Chapter 2 Summary
Data

What did you learn?

Data is information in context

- Who, what, why, where, when, and how.
- We must know at least who (cases), what (variables), and why (how do we treat variables) to be able to say anything useful about the data.

Variables are separated in two basic ways: categorical or quantitative

- Categorical variables group cases into categories and counts them.
- Quantitative variables record measurements or amounts in units.
- Thinking needs to be done as to how to treat the variable.

Statistics are used to help companies market to their customers (*Amazon.com*)

Data needs context: who, what, when, where, why, and how

Context	The context ideally tells *Who* was measured, *What* was measured (in units), *How* the data were collected, *Where* the data were collected, and *When* and *Why* the study was performed.
Data	Systematically recorded information, whether numbers, or labels, together with its context.

Data needs to be organized

Data table	An arrangement of data in which each row represents a case and each column represents a variable.

Who

Case	A case is an individual about whom or which we have data.
Respondents	Individuals who answer a survey
Subjects/ Participants	People on whom we experiment
Experimental units	Animals, plants, web sites, and other inanimate objects upon whom we experiment
Records	The rows in a database, usually identifying cases

What and Why

Area codes were originally designated so that easy numbers to dial on a rotary dial were in high population areas (New York City was 212, L.A. was 213, etc.)

Variable	A variable holds information about the same characteristic for many cases.
Units	A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams.
Categorical variable	A variable that names categories (whether with words or numerals) is called catgeorical.
Quantitative variable	A variable in which the numbers act as numerical values is called quantitative. Quantitative variables always have units.

When, Where, and How also can help understand the data

What can go wrong?

- Don't label a variable as categorical or quantitative without thinking about the question you want it to answer.
- Just because your variable's values are numbers, don't assume that it's quantitative.
- Always be skeptical.

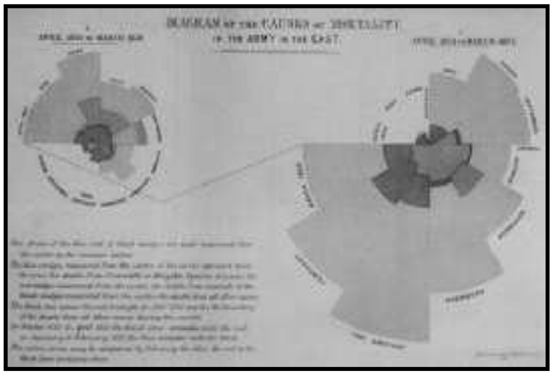
Chapter 3 Summary
Displaying and Describing Categorical Data

What did you learn?
We can summarize categorical data by counting the number of cases in each category, sometimes expressed as percents. Data can be displayed in a bar or circle graph. Categorical relationships can be viewed in a contingency table.

- We examine the marginal distribution of each variable
- We also look at conditional distribution of a variable within each category of the other variable
- We can display these conditional and marginal distributions in bar or circle graphs
- If the conditional distributions of one variable are about the same for every category of the other, the variables are independent

Three Rules of Data Analysis

1. Make a picture – organizes patterns and helps clear thinking about relationships
2. Make a picture – show important features of data and patterns in the data
3. Make a picture – tell others about the data



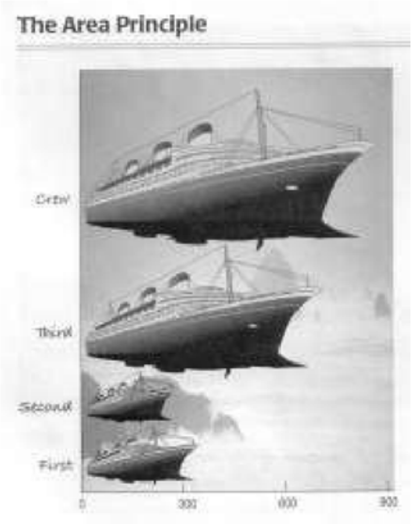
Florence Nightingale used data displays to show more soldiers in the Crimean War (1856) died of illness and infection than of battle wounds

Titanic data

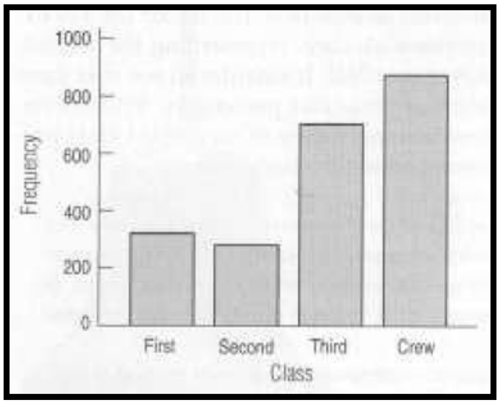
Class	Passenger Count	Percent
First	325	14.77%
Second	285	12.95%
Third	706	32.08%
Crew	885	40.21%

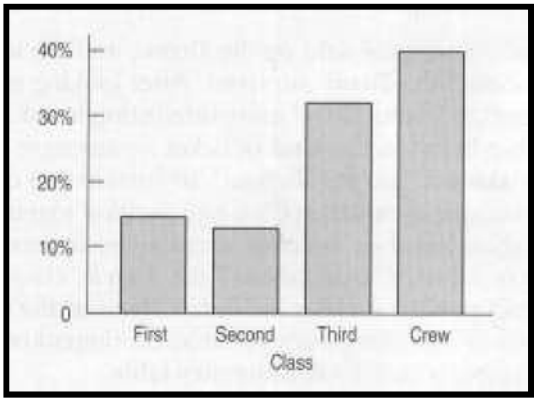
Frequency table	A frequency table lists the categories in a categorical variable and gives the count of observations for each category.
Proportion	A comparison of numbers. For our use, dividing the counts by the total number of cases.
Percentages	Multiplying proportions by 100 to express as percentages, or ratios compared to a total of 100.B41
Relative frequency table	A frequency table lists the categories in a categorical variable and gives the percentage of observations for each category.
Distribution	The distribution of a variable gives the possible values of the variable and the relative frequency of each value.

Area principle	In a statistical display, each data value should be represented by the same amount of area.

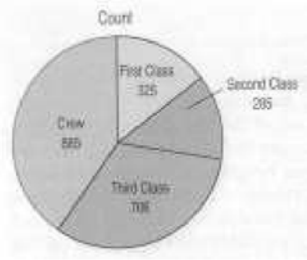


Bar chart	Bar charts show a bar representing the count of each category in a categorical variable.
Relative frequency bar chart	A bar chart that shows the proportion of people in each category rather than counts.





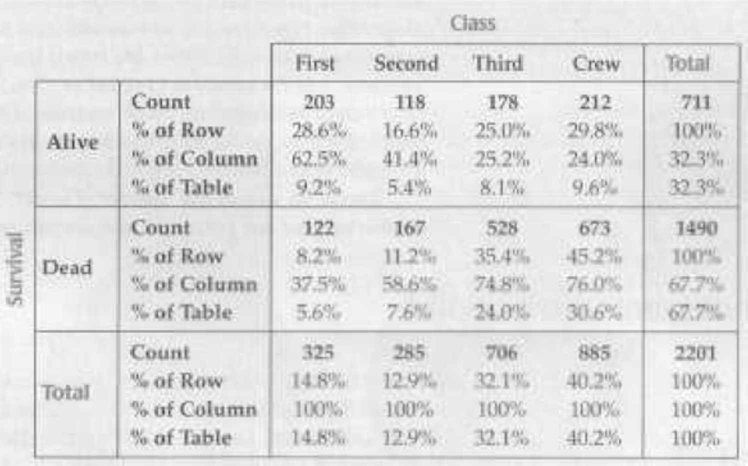
Pie chart	Pie charts show how a "whole" divides into categories by showing a wedge of a circle whose area corresponds to the proportion in each category.



Categorical data condition	Before making a bar or pie chart, be sure that the categorical data is in counts or percentages of individuals.
Contingency table	A contingency table displays counts and, sometimes, percentages of individuals falling into named categories on two or more variables. The table categorizes the individuals on all variables at once, to reveal possible patterns in one variable that may be contingent on the category of the other.

Chapter 3 Summary Continued

Marginal distribution	In a contingency table, the distribution of either variable alone is called the marginal distribution. The counts or percentages are the totals found in the margins (last row or column) of the table.

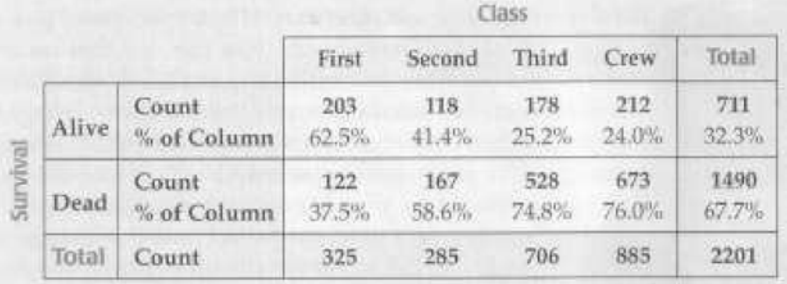


		Class				
		First	Second	Third	Crew	Total
Alive	Count	203	118	178	212	711
	% of Row	28.6%	16.6%	25.0%	29.8%	100%
	% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
	% of Table	9.2%	5.4%	8.1%	9.6%	32.3%
Dead	Count	122	167	528	673	1490
	% of Row	8.2%	11.2%	35.4%	45.2%	100%
	% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
	% of Table	5.6%	7.6%	24.0%	30.6%	67.7%
Total	Count	325	285	706	885	2201
	% of Row	14.8%	12.9%	32.1%	40.2%	100%
	% of Column	100%	100%	100%	100%	100%
	% of Table	14.8%	12.9%	32.1%	40.2%	100%

Survival

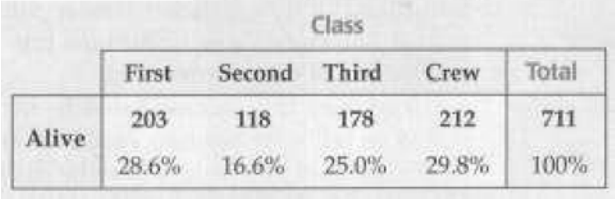
Conditional distribution	The distribution of a variable restricting the *Who* to consider only a smaller group of individuals is called a conditional distribution.

Did the chance of surviving the *Titanic* depend on ticket class?

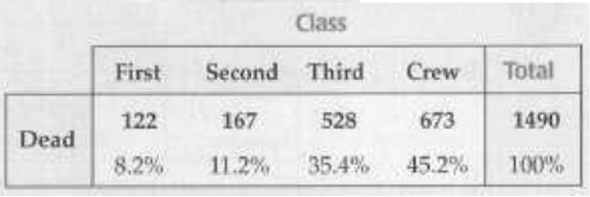


		Class				
		First	Second	Third	Crew	Total
Alive	Count	203	118	178	212	711
	% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
Dead	Count	122	167	528	673	1490
	% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
Total	Count	325	285	706	885	2201

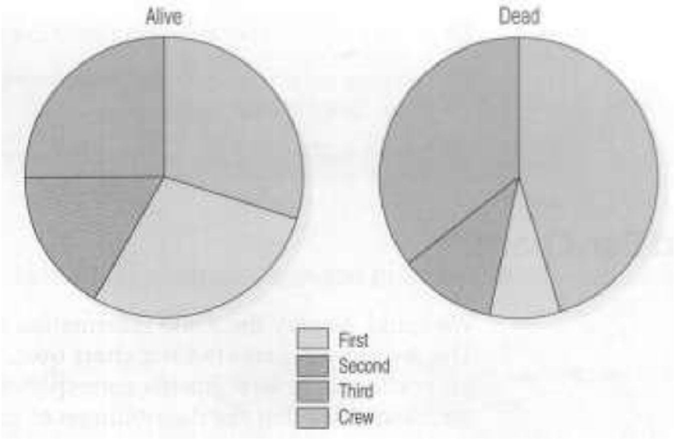
Survival



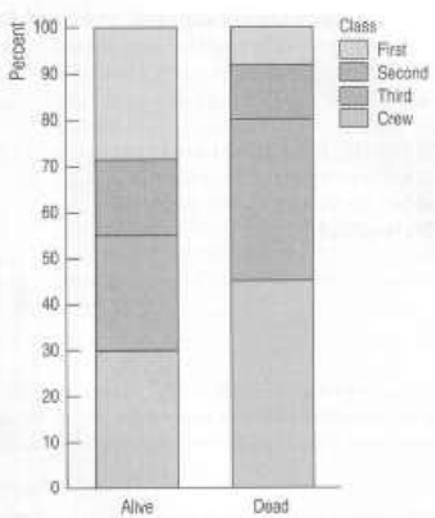
	Class				
	First	Second	Third	Crew	Total
Alive	203	118	178	212	711
	28.6%	16.6%	25.0%	29.8%	100%



	Class				
	First	Second	Third	Crew	Total
Dead	122	167	528	673	1490
	8.2%	11.2%	35.4%	45.2%	100%



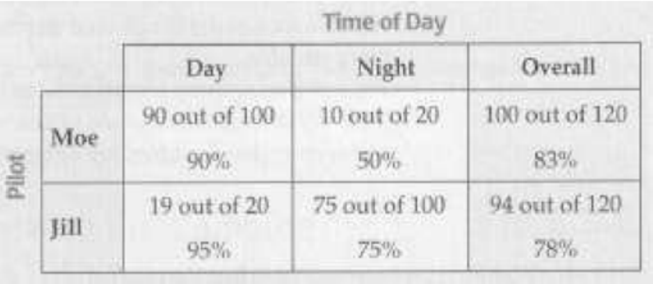
Independence	Variables are said to be independent if the conditional distribution of one variable is the same for each category of the other.
Segmented bar chart	A bar chart where each bar is treated as the "whole" and divides the bar proportionally into segments corresponding to the percentage in each group.



What can go wrong?

- Don’t violate the area principle.
- Keep it honest.
- Don’t confuse similar-sounding percentages.
- Don’t forget to look at the variables separately, too.
- Be sure to use enough individuals.
- Don’t overstate your case.
- Don’t use unfair or silly averages.

Simpson's paradox	When averages are taken across different groups, they can appear to contradict the overall averages. This is known as "Simpson's paradox."



		Time of Day		
		Day	Night	Overall
Pilot	Moe	90 out of 100 90%	10 out of 20 50%	100 out of 120 83%
	Jill	19 out of 20 95%	75 out of 100 75%	94 out of 120 78%

It’s always better to compare percentages or other averages within each level of the other variable.

Chapter 4 Summary
Displaying Quantitative Data

What did you learn?

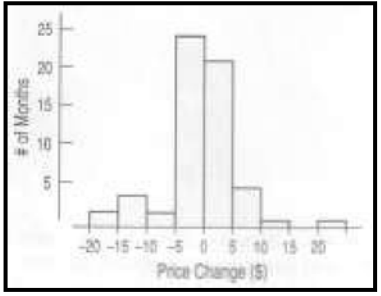
Make a picture for quantitative data to help see the story the data has to tell.

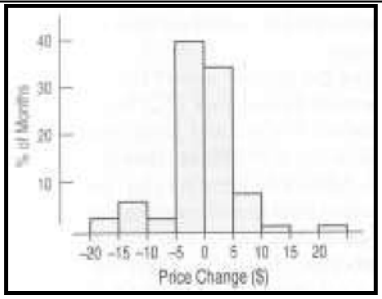
- Distribution of quantitative data can be shown using a histogram, a stem-and-leaf plot, or a dotplot.
- Examine the shape, center, spread, and unusual features of the data.
- We can compare two different groups using displays. If we use the same scale, with can compare them using shape, center, spread, or unusual features of the groups.
- Trends can be viewed on a timeplot of the data that is collected over time.

Enron story – Stock price change:

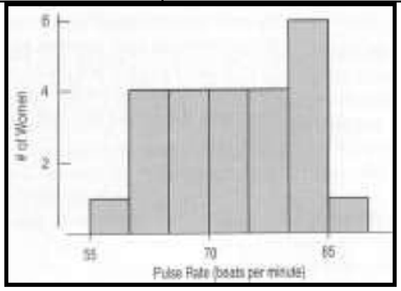
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
1997	-$1.44	-0.75	-0.69	-0.88	0.12	0.75	0.81	-1.75	0.69	-0.22	-0.16	0.34
1998	0.78	0.62	2.44	-0.28	2.22	-0.50	2.06	-0.88	-4.50	4.12	1.16	-0.50
1999	3.28	3.34	-1.22	0.47	5.62	-1.59	4.31	1.47	-0.72	-0.38	-3.25	0.03
2000	5.72	21.06	4.50	4.56	-1.25	-1.19	-3.12	8.00	9.31	1.12	-3.19	-17.75
2001	14.38	-1.08	-10.11	-12.11	5.84	-9.37	-4.74	-2.69	-10.61	-5.85	-17.16	-11.59

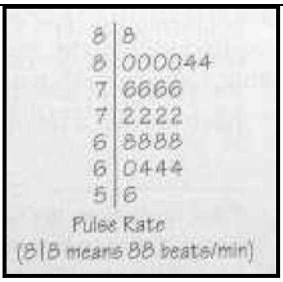
Distribution	The distribution of a variable gives the possible values of the variable and the relative frequency of each value.
Histogram	A histogram uses adjacent bars to show the distribution of values in a quantitative variable. Each bar represents the frequency of values falling in an interval of values.
Relative frequency histogram	A histogram uses adjacent bars to show the distribution of values in a quantitative variable. Each bar represents the relative frequency of values falling in an interval of values.





Stem-and-leaf display	A stem-and-leaf display shows quantitative data values in a way that sketches the distribution of the data.



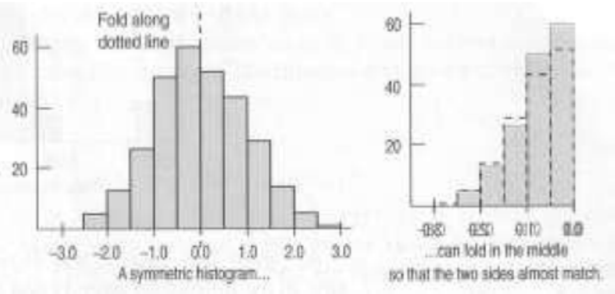


Dotplot	A dotplot graphs a dot for each case against a single axis.
Quantitative data condition	The data are values of a quantitative variable whose units are known.

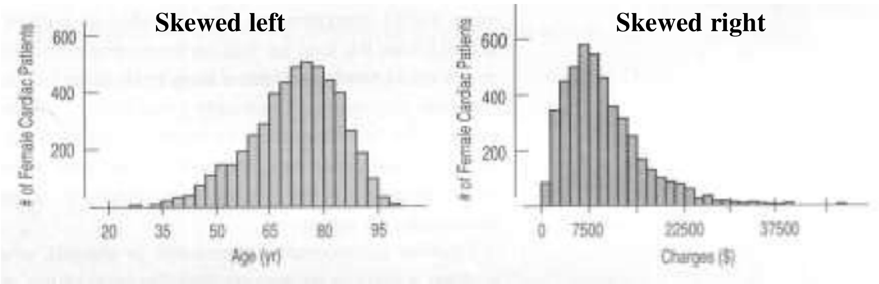
Shape	To describe the shapes of a distribution, look for single versus multiple modes and symmetry versus skewness.
Center	A value that attempts the impossible by summarizing the entire distribution with a single number, a "typical" value.
Spread	A numerical summary of how tightly the values are clustered around the "center."

Modes	A hump or local high point in the shape of the distribution of a variable is called a "mode." The apparent location of modes can change as the scale of a histogram is changed.
Unimodal	Having one mode. This a useful term for describing the shape of a histogram when it's generally mound-shaped.
Bimodal	Distributions with two modes.
Multimodal	Distributions with more than two modes.

Uniform	A distribution that's roughly flat.
Symmetric	A distribution is symmetric if the two halves on either side of the center look approximately like mirror images of each other.



Tails	The tails of a distribution are the parts that typically trail off on either side. Distributions can be characterized as having long tails (if they straggle off for some distance) or short tails (if they don't).
Skewed	A distribution is skewed if it's not symmetric and one tail stretches out farther than the other. Distributions are said to be **skewed left** when the longer tail stretches to the left, and **skewed right** when it goes to the right.

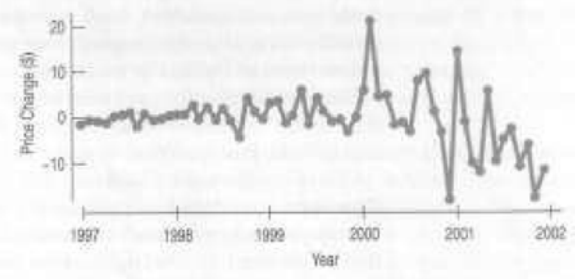


Chapter 4 Summary Continued

Outliers	Outliers are extreme values that don't appear to belong with the rest of the data. They may be unusual values that deserve further investigation, or just mistakes; there's no obvious way to tell. Don't delete outliers automatically - you have to think about them. Outliers can affect many statistical analyses, so you should always be alert for them.

Gaps	Regions of a histogram that have no values for a given data set.

Timeplot	A timeplot displays data that change over time. Often, successive values are connected with lines to show trends more clearly.

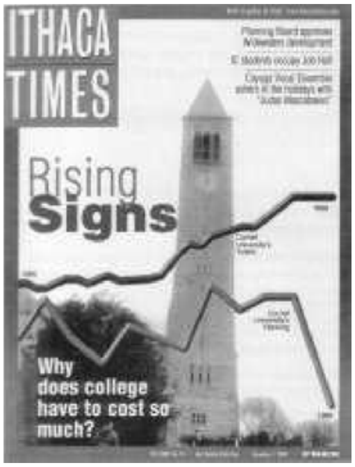


Monthly Enron stock price changes stretched out over time in a timeplot. Now it's easy to see that after being relatively stable, the stock price became somewhat volatile in 1998 and then even more so starting in 2000. **Figure 4.11**

Re-express / Transform	Applying a simple function to a set of data to make a skewed distribution more symmetric.

What can go wrong?

- Don't make a histogram of a categorical variable.
- Don't look for shape, center, and spread of a bar chart.
- Don't use bars in every display – save them for histograms and bar charts.
- Choose a bin width appropriate to the data.
- Avoid inconsistent scales.
- Label clearly.



A plot gone wrong:

- Horizontal scales are inconsistent – one starts in 1965 and the other in 1989.
- Vertical axis isn't labeled, not consistent.
- Vertical scales don't point in the same direction and ranking going lower (from 15th to 6th) should be viewed as an improvement.

Chapter 5 Summary
Describing Distributions Numerically

What did you learn?

Distributions of quantitative variables can be summarized numerically.

- The 5-number summary displays the two quartiles, the median, and the two extremes for a variable.
- Measures of center include mean and median.
- Measures of spread include range, IQR, and standard deviation.
- When the distribution is skewed, we report the median and the IQR. When the distribution is symmetric, we report the mean and the standard deviation (and possibly the median and the IQR as well).

Data distributions can be displayed using boxplots.

- A boxplot reveals some features of a distribution not easily seen in a histogram – the center, the middle 50%, and outliers. Histograms are better at showing the shape of the distribution.
- Boxplots are effective for comparing groups graphically. When discussing group comparisons, we discuss their shape, center, spread, and unusual features.

n always refers to the number of data values.

Typical values for a data set are usually the center value

Center	We summarize the center of a distribution with the mean or the median.
Mean	The mean is found by summing all the data values and dividing by the count.
Midrange	The mean of the minimum and maximum values of a set of data.
Median	The median is the middle value of an organized data set, with half of the data above and half below it.

Spread	We summarize the spread of a distribution with the standard deviation, interquartile range, and range.
Range	The difference between the lowest and highest values in a data set (maximum - minimum).

Quartiles	The median and the quartiles divide data into four equal parts.
Lower Quartile	The lower quartile (Q1) is the value with a quarter of the data below it (the median of the lower half of the data set).
Upper Quartile	The upper quartile (Q3) is the value with the a quarter of the data above it (the median of the upper half of the data set).
Interquartile range (IQR)	The difference between the first and third quartile (IQR = Q3 - Q1).
Percentiles	The *i*th percentile is the number that falls above *i%* of the data.

5-number summary	A 5-number summary for a variable consists of the minimum, the lower quartile, the median, the upper quartile, and the maximum.
Boxplot	A boxplot displays the 5-number summary as a central box with whiskers that extend to the non-outlying data values. Boxplots are particularly effective for comparing groups of different sizes.

Deviations are distances between an individual data point and the ‘center’

y represent individual data of *n* data values and *M* represents the middle, so

$\sum$ represents the difference between each datum and the center

We want the deviations to be 0, so

$\sum \quad = 0$ becomes $\sum \quad \sum$ which becomes

$$\sum \quad + \quad + \cdots + \quad = \quad \text{and } M = \frac{\sum}{n}$$, or the mean

The mean for statistics is referred to as $\bar{y}$ and pronounced “y=bar”

Mean is the balancing point of a histogram

For skewed data, it is better to report the median than the mean as the measure of center

Standard deviation measures the distance each value is from the center and should only be used with symmetric data

Standard deviation	The standard deviation is the square root of the variance.
Variance	The variance is the sum of squared deviations from the mean, divided by the count minus one.

In order for the distances from center to not cancel each other out due to positive and negative differences, we square the difference

Variance is $s^2 = \frac{\sum}{n-} \quad —$ Standard deviation is $s = \sqrt{\frac{\sum}{n-} \quad —}$

What can go wrong?

- Don’t forget to do a reality check.
- Don’t forget to sort the values before finding the median or percentiles.
- Don’t compute numerical summaries of a categorical variable.
- Watch out for multiple modes.
- Be aware of slightly different methods.
- Beware of outliers.
- Make a picture.
- Be careful when comparing groups that have different spreads.

Chapter 6 Summary
The Standard Deviation as a Ruler and the Normal Model

What did you learn?

Data can be easier to understand after shifting or rescaling the data.

- Shifting data by adding or subtracting the same amount from each value affects measures of central and position but not measures of spread.
- Rescaling data by multiplying or dividing every value by a constant changes all of the summary statistics – center, position, and spread.

The power of standardizing data.

- Standardizing uses the standard deviation as a ruler to measure distance from the mean, creating *z*-scores.
- Using *z*-scores allows comparison of values from different distributions or values based on different units.
- A *z*-score can identify unusual or surprising values among data.

A Normal model can sometimes provide a useful way to understand data.

- We can decide whether a Normal model is appropriate by checking the Nearly Normal Condition with a histogram or Normal probability plot.
- Normal models follow the 68-95-99.7 Rules and we can use tables or technology for a more detailed analysis.

Using the standard deviation as a ruler allows the comparison of data sets with different units

Standardizing	We standardize the eliminate units. Standardized values can be compared and combined even if the original variables had different units and magnitudes.
Standardized value	A value found by subtracting the mean and dividing by the standard deviation.
z-score	A *z*-score tells how many standard deviations a value is from the mean; *z*-scores have a mean of zero and a standard deviation of one.

$$z = \frac{y -}{s} \qquad \text{(no units)}$$

Changing center and spread	Changing the center and spread of a variable is equivalent to changing its *units.*
Shifting	Adding a constant to each data value adds the same constant to the mean, the median, and the quartiles, but does not change the standard deviation or IQR.
Rescaling	Multiplying each data value by a constant multiplies both the measures of position (mean, median, and quartiles) and the measures of spread (standard deviation and IQR) by that constant.

Standardizing into *z*-scores does not change the shape of the distribution, but it does change the center (mean = 0) and the spread (standard deviation = 1)

Normal model	A useful family of models for unimodal, symmetric distributions.
Parameter	A numerically valued attribute of a model. For example, the values of μ and σ in a N (μ, σ) model are parameters.
Statistic	A value calculated from data to summarize aspects of the data.
Standard normal model / distribution	A normal model, N (μ, σ), with mean μ = 0 and standard deviation σ = 1.

Using parameters, $z = \frac{y -}{\sigma}$

Normality assumption	When using a normal model, we make the assumption that the distribution of the data is normal.
Nearly normal condition	The shape of the distribution of a data set is unimodal and symmetric.

68-95-99.7 Rule	In a normal model, about 68% of values fall within 1 standard deviation of the mean, about 95% fall within 2 standard deviations of the mean, and about 99.7% fall within 3 standard deviations of the mean.

Fewer than 1 out of a million values have a *z*-score of less than –5 or greater than 5

Remember to make a picture of the distribution for working with Normal models

Normal percentiles	The normal percentile corresponding to a *z*-score gives the percentage of values in a standard normal distribution found at that *z*-score or below.

Example: What proportion of SAT scores fall between 450 and 600?

Think Plan: State the problem
Variables: Name the variable, check conditions and specify Normal model
Show Mechanics: Make a picture of the Normal model. Locate values and shade.
Find *z*-scores for the cut points 450 and 600
Use technology to find the area (or use a table)
Tell Conclusion: Interpret your result in context

Normal probability plot	A display to help assess whether a distribution of data is approximately normal. If the plot is nearly straight, the data satisfy the nearly normal condition.

What can go wrong?

- Don’t use a Normal model when the distribution is not unimodal and symmetric.
- Don’t use the mean and standard deviation when outliers are present.
- Don’t round off too soon.
- Don’t round your results in the middle of a calculation.
- Don’t worry about minor differences in results.