2018

Mathematics in Contemporary Society - Chapter 2 (Spring 2018)

Patrick J. Wallach
Queensborough Community College

Chapter 2

Evaluating a Statistical Study (continued)

Here are some guidelines for evaluating any statistical study:

1) What was the goal of the study? What was the population being studied? What kind of study was it (observational, experiment, case-control)?
2) How was the sample gathered (what sampling method)? When and where was the sample gathered? Do you think the sample is a fair representation of the population?
3) What was the source of the study? Could the researcher have been biased? What does the researcher have to gain from the results of the study?
4) Consider the possibility of selection bias when the researcher chooses the sample. Consider the possibility of participation bias when the sample participants are voluntary.
5) What are the variable(s) of interest in the study? Are these variables easily measured or counted, or are they difficult to define by an exact scale?
6) If a margin of error is given, does the range of possible values lead to vastly different outcomes (such as a candidate winning or losing)?
7) Could there be confounding variables that are affecting the results of the study?
8) Were the sample members questioned fairly using objective questions that did not favor a particular answer?
9) Does the presentation of results seem fair and honest? If there are any graphs or charts (seen in the next section), are they presented fairly?
10) How do you feel about the study overall? Were the original goals of the study obtained, or is more research required? Are the conclusions sensible, meaningful, and of practical significance?

Homework:

For 1)-9), evaluate the following statements based on what you have learned about evaluating statistical studies.

1) A survey of CUNY students revealed that evening students were getting better grades than day students.
2) A recent study by Ferd Motor Company has proven that Ferd automobiles are safer than any other care on the road today.
3) Trent Harmon won the last American Idol because he was the singer that viewers of the show liked the most.
4) It has been found that students taking online courses learn more than students sitting in a classroom.
5) A recent survey of people in midtown Manhattan last Thursday revealed that 80% of New Yorkers voted for Hillary Clinton in 2016.

6) Almost all cold sufferers found that they felt better within a week after taking ZapIt Cold Medicine.
7) A survey of 100 QCC students revealed an average GPA of 3.16.
8) A random sample of 100 CUNY students from the Registrar record showed an average GPA of 2.79.
9) It has been found that people with blue eyes earn an average of $1,400 more annually that people who don't have blue eyes.

For 10)-14), explain why bias may be an issue for the following studies or situations and describe one way to reduce the possibility of bias.

10) Surveys conducted by the Democratic Party in July 2016 showed that Hillary Clinton would win the election for president in 2016 by getting 52% of the vote.
11) Since canceling the show "Evil Zombies" for the fall season, USA Networks has receiving hundreds of phone calls asking for the show to be put back on the air.
12) A survey of supermarket shoppers on a weekday morning has shown that people would pay higher taxes to increase teacher salaries and attract better teachers.
13) QBC News often runs editorials in support of the war in Iraq. QBC news is owned by Apex Inc., a defense contractor.
14) An exit poll plans to predict the outcome of the school budget vote by polling voters between 8 am and 9 am in the morning.

Statistical Tables and Graphs

Frequency Tables

Suppose a teacher gives a five point quiz to twenty-five students. The results are:

3,2,5,4,2,4,4,3,5,1,3,4,3
0,1,3,4,2,1,4,4,5,2,4,5

In this format, what we call a **data set**, we can't get a clear idea of what happened. We can summarize the results in a **frequency table**:

Quiz Score	Frequency
0	1
1	3
2	4
3	5
4	8
5	4
	25

The first column lists the **categories**, which are the six possible grades for this quiz. The second column determines the **frequency**, how many times the grade appears in the list. (Note the total of the frequency adds up to twenty five.) The frequency table is easy to read. We see that more students received 4 than any other score; also, most quiz scores (68%) are in the range of 3-5.

Suppose the same teacher gives an exam to twenty-five students. The results are:

87,95,73,96,88,51,88,99,72,82,41,77,92
82,78,84,91,83,84,77,64,68,91,88,57

In this case, it is less important to know how many students received a particular score. For instance, two students scored a 77. But that's not so important to know—all scores in the 70s are about the same. In this case, we **bin the data** into **categories** (or **classes**) that cover a range of values. For instance, we can use 40-49, 50-59, 60-69 and so on. This is what the frequency table looks like:

Test Score	Frequency
40-49	1
50-59	2
60-69	2
70-79	5
80-89	9
90-99	6
	25

Figure 1

The frequency table is now easier to read. However, individual information is lost. We know that 9 students scored in the 80s, but we don't know what the individual results are without looking at the previous list. Usually, however, we prefer the format that is easier to understand. This frequency table tells us (with just a quick glance) most scores are in the 80s (a B grade) and that 80% (20/25) of the scores are in the range 70-99.

You may wonder how to decide what categories to use. Some general pointers:

1) Categories must include all values from lowest to highest. (41 belongs in 40-49, 99 belongs in 90-99)
2) Categories should not overlap. We wouldn't use 40-50, 50-60, 60-70, etc. because certain scores (like 50 or 60) could belong in multiple categories.
3) Categories should be the same size and spaced evenly apart. The first value in each category (40, 50, 60, 70, etc.) is ten more than the previous value.
4) We usually use 6-10 categories if possible. More categories can get messy and make the frequency table harder to read. For example:

Test Score	Frequency
40-44	1
45-49	0
50-54	1
55-59	1

60-64	1
65-69	1
70-74	2
75-79	3
80-84	5
85-89	4
90-94	3
95-99	3
	25

presents the same data set in a less readable format.

On the other hand, too few categories is not descriptive enough:

Test Score	Frequency
40-59	3
60-79	7
80-99	15
	25

We lose most of the detail we had in the original set.

5) Try to avoid zero frequency categories, particularly at the end. Suppose the first student received a 9 instead of an 87 and our data set was:

9,95,73,96,88,51,88,99,72,82,44,77,92
82,78,84,91,83,84,77,64,68,91,88,57

We may be tempted to use similar categories and create the following:

Test Score	Frequency
0-9	1
10-19	0
20-29	0
30-39	0
40-49	1
50-59	2
60-69	2
70-79	5
80-89	8
90-99	6
	25

But the multiple zero categories are not visually pleasing. This is probably better:

Test Score	Frequency
Less than 40	1
40-49	1
50-59	2
60-69	2
70-79	5
80-89	8

90-99	6
	25

We can use "Less than" or "Greater than" categories at the beginning or end of a frequency table to make it look better.

Recall the previous data set of test scores:

87,95,73,96,88,51,88,99,72,82,41,77,92
82,78,84,91,83,84,77,64,68,91,88,57

We saw these could be arranged in a frequency table as follows:

Test Score	Frequency
40-49	1
50-59	2
60-69	2
70-79	5
80-89	9
90-99	6
	25

Figure 1

Relative Frequency and Cumulative Frequency

Taking the original frequency table *(Figure 1)*, we can add columns for **relative frequency** and **cumulative frequency**:

Test Score	Frequency	Relative Frequency	Cumulative Frequency
40-49	1	1/25 = 0.04 = 4%	**1**
50-59	2	2/25 = 0.08 = 8%	1+2=**3**
60-69	2	2/25 = 0.08 = 8%	1+2+2=**5**
70-79	5	5/25 = 0.2 = 20%	1+2+2+5=**10**
80-89	9	9/25 = 0.36 = 36%	1+2+2+5+9=**19**
90-99	6	6/25 = 0.24 = 24%	1+2+2+5+9+6=**25**
	25	25/25 = 1 = 100%	

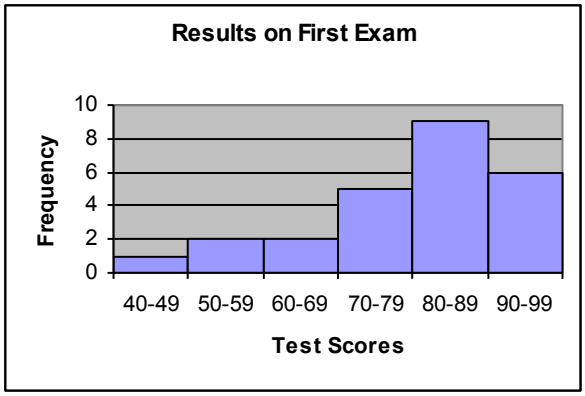
Relative frequency is obtained for each category by taking the frequency of the category and dividing by the total frequency (in this case, 25). It is usually expressed as a decimal or percentage, not a fraction.

Cumulative frequency is obtained by adding the frequencies above and included in the category. For the first row, a cumulative frequency of 1 indicates that 1 score is below 50. For the second row, a cumulative frequency of 3 indicates that 3 scores are below 60. For the third row, a

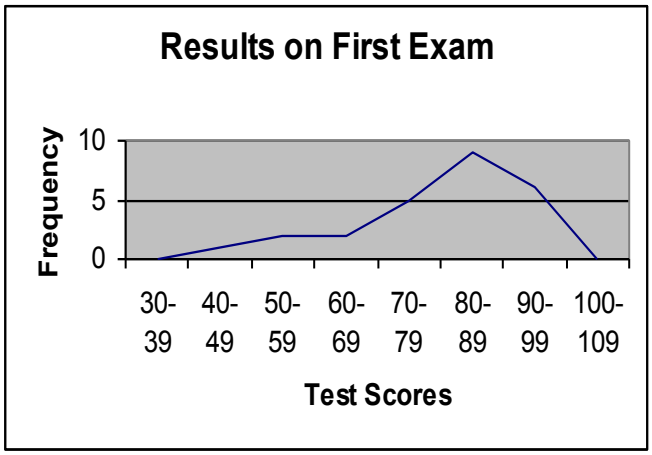
cumulative frequency of 5 indicates that 5 score are below 70. (And so on.) The last cumulative frequency equals the total frequency, since all scores are below 25.

Other Graphic Displays

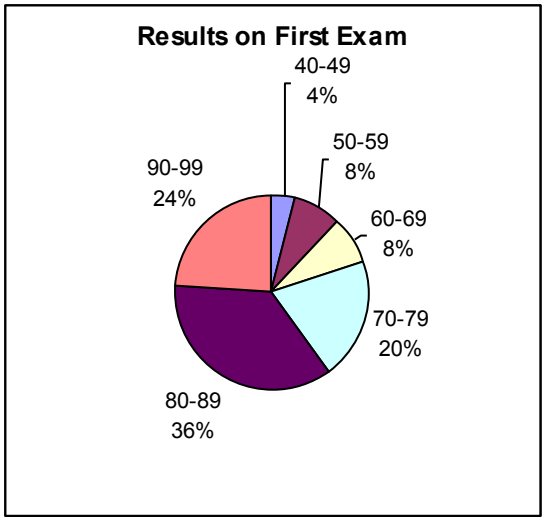
We can use the original frequency table (*Figure 1*) to generate graphic displays to the data set. Some of the more popular displays are:



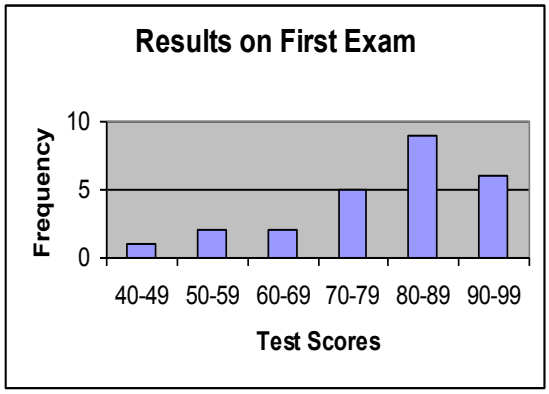
The histogram represents each frequency with the height of a bar. In a histogram there are no spaces between the bars.



The line chart represents each frequency with a point on a line.



The pie chart represents each frequency with a slice of a pie.



The bar graph is like a histogram with spaces between the bars.

We notice here that the histogram, line chart and bar graph all give us related pictures, with frequency represented by height of bars or lines. The pie chart is obviously different, with frequency represented as the size of a slice in a pie. Percentages give us better sense of how big each slice and category actually is.

Graphic displays tell a story about the data set visually. We can easily determine the highest frequency and spot general trends. We can ask some general questions for any set of graphs:

Is there a peak in the graphs (look at the histogram, line chart and bar graph)? Where is it located (in the center, to the right, to the left)?
The peak here is located in the category 80-89, to the right of the graphs. This indicates a higher percentage of high scores than low scores on the exam. The pie chart tells us that 36% of the scores are in the range 80-89.

Is there more than one peak, or no peak at all?
There is only one peak in the graphs. We shall see later that we call this kind of result ***unimodal****.*

Is the graph symmetric in shape (it can be cut down the middle into two equal halves)? *No, these graphs are not really symmetric.*

Some additional notes on making proper graphs:

1) Label everything! The horizontal axis of the histogram, line chart and bar graph should identify the data set that is being represented. The vertical axis usually represents the frequency of the different categories.
2) All graphs should receive a title to further describe the data set.
3) Use proper scaling; this means that the spacing of frequency and category is equal and consistent. This will be easy in Excel, where it is done automatically for us. We need to be more careful when we create graphs by hand.

Homework:

Given the following 40 football player weights:

180 197 193 185 189 284 308 252 176 223 272 183 251 225
168 294 248 215 265 282 183 294 207 239 282 299 189
286 115 196 285 197 338 245 289 245 188 198 344 201

Construct (by hand) a:
a) frequency table with relative frequency and cumulative frequency included

(Note: The frequency table part (a) can be done in a Word file by just using tabs between the columns. It can also be entered into Excel if you find that easier.)

Quantitative and Qualitative Data

Quantitative data is data that can be measured or counted, such as height, weight, age, GPA, family size, income, etc. Quantitative data can always be ordered from smallest value to largest value. It can therefore be binned into categories from lowest values to highest values.

Quantitative data is often displayed in histograms and line charts. The lowest values are always on the left and the higher values are always on the right.

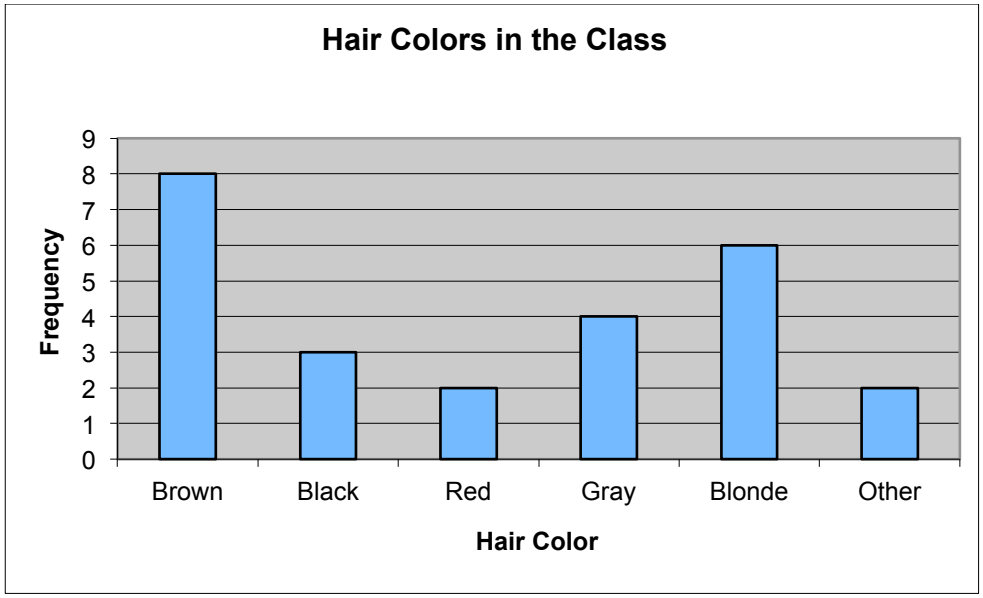
Qualitative data is data that can be divided into different categories, such as hair color (black, brown, red, etc.) or marital status (single, married, divorced, etc.) or favorite ice cream flavor (vanilla, chocolate, strawberry, etc.). Qualitative data cannot be put in any "order" because there is no lowest value or highest value. We often use bar graphs and pie charts with qualitative data.

Question 1: Give three examples of quantitative data. Give three examples of qualitative data.

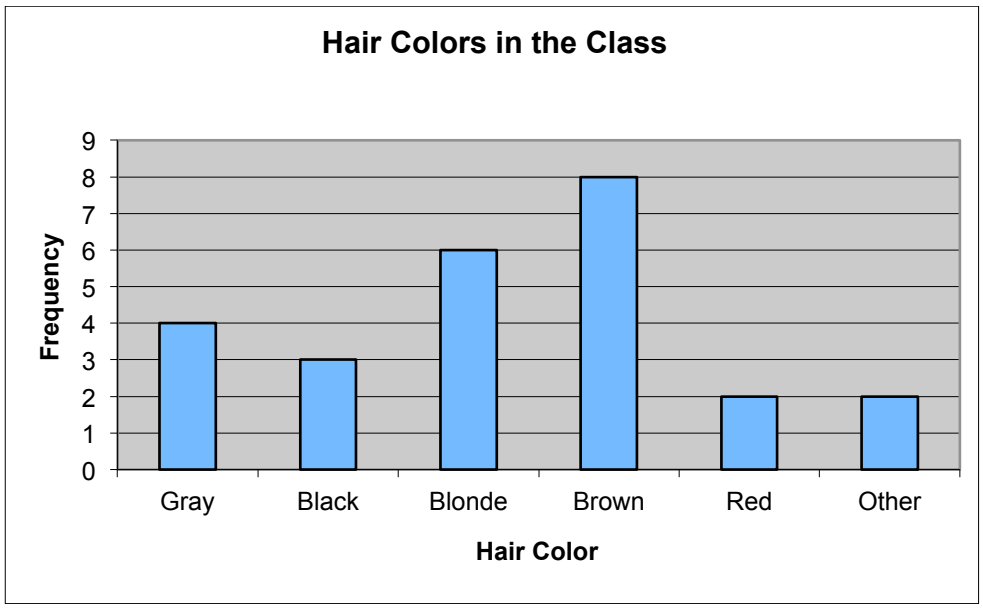
Suppose we had the following frequency table:

Hair Color	Frequency
Brown	8
Black	3
Red	2
Gray	4
Blonde	6
Other	2
	25

The order of the colors (first, second, third…) really doesn't matter. We could create a bar graph like this:



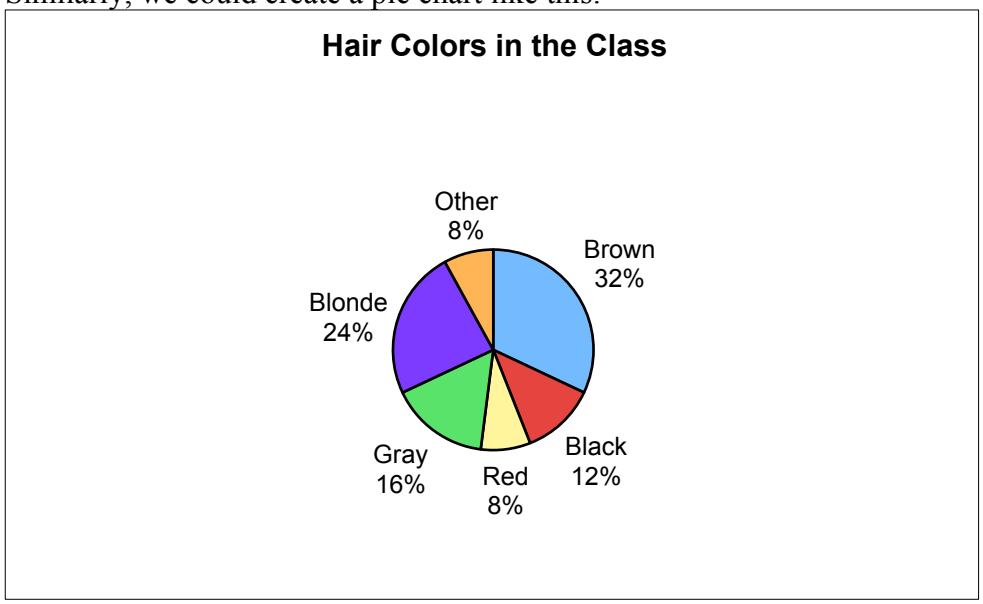
or this:



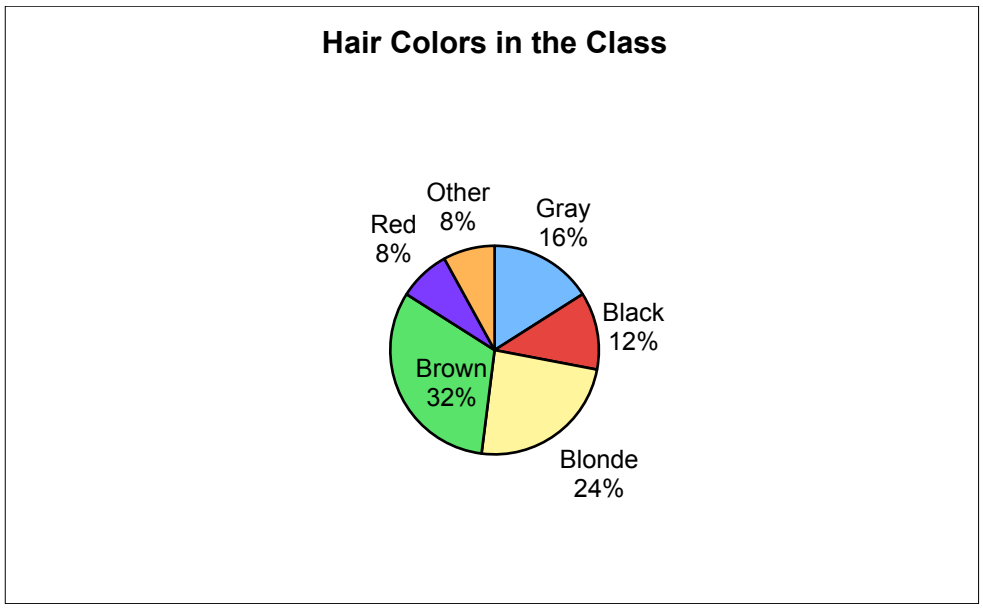
These are all the same really because the height of each corresponding color is the same. The location of the Brown bar could be on the left, right or middle.

Question 2: Does it matter how the bars are arranged in these cases? How would you arrange them?

Similarly, we could create a pie chart like this:



or this:



As long as the slices remain the same size, location is unimportant.

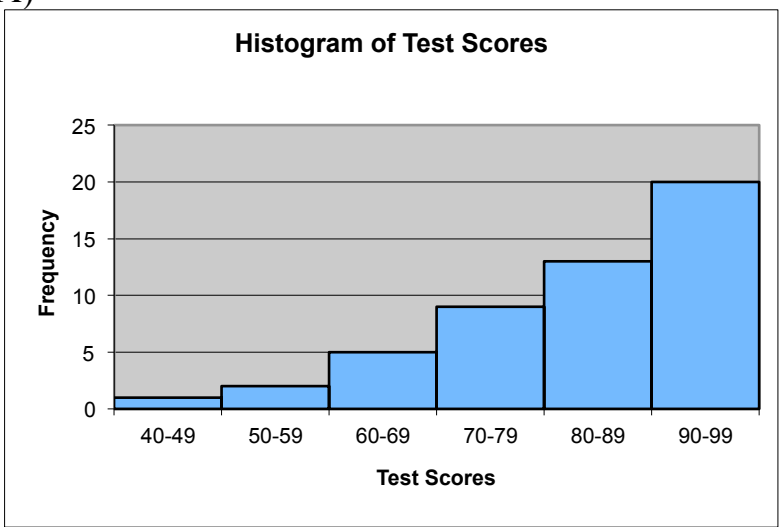
Question 3: Does it matter how the slices are arranged in these cases? How would you arrange them?

We will explore continue to explore graphs in upcoming lessons. Stay tuned!

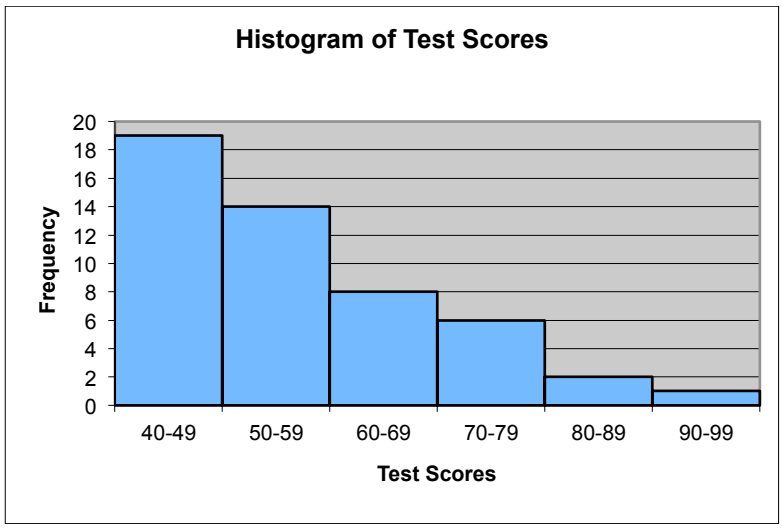
Question 4: Suppose a teacher gives a test to 50 students. From the data set, she creates a frequency table and then a histogram. The histogram tells a story about the test to whoever looks at it.

For the following six histograms, write two statements (12 total) about what each histogram says to you. *(Don't look ahead—use your own ideas!)*

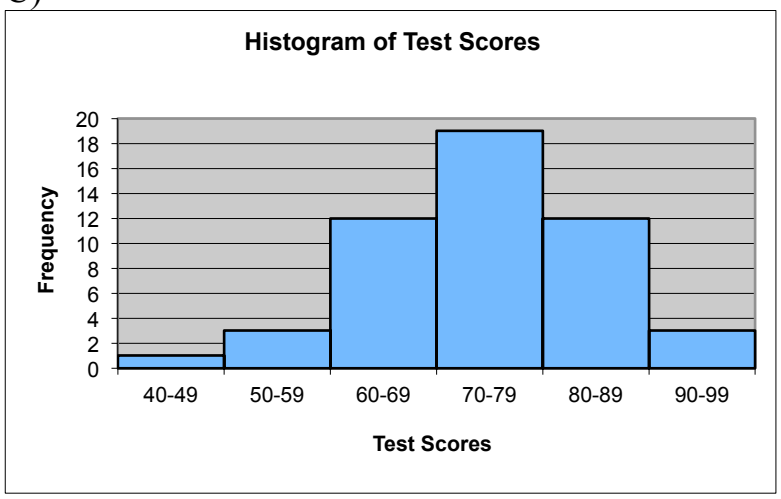
A)



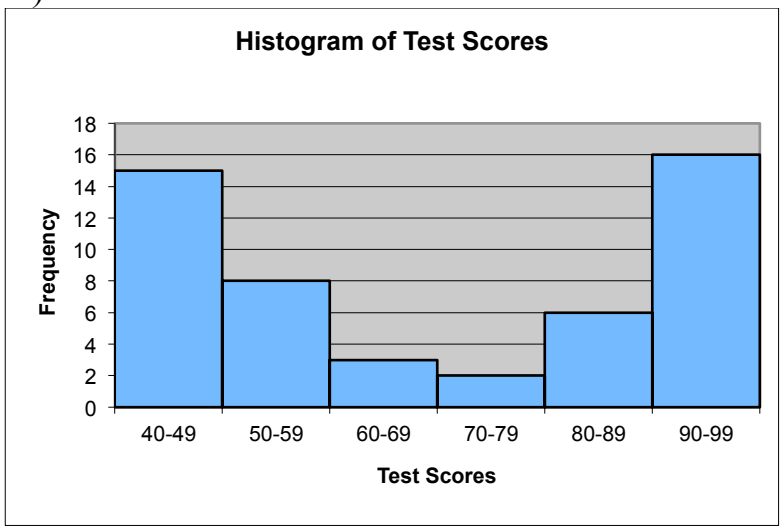
B)


Histogram of Test Scores
Frequency
20
18
16
14
12
10
8
6
4
2
0
40-49
50-59
60-69
70-79
80-89
90-99
Test Scores

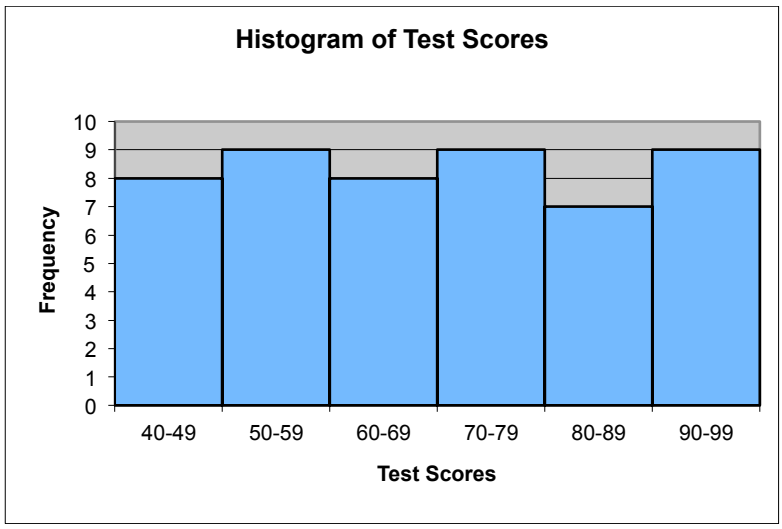
C)


Histogram of Test Scores
Frequency
20
18
16
14
12
10
8
6
4
2
0
40-49
50-59
60-69
70-79
80-89
90-99
Test Scores

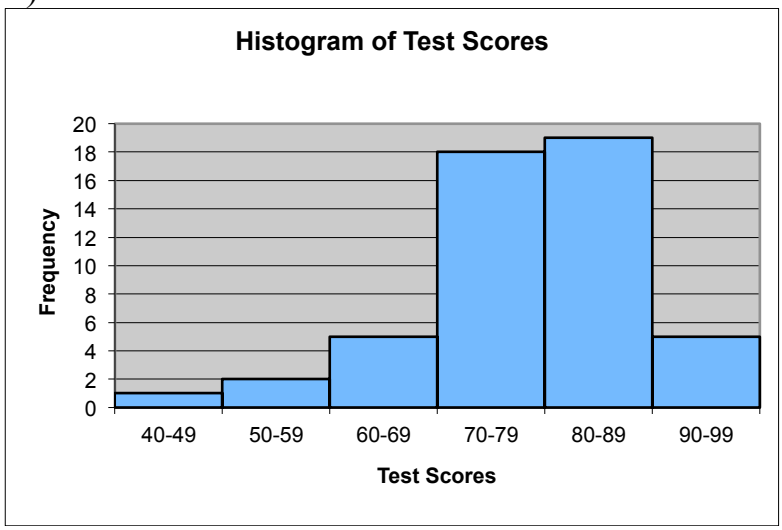
D)


Histogram of Test Scores
Frequency
18
16
14
12
10
8
6
4
2
0
40-49
50-59
60-69
70-79
80-89
90-99
Test Scores

E)



F)



Let's think about your responses:

A) Some of you might have offered statements like:

"Most people scored in the 90s."
"Only a few people failed."
"The highest part of the graph is for scores of 90-99"

These are all observation statements. **Observation statements** (made properly) cannot really be disputed; they are statements that everyone could agree upon.

Some of you may have given statements such as:

"The test was too easy."

"Everyone studied for the test."
"It must have been an open book test."

These are all conclusion statements. **Conclusion statements** analyze the reason behind a particular result and attempt to explain it. Conclusion statements may be true or false; they should be backed up with supporting information.

Question 5: Look back at your 12 responses. Next to each one, put the word "Observation" or "Conclusion" based on what kind of statement you believe was made.

Graph A) has certain characteristics: The peak is on the right side of the graph; most values in the data set are on the right side of the graph; there is a "tail" on the left side for the few results that were lower scores.

We call this kind of graph a **left-skewed distribution**. Graphically, a left-skewed distribution peaks on the right side and trails off (with a tail) on the left side.

B) Some of you might have offered statements like:

"Most people scored in the 40s."
"More than half the class failed with a score below 60."
"The lowest part of the graph is for scores of 90-99"

These are all observation statements.

Some of you may have given statements such as:

"The test was too hard."
"The professor is too difficult."
"Nobody studied for the test."

These are all conclusion statements.

Graph B) has certain characteristics: The peak is on the left side of the graph; most values in the data set are on the left side of the graph; there is a "tail" on the right side for the few results that were higher scores.

We call this kind of graph a **right-skewed distribution**. Graphically, a left-skewed distribution peaks on the left side and trails off (with a tail) on the right side.

C) Some of you might have offered statements like:

"Most people scored in the 70s."
"Most students passed."
"The lowest part of the graph is for scores of 40-49"

These are all observation statements.

Some of you may have given statements such as:

"The test was fair."
"The professor gave an appropriately challenging test."
"This looks like a typical result."

These are all conclusion statements.

Graph C) has certain characteristics: The peak is in the center of the graph; other results are spread evenly on both sides of the peak; there are "tails" on both sides for low scores and high scores. It seems like a "normal" or "average" result.

We call this kind of graph a **normal distribution**. Graphically, a normal distribution is a bell-shaped curve that peaks in the middle of the graph and trails off evenly on both sides. We shall see that the normal distribution is the most common and expected result with data sets of various kinds of measured data (heights, weights, ages, IQ scores, GPA, etc.)

D) Some of you might have offered statements like:

"The highest bar is for scores in the 90s."
"Just about the same number of students scored in the 40s and 90s."
"The lowest part of the graph is for scores of 70-79"

These are all observation statements.

Some of you may have given statements such as:

"The test was completely unfair."
"Half of the students studied and half didn't."
"Part of the class should register for a more challenging course."

These are all conclusion statements.

Graph D) has certain characteristics: There are two peaks (they are almost the same frequency) at different parts of the graph; the frequencies are low between the peaks. It seems like a "split" or "divided" result.

We call this kind of graph a **bimodal distribution**. Graphically, a bimodal distribution has two separate, distinct peaks with usually very little between the peaks. A bimodal distribution usually indicates some division within the data set that tends toward two separate results.

E) Some of you might have offered statements like:

"All categories have about the same frequency."
"Nine students scored in the 50s."
"Half of the class scored below 70."

These are all observation statements.

Some of you may have given statements such as:

"The test was completely unfair."
"The professor designed a test that only half the class could do well on."
"The test was graded in an unfair way."

These are all conclusion statements.

Graph E) has certain characteristics: All categories have roughly the same frequency; there is no peak and no "valley" either.

We call this kind of graph a **uniform distribution**. Graphically, a uniform distribution has roughly the same frequency for all categories, giving it a flat look on

the top of the graph. A uniform distribution usually indicates little difference between the different categories of the data set.

F) Some of you might have offered statements like:

"The graph has one peak."
"Most students scored in the 70s and 80s."
"Few students failed."

These are all observation statements.

Some of you may have given statements such as:

"The test was fair."
"It was hard to do very well on this test."
"This is an expected result."

These are all conclusion statements.

Graph F) may seem to have characteristics of different distributions. It some ways it might look like one result, in some way another.

Some may feel that it is a left-skewed distribution, because it seems that more of the values (the highest bars) are on the right side of the graph. Others may feel it almost looks like a normal distribution, because it is somewhat bell-shaped.

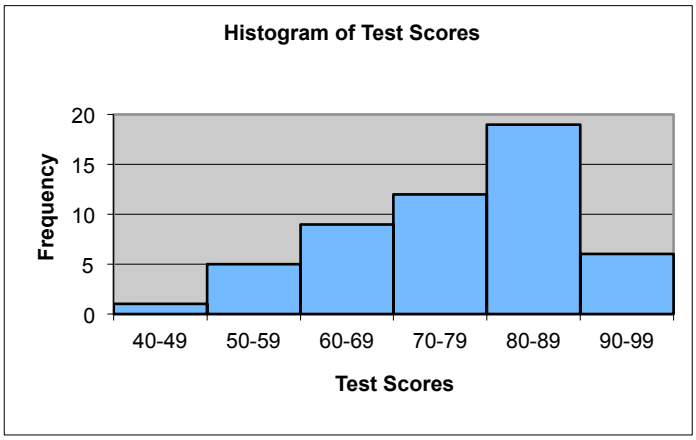
Now that we know more about graphs and the different kinds of results that we can obtain, we can do a couple of additional questions.

Question 6: Choose one of the previous graphs (A-F) and write a short story (about 8-10 sentences) about what happened in the classroom. Your story should help explain why the graph looks the way it does.

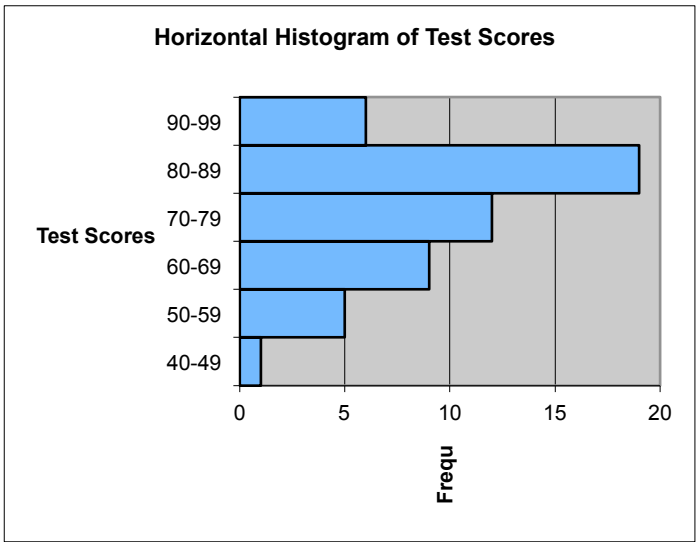
Question 7: Examine the following graphs (1-5), created from the same frequency table. Then answer the questions at the end

Examine the following graphs (1-5), created from the same frequency table. Then consider the questions at the end.

1)

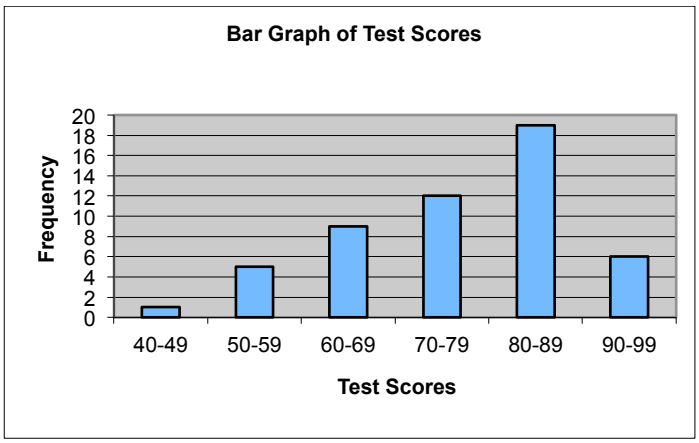


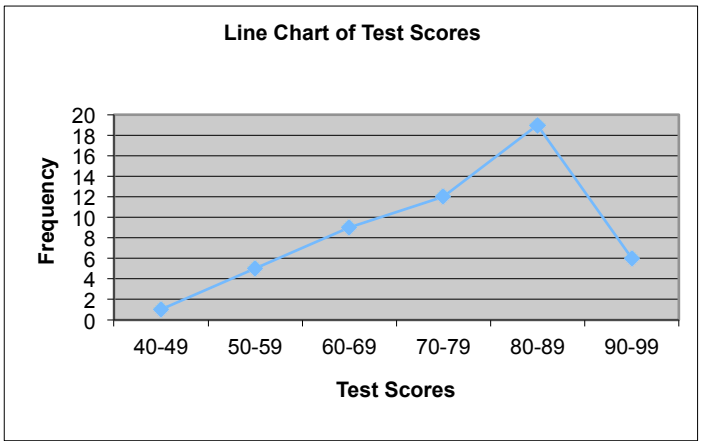
2)



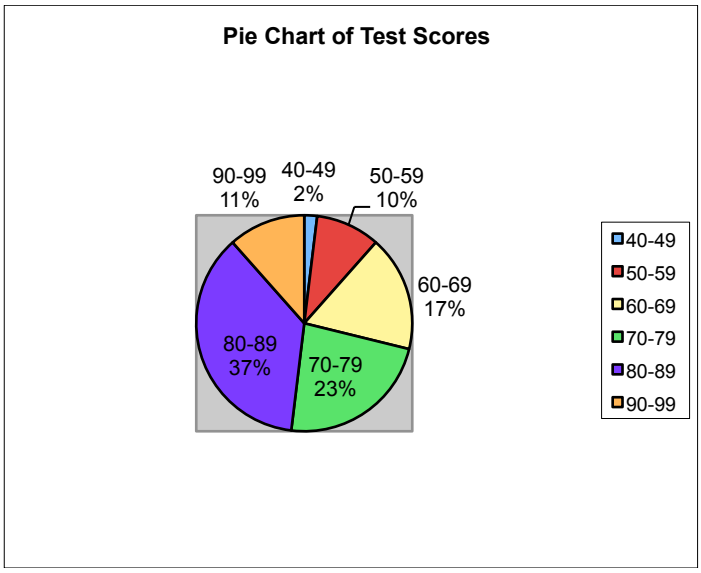
3)

4)


Bar Graph of Test Scores
Frequency
20
18
16
14
12
10
8
6
4
2
0
40-49
50-59
60-69
70-79
80-89
90-99
Test Scores


Line Chart of Test Scores
Frequency
20
18
16
14
12
10
8
6
4
2
0
40-49
50-59
60-69
70-79
80-89
90-99
Test Scores

5)


Pie Chart of Test Scores
90-99
11%
40-49
2%
50-59
10%
60-69
17%
70-79
23%
80-89
37%
40-49
50-59
60-69
70-79
80-89
90-99

Which do you feel is the best representation of the data set? Why?

Which do you feel is the poorest representation of the data set? Why?

Lab Assignment #2— Frequency Tables and Graphs

Due _____

Create a simple frequency table in Excel, from an imaginary sample of fifty student IQ scores. For example:

IQ Score	Frequency
80-89	3
90-99	9
100-109	19
110-119	9
120-129	7
130-139	3
	50

(You should make up your own categories and use your own frequencies.)

*We can use a frequency table to generate graphs in Excel. The **Charts** window makes it very easy.*

To create a bar graph, use the following instructions:

1) Use the mouse to highlight the entire frequency table, except the total on the last line.
2) Click on the **Insert** tab to change the window on top of Excel.
3) Go to the **Charts** window.
4) **Step 1:** Click on the **Insert Column Chart** icon to get a column style bar graph.
5) **Step 2:** Choose from the pictures of various styles of bar graphs. Select a simple style (like **Clustered Column** under the **2-D Column** set of options). The bar graph is immediately generated.

6) **Step 3:** You will also see that you have been shifted to **CHART TOOLS** which you can use to personalize the graph.

a) Under **Design** you can select a style for your bars and background. Click on **Change Colors** to change the color of your bars.

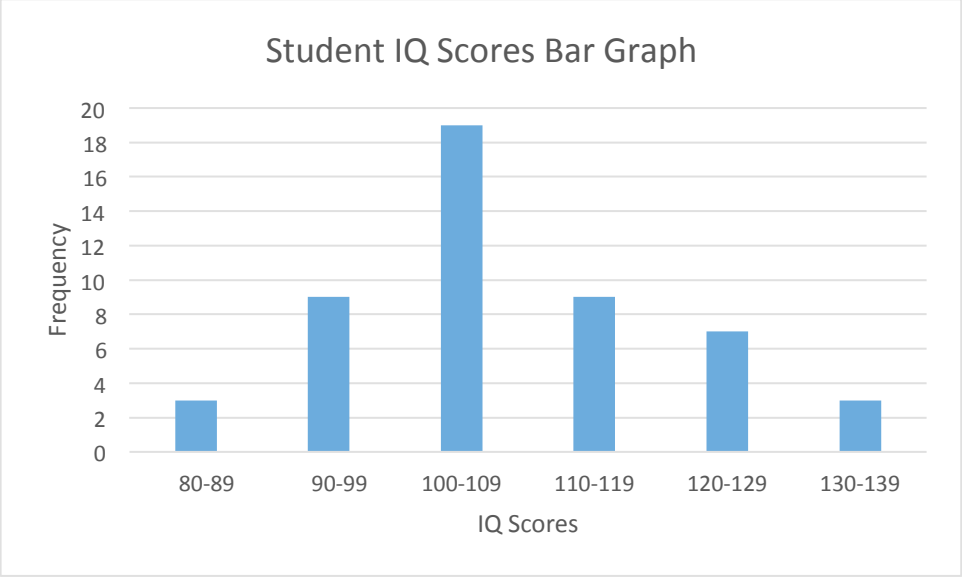
b) If there is already a title on top of the graph ("Frequency" may appear) click on it and change it to a better title, such as "Student IQ Scores Bar Graph." (This can be done in the text box itself or Formula Bar above.) If not, click on **+ (Chart Elements)** and click on **Chart Title** to place one.

c) You can click on **+ (Chart Elements)** and **Axis Titles** to label your x-axis and y-axis by selecting **Primary Horizontal** and **Primary Vertical Axis** titles and changing the horizontal x-axis label to something like "IQ Scores." For the vertical y-axis, change the label to "Frequency."

d) Examine the other options in **Chart Elements** to make further changes in your bar graph. If you don't like something that you changed, change it back. You can also change **Chart Styles** and **Chart Filters** as well (but keep it simple the first time).

e) You can revisit **Design** and **Move Chart Location** to choose to place the graph as an **Object in** Sheet 1 (what it does automatically) or as a **New Sheet**.

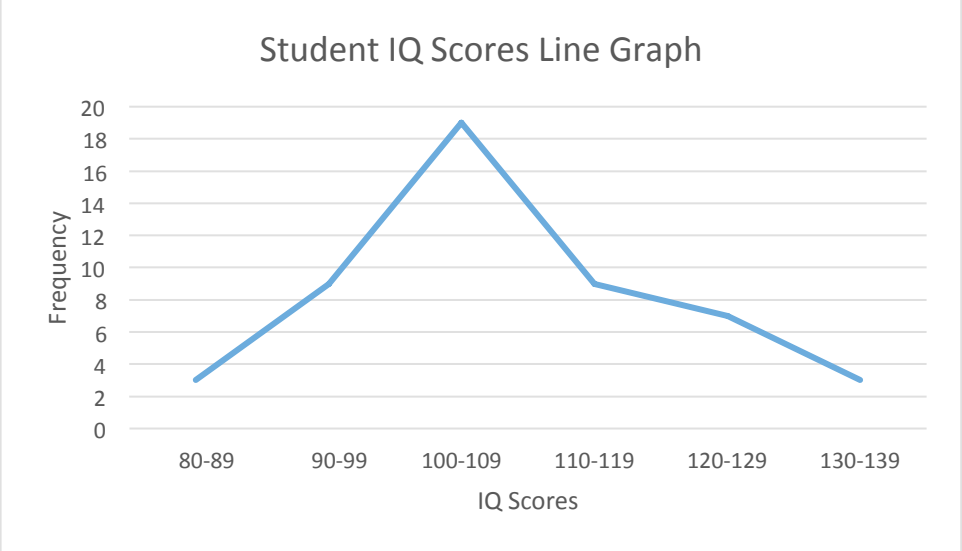
Your result should look something like this:



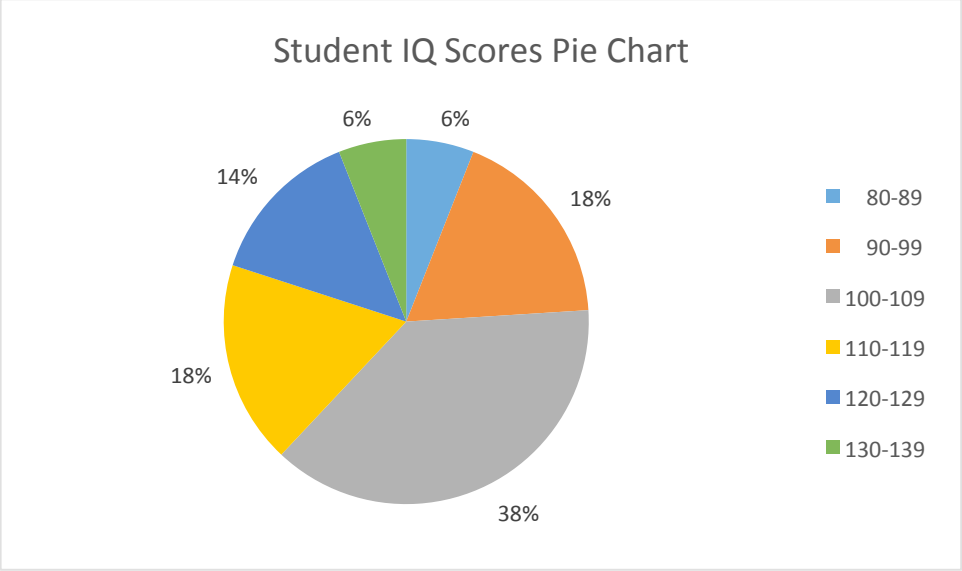
You can now create other graphs by reselecting your frequency table and selecting other options.

From the frequency table, create **three more graphs**:

a) line graph



b) pie chart



c) another **appropriate** graphical display
(There should be a total of 4 graphs.)

Answer the following questions, based on your results:

Which graph is the best representation of your frequency table? Why? Which graph is the worst representation of your frequency table? Why?