

# The Basic Practice of Statistics Ninth Edition

David S. Moore    William I. Notz

## Chapter 1 Picturing Distributions with Graphs

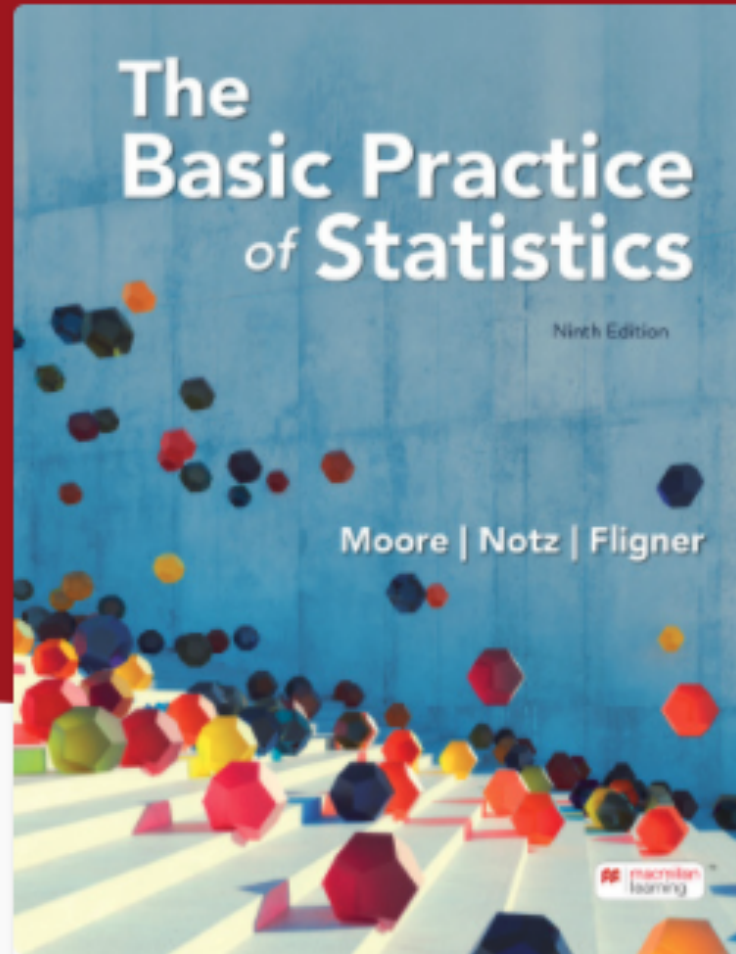
### Lecture Slides

# The Basic Practice of Statistics

Ninth Edition

Moore | Notz | Fligner

 macmillan  
learning



# In Chapter 1 we cover ...

- ❑ Individuals and variables
- ❑ Categorical variables: pie charts and bar graphs
- ❑ Quantitative variables: histograms
- ❑ Interpreting histograms
- ❑ Quantitative variables: stemplots
- ❑ Time plots

# Data Science

- ❓ Data science is a multidisciplinary field ( **Statistics**, Computer Science, Mathematics) with the goals of extracting insight/information from data and making data-driven decisions.
- ❓ Data Science encompasses: data collection, storage, preprocessing, **analysis** and visualisation.

# Statistics

**Statistics** is the science of data. The first step in dealing with data is to organize your thinking about the data.

---

**Individual:** an object described by a set of data

**Variable:** a characteristic of the individual

---

Individuals

	A	B	C	D	E	F
1	PatientID	Sex	Age	T1Diabetes	CHD	Income
2	125	1	45	2	1	60
3	58	1	40	2	2	48
4	148	2	52	1	2	98
5	128	2	56	1	2	100

Variables

# When planning a study ...

... or simply exploring data from someone else's work, ask yourself these questions:

<b>Who?</b>	What <i>individual</i> does the data describe?
<b>What?</b>	How many and what are the exact definitions of the <i>variables</i> in the data? In what unit of measurement is each variable recorded?
<b>Where?</b>	The context of the data collection is always important.
<b>When?</b>	(see previous point.)
<b>Why?</b>	Were the data collected to describe just those individuals or to represent a larger group?

# Types of Variables

A **categorical variable** places individuals into one of several groups or categories.

A **quantitative variable** takes numerical values for which arithmetic operations make sense (usually recorded in a *unit of measurement*).

	A	B	C	D	E	F
1	PatientID	Sex	Age	T1Diabetes	CHD	Income
2	125	1	45	2	1	60
3	58	1	40	2	2	48
4	148	2	52	1	2	98
5	128	2	56	1	2	100

Most data tables follow this format—the data here appear in a **spreadsheet** program.

# Exploratory Data Analysis

An **exploratory data analysis** is the process of using statistical tools and ideas to examine data in order to describe their main features.

---

## EXPLORING DATA

- ? Begin by examining each variable by itself. Then move on to studying the relationships among the variables.
  - ? Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.
-



# Distribution of a Variable

To examine a single variable, we usually want to display its **distribution**.


---


## DISTRIBUTION OF A VARIABLE

- ? The **distribution of a variable** tells us what values it takes and how often it takes these values.
  - ? The values of a categorical variable are labels for the categories. The **distribution of a categorical variable** lists the categories and gives either the count or the percent of individuals who fall in each category.
-

# Displaying Categorical Data

The distribution of a categorical variable lists the categories and gives either the count or the percent of individuals who fall into each category.

 **Pie charts** show the distribution of a categorical variable as a “pie” where the sizes of the slices reflect the counts or percents for the categories.

 **Bar graphs** represent each category as a bar whose height shows the category count or percent.

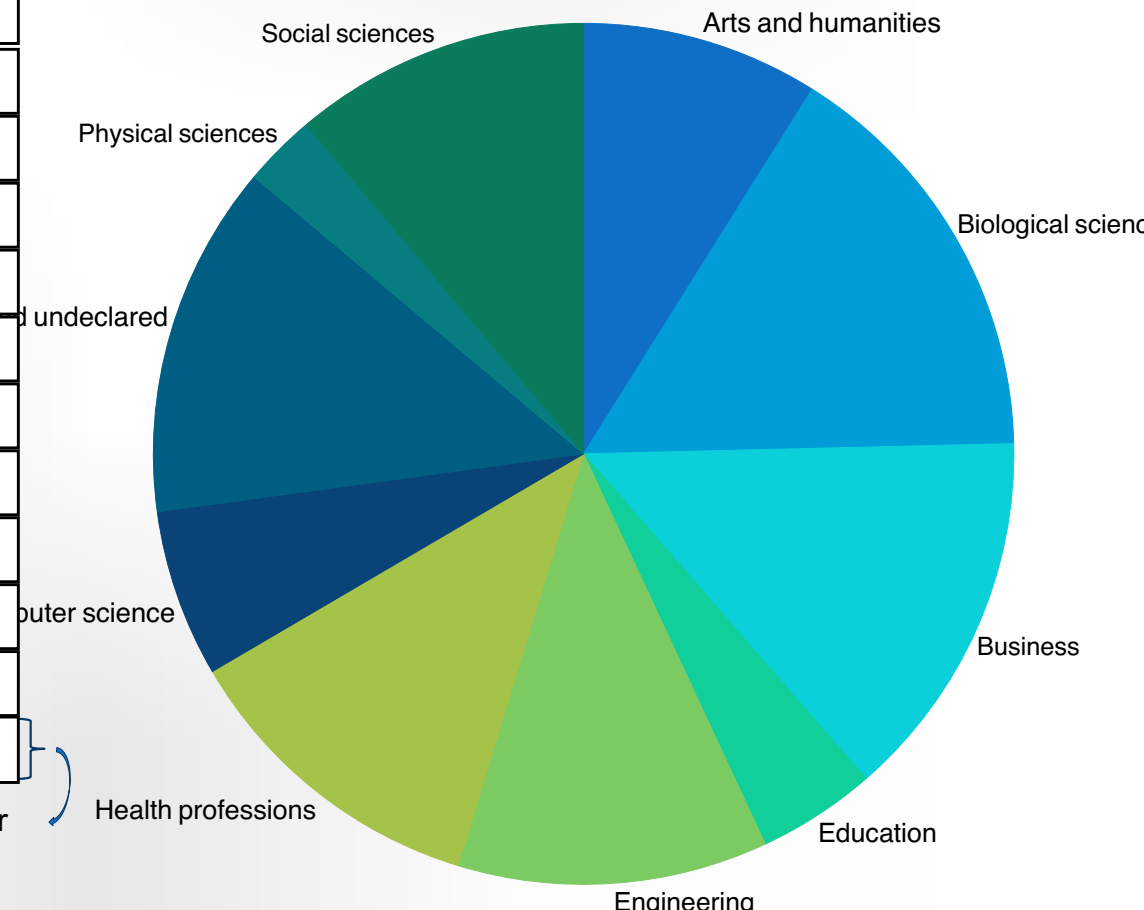
# Pie Chart

EXAMPLE: What do the 1.5 million full-time first-year students plan to study? Here are data on the percents of post-secondary first-year students who plan to major in several discipline areas.

Field of Study	Percent of Student
Biological sciences	15.5
Business	13.8
Health professions	11.7
Engineering	11.5
Social sciences	11
Arts and humanities	8.8
Math and computer science	6.2
Education	4.4
Physical sciences	2.7
Other majors and undeclared	13.1
Total	98.7

rounding error

PERCENT OF STUDENTS



# Lab tutorial – Jupyter Notebook

jupyter PieChart\_Chapter1 Last Checkpoint: a few seconds ago (autosaved)



Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted

Python 3



```
In [11]: import pandas as pd

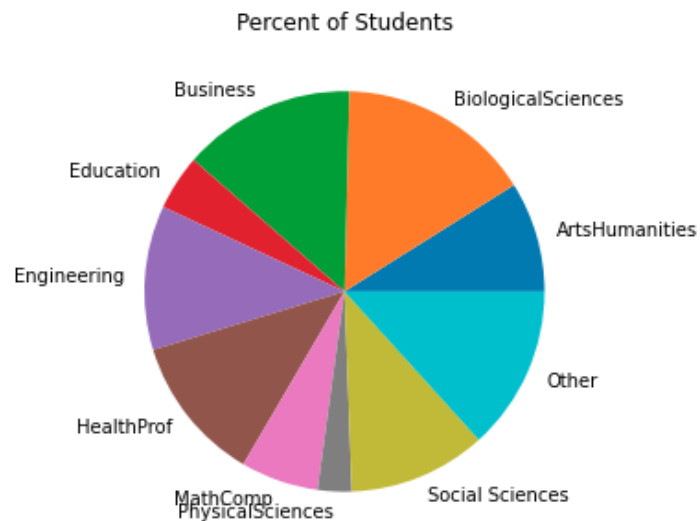
data = pd.read_csv (r'/Users/camiladesouza/OneDrive - The University of Western Ontario/DS1000/csv/chapter_01/eg01-02ma
print(data)
```

	Field of Study	Percent
0	ArtsHumanities	8.8
1	BiologicalSciences	15.5
2	Business	13.8
3	Education	4.4
4	Engineering	11.5
5	HealthProf	11.7
6	MathComp	6.2
7	PhysicalSciences	2.7
8	Social Sciences	11.0
9	Other	13.1

# Lab tutorial – Jupyter Notebook

```
In [20]: df = pd.DataFrame(data, columns = ['Field of Study', 'Percent'])
```

```
In [36]: import matplotlib.pyplot as plt
fig = df.plot.pie(y='Percent', figsize=(5,5),
                  labels=df['Field of Study'], legend=False, title="Percent of Students", ylabel='')
plt.show(fig)
```

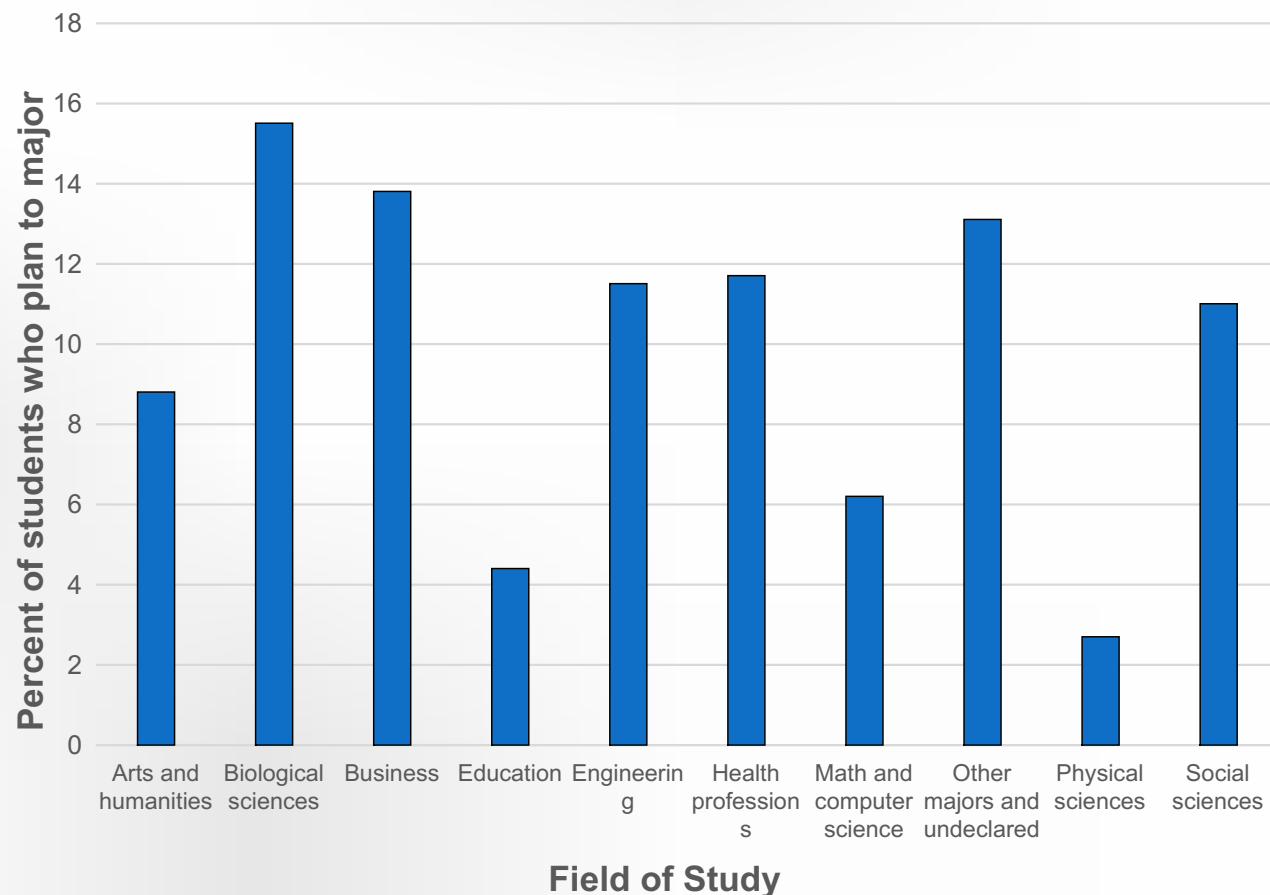


```
In [38]: fig.figure.savefig('piechart.pdf')
```

# Pie Charts or Bar Graphs

EXAMPLE (cont'd): Here are data on the percents of post-secondary first-year students who plan to major in several discipline areas, now alphabetized by field of study.

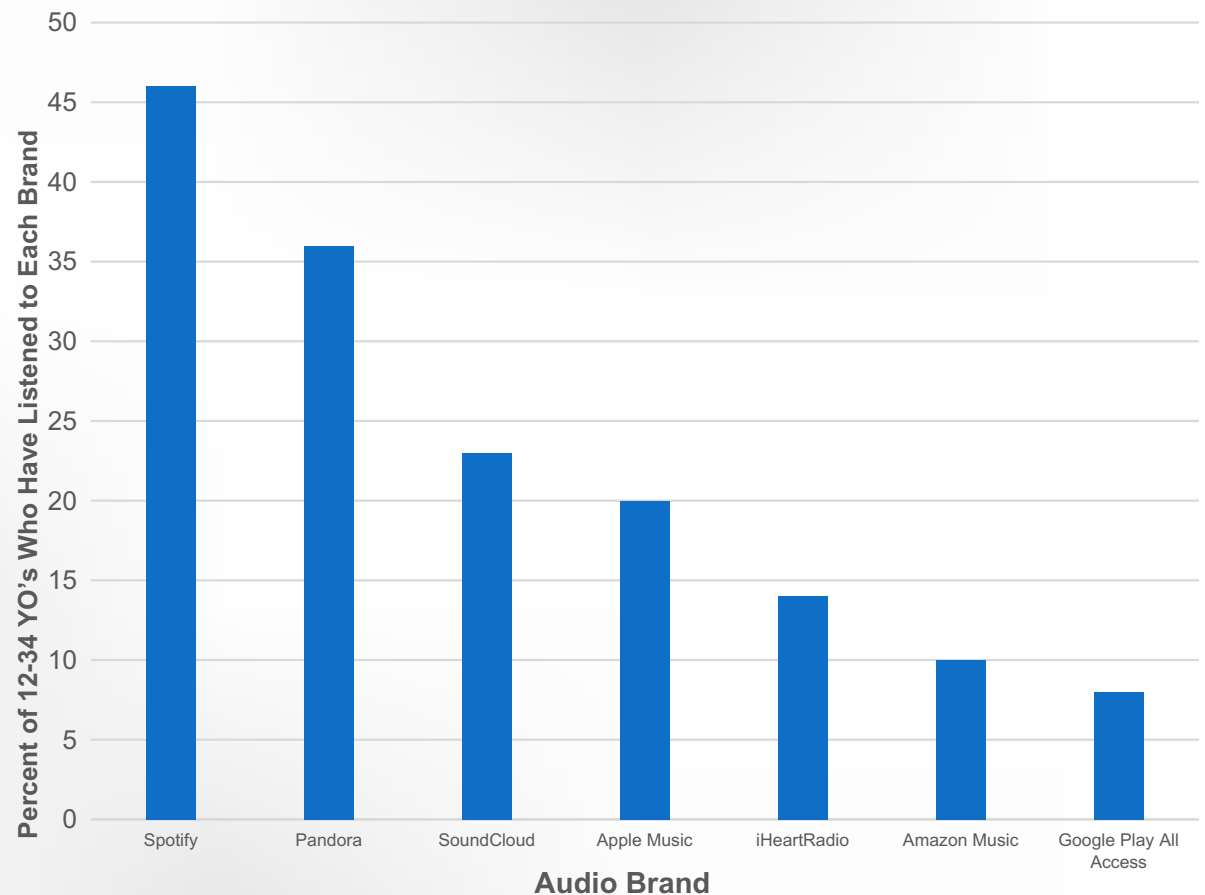
Field of Study	Percent of Student
Biological sciences	15.5
Business	13.8
Health professions	11.7
Engineering	11.5
Social sciences	11
Arts and humanities	8.8
Math and computer science	6.2
Education	4.4
Physical sciences	2.7
Other majors and undeclared	13.1



# Bar Graphs *Only*

EXAMPLE: What sources do Americans aged 12–34 years use to keep up to date and learn about music?

Brand	Percent of 12-34s Who Have Listened to Each Brand
Pandora	36
Spotify	46
iHeartRadio	14
Apple Music	20
Amazon Music	10
SoundCloud	23
Google Play All Access	8



Note: For bar graphs, percents don't ***necessarily*** add to 100.

# Quantitative Data

The distribution of a quantitative variable tells us what values the variable takes on and how often it takes on those values.

? **Histograms** show the distribution of a quantitative variable by using bars where the height of each bar represents the number of individuals who take on a value within a particular class/interval.

? **Stemplots** separate each observation into a stem and a leaf that are then plotted to display the distribution, while maintaining the original values of the variable.



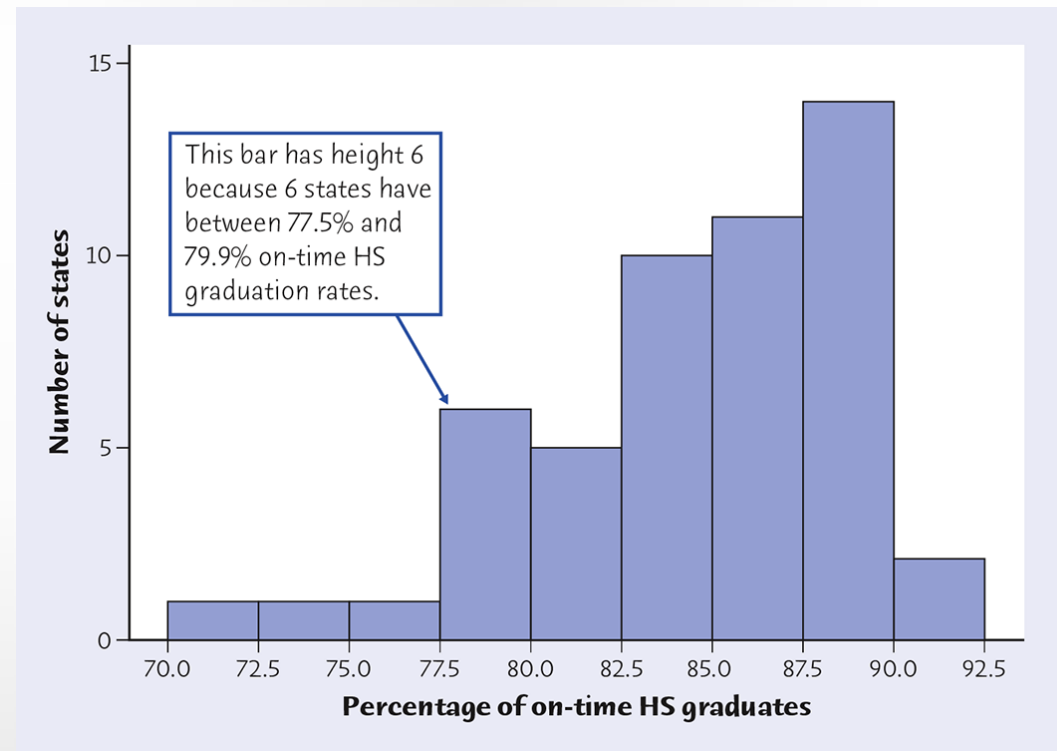
# Histograms (1 of 2)

- ? Are appropriate for quantitative variables that take on many values and/or for large datasets.
- ? Divide the possible values into classes/intervals (equal widths).
- ? Count how many observations fall into each interval (may change to percents).
- ? Draw a picture representing the distribution—bar heights are equivalent to the number (percent) of observations in each interval.

# Histograms (2 of 2)

**?** EXAMPLE: Freshman Graduation Rate, or FGR, Data for 2016–2017

Class	Count	Class	Count
70.0 to < 72.5	1	82.5 to < 85.0	10
72.5 to < 75.0	1	85.0 to < 87.5	11
75.0 to < 77.5	1	87.5 to < 90.0	14
77.5 to < 80.0	6	90.0 to < 92.5	2
80.0 to < 82.5	5		



# Interpreting Histograms

---

## EXAMINING A HISTOGRAM

- ❓ In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.
  - ❓ You can describe the overall pattern by its **shape**, **center**, and **variability**. You will sometimes see variability referred to as **spread**.
  - ❓ An important kind of deviation is an **outlier**, an individual that falls outside the overall pattern.
-

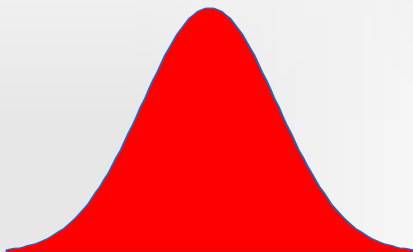
# Overall Shape of Distributions

---

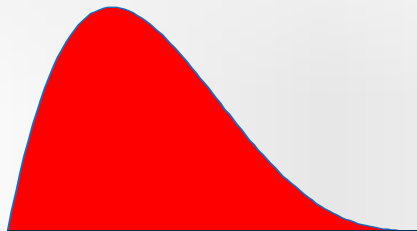
## SYMMETRIC AND SKEWED DISTRIBUTIONS

- ? A distribution is **symmetric** if the right and left sides of the graph are approximately mirror images of each other.
  - ? A distribution is **skewed to the right** (**right-skewed**) if the right side of the graph (containing the half of the observations with larger values) is much longer than the left side.
  - ? A distribution is **skewed to the left** (**left-skewed**) if the left side of the graph is much longer than the right side.
- 

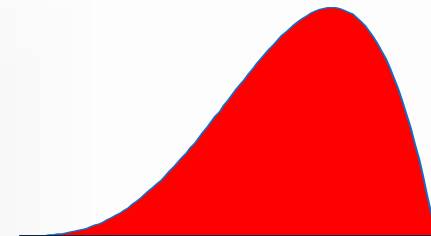
Symmetric



Right-skewed



Left-skewed



# Stemplots (Stem-and-Leaf Plots) (1 of 2)

---

To make a stemplot:

1. Separate each observation into a stem (consisting of all but the final, or rightmost, digit) and a leaf (the final digit). Stems may have as many digits as needed, but each leaf contains only a single digit.
  2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column. Be sure to include all the stems needed to span the data, even when some stems have no leaves.
  3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.
-

# Example of stemplot

5.3, 5.6, 5.3, 5.0, 5.5, 5.9, 5.9, 3.3, 5.0, 3.4, 5.0, 3.4, 4.6, 6.9, 6.0

- Stems are the non-decimals, leaves are decimals and ordered

# Example of stemplot

5.3, 5.6, 5.3, 5.0, 5.5, 5.9, 5.9, 3.3, 5.0, 3.4, 5.0, 3.4, 4.6, 6.9, 6.0

- Stems are the non-decimals, leaves are decimals and ordered

3.3, 3.4, 3.4, 4.6, 5.0, 5.0, 5.0, 5.3, 5.3, 5.5, 5.6, 5.9, 5.9, 6.0, 6.9

# Example of stemplot

5.3, 5.6, 5.3, 5.0, 5.5, 5.9, 5.9, 3.3, 5.0, 3.4, 5.0, 3.4, 4.6, 6.9, 6.0

- Stems are the non-decimals, leaves are decimals and ordered

3.3, 3.4, 3.4, 4.6, 5.0, 5.0, 5.0, 5.3, 5.3, 5.5, 5.6, 5.9, 5.9, 6.0, 6.9

**3 | 344**

**4 | 6**

**5 | 000335699**

**6 | 09**



# Stemplots (Stem-and-Leaf Plots) (2 of 2)

**?** *If there are very few stems* (that is, when the data cover only a very small range of values), then we may want to create more stems by splitting the original stems.

**?** **EXAMPLE:** 3.3, 3.4, 3.4, 4.6, 5.0, 5.0, 5.0, 5.3, 5.3, 5.5, 5.6, 5.9, 5.9, 6.0, 6.9

**3 | 344** **?** **decimals 0 to 4**

**3 |** **?** **decimals 5 to 9**

**4 |**

**4 | 6**

**5 | 00033**

**5 | 5699**

**6 | 0**

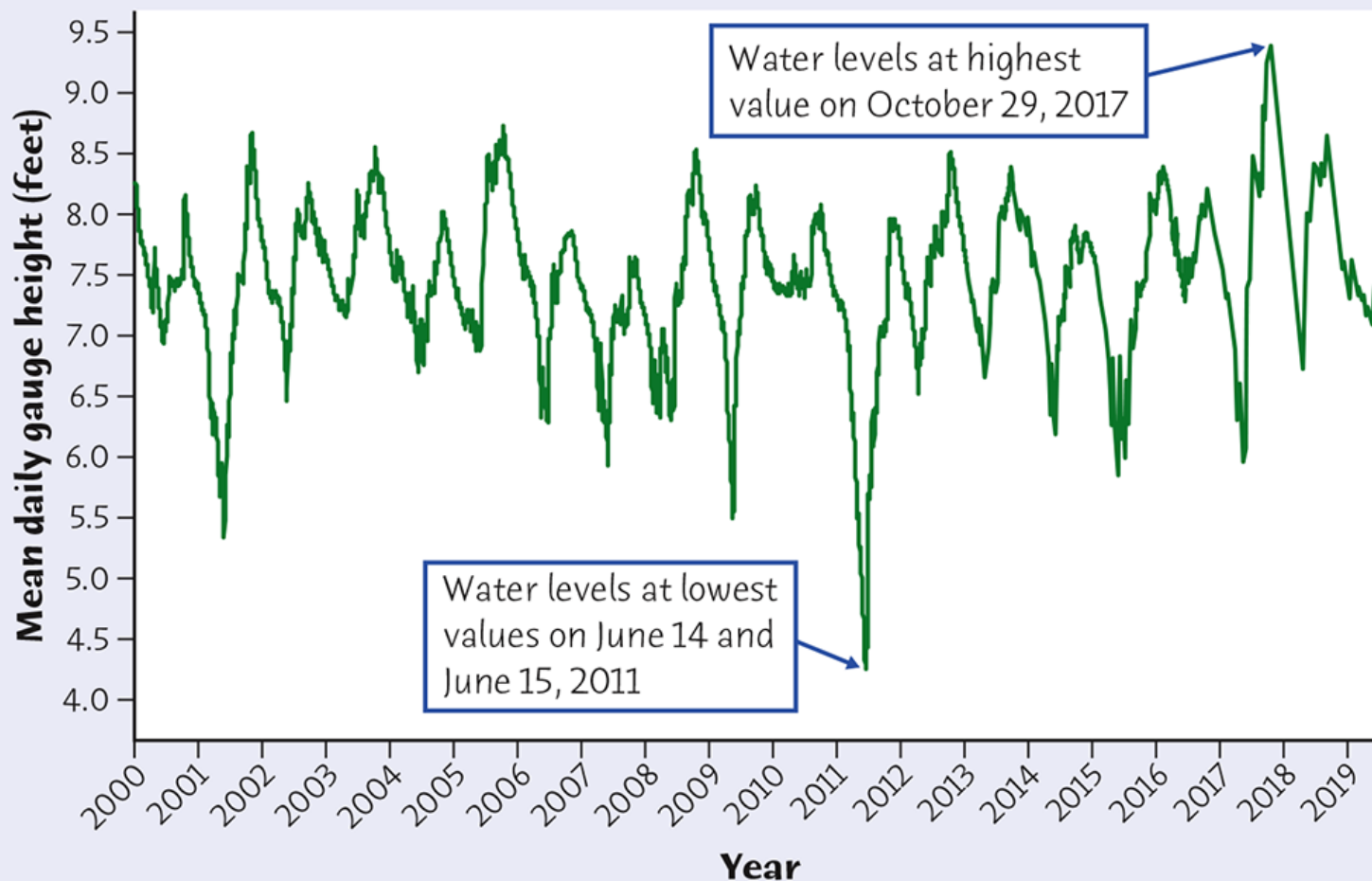
**6 | 9**

# Time Plots (1 of 3)

- ? A time plot shows behavior over time.
- ? Time is always on the horizontal axis, and the variable being measured is on the vertical axis.
- ? Look for an overall pattern (trend) and for deviations from this trend. Connecting the data points by lines may emphasize this trend.
- ? Look for patterns that repeat at known regular intervals (seasonal variations).

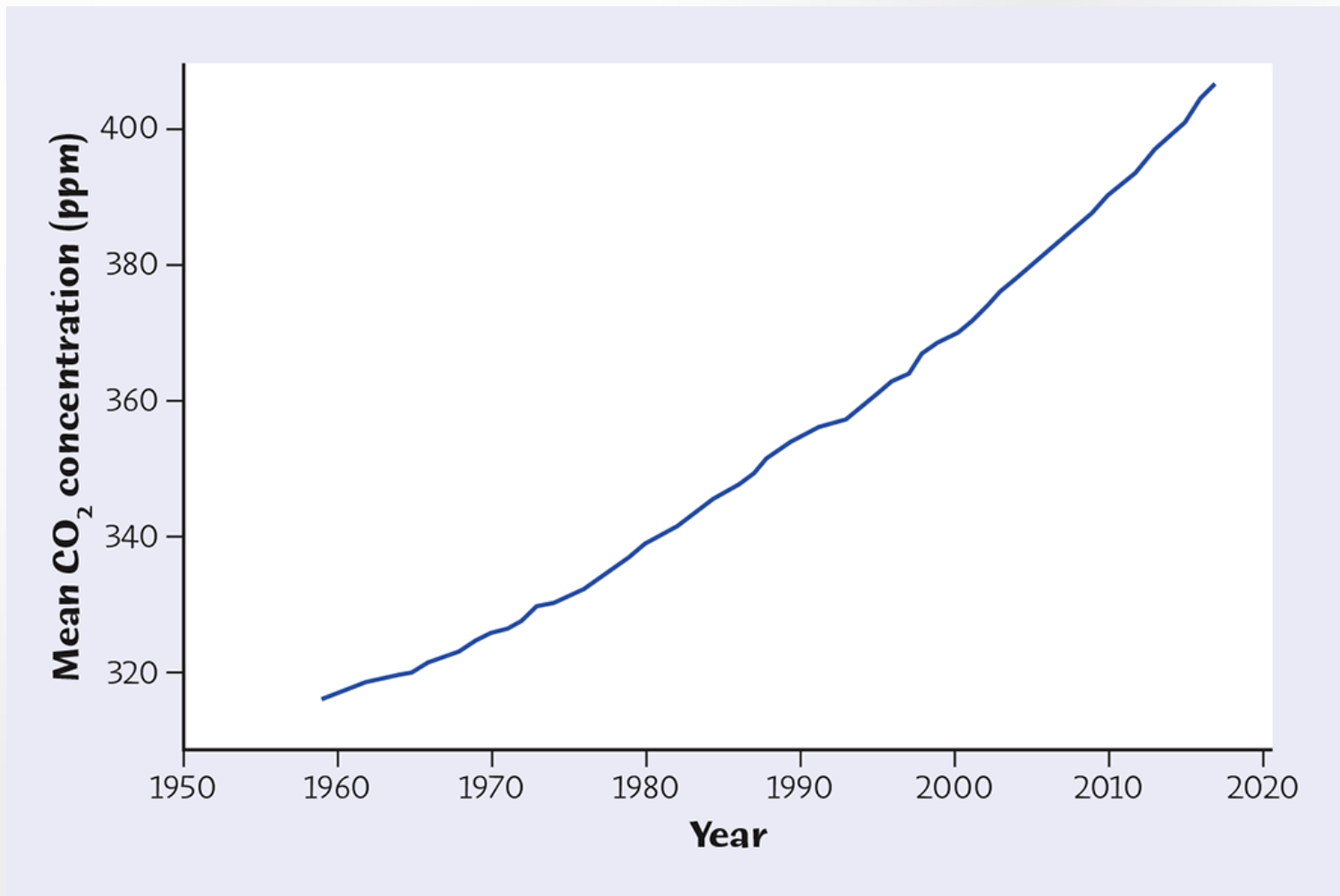
# Time Plots (2 of 3)

Time Plot of Average Gauge Height,  
Everglades National Park Monitoring Station



# Time Plots (3 of 3)

Time Plot of the Mean Atmospheric CO<sub>2</sub> Concentration (ppm)



Florence Nightingale a pioneer on picturing data: Joy of Stats  
with Hans Rosling

[https://www.youtube.com/watch?v=yhX0OR1\\_Vfc](https://www.youtube.com/watch?v=yhX0OR1_Vfc)