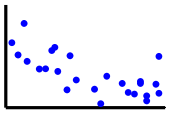


TOPIC D4: COMPARING BATCHES AND RELATIVE STANDING

SPOTLIGHT: WHERE'S THE BEST PLACE TO LIVE?

Where is an ideal place to live in America? Perhaps you wish to live in a dream house in suburbs of a large city. But that might mean that you have a high cost of living and commute one hour to work. Maybe instead you would prefer to live in a small town where you are close to work. But this small town might have limited opportunities for nightlife, restaurants, theater, or activities for children. Many people prefer to live in warm climates such as Florida, but other people like colder climates such as Colorado where there are opportunities to engage in winter sports such as ice skating and skiing.

The book *Cities Ranked & Rated* recognizes that people have different goals, needs, aspirations and interests, and these qualities impact the choice of desirable locations to live. This book considers four broad categories that people may use when evaluating a possible place to relocate: economy, cost of living, climate, and character. Economy refers to the economic health and commercial aspects of a place. Cost of living refers to the costs of housing and necessities and tax burden. The character of a place refers to the area's "look and feel," its activities and services, and any negative aspects such as crime and health problems.

This book gives a rich set of facts and figures for 403 North American cities. Using these measurements, the book gives a numerical rating and ranking of the cities with respect to economy and jobs, cost of living, climate, education, health and healthcare, crime, transportation, leisure, arts and culture, and quality of life. It may not surprise you that New York City has the top rating with respect to leisure and Boston has the top-rated education. But did you know that the metro area with the strongest economy and jobs is Billings, Montana, and the least-expensive metro area is Casper, Wyoming? The top-rated area with respect to quality of life (including physical attractiveness, heritage, friendliness of residents and overall ease of living) is Madison,

Wisconsin. In this book, we will use some of the data collected in this book to compare groups of cities with respect to different characteristics.

PREVIEW

Although methods of graphing and summarizing a single batch of data are useful, much of data analysis is involved with comparison of batches. We know how to compare two individuals by simply taking a difference of their individual values, such as "Susie scored 10 points higher than Joe on the math test." We would like to make similar comparative statements when we compare two or more batches of data, and this topic will describe how that can be done. Also, we'll discuss a method for describing one's relative standing in a collection of quantitative data and introduce a rule of thumb for detecting outliers.

In this topic your learning objectives are to:

- Understand how one can compare two batches of categorical data by the use of graphs and the computation of percentages.
- Understand what it means for one batch of quantitative data to be larger than another batch of data.
- Understand how one can summarize and display a batch of quantitative data by the use of five numbers.
- Understand when it is appropriate and inappropriate to compare two batches of quantitative data.
- Understand how a standardized score measures the relative standing of an observation, and understand how one can identify unusually small or large data values.

COMPARING BATCHES OF CATEGORICAL DATA

In topics D2 and D3, we looked at the country of origin of Major League baseball players who were born in the year 1975. We found that approximately two-thirds of the

players were born in the United States, but a sizeable proportion of players came from Latin America.

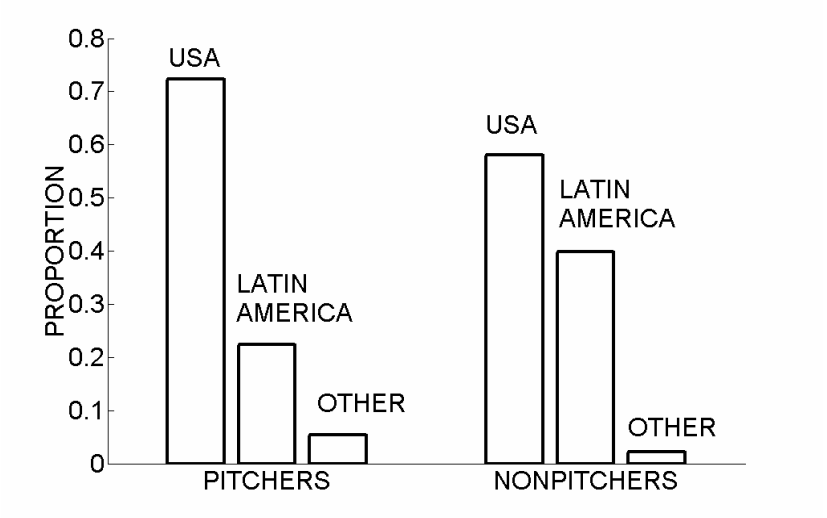
Looking at these data more carefully, it is natural to ask: do players from other countries excel in particular positions in baseball? Specifically, are there differences in the country of birth distribution between pitchers and nonpitchers?

To answer the question, we revisit our data and divide the 205 players into two groups – the 112 players who are pitchers and the 93 players who are nonpitchers. For each group, we categorize the players by the country of origin (USA, Latin American, or Other). We present these two frequency tables below. In addition, for each group, we find the proportion and percentage for each category.

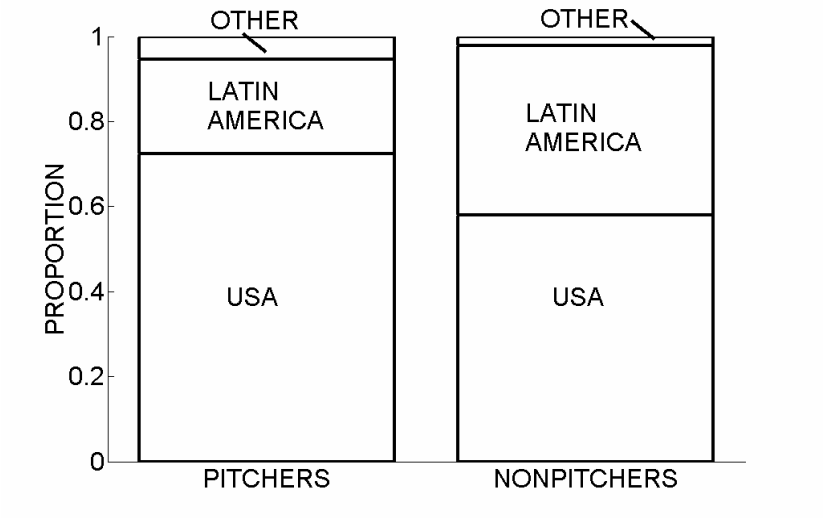
Pitchers				Nonpitchers		
Frequency	Proportion	Percentage	COUNTRY	Frequency	Proportion	Percentage
81	0.723	72	USA	54	0.581	58
25	0.223	22	Latin America	37	0.398	40
6	0.054	5	Other	2	0.022	2
112	1	100	TOTAL	93	1	100

To compare the country of origins of the pitchers and nonpitchers, we'd like first to construct a graph. In topic D2, we used a bar chart and a segmented bar chart to display a single frequency table of categorical data. These same graphs can be used to compare several frequency tables.

We first show side-by-side bar charts of the pitcher and nonpitcher tables. Since the numbers of pitchers and nonpitchers are not equal, it is best to plot the proportions (or percentages) of the two tables – in that way, the total proportion is equal to one for both groups and it is easier to make comparisons. We see differences – there are over twice as many USA pitchers than Latin American pitchers, while the proportions of USA and Latin American nonpitchers are similar in size. It appears that pitchers are more likely to be American than nonpitchers.



A similar graphical comparison can be made by use of segmented bar charts. In the figure below, we show side-by-side segmented bar charts of the category proportions. As in the earlier display, it is best to graph the proportions (instead of the frequencies) since the sum of proportions is equal to one for both tables. From this display, we clearly see the differences in the country of birth between the pitchers and nonpitchers.



Now that we have noticed that there is a difference in the country of birth in the two groups of players, how can we summarize this difference? If we focus on the percentages in the table, we read that 22% of the pitchers are Latin American in origin, and 40% of the nonpitchers are Latin American. If we look at the difference, we can say

that the proportion of Latin Americans among the nonpitchers exceeds the proportion of Latin Americans among the pitchers by 40 – 22 = 18%.

PRACTICE: COMPARING BATCHES OF CATEGORICAL DATA

In the book *Cities Rated & Ranked*, 331 metropolitan areas are ranked from best to worst on the basis of quality of life. The top 30 and the bottom 30 metropolitan areas are listed in the table below. In addition, the table gives the population density (LOW or HIGH compared to the U.S. average) and the population growth in the 1990-2002 (LOW or HIGH compared to the U.S. average).

TOP 30 U.S. Metropolitan Areas

	STATE	DENSITY	GROWTH
Charlottesville	VA	LOW	HIGH
Sante Fe	NM	LOW	HIGH
San Luis Obispo-Alascadero-Paso Robles	CA	LOW	HIGH
Santa Barbara – Santa Maria – Lompoc	CA	LOW	LOW
Honolulu	HI	HIGH	LOW
Ann Arbor	MI	LOW	HIGH
Atlanta	GA	HIGH	HIGH
Asheville	NC	LOW	HIGH
Reno	NV	LOW	HIGH
Corvallis	OR	LOW	LOW
Roanoke	VA	LOW	LOW
Portland – Vancouver	OR-WA	LOW	HIGH
Raleigh-Durham-Chapel Hill	NC	LOW	HIGH
Bryan-College Station	TX	LOW	HIGH
Lynchburg	VA	LOW	LOW
Olympia	WA	LOW	HIGH
Norfolk-Virginia Beach-Newport News	VA-NC	HIGH	LOW
Colorado Springs	CO	LOW	HIGH
Nassau-Suffolk	NY	HIGH	LOW
Pueblo	CO	LOW	HIGH
Eugene-Springfield	OR	LOW	LOW
Austin-San Marcos	TX	LOW	HIGH
Lafayette	IN	LOW	LOW

BOTTOM 30 U.S. Metropolitan Areas

	STATE	DENSITY	GROWTH
Macon	GA	LOW	LOW
Owensboro	KY	LOW	LOW
Jackson	TN	LOW	HIGH
Wheeling	WV-OH	LOW	LOW
Joplin	MO	LOW	HIGH
Racine	WI	HIGH	LOW
Sharon	PA	LOW	LOW
Erie	PA	LOW	LOW
Dutchess Country	NY	LOW	LOW
Dubuque	IA	LOW	LOW
Waterbury	CT	HIGH	LOW
Lewiston-Auburn	ME	LOW	LOW
Brownsville-Harlingen-San Benito	TX	LOW	HIGH
Yuba City-Marysville	CA	LOW	HIGH
Modesto	CA	LOW	HIGH
McAllen-Edinburgh-Mission	TX	LOW	HIGH
Jacksonville	NC	LOW	LOW
New Bedford	MA	HIGH	LOW
Houma	LA	LOW	LOW
Alexandria	LA	LOW	LOW
Fort Smith	AR-OK	LOW	HIGH
Anniston	AL	LOW	LOW
Gadsden	AL	LOW	LOW

Minneapolis-St. Paul	MN-WI	HIGH	HIGH	Pine Bluff	AR-OK	LOW	LOW
Dover	DE	LOW	HIGH	Lawrence	MA-NH	HIGH	LOW
Washington	DC-MD-VA-WV	HIGH	HIGH	Kankakee	IL	LOW	LOW
Fayetteville-Springdale-Rogers	AR	LOW	HIGH	Merced	CA	LOW	HIGH
Pittsburgh	PA	HIGH	LOW	Newburgh	NY-PA	LOW	HIGH
Bloomington	IN	LOW	LOW	Stockton-Lodi	CA	LOW	HIGH
Stamford-Norwalk	CT	HIGH	LOW	Laredo	TX	LOW	HIGH

1. For each group of metropolitan areas, find the number of areas that have a low and high population density and place your frequencies in the table below.

	TOP AREAS			BOTTOM AREAS	
Population density	Frequency	Proportion		Frequency	Proportion
LOW					
HIGH					

2. Construct side-by-side bar charts of the proportions of low and high density areas in the "top" and "bottom" groups.

3. By computing proportions, investigate if the population density differs between the "top" areas and the "bottom" areas.

4. For each group of areas, find the proportion that have a low and high population growth rate. Construct two segmented bar charts to compare the proportions of low and high growth in the two groups.

5. Are the population growth rates different for the "top" and "bottom" metropolitan areas? (Compare proportions to answer the question.)

COMPARING BATCHES OF QUANTITATIVE DATA

Next we discuss how we compare two or more batches of quantitative data. First, let's talk about how we compare things. Suppose someone is interested in comparing your height with your mom's height. Now this person can say "you are taller than your mom," but usually he or she is interested in a more informative comparison like

"you are three inches taller than your mom."

This is a typical kind of comparison -- we say that one measurement is so much greater (or smaller) than another measurement.

Other types of comparisons are in terms of ratios. For example, suppose you wish to compare your income this year with your income last year. You may say that this year's income is, say, $2000 more than last year's income. But it is more common to compare incomes in terms of ratios. For example, pay raises are usually expressed in terms of percent, so you may say that this year's income is 4% higher than last year's income.

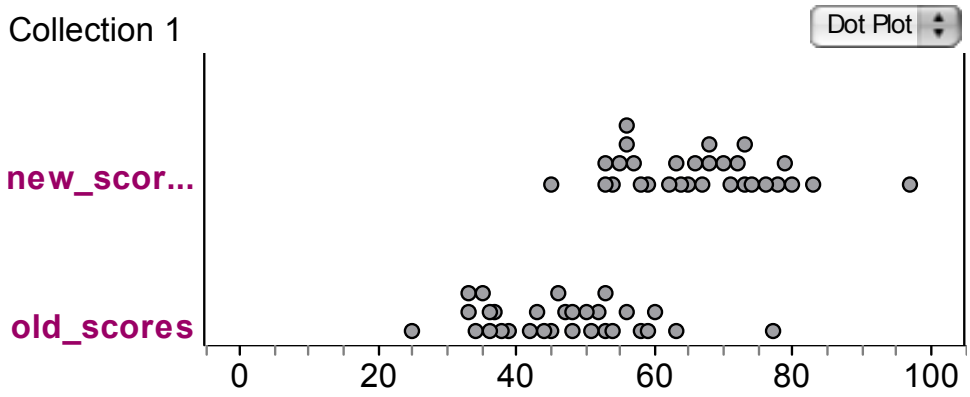
When we compare two batches of data, the easiest type of comparison is to say, for example, that

"one batch tends to be 10 more than another batch."

What does it mean to say that one batch of quantitative data is larger than a second batch?

Suppose you are given a relatively difficult exam in a particular class. The class does poorly – the median grade (out of 100 possible points) is only 46.5 and the quartiles are 37 and 53. The instructor decides to curve the exam grades by adding 20 points to each student's grade. What is the effect of this adjustment on the batch of test scores?

The figure below shows dotplots of the old test scores and the new test scores on the same scale.



How do the two batches of scores differ? Note that both sets of test scores have the same shape and same spread, but different locations. Note that one can get the distribution of new scores by moving the old distribution of scores 20 points to the right.

This example illustrates the situation when we are able to compare two batches. If two batches have approximately the same shape and same spread, then saying

"batch 1 is 10 points larger than batch 2"

means that we can obtain batch 1 by adding 10 points to each value in batch 2. But this statement assumes that the batches have similar spreads. In practice, we have to check this assumption before we can make any comparison.

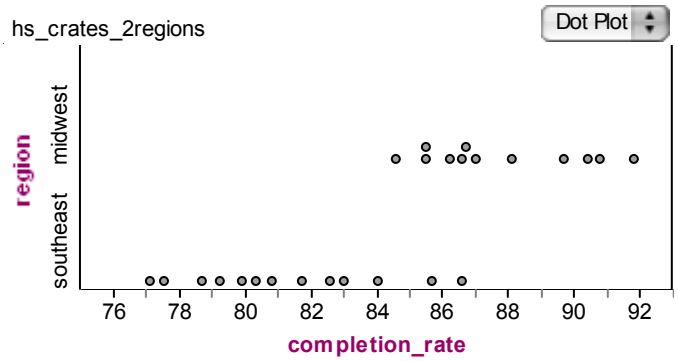
In Section D1, we looked at the high school completion rates for all states. We saw much variability in these completion rates. Some of the smallest completion rates corresponded to the Southeast states and the largest rates for states in the Midwest. That raises the question: Do Southeast states generally have lower high school completion rates than Midwest states?

To start to answer this question, we collect the high school completion rates for the two groups of states.

Midwest States		Southeast States	
Illinois	85.5	Alabama	77.5
Indiana	84.6	Arkansas	81.7
Iowa	89.7	Florida	84
Kansas	88.1	Georgia	82.6
Michigan	86.2	Kentucky	78.7
Minnesota	90.8	Louisiana	80.8
Missouri	86.6	Maryland	85.7
Nebraska	90.4	Mississippi	80.3
North Dakota	85.5	North Carolina	79.2

Ohio	87	South Carolina	83
South Dakota	91.8	Tennessee	79.9
Wisconsin	86.7	Virginia	86.6
		West Virginia	77.1

A good way of graphically comparing the two groups of completion rates is by *parallel dotplots*. On the below figure, we construct a dotplot of the Midwest states values on the Midwest line and a dotplot of the Southeast states values on the Southeast line.



A different graph of the rates for the two groups is ***side-by-side stemplots***, where one uses a common list of stems (as we have done below) and place the leaves for the Midwest states on the right and the leaves for the Southeast states on the left.

SOUTHEAST STATES		MIDWEST STATES
51	77	
7	78	
92	79	
83	80	
7	81	
6	82	
0	83	
0	84	6
7	85	55
6	86	267
	87	0
	88	1
	89	7
	90	48
	91	8

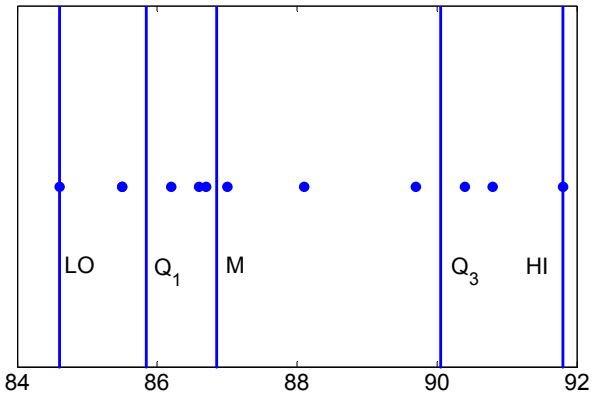
77|1 means 77.1 84|6 means 84.6

To compare two batches, it is useful to first summarize each batch with a five-number summary and then compare the summaries of the two batches. You can confirm that the five-number summary of the high school completion rates for the Midwest states, and the five-number summary of the rates for the Southeast states are given by

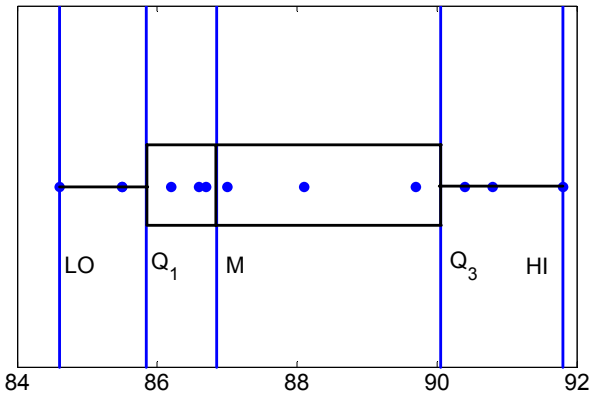
MIDWEST: (LO, Q_L, M, Q_U, HI) = (84.6, 85.85, 86.85, 90.05, 91.8)

SOUTHEAST: (LO, Q_L, M, Q_U, HI) = (77.1, 79.2, 80.8, 83, 86.6).

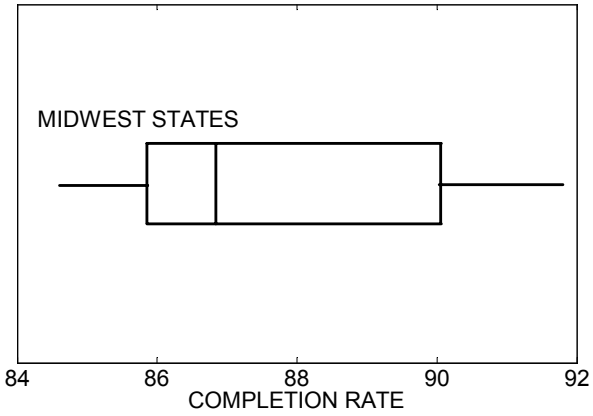
A ***boxplot*** is a graph of a five-number summary. To draw this for the Midwest states rates, we first locate the five numbers (LO, Q_L, M, Q_U, HI), on a number line below.



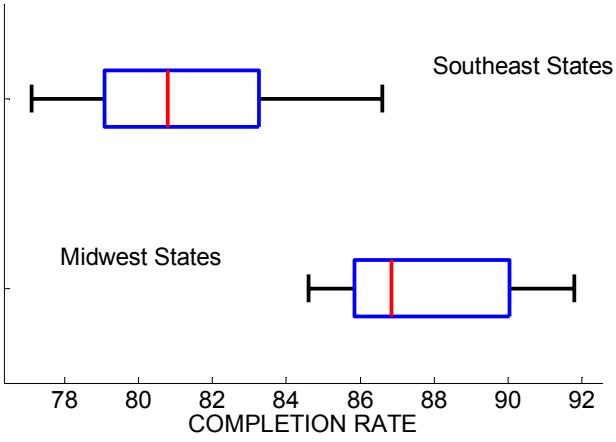
Next we draw a box extending from the lower to upper quartiles with the location of the median represented by a line inside the box. We complete the boxplot by drawing lines (sometimes called whiskers) from the outsides of the box to the locations of the LO and HI observations, respectively.



Removing the labels and the guidelines, we get the final boxplot display:



In similar fashion, we can construct a boxplot of the Southeast states completion rates. If we place both boxplots on the same graph on the same scale, we get the following display:



We can compare the two groups if the spread of the southeast completion rates is about the same as the spread of the Midwest rates. To check this, we place the quartiles and the interquartile spread of each group in the below table. We note that the two IQR's are 4.2 and 3.8, so the two batches have approximately the same spread.

Group	Q_L	Q_U	IQR
Midwest States	85.85	90.05	4.2
Southeast States	79.2	83	3.8

When two batches have equal spreads (as they do here), one can compare the two groups by finding the difference in medians. We note that the median completion rate of the Midwest states is 86.85 compared to a median rate of 80.8 for the Southeast states. This means that the high school completion rates for the Midwest states tend to be 86.85 – 80.8 = 6.05 higher than the completion rates for the Southeast states.

To emphasize what "one batch tends to be 6 units larger than a second batch" means, look at the boxplot that represents the high school completion rates for the Southeast states. Suppose we add 6 points to each of the rates for the Southeast states. When we do this each of LO, Q_L, M, Q_U, HI will increase by 6 points and the five-number summary will change from

(77.1, 79.2, 80.8, 83, 86.6) to (83.1, 85.2, 86.8, 89, 92.6).

This new five-number summary and the corresponding boxplot represent the completion rates for the Midwest states.

SPECIAL NOTE: How can we say that two batches have approximate equal spreads? This is a difficult question to answer in general, but here are some guidelines to use in practice. (These guidelines assume that each batch is approximately mound-shaped.) We use the IQR to measure the spread of a batch and we compare the spreads of the two

batches by means of the ratio IQR(batch2)/IQR(batch1). If each batch has about 50 values, then if the ratio of IQRs is between .75 and 1.35, then we can say the batches have approximately equal spreads. If we have smaller batches, each of size 25, then we can conclude "equal spreads" if the ratio of IQRs is between .6 and 1.6.

PRACTICE: COMPARING BATCHES OF QUANTITATIVE DATA

The book *Cities Rated & Ranked* produces a ranking of the top 30 and bottom 30 metropolitan areas with respect to quality of life. One key variable that may distinguish the top and bottom areas is the unemployment rate. These rates (expressed as a percentage) are shown in the below table.

TOP METROPOLITAN AREAS					BOTTOM METROPOLITAN AREAS				
4.0	3.6	4.8	5.6	2.7	6.9	5.4	10.6	3.4	3.3
7.4	2.6	4.0	4.3	4.7	5.2	5.7	7.4	6.8	8.0
8.1	6.6	4.6	3.9	3.3	7.4	4.6	5.6	8.2	8.4
2.6	3.5	5.8	4.0	4.7	6.3	4.5	5.9	4.1	5.0
4.1	4.3	4.3	5.6	4.9	4.0	4.1	13.6	11.5	9.3
3.2	3.6	3.4	3.2	3.3	4.7	9.2	6.3	4.7	11.0

1. Construct back-to-back stemplots of the unemployment rates for the top and bottom areas using the stems shown below.

```
Bottom areas     | 2 |          Top areas
                 | 3 |
                 | 4 |
                 | 5 |
                 | 6 |
                 | 7 |
                 | 8 |
                 | 9 |
                 |10 |
                 |11 |
                 |12 |
                 |13 |

    3|2 means an unemployment rate of 3.2%
```

2. Find five-number summaries of the rates for each group of areas.

3. Is it reasonable to say that the two groups of unemployment rates have similar spreads? Explain.

4. Is it appropriate to compare the two groups of rates by computing the difference in the medians? Explain.

RELATIVE STANDING

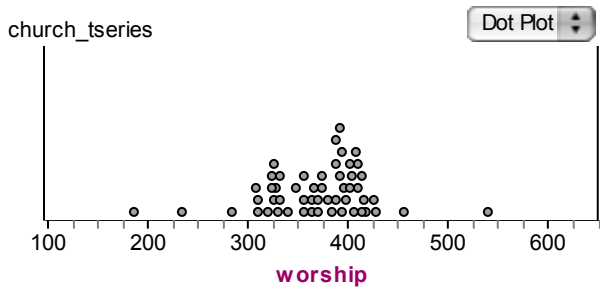
Suppose you are interested in learning about the Sunday worship attendance at a local church. You are told that the attendance for one week (a Sunday in July) was 327. You might wonder if this number represents a typical attendance for this church. You might suspect that this number is lower than average since you know that church attendance is usually smaller in the summer months due to people going on vacation. But how much smaller? Is there a way of talking about this attendance number (327) in the context of the distribution of attendance numbers for this church over many weeks?

To answer these questions, we collect the attendance numbers for this church for all of the weeks this year. This table displays the attendance numbers in chronological order. So, for example, the attendance numbers in the first five weeks in January were 234, 394, 417, 186 and 406.

MONTH	ATT	MONTH	ATT	MONTH	ATT	MONTH	ATT
January	234	April	339	July	309	October	428
January	394	April	369	July	284	October	365
January	417	April	539	July	307	October	414
January	186	April	414	July	327	October	394
January	406	April	388	July	348	October	402
February	397	May	395	August	373	November	410
February	310	May	374	August	332	November	404
February	425	May	384	August	355	November	387
February	319	May	325	August	324	November	392
March	364	June	329	September	324	December	401
March	370	June	355	September	388	December	455

March	364	June	392	September	415	December	408
March	356	June	331	September	409	December	379
						December	326

We graph the attendance numbers using a dotplot. We see a lot of variability in the numbers -- the attendances range from about 186 to 539 and a typical number in the middle of the batch is approximately 370. If the pastor of this church wonders about the summer worship attendance, the July attendance number of 327 looks a little lower than average.



One way of describing the location of 327 relative to the distribution of values is based on a *standardized score*. Using a computer, we find the mean $\overline{x}$ and standard deviation s of the weekly attendance numbers to be $\overline{x} = 368.6$, $s = 54.5$. The standardized score or z-score of a data value x is found by subtracting the mean and dividing by the standard deviation:

$$z = \frac{x - \overline{x}}{s}.$$

Here the standardized score of July's attendance 327 is

$$z = \frac{327 - 368.8}{54.5} = -0.77.$$

The standardized score tells you the number of standard deviations the data value is from the mean. A *positive* z-score indicates the value is *above* the mean, and a *negative* z-score means the value is *below* the mean. Here z = -0.77 which means that the particular

July attendance of 327 is approximately .8 or 4/5 of a standard deviation below the mean Sunday attendance. For a second example, note that the attendance for the first week of October was 428 has a z-score of

$$z = \frac{428 - 368.8}{54.5} = 1.09 .$$

Since this standardized score is a positive value close to one, we can say that this attendance of 428 is approximately one standard deviation above the mean.

FLAGGING POSSIBLE OUTLIERS

When we graphed the weekly attendance numbers, we saw one large cluster of values in the 300-450 range and we also saw a few unusually small and large values. Do these extreme attendance numbers deserve extra attention? Is there a method of identifying possible outliers in a dataset?

There is a useful "rule of thumb" for identifying data values that are unusually small or large. This method is based on the quartiles and the IQR that measures the spread of the middle half of the measurements. We first compute the five-number summary of the attendance numbers:

$$(\text{LO}, Q_L, \text{M}, Q_U, \text{HI}) = (186, 330, 374, 403, 539).$$

We say that an extreme observation is worthy of special attention if it falls further than one step from the lower and upper quartiles, where a step is defined to be

$$\text{STEP} = 1.5 \text{ IQR}.$$

We call such extreme observations *outliers*.

Let's illustrate this rule of thumb for our attendance data:

1. We compute the interquartile range $\text{IQR} = Q_U - Q_L = 403 - 330 = 73$.
2. A step is then equal to $\text{STEP} = 1.5 \text{ IQR} = 1.5 (73) = 109.5$.

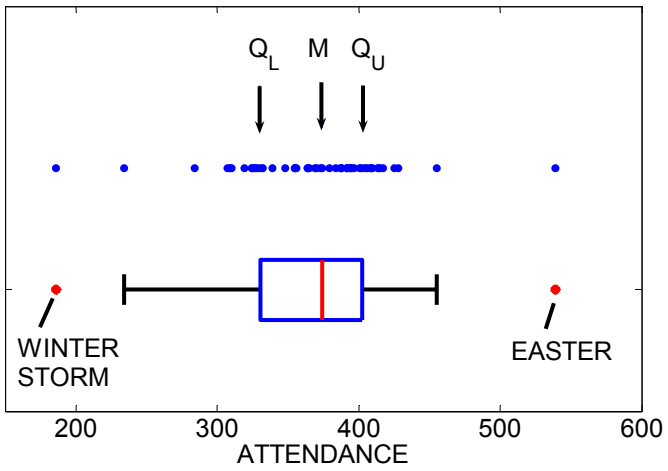
3. We say that an extreme observation deserves extra attention if either it

-- falls below Q_L - STEP = 330 – 109.5 = 220.5

or

-- it falls above Q_U + STEP = 403 + 109.5 = 512.5 .

Does this rule identify any unusually small or large attendance numbers? Looking back at our data, we see that the April attendance of 539 and the January attendance of 186 are outliers using our definition. It is common to draw a boxplot showing these outliers. The process of constructing this *modified boxplot* is displayed below. On the top of the figure, we show the observations with the locations of the median and quartiles indicated. We plot the "box" part of the boxplot as before. We then indicate the outliers by separate plotting points and draw lines (whiskers) from the box to the most extreme points at each end that are not outliers.



This rule of thumb identifies extreme data values that deserve extra attention. What are the possible explanations for unusually high or low attendance numbers? For the Christian faith, there are two major religious holidays, Christmas and Easter that will draw a large number of people for worship. Also, inclement weather may make it harder for people to attend church on particular days, causing small worship attendances. Indeed, in this example, it is not difficult to infer that the single large attendance number

corresponds to an Easter Sunday in April, and the small January attendance number was likely due to a winter storm that made it difficult to travel to church.

PRACTICE: RELATIVE STANDING AND FLAGGING OUTLIERS

Again consider the unemployment rates for the top 30 metropolitan areas as reported by *Cities Rated & Ranked* .

TOP METROPOLITAN AREAS				
4.0	3.6	4.8	5.6	2.7
7.4	2.6	4.0	4.3	4.7
8.1	6.6	4.6	3.9	3.3
2.6	3.5	5.8	4.0	4.7
4.1	4.3	4.3	5.6	4.9
3.2	3.6	3.4	3.2	3.3

1. The mean and standard deviation of the unemployment rates are given by $\bar{x} = 4.36$ % and s = 1.33 %. The unemployment rate for Bloomington, Indiana is 2.7 %. Find Bloomington's standardized score.

2. Interpret Bloomington's standardized score in terms of the number of standard deviations above or below the mean.

3. Find and interpret the standardized score for Portland, Oregon that has an unemployment rate of 8.1 %.

4. Construct a modified boxplot of these rates. Using the rule of thumb, identify any outliers among the unemployment rates among the top areas.

ACTIVITY: COMPARING MEN AND WOMEN IN THE CLASS DATASET

In Topic D1, we collected a number of different variables from all the students in the class. Choose two variables that you believe will be different between men and women in the class. (One obvious variable that would distinguish genders is height.)

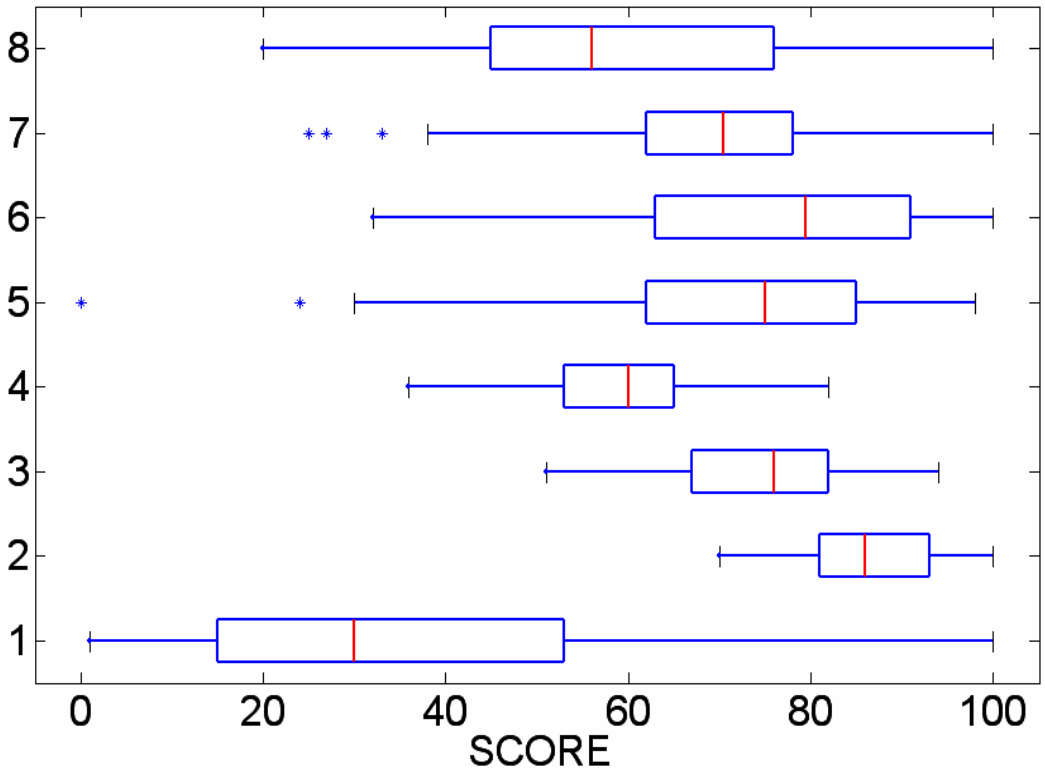
For each variable

1. Construct parallel dotplots or parallel stemplots comparing men and women.
2. Find five-number summaries of each group.
3. Construct parallel boxplots.
4. Find the interquartile spread for each group. Is it reasonable to say that the men and women data have approximately the same spreads?
5. If the answer is "yes" to the previous question, make a comparison between the men and women. (It is not sufficient to say that one group tends to be larger than the second group. You should indicate how much larger.)

ACTIVITY: MATCHING STATISTICS WITH BOXPLOTS

DESCRIPTION: In this activity, we will match up boxplots and their corresponding summary statistics. The relative locations of the mean and median are informative about the shape of the data that affects the relative lengths of the "whiskers" and the two components of the "box" part of the boxplot. In addition, the value of the standard deviation is useful in understanding the length of the boxplot.

Below are the boxplots for eight datasets and following are summary statistics (the mean, median, and standard deviation) for eight datasets.



In the "Boxplot" row of the below table, write down the letter of the matching boxplot.

Dataset	1	2	3	4	5	6	7	8
mean	59.44	60.58	75.00	69.02	86.68	77.12	36.48	71.14
median	60.00	56.00	76.00	70.50	86.00	79.50	30.00	75.00
st dev	9.46	20.79	10.58	16.64	8.42	17.47	26.46	20.74
Boxplot								

ACTIVITY: COUNTING PASTA

DESCRIPTION: Suppose you manage an Italian restaurant and pasta is one of the main items you serve. Employees use different techniques of measuring half-cup servings of pasta. Some workers use a cup measure and fill pasta to the half-cup line, and other workers use half-cup spoons. You notice that these half-cup servings vary in size. As a

manager, you need to find the way of measuring pasta (using a cup or spoon measure) that varies as little as possible so that you can plan and run your business better.

MATERIALS NEEDED: Several boxes of pasta shells. A set of clear one-cup measures and a set of plastic spoon measures.

1. Collecting the data

The class will be divided into an even number of teams consisting of at least two students each. The teams will then be separated into two groups.

Within the first group, each team pours shells into a cup measure and counts the number of shells.

(a) One person pours shells into the cup measure until he or she thinks it is full at the half-cup level.

(b) A second person then counts the number of shells and records it, without informing the first person the number.

(c) Each team repeats the experiment five times and computes the mean ($\bar{x}$) and standard deviation (s) for its measurements.

In the second group, students use a plastic spoon to measure a half cup of shells.

(a) One person pours shells into the spoon until he or she thinks that the spoon is full.

(b) A second person then counts the number of shells and records it, without informing the first person the number.

(c) Each team repeats the experiment five times and computes the mean and the standard deviation (s) for its measurements.

2. Making plots

(a) Construct stemplots of the mean counts of the teams for each group and combine these two plots in a back-to-back stem and leaf diagram.

(b) Construct stemplots for the standard deviations of the teams for each group and combine these in a back-to-back stem and leaf diagram.

3. Analyzing the results

(a) Based on looking at the stem and leaf diagrams for the mean counts for the two groups in Making plots (a), do you see any difference? Does one method of measurement tend to give larger number of counts of shells on average?

(b) Based on looking at the stem and leaf diagrams for the standard deviation of the counts for the two groups in Making plots (b), do you see any difference? Do both measurement processes vary by about the same amount?

Note: We often divide variability in processes into two types:

- Common cause variation is part of the system or process and affects everyone in the system.
- Special cause variation either is not part of the system or process all of the time or does not affect everyone in the system.

(c) Based on your comments in (a) and (b), which of the two measurement methods (cup or spoon) would you recommend and why?

WRAP-UP

In this topic, we were introduced to some basic methods for comparing two or more batches. When the variables are categorical, *side-by-side barcharts* or *segmented bar charts* are useful in comparing frequency distributions and groups can be compared by computing a difference in percentages. When comparing batches of quantitative data, a comparison like "one batch is 10 larger than a second batch" is possible when the two batches have approximately equal spreads. To summarize each batch we compute a *five-number summary*, and *parallel boxplots* are helpful in graphically comparing batches. To understand one's *relative standing* in a batch, it is useful to compute a *standardized score* that indicates the number of standard deviations one falls from the mean. A rule of thumb was described for flagging possible outliers for special attention and a *modified boxplot* displays these possible outliers with special plotting points.

EXERCISES

1. Bird Watching

The below table gives the frequencies of all types of owls spotted by bird watchers in the states of Ohio and Pennsylvania who participated in the 2006 survey of the Great Backyard Bird Count.

	Ohio	Pennsylvania
Type	Frequency	Frequency
Barn Owl	6	6
Eastern Screech Owl	13	30
Great Horned Owl	16	41
Snowy Owl	9	6
Barred Owl	27	13
Long-eared Owl	3	2
Short-eared Owl	14	4

a. For each state, find the percentage of each type of owl spotted.

b. Construct parallel bar charts for the percentages in the two states.

c. Find the mode for each state.

d. Find the types of owls where there is a significant difference in the percentages in the two states.

2. **Teacher's Salary and Proficiency.**

The average teacher's salary and the percentage of 8^{th} graders "above proficiency" on a specific mathematics exam were collected for all states in the *Report Card on American Education* by the American Legislature Exchange Council. The average teacher's salary for a state was classified as "LOW," "MEDIUM" and "HIGH" and the percentage above proficiency (PCT) was broken down into three groups. The table below gives the number of states with a given salary level and range of PCT; for example, we see that 7 states have low teacher salaries and PCT below 24%.

		Percentage (PCT) above proficiency		
		PCT < 24	24 <= PCT < 32	PCT >=32
Teacher's Salary	LOW	7	5	6
	MEDIUM	5	3	9
	HIGH	2	9	4

a. Do you think there is a relationship between teachers' salaries and performance of students on a standardized test? Explain.

b. For each group of teacher's salary, find the proportion of states with a percentage above proficiency in each of the three groups.

c. Construct segmented bar charts of the proportions computed in part b, where each bar corresponds to one group of teachers' salaries.

d. Based on your computations in parts b and c, do you think states with higher salaries tend to have better performances on the mathematics exam? Explain.

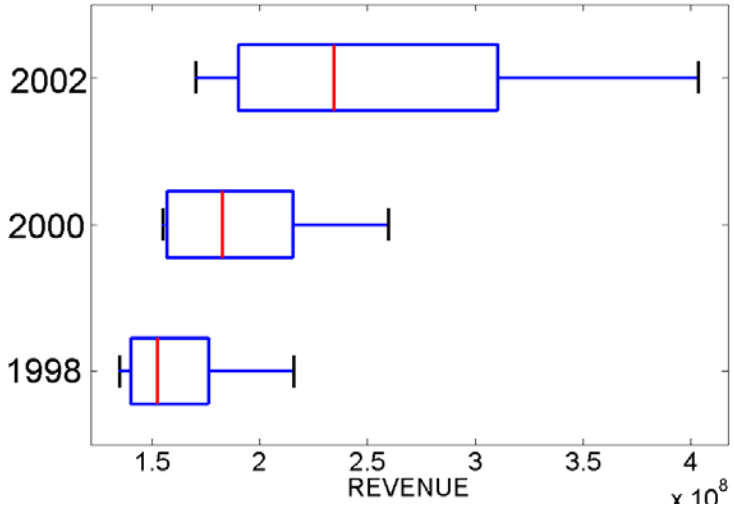
3. **Gross Sales for Top Movies**

The table below lists the top ten movies (in terms of gross revenue) for the years 1998, 2000, and 2002.

Movie	Revenue ($)	Movie	Revenue ($)
Spider-Man (2002)	403706375	Saving Private Ryan (1998)	216119491
The Lord of the Rings: The Two Towers (2002)	340478898	Armageddon (1998)	201573391
Star Wars: Episode II - Attack of the Clones (2002)	310675583	There's Something About Mary (1998)	176483808
Harry Potter and the Chamber of Secrets (2002)	261970615	A Bug's Life (1998)	162792677
My Big Fat Greek Wedding (2002)	241437427	The Waterboy (1998)	161487252
Signs (2002)	227965690	Doctor Dolittle (1998)	144156464
Austin Powers in Goldmember (2002)	213079163	Rush Hour (1998)	141153686
Men in Black II (2002)	190418803	Deep Impact (1998)	140459099
Ice Age (2002)	176387405	Godzilla (1998)	136023813
Chicago (2002)	170684505	Patch Adams (1998)	135014968

How the Grinch Stole Christmas (2000)	260031035
Cast Away (2000)	233630478
Mission: Impossible II (2000)	215397307
Gladiator (2000)	187670866
What Women Want (2000)	182805123
The Perfect Storm (2000)	182618434
Meet the Parents (2000)	166225040
X-Men (2000)	157299717
Scary Movie (2000)	156997084
What Lies Beneath (2000)	155370362

Parallel boxplots of the revenues for the three years are displayed below. (The unit for REVENUE on the graph is in hundreds of millions of dollars.)



a. From the graph, describe the distribution of each batch. This discussion should include statements about shape, typical values, spread, and any unusual characteristics.

b. Using the graph, complete the below table.

Group	Median	Q_L	Q_U	IQR $= Q_U - Q_L$
1998				
2000				
2002				

c. How does the spread of revenues change from 1998 to 2000 to 2002? (Look at the IQR column.)

d. How does the median revenue change from 1998 to 2000 to 2002 movies?

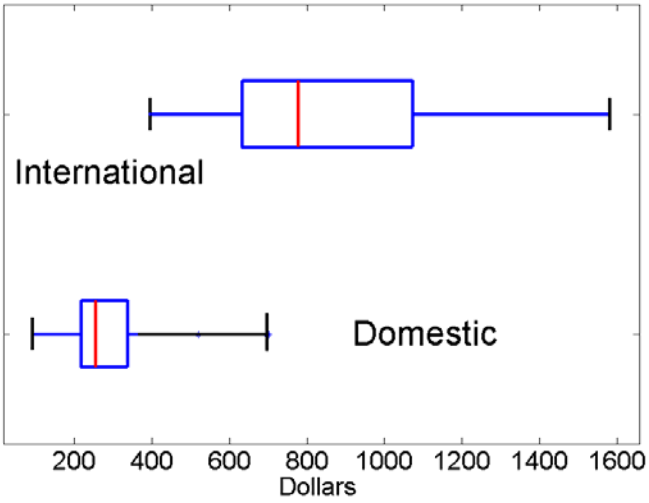
e. Can one say that revenues for one year, say 2002, are a particular dollar amount greater than the revenues for a second year such as 2000? Why or why not?

4. **Fares of National and International Flights**

The following table gives the airfares (in dollars) from Detroit to a number of U.S. and International cities. (The data were collected from orbitz.com in May 2004.)

DOMESTIC		INTERNATIONAL	
CITY	FARE	CITY	FARE
San Francisco	310	London	624
Chicago	92	Cape Town	1582
Miami	252	Beijing	1022
Denver	198	Paris	619
Las Vegas	242	Sydney	1200
Philadelphia	258	Amsterdam	793
Boston	327	Lima, Peru	644
New York	170	Mexico City	397
Fargo	369	Jerusalem	968
San Diego	312	Bangkok	1124
Portland	352	Tokyo	763
Raleigh	224	Milan	657
Kansas City	220		
Honolulu	701		
Fairbanks	696		
Orlando	236		
Sante Fe	521		
New Orleans	280		
Phoenix	204		
Houston	219		

Parallel boxplots of the domestic and international fares are shown in the figure below.



a. From the graph, find the median fare of the domestic flights and the median fare of the international fares.

b. By using a suitable measure from the graph, compare the spreads of the two batches of fares.

c. Would it be appropriate to compare the domestic and international fares by just comparing the medians? Why or why not?

5. **Salaries of Basketball and Baseball Players**

The table below gives the salaries (in millions of dollars) for the players on the 2003-2004 Los Angeles Lakers basketball team and the Los Angeles Dodgers baseball team.

Los Angeles Lakers		**Los Angeles Dodgers**	
PLAYER	SALARY	PLAYER	SALARY
Shaquille O'Neal	26.5	Shawn Green	16.7
Kobe Bryant	13.5	Darren Dreifort	11.4
Brian Cook	0.8	Hideo Nomo	9
Rick Fox	4.5	Todd Hundley	7
Derek Fisher	3.0	Jeff Weaver	6.2
Horace Grant	1.1	Adrian Beltre	5
Devean George	4.5	Eric Gagne	5
Karl Malone	1.5	Odalis Perez	5
Gary Payton	4.9	Paul Lo Duca	4.1
Bryon Russell	1.1	Paul Shuey	3.9
Kareem Rush	1.1	Juan Encarnacion	3.6

Ime Udoka	0.4	Kazuhisa Ishii	2.5
Luke Walton	0.4	Wilson Alvarez	1.5
Jamal Sampson	0.6	Guillermo Mota	1.5
Stanislav Medvedenko	1.5	Tom Martin	1.4
		Alex Cora	1.3
		Robin Ventura	1.2
		Dave Roberts	1.0
		Cesar Izturis	.4
		David Ross	.3
		Duaner Sanchez	.3
		Jayson Werth	.3
		Brian Falkenborg	.3
		Jason Grabowski	.3
		Wilkin Ruan	.3
		Joe Thurston	.3

a. Using the stems below, construct back-to-back stemplots of the Lakers and Dodgers salaries. (The break between the stem and leaf occurs at the decimal point. Since Shaquille O'Neal's salary is so large relative to the remaining salaries, you can place his salary on the "HI" line.)

Lakers Salaries		**Dodgers Salaries**
	0	
	1	
	2	
	3	
	4	
	5	
	6	
	7	
	8	
	9	
	10	
	11	
	12	
	13	
	14	
	15	
	16	
	HI	

b. Find five-number summaries of each dataset. Describe the distribution of each batch of salaries using the summary numbers.

c. Compare the two groups of salaries. Can you say that one group tends to get higher salaries than the other group?

6. **Ice Cream Calories of Two Manufacturers**

The following table gives the calories of a half-cup serving of different flavors made by Ben and Jerry's and Breyers.

Ben and Jerry's		Breyers	
Flavor	calories	Flavor	calories
Brownie Batter	310	carmel fudge	160
Butter Pecan	290	vanilla	140
Cherry Garcia	260	french vanilla	150
Chocolate	260	van/choc/straw	140
Chocolate Chip Cookie Dough	270	butter pecan	170
Chocolate Fudge Brownie	270	chocolate	150
Chubby Hubby	330	mint chocolate chi	160
Chunky Monkey	300	strawberry	120
Coffee	240	rocky road	150
Coffee HEATH Bar Crunch	290	cookies & cream	160
Dublin Mudslide	270	vanilla fudge twirl	140
Everything But The...	320	peach	130
Fudge Central	300	coffee	140
Half Baked	280	cherry vanilla	140
Karamel Sutra	280	chocolate chip	160
Mint Chocolate Cookie	270	chocolate chip cookie dough	170
New York Super Fudge Chunk	310	vanilla & choc fudge checks	170
Oatmeal Cookie Chunk	280	banana fudge chunk	170
One Sweet Whirled	280	vanilla fudge brownie	160
Peanut Butter Cup	380	cherry chocolate chip	150
Phish Food	280	peanut butter & fudge	170
Pistachio Pistachio	280	dulce de Leche	150
Primary Berry raham	270	lactose free vanilla	160
Strawberry	240	mocha almond fudge	170
Uncanny Cashew	290	butter almond	160
Vanilla	240	calcium rich vanilla	130
Vanilla HEATH Bar Crunch	300	carmel praline crunch	180
Vanilla Swiss Almond	280	fresa banana	140
		homemade vanilla	140
		extra creamy vanilla	150
		extra creamy chocolate	140
		take two	150
		take two (sherbet)	130

a. Construct a stemplot of the calorie numbers for the Ben and Jerry's flavors and construct a separate stemplot of the calorie numbers of the Breyers flavors.

b. Find a five-number summary of each batch of calorie numbers.

c. Assuming that the spreads of the two batches of calorie numbers are approximately equal, compare the two batches. Which brand of ice cream tends to have more calories and by how much (on average)?

7. **Basketball Field Goal Percentages**

If a basketball team wins many games, then one would expect that the team would have a high proportion of field goal attempts that are successful. In contrast, teams that lose games tend to have a low proportion of successful field goal attempts. But that raises the interesting question: how much better are good teams than bad teams in making field goals? To answer this question, the table below gives, for the 2003-2004 basketball season, the team field goal proportions (FG) for the top 20 ranked college basketball teams and the team field goal proportions for the bottom 20 ranked teams.

BOTTOM TEAMS		TOP TEAMS	
	FG		FG
Cleveland State	.408	Duke	.472
Harvard	.394	Kentucky	.470
Norfolk State	.392	St. Joseph's	.478
Albany NY	.401	Mississippi State	.468
Eastern Illinois	.444	Connecticut	.481
VMI	.403	Pittsburgh	.482
Western Illinois	.429	Oklahoma State	.517
Florida International	.404	Stanford	.485
Campbell	.394	Texas	.445
Navy	.394	Cincinnati	.457
Charleston Southern	.423	Syracuse	.465
Md. Eastern Shore	.382	Florida	.483
Bethune-Cookman	.391	Georgia Tech	.470
Army	.362	Gonzaga	.521
Howard	.399	Wisconsin	.446
Loyola-Maryland	.378	North Carolina State	.452
North Carolina A&T	.361	Kansas	.462
Nicholls State	.394	North Carolina	.461
Arkansas-Pine Bluff	.344	Maryland	.443
Dartmouth	.407	Providence	.453

a. For the Top Teams, the mean and standard deviation of the shooting proportions are $\bar{x}$ =.471 and s = .021, respectively. Find the standardized scores of North Carolina and Gonzaga and explain what these scores tell you about the relative standing of these two schools with respect to field goal shooting.

b. Using the rule of thumb, determine if there are any outliers in field goal proportion among the Top Teams, and also determine if there are any outliers in shooting among the Bottom Teams.

c. For each batch of shooting proportions, construct a suitable graph and summarize the batch by computing a five-number summary. Write a short paragraph about the distribution of proportions for each batch. Construct parallel boxplots of the two batches. From your work, compare the shooting proportions of the Bottom and Top teams.

8. **Gas Prices**

During the spring of 2004, gasoline prices rose sharply. Consumers were interested in the variation of gas prices across states and the cause of this variation. The table below gives the average gas price for Eastern States and Western States, where the dividing line for west/east was the Mississippi River.

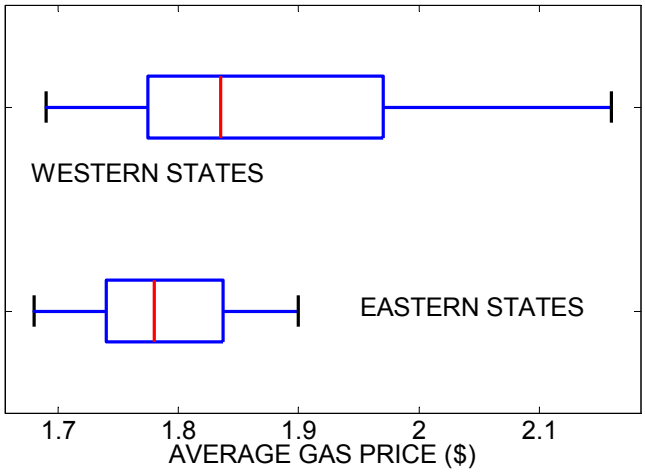
EASTERN STATES		WESTERN STATES	
State	gas_price	state	gas_price
Alabama	1.73	Alaska	1.92
Connecticut	1.85	Arkansas	1.74
District of Columbia	1.85	Arizona	1.98
Delaware	1.76	California	2.16
Florida	1.82	Colorado	1.85
Georgia	1.7	Hawaii	2.16
Illinois	1.89	Iowa	1.77
Indiana	1.83	Idaho	1.96
Kentucky	1.77	Kansas	1.8
Massachusetts	1.78	Louisiana	1.73
Maryland	1.78	Minnesota	1.8
Maine	1.79	Missouri	1.72
Michigan	1.85	Montana	1.89
Mississippi	1.74	North Dakota	1.85
North Carolina	1.74	Nebraska	1.82
New Hampshire	1.73	New Mexico	1.78
New Jersey	1.69	Nevada	2.11
New York	1.9	Oklahoma	1.69
Ohio	1.83	Oregon	2.06
Pennsylvania	1.79	South Dakota	1.81
Rhode Island	1.82	Texas	1.7
South Carolina	1.68	Utah	1.94
Tennessee	1.74	Washington	2.03
Virginia	1.71	Wyoming	1.79
Vermont	1.76		

Wisconsin	1.89
West Virginia	1.84

a. The mean and standard deviation of the gas prices for the eastern states are $\bar{x}$ = \$1.79 and s = \$0.063, respectively. Find the standardized scores for the gas prices of South Carolina and New York and give an interpretation of these scores.

b. Using the rule of thumb, determine if there are any outliers among the gas prices for the western states.

Parallel boxplots of the gas prices from the western and eastern states are displayed in the figure below.



c. Compare the two batches of gas prices with respect to "average" and spread.

d. On the average, how much more expensive is gas from a western state than from an eastern state?

e. Would it be accurate to say that *all* western states have more expensive gas than *all* eastern states? If this is not true, find an eastern state that has more expensive gas than a western state.

9. **Cost of Grocery Shopping**

How much money does a consumer spend on a single trip at the grocery store? How has the single-trip cost at a grocery store changed from 2001 to 2003?

Below is a stemplot of the grocery costs (in dollars) of 50 visits in 2001 and 34 visits in 2003 made by the author. (The smallest value in the first dataset corresponds to a purchase of $10.)

Make a comparison of the two batches by (a) finding 5-number summaries of each batch, and (b) constructing parallel boxplots. Compare the two datasets with respect to averages (medians) and spreads (quartile spreads). Make a comparison by comparing medians.

PURCHASE AMOUNTS IN 2001		**PURCHASE AMOUNTS IN 2003**
0	1	3
7	1	6778
3322	2	12
98	2	79
43332	3	13
998777665	3	559
43100	4	1
998	4	5799
4442100	5	12
76	5	5557
300	6	
77655	6	56788
0	7	1
7	7	8
	8	12
	8	
1	9	

1|0 means $10 1|3 means $13

10. **Car Mileages**

The table below gives the mileage (miles per gallon) for a selection of 2004 model cars.

CAR	TYPE	MPG	CAR	TYPE	MPG
Chrysler PT Cruiser	Sedan	18	Acura MDX	SUV	17
Ford Focus	Sedan	24	Buick Rendezvous	SUV	16
Honda Civic	Sedan	36	Ford Escape	SUV	17
Hyundai Accent	Sedan	26	Honda Element	SUV	20
Mitsubishi Lancer	Sedan	20	Hyundai Santa Fe	SUV	18
Pontiac Vibe	Sedan	26	Mazda Tribute	SUV	18
Saturn Ion	Sedan	24	Nissan Murano	SUV	19
Subaru Impreza	Sedan	20	Saturn Vue	SUV	18
Toyota Corolla	Sedan	29	Subaru Forester	SUV	21
Toyota Matrix	Sedan	24	Toyota Rav4	SUV	22
Volkswagen Golf	Sedan	41	Volkswagen Toureg	SUV	15
Volkswagen New Beetle	Sedan	25	Volvo XC90	SUV	18

a. Looking at the entire collection of cars, use the rule of thumb to find any possible outliers in the mileage values.

b. Draw a modified boxplot of the mileages.

c. By the use of graphs and appropriate summary statistics, compare the mileages of the Sedans with the mileages of the SUVs. Can you explain why there are substantial differences in the mileages between the two groups of cars?

11. **State Population Changes**

The table below gives the percentage change in population from 1990 to 2000 for all states in the United States. The first column indicates if the state is East or West of the Mississippi River.

	State	%change		State	%change		State	%change		State	%change
East	AL	10	East	IL	8.7	West	MT	12.9	East	RI	4.5
West	AK	14	East	IN	9.7	West	NE	8.4	East	SC	15.1
West	AZ	40	West	IA	5.4	West	NV	66.4	West	SD	8.3
West	AR	13.7	West	KS	8.5	East	NH	11.4	East	TN	16.6
West	CA	13.8	East	KY	9.7	East	NJ	8.8	West	TX	22.8
West	CO	30.6	East	LA	5.9	West	NM	20.1	West	UT	29.7
East	CT	3.6	East	ME	3.8	East	NY	5.5	East	VT	8.2
East	DE	17.6	East	MD	10.8	East	NC	21.4	East	VA	14.4
East	DC	-5.6	East	MA	5.5	West	ND	0.6	West	WA	21.1
East	FL	23.5	East	MI	6.9	East	OH	4.7	East	WV	0.8
East	GA	26.4	West	MN	12.4	West	OK	9.7	East	WI	9.7
West	HI	9.3	East	MS	10.5	West	OR	20.4	West	WY	8.8
West	ID	28.5	West	MO	9.3	East	PA	3.4			

a. Explain (before looking at the data) how you think the United States has grown in recent years. Are there particular areas of the country that have experienced high growth? Is your state one of the high growth states?

b. Suppose we define a population change as being HIGH if the percentage change exceeds 10 % and LOW if the percentage change is 10 % or lower. Compare the proportion of LOW and HIGH population changes for the Eastern and Western states using a graph and appropriate summary statistic.

c. Using parallel boxplots, compare the percentage changes for the Eastern and Western states.

12. **City Temperatures**

The table below gives the average temperature for eight cities for each month of the year.

Month	San Francisco	Vero Beach	Duluth	Albuquerque	San Diego	Philadelphia	Honolulu	Indianapolis
Jan	48.7	61.6	7.0	34.2	57.4	30.4	72.9	25.5
Feb	52.2	62.7	12.3	40.0	58.6	33.0	73.0	29.6
Mar	53.3	67.2	24.4	46.9	59.6	42.4	74.4	41.4
Apr	55.6	71.3	38.6	55.2	62.0	52.4	75.8	52.4
May	58.1	75.8	50.8	64.2	64.1	62.9	77.5	62.8
Jun	61.5	79.5	59.8	74.2	66.8	71.8	79.4	71.9
July	62.7	81.1	66.1	78.5	71.0	76.9	80.5	75.4
Aug	63.7	81.3	63.7	75.9	72.6	75.5	81.4	73.2
Sep	64.5	80.1	54.2	68.6	71.4	68.2	81.0	66.6

Oct	61.0	75.5	43.7	57.0	67.7	56.4	79.6	54.7
Nov	54.8	69.3	28.4	44.3	62.0	46.4	77.2	43.0
Dec	49.4	63.7	12.8	35.3	57.4	35.8	74.1	30.9

a. Suppose we say that a month is "HOT" if the average temperature exceeds 70; otherwise the month is "COLD." For each city, find the proportion of HOT and COLD months for each city. Graph these proportions and discuss any differences you see between cities.

b. For four cities of your choice, compare the monthly temperatures. Construct a graph that is helpful in comparing cities and find appropriate summary statistics for comparing cities with respect to monthly temperature.

13. **Church Attendance**

This table gives the worship attendance at a church for all weeks during the four seasons of the year. For example, the attendances during the first two weeks of the Winter season were 439 and 349.

WINTER	SPRING	SUMMER	FALL
439	426	375	429
349	418	342	470
388	535	416	438
421	352	342	407
363	522	332	395
362	398	372	355
406	427	293	402
343	384	343	455
479	409	322	422
399	472	357	425
381	377	299	426
289	344	348	399

a. Suppose we say that attendance is "HIGH" for a particular week if it is 350 or higher, otherwise it is "LOW." Find the proportion of HIGH and LOW attendance numbers for each season of the year.

b. Construct a suitable graph comparing the proportions you computed in part a. What differences do you find between seasons?

c. Compare the four seasons of attendance numbers by use of five-number summaries and parallel boxplots. Do you reach the same conclusions as you found in part b?

14. **Men and Women Professional Golfers**

The datasets pgastats.txt and lpgastats.txt contain statistics for the top 30 men and top 30 women professional golfers for the 2002 season. Some of the variables included on these datasets include.

DRIVING_AVG – the average (mean) length of a drive

DRIVING_ACC – the percentage of drives that land in the fairway

GREEN_PCT – the percentage of greens that are hit in regulation

PUTTS – the average number of putts for a 18-hour round

For two of these variables, compare the men and women professional golfers using the methods described in this topic.