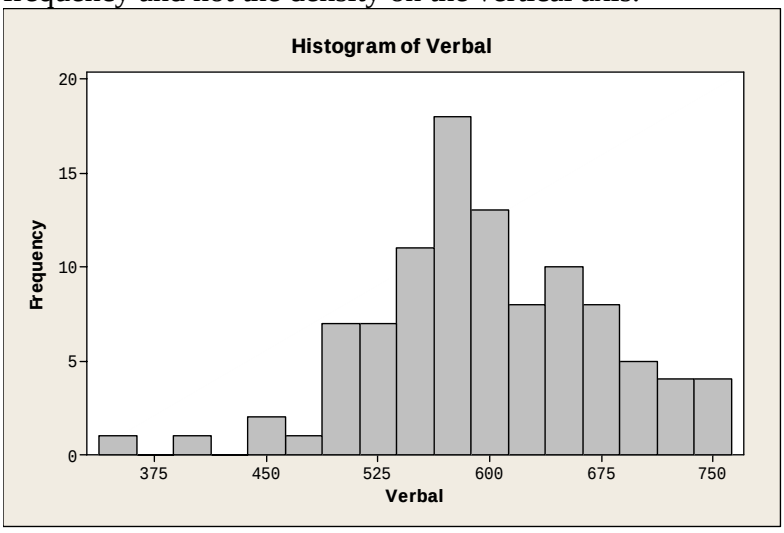
Lab #2 – Mathematics 140
Statistics, Normal Probability, Correlation, and Regression

In the first assignment, we learned a little about Minitab. You should know how to put numbers into the software, how to find simple statistics like the average and the median.

1. Open the Minitab worksheet called **Ga.** This data can be downloaded from our course webpage http://www.csub.edu/~rpeck/Math140/. Go to the course webpage, click on the link for **Labs** and then the link for **GA.mtw**. This data file contains data collected to study the relationship between SAT scores (Scholastic Aptitude Tests, often used as college admissions or course placement criteria) and grade point averages (GPA on a 0.0 to 4.0 scale). The data set contains the GPA, verbal SAT score, and mathematical SAT score for a random sample of 100 students from a northeastern university.

   a. Make a Histogram for verbal SAT score (column C1). Notice Minitab, plots the frequency and not the density on the vertical axis.



   b. Based on the histogram, what are the minimum and maximum values for verbal SAT score? Is your answer an estimate or an exact value?
   Tough to say… the groups are about 25 wide, so…
   minimum about 350 maximum about 762

   c. Make a stem-and-leaf display for the verbal SAT score. Using the stem-and-leaf display make a Histogram **BY HAND** for this data. Use categories 350 – 399, 400 – 449, 450 – 499, etc.

```
3 | 6
4 | 0
4 | 5688899
5 | 0001222333334
5 | 55555566666777777788888888999999
6 | 000011112223333444444
6 | 556666668888999
7 | 00111244
7 | 55
```

   The histogram made by hand should look exactly like this stem-and-leaf plot.

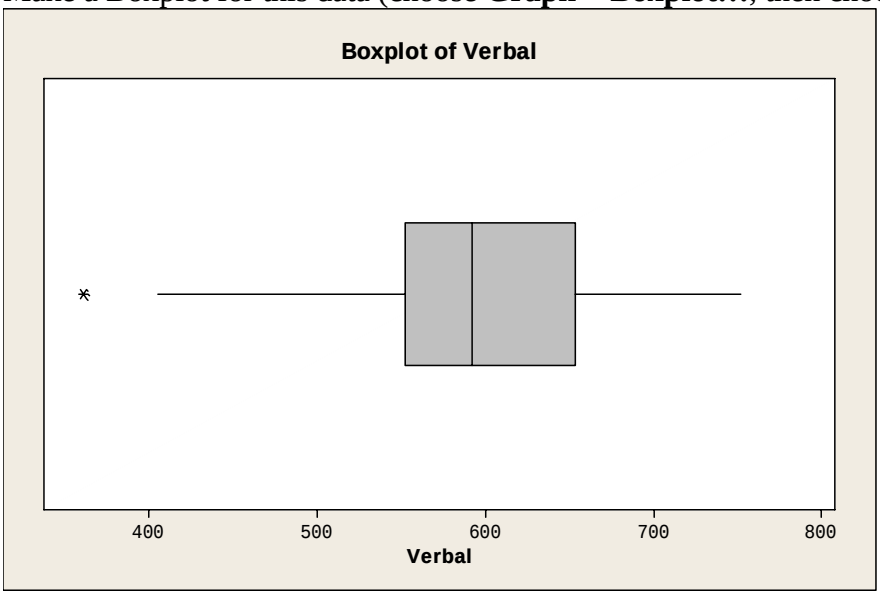
   d. What percent of the students had a verbal SAT score of 700 or higher?
   There are ten numbers on the last two stems above; thus 10/100 = 10%

e. Determine the mean and standard deviation for verbal SAT score.

f. Determine the median, and lower and upper quartiles for verbal SAT score.

Variable	Mean	StDev	Q1	Median	Q3	Maximum
Verbal	598.49	76.03	552.00	592.50	653.25	752.00

g. Make a Boxplot for this data (choose **Graph > Boxplot…**, then choose the **Simple** icon).



h. Are there any outliers? If so, what are there value(s). What is the shape of the distribution?

The minimum verbal SAT score—361 – is an outlier. The distribution is left-skewed.

i. About what percentage of the verbal SAT scores do we expect (according to the 68-95-99.7 rule, also known as the empirical rule) to be within 1 standard deviation of the mean? Count the number of verbal SAT scores within one standard deviation of the mean (use the numbers in part e).

We expect 68% of the scores to be within 1 standard deviation of the mean

$\bar{x} - s = 522.46 \qquad \bar{x} + s = 674.52$

67 of the 1100 numbers (67%) are in the interval [523, 674]

j. About what percentage of the verbal SAT scores do we expect to be within 2 standard deviations? Calculate the actual percentage of verbal SAT scores that are within two standard deviations of the mean.

We expect 95% of the scores to be within 2 standard deviation of the mean

$\bar{x} - 2s = 446.43 \qquad \bar{x} + 2s = 750.55$

96 of the 1100 numbers (96%) are in the interval [447, 750]

k. About what percentage of verbal SAT scores do we expect to be between 1 standard deviation above the mean and 2 standard deviations above the mean? Calculate the actual percentage of verbal SAT scores that are in this interval.

We expect 13.5% of the scores to be between 1 and 2 standard deviation above the mean. This is half of 95% – 65%

$\bar{x} + s = 674.52 \qquad \bar{x} + 2s = 750.55$

15 of the 1100 numbers (15%) are in the interval [675, 750]

2. You take three exams; your score, the mean score, and the standard deviation of the scores for each exam are given below. Which score has the better relative position? Explain your answer.

Your Score	$\bar{x}$	s	$Z = \frac{x - \bar{x}}{x}$
26.6	23.0	9.0	**0.40**
4.2	2.8	1.3	**1.08**
472.5	423.0	55.0	**0.90**

The second exam has the largest z-score for these three exams and is therefore the exam for which you did relatively the best. A z-score of 1.08 indicates that you scored 1.08 standard deviations above the mean.

3. Assume that the random variable Z follows a Standard Normal distribution (that is, N(0, 1)). Find the following probabilities:

   a. P(Z < 2.46) = **0.993053**
   b. P(Z > -1.35) = **1 – 0.088508 = 0.911492**
   c. P(-0.3 < Z < 0.56) = **0.712260 – 0.382089 = 0.330171**

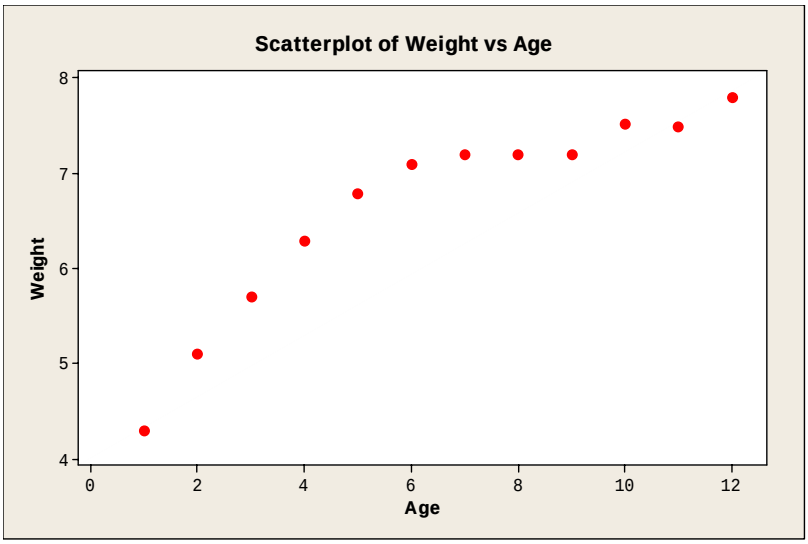
4. Find the following.

   a. Find the number Z such that the proportion of observations that are less than Z in a standard normal is 0.8.
      0.841621
   b. Find a number Z such that 35% of all observations from a standard normal are greater than Z.
      0.385320
   c. Find the 90th percentile for students taking the SAT exam. Scores on the SAT exam follow a Normal distribution with mean 1026 and standard deviation 209.
      1293.84

5. A study of nutrition in developing countries collected data from the Egyptian village of Nahya. Here are the mean weights (in kilograms) for 170 infants who were weighed each month during the first year of life. (Data from Zeinab E. M. Afifi, "Principal components analysis of growth of Nahya infants: size, velocity, and other factors," *Human Biology*, **57**, (1985), pp. 659 – 669.)

Age	1	2	3	4	5	6	7	8	9	10	11	12
Weight	4.3	5.1	5.7	6.3	6.8	7.1	7.2	7.2	7.2	7.52	7.5	7.8

   a. Make a scatter plot of weight (Y) against age (X) (choose **Graph > Scatterplot…**, then choose the icon for **Simple**).



b. Determine the correlation between X and Y (choose **Stat > Basic Statistics > Correlation…**).
`Pearson correlation of Age and Weight = 0.916`

c. What is the regression line predicting weight from age (choose **Stat > Regression > Regression…**)?
`The regression equation is Weight = 4.86 + 0.275 Age`

d. What is the coefficient of determination, R^2, for this regression?
83.9%

e. Compute the correlation coefficient from the coefficient of determination.
$r = +\sqrt{R^2} = \sqrt{0.839} = 0.916$