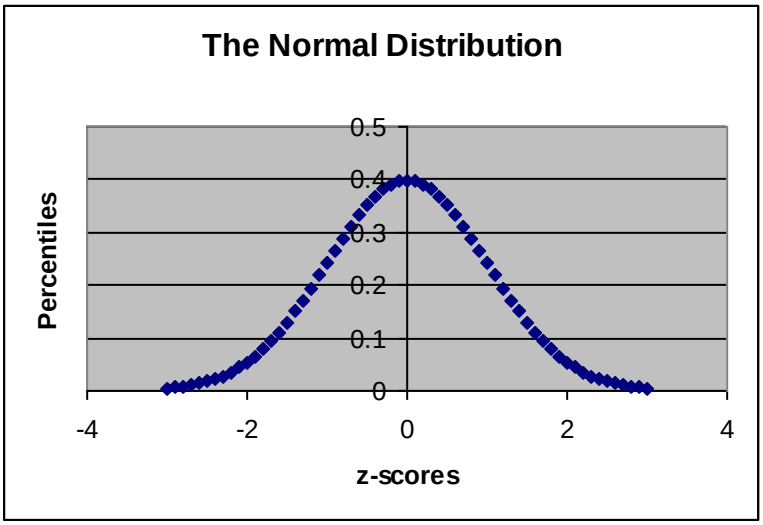
2015

Mathematics in Contemporary Society Chapter 7

Patrick J. Wallach
Queensborough Community College

Chapter 7

Understanding the Normal Distribution



Most distributions of data from a large population (heights, weights, IQ scores, temperatures, stock prices, GPA, income, etc.) can be represented by a **normal distribution**. The normal distribution:

1) Is centered at the mean=median=mode of the population.
2) Is symmetric and bell-shaped with one peak.
3) Is the highest near the mean, where most values of the population are clustered.
4) Has tapering tails on both sides, where large deviations from the mean (outliers) become increasingly rare (but not "impossible").
5) Has values distributed according to the **standard deviation**.

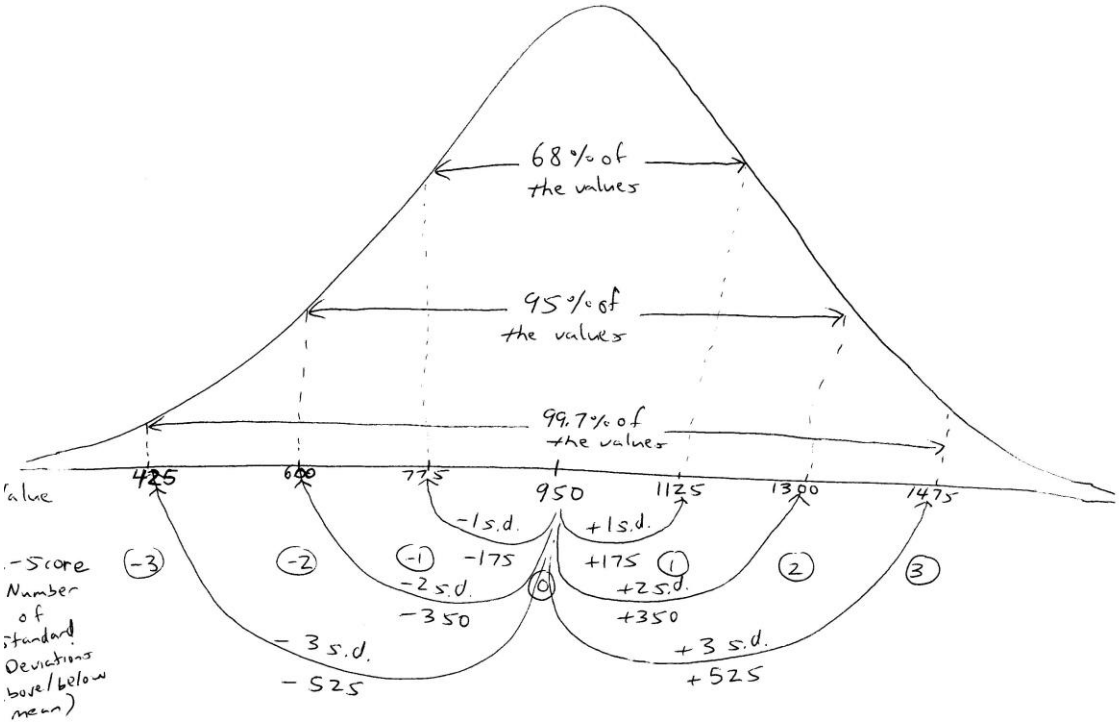
Question 1: Identify a data set (other than the ones mentioned above) that you believe is normally distributed.

The 68-95-99.7 Rule

When we have a normal distribution and the mean and standard deviation are known, we can apply the **68-95-99.7 Rule** to the Normal Distribution. It is always true that:

68% of all population values are within 1 standard deviation of the mean
95% of all population values are within 2 standard deviations of the mean
99.7% of all population values are within 3 standard deviations of the mean (*)

We can apply the 68-95-99.7 Rule to a normal distribution of SAT scores with a mean of 950 and a standard deviation of 175:



We place the mean of 950 at the center of the normal distribution. Remember the mean=median=mode. This essentially means 50% of the values are below 950 and 50% of the values are above 950.

68% of all population values are between 950 – 175 and 950 + 175 (775 to 1125). We get this by adding and subtracting one standard deviation (175) from the mean.

95% of all population values are between 950 – 2•175 and 950 + 2•175 (950 – 350 and 950 + 350 is 600 to 1300). We get this by adding and subtracting two standard deviations (175·2=350) from the mean.

99.7% of all population values are between 950 – 3•175 and 950 + 3•175 (950 – 525 and 950 + 525 is 425 to 1475). We get this by adding and subtracting three standard deviation (175·3=525) from the mean.

Using the example, we can measure results in terms of standard deviations:

1125 is +1 standard deviations above the mean.
1300 is +2 standard deviations above the mean.

1475 is +3 standard deviations above the mean.
775 is -1 standard deviations below the mean.
600 is -2 standard deviations below the mean.
425 is -3 standard deviations below the mean.

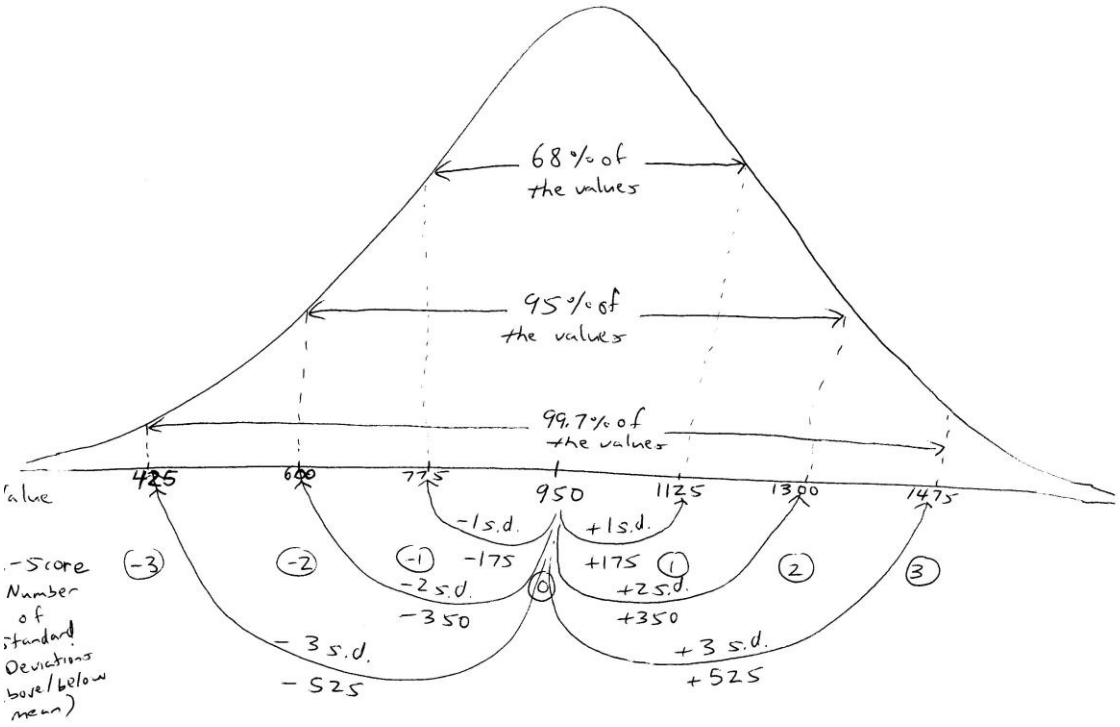
These numbers (+1, +2, etc.) are **z-scores**, which we will discuss soon.

Other values fall somewhere in between. 1200 is between +1 and +2 standard deviations above the mean.

<u>Question 2:</u> Apply the 68-95-99.7 Rule to a normal distribution of IQ scores with a mean of 105 and a standard deviation of 4.6, as seen previously.

a) Which value is +1 standard deviations above the mean?
b) Which value is +2 standard deviations above the mean?
c) Which value is +3 standard deviations above the mean?
d) Which value is -1 standard deviations below the mean?
e) Which value is -2 standard deviations below the mean?
f) Which value is -3 standard deviations below the mean?
g) Give a value that is between +2 and +3 standard deviations above the mean.

Other problems



Using the symmetry of the normal distribution, we can answer other questions as well:

1) What percentage of scores are below 950?
   50%. Because the normal distribution is symmetric, half of the scores are above the mean and half of the scores are below the mean.
2) What percentage of scores are above 950?
   50% as well**.**
3) What percentage of scores are between 950 and 1125?
   34%. The normal distribution is symmetric; if 68% percent of the scores are between 775 and 1125, then 68%/2=34% are between 950 and 1125.
4) What percentage of scores are between 600 and 950?
   47.5%. The normal distribution is symmetric; if 95% percent of the scores are between 600 and 1300, then 95%/2=47.5% are between 600 and 950.
5) What percentage of scores are between 950 and 1475?
   49.85%. The normal distribution is symmetric; if 99.7% percent of the scores are between 425 and 1475, then 99.7%/2=49.85% are between 950 and 1475.
6) What percentage of scores are above 1125?
   16%. This is a slightly different problem. We know that 50% of the scores are above 950. We know (from 3) that 34% of the scores are between 950 and 1125. What's left? 50% - 34% = 16%.
7) What percentage of scores are below 600?
   2.5%. We know that 50% of the scores are below 950. We know (from 4) that 47.5% of the scores are between 600 and 950. What's left? 50% - 47.5% = 2.5%.
8) What percentage of scores are above 1475?
   0.15%. 50% of the scores are above 950. 49.85% (from 5) of the scores are between 950 and 1475. What's left? 50% - 49.85% = 0.15%.
9) What percentage of scores are between 600 and 1125?
   81.5%. This is also a different problem, but related to what we already know. We know (from 4) that 47.5% of the scores are above between 600 and 950. We know (from 3) that 34% of the scores are between 950 and 1125. If we put them together we get 47.5% + 34% = 81.5% of the scores between 600 and 1125.
10) What percentage of scores are between 600 and 1475?
    97.35%. Again, this is related to what we already know. We know (from 4) that 47.5% of the scores are above between 600 and 950. We know (from 5) that 49.85% of the scores are between 950 and 1475. If we put them together we get 47.5% + 49.85% = 97.35% of the scores between 600 and 1475.

These methods work fine if we're dealing with values that are an integer number of standard deviations above or below the mean. Recall that:

1125 is +1 standard deviations above the mean.
1300 is +2 standard deviations above the mean.
1475 is +3 standard deviations above the mean.
775 is -1 standard deviations below the mean.
600 is -2 standard deviations below the mean.
425 is -3 standard deviations below the mean.

These are the values that correspond with the 68-95-99.7 Rule. But what about other values, like 1200? For the moment, all we can say is that 1200 is between +1 and +2 standard deviations above the mean. We need something more than this.

Question 3: Following the answers above, answer the following:

a) What percentage of scores are between 775 and 950?
b) What percentage of scores are between 950 and 1300?
c) What percentage of scores are between 425 and 950?
d) What percentage of scores are below 775?
e) What percentage of scores are above 1300?
f) What percentage of scores are below 425?
g) What percentage of scores are between 775 and 1300?
h) What percentage of scores are between 425 and 1300?

z-scores

Suppose you are told that your value is 10 units above average for the population. Is that good or bad?

It depends on the population we're talking about. Consider the following populations that you could be a part of:

a) **Final exam scores for MA-321.** If your final exam score is 10 points above average, then you probably did fairly well on the exam. This seems like a very good result.
b) **SAT scores for 2010**. Obviously, it's good to be above average, but since SAT scores range from 600-2400, it's less interesting for your SAT score to 10 points above the average score.
c) **Height of QCC students.** 10 inches is almost a foot. It seems to be quite significant to be 10 inches taller than the average person at QCC. Is being that much taller a good thing or a bad thing? It depends.
d) **Weight of QCC students.** Weight has a much larger range. Therefore, 10 pounds over the average weight is less significant than 10 inches over the average height. Still, being 10 pounds over the average weight might be favorable or unfavorable depending on the person. For example, if you're a bodybuilder, you may prefer to be above the average weight.
e) **Annual income of 30 year old adults.** It good to be above average, of course, but $10 above an average in the thousands doesn't seem all that significant.
f) **Adult IQ scores.** IQ scores often range between 90 and 140 for most adults. 10 points above average sounds like a meaningful result.
g) **Credit card debt of the college student.** $10 above average isn't so much for a number that is probably in the hundreds or thousands. But I don't think this is something you ever want to be above average in anyway!

As you can see, being above average is relative to the population we're

talking about. In some cases, it is not ideal to be above average anyway.

Question 4: Suppose you are told that your value is 200 units below average for the population. Choose three different populations and determine the following:
a) Is your value statistically significant?
b) Is this a "good" or "bad" score to have?

We can see that what we need is a way to measure the statistical significance of particular scores. We do!

The z-score (or standard score) of a particular data value in a population or data set is the number of standard deviations that value is above or below the mean. The formula is as follows:

$$z = \frac{(\text{data value} - \text{mean})}{\text{standard deviation}}$$

Example 1

a) Suppose your SAT score of 1050 is part of a population of SAT scores with a mean of 875 and a standard deviation of 150. What is your z-score?

We plug in values into the formula:

$$z = \frac{(\text{data value} - \text{mean})}{\text{standard deviation}}$$

$$= \frac{(1050 - 875)}{150}$$

$$= \frac{175}{150}$$

$$= 1.1666666667$$

$$= 1.2 \text{ (rounded to nearest tenth)}$$

This means your SAT score is 1.2 standard deviations above the mean.

b) Suppose Joe's has an SAT score of 840 (same mean and standard deviation). What is his z-score?

We plug in values into the formula:

$$z = \frac{(\text{data value} - \text{mean})}{\text{standard deviation}}$$

$$= \frac{(840 - 875)}{}$$

150

$= \frac{-35}{150}$

$= -0.2333333333$

$= -0.2$ (rounded to nearest tenth)

This means his SAT score is 0.2 standard deviations below the mean.

**We see in these examples that values below the mean produce negative z-scores and values above the mean produce positive z-scores.

Question 5: Using the same mean and standard deviation as above, calculate the z-score of the following SAT results:
a) 1200
b) 700
c) 1000
d) 1300

What do we do with z-scores?

We know that almost all of the population (99.7%) is within 3 standard deviations of the mean. Therefore, almost all z-scores results will be between -3 and +3.

The larger the z-score (positive or negative), the more statistically significant the result.

Question 6: Using the same mean and standard deviation as above, what is more statistically significant:

a) An SAT score of 1320.
b) An SAT score of 420.

More on z-scores

We saw how to convert any data value to a z-score (or standard score) that represents how many standard deviations the data is above (or below) the mean. But how is that useful?

For one thing, the larger the z-score (whether positive or negative) is, the more statistically significant the result. A z-score of -2.1 is more interesting to a researcher than a z-score of 1.2, statistically speaking.

Consider the 68-95-99.7 rule. If 68% of the values in a normal distribution are within 1 standard deviation, 68% of the z-scores are between –1 and +1. These are common scores, and not all that statistically interesting.

95% of the scores are within 2 standard deviations. Therefore, we expect most (19 out of 20) values to have a z-score between –2 and +2. Results with z-scores between -2 and +2 make up most of a normally distributed population. z-scores between –2 and –1 or +1 and +2 are more statistically interesting than scores between –1 and +1, but not that fantastic.

The remaining 5% is the most interesting statistically. z-scores below –2 or above +2 represent the values of the population (or data set) that are particularly uncommon and rare. These are the results that warrant further examination and explanation.

Finally, anything with a z-score below –3 or above +3 is considered to be extraordinary and fantastic. Only 0.3% of the values in a normal distribution belong in this category. This means, that if you selected 1,000 values from the population, only 3 values would have a z-score below –3 or above +3. That makes them special results.

In the end, the larger the z-score, the more statistically significant it is.

Example 1

One of the things we can do with z-scores is compare results from related examples using different data sets. Consider the following test score results for a student named Susan:

Susan's Class	Susan's Test Score	Class Mean	Class Standard Deviation
History	85	78.6	5.2
English	88	81.4	4.3
Mathematics	82	75.3	3.1
Chemistry	91	85	8.7

Remember that: $z = \frac{\text{(data value – mean)}}{\text{Standard deviation}}$

Question 7:

a) Calculate the four z-scores for each of Susan's test scores.
b) Which result is the most statistically significant?
c) Explain the result in b) to a non-statistics person.

So there are a few things we can do with z-scores by themselves. To make z-scores more useful though, we need to know about percentiles.

Percentiles

Suppose you receive your SAT score from the testing service and are told that your score places you in the 80^{th} percentile. What does that mean? Does that mean your score is an 80? No. Does that mean you answered 80% of the questions correctly? No. It means that

your result is better that 80% of the SAT scores for everyone who took the exam with you.

The nth **percentile** of a data set is the score that is better than n% of the values in the data set.

If your SAT score is at the 40^{th} percentile, your score is better than 40% of all of the SAT scores. If your SAT score is at the 95^{th} percentile, your score is better than 95% of the SAT scores. And so on.

In a situation like SAT scores, the higher the percentile is the better. But there are other cases where a lower percentile is desirable. We wouldn't want our car insurance rates to be in the 90^{th} percentile, higher than 90% of the rates everyone else is paying!

To relate z-scores to percentiles, we have a very nice table:

z-score	Percentile	z-score	Percentile	z-score	Percentile
-4.0	0.003%	-1.1	13.57%	1.4	91.92%
-3.5	0.02%	-1.0	15.87%	1.5	93.32%
-3.4	0.03%	-0.9	18.41%	1.6	94.52%
-3.3	0.05%	-0.8	21.19%	1.7	95.54%
-3.2	0.07%	-0.7	24.20%	1.8	96.41%
-3.1	0.10%	-0.6	27.43%	1.9	97.13%
-3.0	0.13%	-0.5	30.85%	2.0	97.72%
-2.9	0.19%	-0.4	34.46%	2.1	98.21%
-2.8	0.26%	-0.3	38.21%	2.2	98.61%
-2.7	0.35%	-0.2	42.07%	2.3	98.93%
-2.6	0.47%	-0.1	46.02%	2.4	99.18%
-2.5	0.62%	0.0	50.00%	2.5	99.38%
-2.4	0.82%	0.1	53.98%	2.6	99.53%
-2.3	1.07%	0.2	57.93%	2.7	99.65%
-2.2	1.39%	0.3	61.79%	2.8	99.74%
-2.1	1.79%	0.4	65.54%	2.9	99.81%
-2.0	2.28%	0.5	69.15%	3.0	99.87%
-1.9	2.87%	0.6	72.57%	3.1	99.90%
-1.8	3.59%	0.7	75.80%	3.2	99.93%
-1.7	4.46%	0.8	78.81%	3.3	99.95%
-1.6	5.48%	0.9	81.59%	3.4	99.97%
-1.5	6.68%	1.0	84.13%	3.5	99.98%
-1.4	8.08%	1.1	86.43%	4.0	99.997%
-1.3	9.68%	1.2	88.49%		
-1.2	11.51%	1.3	90.32%		

Every z-score has a corresponding percentile:

1.6 corresponds to a percentile of 94.52%
-0.8 corresponds to a percentile of 21.19%
0.4 corresponds to a percentile of 65.54%
-2.2 corresponds to a percentile of 1.39%

If you know your z-score, you know the percentile you belong to. With this table, we can begin to answer normal distribution problems.

Example 2

Suppose the IQ scores of students at CUNY are normally distributed with a mean of 105.7 and a standard deviation of 6.8. This is all we need to answer the following questions:

1) What percentage of students has an IQ below 115?

    a) First we calculate the z-score using the formula:

$$z = \frac{(\text{data value} - \text{mean})}{\text{standard deviation}}$$

$$= \frac{(115 - 105.7)}{6.8}$$

$$= \frac{9.3}{6.8}$$

$$= 1.367647059$$

$$=1.4 \text{ (rounded to nearest tenth)}$$

b) We can then look up z = 1.4 as a percentile. (See the table.) We find the corresponding percentile is **91.92%.** This means that **91.92%** of the IQ scores are below 115.

2) What percentage of students has an IQ below 95?

a) First we calculate the z-score using the formula:

$$z = \frac{(\text{data value} - \text{mean})}{\text{standard deviation}}$$

$$= \frac{(95 - 105.7)}{6.8}$$

$$= \frac{-10.7}{6.8}$$

$$= -1.573529412$$

$$=-1.6 \text{ (rounded to nearest tenth)}$$

b) We can then look up z = -1.6 as a percentile. The corresponding percentile is **5.48%.** This means that **5.48%** of the IQ scores are below 95.

**We see that we can answer percentage below problems by calculating the z-score and looking up the percentile for an answer. What about percentage above problems?

3) What percentage of students has an IQ above 100?

a) First we calculate the z-score using the formula:

$$z = \frac{(100 - 105.7)}{6.8}$$

$$= \frac{-5.7}{}$$

6.8

= -0.838235294

= -0.8 (rounded to nearest tenth)

b) We can then look up z = -0.8 as a percentile. We see 21.19%. This means that 21.19% of the scores are below 100. But we want the percentage above. If 21.19% of the values are below 100, then **100.00% - 21.19% = 78.81%** of the IQ scores are above 100.

4) What percentage of students has an IQ above 113?

a) First we calculate the z-score using the formula:

$$z = \frac{(113 - 105.7)}{6.8}$$

$$= \frac{7.3}{6.8}$$

= 1.1 (rounded to nearest tenth)

b) We can then look up z = 1.1 as a percentile. We see 86.43%. This means that 86.43% of the scores are below 113. If 86.43% of the values are below 100, then **100.00% - 86.43% = 13.57%** of the IQ scores are above 113.

**We see that we can answer percentage above problems by calculating the z-score and looking up the percentile and then subtracting it from 100% for an answer. What about percentage between problems?

5) What percentage of students has an IQ score between 90 and 110?

a) We have two values, so we need to calculate two z-scores:

For data value=90 $z = \frac{(90 - 105.7)}{6.8}$

$$= \frac{-15.7}{6.8}$$

= -2.3 (rounded to nearest tenth)

For data value=110 $z = \frac{(110 - 105.7)}{6.8}$

$$= \frac{4.3}{6.8}$$

$$= 0.6 \text{ (rounded to nearest tenth)}$$

b) We can then look up both percentiles. For z = -2.3, 1.07% of the scores are below 90; for z = 0.6, 72.57% of the scores are below 110. If 72.57% of the scores are below 110, and 1.07% of that represents scores below 90, then we can take the difference between the two. 72.57% - 1.07% = 71.50% of the IQ scores are between 90 and 110.

**We see that we can answer percentage between problems by calculating two z-scores (one for each value), looking up the two percentiles and subtracting the smaller from the larger for an answer.

Question 8: Answer the following additional questions:

a) Find the percentage of scores below 118.
b) Find the percentage of scores above 102.
c) Find the percentage of scores between 93 and 109.

Discussion of Statistical Project, Part I:

You should have gathered a data set of 50 values for the Statistical Project. Before you consider Lab #6, pair up with another student and discuss your data set.

Answer the following questions about your data set:

1) a) What is the population you are examining?

   b) What population parameter is being examined?

2) Why are you interested in this population?

3) Where did the sample come from?

4) Why did you select the 50 values for your sample (as opposed to other available values)?

5) What do you believe your sample will indicate about the population?

Lab Assignment #6 – Statistical Project, Part I
Due _____

Use the data set of the 50 values that you have gathered for your project. More details about the data set were given in class.

1) Enter the values into Excel in column A.
2) From your data set you can create a frequency table. Bin your data appropriately. Choose 6-10 equally spaced categories.
3) From the frequency table, construct a(n):
   a) histogram (bar graph)
   b) line chart
   c) pie graph
   d) another display you consider appropriate for your set

4) From the values themselves, calculate the (Consult Lab #6 for details):
   a) mean
   b) median
   c) mode
   d) range (Calculated by subtracting the lowest value from the highest value)
   e) standard deviation (Calculated using the same procedure for mean, median, mode with Excel's **STDEV** or **STDEV.S** function, in the Statistical category)
   f) five-number summary
   (This is somewhat more complicated since there are five values to be determined. We use Excel's **QUARTILE** or **QUARTILE.INC** function to find them. You can access QUARTILE with the other Statistical functions. To calculate the minimum value of the set, select the data set (A1:A50) as the **Array** and input the value 0 in the field **Quart.** To calculate the lower or first quartile, repeat the procedure but enter 1 in the field **Quart**. Repeat the procedure again, using 2, 3 and 4 to give you the median, upper or third quartile and the maximum value. See me for details if this is still confusing.)

5) Answer the following questions (this will eventually help you write your report):

1) What is your population? What is the population parameter you are examining?
2) How did you gather your sample?
3) Looking at the bar graph or line chart, what kind of data distribution do you seem to have? (Is it normal, left-skewed, right-skewed, uniform, bimodal or something else?)
4) Given your measures of sample "average," what do you think the average is for the entire population represented by your sample? What kind of data distribution do you expect from the entire population? (What I'm really asking is whether or not your sample of values is a fair representation of the entire population of values.)
5) If your data set consisted of 100 values instead of 50, would your results (mean, median, mode and graphs) be different in any way, or roughly the same?