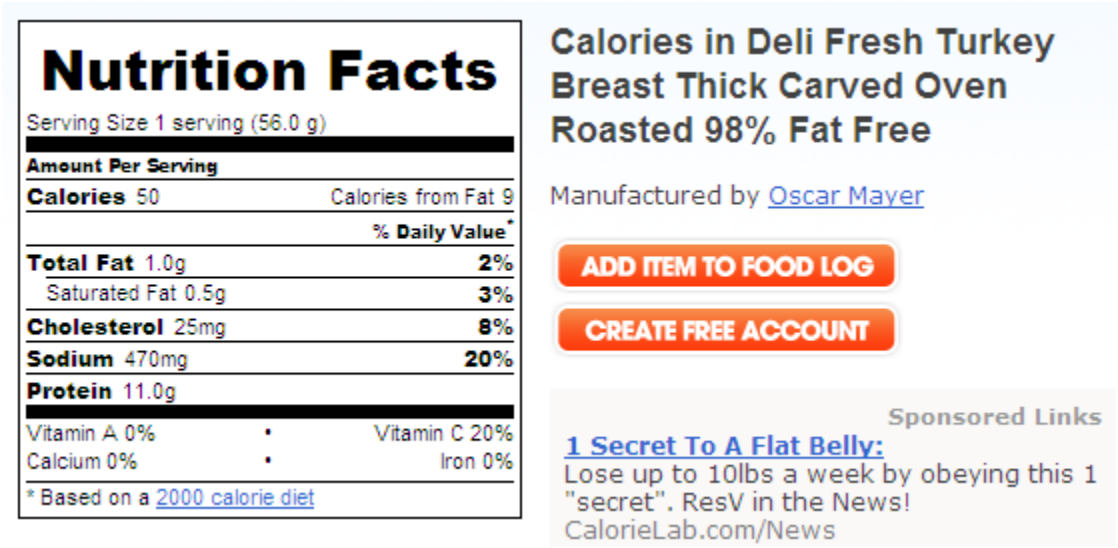
1 – Uni-Variate Data Analysis

Mark Twain – "If you don't read the newspaper you are uninformed, if you do read the newspaper you are misinformed". http://en.wikiquote.org/wiki/Talk:Mark_Twain, 12/27/11

Problem set 1-1

1. How can Oscar Mayer claim that the turkey is 98% fat free when almost 20% of the calories come from fat?



Nutrition Facts

Serving Size 1 serving (56.0 g)

Amount Per Serving	
Calories 50	Calories from Fat 9
	% Daily Value*
Total Fat 1.0g	**2%**
Saturated Fat 0.5g	**3%**
Cholesterol 25mg	**8%**
Sodium 470mg	**20%**
Protein 11.0g	

Vitamin A 0% • Vitamin C 20%
Calcium 0% • Iron 0%

* Based on a 2000 calorie diet

Source: CalorieCount, 98% Fat Free Turkey, http://caloriecount.about.com/calories-oscar-mayer-deli-fresh-turkey-i133863, 5/31/09

One serving has 1 gram of fat and one serving is 56 grams. $1/56 \approx 0.02$ or 2%

2. The article, “El Nino Seen As Trigger For Violence In The Tropics”, reports that from 1950 to 2004 civil conflict was more likely to occur in tropical countries during El Niño years. (El Niño, the opposite of La Niña, results in most tropical countries experiencing hotter, drier weather). There was a 3% chance of conflict during La Niña and that rose to 6% during El Niño.

   This change can be exaggerated or diminished depending on the way you compare 3% and 6%. Use subtraction, division, and percent change to create three comparative statistics.

   Source: NPR, El Nino Seen As Trigger For Violence In The Tropics, http://www.npr.org/2011/08/24/139914440/el-nino-seen-as-trigger-for-violence-in-the-tropics, 8/24/11

   Subtraction: 6% - 3% = 3%. Conflict rose by 3%.
   Division: 6% / 3% = 2. Conflict doubled.
   Percent change: (6% - 3%) / 3% = 1 = 100%. Conflict rose by 100%.

   For 3-6, refer to the table at the bottom of the page.

3. How many crimes were perpetrated in the US in 1995? **13,867,000 crimes**
4. What percentage of the total crimes were violent crimes?

   1,799 / 13,867 ≈ 13%

5. What percentage of the Property Crime in Metropolitan Areas was Motor Vehicle Theft? **1,381 / 10,426 ≈ 13%**

6. Write a complete sentence that fully describes what the number 234, found in the Violent Crime row, means. Write the sentence in the style of a newspaper article. **According to the FBI, 234 violent crimes were committed per 100,000 people in rural areas in the US in 1995.**

No. 314. Crimes and Crime Rates, by Type and Area: 1995

[**In thousands, except rate.** Rate per 100,000 population; see headnote, table 313. Estimated totals based on reports from city and rural law enforcement agencies representing 96 percent of the national population. For definitions of crimes, see text, section 5]

TYPE OF CRIME	UNITED STATES		METROPOLITAN AREAS [1]		OTHER CITIES		RURAL AREAS	
	Total	Rate	Total	Rate	Total	Rate	Total	Rate
Total	**13,867**	**5,278**	**12,045**	**5,761**	**1,158**	**5,315**	**664**	**2,083**
Violent crime	1,799	685	1,619	774	105	484	75	234
Murder and nonnegligent manslaughter	22	8	19	9	1	5	2	5
Forcible rape	97	37	81	39	8	38	8	25
Robbery	581	221	560	268	16	72	5	17
Aggravated assault	1,099	418	959	459	80	369	60	187
Property crime	12,068	4,593	10,426	4,986	1,053	4,833	590	1,850
Burglary	2,595	988	2,192	1,048	201	924	202	634
Larceny-theft	8,001	3,045	6,853	3,278	799	3,669	348	1,091
Motor vehicle theft	1,473	561	1,381	660	52	240	40	125

[1] For definition, see Appendix II.

Source: U.S. Federal Bureau of Investigation, *Crime in the United States*, annual.

For 7-8, refer to the table at the bottom of the page.

7. How many families received child support that had incomes of $15,000 and over?
   4,421,000 – 1,054,000 = 3,367,000 families

8. Consider the unemployment compensation of white families.
   a. What percentage of white families receiving specified sources of income received unemployment compensation?
      4, 336 / 58,872 ≈ 7%

   b. What percentage of families who received unemployment compensation were white?
      4, 336 / 5,022 ≈ 86%

No. 581. Number of Families Receiving Specified Sources of Income, by Characteristic of Householder and Family Income: 1995

[In thousands. Families as of March 1996. Based on Current Population Survey; see text, sections 1 and 14, and Appendix III]

SOURCE OF INCOME	Total families [1]	Under 65 years old	65 years old and over	White	Black	Hispanic origin [2]	Under $15,000	$15,000 to $24,999	$25,000 to $34,999
Total	**69,597**	**58,292**	**11,306**	**58,872**	**8,055**	**6,287**	**9,723**	**10,040**	**9,828**
Earnings	59,055	54,301	4,753	50,186	6,555	5,406	5,358	7,367	8,279
Wages and salary	57,324	52,965	4,359	48,589	6,480	5,276	4,991	7,050	7,937
Social Security, railroad retirement	16,356	5,862	10,494	14,370	1,592	915	2,716	3,885	3,116
Supplemental Security Income (SSI)	2,421	1,921	500	1,592	669	360	1,026	591	323
Public assistance	3,616	3,530	86	2,153	1,262	767	2,493	594	251
Veterans payments	1,735	1,054	681	1,507	172	55	163	247	278
Unemployment compensation	5,022	4,807	215	4,336	514	503	521	732	850
Workers compensation	1,571	1,458	114	1,337	165	137	122	210	265
Retirement income	10,001	4,208	5,792	9,106	697	339	473	1,786	2,019
Private pensions	6,328	2,259	4,069	5,810	410	211	337	1,314	1,425
Military retirement	956	673	283	851	79	29	8	83	131
Federal employee pensions	1,182	445	737	1,030	125	27	47	150	226
State or local employee pensions	1,911	785	1,126	1,746	127	57	70	260	361
Other income	10,322	9,895	427	8,496	1,393	776	1,761	1,509	1,472
Alimony	248	237	11	210	33	12	38	31	46
Child support	4,421	4,378	43	3,645	664	307	1,054	801	774
Education assistance	4,784	4,648	137	3,895	642	371	603	596	591

[1] Includes other items not shown separately. [2] Persons of Hispanic origin may be of any race.

Source: U.S. Bureau of the Census, "Current Population Survey, Annual Demographic Survey, March Supplement"; published 18 November 1996; <http://ferret.bls.census.gov/macro/031996/faminc/09000.htm>.

For 9-10, refer to the table at the bottom of the page.

9. How many pounds of Waste were generated in the US in 1995?
   208.1 million tons * 2,000 pounds/ton = 416,200 million pounds =

   416,200,000,000 pounds = 416.2 billion pounds

10. In 1995, were “Other Nonferrous Metals” a problem as far as recovery is concerned? For a. and b. below, write as if the sentences will appear in a newspaper article.
    a. Write a sentence that convinces that Other Nonferrous Metals are recovered the most of all materials. **In the US in1995, 69.5% of “other non-ferrous materials” generated were recovered, more than any other waste category.**

    b. Write a sentence that convinces that Other Nonferrous Metals are recovered the least of all materials. **In the US in1995, 0.9 million tons of “other non-ferrous materials” generated were recovered, less than any other waste category.**

No. 385. Generation and Recovery of Selected Materials in Municipal Solid Waste: 1970 to 1995

[In millions of tons, except as indicated. Covers post-consumer residential and commercial solid wastes which comprise the major portion of typical municipal collections. Excludes mining, agricultural and industrial processing, demolition and construction wastes, sewage sludge, and junked autos and obsolete equipment wastes. Based on material-flows estimating procedure and wet weight as generated]

ITEM AND MATERIAL	1970	1980	1985	1990	1991	1992	1993	1994	1995
Waste generated, total	**121.9**	**151.5**	**164.4**	**197.3**	**196.9**	**202.2**	**205.4**	**209.6**	**208.1**
Paper and paperboard	44.2	54.7	61.5	72.7	71.0	74.3	77.4	80.8	81.5
Ferrous metals	12.6	11.6	10.9	12.6	12.7	12.1	11.9	11.8	11.6
Aluminum	0.8	1.8	2.3	2.8	2.8	2.9	2.9	3.0	3.0
Other nonferrous metals	0.7	1.1	1.0	1.1	1.1	1.1	1.1	1.4	1.3
Glass	12.7	15.0	13.2	13.1	12.6	13.1	13.6	13.4	12.8
Plastics	3.1	7.9	11.6	17.1	17.7	18.4	19.0	19.3	19.0
Yard waste	23.2	27.5	30.0	35.0	35.0	35.0	33.3	31.5	29.8
Other wastes	24.6	31.9	33.9	42.8	44.0	45.3	46.2	48.5	49.1
Materials recovered, total	**8.6**	**14.5**	**16.4**	**33.9**	**37.7**	**41.4**	**44.8**	**52.0**	**56.2**
Paper and paperboard	7.4	11.9	13.1	20.2	22.5	24.5	25.5	29.5	32.6
Ferrous metals	0.1	0.4	0.4	2.6	3.1	3.4	3.9	4.1	4.2
Aluminum	-	0.3	0.6	1.0	1.0	1.1	1.1	1.2	1.0
Other nonferrous metals	0.3	0.5	0.5	0.7	0.7	0.7	0.7	1.0	0.9
Glass	0.2	0.8	1.0	2.6	2.6	2.9	3.0	3.1	3.1
Plastics	-	-	0.1	0.4	0.5	0.6	0.7	0.9	1.0
Yard waste	-	-	-	4.2	4.8	5.4	6.9	8.0	9.0
Other wastes	0.6	0.6	0.7	2.1	2.6	2.9	3.1	4.2	4.3
Percent of generation recovered, total	**7.1**	**9.6**	**10.0**	**17.2**	**19.1**	**20.5**	**21.8**	**24.8**	**27.0**
Paper and paperboard	16.7	21.8	21.3	27.8	31.7	33.0	32.9	36.5	40.0
Ferrous metals	0.8	3.4	3.7	20.4	24.1	27.7	32.8	35.0	36.5
Aluminum	-	16.7	26.1	35.9	35.6	38.7	35.8	37.8	34.6
Other nonferrous metals	42.9	45.5	50.0	66.4	65.5	63.4	63.1	73.3	69.5
Glass	1.6	5.3	7.6	20.0	20.3	22.0	22.1	23.3	24.5
Plastics	-	-	0.9	2.2	2.5	3.3	3.5	4.9	5.3
Yard waste	-	-	-	12.0	13.7	15.4	20.8	25.4	30.3
Other wastes	2.4	1.9	2.1	4.9	5.8	6.4	6.8	8.6	8.7

- Represents zero.

Source: Franklin Associates, Ltd., Prairie Village, KS, *Characterization of Municipal Solid Waste in the United States: 1995.* Prepared for the U.S. Environmental Protection Agency.

11. Make sure that Fathom is installed on your laptop.

Problem set 1-2

Mark Twain – "Get your facts first, and then you can distort them as much as you please".
http://www.quotationspage.com/quote/286.html, 12/27/11

1. A family got together for Thanksgiving and the ages of everyone present are given in the stemplot below.
   a. How many people came for Thanksgiving? **22**
   b. What is the maximum and what does it mean with in the context of this data set? **71; The oldest person who came for Thanksgiving was 71.**
   c. What does 6|9 represent? **Stem 6 + Leaf 9 represents the person who was 69.**
   d. What does 5 7 (found in the middle of the stemplot) represent? **57 does not represent anything. 5 is a leaf of 3 so it is part of 35, the person who was 35; 7 is a leaf of 3 so it is part of 37, the person who was 37**
   e. Which age occurs most frequently? **14**
   f. Describe the shape of the distribution and what information does this provide within the context. **The distribution is skewed to the right which tells us that most attendees were young and there were only a few elderly attendees.**

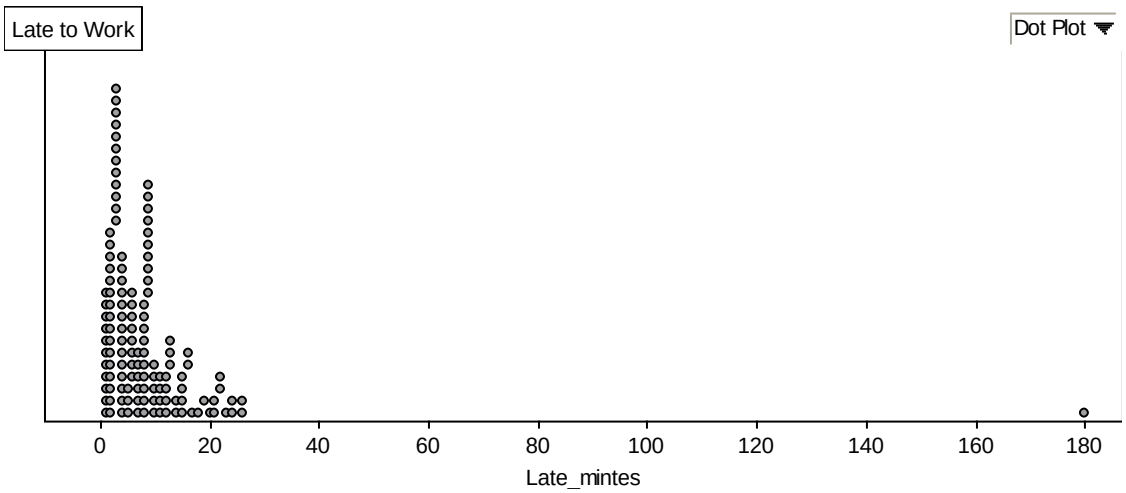
0	0 1 3 3 5 9
1	1 4 4 4
2	7 8 9 9
3	5 7
4	2 3 5 5
5	
6	9
7	1

Legend
3|5 = 35

2. Consider the ages of Deerfield Academy students. Would a stem and leaf diagram be appropriate for this data set? Explain.
   No; two reasons. There would be only one stem and too many leaves.

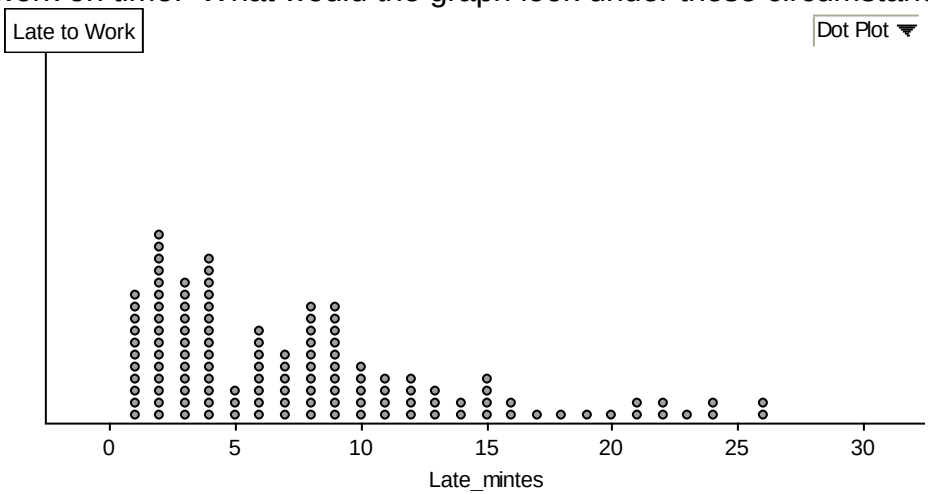
3. 💻 A company with 278 employees was concerned that there was a problem with their employees showing up late to work so management recorded how many minutes each person was late. If they showed up early or on time no information was collected for that employee. Open the Fathom file called "Late to Work" to see the data.

a. Create a dotplot (you do not need to copy the dot plot to your notebook).



b. Six people were late the same number of minutes. How many minutes late were these six people? **7**

c. Let's say that the next day the lateness data is the same with one exception, the person who was the latest (180 minutes late) arrived at work on time. What would the graph look under these circumstances?



d. Using what you learned in (c), under what circumstances is a dotplot an inappropriate graph? **A dotplot is inappropriate when the range of data become too large.**

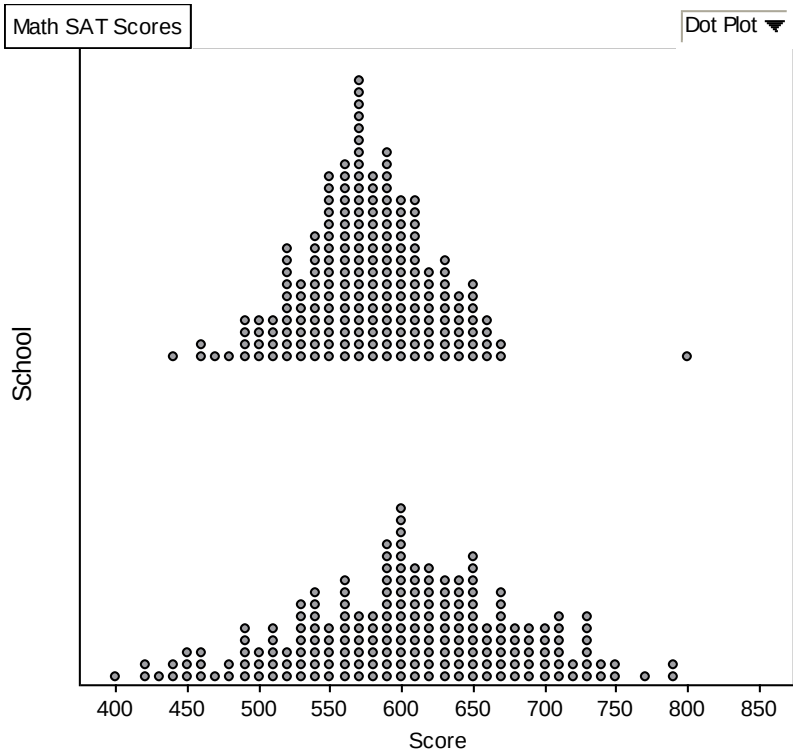
4. 💻 Two schools reported the Math SAT- I scores of their seniors. Open the Fathom file called "SAT Scores" to see the data.

a. How many seniors were in each school?

Math SAT Scores	Summary Table		
⇩	⇨	**School**	
		School 1	**School 2**	Row Summary
		197	201	398

S1 = count ( )

b. Create a dotplot for each school (you do not need to copy the dot plot to your notebook). What was the range of SAT scores for each school?
Range for School 1 is 790 – 400 = 390; Range for School 2 is 800 – 440 = 360



c. Describe the shape of the distribution for each school.
The distributions for both schools are relatively symmetrical.

d. Which school's seniors did better on the SAT? Support your answer.
Sample answer: School 1 did better because 36 out of 197 students scored above 670. Only one student in School 2 scored above 670.

e. You explained why one school did better on the SATs in (c). Now suppose you were the headmaster of the other school. Write a few sentences trying to convince someone that your SAT scores are better than the school you chose in (c). **Sample answer: School 2 did better because the results are much more consistent. Not one student scored below 440. Also, School 2 had a student with a score of 800 while no student at School 1 had a perfect score.**

Important technical note: You should note that a number of problems have a computer icon (💻) in front of the problem. For all 💻 problems you are expected to solve the problem with a computer and save a file you can access in class that represents the solution. You should also transcribe your results to your notebook.

Problem set 1-3

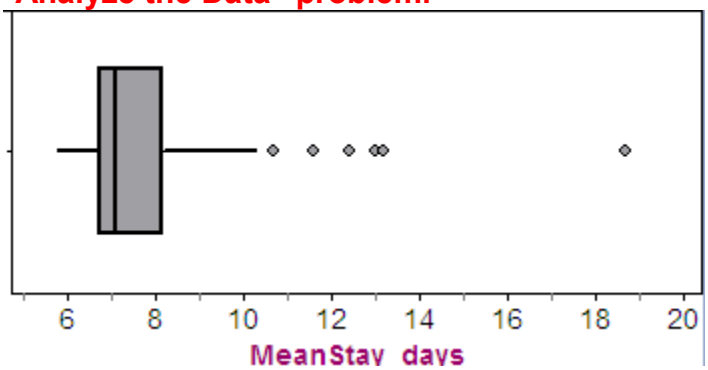
1. 💻 You got an 82 on your first math test. The scores of the class (including yours) are 82, 91, 87, 100, 28, 72, 83, 77, 88, 84, 86, 84. This data set is called "Math Test Scores".
   a. Write a sentence or two that convinces your parents that you did really well on this test. **I did really well, the mean was approximately 80% and I got an 82%.**
   b. Now write as if you are your parents and convince your son/daughter that they didn't do so well after all. **I think you need to work a little harder; one person only earned 28% which brought down the mean. The median was 84% and you did not do as well as that.**

2. In problem (1), assign each score to the variable *s* with the index *i*, that is s_1=82, s_2=91, …, s_{12}=84.
   a. Write an expression for the sum of the scores using sigma notation.
   $$\sum_{i=1}^{12} s_i$$
   b. Write an expression for the mean of the scores using sigma notation.
   $$\frac{\sum_{i=1}^{12} s_i}{12}$$
3. Make up a data set of five data values in which mode is negative and $\bar{x}$ is positive.
   Sample: -1, -1, 0, 1, 2
4. Make up a data set of five data values in which $\bar{x}$ is negative and the mode is positive.
   Sample: -2, -1, 0, 1, 1
5. 💻 In the 1996-97 basketball season, Michael Jordan earned $30,140,000. The rest of the team's salaries can be seen by opening the Fathom file called "Chicago Bulls".
   a. What is the mean salary of the team?
   about $4.2 million
   b. What is the median salary of the team?
   about $1.2 million
   c. If you read a newspaper story about the Chicago Bulls in 1996-97 that used the mean as a measure of center, what type of "spin" would the writer be trying to use? **A newspaper writer might report the mean salary of about $4.2 million if they were trying to make it sound as if all the Bulls are earning a lot.**
   d. If you read a newspaper story about the Chicago Bulls in 1996-97 that used the median as a measure of center, what type of "spin" would the writer be trying to use? **A newspaper writer might report the median salary of about $1.2 million if they were trying to make it sound as if the Bulls were not earning that much.**

Problem set 1-4

1. The 5-number summary is usually listed as min, Q_1, median, Q_3, max. Each one of the 5 has at least one synonym. List one for each.
   min (0^{th} percentile), Q_1(25^{th} percentile), median(50^{th} percentile or Q_2), Q_3(75^{th} percentile), max(100^{th} percentile)

2. 💻 Open the "Mean Hospital Stay" data set in Fathom.
   a. "Analyze the Data".
      Everything below should be written in your notebook for every "Analyze the Data" problem.



   2. The distribution is skewed to the right.
   3. Five number summary: min=5.8, Q1=6.7, med=7.1, Q3=8.2, max=18.7
   4. Identify any outliers (1.5 X IQR criterion)
      IQR=8.2 - 6.7 = 1.5
      1.5 x IQR = 2.25
      Upper cutoff = Q3 + 1.5 x IQR = 8.2 + 2.25 = 10.45. All number above 10.45 are outliers.
      MA, NJ, SD, ND, PA, MI, and MT are upper outliers
      Lower cutoff = Q1 - 1.5 x IQR = 6.7 - 2.25 = 4.45. All number below 4.45 are outliers.
      No lower outliers
   5. Mean=7.994
   6. Population standard deviation is approx. 2.49

   b. Write a complete sentence describing what Q_1 means within the context of this problem. **In 1993, 25% of the states had a mean stay in the hospital of less than 6.7 days.**
   c. What percentage of the states have a mean stay between Q_1 and Q_3? **50%**
   d. Which measure of center, mean or median, seems more appropriate with this data set and why? **Median, because of upper outliers.**
   e. Can you make any geographic generalizations about the states that have comparatively long vs. short mean stays in hospitals. **No, it does not seem possible in this case to make a geographic generalization.**

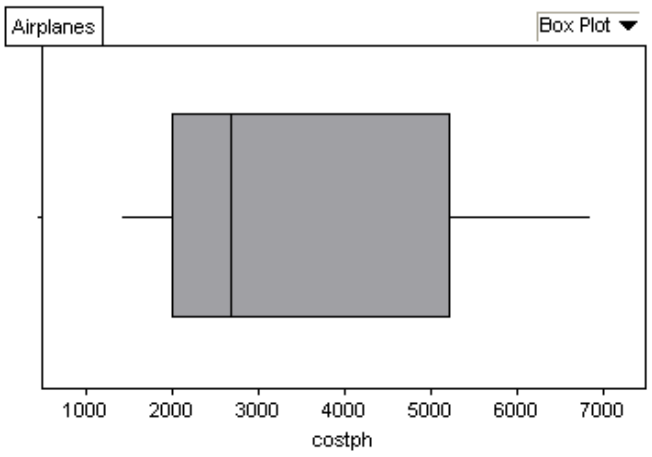
3. Give an example of a data set where a box plot is not appropriate.
   The batting averages of the starters on a baseball team. There are too few data values to use a box plot.

4. What shape of a box and whisker plot or dot plot will result in the median being appreciably different from the mean, even when there are no outliers? **One that is not symmetrical – one that is skewed.**
5. 💻 Open the Fathom data set called "Airplanes"
   a. Create a box and whisker plot of the attribute "costph": cost of operating in dollars per hour. Without doing any calculations, predict which will be bigger, the mean or the median, just by looking at the graph. Write a sentence which explains your reasoning. **From the box plot it seems that the mean is higher than the median because the data is skewed (spread out) to the right.**
   b. "Analyze the data" for the attribute "costph": cost of operating in dollars per hour. Check to see if your conjecture in (a) was correct.

New Collection

Airplanes

	costph
1	6859
2	6447
3	3720
4	5281
5	5237
6	6078
7	3558
8	2675
9	3348
10	2177
11	2504
12	2124
13	2087
14	1918
15	1923
16	1904
17	1988
18	2002
19	1409
20	4241
21	6406



Airplanes — Summary Table

	costph
	3518.381
	1811.309

S1 = mean ( )
S2 = sampleStdDev ( )

1. Box Plot
2. The distribution is slightly skewed to the right
3. Five number summary: min=1409, Q1=2002, med=2675, Q3=5237, max=6859
4. Identify any outliers (1.5 X IQR criterion)
   IQR=5237 - 2002 = 3235
   1.5 x IQR = 4852.5
   Upper cutoff = Q3 + 1.5 x IQR = 5237+ 4852.2 = 10,089.5. All number above 10,089.5 are outliers.
   There are no upper outliers
   Lower cutoff = Q1 - 1.5 x IQR = 2002 - 4852.2 =-2850.5. All number below -2850.5 are outliers.
   There are no lower outliers
5. Mean=3,518.4
6. Sample standard deviation is approx. 1,811.3

6. 💻 Open the Fathom data set called "2010 Midsize car fuel economy".
   a. Create a box and whisker plot of fuel economy. Based on the box and whisker plot alone, are there any outliers? If so, which car(s)?
   b. Test for outliers using the 1.5 x IQR and verify algebraically your results in part (a). **This problem is left for you to do; verify your results in class.**

Problem set 1-5

1. Standard deviation is a measure of _____ **spread** _____ (fill in the blank).

2. Which set of numbers has a higher standard deviation?
   a. Set 1: 8, 12, 9, 11, 7, 13 — Set 2: 2, 18, 1, 19, 3, 17 **Set 2**

   b. Set 1: 108, 112, 109, 111, 107, 113 — Set 2: 2, 18, 1, 19, 3, 17 **Set 2**

   c. Set 1: 108, 112, 109, 111, 107, 113 — Set 2: 8, 12, 9, 11, 7, 13 **Same std dev**

3. Estimate the standard deviation of the following set of numbers.

a.	90	90	90	90	90	90	90	90	90	90	**= 0**
b.	85	85	85	85	85	95	95	95	95	95	**= 5**
c.	90	91	89	90	88	92	90	89	90	91	**≈ 1**
d.	88	92	79	86	94	91	94	86	90	93	**≈ 4.5**
e.	79	99	97	89	76	87	93	65	97	95	**≈ 10.5**

   For a, b, and c, any number close to the given answer is fine.

4. According to WebMD, "Your blood pressure consists of two numbers: systolic and diastolic. Someone with a systolic pressure of 120 and a diastolic pressure of 80 has a blood pressure of 120/80, or "120 over 80." The systolic number shows how hard the blood pushes when the heart is pumping. The diastolic number shows how hard the blood pushes between heartbeats, when the heart is relaxed and filling with blood. High blood pressure (hypertension) is 140/90 or higher".

   The World Health Organization collected blood pressure for samples of various countries and some the data is in the table below.

	Systolic		Diastolic	
Country	Mean	Std Dev	Mean	Std Dev
China	129	24	80	13
Italy	129	19	82	10
Switzerland	128	19	77	11
Yugoslavia	135	19	86	11

   a. Of the countries listed, which has the biggest problem with hypertension?
      Yugoslavia. They have the highest systolic and diastolic means.
   b. Compare the systolic values for China and Italy.
      On average the two countries have the same systolic values, but China has a much bigger variability (higher spread of the data).

c. Compare the systolic values for Switzerland and Yugoslavia.
The two countries have the same variability of systolic values (same spread of the data), but Yugoslavia has a much larger average.

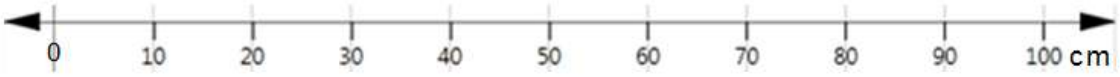
Sources: WebMD, Hypertension, http://www.webmd.com/hypertension-high-blood-pressure/tc/high-blood-pressure-hypertension-overview, 12/23/12
World Health Organization, Blood Pressure by Country, http://www.ktl.fi/publications/monica/bp/table8.htm, 12/23/12

5. The Journal of the American Medical Association published an article in Sept. of 2011 titled "Lesbian, gay, bisexual, and transgender-related content in undergraduate medical education". The study sampled US medical students asking how many hour they spent on LGBT health. The mean was 7 hours with a standard deviation of 6.5 hours.
   a. Create a data set with 10 numbers that you think would have a mean of 7 hours and a standard deviation of 6.5 hours.
   b. Now enter these numbers into Fathom and see if you were close. If not, change the numbers until the mean is 7 and the standard deviation is 6.5.
   Sample data set. Note, no matter what data set you ended up with, students either had little/no education on LGBT issues or a lot.

   0 0 1 2 5 6 7 13 18 17

American Medical Student Association, On Call, http://www.amsa.org/AMSA/Homepage/TakeAction/AMSAOnCall/11-09-14/Are_Med_Students_Receiving_Adequate_Education_on_LBGT_Issues.aspx, 12/23/11

6. Below are images of Barracuda of various lengths. Estimate the mean and standard deviation length of the fish. Zoom out to see entire page. **Mean ≈ 67, std. dev ≈ 9**
Image source: Florida Fishing Info, Naples Fishing, http://www.floridafishinginfo.net/naples_florida_fishing.html, 11/20/11

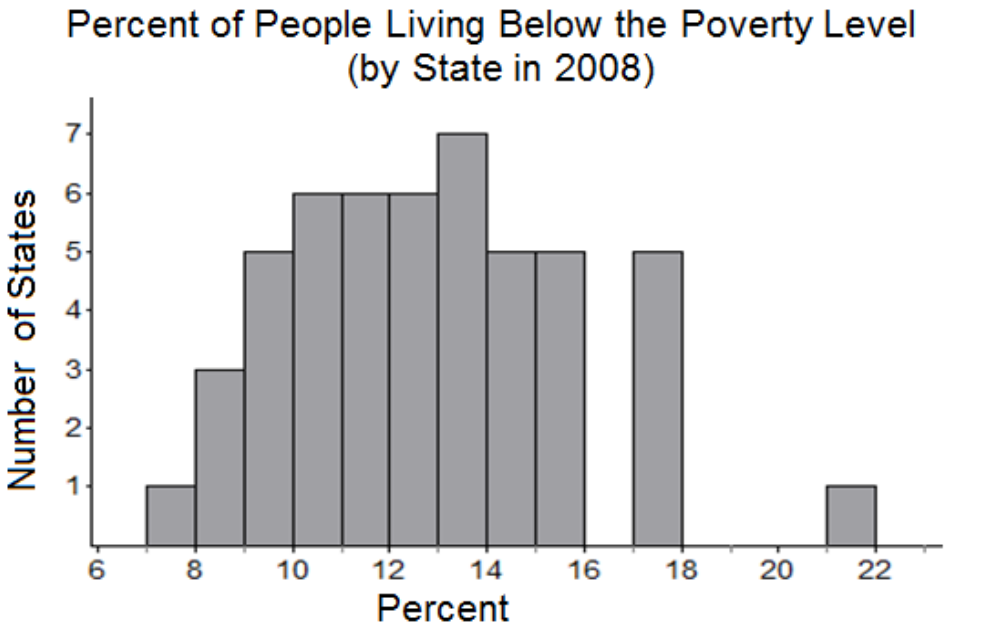


Problem set 1-6

Note: For this problem set assume that histograms include the left endpoint of the interval and not the right. This is the default for most software.

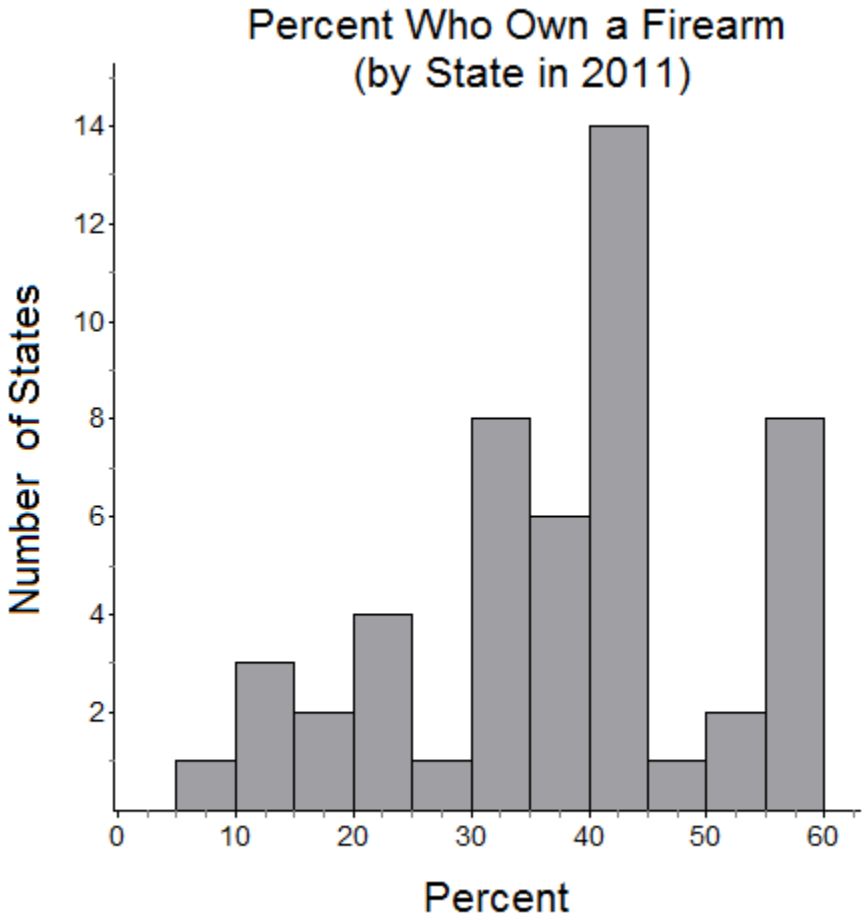
1. The histogram below gives the percent of people living below the poverty level for each state in the US. In 2008, 13.2% of people in the US were living below the poverty level with Mississippi at 21.2% and New Hampshire at 7.6%.
   a. In what interval of the histogram does Mississippi lie?

      Give your answer in double inequality notation. $21 \le x < 22$
   b. Write a sentence that gives meaning to the tallest bar of the histogram.
      In 7 of 50 states (14%), the percentage of people that lived in below the poverty level was in the interval $13 \le x < 14$.
   c. What percent of states have a poverty level ≥ 17%?
      $\frac{6}{50} = 12\%$
   d. Estimate the standard deviation of percent that live below the poverty level for all 50 states. **About 3%; any answer between 2% and 5% is OK.**



Source:Census Bureau, Poverty, http://www.census.gov/compendia/statab/2012/ranks/rank34.html, 12/26/11

2. The histogram below gives the percent of people who own a firearm for each state in the US.
   a. If in a state exactly 30% of the population owned a firearm, in what interval would it lie? Give your answer in double inequality notation. $30 \le x < 35$
   b. In what percent of states did at least 40% of their citizens own a firearm? $\frac{25}{50} = 50\%$
   c. Estimate the standard deviation of the percent who own a firearm by state for all 50 states. **About 13%; any answer between 10% and 16% is OK.**
   d. Wyoming, Alaska, Montana, South Dakota, West Virginia, Arkansas, Idaho, and Mississippi are the eight states that have firearm ownership abo e 55%. If these states all had a reduction in firearm ownership and other states stayed the same, what would happen to the standard deviation and why? **The standard deviation would go down because the data would be less spread out.**

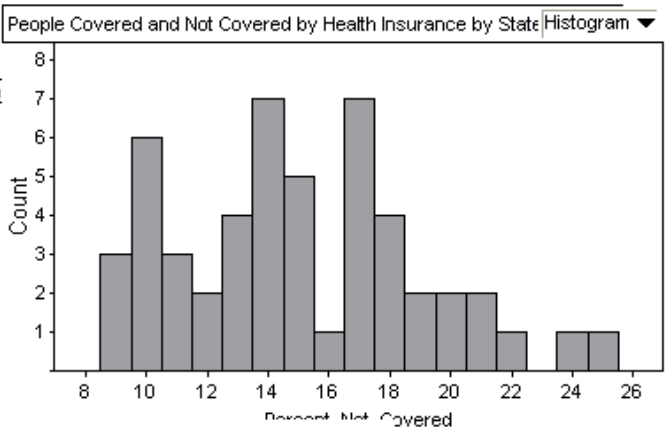


Source: NC Health Statistics, Firearms, http://www.schs.state.nc.us/SCHS/brfss/2001/us/firearm3.html, 12/26/2011

3. 💻 Open the Fathom file called “US Health Ins. Coverage”.
   a. “Analyze the Data” for the variable “Percent not Covered”.

People Covered and Not Covered by Health Insu...

	State	Total_1000s
1	Nebraska	1716
2	Iowa	2837
3	Minnesota	4833
4	Vermont	593
5	Hawaii	1201
6	Rhode Island	968
7	Kansas	2616
8	Massachusetts	6117



1. Histogram
2. The distribution is skewed to the right
3. Five number summary: min=9.0, Q1=11.8, med=14.7, Q3=17.5, max=24.5
4. Identify any outliers (1.5 X IQR criterion)
   IQR=17.5 - 11.8 = 5.7
   1.5 x IQR = 8.55
   Upper cutoff = Q3 + 1.5 x IQR = 17.5 + 8.55 = 26.05. All number above 26.05 are outliers.
   No upper outliers
   Lower cutoff = Q1 - 1.5 x IQR = 11.8 - 8.55 = 3.25. All numbers below 3.25 are outliers.
   No lower outliers
5. Mean=15.1
6. Population standard deviation = 3.9

People Covered and Not Covered by Healt...

⇩	⇨ Percent_Not_Covered
	15.058824
	3.9013912

S1 = mean ( )
S2 = popStdDev ( )

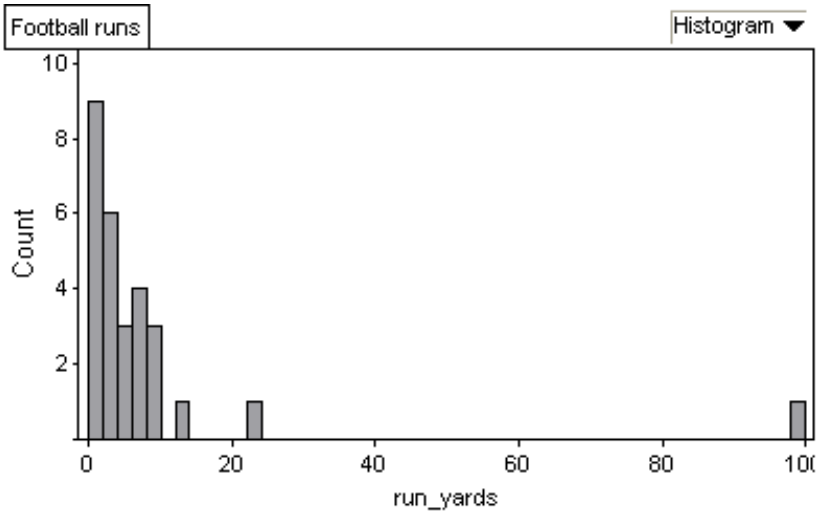
   b. Create a histogram for the variable “Percent not Covered”. What is the appropriate minimum and maximum bin width (there is no exact answer to this question)? **min ≈ 1.0 (or maybe 0.5), max ≈ 2.0**
   c. If your intention is to point out that Texas and Arizona have a problem with a high percentage of people not covered by health insurance, would you use the minimum or maximum bin width? Why? **Use the minimum because it sets Texas and Arizona apart from the others.**

4. Open up the Fathom file called “Football Runs”.
   a. “Analyze the Data”.

Football runs

Football runs

	run_yards
1	0
2	0
3	1
4	1
5	1
6	1
7	1
8	1



1. Histogram
2. The distribution is skewed to the right
3. Five number summary: min=0, Q1=1.0, med=2.5, Q3=6.5, max=99
4. Identify any outliers (1.5 X IQR criterion)
   IQR=6.5 - 1.0 = 5.5
   1.5 x IQR = 8.25
   Upper cutoff = Q3 + 1.5 x IQR = 6.5 + 8.25 = 14.75.
   All number above 14.75 are outliers.
   22 and 99 are upper outliers
   Lower cutoff = Q1 - 1.5 x IQR = 1.0 - 8.25 = -7.25.
   All numbers below -7.25 are outliers.
   No lower outliers
5. Mean=7.8
6. Population standard deviation = 18.1

Football runs — Summary Table

	run_yards
	7.7857143
	18.142998

S1 = mean ( )
S2 = popStdDev ( )

   b. If you were NMH and trying to downplay how well Dan Schribman did, what measure of center would you use and why? **NMH would use the median because it is not affected by the outliers 22 and 99.**
   c. What is an example of a cell width that does not make sense for this data? **If the bin width is 30, all but one run is in the same bin.**
   d. What seems to be the optimal cell width for this data? **1 or 2**

Problem set 1-7

1. Under what circumstances is it helpful to know the z-score?
   When you do not know the mean and standard deviation of the data set.

2. The mean and standard deviation weight for a 6 month old boy are 18.5 pounds and 2.4 pounds respectively. If a 6 month old boy weighs 11.9 pounds, what is his z-score? Source: Child-Specific Exposure Factors Handbook, Infant weights, oaspub.epa.gov/eims/eimscomm.getfile?p_download_id=36528, 12/27/11.

3. The mean and standard deviation weight for a 6 month old girl are 17.0 pounds and 1.8 pounds respectively. If a 6 month old girl has a z-score of 0.3, how much does she weigh?

4. A 2009 study published in the Journal for Human Factors and Ergonometric Society measured the reaction time for drivers who were not texting compared to those who were texting. The results:

Driving Brake Onset Time in milliseconds		
	Mean (ms)	Std Dev (ms)
Not Texting	881	349
Texting	1,077	380

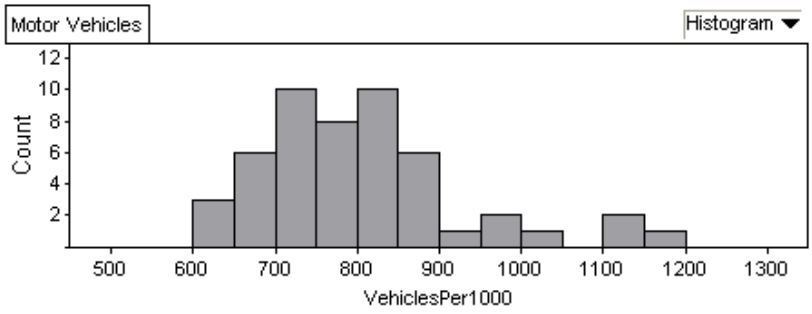
Source: SAGE, Text Messaging During Simulated Driving, http://hfs.sagepub.com/content/51/5/762.full.pdf?keytype=ref&siteid=sphfs&ijkey=gRQOLrGlYnBfc, 12/27/11

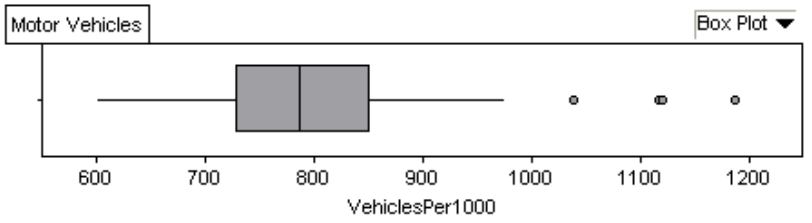
   a. Would a driver rather have a positive or negative z-score for brake onset time? Explain.
   b. If a driver who was not texting had a z-score of +1.0, what is the corresponding break onset time?
   c. If a driver who was texting had a z-score of +1.0, what is the corresponding break onset time?
   d. For a driver who was texting, what z-score would he/she need to have in order to achieve the mean not texting brake onset time of 881 ms?

5. 💻 Open the Fathom data set called “Motor Vehicles” and “Analyze the Data”.

Motor Vehicles

	State	VehiclesPer1000
1	WY	1187
2	SD	1121
3	MT	1117
4	ND	1039
5	TN	975
6	IA	971
7	ID	930
8	NE	892
9	NM	879
10	OR	865
11	OK	857
12	NH	853





1. Relative Frequency Histogram and Box plot
2. The distribution is skewed to the right
3. Five number summary: min=600, Q1=728, med=786.5, Q3=851 max=1187
4. Identify any outliers (1.5 X IQR criterion)
   IQR=851 - 728 = 123
   1.5 x IQR = 184.5
   Upper cutoff = Q3 + 1.5 x IQR = 851 + 184.5 = 1,035.5. All number above 1,035.5 are outliers.
   WY, SD, MT, ND upper outliers
   Lower cutoff = Q1 - 1.5 x IQR = 728 - 184.5 = 543.5. All numbers below 543.5 are outliers.
   No lower outliers
5. Mean=805.24
6. Population standard deviation = 123

Motor Vehicles | Summary Table

VehiclesPer1000	805.24
	122.9796

S1 = mean ( )
S2 = popStdDev ( )

6. 💻 Refer to the data set for Motor Vehicles:

	St...	VehiclesPer1000
1	WY	1187
2	SD	1121
3	MT	1117
4	ND	1039
5	TN	975
6	IA	971
7	ID	930
8	NE	892
9	NM	879
10	OR	865
11	OK	857
12	NH	853
13	CO	851

Motor Vehicles

VehiclesPer1000	805.24
	50
	124.22816
	17.568515
	0

S1 = mean ( )
S2 = count ( )
S3 = stdDev ( )
S4 = stdError ( )
S5 = count (missing ( ))

a. What is the z-score for Wyoming? What does that z-score mean?
(1187 - 805.24)/124.23 ≈ 3.1. Wyoming has a z-score of about 3.1 which means the state's Vehicles per 1000 people is 3.1 standard deviations above the national mean.
b. What state got the lowest z-score? Write a complete sentence describing what that z-score means in the context of motor vehicles?
(600 - 805.24)/124.23 ≈ -1.7. New York has a z-score of about -1.7 which means the state's Vehicles per 1000 people is 1.7 standard deviations below the national mean.
c. What is true about all states that have negative z-scores?
They are below the mean.

7. 💻 You are taking Physics and you just took your 1st test. The scores of everyone in the class are 81, 80, 80, 81, 81, 82, 80, 81, 82, 80, 80, 81. (The data set is called "Physics Test 1".)
   a. Was this a fair test? Why or why not? **No. There was not enough range of difficulty. About 80% of the test was too easy and 20% was too hard.**
   b. Your friend in the class is bummed because he got an 80. What would you tell him? **Sample answers – You only scored 2% less than the maximum. The scores were bi-modal and your score was one of the modes.**

8. 💻 The scores on the 2nd test were 79, 79, 85, 79, 89, 87, 94, 97, 72, 74, 79, 79. You got a 79. Write a short paragraph on how you did. (The data set is called "Physics Test 1".) **I earned both the mode and median, but scored lower than the mean of 82.75. My z-score was about -0.5. I did "OK".**

9. 💻 Three people got an 88% on a test, one in period 3, one in period 5, and one in period 7. The scores of all the students in the three classes are shown below. Did they all do the same relative to their classmates? Who did the best? Who did the worst? Explain. (The data set is called "Test Scores-Periods 3,5,7".)

   Period 3: 90, 77, 77, 83, 79, 77, 92, 84, 83, 83, 88
   Period 5: 88, 62, 75, 83, 78, 74, 96, 84, 83, 94, 96
   Period 7: 82, 82, 82, 82, 82, 83, 83, 83, 83, 83, 88
   The student in period 7 did the best, earning a z-score of over 3. The student in period 5 did the worst of the 3 students, earning a z-score of 0.5. Note: you can't always just compare to the mean. The mean in all three classes is the same.

Test Scores - Periods 3, 5, and 7

	Class			Row Summary
	Period 3	Period 5	Period 7	
Score	11	11	11	33
	5.2915026	10.469002	1.7320508	6.6285368
	1.5954481	3.1565228	0.52223297	1.1538801
	0	0	0	0
	83	83	83	83

S1 = count ( )
S2 = stdDev ( )
S3 = stdError ( )
S4 = count (missing ( ))
S5 = mean ( )

10. Fill in the following table with Yes, No, NA (Not Applicable), and comments as appropriate.

Characteristic / Display	Shown on Display?				Appropriate for?		Important Characteristics:
	Individual Data Points	5-Number Summary	Mean	Std. Dev.	Large Data Sets	Small Data Sets	
Dot Plot	Yes	No	No	No	No n<150	Yes	Displays freq. distribution of small data sets & mode
Stem Plot	Yes	No	No	No	No n<150	Yes	Displays freq. distribution of small data sets & min/max
Bar Graph	Maybe	N.A. (Not Applicable)	N.A.	N.A.	Yes	Yes	One variable is a category, other is frequency
Frequency Histogram	No	No	No	No	Yes	OK	Displays number of values in each interval
Relative Frequency Histogram	No	No	No	No	Yes	No	Displays the percent in each interval
Box Plot	No	Yes	No	No	Yes	No n>20	Shows the 5-number summary

11. Which display(s) would not be appropriate for the set of heights of all students at Deerfield Academy? Explain why. Which display(s) would be appropriate?
Not appropriate – Bar, dot, stem
Appropriate – Frequency histogram, relative frequency histogram, box plot

12. Which display(s) would not be appropriate for the set of ages of students in this class? Explain why. Which display(s) would be appropriate?
    Not appropriate – Box plot, Bar
    Appropriate – Frequency histogram, relative frequency histogram, dot, stem

Problem set 1-8

Project 1 – Global Warming

We will have an on-line discussion about whether global warming is "generally exaggerated" or not. You will be graded on how convincing you are, not your opinion. We will have three rounds of discussion and every student will make one post and reply to another student's post in each round. The first round is a practice round (otherwise identical to round one) in which you will get used to the etiquette of communicating in an on-line discussion and I will critique your work in class so you get a better idea of what is expected. During the second round, half of the students in the class will randomly be assigned to argue that global warming is generally exaggerated and the other half will argue that it is not generally exaggerated. During the third round you will switch "sides" of the argument. An example of the type of writing that is expected is posted in the "Fracing" on-line discussion. Follow the following guidelines:

1. Base your writing on scientific, statistical evidence.
2. Give a complete citation for your source.
3. Be convincing. After reading your post, the reader should be convinced that you are right. After reading your reply, the reader should be convinced that the writer of the post is wrong and you are right.
4. Be succinct and concise. Each post and reply should be about one paragraph with one or two major points.
5. You can include a graphic, but if you do the graphic should be used to strengthen your argument. A graphic that is simply inserted without clear integration into the text detracts from your argument (and your grade).

Project 2 - TransFats

💻 The New England Journal of Medicine published an article (354;15, P. 1651) in April of 2006 which included, among other things, information about trans fat in french fries for many cities around the world. The article has a table which contains "a comparison of the amounts of industrially produced trans fatty acids in a large serving" of french fries from McDonalds and KFC. The values in the table are trans fatty acids as a percent of the total fat. According to information provided in the article, "daily intake of about 5 g of trans fat is associated with a 25% increase in the risk of ischemic heart disease. For this reason, it is recommended that the consumption of trans fat be as low as possible".

The data published in the New England Journal of Medicine published are found in a Fathom file in Chapter 1→ 1-8: "Spin the Data" Project →French fries (Fathom).

You will write two articles that might be published in a professional journal in which readers all have taken statistics (assume they know everything you learned in FST in Chapter 1). The first article will be written with a spin that makes McDonalds french fries look healthier than KFC french fries. The second

article will be written with a spin that makes KFC french fries look healthier than McDonalds french fries.

The report for each fast food chain should be a maximum of one page in length that includes summary statistics, graphs, and a convincing discussion of why the french fries from that chain are healthier.

(1) New England Journal of Medicine, 354;15, P. 1651, April 13, 2006

Problem set 1-9

💻 “Who wrote the Federalist papers?” This is a statistics lab about The Federalist Papers, which were written between 1787 and 1788 to persuade the citizens of the State of New York to ratify the Constitution.

The idea for this lesson comes from FST 2nd edition, UCSMP, Usiskin et all, Scott Foresman, Chapter 1.