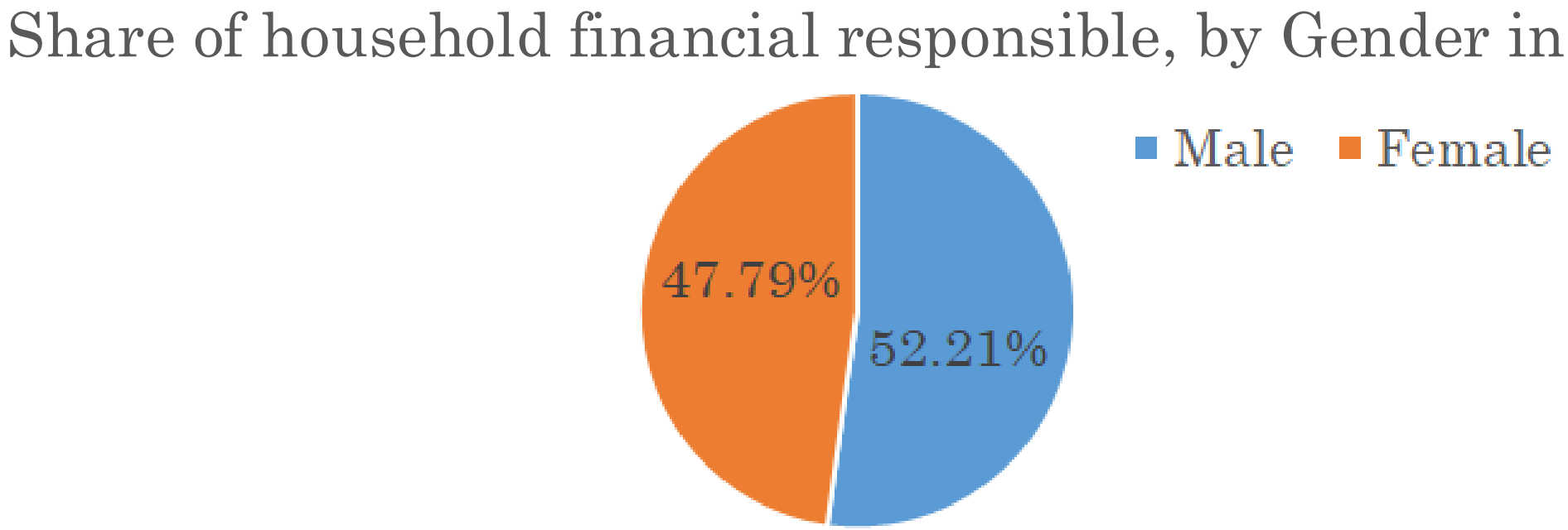
Appropriate for: Qualitative, Quantitative Discrete (few values)



Graphs – Pie Chart

A circle divided into portions that represent the relative frequencies or percentages of each category/value.

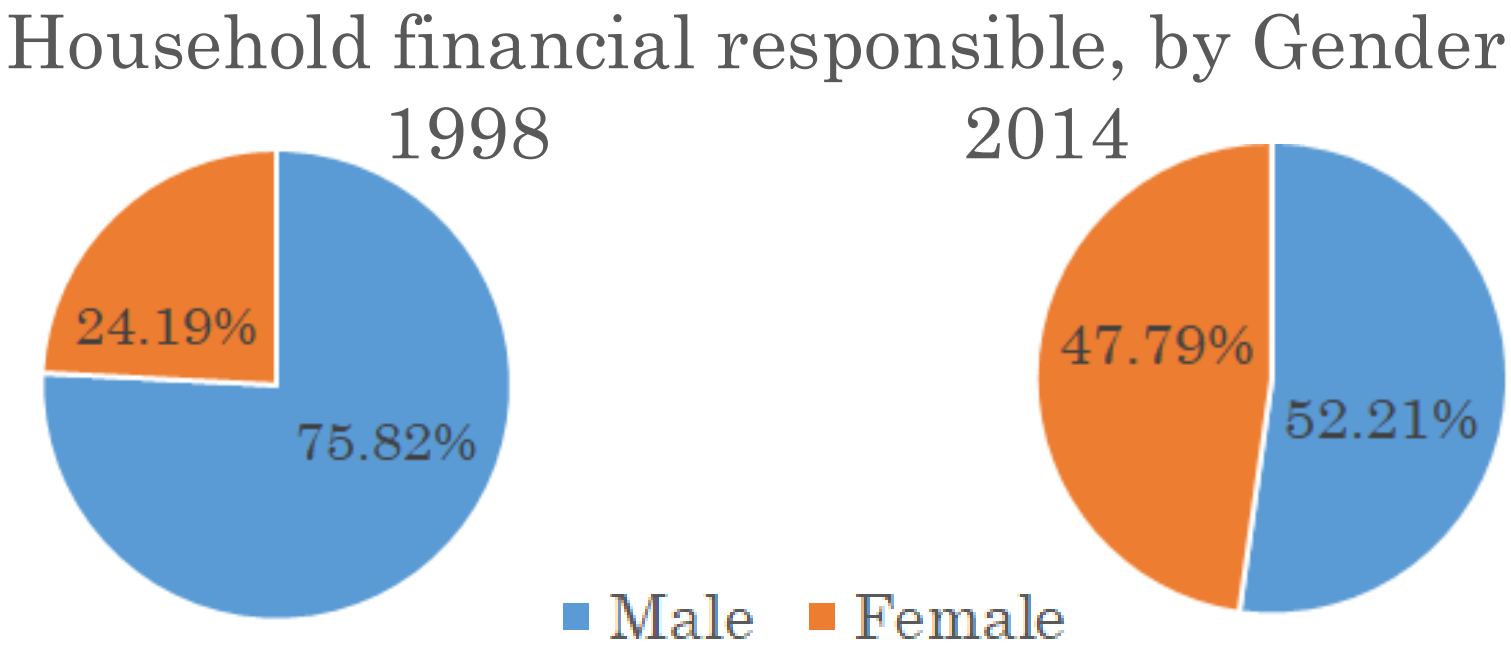
Appropriate for: Qualitative, Quantitative Discrete (few values)



Graphs – Pie Chart

A circle divided into portions that represent the relative frequencies or percentages of each category/value.

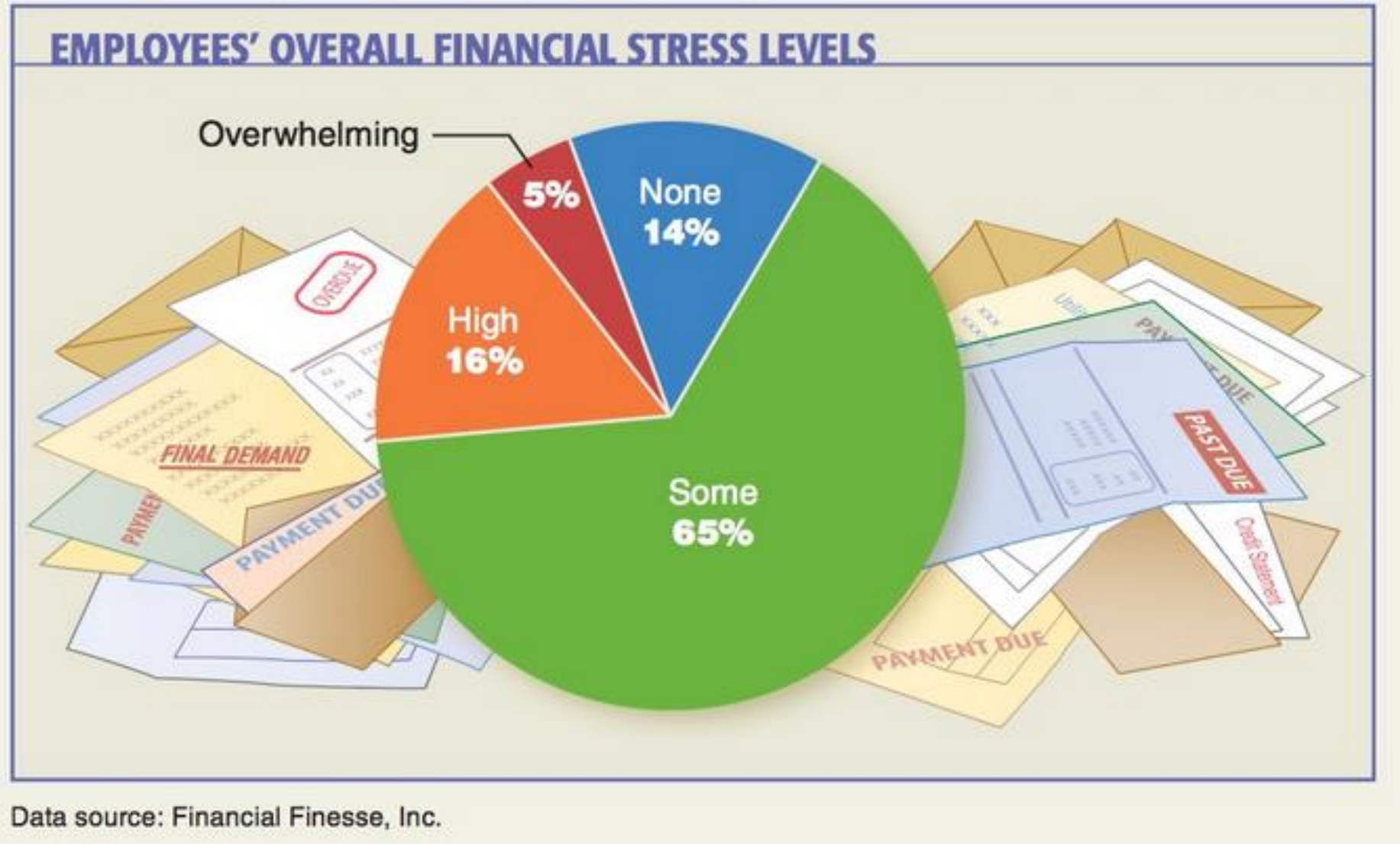
Appropriate for: Qualitative, Quantitative Discrete (few values)



Graphs – Pie Chart

Data source: Gallup poll of U.S. adults aged 18 and older conducted July 9–12, 2012

Graphs – Pie Chart



Graphs – Pie Chart

- ***Pie chart:*** a graph showing the differences in frequencies or percentages among categories of a **nominal** or an **ordinal** variable. The categories are displayed as segments of a circle whose pieces add up to 100 percent of the total frequencies.

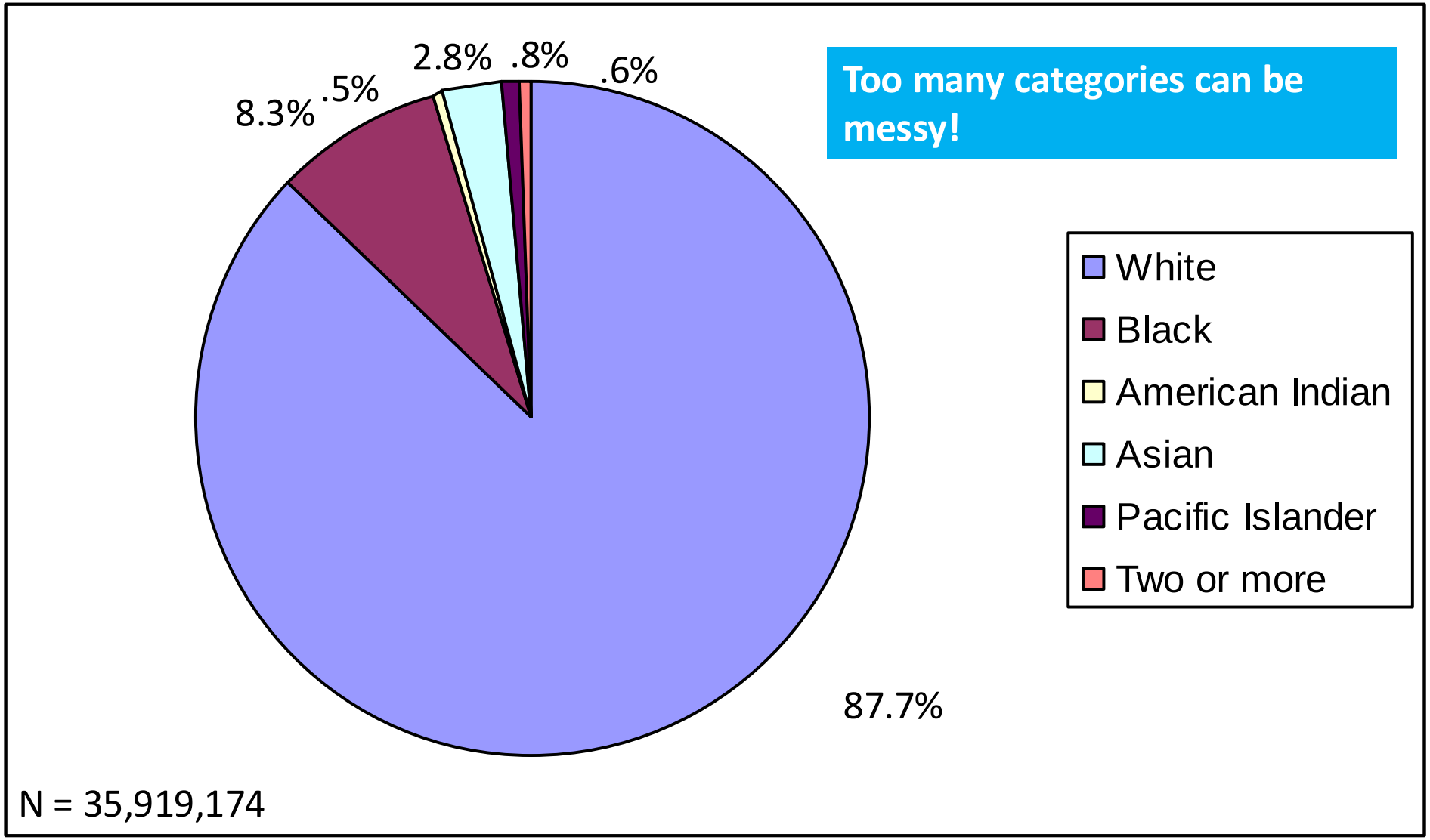


Figure 3.1 Annual Estimates of U.S. Population 65 Years and Over by Race, 2003

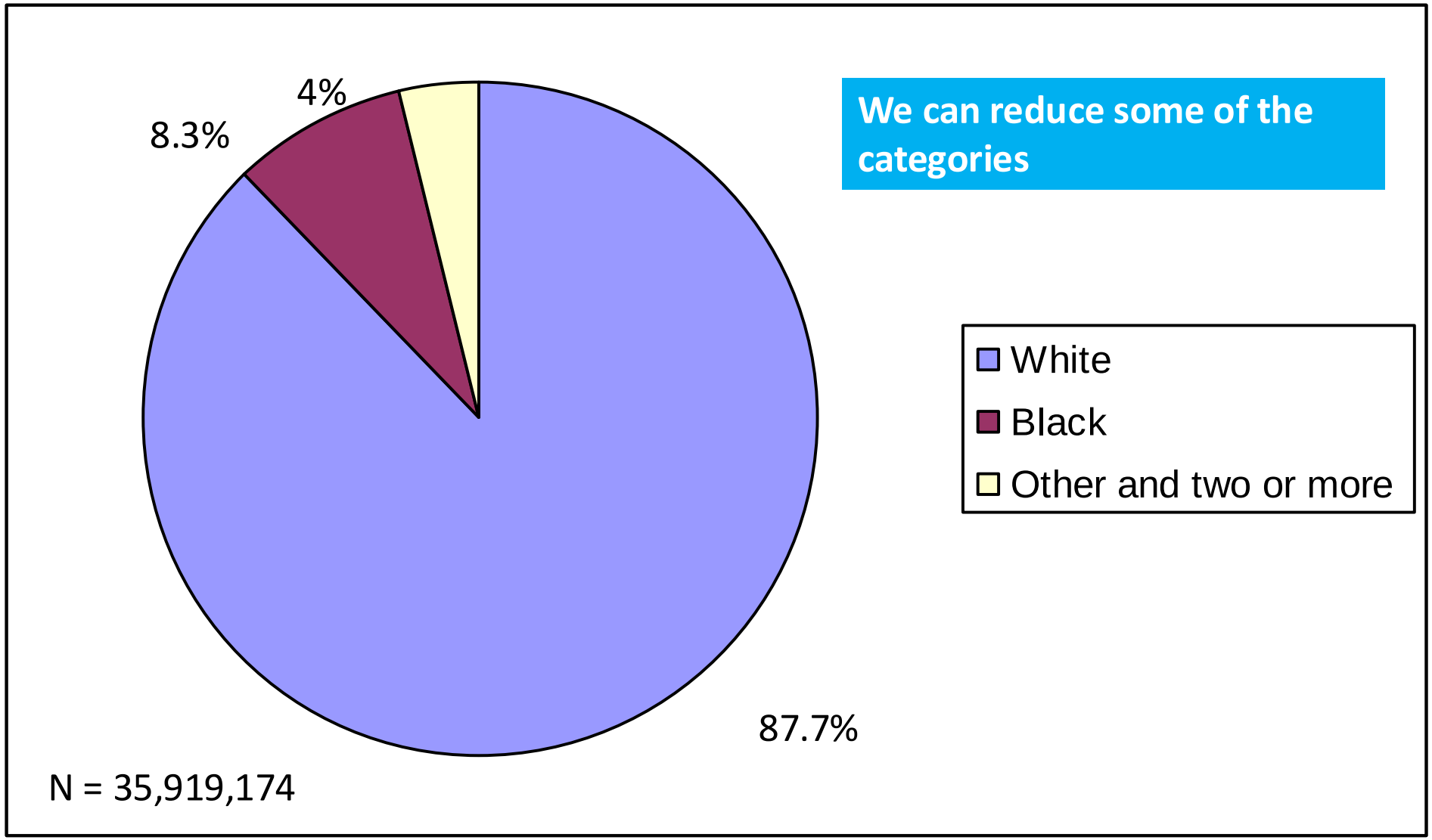
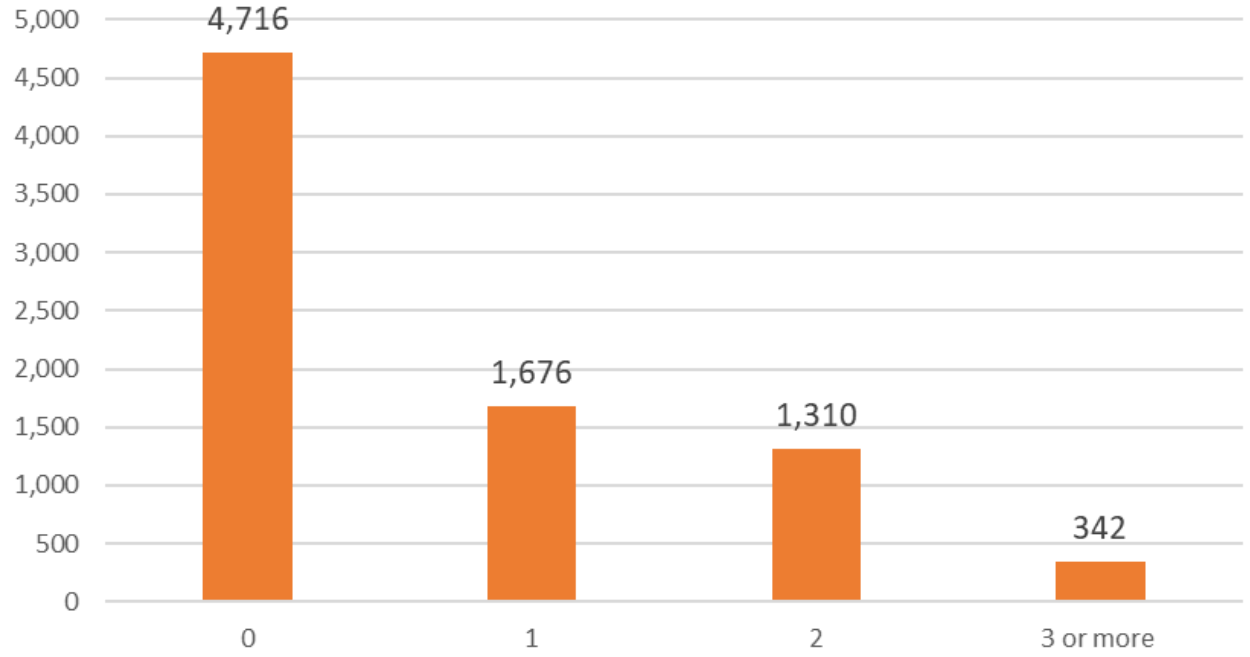


Figure 3.2 Annual Estimates of U.S. Population 65 Years and Over, 2003

Graphs – Bar Chart

Each bar's height represents the frequencies (absolute, relative, percentage) of each category/value.

Appropriate for: Qualitative, Quantitative Discrete

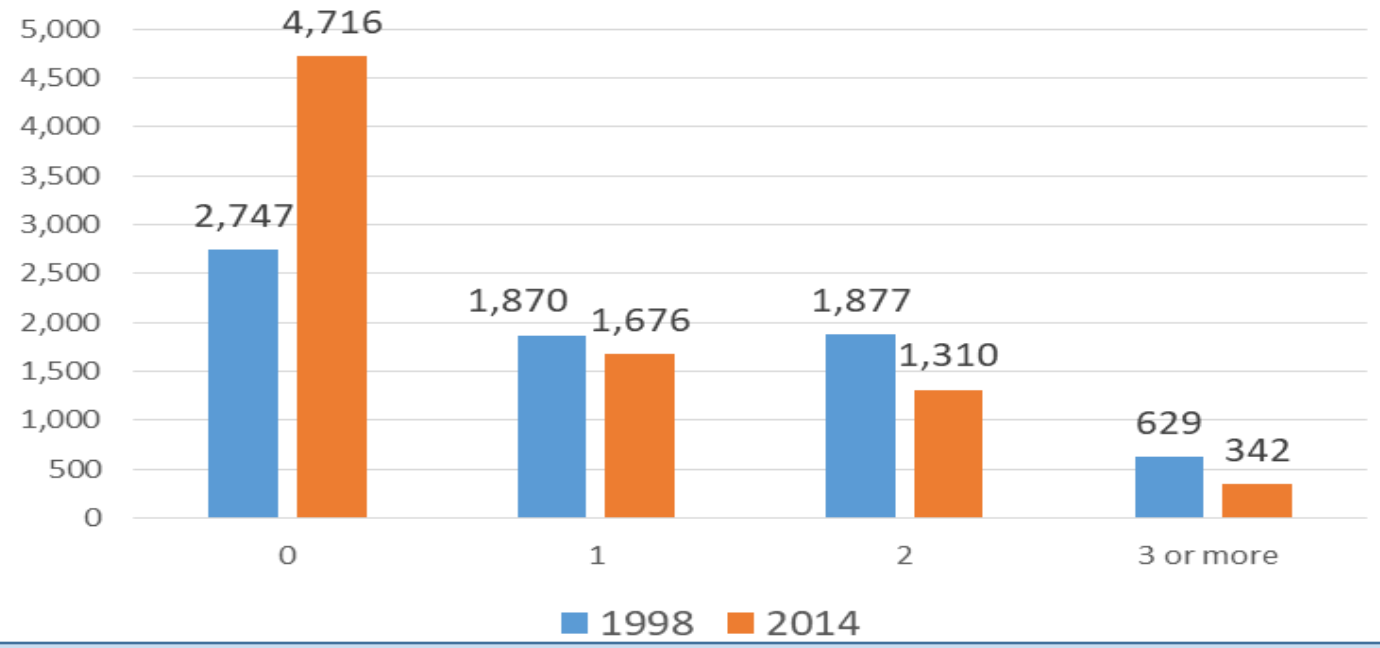


Graphs – Bar Chart

Each bar's height represents the frequencies (absolute, relative, percentage) of each category/value.

Appropriate for: Qualitative, Quantitative Discrete

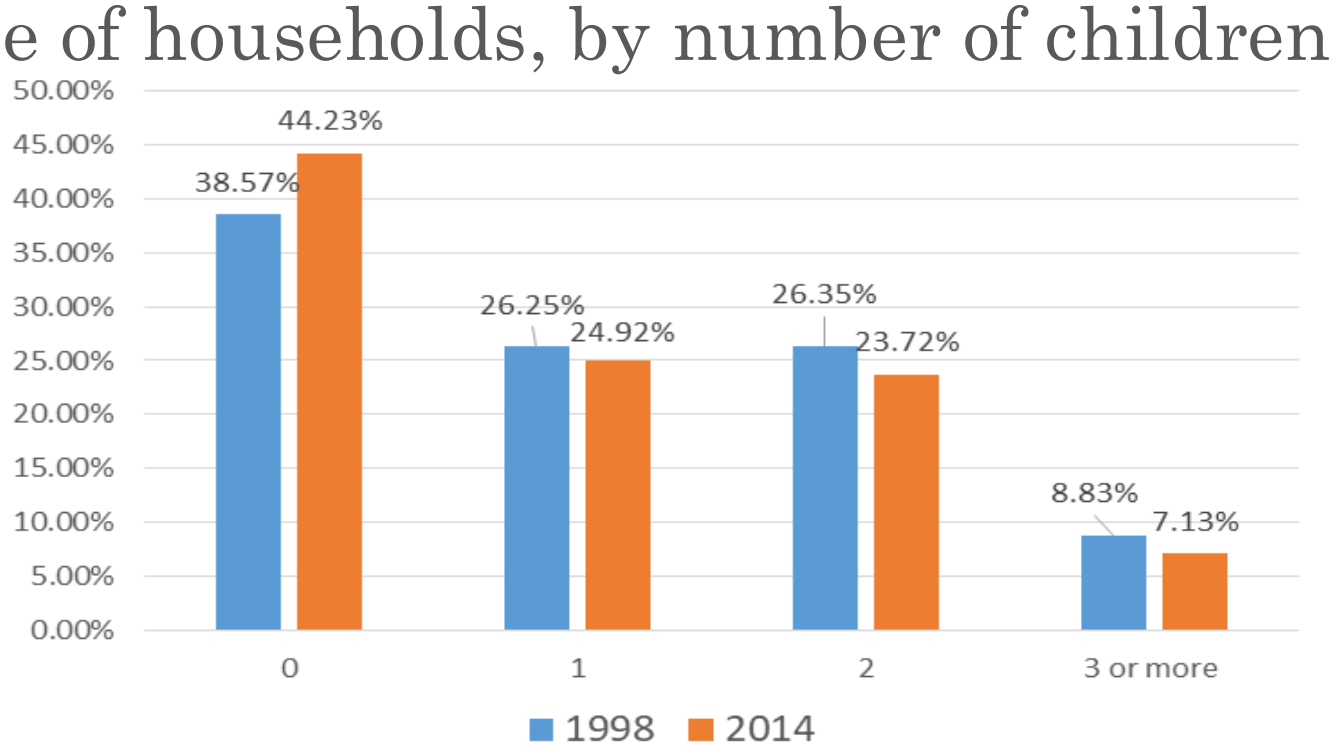
Number of households, by number of children: 1998 and 2014



Graphs – Bar Chart

Each bar's height represents the frequencies (absolute, relative, percentage) of each category/value.

Appropriate for: Qualitative, Quantitative Discrete



Graphs – Bar Chart

- ***Bar graph:*** a graph showing the differences in frequencies or percentages among categories of a **nominal** or an **ordinal** variable. The categories are displayed as rectangles of equal width with their height proportional to the frequency or percentage of the category.

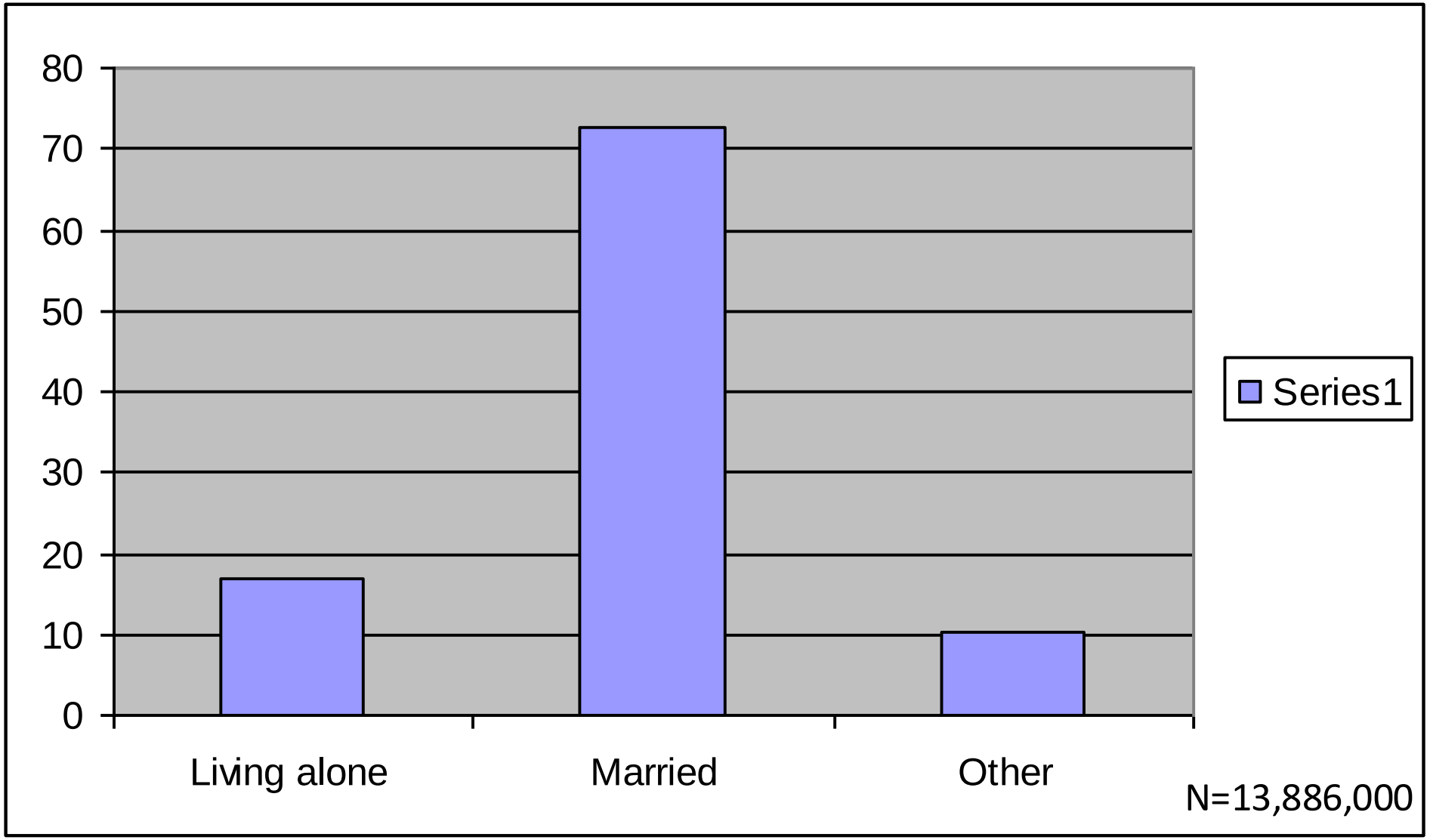


Figure 3.3 Living Arrangements of Males (65 and Older) in the United States, 2000

Can display more info by splitting sex

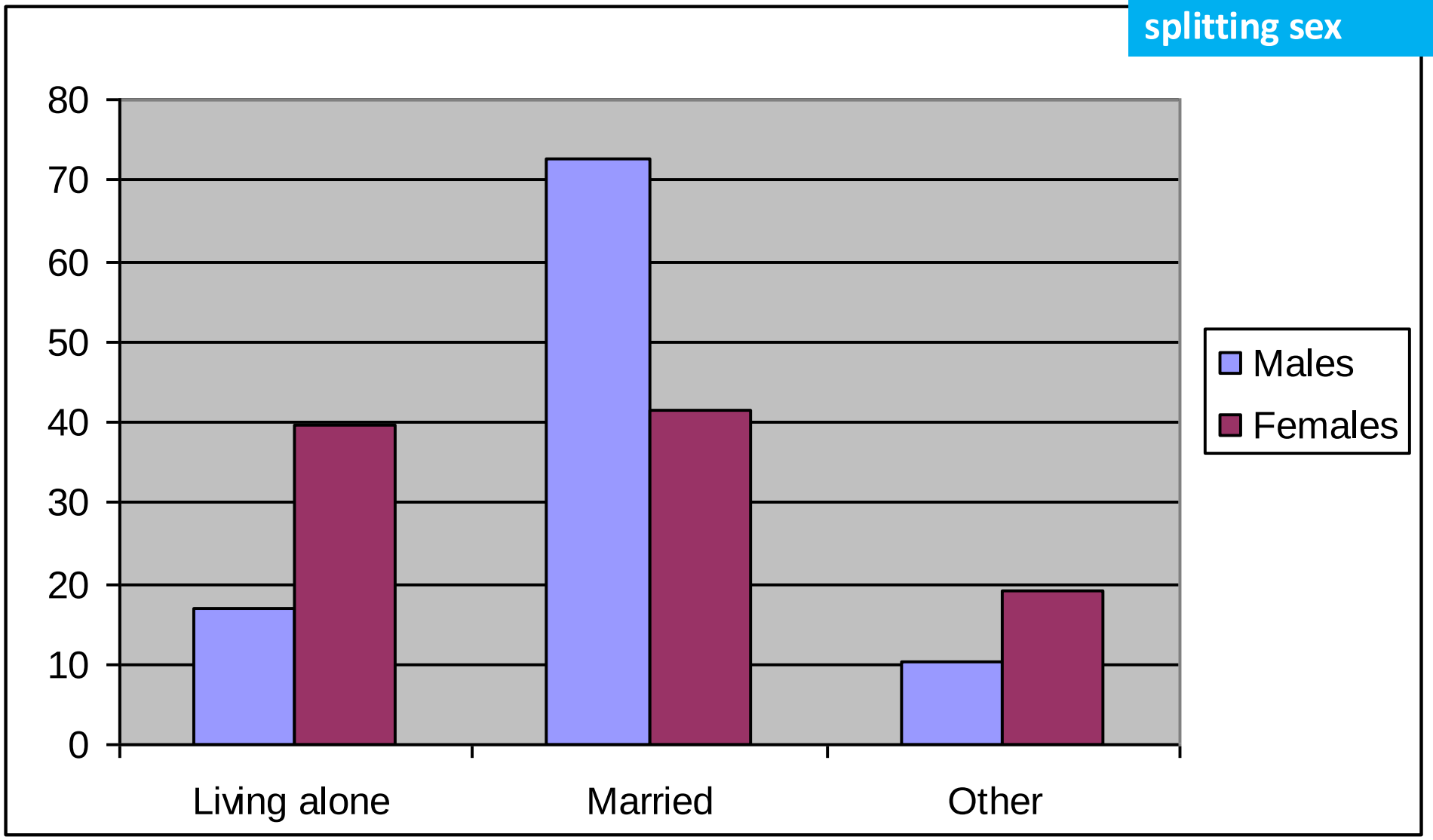
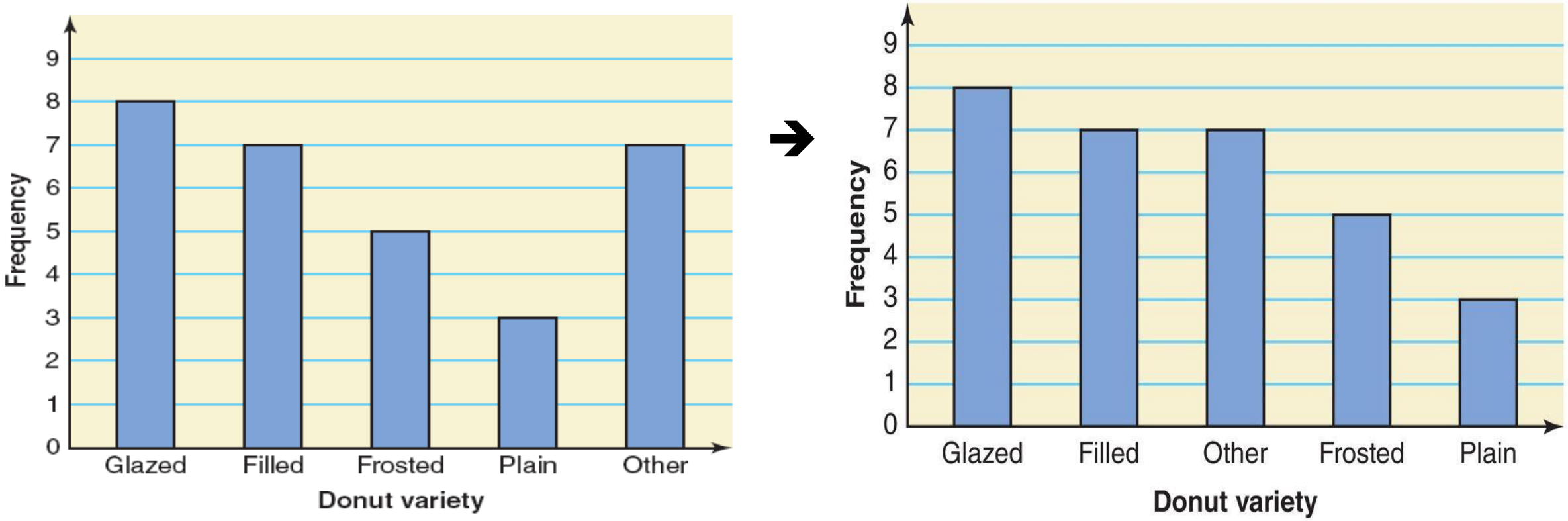


Figure 3.4 Living Arrangement of U.S. Elderly (65 and Older) by Gender, 2003

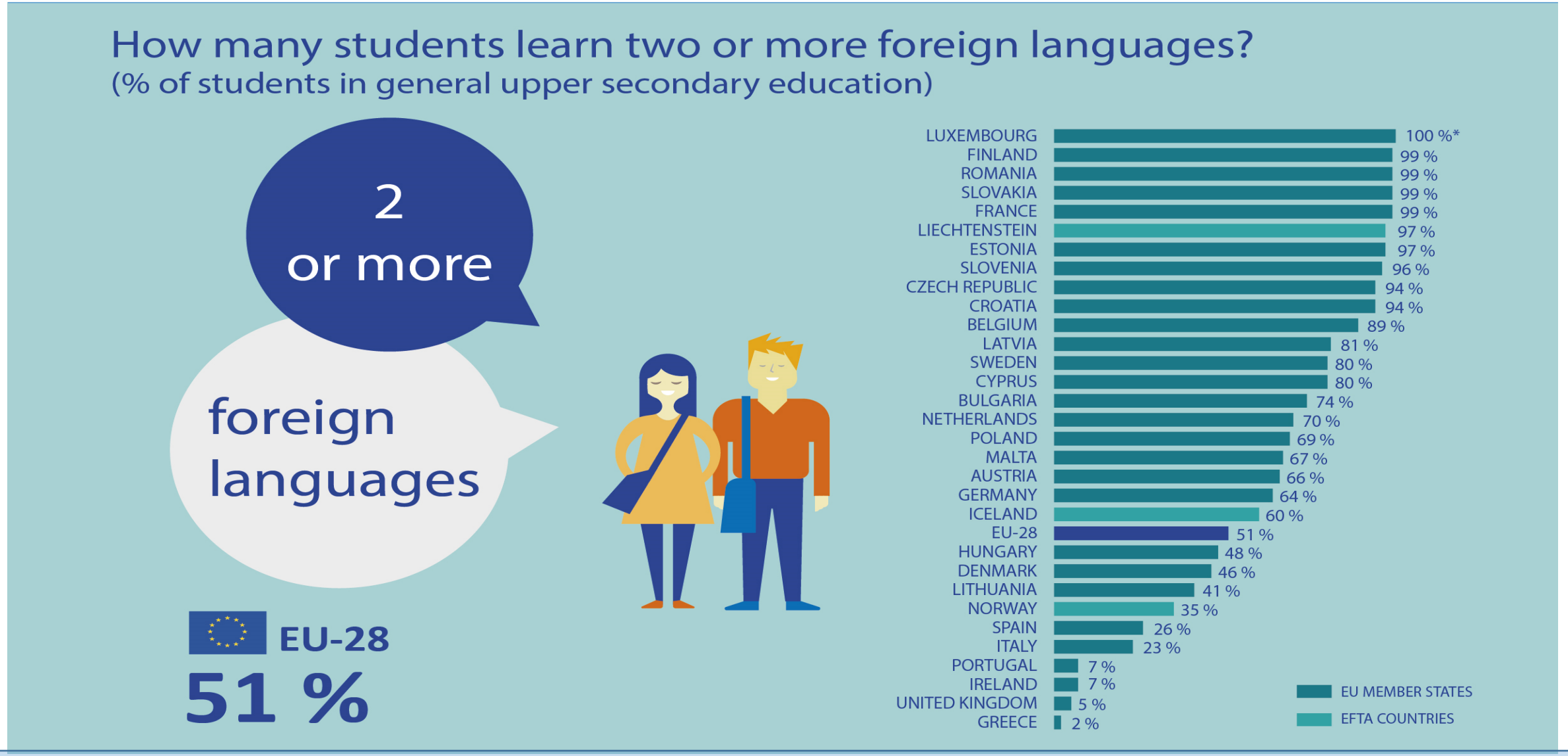
Graphs – Pareto Chart

A Bar chart where the bars are in descending order

Ex: Bar Chart ➔ Pareto Chart



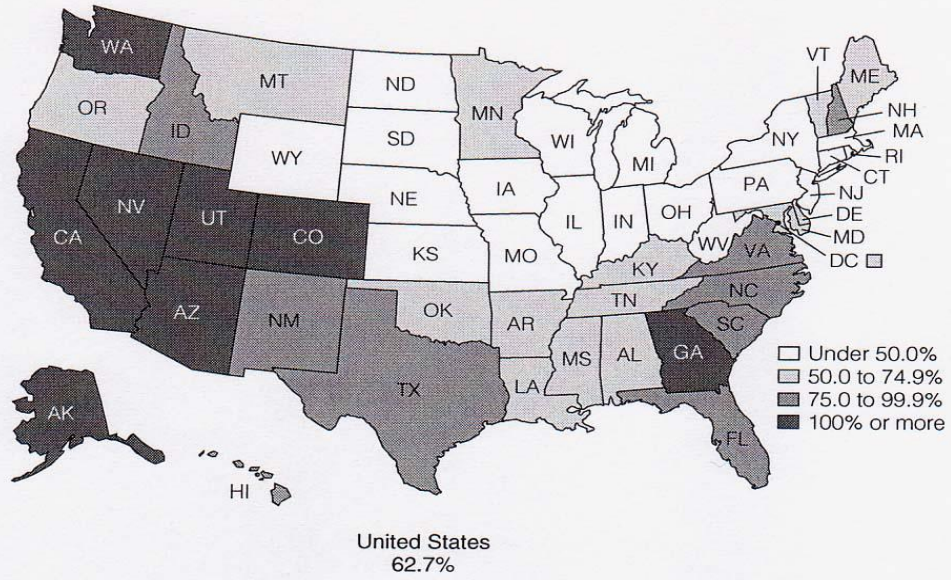
Graphs – Pareto Chart



The Statistical Map

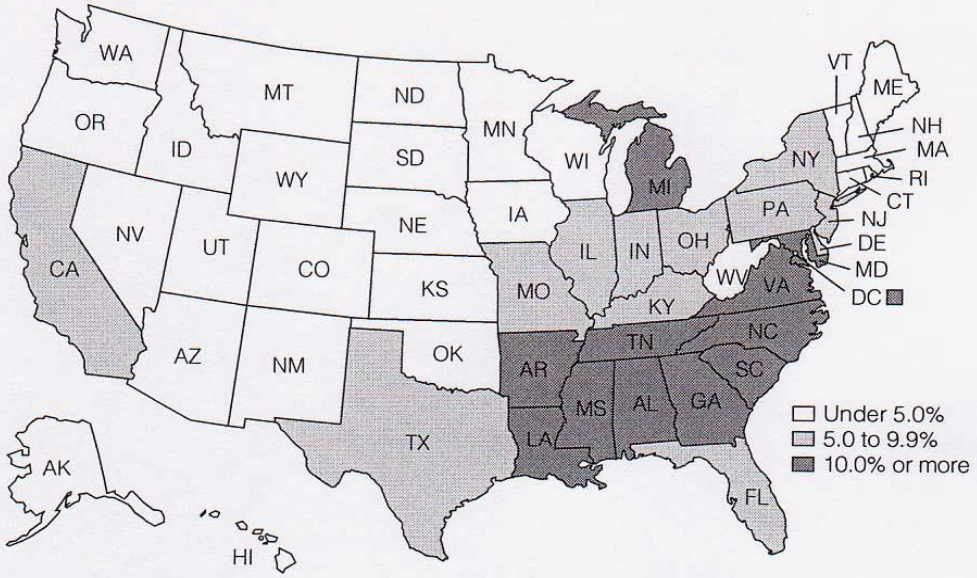
We can display dramatic geographical changes in American society by using a statistical map. Maps are especially useful for describing geographical variations in variables, such as population distribution, voting patterns, crimes rates, or labor force participation.

Figure 3.6 **Percentage Increase in Population 65 Years and Over, 1993 to 2020**



Source: U.S. Bureau of the Census, 1993 from 1994 Press Release, *Updated National/State Population Estimates,* CB94-43; 2020 from "Population Projections for States, by Age, Sex, Race, and Hispanic Origin: 1993 to 2020," *Current Population Reports,* P25-111, U.S. Government Printing Office, Washington, DC, 1994.

Figure 3.7 **Percentage Black of Total State Population 65 Years and Over, 1991**

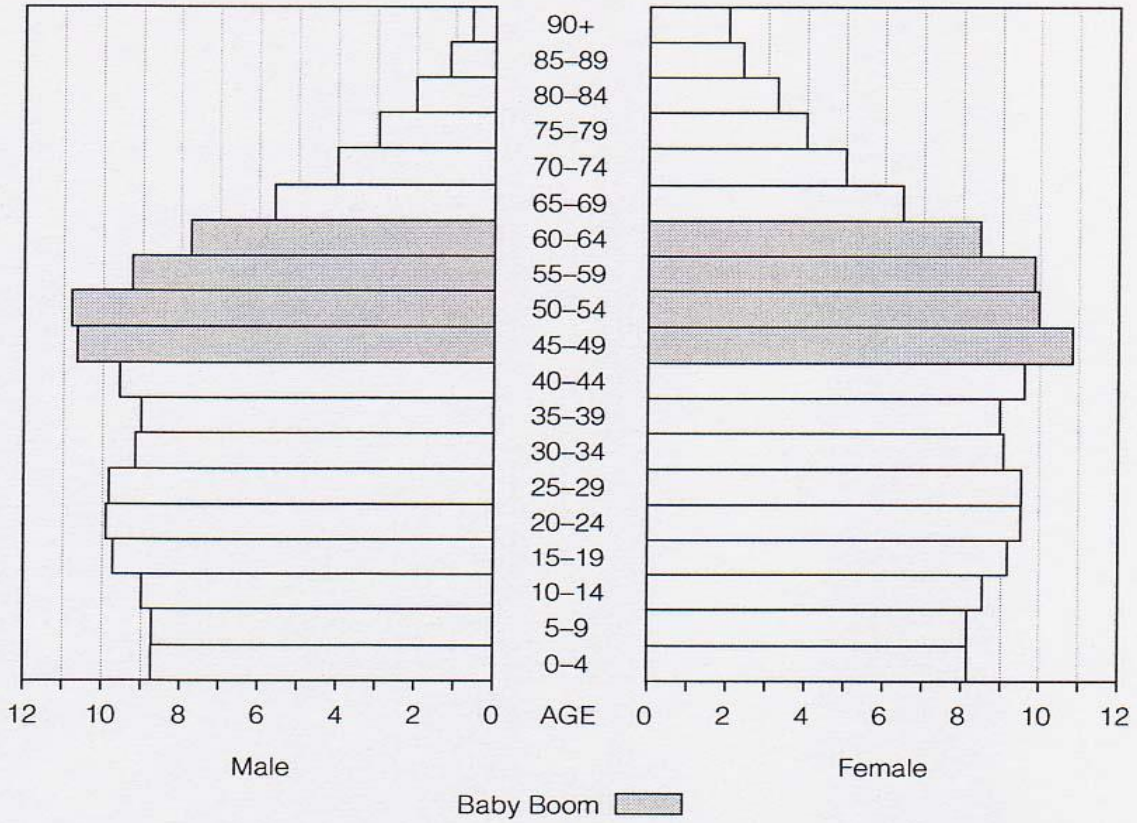


Source: U.S. Bureau of the Census, "1991 Estimates of the Population of States by Age, Sex, Race, and Hispanic Origin," PE-16.

Graphs – Histogram

The Histogram is a graph showing the differences in frequencies or percentages among categories of an **interval-ratio** variable. The categories are displayed as contiguous bars, with width proportional to the width of the category and height proportional to the frequency or percentage of that category.

Figure 3.11 **U.S. Population by Gender and Age, 2010 (in millions)**



Source: U.S. Bureau of the Census, *Current Population Reports*, 1992, P23-178.

Graphs – Histogram

A bar chart in which each contiguous bar represents a class:

1. the base is proportional to the class width
2. the area is proportional to the relative frequency, rf_i
3. ➔ the height is given by the *density*, h_i

Appropriate for: Quantitative Continuous (in class)

Steps:

1) Compute the relative frequency of each class, $rf_i = \frac{f_i}{n}$
2) Compute the width of each class, $W_i = upper\ limit\ - lower\ limit$
3) Derive the density, as $h_i = \frac{rf_i}{W_i}$

Graphs – Histogram: Example

Sample of 400 households, by weekly fuel expenses.

Weekly fuel expenses (in €)	Absolute Frequency fi
0|-20	20
20|-50	80
50|-100	210
100|-200	50
200|-300	25
300|-350	15
Total	400

Graphs – Histogram: Example

Sample of 400 households, by weekly fuel expenses.

Weekly fuel expenses (in €)	Absolute Frequency fi	Relative Frequency rfi	Width Wi	Density hi= rfi/Wi
0|-20	20	20/400=	20-0=	0.05/20 =
20|-50	80			
50|-100	210			
100|-200	50			
200|-300	25			
300|-350	15			
Total	400			

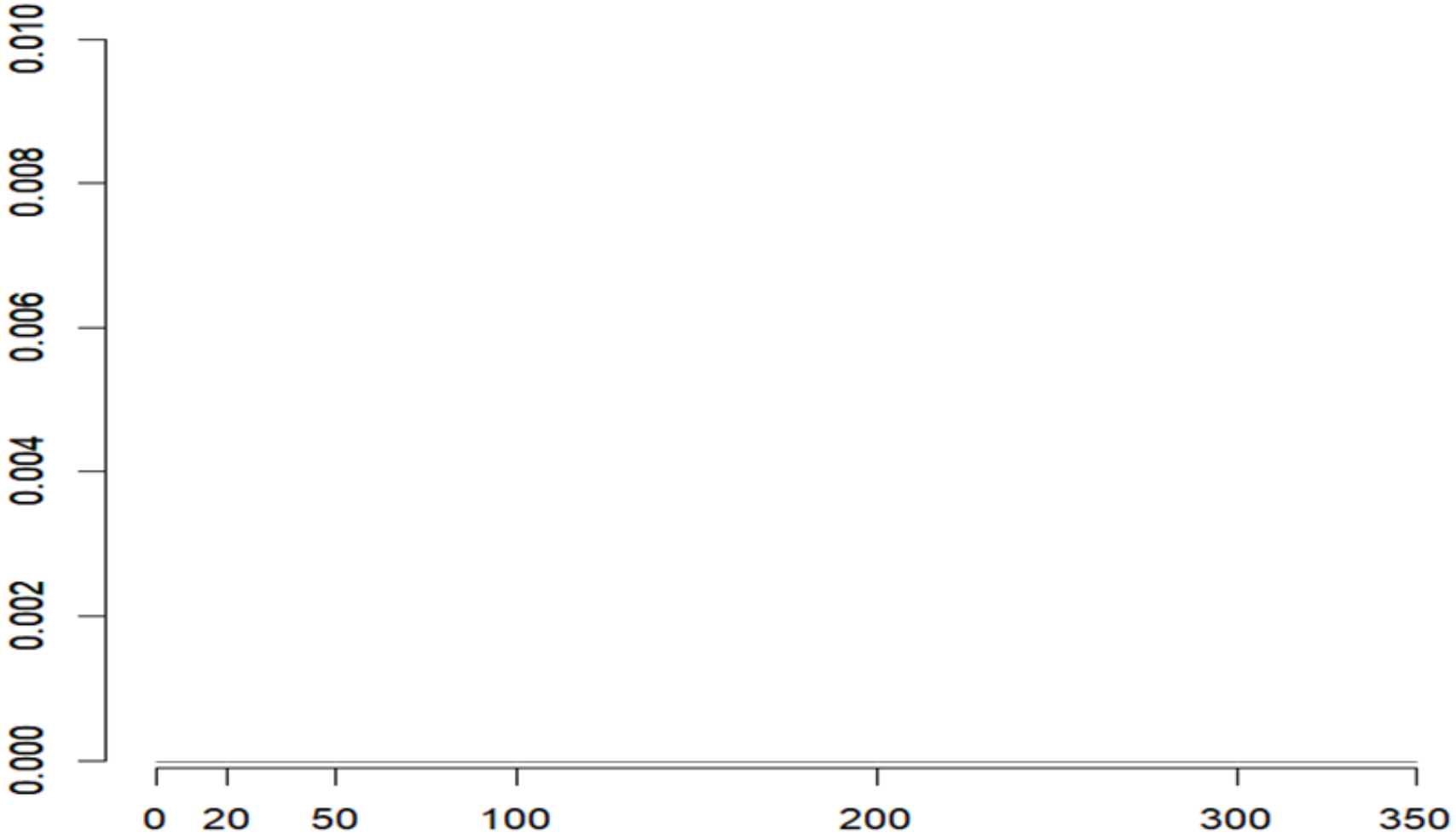
Graphs – Histogram: Example

Sample of 400 households, by weekly fuel expenses.

Weekly fuel expenses (in €)	Absolute Frequency fi	Relative Frequency rfi	Width Wi	Density hi
0|-20	20	20/400= 0.0500	20-0= 20	0.0025
20|-50	80	80/400= 0.2000	50-20= 30	0.0067
50|-100	210	210/400= 0.5250	100-50 50	0.0105
100|-200	50	50/400= 0.1250	200-100 100	0.0013
200|-300	25	25/400= 0.0625	300-200= 100	0.0006
300|-350	15	15/400= 0.0375	350-300= 50	0.0008
Total	400	1.00		

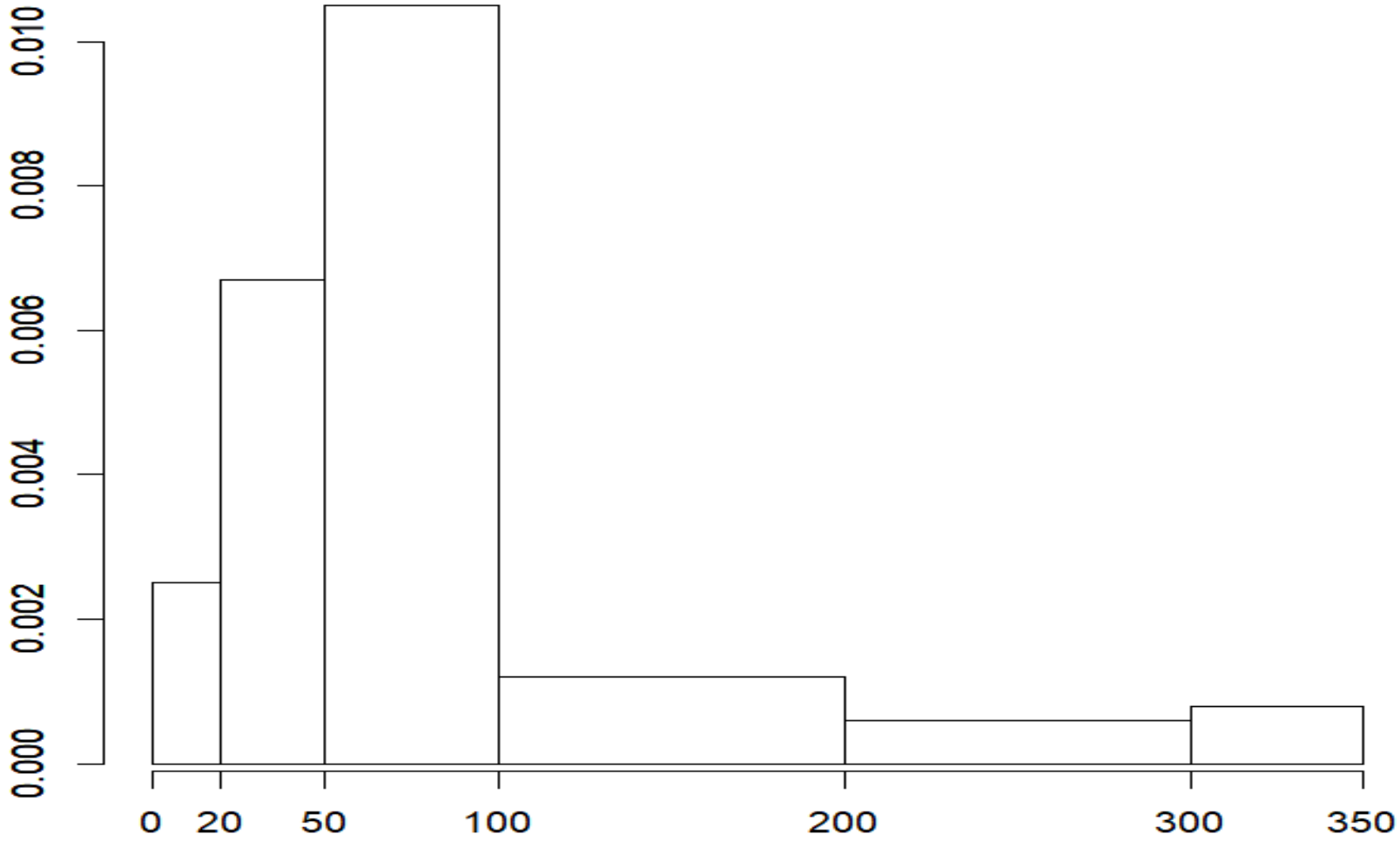
Graphs – Histogram: Example

Sample of 400 households, by weekly fuel expenses.



Graphs – Histogram: Example

Sample of 400 households, by weekly fuel expenses.



Graphs – Histogram: another example

Draw the histogram for the following distribution:

Weekly food expenses (in €)	Absolute Frequency fi
0|-10	20
10|-50	120
50|-80	90
80|-100	20
Total	250

Graphs – Histogram: another example

Steps: derive the relative frequency, the width and the density

Weekly food expenses (in €)	Absolute Frequency fi	Relative Frequency rfi	Width Wi	Density hi
0|-10	20			
10|-50	120			
50|-80	90			
80|-100	20			
Total	250			

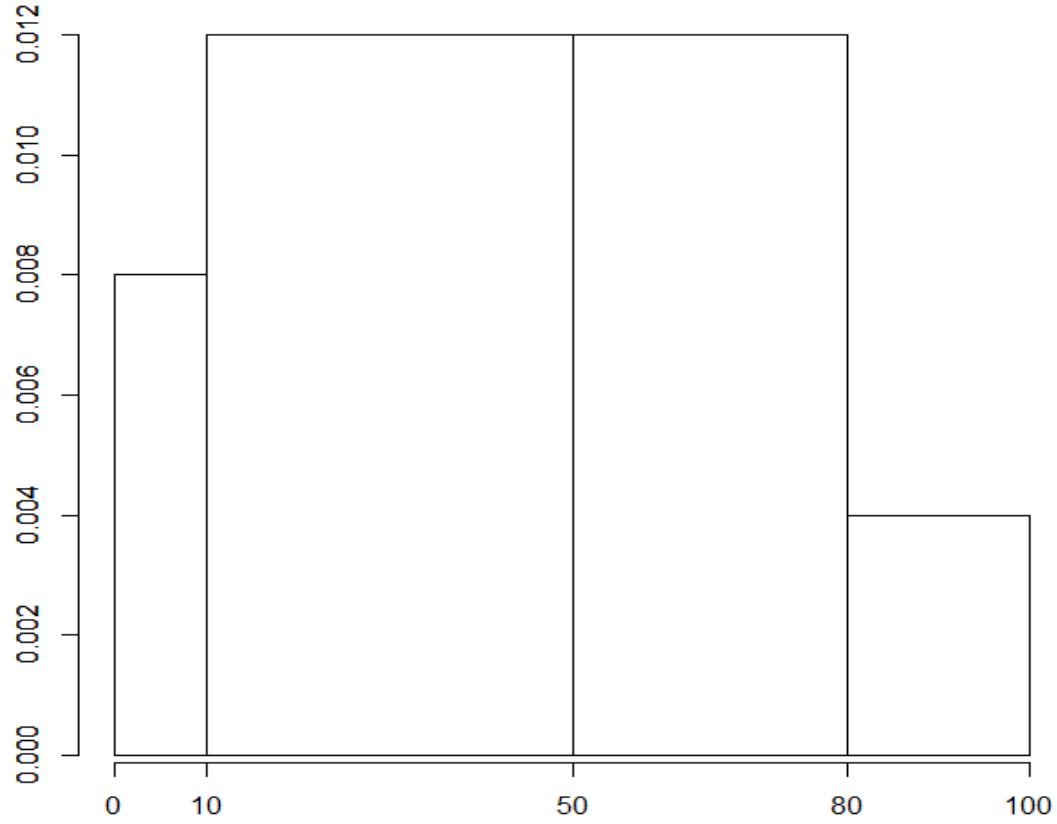
Graphs – Histogram: another example

Steps: derive the relative frequency, the width and the density

Weekly food expenses (in €)	Absolute Frequency fi	Relative Frequency rfi	Width Wi	Density hi
0|-10	20	0.0800	10	0.0080
10|-50	120	0.4800	40	0.0120
50|-80	90	0.3600	30	0.0120
80|-100	20	0.0800	20	0.0040
Total	250	1.00		

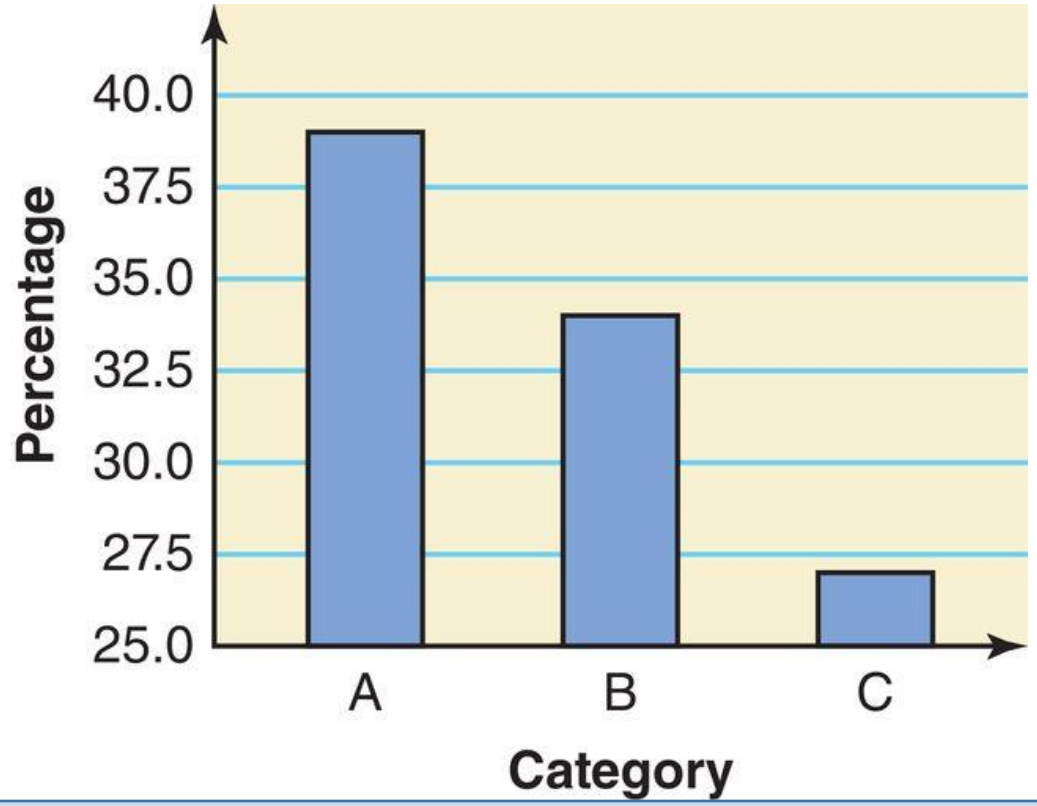
Graphs – Histogram: Example

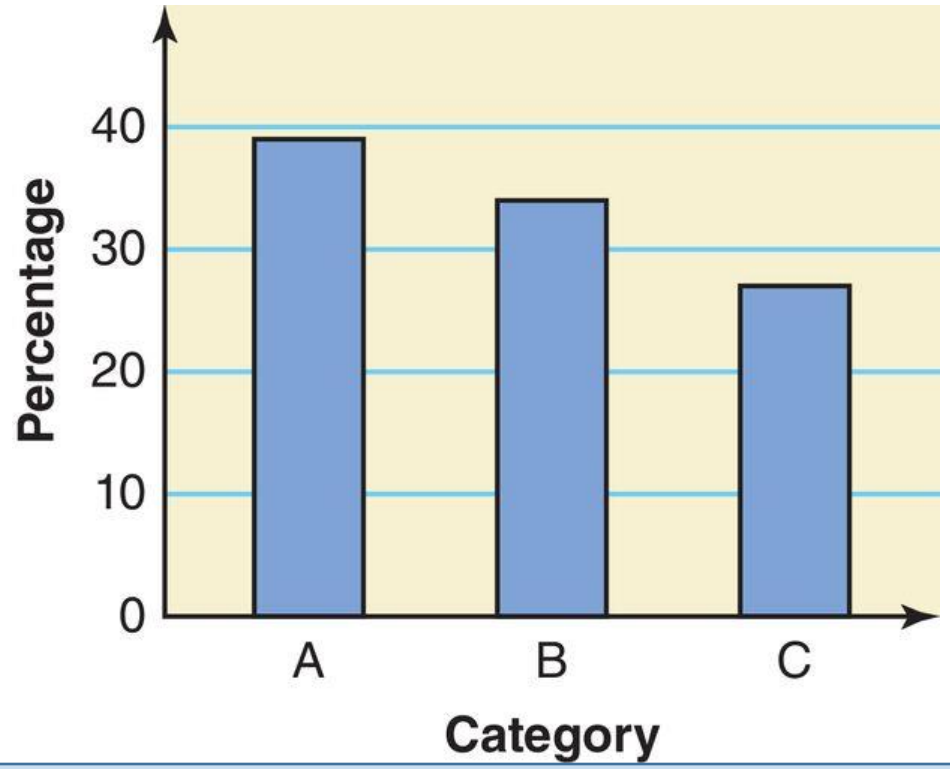
Sample of 250 individuals, by weekly food expenses.



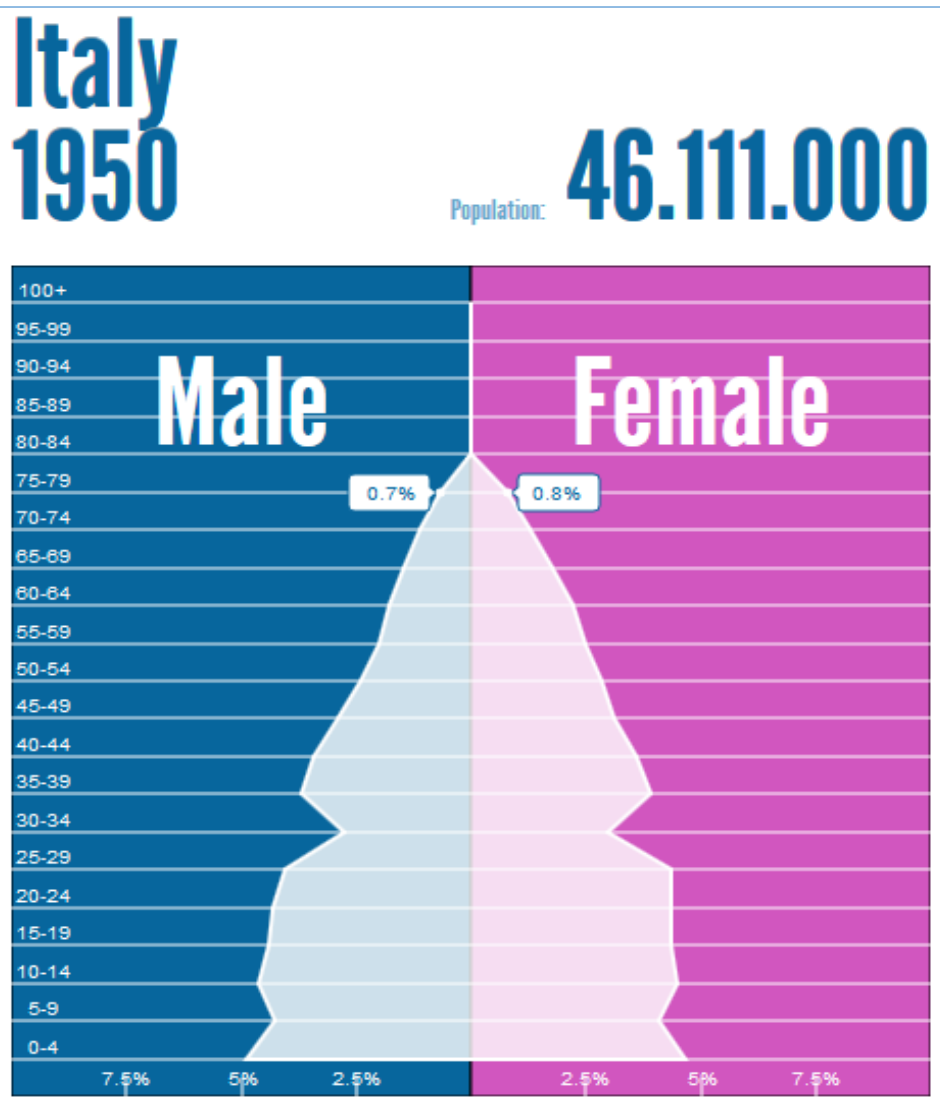
Graphs – Warnings

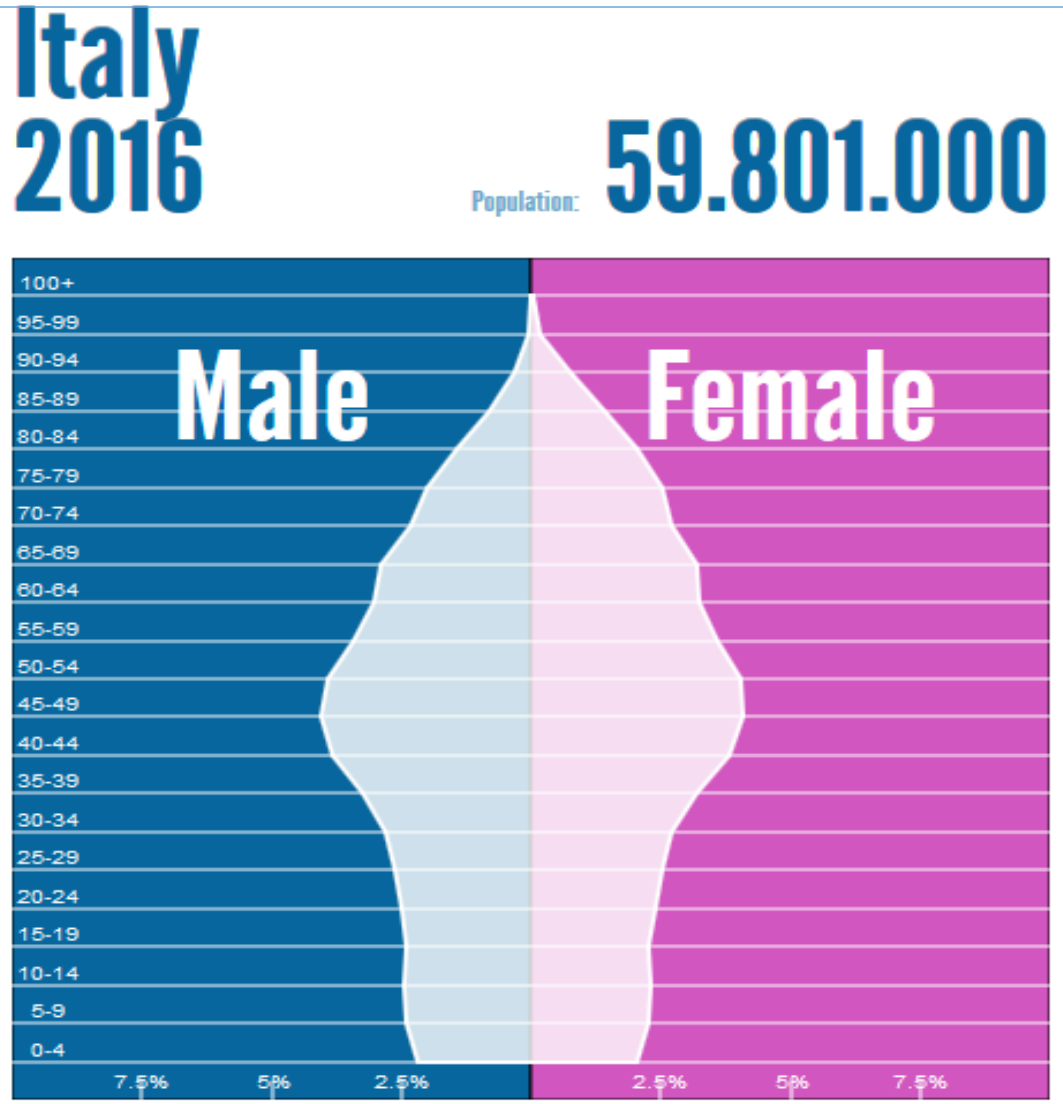
1. Use the most appropriate type of graph
2. Always apply labels to the axes
3. Pay attention to the scales!!!





Exploring data: bar graph for age-classes



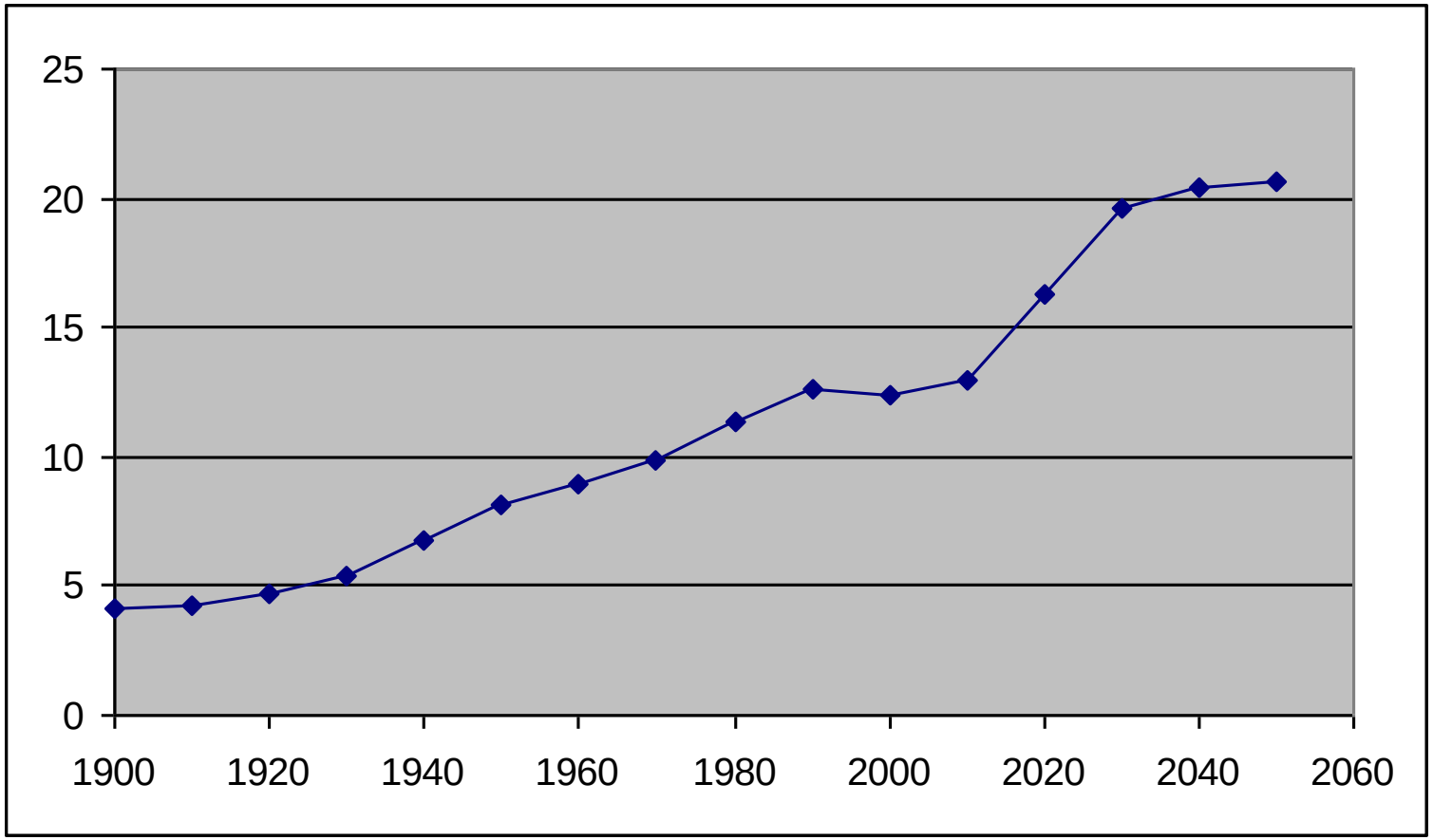


Time Series Charts

Time series chart: a graph displaying **changes** in a variables at **different points in time**. It shows time (measured in units such as years or months) on the horizontal axis and the frequencies (percentages or rates) of another variable on the vertical axis.

Time Series Charts

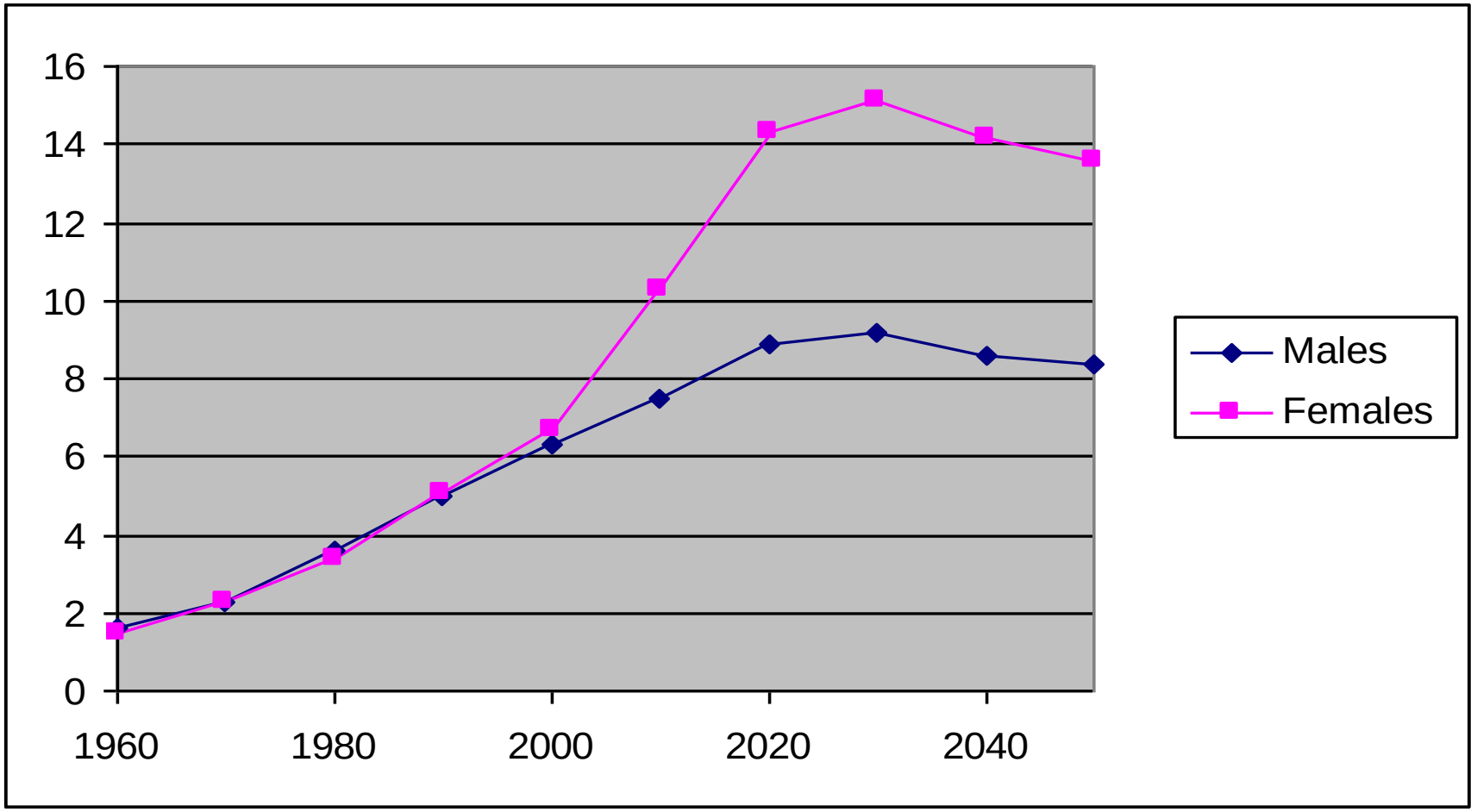
Figure 3.12 Percentage of Total U. S. Population 65 Years and Over, 1900 to 2050



Source: Federal Interagency Forum on Aging Related Statistics, *Older Americans 2004: Key Indicators of Well Being*, 2004.

Time Series Charts

Figure 3.13 Percentage Currently Divorced Among U.S. Population 65 Years and Over, by Gender, 1960 to 2040



Source: U.S. Bureau of the Census, "65+ in America," Current Population Reports, 1996, Special Studies, P23-190, Table 6-1.

Distortions in Graphs

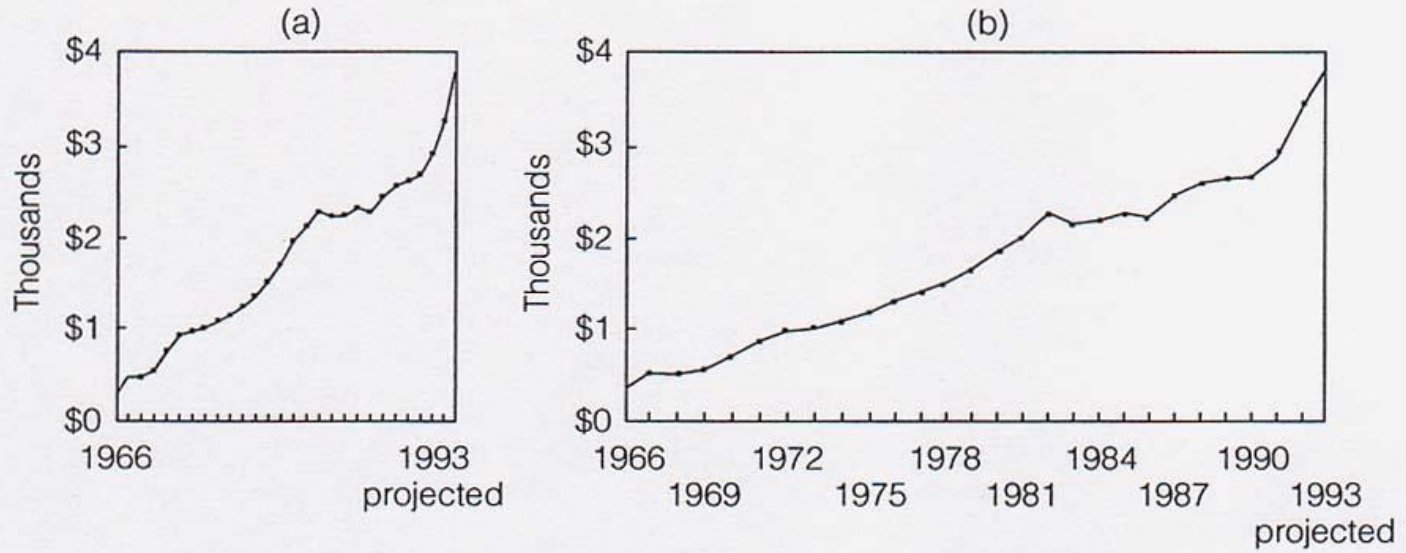
Graphs not only quickly inform us; they can quickly **deceive** us.

Because we are often more interested in general impressions than in detailed analyses of the numbers, we are **more vulnerable** to being swayed by **distorted graphs**.

Distortions in Graphs

Shrinking an Stretching the Axes: Visual Confusion

Figure 3.16 **Cost per Child Enrolled in Head Start Program**

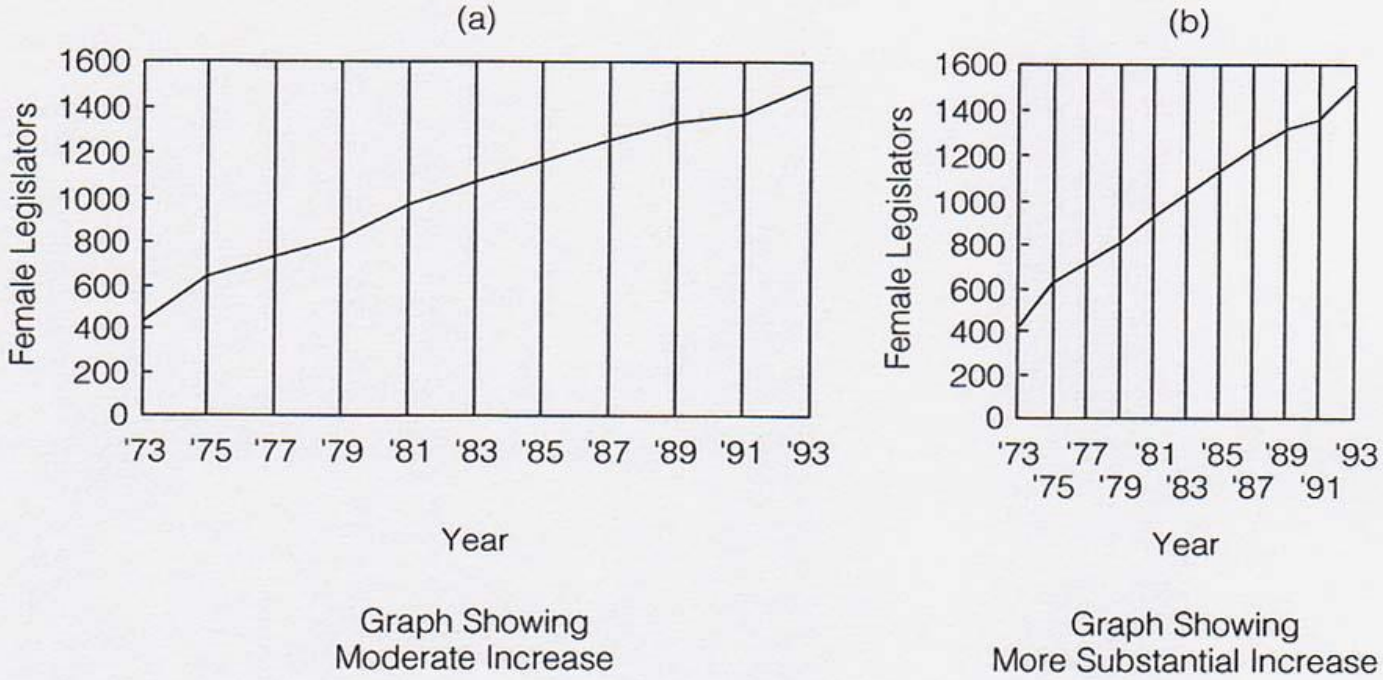


Source: Adapted from *USA TODAY*, March 15, 1993. Copyright 1993 USA TODAY. Reprinted with permission.

Distortions in Graphs

Shrinking an Stretching the Axes: Visual Confusion

Figure 3.17 **Women in U.S. Legislatures, 1973 to 1993**



Graph Showing Moderate Increase

Graph Showing More Substantial Increase

Source: Adapted from Marty Baumann, *USA TODAY*, February 12, 1993. Copyright 1993 USA TODAY. Reprinted with permission.

Why use charts and graphs?

- **What do you lose?**
  - –ability to examine numeric detail offered by a table
  - –potentially the ability to see **additional** relationships within the data
  - –potentially **time**: often we get caught up in selecting colors and formatting charts when a simply formatted table is sufficient

- **What do you gain?**
  - –ability to **direct readers' attention** to one aspect of the evidence
  - –ability to **reach readers** who might otherwise be intimidated by the same data in a tabular format
  - –ability to focus on **bigger picture** rather than perhaps minor technical details

Shapes of distributions

The most common shapes of the distributions are:

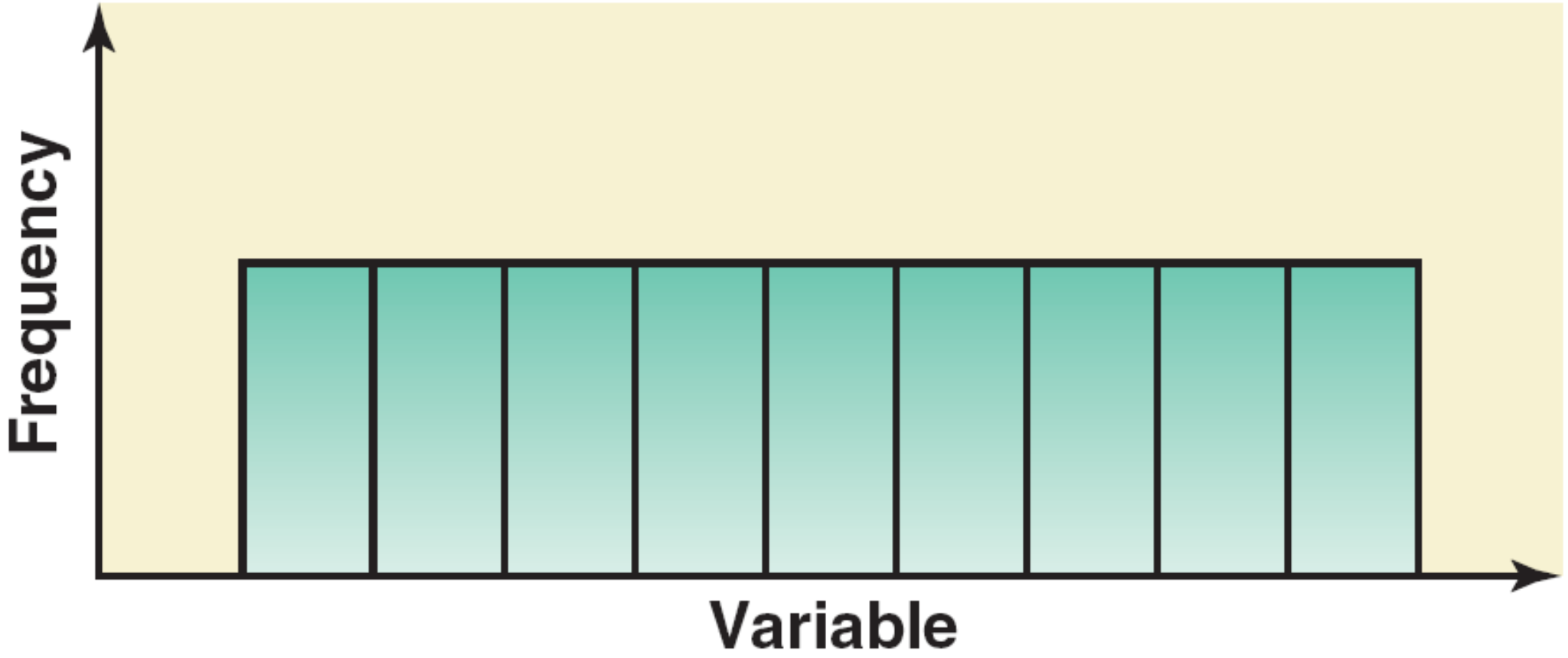
1) Uniform

2) Symmetric (but not uniform)

3) Skewed (to the left, to the right)

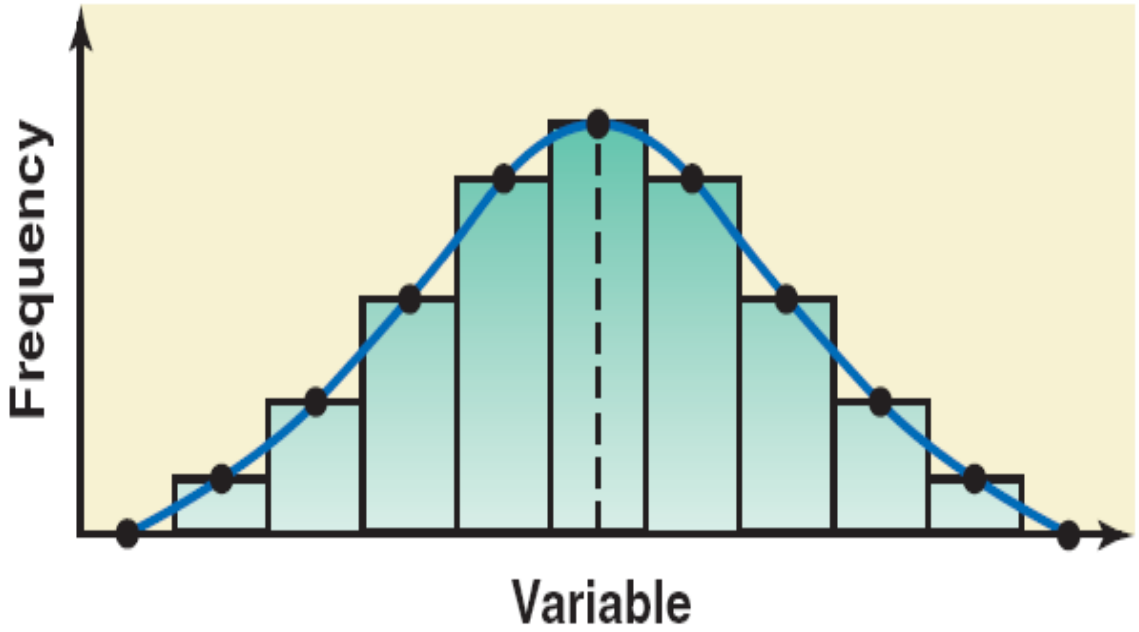
Shapes of distributions: Uniform

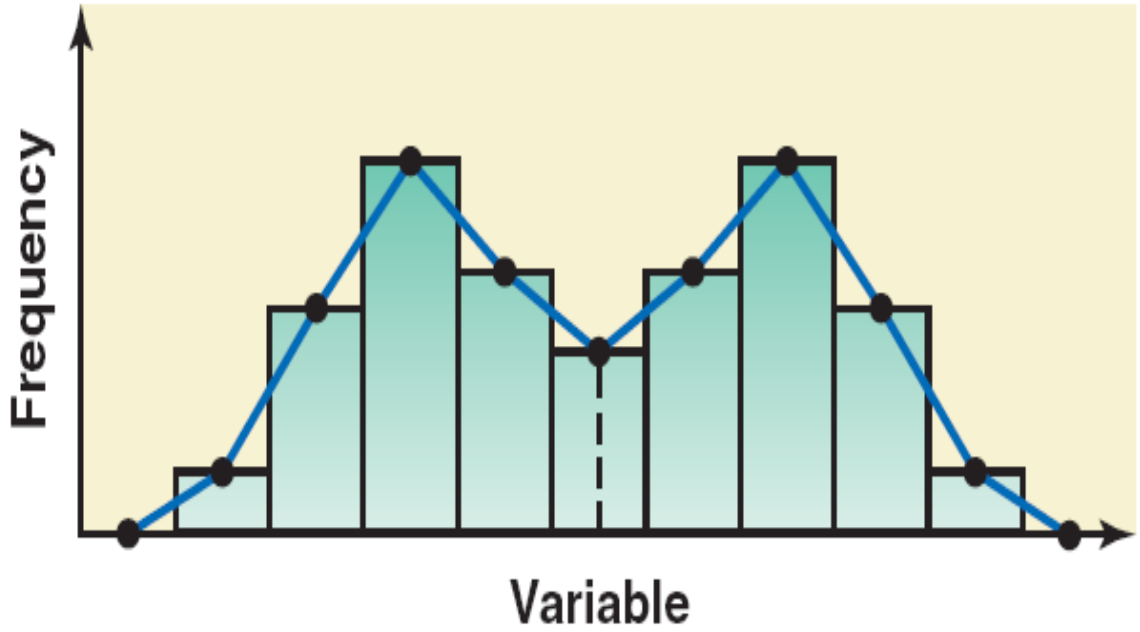
Uniform or Rectangular: same frequency for each category/value/class



Shapes of distributions: Symmetric

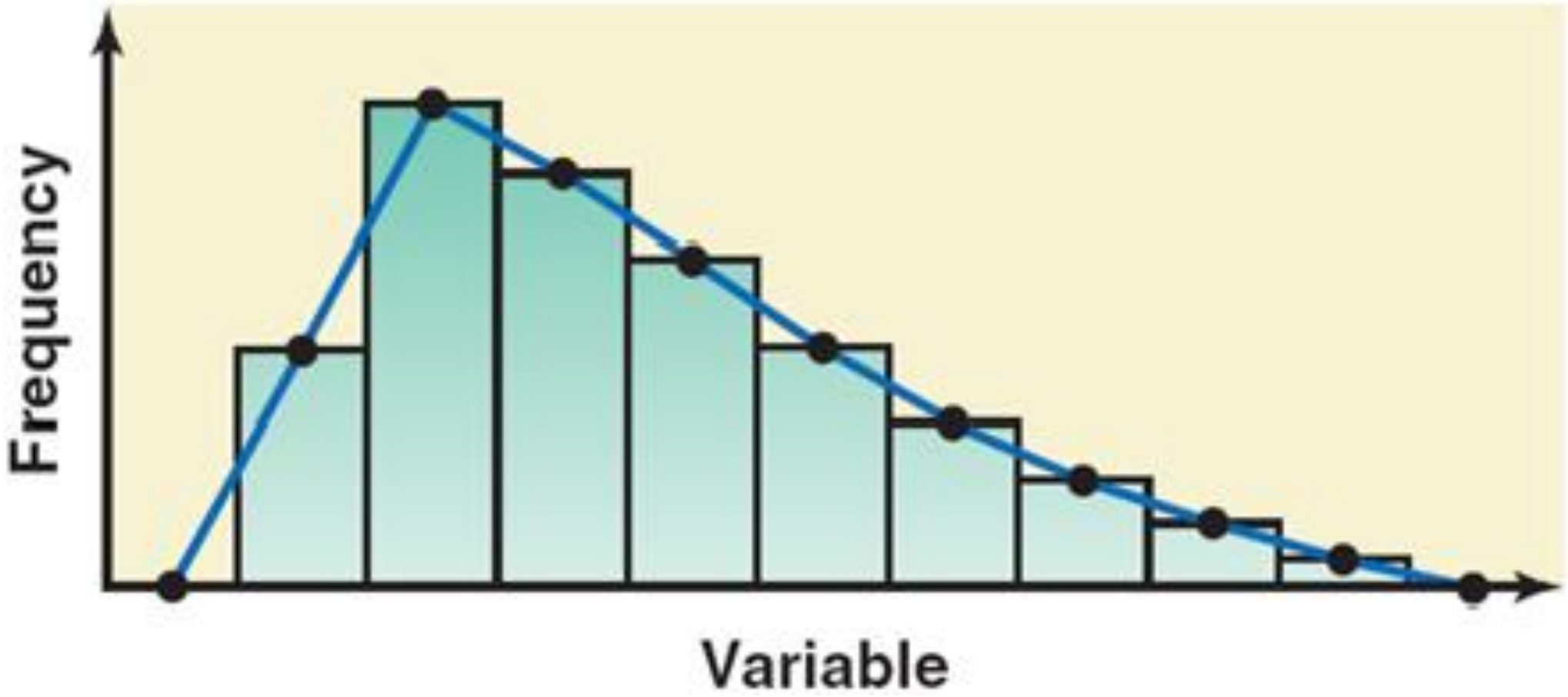
Identical on both sides of its central point





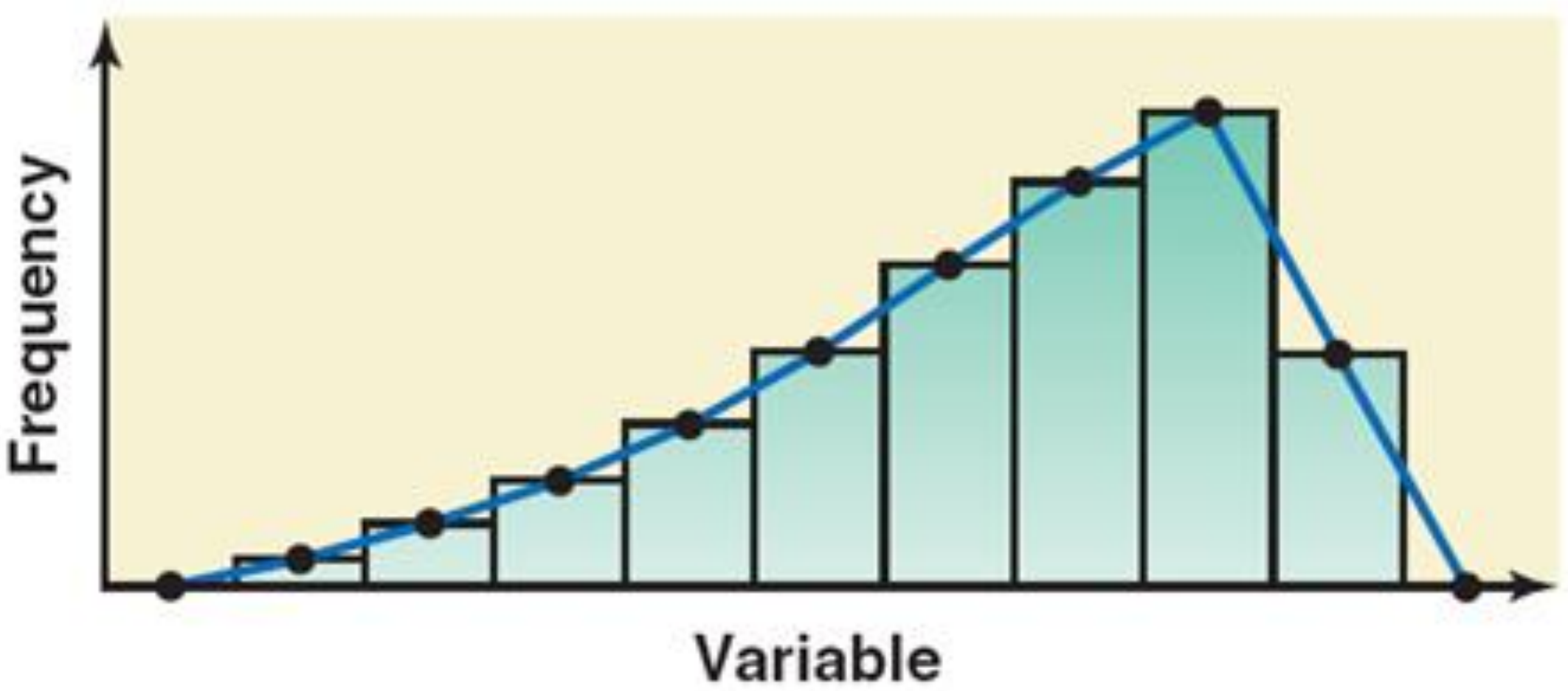
Shapes of distributions: Skewed (right)

Non symmetric, with the right tail longer than the left one



Shapes of distributions: Skewed (left)

Non symmetric, with the left tail longer than the right one



How do you expect the distribution of income?

How do you expect the distribution of income?