

City University of New York (CUNY)

CUNY Academic Works

Open Educational Resources

Queensborough Community College

2015

Mathematics in Contemporary Society Chapter 5

Patrick J. Wallach

Queensborough Community College

[How does access to this work benefit you? Let us know!](#)

More information about this work at: https://academicworks.cuny.edu/qb_oers/7

Discover additional works at: <https://academicworks.cuny.edu>

This work is made publicly available by the City University of New York (CUNY).

Contact: AcademicWorks@cuny.edu

Chapter 5

We have an exam next week. Since we will be reviewing for the test, there is only one hour of new material, which reinforces and builds upon what you already learned.

Test Topics for Exam #1

There will be writing on the exam!!!

The test will consist of 6-8 short answer questions and 3 or 4 problems.

Topics include:

Population, sample, population parameters, sample statistics

Observational study

Sampling techniques

Experiment: treatment group, control group, placebo

Case-control study

Evaluating a statistical study

Selection bias, participation bias

Creating or reading a: frequency table (with relative and cumulative frequency), histogram, pie chart, line chart, time-series diagram

Quantitative and Qualitative data

Using a multiple bar graph, stack plot, geographical display

Graphic distortion

Positive correlation, negative correlation, no correlation

Creating a scatter plot or scatter diagram, identifying correlation

Causality

Calculating mean, median, mode

Outliers

Describing distributions by peaks, symmetry, skewness and variation

Identifying the shape of a distribution (normal, left-skewed, right-skewed, uniform, bimodal, etc.)

Blackboard Assignment!

(if Blackboard is being used for the class)

To help you study for the test, everyone should participate in Blackboard's Discussion Board. The instructions are very simple:

- a) Post two **separate** review questions for the exam as a message.
- b) Answer two other review questions posted by someone else. (You can add to someone else's answer if you make a **meaningful** contribution to an incorrect or incomplete answer.)

(Some of the following information was presented in the notes previously—you should remember it. We have new information to consider with it.)

Shapes of a Distribution

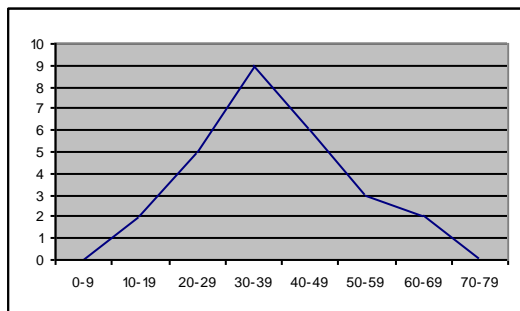
We often refer to the graph of a data set as a **distribution** of the data set. The overall shape of the distribution is often of interest to the researcher.

Now that we know about the mean, median and mode, we want to reexamine the different kinds of graphs we can obtain from a data set.

How Many Peaks?

The most immediately noticeable quality of a distribution is the number of peaks that it has. There are different possibilities:

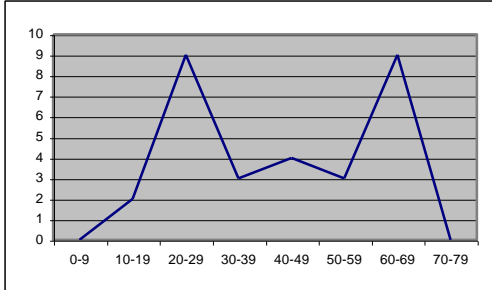
Example 1:



A distribution may have one peak or be a **unimodal distribution**. We say that the peak of the graph represents the mode of the distribution, the value(s) that occur(s) the most frequently. It makes sense, because the peak represents the value(s) with the highest frequency. Highest frequency is what the mode is all about.

Question 1: Identify a data set that could be a unimodal distribution.

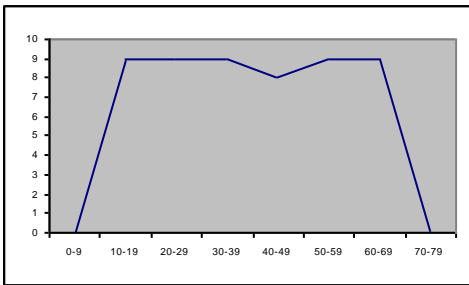
Example 2:



A distribution may have two peaks and be a **bimodal distribution**. In this case, there are two separate modes that occur with roughly equal frequency.

Question 2: Identify a data set that could be a bimodal distribution. What makes it bimodal?

Example 3:

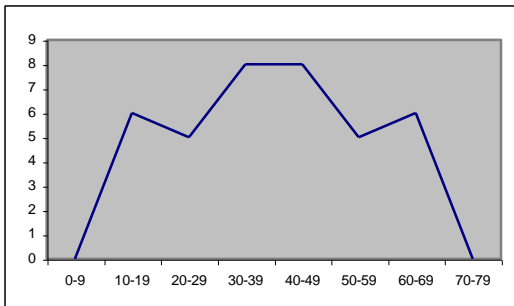


A distribution may have no real peak and be a **uniform distribution**. Generally, we say there is no mode here—all values have roughly the same frequency.

Question 3: Identify a data set that could be a uniform distribution.

Does it Have Symmetry?

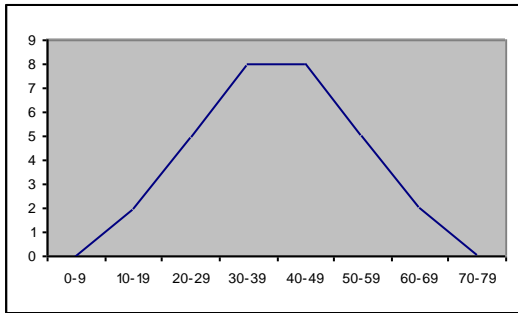
Example 4:



A distribution that can be divided down the middle into two halves that are mirror images of the other is said to be **symmetric**. A symmetric graph shows that the data set is balanced evenly. The median of a symmetric data set is in the center of the graph, because you have can divide the graph into two equal halves down the middle.

A symmetric graph has the mean and median at the center, but not necessarily the mode.

Example 5:

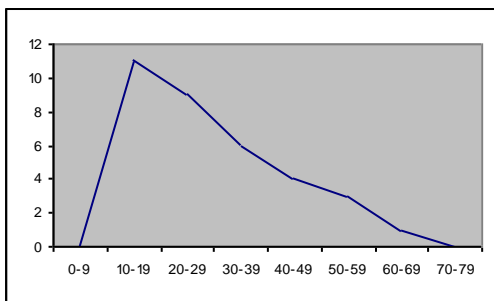


The **normal distribution** has a bell-shaped appearance and is symmetric. In the normal distribution, we say that mean=median=mode. The average of all values is at the center of the graph. The middle term is at the center of the graph. The value that occurs with the most frequency is at the center of the graph.

Question 4: Identify a data set that should be a normal distribution.

Is it a Skewed Graph?

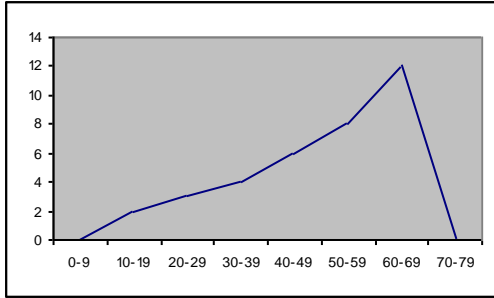
Example 6:



A **right-skewed distribution** has a peak on the left side and a “tail” on the right. The outliers (values that are unusually high in comparison to most of the data) of the distribution are on the right side. We say that mode<median<mean. The mode is the smallest value, located at the peak of the graph. The median, the middle term, is higher than the mode. (Remember, the median has to divide the set into 2 equal halves—most of the values are low here.) The mean is the largest value because outliers raise the value of the numerical average.

Question 5: Identify a data set that could be a right-skewed distribution.

Example 7:



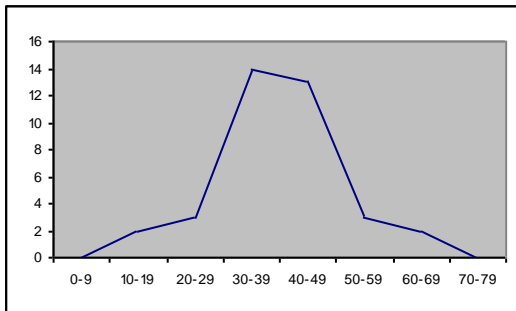
A **left-skewed distribution** has a peak on the right side and a “tail” on the left. The outliers (values that are unusually low in comparison to most of the data) of the distribution are on the left side. We say that $\text{mean} < \text{median} < \text{mode}$. The mode is the highest value, located at the peak of the graph. The median, the middle term, is lower than the mode. (Remember, the median has to divide the set into 2 equal halves—most of the values are high here.) The mean is the smallest value because outliers lower the value of the numerical average.

Question 6: Identify a data set that could be a left-skewed distribution.

Variation

Variation measures how closely or how far apart the values of a distribution are overall:

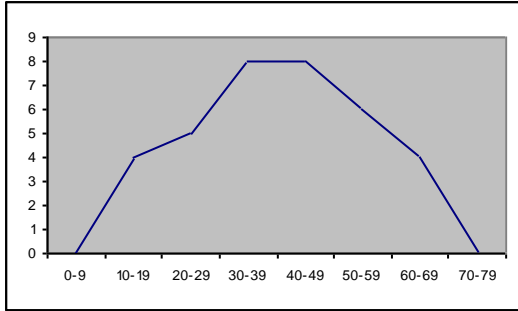
Example 8:



If most values are toward the center, with low frequency elsewhere, the distribution has **low variation**.

Question 7: Identify a data set that could have low variation.

Example 9:



If the values are spread out (generally high frequencies for most categories), the distribution has **high variation**.

Question 8: Identify a data set that could have high variation.

More on Shapes of a Distribution

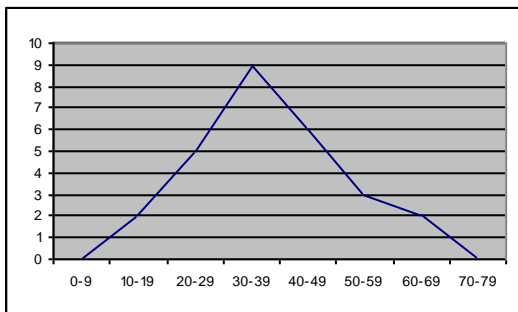
What makes a distribution look the way it does? The values of the data set are distributed in such a way to give the distribution a particular appearance. This appearance tells us much information about the nature of the data set itself.

Today's notes try to give further explanations about the results we may obtain when gathering our own data sets and creating our graphs from them. *(Some of the notes may be repetitive to reinforce the points previously given.)*

Peaks in a Distribution

A distribution may have one peak or be a **unimodal distribution**.

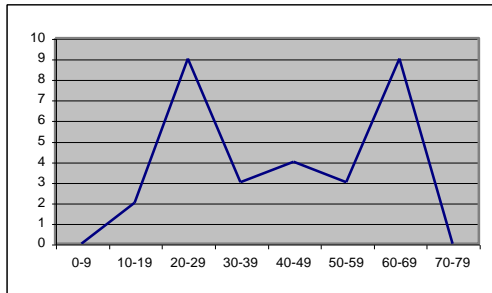
Example 1:



Without knowing what data set we're dealing with here, we can certainly say that there are more values in the category 30-39 than any other category. The mode is a little different here; I don't know which specific value occurs the most often (in fact, the mode of the actual data set may be higher than 39 or lower than 30). Graphically, the peak of the graph represents the mode of the distribution.

We may expect that most data sets will have unimodal distributions. The peak of the graph identifies which category of values best describes the data set.

Example 2:

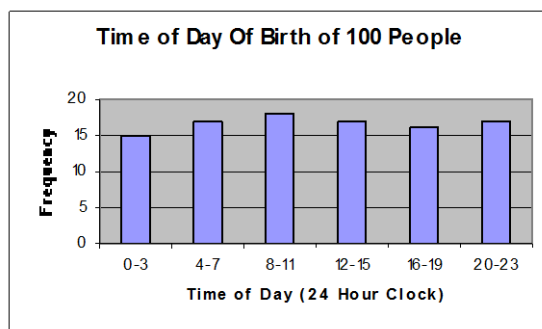


A distribution with two peaks is a **bimodal distribution**. In this case, there are two separate modes that occur with roughly equal frequency.

When a bimodal distribution occurs, it suggests that the data set is divided in such a way that there are two subgroups within the data set, each with its own mode.

If the above example was a set of test scores, it tells us that the data set is divided in such a way as to produce separate results. Perhaps the class in question was made up of different kinds of students (maybe half of them came into the course by passing a placement exam and the other half came from a prerequisite course). Maybe a surprise test was given and half the class was prepared and the other half was not. A bimodal distribution is certainly not an expected result.

Example 3:



A distribution may have no real peak and be a **uniform distribution**. Generally, we say there is no mode here—all values have roughly the same frequency.

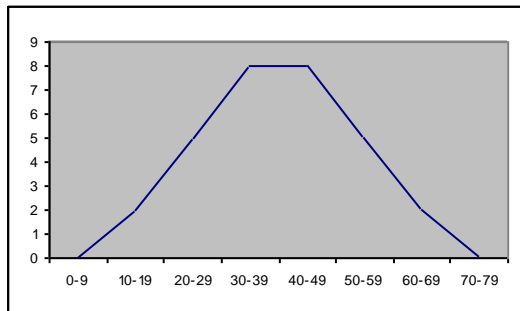
For the above bar graph, 100 people were asked what time of day they were born. 0-3 represents 12:00 AM to 3:59 AM, 4-7 represents 4:00 AM to 7:59 AM, and so on.

In this situation, we really don't expect there to be a time slot that is more frequent than any other time slot. The time of day that we are born is random; there shouldn't be any one time that occurs more often than any other. A person has just as much chance as being born between 8:00 AM and 11:59 AM as 4:00 PM to 7:59 PM.

When a data set has a uniform distribution, it suggests there are a set of values with no overall defining characteristics. While the mean and median may be right in the middle of these values, the lack of a mode tells us that one value is just as good as any other.

Symmetry

Example 4:



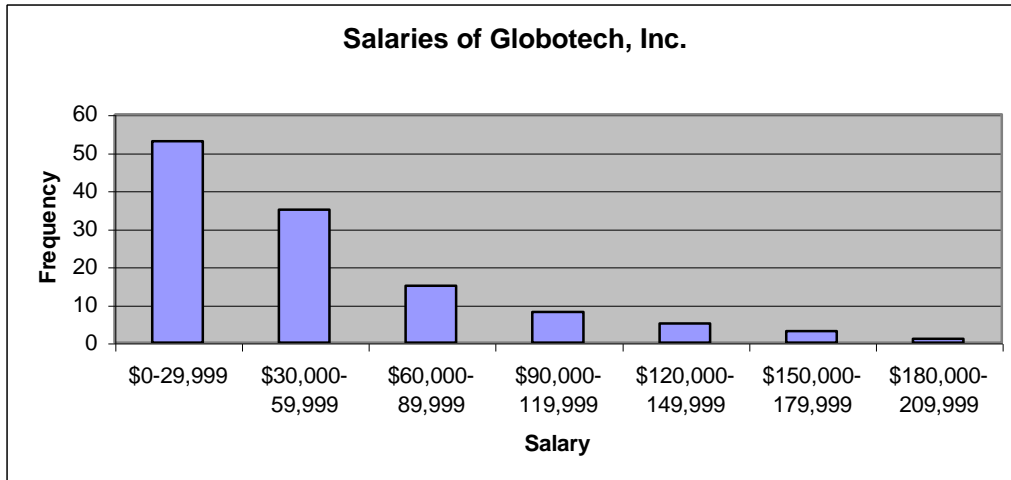
A distribution that can be divided down the middle into two halves that are mirror images of the other is said to be **symmetric**. A symmetric graph shows that the data set is balanced. With the median in the middle, there are just as many values above it as there are below, distributed in like fashion.

In terms of symmetric graphs, we are most interested in the normal distribution. The **normal distribution** has a bell-shaped appearance and is symmetric. In the normal distribution, we say that mean=median=mode. The average of all values is at the center of the graph. The middle term is at the center of the graph. The value that occurs with the most frequency is at the center of the graph.

Many data sets in the real world (height, weight, IQ scores, GPA, shoe size, SAT scores, automobile mileage, cost of shipping a package, etc.) correspond to the normal distribution. When examining a data set, we expect to obtain a normal distribution unless there is a significant reason (some have already been identified) for a different result.

Skewness

Example 5:



A **right-skewed distribution** has a peak on the left side and a “tail” on the right. The outliers (values that are unusually high in comparison to most of the data) of the distribution are on the right side. We say that $\text{mode} < \text{median} < \text{mean}$. The mode is the smallest value, located at the peak of the graph. The median, the middle term, is higher than the mode. (Remember, the median has to divide the set into 2 equal halves—most of the values are low here.) The mean is the largest value because outliers raise the value of the numerical average.

The above example describes the salaries of 120 employees of Globotech, Inc. As you can see, 75% of the employees earn less than \$60,000. But there are a few who make more. We can imagine that the employees earning more than \$100,000 represent the executives of the corporation, with the one employee (probably the president) making over \$180,000.

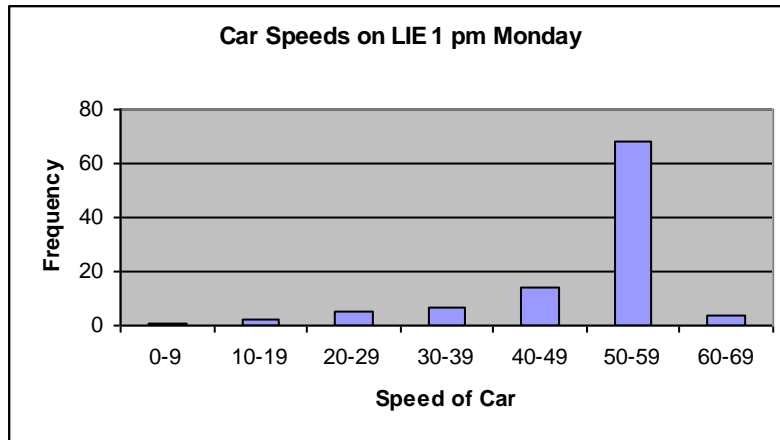
Is salary distributed evenly here? Obviously not.

If you were told that the mean salary of the company was \$55,000, you would argue that considerably more than half of the employees make less than \$55,000. The high executive salaries cause the mean to be much higher than the 53 employees earning salaries below \$30,000. The mean is not the most useful measure of average here.

If the mean is less useful, we could consider the median salary of Globotech, Inc. If the median salary is \$33,500 that tells us that 50% of the employees earn less than \$33,500 and 50% earn more than \$33,500. The median gives the middle salary here. Still, the median does not give much other detail.

Perhaps the mode category of \$0 - \$29,999 is the most useful piece of information. Almost half of the employees of Globotech make less than \$30,000. This may be the most honest representation of “average” employee salary.

Example 6:



A **left-skewed distribution** has a peak on the right side and a “tail” on the left. The outliers (values that are unusually low in comparison to most of the data) of the distribution are on the left side. We say that $\text{mean} < \text{median} < \text{mode}$. The mode is the highest value, located at the peak of the graph. The median, the middle term, is lower than the mode. (Remember, the median has to divide the set into 2 equal halves—most of the values are high here.) The mean is the smallest value because outliers lower the value of the numerical average.

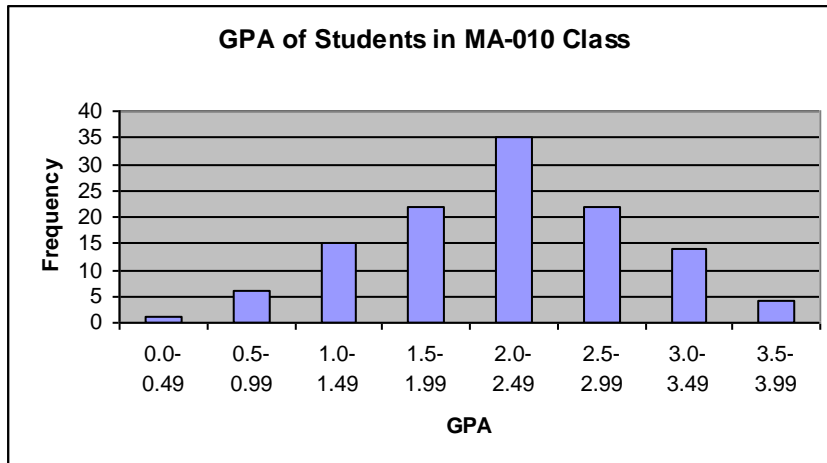
In the above example, we see that most of the cars are traveling at a speed of 50-59 miles per hour (Most people obey the speed limit, right?). However, there are a few slower moving vehicles. Some people prefer lower speeds; there may be construction vehicles; some cars may have their flashers on, etc.

In this case, the mean speed may end up being less than 50-59 mph. But we can clearly see that most drivers are traveling near the speed limit of 55 mph. The mode or median may be more useful measures of “average” speed in this case.

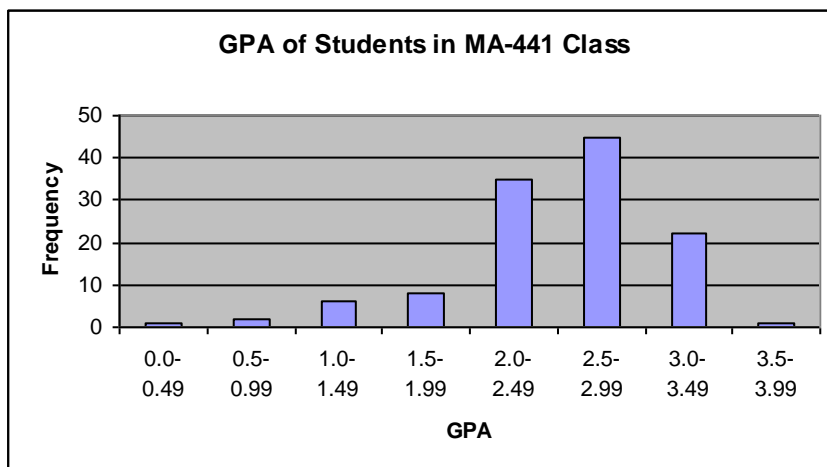
Variation

Variation measures how closely or how far apart the values of a distribution are overall: Suppose we examine the GPAs of students in two different classes:

Example 7:



Example 8:



Both distributions describe data sets of GPAs. The first distribution has greater variation than the second distribution.

Why is this true? MA-010 is an introductory mathematics class that many students have to take, while MA-441 is a calculus class for mathematics and science majors, who have had to pass prerequisites and have been at QCC for one or more semesters. We would expect more diversity among MA-010 students than MA-441 students.

Variation is not always easy to determine from one data set or distribution—it is often easier when making comparisons between graphs.

It is somewhat subjective, but we can say a distribution has low variation if the frequencies of two or three adjacent categories is much higher than all other categories and high variation if the frequencies are closer together overall.

Homework:

For 1)-10), consider the following data sets:

- a) Describe the kind of distribution (bimodal, uniform, normal, left-skewed, right-skewed) you would expect to get from a graph of the data set.
- b) Give a brief reason why you expect that kind of distribution.
 - 1) The salaries of New York Yankees baseball players
 - 2) The salaries of computer programmers in New York City
 - 3) The SAT scores of May 2004
 - 4) The weights of 100 football players
 - 5) The GPAs of 100 students in the Honors Program at QCC
 - 6) The closing prices of 100 random stocks this Friday
 - 7) The last digit of the social security number of 500 people
 - 8) The CPE exam scores of 200 QCC students this fall
 - 9) The annual income of a sample of 50 day students and 50 evening students
 - 10) The amount paid by the first 100 visitors to the Museum of Natural History, where the suggested donation is \$10

For 11)-20), consider the following data sets:

- a) Give an example of such a data set with low variation
- b) Give an example of such a data set with high variation

(For example, if the data set was “Height”. I might say that 50 members of the QCC male basketball team would have low variation (everyone is tall and over 6 feet) as compared to 50 students chosen randomly from the parking lot outside the Humanities Building.)

- 11) Annual income
- 12) GPA
- 13) Temperature
- 14) Weight
- 15) Age
- 16) Closing price of stock
- 17) IQ score
- 18) SAT score
- 19) Home Runs
- 20) Miles per Hour

That’s it for now—study for the exam!!!