# *Stat 101 Problem Set #2 -- Due Thursday Sept. 18*

1   **Thinking about Shape** Would you expect distributions of these variables to be uniform, unimodal, or bimodal? Symmetric or skewed? Explain why.
   a)   The number of speeding tickets each student in the senior class of a college has ever had.
   b)   Players' scores (number of strokes) at the U.S. Open golf tournament in a given year.
   c)   Weights of female babies born in a particular hospital over the course of a year.
   d)   The length of the average hair on the heads of students in a large class.

2   **iPod songs.** Investigate how many iPod songs Stat 101 has on its iPods (from the data set Survey). Using appropriate graphical and numerical summaries, write up a few sentences describing what you find.

3   **Zipcodes** Holes R Us, an Internet company that sells piercing jewelry keeps transaction records on their sales. At a recent sales meeting, one of the staff presented a histogram of the zipcodes of their last 500 customers to understand where their sales are coming from. Comment on the usefulness and appropriateness of the display.
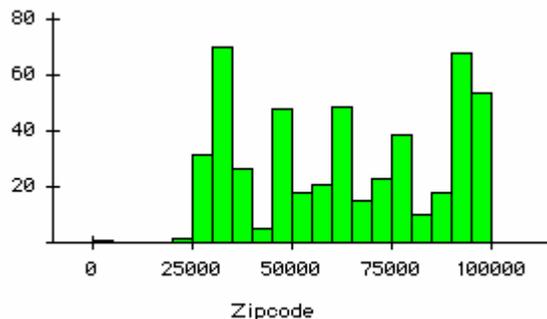
4   **Assets** Here is a histogram of the assets (in millions of dollars) of 79 companies chosen from the *Forbes* list of the nation's top corporations*.*
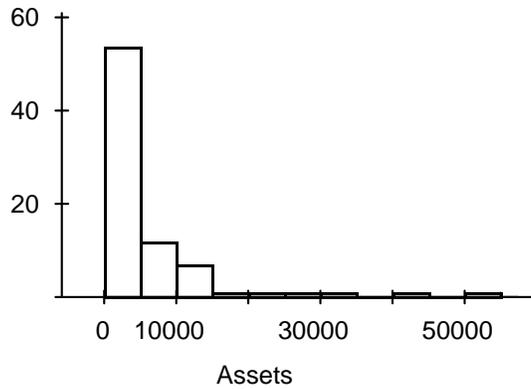
Assets

a) What aspect of this distribution makes it difficult to summarize, or to discuss center and spread?

b) Here are the same data after reexpressions as the square root of assets (on the left) and the logarithm of assets (on the right) Which reexpression do you prefer? Why?
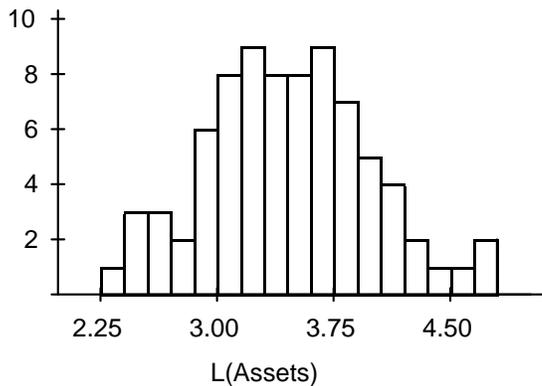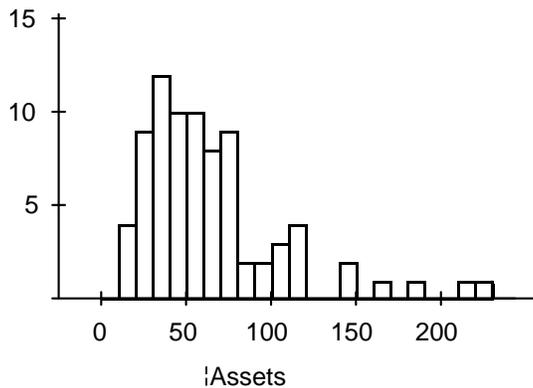
√Assets

L(Assets)

c) In the square root reexpression, what does the value 50 actually indicate about the company's assets?

d) In the logarithm reexpression, what does the value 3 actually indicate about the company's assets?

5   **Wild card Summer Olympics** (Data on Blackboard) Seventy one swimmers finished the qualifying first day of the men's 100 m swim in Sydney in 2000. The mean time was 52.65 seconds with a standard deviation of 7.66 seconds. The median time was 51.34 seconds and the IQR was 2.58 seconds.

a) What shape would expect the distributions of times to have?

b) What might account for the difference between these statistics?
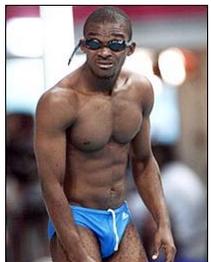
c) Write up a couple of sentences analyzing the times. You might want to read the text box to the right.

6 **More Olympics.** In running, the best runners (either 2 or 4) in each heat qualify for the next round of the event. In swimming the qualifying swimmers are the ones with the fastest times in the previous round. Look at the times from the 2004 Olympics for the Women's 400m event in each sport. Describe what's going on in terms of the times for each heat. Do you think athletes are randomly assigned to each heat? Explain. (Data on Blackboard – Womens 400m Swimming Heats and Womens 400m Running Heats).

7 **Hams** A specialty foods company mails out "gourmet hams" to customers willing to pay a gourmet price. The hams vary in size from 4.15 to 7.45 pounds, with a mean weight of 6 pounds and standard deviation 0.65 pounds. The quartiles and median weights are 5.6, 6.2, and 6.55 pounds.
   a) Find the range and the IQR of the weights.
   b) Do you think the distribution of the weights is symmetric or skewed? If skewed, which way? Why?
   c) If these weights were expressed in ounces (1 pound = 16 ounces) what would the mean, standard deviation, quartiles, median, IQR, and range be?
   d) When the company ships these hams the box and packing materials add 30 ounces. What are the mean, standard deviation, quartiles, median, IQR, and range of weights of boxes shipped (in ounces)?
   e) One customer made a special order of a 10-pound ham. Which of these summary statistics might not change if that data value is added to the distribution?

8 **Professors** A friend tells you about a recent study dealing with the number of years of teaching experience among current college professors. He remembers the mean, but can't recall whether the standard deviation was 6 months, 6 years, or 16 years. Tell him which one it must have been, and why.

9 **TV watching** (data on Blackboard) A survey of 200 college students at Williams College conducted during the week of March 15 showed the following distribution of the number of hours of TV watched per week

   a) According to the normal model, what percentage of students will watch fewer than 1 standard deviation below the mean number of hours?
   b) For these data, what does that mean? Explain.

c) Explain the problem in using the normal model for these data.

10 **Stem-and-leaf** Another invention of J.W. Tukey is the stem-and-leaf display (see page 51 in *Intro Stats*). They are most useful as a "do it by hand" analysis tool – I usually start scratching down a stem-and-leaf, for example, as I grade midterms, to see how the class is doing as I go along. Their usefulness is, admittedly, getting less as statistical software (and even programs like Excel) become ubiquitous. You can actually do one in JMP. They're found under **Analyze → Distributions.** Then under the red triangle next to the variable name you'll see Stem and Leaf. Look at the times from the Wild Card swim times in problem 5 and produce the stem-and-leaf display of the times. What does 4|9 mean? Now set aside the outlier and redo the stem-and-leaf. What does 49 | 2 mean now? Comment.