Chapter 1

Looking at Data—Distributions

Introduction to the Practice of STATISTICS EIGHTH EDITION

Moore / McCabe / Craig

Lecture Presentation Slides

Chapter 1
Looking at Data—Distributions

Introduction

1.1 Data

1.2 Displaying Distributions with Graphs

1.3 Describing Distributions with Numbers

1.4 Density Curves and Normal Distributions

1.1 Data

- Cases, labels, variables, and values
- Categorical and quantitative variables

What Is Statistics?

- **Statistics** is the science of learning from data.
- The first step in dealing with data is to organize your thinking about the data.
- An **exploratory data analysis** is the process of using statistical tools and ideas to examine data in to discover or describe their main features.

Cases, Labels, Variables, and Values

- ✓ **Cases** are the objects described by a set of data. Cases may be customers, companies, experimental subjects, or other objects.
- ✓ A **variable** is a special characteristic of a case.
- ✓ A **label** is used in some data sets to provide additional information about a variable.
- ✓ Different cases can have different **values** of a variable.

Categorical and Quantitative Variables

- A **categorical** variable places each case into one of several groups, or categories. Sometimes also referred to as a nominal variable (i.e., used for naming). Example: first language, ethnicity

- An ordinal variable organizes the cases in a particular order. Example: socio-economic status, highest educational attainment

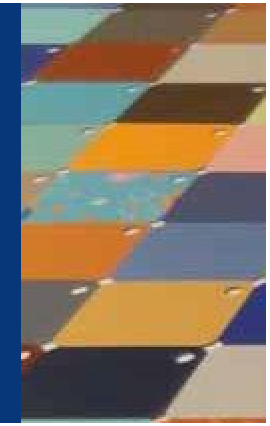
- A **quantitative** variable takes numerical values for which arithmetic operations such as adding and averaging make sense.

  - An interval variable is a variable where differences between numbers are meaningful. Example: IQ, GPA, degree Celsius

  - A ratio variable is a variable which has an absolute zero (i.e., zero is meaningful. Example: height, income, degree Kelvin

- The **distribution** of a variable tells us the values that a variable takes and how often it takes each value.

1.2 Displaying Distributions with Graphs



- Variables
- Examining distributions of variables
- Graphs for categorical variables
  - Bar graphs
  - Pie charts
- Graphs for quantitative variables
  - Histograms
  - Stemplots

Variables

Exploring Data

- Begin by examining each variable by itself. Then move on to study 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 graphically display its **distribution.**

- The distribution of a variable tells us what values it takes and how often it takes these values.
- Distributions can be displayed using a variety of graphical tools. The proper choice of graph depends on the nature of the variable.

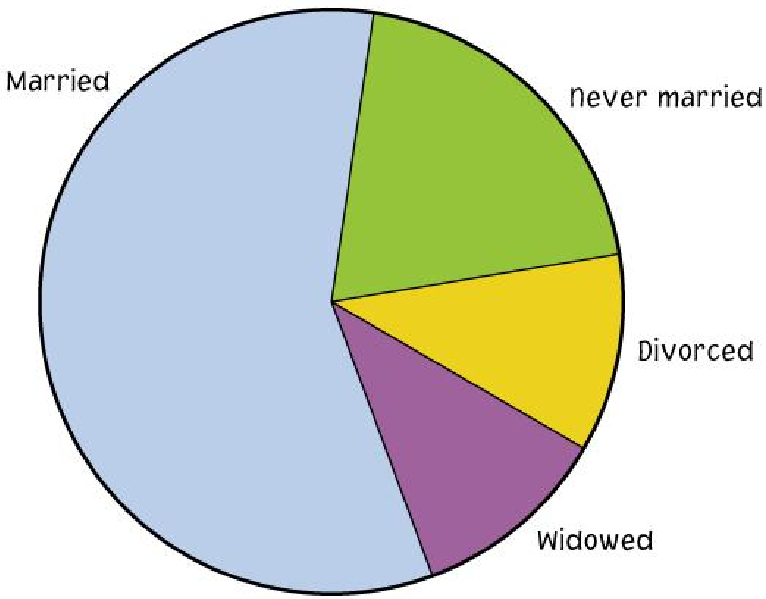
Categorical variable
Pie chart
Bar graph

Quantitative variable
Histogram
Stemplot

Categorical Variables

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

- **Pie charts** show the distribution of a categorical variable as a "pie" whose slices are sized by the counts or percents for the categories.
- **Bar graphs** represent categories as bars whose heights show the category counts or percents.



Quantitative Variables

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

- **Histograms** show the distribution of a quantitative variable by using bars. The height of a bar represents the number of individuals whose values fall within the corresponding class.
- **Stemplots** (or stem-and-leaf plots) 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.

Stemplots

Example: Weight Data — Introductory Statistics Class

192	110	195
152	120	170
135	185	120
110	165	185
128	212	175
180	119	203
260	165	185
170	210	123
165	186	139
150	100	106

Stems **Leaves**

10	
11	
12	
13	5
14	
15	2
16	
17	
18	
19	2
20	
21	
22	
23	
24	
25	
26	

Key

20|3 means 203 pounds

Stems = 10s
Leaves = 1s

Histograms

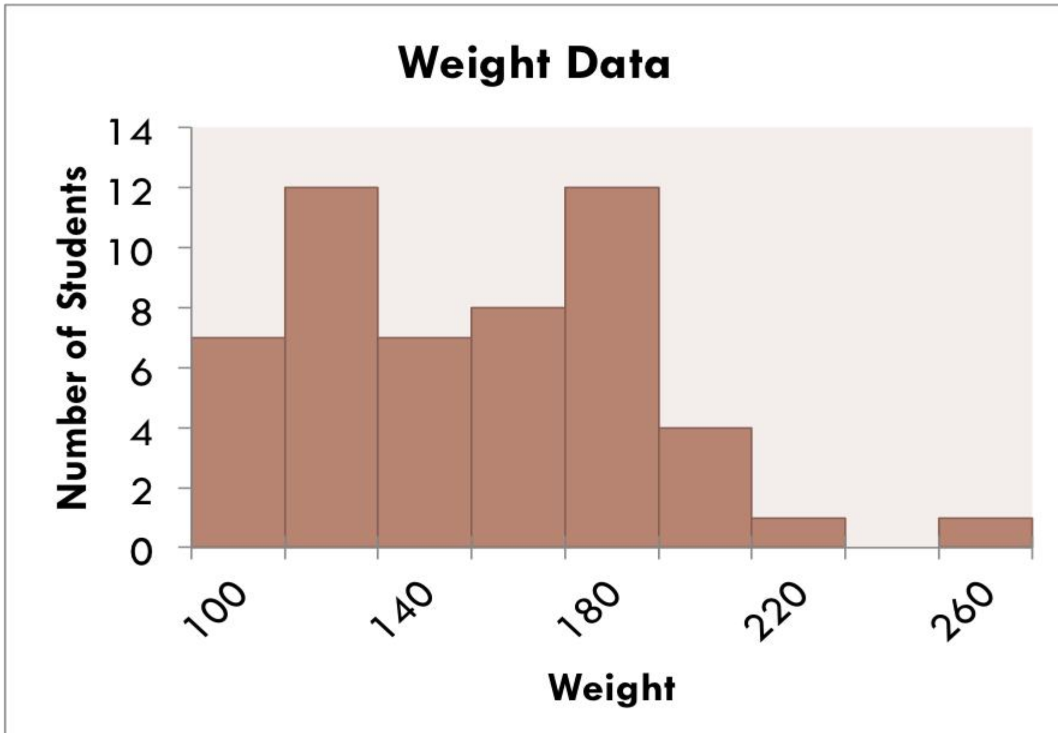
When quantitative variables take many values, a histogram may be more appropriate.

In a histogram, we:

- Divide the possible values into **classes,** or intervals of equal widths.
- Count how many observations fall into each interval. Instead of counts, one may also use percents.
- Draw a picture representing the distribution—each bar height is equal to the number (percent) of observations in its interval.

Histograms

Example: Weight Data—Introductory Statistics Class



Weight group	Count
100 - <120	7
120 - <140	12
140 - <160	7
160 - <180	8
180 - <200	12
200 - <220	4
220 - <240	1
240 - <260	0
260 - <280	1

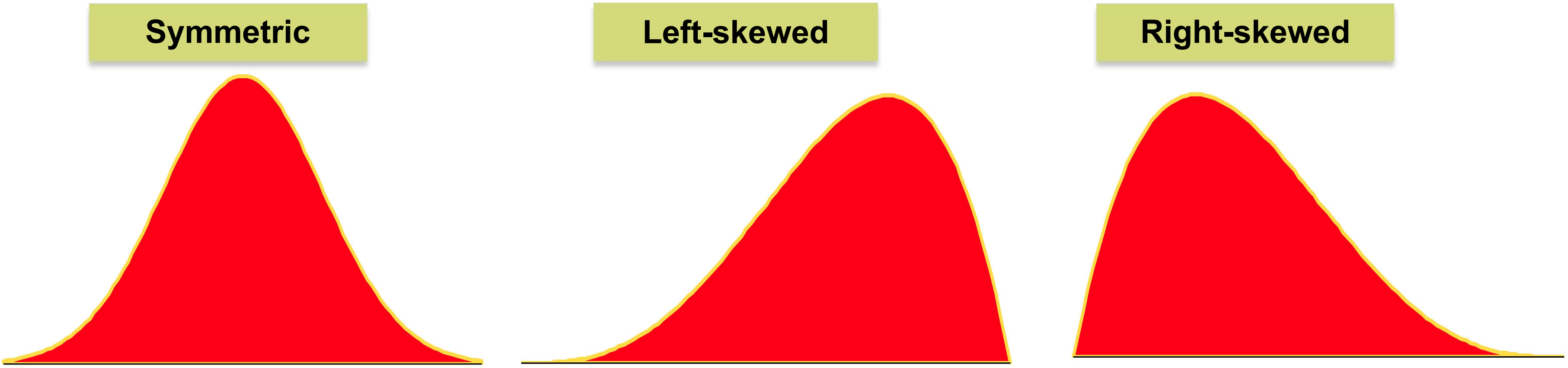
Examining Distributions

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 **spread.**
- An important kind of deviation is an **outlier,** an individual that falls outside the overall pattern.

Examining Distributions

- A distribution is **symmetric** if the right and left sides (or tails) 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.
- It is **skewed to the left** (left-skewed) if the left side of the graph is much longer than the right side.

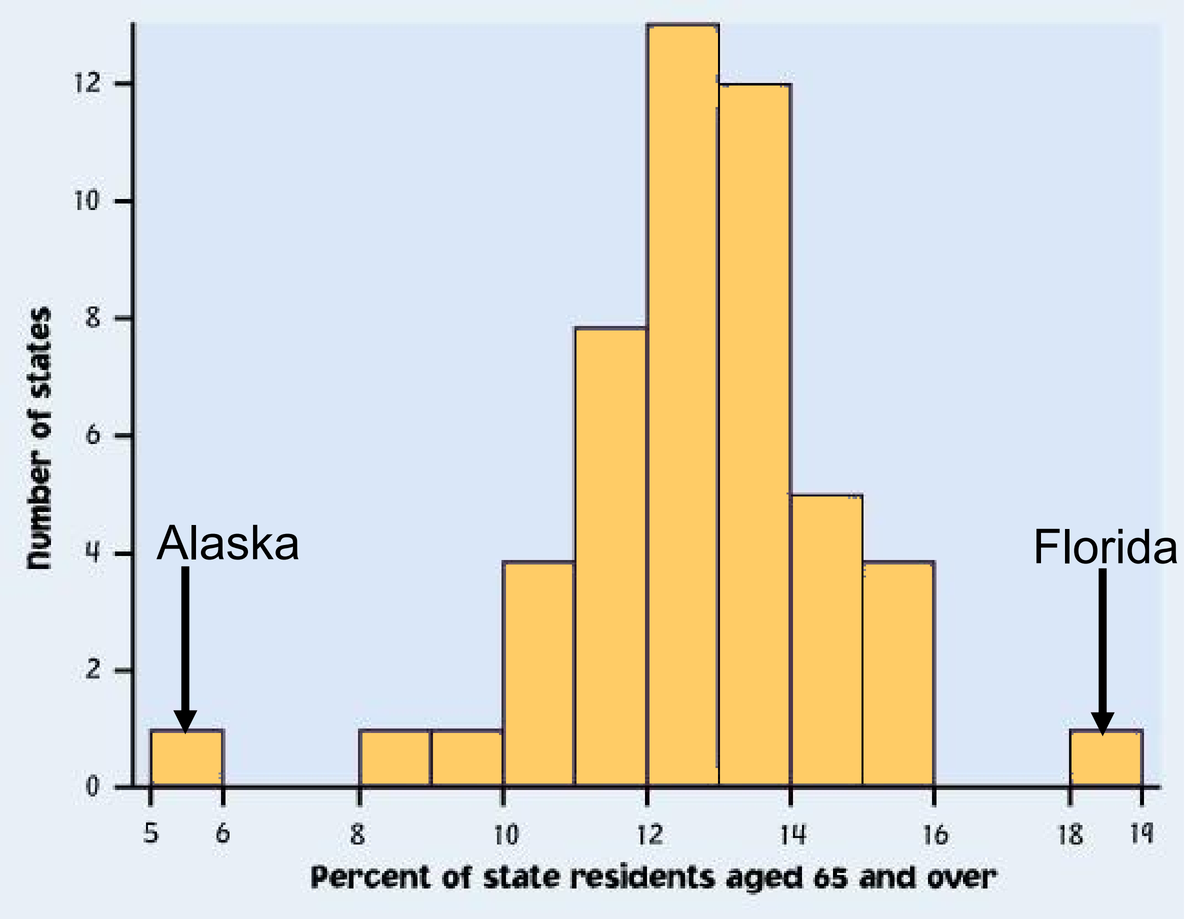


Outliers

An important kind of deviation is an **outlier.** Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.

The overall pattern is fairly symmetric except for two states that clearly do not belong to the main pattern. Alaska and Florida have unusually small and large percentages, respectively, of elderly residents in their populations.

A large gap in the distribution is typically a sign of an outlier.



1.3 Describing Distributions with Numbers

- Measures of center: mean, median
- Mean versus median
- Measures of spread: quartiles, standard deviation
- Five-number summary and boxplot
- Choosing among summary statistics
- Changing the unit of measurement

Measuring Center: The Mean

The most common measure of center (or central tendency) is the arithmetic average, or mean.

To find the **mean** $\bar{x}$ (pronounced "*x*-bar") of a set of observations, add their values, and divide by the number of observations. If the n observations are $x_1, x_2, x_3, \ldots, x_n$, their mean is:

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

In more compact notation: $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum x_i$$

Measuring Center: The Median

Because the mean is highly susceptible to the influence of extreme observations, it is not a **resistant measure** of center.

Another common measure of center is the **median.**

The **median *M*** is the midpoint of a distribution, the number such that half of the observations are smaller and the other half are larger.

To find the median of a distribution:

1. Arrange all observations from smallest to largest.
2. If the number of observations n is odd, the median M is the center observation in the ordered list.
3. If the number of observations n is even, the median M is the average of the two center observations in the ordered list.

Measuring Center: Example

Use the data below to calculate the mean and median of the commuting times (in minutes) of 20 randomly selected New York workers.

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

$$\bar{x} = \frac{10 + 30 + 5 + 25 + \ldots + 40 + 45}{20} = 31.25 \text{ minutes}$$

0	5
1	005555
2	00**05**
3	00
4	005
5	
6	005
7	
8	5

Key: 4|5 represents a New York worker who reported a 45-minute travel time to work.

$$M = \frac{20 + 25}{2} = 22.5 \text{ minutes}$$

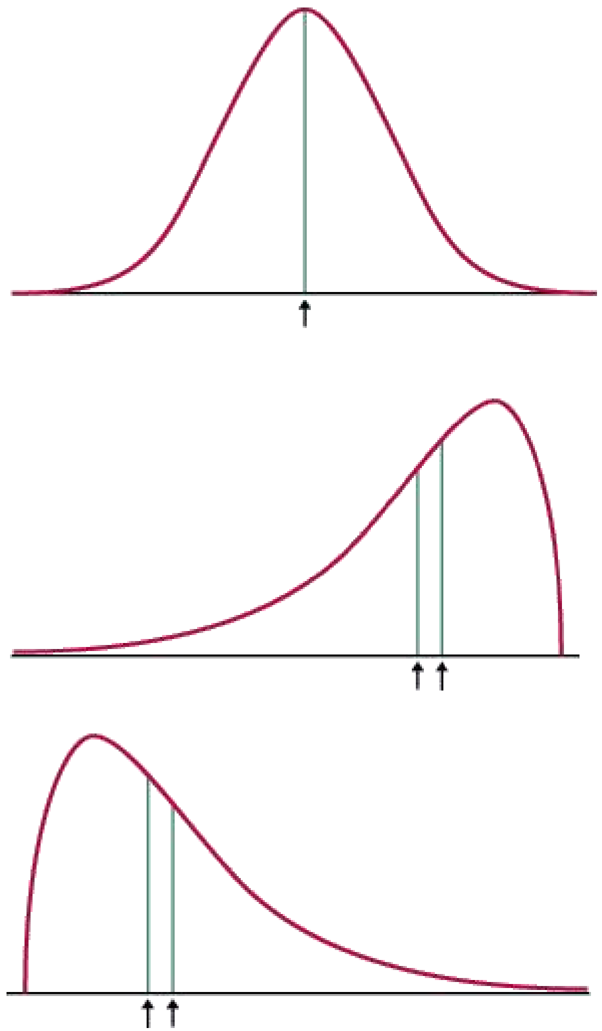
Comparing Mean and Median

The mean and median measure center in different ways, and both are useful.

The mean and median of a roughly **symmetric** distribution are close together.

If the distribution is exactly **symmetric,** the mean and median are exactly the same.

In a **skewed** distribution, the mean is usually farther out in the long tail than is the median.



Measuring Spread: The Quartiles

- A measure of center alone can be misleading.
- A useful numerical description of a distribution requires both a measure of center and a measure of spread (or variability).

How to Calculate the Quartiles and the Interquartile Range

To calculate the **quartiles:**

- Arrange the observations in increasing order and locate the **median *M***.
- The **first quartile Q_1** is the median of the observations located to the left of the median in the ordered list.
- The **third quartile Q_3** is the median of the observations located to the right of the median in the ordered list.
- The **interquartile range (*IQR*)** is defined as: $IQR = Q_3 - Q_1$.

The Five-Number Summary

The minimum and maximum values alone tell us little about the distribution as a whole. Likewise, the median and quartiles tell us little about the tails of a distribution.

To get a quick summary of both center and spread, combine all five numbers.

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest.

Minimum $\quad Q_1 \quad M \quad Q_3 \quad$ ***Maximum***

Boxplots

The median and quartiles divide the distribution roughly into quarters. This leads to a new way to display quantitative data, the **boxplot.**

How to Make a Boxplot

- Draw and label a number line that includes the range of the distribution.
- Draw a central box from Q_1 to Q_3.
- Note the median M inside the box.
- Extend lines (whiskers) from the box out to the minimum and maximum values that are not outliers.

Suspected Outliers: 1.5 × IQR Rule

In addition to serving as a measure of spread, the interquartile range (IQR) is used as part of a rule of thumb for identifying outliers.

The 1.5 × IQR Rule for Outliers

Call an observation an outlier if it falls more than 1.5 × IQR above the third quartile or below the first quartile.

In the New York travel time data, Q_1 = 15 minutes, Q_3 = 42.5 minutes, and so IQR = 27.5 minutes.

For these data, $1.5 \times IQR = 1.5(27.5) = 41.25$

$Q_1 - 1.5 \times IQR = 15 - 41.25 = \mathbf{-26.25}$

$Q_3 + 1.5 \times IQR = 42.5 + 41.25 = \mathbf{83.75}$

Any travel time shorter than –26.25 minutes or longer than 83.75 minutes is considered an outlier.

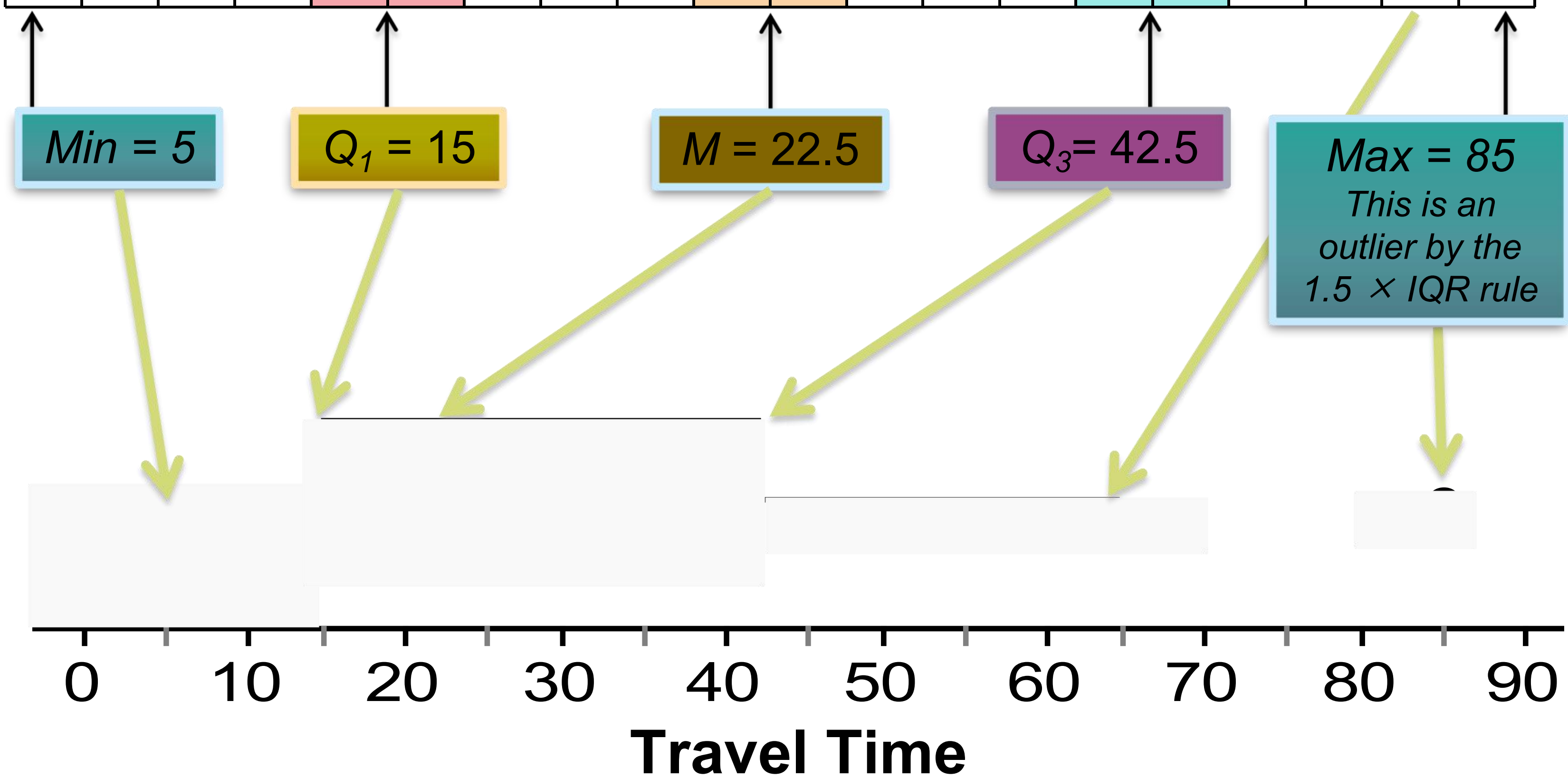
0	5
1	005555
2	0005
3	00
4	005
5	
6	005
7	
8	**5**

Boxplots

Consider our New York travel times data. Construct a boxplot.

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

5	10	10	15	**15**	**15**	15	20	20	**20**	**25**	30	30	40	**40**	**45**	60	60	65	85



Measuring Spread: The Standard Deviation

The most common measure of spread looks at how far each observation is from the mean. This measure is called the **standard deviation.**

The **standard deviation** s_x measures the average distance of the observations from their mean. It is calculated by finding the "average" of the squared distances and then taking the square root. This average squared distance is called the **variance.**

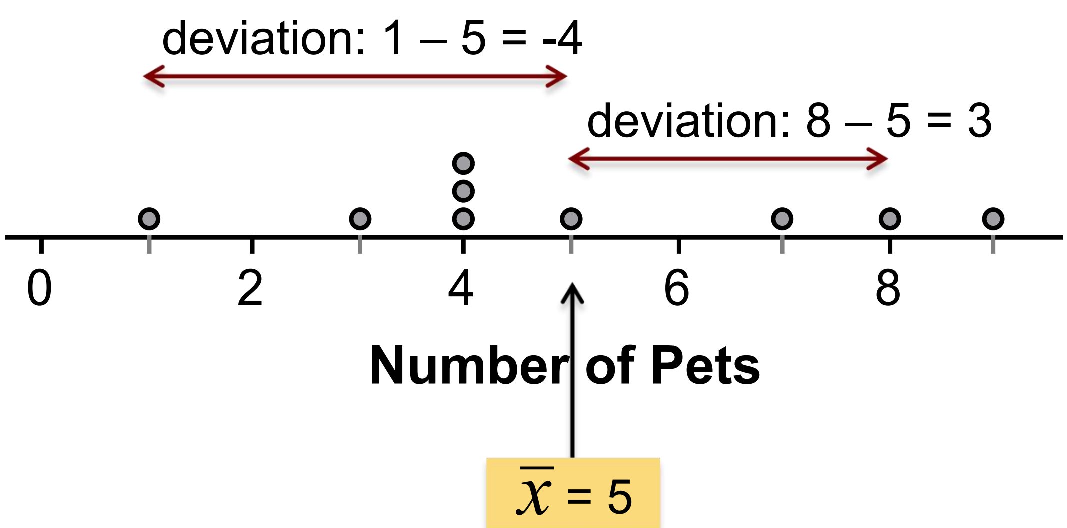
$$\text{variance} = s_x^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

$$\text{standard deviation} = s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

Calculating the Standard Deviation

Example: Consider the following data on the number of pets owned by a group of nine children.

1. Calculate the mean.
2. Calculate each *deviation.*
   deviation = *observation – mean*



Calculating the Standard Deviation

3. Square each deviation.

4. Find the "average" squared deviation. Calculate the sum of the squared deviations divided by (n – 1). This is called the **variance.**

5. Calculate the square root of the variance. This is the **standard deviation.**

x_i	$(x_i\text{-mean})$	$(x_i\text{-mean})^2$
1	1 – 5 = –4	$(-4)^2 = 16$
3	3 – 5 = –2	$(-2)^2 = 4$
4	4 – 5 = –1	$(-1)^2 = 1$
4	4 – 5 = –1	$(-1)^2 = 1$
4	4 – 5 = –1	$(-1)^2 = 1$
5	5 – 5 = 0	$(0)^2 = 0$
7	7 – 5 = 2	$(2)^2 = 4$
8	8 – 5 = 3	$(3)^2 = 9$
9	9 – 5 = 4	$(4)^2 = 16$
	Sum = ?	**Sum = ?**

"Average" squared deviation = 52/(9 – 1) = 6.5. This is the **variance.**

Standard deviation = square root of variance = $\sqrt{6.5} = 2.55$

Properties of the Standard Deviation

- *s* measures spread about the mean and should be used only when the mean is the measure of center.
- $s = 0$ only when all observations have the same value and there is no spread. Otherwise, $s > 0$.
- *s* is not resistant to outliers.
- *s* has the same units of measurement as the original observations.

Choosing Measures of Center and Spread

We now have a choice between two descriptions for center and spread:

- ✓ Mean and standard deviation
- ✓ Median and interquartile range

Choosing Measures of Center and Spread

The median and *IQR* are usually better than the mean and standard deviation for describing a skewed distribution or a distribution with outliers.

Use mean and standard deviation only for reasonably symmetric distributions that don't have outliers.

NOTE: Numerical summaries do not fully describe the shape of a distribution. *ALWAYS PLOT YOUR DATA!*

Changing the Unit of Measurement

Variables can be recorded in different units of measurement. Most often, one measurement unit is a **linear transformation** of another measurement unit: $x_{new} = a + bx$.

Linear transformations do not change the basic shape of a distribution (skew, symmetry, multimodal). But they do change the measures of center and spread:

- Multiplying each observation by a positive number *b* multiplies both measures of center (mean, median) and spread (IQR, *s*) by *b*.
- Adding the same number *a* (positive or negative) to each observation adds *a* to measures of center and to quartiles, but it does not change measures of spread (IQR, *s*).

Note: Linear transformation can only be applied to quantitative variables.

1.4 Density Curves and Normal Distributions

- Density curves
- Measuring center and spread for density curves
- Normal distributions
- The 68-95-99.7 rule
- Standardizing observations
- Using the standard Normal Table
- Inverse Normal calculations
- Normal quantile plots

Exploring Quantitative Data

We now have a kit of graphical and numerical tools for describing distributions. We also have a strategy for exploring data on a single quantitative variable. Now we'll add a fourth step to the strategy.

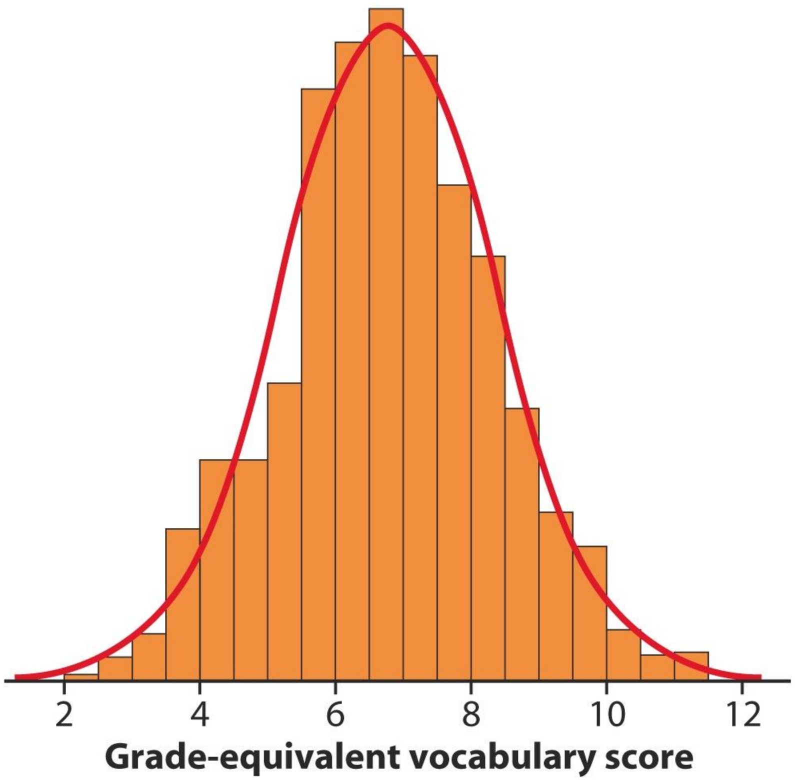
Exploring Quantitative Data

1. Always plot your data: make a graph.
2. Look for the overall pattern (shape, center, and spread) and for striking departures such as outliers.
3. Calculate a numerical summary to briefly describe center and spread.
4. Sometimes, the overall pattern of a large number of observations is so regular that we can describe it by a smooth curve.

Density Curves

Example: Here is a histogram of vocabulary scores of 947 seventh graders.

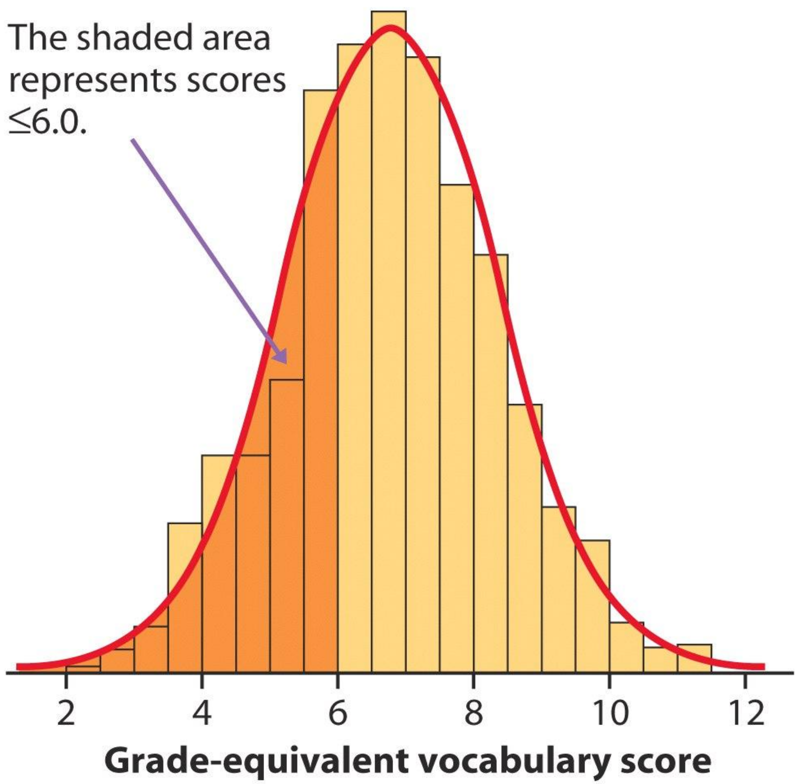
The smooth curve drawn over the histogram is a **mathematical model** for the distribution.



Density Curves

The areas of the shaded bars in this histogram represent the proportion of scores in the observed data that are less than or equal to 6.0. This proportion is equal to 0.303.

Now the area under the smooth curve to the left of 6.0 is shaded. If the scale is adjusted so the total area under the curve is exactly 1, then this curve is called a **density curve.** The proportion of the area to the left of 6.0 is now equal to 0.293.



Density Curves

A **density curve** is a curve that:

- is always on or above the horizontal axis
- has an area of exactly 1 underneath it

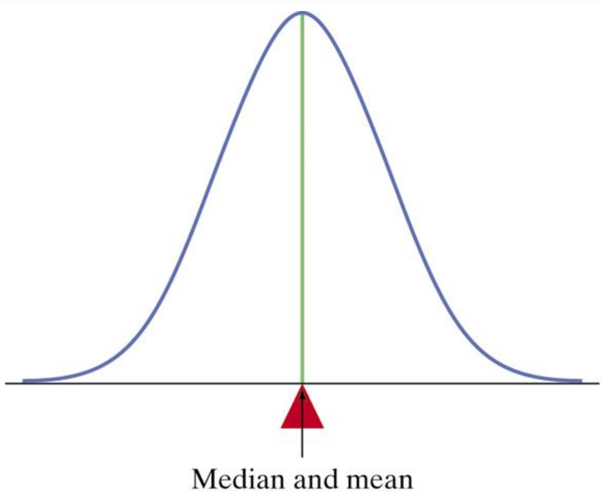
A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values on the horizontal axis is the proportion of all observations that fall in that range.

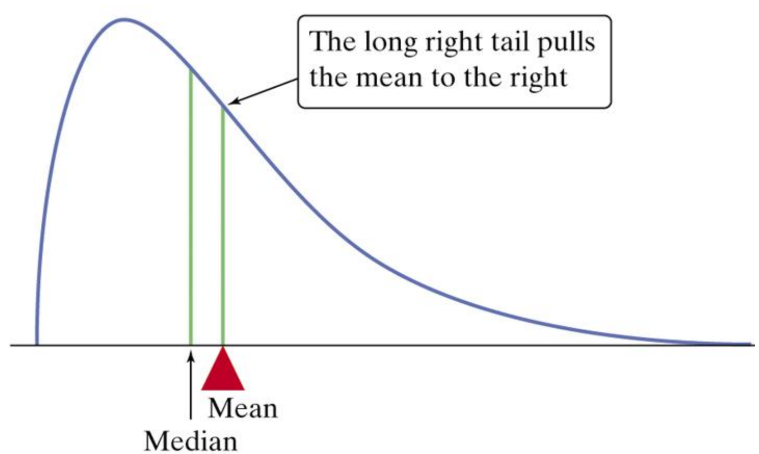
Density Curves

Our measures of center and spread apply to density curves as well as to actual sets of observations.

Distinguishing the Median and Mean of a Density Curve

- The **median** of a density curve is the equal-areas point—the point that divides the area under the curve in half.
- The **mean** of a density curve is the balance point, that is, the point at which the curve would balance if made of solid material.
- The median and the mean are the same for a symmetric density curve. They both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.





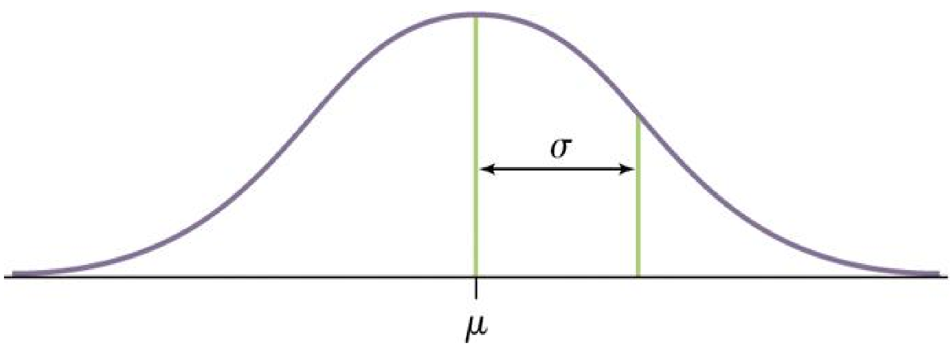
Density Curves

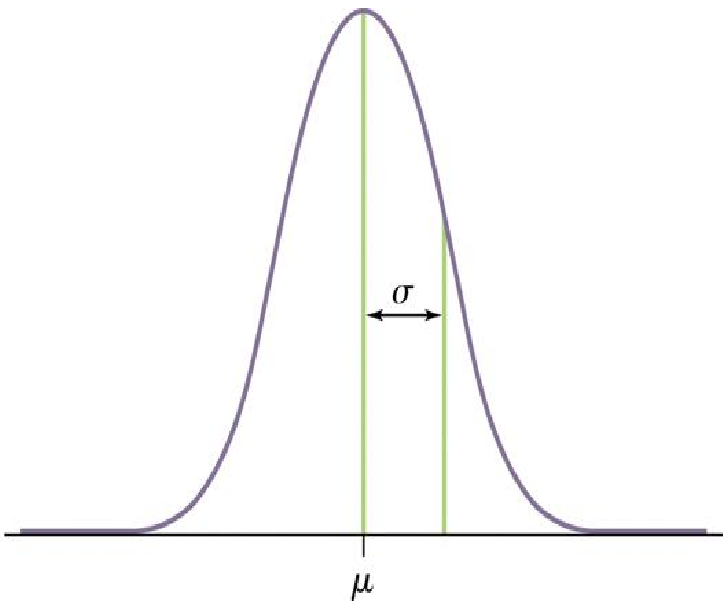
- The mean and standard deviation computed from actual observations (data) are denoted by $\bar{x}$ and *s*, respectively.
- The mean and standard deviation of the actual distribution represented by the density curve are denoted by μ ("mu") and σ ("sigma"), respectively.

Normal Distributions

One particularly important class of density curves is the class of Normal curves, which describe Normal distributions.

- All Normal curves are symmetric, single-peaked, and bell-shaped.
- A specific Normal curve is described by giving its mean μ and standard deviation σ.





Normal Distributions

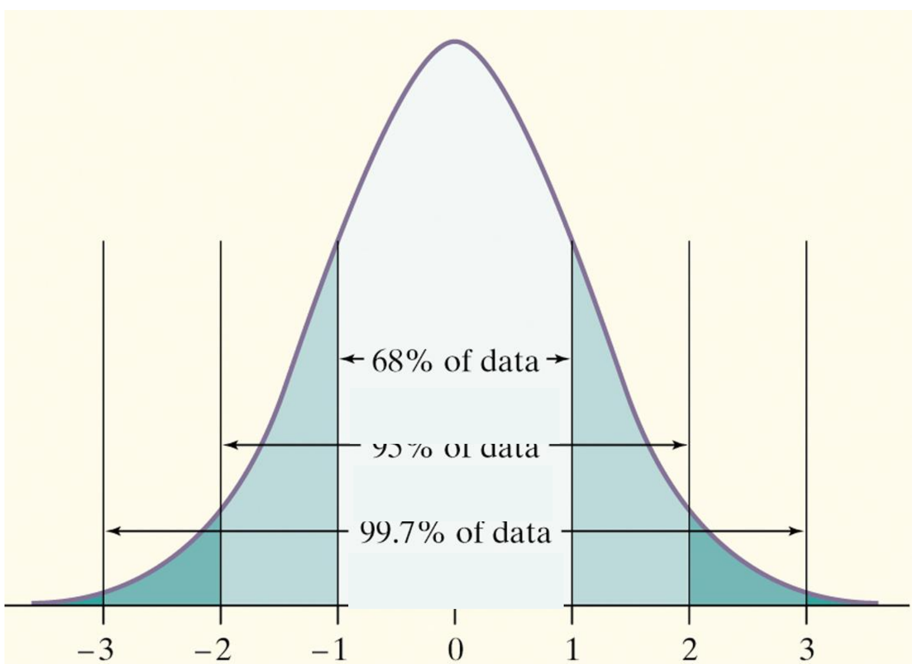
A **Normal distribution** is described by a Normal density curve. Any particular Normal distribution is completely specified by two numbers: its mean μ and standard deviation σ.

- The mean of a Normal distribution is the center of the symmetric **Normal curve.**
- The standard deviation is the distance from the center to the change-of-curvature points on either side.
- We abbreviate the Normal distribution with mean μ and standard deviation σ as $N(\mu,\sigma)$.

The 68-95-99.7 Rule

In the Normal distribution with mean μ and standard deviation σ:

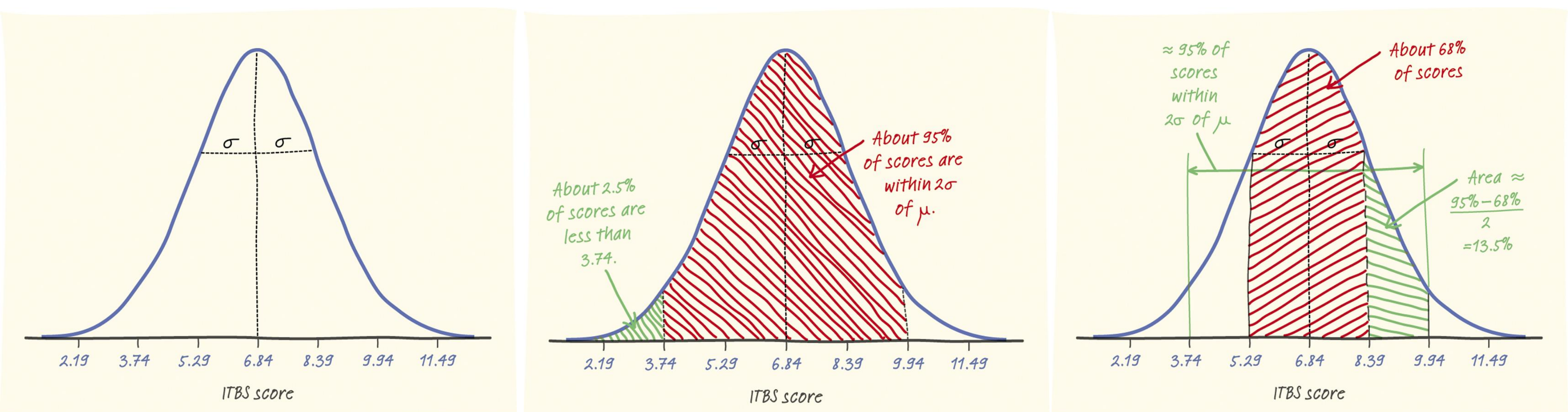
- Approximately **68%** of the observations fall within σ of μ.
- Approximately **95%** of the observations fall within 2σ of μ.
- Approximately **99.7%** of the observations fall within 3σ of μ.



The 68-95-99.7 Rule

The distribution of Iowa Test of Basic Skills (ITBS) vocabulary scores for 7th-grade students in Gary, Indiana, is close to Normal. Suppose the distribution is *N*(6.84, 1.55).

- ✓Sketch the Normal density curve for this distribution.
- ✓What percent of ITBS vocabulary scores are less than 3.74?
- ✓What percent of the scores are between 5.29 and 9.94?



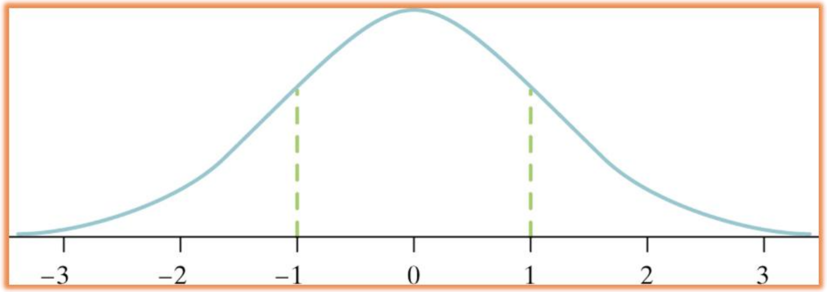
Standardizing Observations

If a variable *x* has a distribution with mean μ and standard deviation σ, then the **standardized value** of *x*, or its **z-score,** is

$$z = \frac{x - \mu}{\sigma}$$

All Normal distributions are the same if we measure in units of size σ from the mean μ as center.

The **standard Normal distribution** is the Normal distribution with mean 0 and standard deviation 1. That is, the standard Normal distribution is *N*(0,1).



The Standard Normal Table

Because all Normal distributions are the same when we standardize, we can find areas under any Normal curve from a single table.

The Standard Normal Table

Table A is a table of areas under the standard Normal curve. The table entry for each value *z* is the area under the curve to the left of *z*.

TABLE A Standard normal probabilities

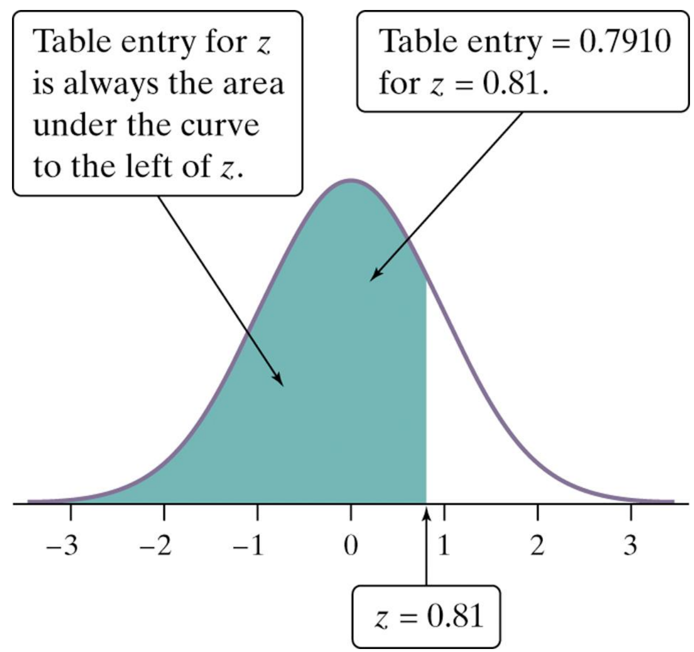
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681

The Standard Normal Table

Suppose we want to find the proportion of observations from the standard Normal distribution that are less than 0.81.
We can use Table A:

z	.00	.01	.02
0.7	.7580	.7[illegible]1	.7642
0.8	.7881	.7910	.7939
0.9	.8159	.8186	.8212

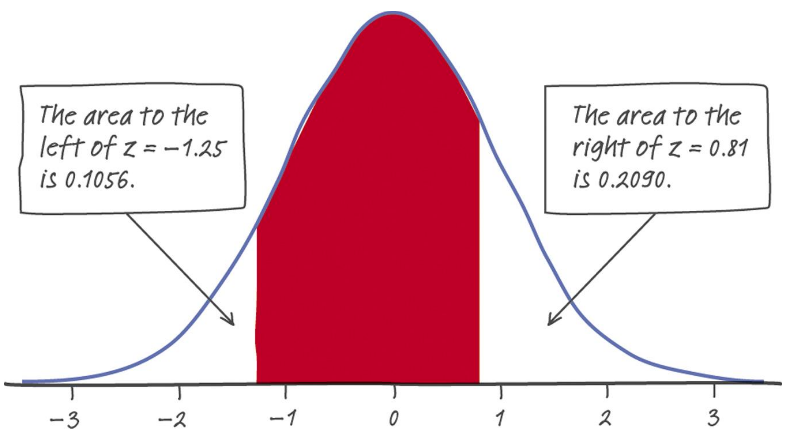
$P(z < 0.81) = .7910$



Normal Calculations

Find the proportion of observations from the standard Normal distribution that are between –1.25 and 0.81.

Can you find the same proportion using a different approach?



$$1 - (0.1056+0.2090) = 1 - 0.3146$$
$$= \mathbf{0.6854}$$

Normal Calculations

How to Solve Problems Involving Normal Distributions

Express the problem in terms of the observed variable *x*.

Draw a picture of the distribution and shade the area of interest under the curve.

Perform calculations.

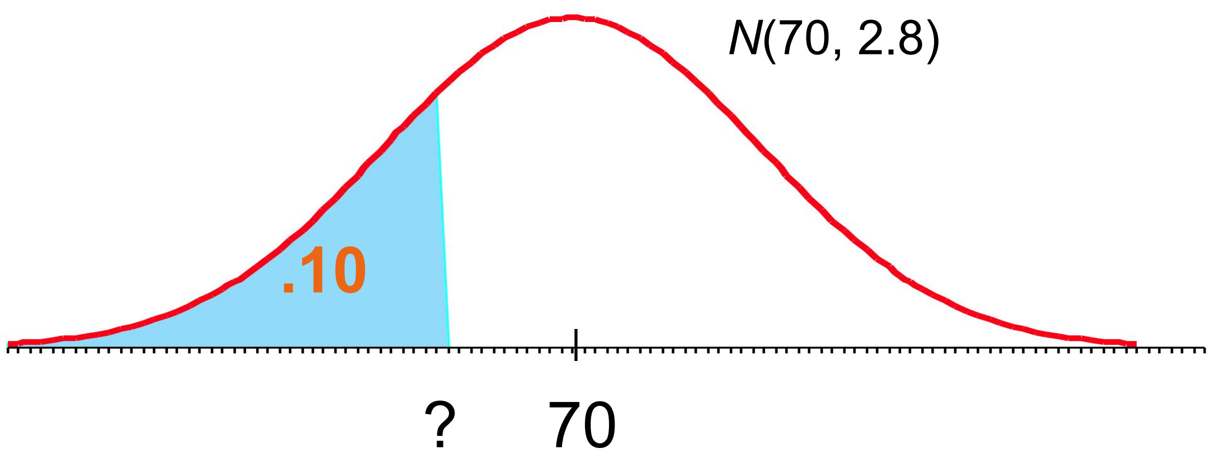
- ***Standardize*** *x* to restate the problem in terms of a standard Normal variable *z*.
- ***Use Table A*** and the fact that the total area under the curve is 1 to find the required area under the standard Normal curve.

Write your conclusion in the context of the problem.

Normal Calculations

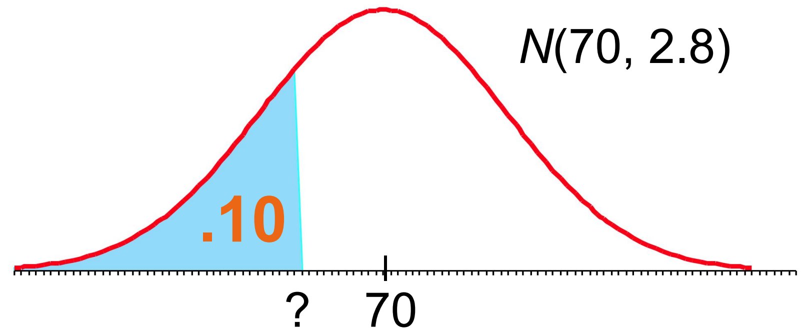
According to the Health and Nutrition Examination Study of 1976–1980, the heights (in inches) of adult men aged 18–24 are *N*(70, 2.8).

If exactly 10% of men aged 18-24 are shorter than a particular man, how tall is he?



Normal Calculations

How tall is a man who is taller than exactly 10% of men aged 18–24?



Look up the probability closest to 0.10 in the table.

Find the corresponding **standardized score.**

The value you seek is that many standard deviations from the mean.

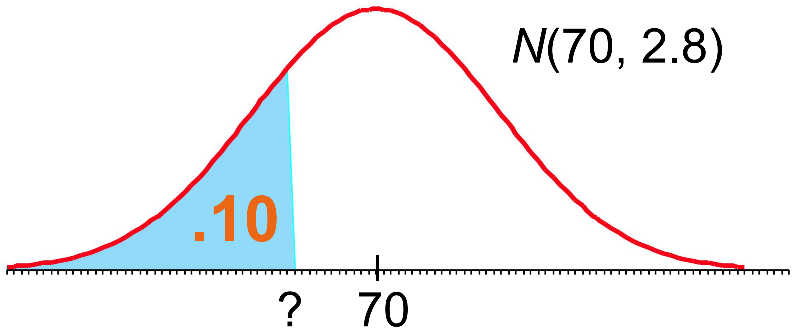
z	.07	.08	.09
–1.3	.0853	.0[illegible]8	.0823
–1.2	.1020	.1003	.0985
–1.1	.1210	.1190	.1170

Z = –1.28

Normal Calculations

How tall is a man who is taller than exactly 10% of men aged 18–24?

Z = –1.28



We need to "unstandardize" the *z*-score to find the observed value (*x*):

$$z = \frac{x-\mu}{\sigma} \Leftrightarrow x = \mu + z\sigma$$

$$\begin{aligned} x &= 70 + z(2.8) \\ &= 70 + [(-1.28) \times (2.8)] \\ &= 70 + (-3.58) = \underline{\mathbf{66.42}} \end{aligned}$$

A man would have to be approximately 66.42 inches tall or less to be in the lower 10% of all men in the population.

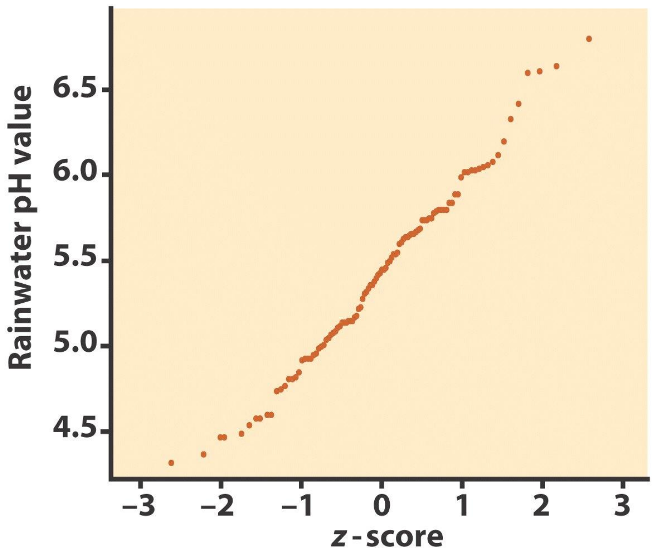
Normal Quantile Plots

One way to assess if a distribution is indeed approximately Normal is to plot the data on a **Normal quantile plot.**

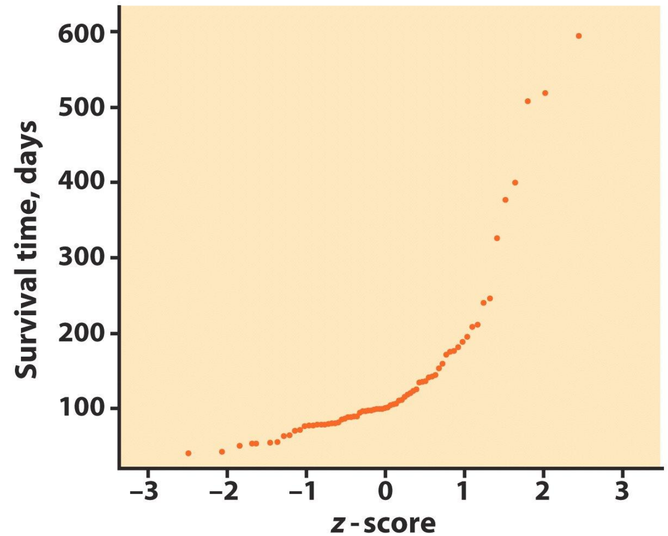
The data points are ranked and the percentile ranks are converted to *z*-scores with Table A. The *z*-scores are then used for the *x*-axis against which the data are plotted on the *y*-axis of the Normal quantile plot.

- If the distribution is indeed Normal, the plot will show a straight line, indicating a good match between the data and a Normal distribution.
- Systematic deviations from a straight line indicate a non-Normal distribution. Outliers appear as points that are far away from the overall pattern of the plot.

Normal Quantile Plots



Good fit to a straight line: The distribution of rainwater pH values is close to Normal.



Curved pattern: The data are not Normally distributed. Instead, the data are right skewed: A few individuals have particularly long survival times.

Normal quantile plots are complex to do by hand, but they are standard features in most statistical software.

Chapter 1 Looking at Data—Distributions

Introduction

1.1 Data

1.2 Displaying Distributions with Graphs

1.3 Describing Distributions with Numbers

1.4 Density Curves and Normal Distributions