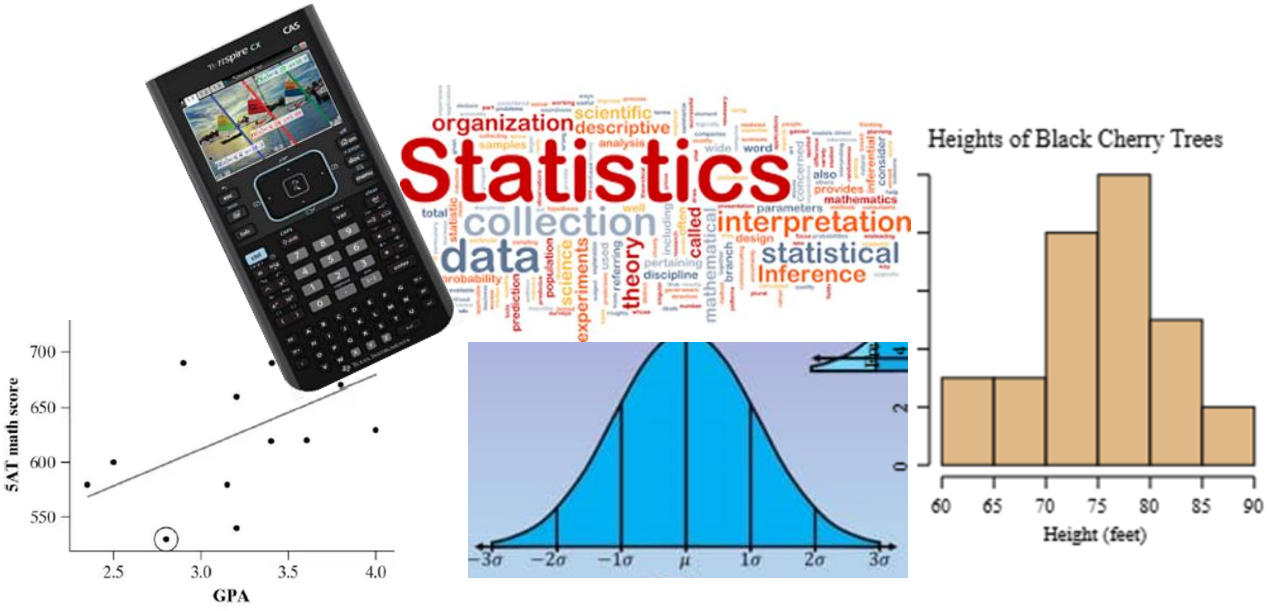
AP Statistics

Summer Packet - 2026



Due Thursday, August 6, 2026

This Summer Packet counts in the Minor Category.

Dreher High School
AP Statistics

Congratulations on your decision to take AP Statistics! The first few pages of this packet will serve as a quick overview of the year ahead including information about the TI-Nspire CX graphing calculator. Like a few other AP classes, there is some amount of summer work I'd like you to take on. Overall, you should not need to invest much more than 5 hours. I suggest that you spread the work of the packet out over the course of the summer so that you can still enjoy your summer. After all, you've earned it!

Here is an invitation code to my Summer 2026 Schoology group: **7X2B-KFDX-VT87W**

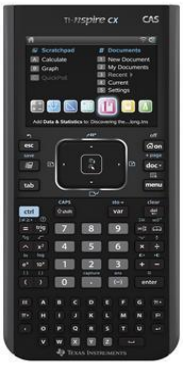
I **strongly** encourage you to join this group ASAP so you do not miss any important information.

Once you are a member of this Schoology Group, you can check for updates and even discuss problems you are having with any of the problems on the AP Statistics Summer Packet.

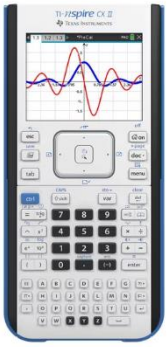
Once the school year begins, we will use our Schoology Course site exclusively and you will no longer need to return to the Summer 2026 Schoology Group.

Calculator Information

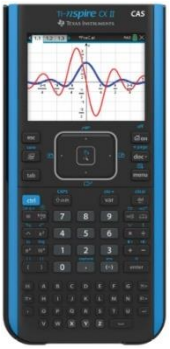
During their two years of calculus, (AP Calculus AB and AP Calculus BC), as well as AP Statistics, students will use either the Texas Instruments® TI-Nspire CX or the TI-Nspire CX CAS graphing calculator.



TI-Nspire CX CAS



TI-Nspire CX II CAS



TI-Nspire

Dreher High School charges a $25 rental fee each year for the use of the TI-Nspire or TI-Nspire CX CAS calculator for students who do not own their own calculator. If you are interested in purchasing your own calculator, it can be very advantageous to do so. The majority of students who enroll in AP Statistics as well as AB and BC Calculus purchase their own and may find it very beneficial in both high school and college in their mathematics and science courses.

Please note, that while able to use on the College Board AP Exams, students may not use a CAS (Computer Algebra System) graphing calculator on the ACT, the SAT, and the PSAT.

Even if the colleges don't allow the use of the calculator on exams, former students have communicated to us they use them extensively to help them better understand assignments and lecture notes. The TI-Nspire is an invaluable tool to use in statistics, chemistry and physics. There are two versions of the TI-Nspire CX calculator (CAS and non-CAS). As you can see from the images above, there are two models of the CAS version. All models are acceptable at DHS, but the TI-Nspire CX II CAS has a couple other nice helpful features that students tend to enjoy along with a slightly quicker processor. The two models are virtually the same price.

As of May 2026, the best price ($139.99) I have seen for the TI-Nspire CX can be found at https://www.walmart.com/ip/One-TI-Nspire-with-Recharge-Battery/344727554?wmlspartner=wlpa&selectedSellerId=101324932&adid=22222222227344727554_101324932_14069003552_202077872&wl0=&wl1=g&wl2=c&wl3=42423897272&wl4=pla-2449037643288&wl5=9010372&wl6=&wl7=&wl8=&wl9=pla&wl10=736905370&wl11=online&wl12=344727554_101324932&veh=sem&gad_source=4&gad_campaignid=202077872&gbraid=0AAAADmfBIqCHI93L3yzl-lqPiXCvtfO0&gclid=CjwKCAjwwpDQBhAuEiwAa-4WoxPq_636y8cCK6u5F4tIVxaCSnH1u2PNIAjDwdxtlrvOmubQNCnnlxoCrWwQAvD_BwE

AP Stat: Notes Chapter 1
Stats Starts Here

What is Statistics?

Statistics is the science of data. Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world. Statistics is largely about trying to find explanations for why data vary while acknowledging that some amount of the variation will always remain a mystery. Don't confuse this study of data with a statistic (a calculation made from data).

In essence, Statistics is about variation.

- People have different opinions about important issues. It can be important to see how their answers vary.
- When we take measurements in an experiment, we expect individuals to be slightly different. How much difference is simply due to random variation? And when is a difference so large that we believe something other than random variation is at work?

There are three steps to doing Statistics:

1. ***Think*** first. Know where you're headed and why!

2. ***Show***: the mechanics of calculating statistics and making displays

3. ***Tell*** what you've learned. Explain your results so that someone else can understand your conclusions.

The big idea of Statistics is that we have a question about some large group (the population) that can be answered through measurement. We use the word *measure* in a very liberal sense…you're probably thinking that *measure* means to write down a number. That is true, but not the whole story. If you write down *tall*, that's a measurement. If you take a picture, that's a measurement. While there are lots of ways to measure, we'll stick to words and numbers.

The characteristic which we measure is called the **variable**. Perhaps we want to know the average mass of a kumquat—*mass* is the variable. Perhaps we want to know the proportion of pink VW's in the U.S.—*color* is the variable. The things that we measure (kumquats, VW's) are called **individuals**. The collection of all individuals is called the **population**.

The measurements that we take are the **data** (a collection of numbers, characters, images, or other items that provide information about something). In essence, data are values along with their context.

What are data?

Data are systematically recorded information, whether numbers or labels, together with its context. Please notice that datum is the singular for data. Data values, no matter what kind, are useless without their context. For the context of the data, consider the 5 W's: Who, What, When, Where, and (if possible) Why. Often, we add How to the list. Answering these questions can provide the context for data values—in fact, the answers to the first two questions are essential!

In order to determine the context of data, consider the "W's"
 Who – *the cases (about whom the data was collected). People are referred to as* ***respondents****,* ***subjects,*** *or* ***participants****, while objects are referred to as* ***experimental units.***
Individuals who answer a survey are called respondents
People on whom we experiment are subjects or participants
Animals, plants, websites, and other inanimate subjects are called experimental units.
 What (and in what units) – *the variables (the characteristics) recorded about each individual. The variables should have a name that identify What has been measured. To understand variables, you must Think about what you want to know.*
 When – *when the data was collected.*
 Where – *where the data was collected.*
 Why – *why the data was collected. This can determine whether a variable is treated as* ***categorical*** *or* ***quantitative****.*
 How – *how the data was collected.*

Note: the answers to "who" and "what" are essential.

DEFINITION Statistical problem-solving process

- **Formulate questions:** Clarify the research problem and ask one or more valid statistical investigative questions.
- **Collect data:** Design and carry out a plan to collect appropriate data.
- **Analyze data:** Use appropriate graphical and numerical methods to analyze the data.
- **Interpret results:** Draw conclusions based on the data analysis. Be sure to answer the investigative question(s)!

This section begins with a closer look at the first step in the statistical problem-solving process: formulating questions.

A statistical study starts with an investigative question. But not just any question will do. Unlike most mathematical questions, a valid statistical investigative question is based on data that vary. For instance, "Can Joy Milne correctly identify Parkinson's disease status by smell for more than 50% of all shirts like the ones in this experiment?" is a valid investigative question because Joy may identify some shirts correctly and other shirts incorrectly. However, "How much screen time did you have yesterday?" is not a valid investigative question because the question can be answered with a single data value—Yes or No.

Example 1:

Two friends go off campus every day to eat lunch. Because the lunch period is short, they wonder whether it would be faster to order inside their favorite fast-food restaurant or to use the drive-thru. Each day, they flip a coin to determine which method (inside or drive-thru) to use and record the total length of time it takes from the moment they enter the parking lot to the moment they receive their food. After several weeks of collecting data, they analyze their results and determine that ordering inside took about 2 minutes and 34 seconds less than using the drive-thru, on average. Their conclusion: it's faster to go inside.

Determine the investigative question in this statistical study.

SOLUTION:

During all possible lunch visits to this restaurant, is the average time to receive food when ordering inside different than the average time to receive food when ordering at the drive-thru? This is a valid investigative question because the length of time it takes these friends to enter the parking lot, place their order, and receive their food will vary from day to day.

Example: Your Turn

Boarding time Airlines are interested in finding ways for passengers to board their flights more quickly and efficiently. Researchers tested different boarding methods using a group of 72 volunteer passengers of varying ages. The researchers compared the "back-to-front" boarding method used by many airlines at the time with a modified version of this method proposed by astrophysicist Jason Steffen. In the modified method, passengers lined up in advance at the gate in a predetermined order, so that all passengers with an even-numbered window seat would board first from back to front, followed by those with an odd-numbered window seat, then those with an even-numbered middle seat, and so on. On average, the Steffen method was twice as fast as the traditional boarding method. Determine the investigative question in this statistical study.

_____

_____

_____

_____

_____

_____

Populations and Samples

Suppose we want to find out what percentage of young drivers in the United States text while driving. To answer this question, we will survey 16- to 20-year-old drivers who live in the United States. Ideally, we would ask them all by conducting a census. Of course, contacting

every driver in this age group wouldn't be practical—it would take too much time and cost too much money. Instead, we pose the question to a sample chosen to represent the entire population of young drivers.

DEFINITION Population, Census, Sample

The **population** in a statistical study is the entire group of items or individuals we want information about.

A **census** collects data from every item or individual in the population.

A **sample** is a subset of items or individuals in the population from which we collect data.

Example 2:
Identify the population and the sample in each of the following settings.

a. The quality control manager at a factory selects 10 computer monitors from the 50 monitors produced during a particular hour and inspects each monitor for defects in construction and performance.
b. Prior to an election, a news organization surveys 1000 registered voters to predict the percentage of voters who prefer candidate A for president.

SOLUTION:
a. The population is all 50 computer monitors produced in this factory during that hour. The sample is the 10 monitors selected and inspected for defects.
b. The population is all registered voters. The sample is the 1000 registered voters surveyed.

Example: Your Turn
Identify the sample and the population of interest.

a. *Consumer Reports* included an evaluation of 126 tablets from a variety of manufacturers.

_____

_____

_____

b) To estimate the average length of time a patient must wait in the emergency room (ER) of a hospital during a given week, hospital staff record how many minutes it takes for

each of 50 patients to be treated by an ER doctor or nurse from the moment the patient enters the ER.

c) A company is hired to check lifeguard readiness at a local community pool. The company obtains the schedule of lifeguard shifts for the next 2 months. From that schedule, the company selects 20 different shifts during which the lifeguards will be given an unannounced test of their skills.

In a statistical study, the sample size is represented by **n** and the population size is represented by **N**. In part (a) of example 2, the quality control inspector selected a sample of **n** = 10 computer monitors from the population of **N** = 50 monitors produced that hour.

Most statistical studies collect data from samples to answer investigative questions about larger populations. In part (b) of the example, the news organization wants to answer the investigative question "What are the plausible (believable) values for the percentage of all registered voters who prefer candidate A for president?" We refer to this unknown percentage as the population parameter. Suppose that 528 of the 1000 registered voters surveyed prefer candidate A for president. That's 52.8%, which we refer to as the sample statistic. Remember **p** and **s**: **p**arameters come from **p**opulations and **s**tatistics come from **s**amples.

DEFINITION Parameter, Statistic

A **parameter** is a number that describes some characteristic of a population.

A **statistic** is a number that describes some characteristic of a sample.

We use sample statistics to estimate population parameters. In the news organization survey, 52.8% of the sample of registered voters prefer candidate A for president. That's our best guess for the unknown population percentage. Should we conclude that exactly 52.8% of the population of registered voters prefer candidate A for president? No. If another sample of 1000 registered voters was surveyed, the percentage who prefer candidate A for president would

probably yield a different sample statistic. Can we at least say that the actual population parameter is "close" to 52.8%? As you will learn in future units, that depends on what we mean by "close" and how the sample was selected.

Observational Units and Variables

Most data tables follow this format: each row describes an observational unit, and each column contains the values of a variable. The entry in a single cell of the data table, representing the value of one variable for a specific observational unit, is called a datum—the singular form of the word data. Sometimes the observational units in a data set are called individuals or cases. Note that observational units can be people, animals, or things, such as photographs, sounds, or videos.

Data Tables

The following **data table** clearly shows the context of the data presented:

Order Number	Name	State/Country	Price	Area Code	Download	Gift?	ASIN	Artist
105-2686834-3759466	Katherine H.	Ohio	0.99	440	Amsterdam	N	B0000015Y6	Cold Play
105-9318443-4200264	Samuel R	Illinois	1.99	312	Detroit	Y	B000002BK9	Red Hot, Chili Peppers
105-1372500-0198646	Chris G.	Massachusetts	0.99	413	New York, New York	N	B000068ZVQ	Frank Sinatra
103-2628345-9238664	Monique D.	Canada	0.99	902	Los Angeles	N	B0000010AA	Blink 182
002-1663369-6638649	Will S.	Ohio	0.99	567	Beverly Hills	N	B002MXA7Q0	Weezer

Notice that this data table tells us the *What* (column) and *Who* (row) for these data.

DEFINITION Observational unit, Variable

An **observational unit** is an item or individual described in a data set or statistical study.

A **variable** is a characteristic that can take different values for different observational units.

Classification of Variables

Variables come in two basic categories—**quantitative** and **qualitative (a.k.a. categorical)** (this isn't the only way to classify variables—just the only distinction that's important to us).

Qualitative variables measure *qualities*—color, flavor, opinions, etc. Qualitative variables are also known as categorical variables.

Identifier variables (categorical variables that assign a unique value for each case, used to name or identify it) are typically used to identify single cases, not to look for patterns in collection of data. For example, social security numbers and student ID numbers

We usually use a frequency table for categorical variables.

For example, here is a frequency table for the question "What shipping method was chosen?"

Shipping Method	Number of Purchases
Ground	20,345
Second-day	7,890
Overnight	5,432

Quantitative variables measure *quantities*—mass, time, charge, number, length, etc. Please note that quantitative variables always have units!

Quantitative variables can be broken down into two further categories—**discrete** and **continuous**.

Discrete variables have gaps in their possible values—they can only take on discrete (certain) values. The set of Integers is an example of a discrete set. Discrete variables will almost always (for us) measure the *number* of some thing—the number of houses; the number of people; the number of cars, etc.

Continuous variables have no gaps in their possible values. The set of Real numbers is an example of a continuous set. Continuous variables will typically measure physical phenomena—mass, length, volume, ratio, etc.

Please note that, just because your variable's values are numbers, don't assume that it's quantitative!!! For example, consider your student ID number!

In many data sets, the variable "year" is treated as categorical. But it depends on how the data are being used. Consider a data set about cars, in which one of the variables recorded is model year. If we want to know what percentage of cars on the road are 2026 models, we treat year as categorical. If we want to know the average age of cars on the road, we would convert model year to age (in years) and treat this variable as quantitative.

As you will learn, the proper method of data analysis depends on whether a variable is categorical or quantitative. For that reason, it is important to distinguish between these two types of variables. Be sure to include any units of measurement for a quantitative variable (such as centimeters for height).

Example 3:

Census at School is an international project that collects data about primary and secondary school students using online surveys. Since its launch in 2000, students from Australia, Canada, Ireland, Japan, New Zealand, South Africa, the United Kingdom, and the United States have taken part in the project.5 We selected a random sample of 50 U.S. high school students who completed the survey in a recent year. The table displays data from some of the survey questions for the first 10 students in the sample.

State	Birth month	Age (years)	Handedness	Height (cm)	Number of home occupants	Allergies	Preferred communication method
WI	11	17	Right	175	4	Yes	Internet chat/IM
IN	6	16	Right	175.5	5	No	In person
NY	6	17	Right	157	5	Yes	In person
NC	6	17	Right	169	3	No	Internet chat/IM
MA	6	18	Right	169	3	Yes	Phone call
MO	10	18	Right	170	5	No	Text messaging
PA	5	14	Right	170	6	No	Text messaging

Identify the observational units and variables in this data set.

Classify each variable as categorical or quantitative.

SOLUTION:

Observational units: 50 randomly selected U.S. high school students who completed the Census at School survey. Variables: State where student lives, birth month, age (years), handedness, height (cm), number of home occupants, whether the student has allergies, preferred communication method.

Categorical: State where student lives, birth month, handedness, whether the student has allergies, preferred communication method. Quantitative: Age (years), height (cm), and number of home occupants.

Example: Your Turn

Identify the following cases as qualitative or quantitative variables. If the variable is quantitative, identify if it is discrete or continuous.

a) A polling organization asks people if they will be voting for the incumbent candidate.

_____

b) An eye doctor measures the distance at which a patient can read an eye chart.

_____

c) A car company wants to know what colors of cars to make for next year, so they survey consumers about their favorite color.

_____

d) The Census Bureau wants to know how many people are living in each home.

_____

Example: Your Turn Part II

Determine if the variables listed below are *quantitative* or *categorical*. Neatly print "Q" for quantitative and "C" for categorical.

_____ 1. Time it takes to get to school	_____ 8. Height
_____ 2. Number of shoes owned	_____ 9. Amount of oil spilled
_____ 3. Hair color	_____ 10. Age of Oscar winners
_____ 4. Temperature of a cup of coffee	_____ 11. Type of pain medication
_____ 5. Teacher salaries	_____ 12. Jellybean flavors
_____ 6. Gender (1 = female; 2 = male)	_____ 13. Country of origin
_____ 7. Facebook user	_____ 14. Student ID

Example 4:
Identify the 5 W's, including "How," and the type of variable for the "What" for:

Medical researchers at a large city hospital investigating the impact of prenatal care on newborn health collected data from 882 births during 1998-2000. They kept track of the mother's age, the number of weeks the pregnancy lasted, the type of birth (cesarean, induced, natural), the level of prenatal care the mother had (none, minimal, adequate), the birth weight and sex of the baby, and whether the baby exhibited health problems (none, minor, major).

SOLUTION

Who: 882 births which were evaluated
When: 1998-2000
Where: a large city hospital
What (variables): the mother's age, the number of weeks the pregnancy lasted, the type of birth, the level of prenatal care, the birth weight of the baby, the gender of the baby, whether the baby had health issues
Why: Researchers were investigating the impact of prenatal care

How: Not given (perhaps, looking at medical records)
Look at the "what" for the type of variable

Qualitative	Quantitative
-type of birth	-mother's age
-level of prenatal care	-number of weeks the pregnancy lasted
-gender	-birth weight
-whether the baby had health issues	

Example: Your Turn

In June 2003 *Consumer Reports*, a magazine published in the United States, published an article on some SUVs they had tested recently. They reported some basic information about each of the vehicles and the results of some tests conducted by their staff. Among other things, the article told the brand of each vehicle, its price, and whether it had a standard or automatic transmission. They reported the vehicle's fuel economy, its acceleration (number of seconds to go from 0 to 60 mph), and its braking distance to stop from 60 mph. The article also rated each vehicle's reliability as much better than average, better than average, average, worse than average, or much worse than average. Describe the W's, if the information is given and the "how." List the variables—indicate whether each variable is qualitative or quantitative.

Example: Your Turn

Malena is a car buff who wants to find out more about the cars that high school students drive. The principal of a large local high school gives Malena permission to go to the student parking lot and collect some data. Malena selects a random sample of 50 cars from the lot and creates a spreadsheet with each car's license plate, model, year, number of stickers on the car, color, weight (in kilograms), whether it has a navigation system, and highway gas mileage (in miles per gallon).

a. Identify the sample and the population in this statistical study.
b. Identify the observational units and variables. Classify each variable as categorical or quantitative.
c. Formulate a valid investigative question that Malena could ask about the number of stickers on students' cars.

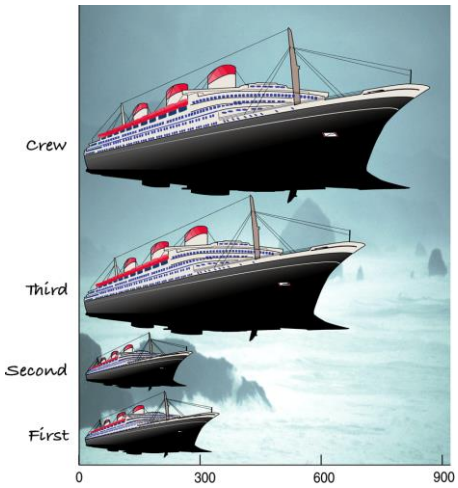
AP Stat: Notes Chapter 2
Displaying and Describing Categorical Data

The main idea to displaying data is to make a picture! This picture will help you to think clearly about the patterns and relationships that may be hiding in the data; will show the important features and patterns in the data; and will tell others about your data.

As we saw last chapter, a **categorical or qualitative** variable is a variable that names categories (whether with words or numerals). We will study how to display these variables in this chapter.

Please be careful, though, when examining displays or constructing displays to adhere to the area principle. The area principle says that the area occupied by a part of the graph should correspond to the magnitude of the value it represents. Violations of the area principle are a common way to err with Statistics.

You might think that a good way to show the Titanic data is with this display:



The ship display makes it look like most of the people on the Titanic were crew members, with a few passengers along for the ride. When we look at each ship, we see the area taken up by the ship, instead of the length of the ship.

The ship display violates the **area principle**: the area occupied by a part of the graph should correspond to the size of the value it represents.

Frequency Table

We can "pile" the data by counting the number of data values in each category of interest. We can organize these **counts** into a **frequency table,** a table that organizes counts into totals and category names.

For example, for the Titanic data:

Class	Count
First	325
Second	285
Third	706
Crew	885

Example 1

A game uses a six-sided cube that does not have numbers on the sides—rather, each side has a color. The colors of the sides are white, black, red yellow, green and blue. The cube is rolled twenty times and the color noted for each roll. Create an appropriate display:

Black	Black	Black	Yellow	Green	Green	Black	Red	Blue	Yellow
Blue	White	Blue	Red	Green	Blue	Blue	Black	Green	Red

SOLUTION

Color	White	Black	Red	Yellow	Green	Blue
Count (Frequency)	1	5	3	2	4	5

Counts are useful, but sometimes we want to know the fraction or proportion of the data in each category. Usually, we express these proportions as percentages (by multiplying the proportions by 100). A **relative frequency table** displays the percentages of the values in each category. Both types of tables show how the cases are distributed across the categories. Please note that the percentages should total 100%, but the sum can be a little too high or too low if the individual category percentages have been rounded.

For example, for the Titanic data:

Class	First	Second	Third	Crew
%	14.77	12.95	32.08	40.21

Both types of tables show how the cases are distributed across the categories. They describe the distribution of a categorical variable because they name the possible categories and tell how frequently each occurs.

Example 2

Construct a relative frequency table (using the table from Example 1).

SOLUTION

Color	White	Black	Red	Yellow	Green	Blue
Relative Frequency	$\frac{1}{20} = 5\%$	$\frac{5}{20} = 25\%$	$\frac{3}{20} = 15\%$	$\frac{2}{20} = 10\%$	$\frac{4}{20} = 20\%$	$\frac{5}{20} = 25\%$

Bar Chart

A bar chart displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison. There are, though, several varieties! One word of caution—**never ever, under any circumstances, construct a 3D bar chart**.

For all bar charts, be sure to scale and label each axis. *Scale* means to write out the values of the variable along the axis; *label* means to tell the name of the variable being measured along that axis.

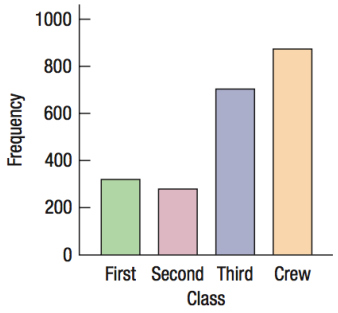
HOW TO MAKE A BAR CHART

1. **Draw and label the axes.** Put the name of the categorical variable under the horizontal axis. To the left of the vertical axis, indicate whether the graph shows the frequency (count) or relative frequency (percentage or proportion) of observational units in each category.
2. **"Scale" the axes.** Write the names of the categories in a logical order at equally spaced intervals under the horizontal axis. On the vertical axis, start at 0 and place tick marks at equal intervals until you equal or exceed the largest frequency or relative frequency in any category.
3. **Draw bars** above the category names. Make the bars equal in width and leave gaps between them. Be sure that the height of each bar corresponds to the frequency or relative frequency of observational units in that category.

Standard

The *Standard Bar Chart* lists values of the variable on the horizontal axis and frequency on the vertical axis (there are some people who reverse those axes…). A bar is drawn for each variable value, and the bars do not touch one another. The order of the values on the horizontal axis does not matter.

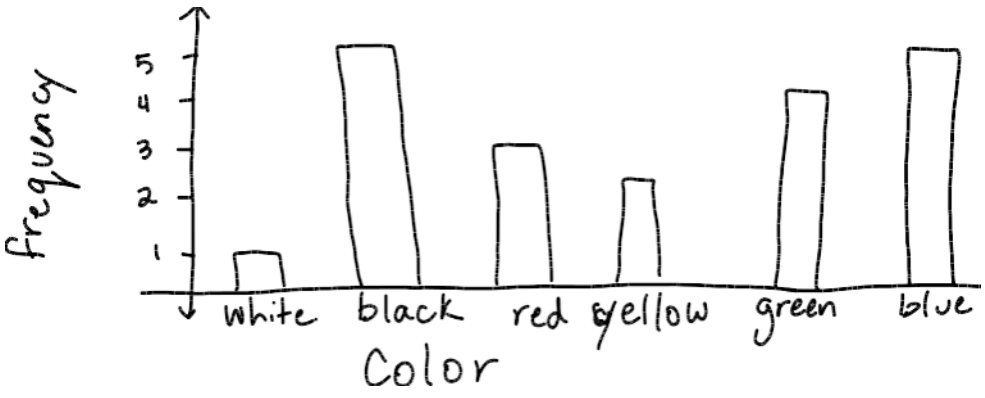
For example, a bar chart looks like:



Example 3

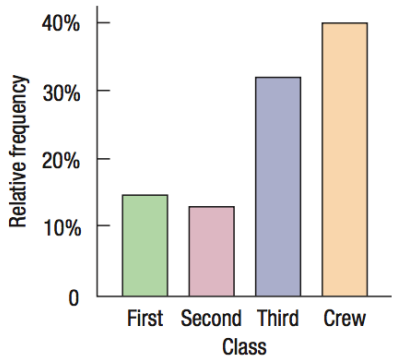
Construct a bar chart (using the table from Example 1)

SOLUTION



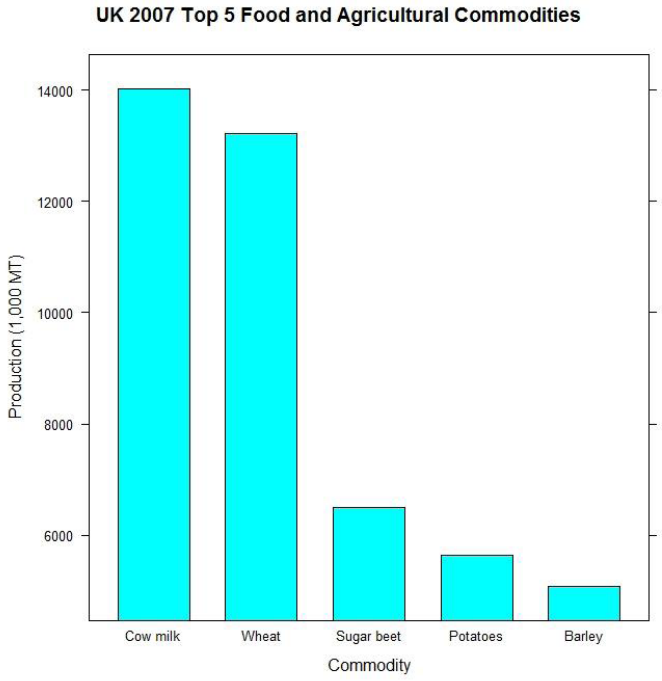
If we want to draw attention to the relative proportion in each category, we could replace the counts with percentages and use a **relative frequency bar chart**.

For example, using the Titanic data:



Pareto Charts

A *Pareto Chart* is a bar chart where the values of the variable are arranged so that the bars decrease in height from left to right



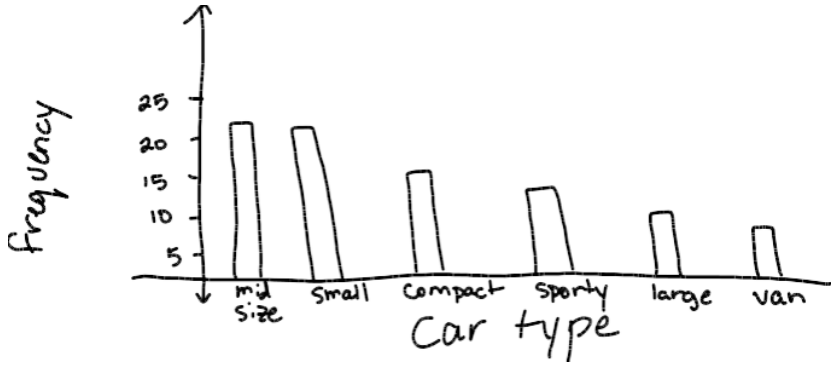
Example 4

A group of cars were classified according to type. The results are shown below:

Compact	Large	Mid-Size	Small	Sporty	Van
16	11	22	21	14	9

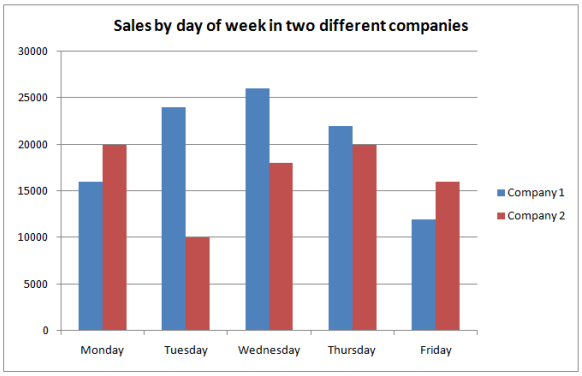
Construct a Pareto Chart.

SOLUTION



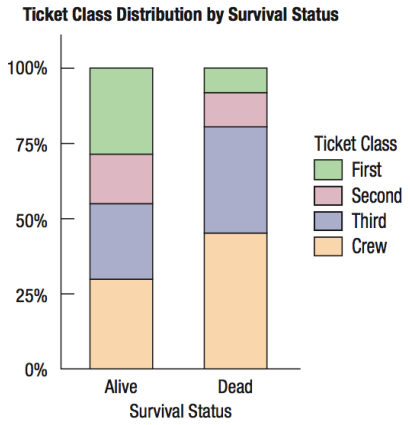
Side-by-side

Sometimes we have two groups of measurements using the same variable—for example, the colors of cars and the colors of trucks. Side-by-side bar charts have two bars where the standard bar chart has only one—one bar for each group from which measurements were taken.



Segmented

A segmented bar chart (also known as a *stacked bar chart*) is used when we have two different qualitative variables in a single frequency table. One of the variables is listed along the horizontal axis, and the values of the other variable make up segments of each bar vertically. A **segmented bar chart** displays the same information as a pie chart, but in the form of bars instead of circles. Each bar is treated as the "whole" and is divided proportionally into segments corresponding to the percentage in each group.

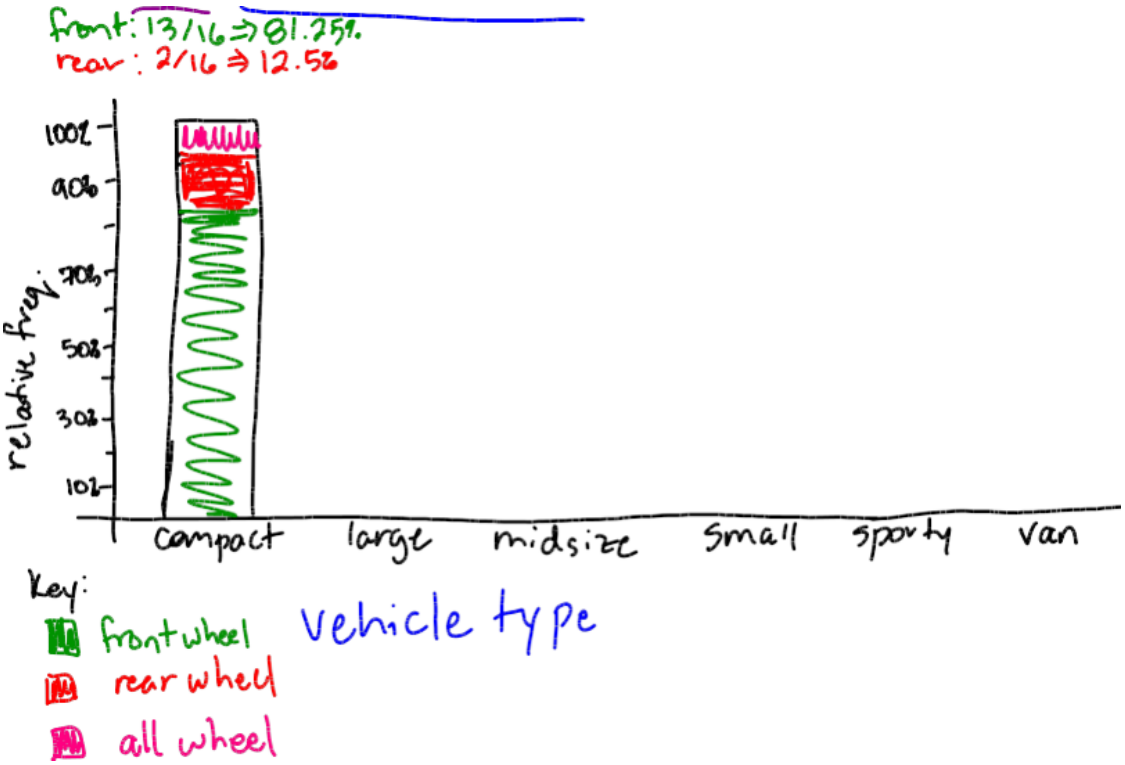


Example 5

As it turns out, there were more measurements on those cars…in particular, each car was also classified by its drive type (front wheel drive, rear wheel drive, all wheel drive). Here are the updated data:

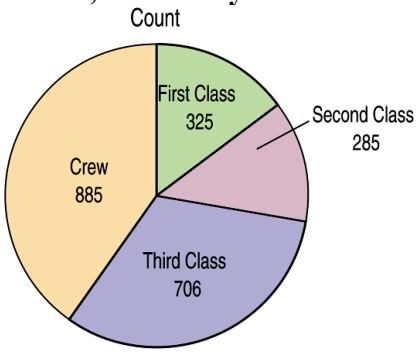
Drive Type	*Compact*	*Large*	*Midsize*	*Small*	*Sporty*	*Van*
Front	13	7	17	19	7	4
Rear	2	4	5	0	5	0
All-Wheel	1	0	0	2	2	5

Construct a segmented bar chart of Drive Type by Vehicle Type.



Pie Chart

A pie chart slices a circle into pieces whose sizes are proportional to the fraction of the whole in each category. To construct one, extend your frequency table so that you can measure relative frequency—the percent of the total. Convert those percents to angles, and then measure out the pie slices accordingly. When you're doing this by hand, just make the angle measures close—there's no need to go and use a protractor. If you need something that exact, use software. Actually, you should probably try to avoid using pie charts in the first place. They are almost always a poor choice when trying to display data. If you do use one, make sure to label each pie slice in some manner…**and never ever, under any circumstances, construct a 3D pie chart**!



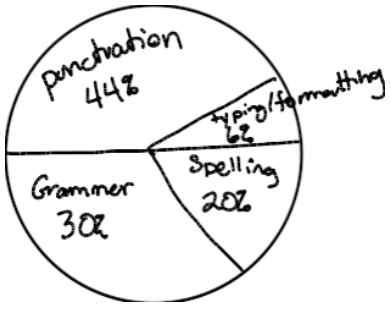
Example 6

A teacher keeps track of the number and types of errors that her students make on their research papers. The results are shown below:

Error	**Count**
Punctuation	22
Grammar	15
Spelling	10
Typing/Formatting	3

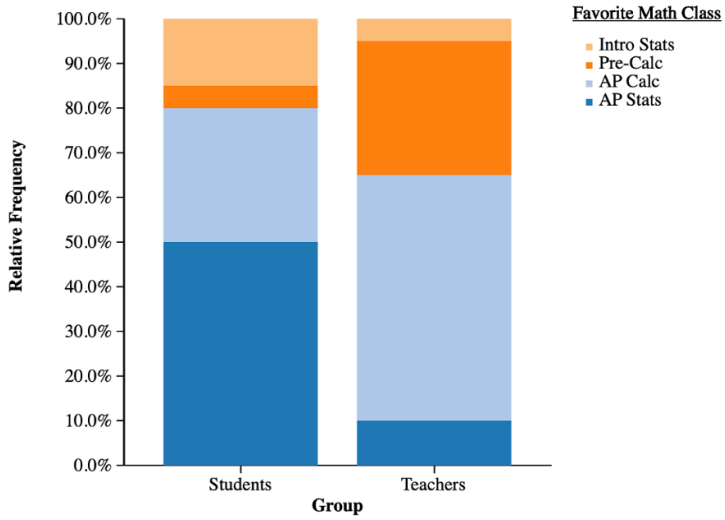
Construct a pie chart of these data.

Error	**Count**	Rel. freq.	Angle
Punctuation	22	$\frac{22}{50}=0.44$	$0.44(360^\circ)=158.4^\circ$
Grammar	15	$\frac{15}{50}=0.3$	$0.3(360^\circ)=108^\circ$
Spelling	10	$\frac{10}{50}=0.2$	$0.2(360^\circ)=72^\circ$
Typing/Forma tting	3	$\frac{3}{50}=0.06$	$0.06(360^\circ)=21.6^\circ$

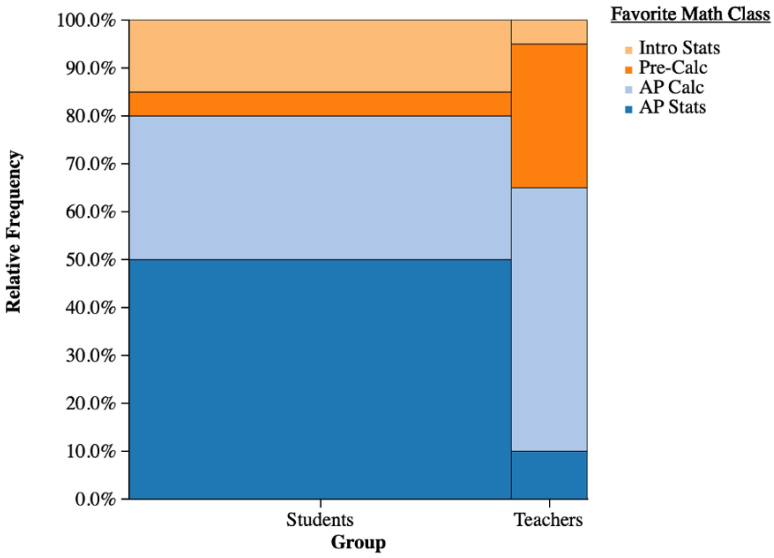


Mosaic Plot

A mosaic plot is a graphical visualization used to display the relationship between two or more categorical variables by representing contingency tables as tiles. The area of each rectangle (tile) is proportional to the frequency of the data, allowing for easy identification of relationships between variables. Start with a **categorical variable**, such as favorite math class, but for two **different groups** (Dreher Senior students and Dreher teachers). Make a segmented bar graph for each group.



The segmented bar graph does well to inform us about the percent of each category within each group. The information that is missing is the size of each group. At Dreher, this sample included 100 Senior students and 20 teachers (we'll let you guess which two teachers picked AP Stats). These vastly different group sizes are not at all represented in the segmented bar graph. A mosaic plot allows us to see these group sizes by scaling on the x-axis!



Before you make a bar chart, pie chart, or any of the charts discussed, always check the ***Categorical Data Condition: the data are counts or percentages of individuals in categories.***

Example: Your Turn

Ms. Wallulis asked her students which fast-food restaurant has the best chicken sandwich. Here are their responses:

Chick-fil-A	Shake Shack
Popeye's	Popeye's
Popeye's	Shake Shack
Chick-fil-A	Popeye's
KFC	Shake Shack
Popeye's	Popeye's
Popeye's	Popeye's
Popeye's	Shake Shack
Popeye's	Chick-fil-A
Chick-fil-A	Shake Shack
Popeye's	KFC
Popeye's	Popeye's
Chick-fil-A	Popeye's
Shake Shack	

(a) Make a frequency table and a relative frequency table to summarize the distribution of fast-food restaurant with the best chicken sandwich.

(b) Do these data support the claim that a majority of Ms. Wallulis's students think that Popeye's has the best chicken sandwich? Justify your answer.

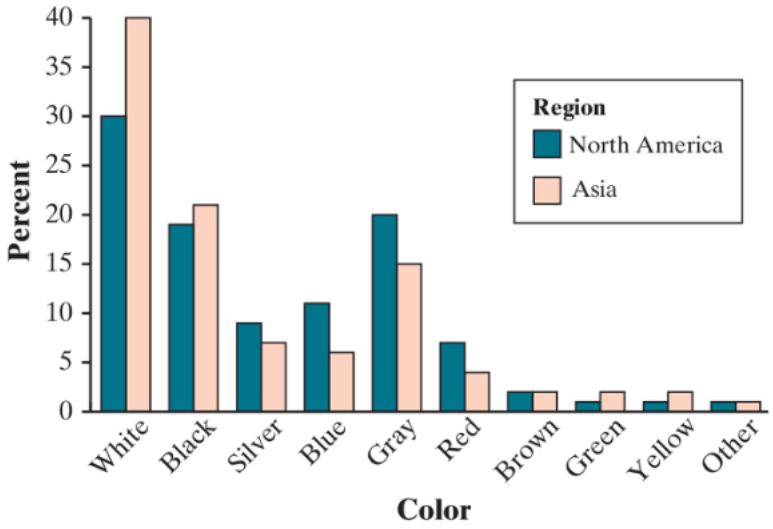
Example: Your Turn

A total of 55 modern Olympic Games have taken place since their inception in 1896. The Summer and Winter Olympic Games are usually held about every four years, excluding the Games cancelled by World War I and World War II. The frequency table summarizes the distribution of region of the world in which these events were hosted. Make a relative frequency bar chart to display the data. Describe what you see.

Region	Olympic Games hosted
Asia	8
Europe	32
North America	12
Oceania	2
South America	1
Total	**55**

Example: Your Turn

Favorite car colors may differ among regions of the world. The side-by-side bar chart displays data on the most popular car colors in a recent year for North America and Asia. Describe similarities and differences in the distributions for these two regions.



What is Joint Frequency?

A joint frequency is how many times a combination of two conditions happens together. For example:

- Pet owners (condition 1) who are women (condition 2),
- Democrats (condition 1) who are married (condition 2),
- Astronauts (condition 1) who are allergic to peanuts (condition 2),
- Lottery winners (condition 1) who go bankrupt (condition 2),
- Hurricanes (condition 1) that are category 5 (condition 2).

This kind of data is called bivariate data (data that has two inputs or variables). So it's sometimes called **bivariate joint frequency**.

Bivariate Data in Tables
Bivariate joint frequencies are displayed in the center of a frequency distribution table.

	Cats	Fish	Dogs	
Men	2	4	6	12
Women	5	3	2	10
	7	7	8	22

Joint frequencies highlighted in a frequency table.

It's called a "joint" frequency because you're looking at where two variables join. For example, the table above shows how many people (women and men) own which pets. The value at the "joint" of women cats is 5. Therefore, the joint frequency of women who own cats is 5. The edges, or totals of the table are called marginal distributions (i.e. they appear in the margins).

What is Joint Relative Frequency?
Joint relative frequency is the ratio of the frequency in a certain category and the total number of data points in that category. In the above table, 7 people own cats, and two of those are men. So the joint relative frequency of male cat owners is 2/7. Other information you can get from the table includes:

- Fish owners who are men is 4/7,
- Cat owners who are women is 5/7,
- Dog owners who are men is 6/8 = 1/4

Contingency Table

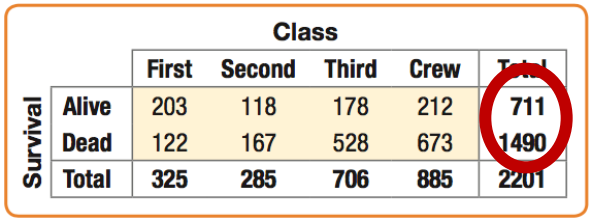
To look at two categorical variables together, we often arrange the counts in a two-way table. Because the table shows how the individuals are distributed along each variable, contingent on the value of the other variable, such a table is called a contingency table. The marginal distribution of one of the categorical variables in a two-way table of counts is the distribution of values of that variable among all individuals described by the table. The counts or percentages are the totals found in the margins of the table!

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

Marginal Distribution

The margins of the table, both on the right and at the bottom, give totals. The bottom line of the table is just the frequency distribution of ticket *Class*. The right column of the table is the frequency distribution of the variable *Survival*. These are the marginal distributions of the two variables.

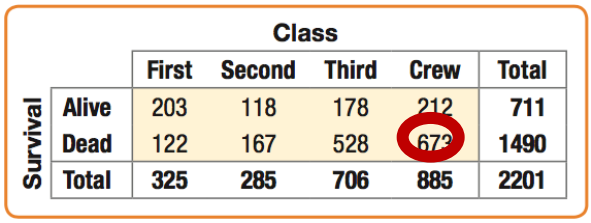
The marginal distribution of *Survival* is:



		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

Each **cell** of the table gives the count for a combination of values of the two values.

For example, the second cell in the crew column tells us that 673 crew members died when the *Titanic* sunk.



		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

Example 7

A survey of 4826 randomly selected young adults (aged 19 to 25) asked, "What do you think are the chances you will have much more than a middle-class income at age 30?" The table shows the responses, omitting a few people who refused to respond or who said they were already rich.

Opinion	Female	Male	Total
Almost no chance	96	98	194
Some chance but probably not	426	286	712
A 50-50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
Total	2367	2459	4826

SOLUTION

a) Calculate the marginal distribution (in percents) of opinions.

Almost no chance: $\frac{194}{4826} = 4.0\%$

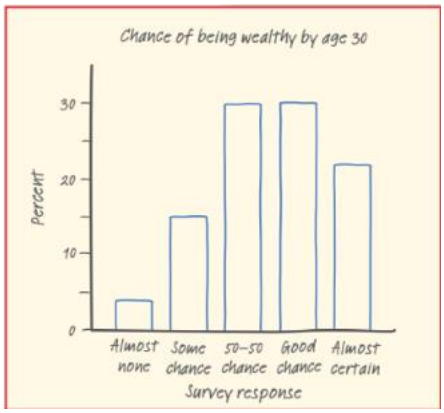
Some chance but probably not: $\frac{712}{4826} = 14.8\%$

50-50 chance: $\frac{1416}{4826} = 29.3\%$

A good chance: $\frac{1421}{4826} = 29.4\%$

Almost certain: $\frac{1083}{4826} = 22.4\%$

b) Construct a bar graph of the marginal distribution. Describe what you the marginal distribution tells you.



It seems that many young people are optimistic about their future income—over 50% of those who responded to the survey felt that they had a "good chance" or were "almost certain" to be rich by age 30.

Conditional Distribution

A conditional distribution of a variable describes the values of that variable among individuals who have a specific value of another variable. There is a separate conditional distribution for each value of the other variable.

Using our previous table,

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

What if we wanted to see the distribution of class under the condition of surviving or not. We will restrict the Who first to survivors and make a pie chart for them. Then we will refocus the Who on the non-survivors and make their pie chart.

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203 28.6%	118 16.6%	178 25.0%	212 29.8%	711 100%
	Dead	122 8.2%	167 11.2%	528 35.4%	673 45.2%	1490 100%

Example 8

Using the two-way table from Example 7, construct the conditional distribution of (a) opinion among women and (b) opinion among men.

SOLUTION

a.

Conditional distribution of opinion among women

Response	Female
Almost no chance	$\frac{96}{2367} = 4.1\%$
Some chance	$\frac{426}{2367} = 18.0\%$
A 50-50 chance	$\frac{696}{2367} = 29.4\%$
A good chance	$\frac{663}{2367} = 28.0\%$
Almost certain	$\frac{486}{2367} = 20.5\%$

b.

Conditional distribution of opinion among men

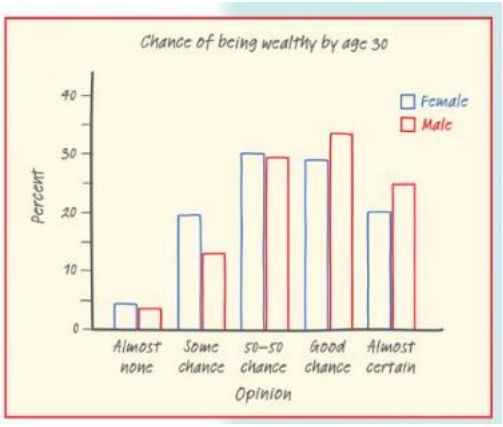
Response	Male
Almost no chance	$\frac{98}{2459} = 4.0\%$
Some chance	$\frac{286}{2459} = 11.6\%$
A 50-50 chance	$\frac{720}{2459} = 29.3\%$
A good chance	$\frac{758}{2459} = 30.8\%$
Almost certain	$\frac{597}{2459} = 24.3\%$

Example 9

Using Example 8, based on the survey data, can we conclude that young men and women differ in their opinions about the likelihood of future wealth?

SOLUTION

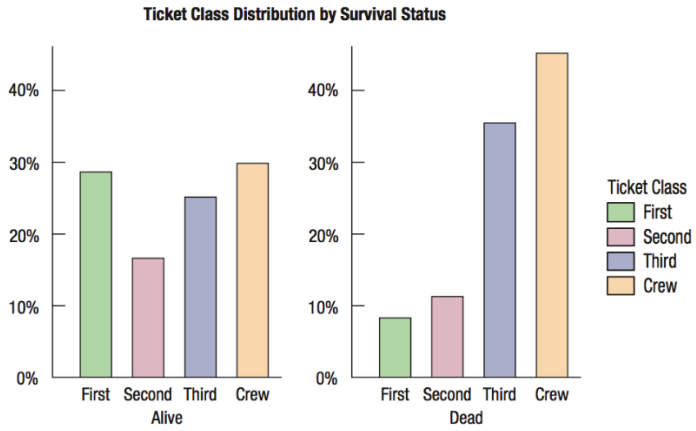
We suspect that gender might influence a young adult's opinion about the chance of getting rich. Thus, we will compare the conditional distributions of response for males and for females. Using the side-by-side bar graph, men seem somewhat more optimistic about their future income than women. Men (11.6%) were less likely to say that they have "some chance but probably not" than women (18%). Men (30.8%) were more likely to say that they have a "good chance" than women (28%) as well as men (24.3%) were most likely to say "almost certain" than women (20.5%).



Response	Female	Male
Almost no chance	$\frac{96}{2367} = 4.1\%$	$\frac{98}{2459} = 4.0\%$
Some chance	$\frac{426}{2367} = 18.0\%$	$\frac{286}{2459} = 11.6\%$
A 50-50 chance	$\frac{696}{2367} = 29.4\%$	$\frac{720}{2459} = 29.3\%$
A good chance	$\frac{663}{2367} = 28.0\%$	$\frac{758}{2459} = 30.8\%$
Almost certain	$\frac{486}{2367} = 20.5\%$	$\frac{597}{2459} = 24.3\%$

Variables can be associated in many ways and to different degrees. The best way to tell whether two variables are associated is to ask whether they are not. In a contingency table, when the distribution of one variable is the same for all categories of another, we say that the variables are independent. That tells us there's no association between these variables.

Back to our Titanic example, we see that the distribution of *Class* for the survivors is different from that of the non-survivors. This leads us to believe that *Class* and *Survival* are associated, that they are not independent. The variables would be considered **independent** when the distribution of one variable in a contingency table is the same for all categories of the other variable.



Example 10:

Using the date from Examples 8 and 9, are gender and opinion about future wealth independent?

SOLUTION:
We need to compare the conditional distribution of opinion among males to the conditional distribution of opinion among females. For example, 4.0% of males responded "almost no chance" while 4.1% of females also answered "almost no chance." Similarly, 11.6% of males said that they had "some chance" while 18.0% of females said that they had "some chance." Likewise, 29.3% of males and 29.4% of females said that

they had a "50-50 chance." 30.8% of males responded that they had "a good chance" while only 28.0% of females felt that they had "a good chance." Finally, 24.3% of males were "almost certain" while 20.5% of females were "almost certain." As each opinion type is different among males and females, gender and opinion about future wealth are NOT independent.

Response	Female	Male
Almost no chance	$\frac{96}{2367} = 4.1\%$	$\frac{98}{2459} = 4.0\%$
Some chance	$\frac{426}{2367} = 18.0\%$	$\frac{286}{2459} = 11.6\%$
A 50-50 chance	$\frac{696}{2367} = 29.4\%$	$\frac{720}{2459} = 29.3\%$
A good chance	$\frac{663}{2367} = 28.0\%$	$\frac{758}{2459} = 30.8\%$
Almost certain	$\frac{486}{2367} = 20.5\%$	$\frac{597}{2459} = 24.3\%$

Example: Your Turn

Examine the data about gender and favorite fast-food restaurant among a large Statistics class.

	Chick-fil-A	McDonalds	Burger King	Total
Male	72	17	24	113
Female	32	11	51	94
Total	104	28	75	207

1. What percent of the class are females who like Chick-fil-A?

2. What percent of those preferring Chick-fil-A are female?

3. What percent of the females prefer Chick-fil-A?

4. What is the marginal distribution of gender?

5. What is the marginal distribution of favorite fast-food restaurant?

6. The conditional distribution of favorite fast-food restaurant among males:

7. The conditional distribution of favorite fast-food restaurant among females:

8. What is the conditional relative frequency distribution of gender among those preferring Chick-fil-A?

If the conditional distributions are the same, we can conclude that the variables are not associated. Therefore, they are independent of one another.

If the conditional distributions differ, we can conclude that the variables are somehow associated. Therefore, they are not independent of one another.

9. Are gender and favorite fast-food restaurant independent? Explain.

Example: Your Turn

The Pew Research Center asked a random sample of 2024 adult cell-phone owners from the United States which type of cell phone they own: iPhone, Android, or other (including non-smartphones). Here are the results, broken down by age category:

		Age			
		18–34	35–54	55 +	Total
Type of phone	iPhone	169	171	127	467
	Android	214	189	100	503
	Other	134	277	643	1054
	Total	517	637	870	2024

(a) Find the distribution of type of cell phone for each age group. Make a segmented bar graph to compare these distributions.

(b) Describe what the graph in (a) reveals about the association between age and type of cell phone for adult cell-phone owners.

Simpson's Paradox

When averages are taken across different groups, they can appear to contradict the overall averages

Example

Consider this investigation of admission rates for men and women at the University of California at Berkeley's graduate schools. As reported in an article in *Science*, about 45% of male applicants were admitted, but only about 30% of female applicants got in. Looks like a clear case of discrimination, right? But let's break down the data by school (Engineering, Law, Medicine, etc). It turned out that, within each school, the women were admitted at nearly the same, or in some cases, much higher rates than the men. How could this be? Women applied in large numbers to schools with very low admission rates (Law and Medicine). Men tended to apply to Engineering and Science—schools that have admission rates above 50%. When the average was taken, the women had a much lower overall rate, but the average didn't really make sense.