Chapter 1
Statistics: The Art and Science of Learning from Data

SECTION 1.1: PRACTICING THE BASICS

1.1. Aspirin and heart attacks

a) Aspects of the study that have to do with design include the sample, the randomization of the halves of the sample to the two groups (aspirin and placebo), and the plan to obtain percentages of each group that have heart attacks.
b) Aspects having to do with description include the actual percentages of the people in the sample who have heart attacks (i.e., 0.9% for those taking aspirin and 1.7% for those taking placebo).
c) Aspects that have to do with inference include the use of statistical methods to conclude that taking aspirin reduces the risk of having a heart attack.

1.2 Poverty and race

a) The aspects referring to description are the percentages of the 50,000 households (14.7% of whites, 30.4% of blacks, and 11.1% of Asians) who had incomes below the poverty level.
b) The statistical method that predicted that the percentage of *all* black households in the United States that had income below the poverty level was between 28.6% and 32.2% is an example of inference.

💻1.3 GSS and heaven

Yes, definitely: 64.6%; Yes, probably: 20.8%; No, probably not: 8.7%; No, definitely not: 5.9%

💻1.4 GSS and heaven and hell

a) Yes, definitely: 64.3%; Yes, probably: 20.8%; No, probably not: 8.8%; No, definitely not: 6.0%
b) Yes, definitely: 52.6%; Yes, probably: 20.3%; No, probably not: 14.8%; No, definitely not: 12.3%; The percentage of "yes, definitely" responses was higher for belief in heaven in 2008.

💻1.5 GSS for subject you pick

The results for this item will be different depending on the topic that you chose.

SECTION 1.2: PRACTICING THE BASICS

1.6 Description and inference

a) With description, we are summarizing a group of numbers. We can use description with either samples or populations. With inferences, we use data from samples to make conclusions or predictions about populations. For example, if we ask a sample of adults how many pets they own, and take the mean number of pets, that number is a description. If we use that number to predict the mean number of pets owned by the whole population, the predicted mean (or the predicted range for the mean) would be an inference.
b) Descriptive statistics would be useful to summarize data from a population. With a census, it would be unwieldy to examine everyone's ages, for example, but it would be useful to know a mean age. Inferential statistics are not needed, however, because we already have information about the population; we don't need to predict it.

1.7 Number of good friends

a) The sample is the 840 respondents to the General Social Survey question, "About how many good friends do you have?"
b) The population is the American adult public.
c) The statistic reported is the percentage of respondents having only 1 good friend (i.e., 6.1%).

1.8 Concerned about global warming?

a) The sample is the set of polled Floridians. The population is the set of all adult Florida residents.
b) The percentages quoted are statistics since they are summaries of the sample.

1.9 EPA

a) The subjects in this study are cars – specifically, new Honda Accords.
b) The sample is the few new Honda Accords that are chosen for the study on pollution emission and gasoline mileage performance.
c) The population is all new Honda Accords.

1.10 Babies and social preference

a) The sample includes the 16 six-month-old infants who were shown the sequence of videos.
b) The population would be all six-month-old infants.
c) The inference is that six-month-old infants have both the ability to recognize and the preference to align themselves with the helpful figure from the video.

1.11 Graduating seniors' salaries

a) These are descriptive statistics. They are summarizing data from a population – all graduating seniors at a given school.
b) These analyses summarize data on a population – all graduating seniors at a given school; thus, the numerical summaries are best characterized as parameters.

1.12 At what age did women marry?

a) The mean age of 24.1 years for this sample is descriptive.
b) The historian estimates the age for the whole population of brides in early 19th century New England, estimating the average age to fall between 23.5 and 24.7. This is inferential.
c) The inference refers to the population of all brides between the years of 1800 and 1820.
d) The average of 24.1 years is based on a sample and is therefore a statistic.

1.13 Age pyramids as descriptive statistics

a) The bar graph for 1750 shows shorter and shorter bars as age increases indicating that there were few Swedish people who were old in 1750.
b) For every age range, the bars are much longer for both men and women in 2010 than in 1750.
c) The bars for women in their 70's and 80's in 2010 are longer than those for men of the same age in the same year.
d) The first manned space flight took place in 1961 so that people born during this era would fall in the 45-49 year old category. This is the largest five-year group for both men and women.

1.14 Gallup polls

Responses to this exercise will differ depending on the studies that students choose. (a) The descriptive statistic will be a summary of data, without any prediction or population estimate. It might be a mean rating for a given attitude, for example. (b)The inferential statistical analysis will have some kind of prediction or estimation; for example, the inferential statistic might include the margin of error for a mean, indicating that the population mean likely falls somewhere in a given range.

1.15 National service

a) The populations are the same for the two studies. Two separate samples are taken from the same population.
b) The sample proportions are not necessarily the same because the two random samples may differ by chance.

1.16 Samples vary less with more data

a) It would be more surprising to flip a coin 500 times and observe all heads.

b) As the sample size increases, the amount by which sample proportions tend to vary decreases. The estimates from larger samples, therefore, tend to be more accurate than estimates from smaller samples. When the coin is flipped just 5 times, it's easy to see that we could get a sample with all heads. However, when the number of flips is increased to 500, it is much more likely that the sample proportion is near the population proportion of 0.5. It would be extremely unlikely to observe very few heads or almost all heads in 500 flips of a fair coin.

SECTION 1.3: PRACTICING THE BASICS

1.17 Data file for friends

The results for this exercise will be different for each person who does it. The data files, however, should all look like this:

Friend	Characteristic 1	Characteristic 2
1		
2		
3		
4		

For each friend, you'll have a number or label under characteristics 1 and 2. For example, if you asked each friend for gender and hours of exercise per week, the first friend might have m (for male) under Characteristic 1, and 6 (for hours exercised per week) under Characteristic 2.

1.18 Shopping sales data file

Customer	Clothes	Sporting goods	Books	Music CDs
1	$49	$0	$0	$16
2	$0	$0	$0	$0
3	$0	$0	$0	$0
4	$0	$0	$92	$0
5	$0	$0	$0	$0

1.19 Sample with caution

A sample of individuals with children who read the Ann Landers column is not a random sample of individuals with children because every member of the population does not have the same chance of being in the sample. Many individuals with children may not read Ann Landers while others who do read the column may choose not to participate in the survey.

💻1.20 Create a data file with software

Your MINITAB data (from exercise 1.18) will be in the following format, although it will reside in the cells of the MINITAB worksheet.

Customers	Clothes	Sporting Goods	Books	Music CDs
1	49	0	0	16
2	0	0	0	0
3	0	0	0	0
4	0	0	92	0
5	0	0	0	0

💻1.21 Use a data file with software

See solution for 1.20 for format of data in MINITAB.

1.22 Simulate with the *sample for a population* applet

a) These will be different each time this exercise is completed.
b) Regardless of the specific graphs constructed in part a, you will see that the amounts by which sample percentages tend to vary get smaller as the sample size *n* gets larger.
c) The practical implication of this is that larger sample sizes tend to provide more accurate estimates of the true population percentage value.

1.23 Is a sample unusual?

It would be surprising to get a percentage that's more than 20 points from the true population percentage with a sample of 50 people. If you use the applet to conduct a simulation, you'll see that most of the time, the samples fall within 14 points of the true population percentage – from about 56 to 84.

CHAPTER PROBLEMS: PRACTICING THE BASICS

1.24 UW Student survey

a) The population is the entire UW student body of 40,858. The sample is the 100 students who were asked to complete the questionnaire.
b) This value would not necessarily equal the value for the entire population of UW students. It is quite possible that the sample of 100 is not exactly representative of the whole student body. This percentage is only an estimate of the percentage of all students who would respond this way. It is unlikely that any single sample of 100 would have a percentage that was exactly the percentage of the entire population.
c) The numerical summary is a sample statistic because it only summarizes for a sample, not for a population.

1.25 ESP

a) The population of interest is all American adults (the population from which this sample was taken).
b) The sample data are summarized by giving a proportion of all subjects (0.638) who said that they had at least one such experience, rather than giving the individual data points for all 3887 sampled subjects.
c) We might want to make an inference about the population with respect to the proportion who had had at least one ESP experience. We would use the sample proportion to estimate the population proportion.

1.26 Presidential popularity

This is an inferential statistic because Gallup was using the 45% of the sample who approved of how Obama is handling the presidency, along with the margin of error, to make a prediction about the population – how many Americans in general approved of how Obama was handling the Presidency.

1.27 Breaking down Brown versus Whitman

a) The results summarize sample data because not every voter in the 2010 California gubernatorial election was polled.
b) The percentages reported here are descriptive in that they describe the exact percentages of the sample polled who were Democrat and voted for Brown, who were Republican and voted for Brown and who were Independent and voted for Brown.
c) The inferential aspect of this analysis is that the exit poll results were used to predict what percentage of each of the three parties (Democrat, Republican and Independent) voted for Brown in the 2010 California gubernatorial election. The margins of error give a likely range for the population percentages for each of the three parties.

1.28 Reducing stress

a) The sample is the 100 students who were asked if they preferred to have a several-day period between the end of classes and the start of final exams. The population is all students in this school.
b) In this study, (i) descriptive statistics would give us information about the preferences of the 100 students in the sample, whereas (ii) inferential statistics would allow us to draw a conclusion about the preferences of the student body in general.

1.29 Marketing study

a) For the study on the marketing of digital media, the population is all Facebook users, and the sample is the 1000 Facebook users to whom the ad was displayed.

b) Example 4 suggests that we might determine that the average sales per person equaled $0.90. This would be a descriptive statistic in that it describes the average sales per person in the sample of 1000 potential customers. If one were to use this information to make a prediction about the population, this would be an inferential statistic.

1.30 Multiple choice: Believe in reincarnation?

b) inferential statistics.

1.31 Multiple choice: Use of inferential statistics?

c) to make predictions about populations using sample data

1.32 True or false?

False. We often want to describe the sample AND make inferences about the population.

CHAPTER PROBLEMS: CONCEPTS AND INVESTIGATIONS

1.33 Statistics in the news

If your article has numbers that summarize for a given group (sample or population), it's using descriptive statistics. If it uses numbers from a sample to predict something about a population, it's using inferential statistics.

1.34 What is statistics?

This answer will be different depending on the question chosen by the student.

1.35 Surprising ESP data?

This result would be very surprising with such a large sample. You'll notice that when you use the applet to simulate this study, you will get a sample proportion as large as 0.638, when the true proportion is 0.20, only VERY rarely. With such a large sample, if randomly selected, you'd expect a sample proportion very close to the population proportion.

1.36 Create a data file

See solution for Exercise 1.20 for format of data in MINITAB.

CHAPTER PROBLEMS: STUDENT ACTIVITIES

1.37 Activity 3

Answers will vary.

Chapter 1 Statistics: The Art and Science of Learning from Data

1.1 Using Data to Answer Statistical Questions

What Is Statistics?
The art and science of learning from data – designing studies, analyzing the data, and translating data into knowledge and understanding of the world around us.

Who Uses Statistics?
Everyone.

- Marketing – What color package sells the most? On which shelf should we place kid's cereal in a supermarket?
- Medical Studies – Is the new treatment for a disease better than the old one? Is drinking alcohol in moderation good or bad for your blood pressure?
- Government – What is the unemployment rate for each state? Who will win the next election?
- Social Scientists – What percentage of people think that abortion, or homosexual marriage should be legal? Which groups are more likely to think that way?
- Environmental Studies – How many animals of an endangered species are left in this region? What conditions can be changed to increase their numbers?
- Engineering – How many miles per gallon does a car model get? How much variability is there in the contents of 12 oz cans of soda?
- Sports – Who is the best baseball pitcher? Are there "hot streaks" in basketball?

Why Use Statistical Methods?

- **Design** – plan how to obtain the data (Ch 4).
- **Description** – summarize the data with graphs and numerical summaries (Ch 2-3).
- **Inference** – using data from a **random** and **representative sample** to draw conclusions about the **population** of interest. We need to understand some probability concepts to do statistical inference – this allows us to attach a measure of reliability to our inference or conclusion. (We will learn about probability in Ch 4-5 and about statistical inference in Ch 8, 9 and 10.)

1.2 Sample vs. Population

Data sets consist of:
Subjects - persons, animals, or objects in our study / experiment.
Variables - the characteristics that we measure on each subject. We call them variables because they can take on different values for each individual.

Population – all subjects of interest.
Sample – subjects for whom we have data.

Random Sampling – each member of the population has the same chance of being included in the sample. Random samples tend to be representative of the population, so we can draw better conclusions about it.

Parameters – numerical summary of the **population**.
Statistics – numerical summary of the **sample**.

Example: Internet sites report that about 13% of Americans are left-handed. Is this true for students at your university? During a statistics exam, the instructor walks around the room and counts 15 left-handed students out of 98 students in the class.

a) Identify the following:

Variable: dominant hand – left or right

Population: all students at your university

Sample: 98 students in the class

Parameter: 13% of the population is assumed to be left-handed

Statistic: 15/98 = 15% of the sample is left-handed

b) Was random sampling used?

Not quite. All students in the same class may have something in common – major, year in school, interests. Also, some theories associate left-handedness with mathematical genius – although other theories associate it with artistic or sports abilities, as well as other characteristics. So, the one class (sample) may not be representative of the whole population (university) and any inferences or conclusions we draw from this sample may be misleading.

Chapter 2 Exploring Data with Graphs and Numerical Summaries

2.1 Different Types of Data

Categorical Variables – place each observation into groups and they are usually summarized by the **percentage** of observations in each group.

Quantitative Variables– take on numerical values. Key features are the center (**average**) and spread (**variability**) of the data. Quantitative variables can be further split up into:

- **Discrete –** take only a finite list of possible outcomes, such as a count (0, 1, 2, 3, etc.)
- **Continuous –** has an infinite list of possible values that form an interval, even though sometimes we are limited in our ability to measure them

Example: The class list below has all the information for each student in a class at the end of the semester, including their year in school, major, exam grades, project grades, number of absences, their average in the class and their final letter grade in the class.

Student ID#	Name	Yr	Major	Ex1	Ex2	Pr1	Pr2	Abs	Avg	Grade
46895382	Aiken, John	1	Psych	78	82	20	24	2	81.6	B
21657845	Bailey, Kim	2	PolSci	62	74	15	19	10	68.0	D
13695544	Carr, May	2	BusAdm	95	92	25	24	0	94.4	A
	…etc.									

Which of the previous variables are:

Categorical	Discrete Quantitative	Continuous Quantitative
Major	Year in School	Exams
Grade	Absences	Projects
		Average

Note that Name and Student ID # are NOT variables, but student identifiers. We could compute the average Student ID # in the class, but what for?

2.2 Graphical Summaries of Data

The type of graph used depends on the type of variable.
Most graphs are done with a computer, particularly for large data sets.

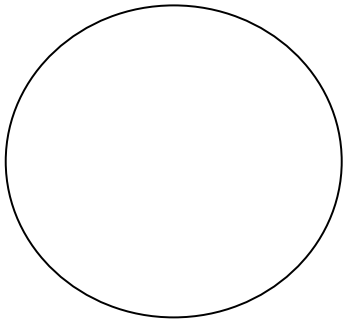
Graphs for Categorical Variables: Bar Charts and Pie Charts

Example: Year in school for students in classroom
NOTE: Collect data in your OWN class – answers will vary.

	Frequency (Count)	Proportion	Percentage
Freshman			
Sophomore			
Junior			
Senior			
Total			

Bar Chart

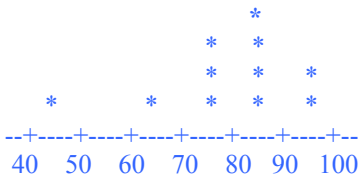
Pie Chart



Graphs for Quantitative Variables: Dotplots, Stem-and-Leaf Plots, and Histograms

Example: Grades on an exam for a small class:
82, 76, 65, 94, 72, 80, 91, 45, 72, 86, 89

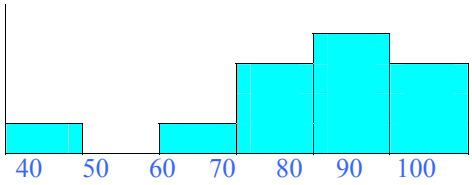
Dotplot



Stemplot

4 | 5
5 |
6 |5
7 | 2 2 6
8 | 0 2 6 9
9 | 1 4

Histogram



For this data set, what can you say about:

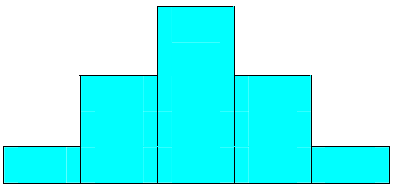
a) the center of the distribution? In the low 80's, upper 70s

b) the spread of the distribution? From 45 to 94 (or 40's to 90's)

c) the shape of the distribution? Roughly bell-shaped

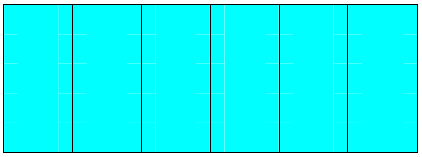
d) unusual observations? 45 seems unusually low

Some common shapes:
Mound or Bell-shape



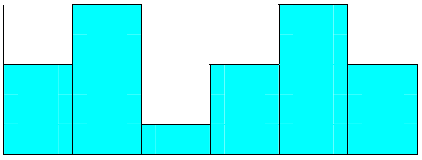
Unimodal =one bump, roughly symmetric
Ex: heights, weights, lengths

Uniform or Rectangular



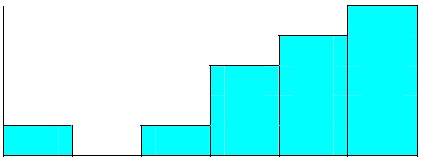
Flat – everything equally likely
Ex: rolling dice

Bimodal



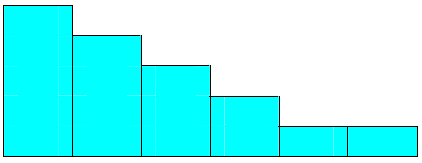
Bimodal = 2 bumps
Combining two groups
Ex: heights of men and women

Skewed Left



Long tail to the left side
Limit on the right side but none on left
(or left limit is very far away)
Ex: grades on an easy exam

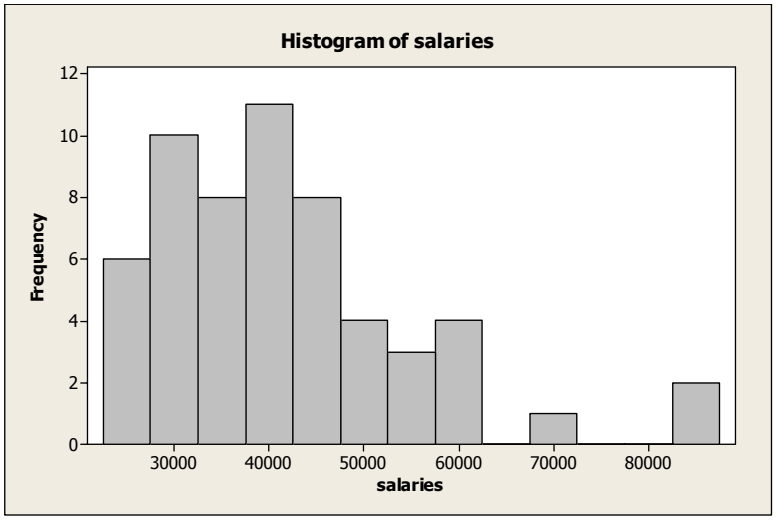
Skewed Right



Long tail to the right side
Limit on the left side but none on right
(or right limit is very far away)
Ex: salaries

Examples: For the following graphs, describe their distributions in terms of shape, center, and spread.

1. Graph of starting annual salaries for engineering graduates.

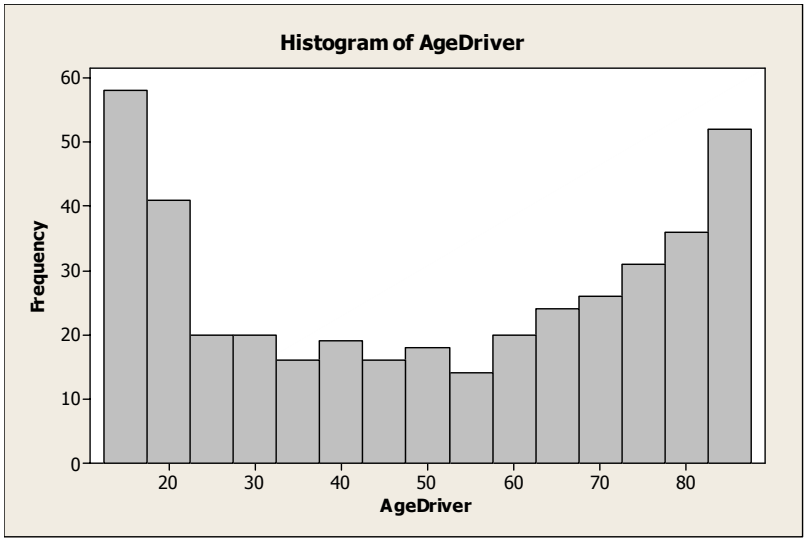


Shape: Skewed Right (long RIGHT tail)

Center: average salary is about $40,000
Spread: salaries vary from about $25,000 to $85,000

Outliers: there are some high salaries, separated from the rest of the distribution by gaps. These are not necessarily outliers, but the long tail expected in a skewed right distribution. We could look into them to determine.

2. What age group has the most dangerous drivers? The following graph represents the number of driver deaths (per 100,000 licensed drivers) by age group, in one year.

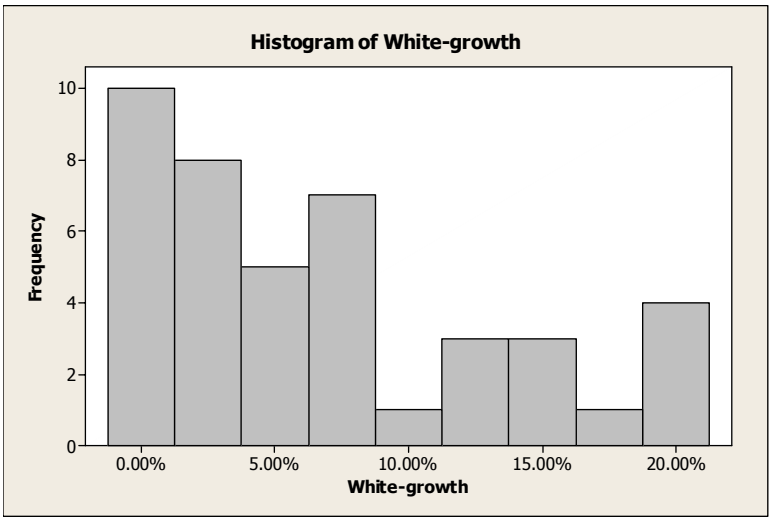


Shape: Bimodal

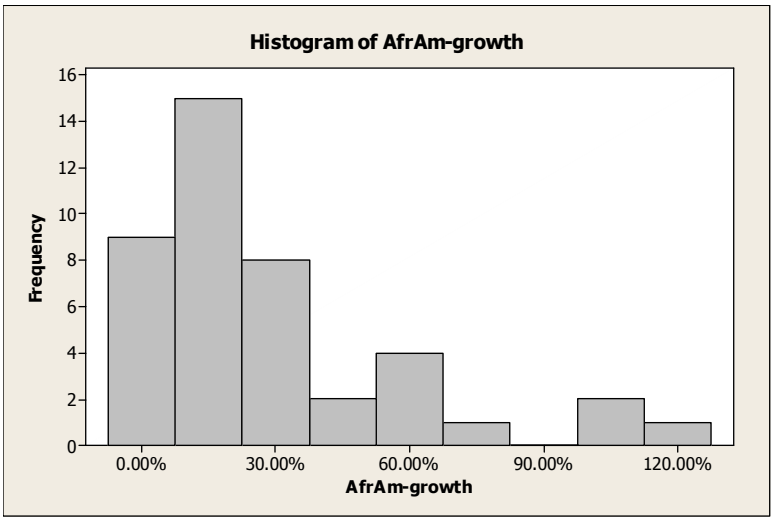
Center: average age is in the 50's, but that is not a typical value – better to use the modes – teenagers and older drivers have the most deaths.

Spread: age goes from teens to over 80 yrs old

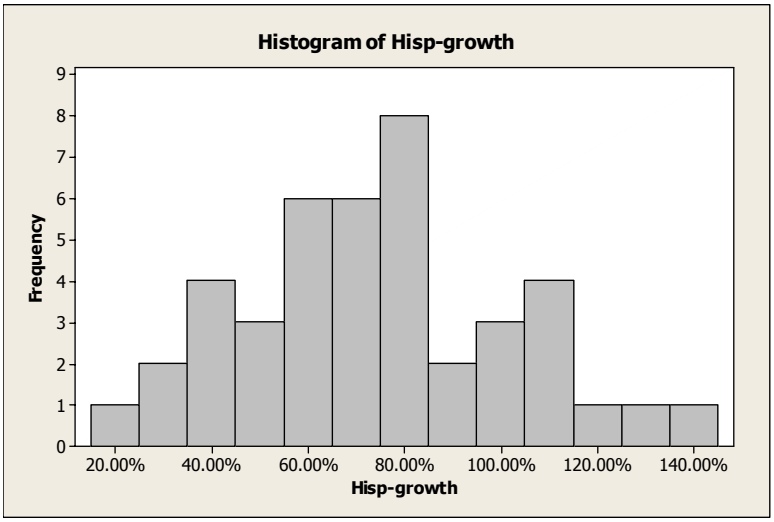
Examples: The following graphs present the population growth, by race, for each state when comparing the numbers from the 2010 census to the 2000 census. The data for each state appears on the table in the next page. Interpret the graphs' findings.



Most states have very small or no growth in their percentage of whites in the population, with an average of around 5%. The distribution is skewed right, with a range from 0% (or negative growth in some cases) to about 20%. There are four states with white growth around 20%: Nevada, Texas, Arizona and Utah



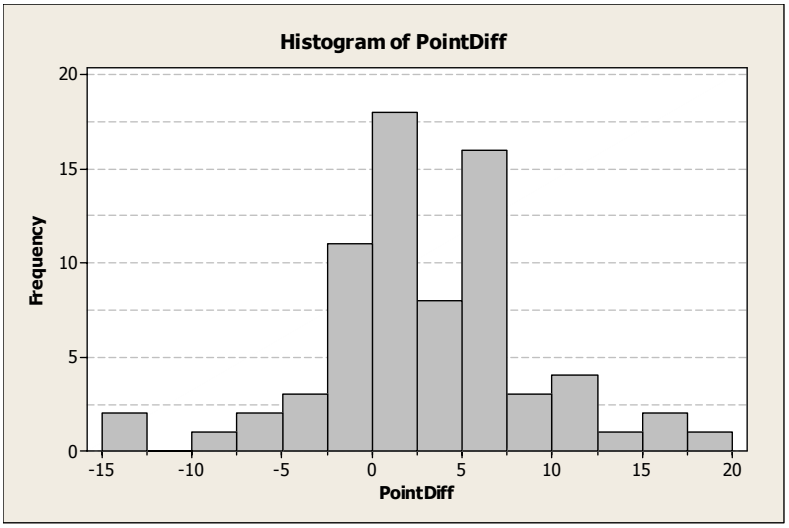
Most states saw an increase in their African American population, with the average increased being around 15%. Most states saw increases between 0% (little or no growth, or even slight decrease for states like Hawaii and Illinois) to 30%. The distribution is skewed right. A few states have doubled or more their percentage of African Americans, but they may have had very small percentages to begin with (South Dakota, North Dakota and Vermont.)

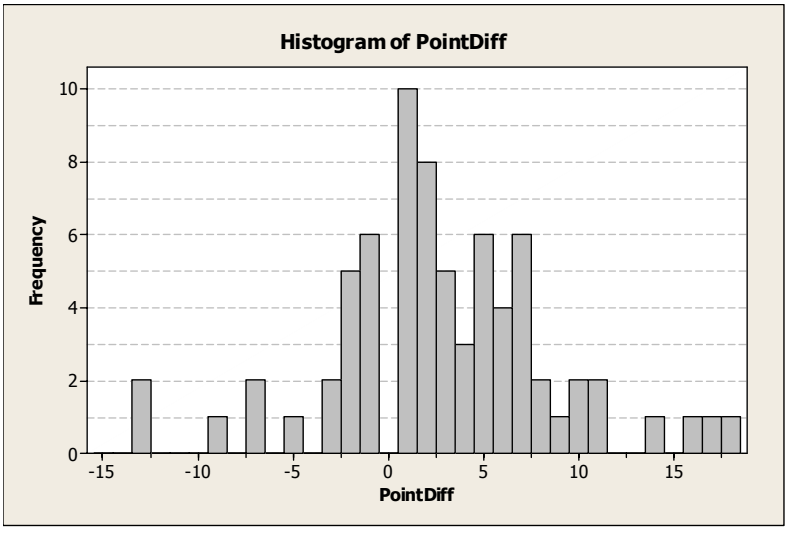


All states saw an increase in their Hispanic population , with the percentage of Hispanics per state almost doubling on average (increased by 80%). Increases ranged from 20% to 140% in a fairly symmetric distribution.

State	White-growth	AfrAm-growth	Hisp-growth	State	White-growth	AfrAm-growth	Hisp-growth
Alabama	3.60%	8.30%	144.80%	Missouri	4.40%	10.20%	79.20%
Alaska	9.00%	6.80%	51.80%	Montana	8.30%	49.60%	58.00%
Arizona	20.50%	63.00%	46.30%	Nebraska	2.60%	20.90%	77.30%
Arkansas	5.00%	7.40%	114.20%	Nevada	19.00%	61.40%	81.90%
California	6.40%	1.60%	27.80%	New Jersey	-1.20%	5.50%	39.20%
Colorado	14.90%	22.20%	41.20%	New Mexico	15.90%	23.90%	24.60%
Connecticut	-0.30%	16.90%	49.60%	North Carolina	12.50%	17.90%	111.10%
Delaware	5.80%	27.30%	96.40%	North Dakota	2.10%	103.30%	73.00%
Florida	13.20%	28.40%	57.40%	Ohio	-1.10%	8.20%	63.40%
Georgia	8.60%	25.60%	96.10%	Oklahoma	3.00%	6.40%	85.20%
Hawaii	14.40%	-2.60%	37.80%	Oregon	8.20%	24.30%	63.50%
Idaho	18.60%	79.80%	73.00%	Pennsylvania	-0.70%	12.50%	82.60%
Illinois	0.60%	-0.60%	32.50%	South Dakota	4.50%	117.90%	102.90%
Indiana	2.80%	16.00%	81.70%	Tennessee	7.90%	13.30%	134.20%
Iowa	1.20%	44.10%	83.70%	Texas	19.60%	23.90%	41.80%
Kansas	3.30%	8.90%	59.40%	Utah	19.40%	65.90%	77.80%
Kentucky	4.60%	14.00%	121.60%	Vermont	1.20%	104.90%	67.30%
Louisiana	-0.70%	0.00%	78.70%	Virginia	7.20%	11.60%	91.70%
Maryland	-0.90%	15.10%	106.50%	Washington	7.80%	26.20%	71.20%
Minnesota	2.80%	59.80%	74.50%	Wisconsin	2.80%	18.00%	74.20%
Mississippi	0.50%	6.20%	105.90%	Wyoming	12.50%	27.60%	58.60%

Example: The following histograms represent use different bin sizes on the x-axis to describe the point differential (points scored by the Gators minus points scored by the opponent) for each game played by the Florida Gators Baseball team during the 2011 season.





a) How many games did the Gators play? About 72

b) How many games did the Gators lose? 19

c) How many games resulted in a tie? None – in baseball you go to extra innings until someone wins.

d) How many games did the Gators win by 10 runs or more? 8

e) How many games did the Gators lose by 10 runs or more? 2

2.3 Measuring the Center of Quantitative Data

Measures of Center

- **Mean** – the average of all the observations: $\bar{x}$
- **Median** – observation "right in the middle": M
- **Mode** – most frequently occurring value

Mathematical Definitions:
Notation

n = # of observations in the data set
$x_1, x_2, x_3, \ldots, x_n$ = first, second, third, …, last observation
summation notation (capital sigma)

Formulas:

Mean: Answer: $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x}{n} = \frac{1}{n}\sum x$

Median:

- Order the observations from smallest to largest.
- The median is the observation in the middle position – that is position (n+1)/2.

Example: Number of "friends" in Facebook (an online social networking service for college students), for a sample of 9 female Facebook members.

288 254 476 329 191 121 404 184 505

a) Find the mean: (288 + 254 + . . . 505) / 9 = **305.78**

b) Find the median: 121 184 191 254 | **288** | 329 404 476 505
n=9 so the median is in position 10/2=5 | **Median** |

Example: Number of "friends" in Facebook (an online social networking service for college students), for a sample of 10 male Facebook members.

65 342 345 172 46 458 106 153 244 810

a) Find the mean: (65 + 342 + . . . + 810) / 10 = **274.1**

b) Find the median: 46 65 106 153 172 | 244 342 345 458 810
n=10 so the median is in position 11/2=5.5 **Median** = (172 + 244) / 2 = **208**

Example: The following is a stemplot done in Minitab of self-reported college GPAs of students enrolled in a large introductory statistics course. Find the median of this data set.

Stem-and-Leaf Display: cGPA

```
Stem-and-leaf of cGPA  N  = 114
Leaf Unit = 0.10

 2    2  01
 3    2  2
 5    2  55
 8    2  677
 16   2  88899999
 39   3  000000000000000011111111
 53   3  22222222223333
(24)  3  444444444445555555555555
 37   3  66666666777777777777
 17   3  8888888999
 7    4  0000000
```

Median = 3.4 (median in position 115/2=57.5, so average of observations in positions 57 and 58)

Notes:

- First column represents the cumulative counts from top and bottom.
- The line with the parentheses contains the median.
- Second column contains the stems.
- Rest of the columns contain the leaves.

- Leaf unit:

  0.10 decimal between stem and leaf
  Example from the first row: 2 01 means 2.0 and 2.1

  1.0 no decimal
  Example from the first row: 2 01 means 20 and 21

  10.0 add a zero after the leaf
  Example from the first row: 2 01 means 200 and 210

  0.01 move decimal to left of stem
  Example from the first row: 2 01 means 0.20 and 0.21

2.4 Measuring the Variability of Quantitative Data

Measures of Variability (Also Called Measures of Spread or Dispersion)

Range = maximum – minimum

Example: Find the range of the distributions of the number of "friends" in Facebook for the samples of female and male Facebook members. Data appears below in order.

Female: 121 184 191 254 288 329 404 476 505

Range = 505 - 121 = 384

Male: 46 65 106 153 172 244 342 345 458 810

Range = 810 – 46 = 764

Variance and Standard Deviation

Measures of spread around the mean, particularly useful for bell-shaped and symmetric distributions.

Variance:

- averaged squared deviation from the mean: s^2
- its units of measurement are those of the original data squared
- we need to take the square root before we interpret

Standard Deviation:

- square root of the variance: s
- its units of measurement are the same as those of the original data

Mathematical Definitions:

Recall Notation:

n = # of observations in the data set

$x_1, x_2, x_3, \ldots, x_n$ = first, second, third, …, last observation

summation notation (capital sigma)

Formulas:

Variance: Answer: $s^2 = \dfrac{\sum (x - \bar{x})^2}{n-1}$

Standard Deviation: Answer: $s = \sqrt{s^2}$

Why n-1 in the denominator? The denominator in the formula for the variance, n-1, represents the number of degrees of freedom (df). This is the number of independent quantities we are adding up in the numerator. Since $\bar{x}$ is computed from those n observations, only n-1 of the distances from the observations to the mean are independent of each other.

Example: Suppose your whole grade in a class is based on four exams, worth 100 points each. Here are the grades you know:

Ex 1	Ex2	Ex3	Ex4	Avg
82	76	85	??	80.5

How many possible values of Exam 4 grade are there?

Only one, so the n^{th} (last) observation is not free to vary – we have lost one degree of freedom by knowing the average of the observations.

Examples: Two very simple data sets. For each one, make a quick plot of their distribution, find the mean, median, and range. Compare the two distributions. Then find the standard deviation using the formula, and also using the statistical functions of your calculator.

Data Set 1: 1 1 1 4 7 7 7

Make Dotplots.

Mean: $\bar{x} = 4$
Median: M=4
Range = 6
Shape: Bimodal – points on the extremes. Expect larger stdev.

x	$x-\bar{x}$	$(x-\bar{x})^2$
1	-3	9
1	-3	9
1	-3	9
4	0	0
7	3	9
7	3	9
7	3	9
	0	54

Variance: $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$ = 54/6 = 9

Standard Dev: s = 3

Data Set 2: 1 3 4 4 4 5 7

Make Dotplots.

Mean: $\bar{x} = 4$
Median: M=4
Range = 6
Shape: bell shaped and symmetric – points close to mean. Expect smaller stdev.

x	$x-\bar{x}$	$(x-\bar{x})^2$
1	-1	1
3	-3	9
4	0	0
4	0	0
4	0	0
5	1	1
7	3	9
	0	20

Variance: $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$ = 20/6 = 3.333

Standard Dev: s = **1.83**

Interpreting the Standard Deviation, s:

- The larger the standard deviation, s, the more spread out the data set is.
- s can never be negative.
- s can only be zero if there is no variability in the data – all the observations are identical.
- s is very much affected by outliers.
- s works best for bell-shaped and symmetric distributions.

Empirical Rule - in any bell-shaped and symmetric distribution you will find approximately:

- 68% of the observations within one stdev of the mean.
- 95% of the observations within two stdev of the mean.
- 99.7% of the observations within three stdev of the mean.

Example: One of the many different scales for IQ scores states that the IQ values for the whole population follow a bell-shaped and symmetric distribution with mean 100 points and standard deviation 16 points.

a) Sketch a graph of this distribution.

Draw a bell-shaped distribution, mark the mean at 100 and mark off 1 stdev on either side of the mean (84, 116), 2 stdev (68, 132) and 3 stdev (52, 148).

b) Between what two values will you find the central:

68% of IQs? 84 -116

95%? 68 - 132

99.7%? 52 -148

c) What percentage of people have IQs over 148?

0.15% or 15 out of 10,000

Example: Heights of college-aged women are approximately bell-shaped and symmetric, with almost all of them (95%) falling between 4'11" (59") and 5'11"(71").

a) Sketch a graph of this distribution.

Draw a bell-shaped distribution and mark off 2 stdev away from the mean.

b) Find the approximate values of the mean and standard deviation.

Mean: 65" Stdev: 3"

2.5 Describing Distributions Using Percentiles and Boxplots

Quartiles divide the data set into four quarters:

- Q1 = Lower Quartile (Q_L)
  = 25% of observations below it = 25th percentile
- Q2 = Median
  = 50% of observations below it = 50th percentile
- Q3 = Upper Quartile (Q_U)
  = 75% of observations below it = 75th percentile

There are many ways to find the quartiles, each resulting in slightly different numbers.
For our purposes,

- Q1 will be the median of the lower half of the data.
- Q3 will be the median of the upper half of the data.

Example: Find the quartiles of the distributions of the number of "friends" in the Facebook for the samples of female and male Facebook members. Data appears below in order.

Female: 121 184 191 254 288 329 404 476 505
Q1 = 187.5 M=Q2 Q3 = 440

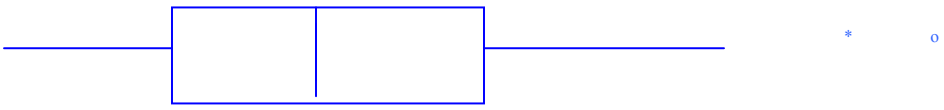
Male: 46 65 106 153 172 244 342 345 458 810
Q1 M=Q2=208 Q3

InterQuartile Range – measures the spread of the central 50% of the data.
IQR = Q3 – Q1

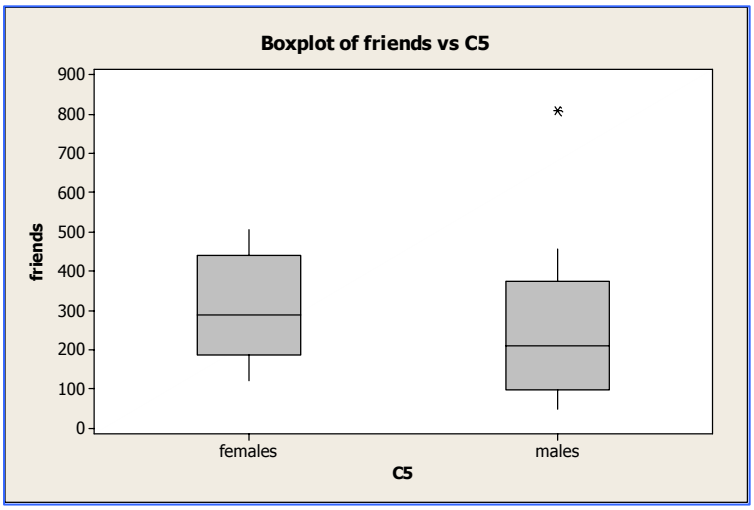
Five Number Summary of Positions:
Minimum, Q1, Median, Q3, Maximum

Boxplots - graphs based on the five number summary.

- Box contains the central 50% of the data (going from Q1 to Q3). Line crossing the box represents the median. Whiskers extend to the minimum and maximum observations.
- In Minitab (and many books and computer packages) the whiskers only extend to the smallest and largest observations that are not considered outliers. The outliers are plotted individually, with an asterisk if it's a "mild" outlier, and an open circle for an "extreme" one.

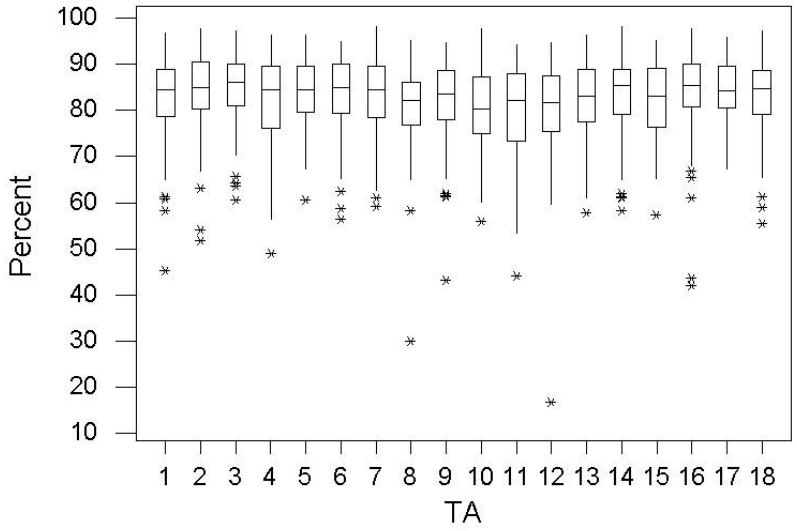


Example: Construct boxplots for distributions of the number of "friends" in Facebook for the samples of female and male Facebook members. Compare the distributions in terms of their most important features.



On average, females have slightly more "friends" than males (females: median~300, males: median~200). The number of "friends" for females is a bit less variable than that for males. Both distributions are fairly symmetric, except for the outlier -- one male who had an exceptionally large number of friends (~800).

Example: A large statistics course has 18 TAs in charge of the discussion sections. Do final grades in the class depend on who the TA for the section was? The boxplots below show the final percentage grade in the class for all students who took the final exam, by TA in charge of the section.



All TA's have similar median grades, in the 80's, with fairly similar variabilities. All of them have a few low outliers. Therefore, it does not matter who the TA for each section happens to be, since it will not have a large influence on final grades in the class.