Statistical Graphs

There are many ways to organize data pictorially using statistical graphs. There are line graphs, stem and leaf plots, frequency tables, histograms, bar graphs, pictographs, circle graphs and box and whisker plots to name a few. All these graphs allow us to look at a picture, rather than a bunch of numbers, get a general idea and draw some conclusions.

Statistics can best be defined as a collection and analysis of numerical information.

Often times we look at data and arrange it so it's easy to read and understand. The first statistic that many of us were formally introduced to is called a percent. For instance, if I told you that you scored 14 correct out of 17 problems, what would that mean to you?

What most of us might do is convert that to a percent. That's approximately 82%. From there, we might assign a letter grade depending upon how we set our scale.

Using a percent or a letter grade allows us a very easy way to analyze our performance. Not a big deal, just something we do regularly. Graphs help us in the same way.

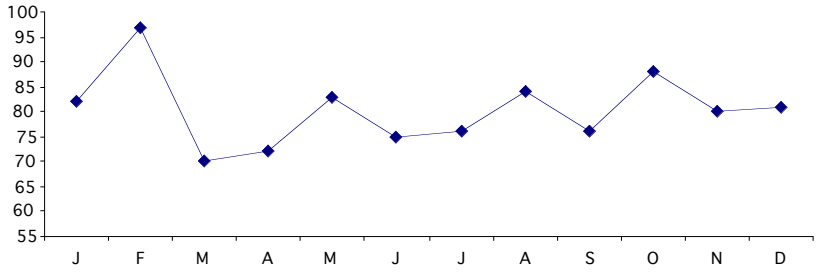
Sec. 1 Line Graph

A line plot consists of a line, on which each score is denoted by an "X" or a dot above the corresponding value on the number line that are connected.

Example 1 Let's say I have a class and I wanted to watch their performance over a 12-month period. The following scores represent the average per month.

82, 97, 70, 72, 83, 75, 76, 84, 76, 88, 80, 81

Arranging that data on a line plot, we have:



Pretty easy, don't you think? That squiggle at the end means I didn't start plotting my scores at zero, the bottom. Without that, I can distort information to unfairly portray information.

Another type of graph is called the Stem & Leaf Graph.

Sec. 2 Bar Graphs

Bar graphs are used when the information under consideration is separated into distinct categories; like months of the year and grades.

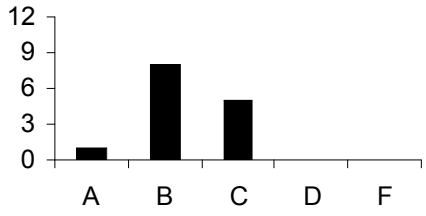
Using the same information, let's construct a bar graph so we know how many A, B, C, D, and F's there are. We define the A's as 90 and above, B's from 80 to 89, C's 70 to 79, etc.

Example 3 Use the following test scores to construct a bar graph

82, 97, 70, 72, 83, 75, 76, 84, 76, 88, 80 81, 81, 82

We have 5 C's, 8 B's and 1 A

Graphing, we have



Another way to represent data is through the use of a histogram.

Sec. 3 Histogram

A histogram looks very much like a bar graph, the difference is bar graphs represent categorical data and histograms represent continuous numerical data.

Like a bar graph, a histogram is made up of adjoining vertical rectangles or bars. If we rotated the last stem and leaf graph 90 degrees and made the rectangles as high as the left portion, we would have a histogram. A histogram looks just like a bar graph, except the rectangles are connected.

A histogram would typically identify what you are talking about on the horizontal axis, the vertical axis describes the frequency of those observations. Two problems you might encounter on a histogram, one is when data falls on the line that divides two rectangles. In which rectangle do you count the data? Another problem is the width of the rectangles, how wide do you want them?

Let's actually do a problem using the information from the previous examples.

Example 4 Use the following test scores to construct a histogram

82, 97, 70, 72, 83, 75, 76, 84, 76, 88, 80 81, 81, 82

One way to determine the width is to first find the range, the difference in the largest score and the smallest.

70, 72, 75, 76, 76, 82, 83, 84, 88, 80, 81, 81, 82, 82 and 97

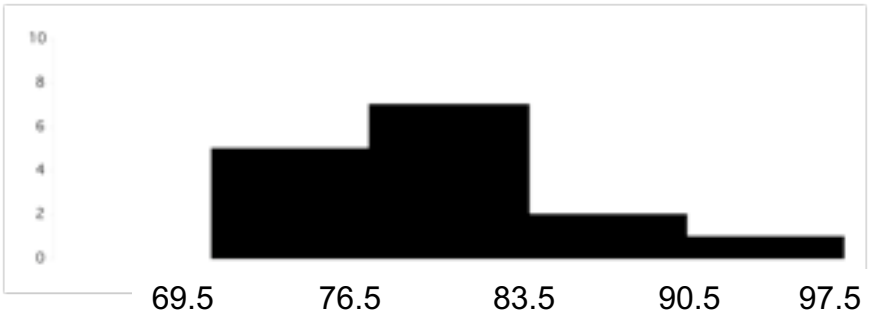
Using the data from the example, the range is 97 – 70 = 27

Now determine the number of categories you'd like to have in your histogram. If you wanted three categories, you divide 27 by three, then each width would be about nine. If you wanted four categories, you'd divide 27 by 4, then the width would be a little

bigger than 6. It's your decision based on how you would like to look t the data.

That takes care of the width problem, now what about if something falls on a line that separates the rectangles. Do we count it in the left or right rectangle? Well, we just won't let that happen. We'll expand the range by one half, then no score can fall on a line. Don't you just love how easy that was to take care of? Or, you could decide if a number falls on a line, it should be counted in the next group up. Either way, it's your decision. Just make one so the interpretation of the graph is simplest.

So, I'm deciding to have four groups, the width is a little more than 6, I'll say seven. And I'm going to begin at 69.5 rather than 70. That should result in all my data falling within a rectangle. I love it. Let's see what it looks like.



This example uses the same data as the 3 previous examples, look at each of their graphs and do some comparing – analyzing. Which graph better represent the data so it is most understandable to an audience.

Sec.4 Stem & Leaf Plots

The stem & leaf plot is nothing more than a bar graph

When we present our information, it will be in two parts, the stem and leaf. Let's say I had two scores; 57 and 54. The way I would write that in a stem & leaf plot is to use the 5| 47 notation Reading that information then, I have a 57 and a 54.

Example 2 Let's use the following test scores to construct a stem and leaf plot.

82, 97, 70, 72, 83, 75, 76, 84, 76, 88, 80, 81, 81, 82

We first determine how the stems will be defined. In our case, the stem will represent the tens column in the scores, the leaf will be represented by the ones column.

Knowing our lowest score is in the 70's and the highest is in the 90's, our stem will consist of 7, 8, and 9, representing the tens column. Usually, the smaller stems are placed on top. You could also graph these vertically.

You can make the decision for yourself. Another decision you can make is whether or not you put the scores in order in the leaf portion. As you can see, I didn't just as an illustration. Normally, I would put the numbers in order because it would make it easier to double check my work.

7 | 02566
8 | 23480112
9 |7

Notice that leaf part of the graph did not have to be in any particular order. So a person reading this plot would know the scores are 70, 72, 75, 76, 76, 82, 83, 84, 88, 80, 81, 81, 82, 82 and 97. What could be easier?

If we were to rotate the stem and leaf plot 90 degrees, make a quarter turn, the graph would resemble a bar graph. Knowing that, what do you think we'll discuss next?

Another type of commonly used graph is the circle graph or pie chart.

Sec. 5 Circle Graph/Pie Chart

A circle graph consists of a circular region partitioned into disjoint sections, each section representing a percentage of the whole.

Example 5 A family weekly income of $200 is budgeted in this manner; $60 food, $50 rent, $20 clothing, $20 books, $30 entertainment and $20 other. Construct a pie chart to illustrate that information.

A circle has a total of 360º, therefore 360 represents the total amount of the budget or 100% of the expenses. To fill in the pie chart, we have to determine what percent is spent for each expense.

To find that percent, I divide the expense by the total budgeted for the week. 60 out of $200 is budgeted for food. Converting that to a percent, we have 60/200 = 30%. So let's do percents.

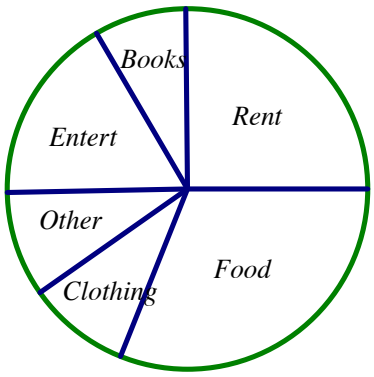
Food	- 60/200 or 30%
Rent	- 50/200 or 25%
Clothing	- 20/200 or 10%
Books	- 20/200 or 10%
Entert.	- 30/200 or 15%
Other	- 20/200 or 10%

The reason we converted those to percents is so we know how much of the circle to partition for each expense.

Since food represents 30% of the pie, we find 30% of 360º, (.30 x 360˚ = 108˚) Doing the same for the other categories, we have

Rent is .25 x 360 = 90˚
Clothing is .10 x 360˚ = 36˚
Books are .10 x 360˚ = 36˚
Entertainment is .15 x 360˚ = 54˚
Other is .10 x 360˚ = 36˚

Let's see what the pie chart looks like using those degree equivalents.



That wasn't so bad, was it?

All these different graphs do is allow us to look quickly at data to give and have some idea of what is occurring. In the above graph, we can quickly determine that most of our money is spent on food, followed by rent. There are other graphs that we didn't discuss, but are just as easily used.

Another type of graph is a box and whisker plot.

Let's look at it.

Sec. 6 Box & Whisker Plot

The box and whisker allows us to look at information broken into four groups – quartiles. To graph this information, we first divide the data in half, actually, we find the median (the middle score). The median splits the information into two groups. Next, we find the median of the top half, then we find the median of the bottom 50%. Those three medians form a box. We'll see how this works in the next example.

Example 6 Construct a Box & Whisker Plot for the following information

70, 72, 75, 76, 76, 82, 83, 84, 88, 80, 81, 81, 82, 82 and 97.

Remember, to find the median, the data must be either ascending or descending order. Since there are 15 scores, the middle score is the median. In our case that's the eighth number which happens to

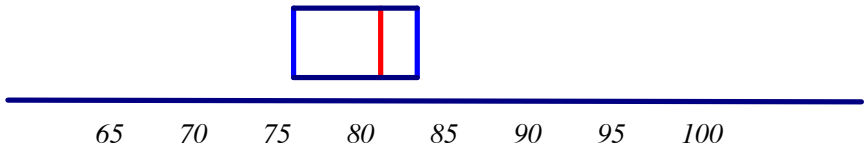
be – oh wow the scores are not in ascending or descending order. So, rewriting them, we have

70, 72, 75, 76, 76, 80, 81, 81, 82, 82, 82, 83, 84, 88 and 97.

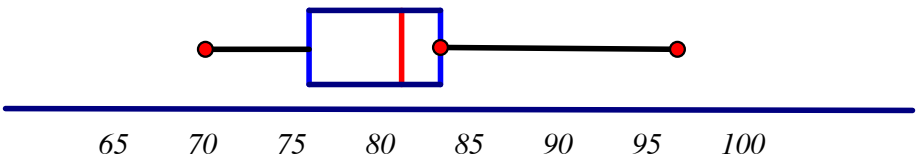
Therefore, the median is 81. That's marked with the red line below.

Now, that results in seven scores being located below the median and seven scores above the median. Piece of cake so far, don't you think? Let's go ahead and find the median for the bottom half. It's 76. Dividing the top group in two, we find the median of that group is 83. Those lines are in blue below.

The medians of those two groups make up the box, the median of the whole group just puts a divider in the box.



That's pretty nifty. Now for the whiskers. To make the whiskers, all we do is put a dot on the lowest score and on the highest and connect those dots to the box. We now have a box & whisker plot – well almost.



We are not quite done, we need to check for something called outliers.

The data in the box represent the Inter Quartile Range – abbreviated IQR, the average, the middle 50%.

The whisker on the left represents the bottom quartile, the bottom 25%, the whisker on the right represents the top 25%.

The difference between the upper and lower quartiles is called the "interquartile range" (IQR). A statistic useful for identifying extremely large or small values of data are called "outliers".

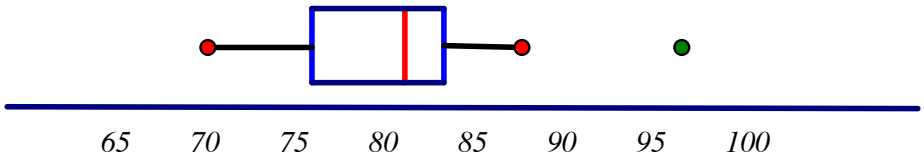
An outlier is commonly defined as any value of the data that lies more than 1.5 IQR units below the lower quartile or more than 1.5 IQR units above the upper quartile.

In our example the lower quartile was at 76, the upper at 83. Using that the IQR = 83 – 76 = 7. Multiplying 7 by 1.5, we have

$$(1.5)(7) = 10.5$$

Therefore, any score below 76 – 10.5 = 65.5 is an outlier as is any score above 83 + 10.5 = 93.5. There are no points below 65.5 so we are OK on the left. But, the 97 is above 93.5, so that score is an outlier. Outliers are indicated by using an asterisk. When there are outliers, the whiskers end at the value farthest away from the box that is within 1.5 IQR from the end. In our example that would be 88.

Now, let's see what that does to the graph.



I know what you are thinking; this is too easy. That's the problem with math, you just can't make it hard!

So,all we do in the box & whisker plots is

Algorithm for Constructing a Box and Whisker Plot

(1) find the median of the entire set of data,
(2) find the medians for the top 50% and the bottom 50% to form the box
(3) make whiskers to the largest and smallest data points from the box
(4) check for outliers.

Have you ever wondered why there were so many different types of graphs? Probably not.

But there is a simple answer. Line graphs are normally used to demonstrate trends of a single variable. Bar graphs are used for more than one variable. Like bar graphs, histograms show relationships in more than one variable but are typically continuous – like time or number grades. Bar graphs are discontinuous, like letter grades or months in a year.

A stem and leaf is very much like a frequency polygon turned vertically. It's great for comparing data in much the same way as bar graphs.

The circle graph is best used for comparisons. We can compare one piece to other pieces or to the whole and get a feeling for what's biggest, smallest or whatever.

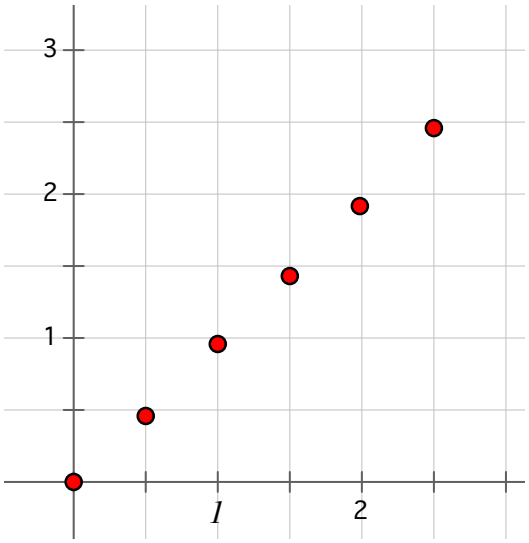
The box and whisker is used to keep people busy. Actually, it's not. A box and whisker allows us to see the median very quickly how the scores are dispersed and it also divides the data into quartiles. The smaller the boxes or whiskers, the more closely the scores are to that median. The asterisks are often referred to as outliers.

Sec. 7 Scatter Plots

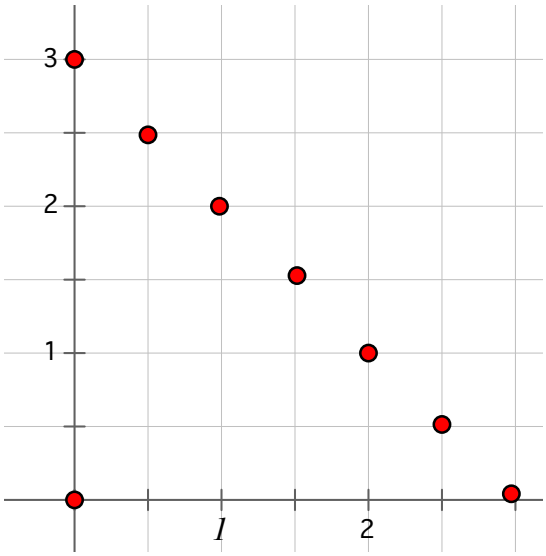
Scatter plots are like line graphs in that they use horizontal and vertical axes to plot data points. However, they have a very specific purpose. Scatter plots show how much one variable is affected by another. The relationship between two variables is called their **correlation**.

Scatter plots usually consist of many data points. The closer the data points come when plotted to making a straight line, the higher the correlation between the two variables, or the stronger the relationship.

If the data points make a straight line going from the origin out to high x- and y-values, then the variables are said to have a **positive correlation**. If the line goes from a high-value on the y-axis down to a high-value on the x-axis, the variables have a **negative correlation**.



Positive Correlation



Negative Correlation

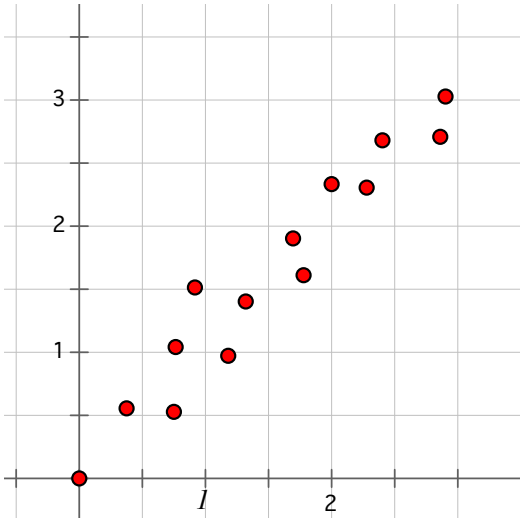
A perfect positive correlation is given the value of 1. A perfect negative correlation is given the value of –1. If there is absolutely no correlation present the value given is 0. The closer the number is to 1 or –1, the stronger the correlation, or the stronger the relationship between the variables. The closer the number is to 0, the weaker the correlation. So, something that seems to kind of correlate in a positive direction might have a value of 0.68, whereas something with an extremely weak negative correlation might have the value –.18.

Example 1 A situation where you might find a perfect positive correlation, as we have in the graph on the left above, would be when you compare the total amount of money spent on tickets to a movie with the number of people who go. This means that every time that "x" number of people go, "y" amount of money is spent on tickets without variation. An example of a direct variation.

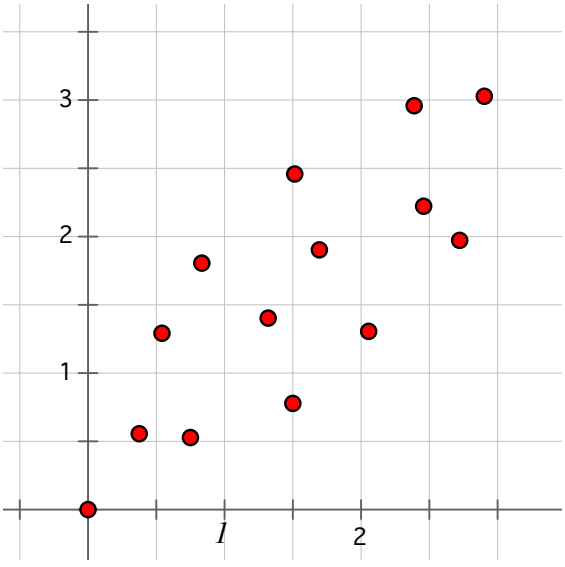
Example 2 A situation where you might find a perfect negative correlation, as in the graph on the right above, would be if you were comparing the amount of time it takes to reach a destination with the distance of a car traveling at constant rate from that destination. An indirect variation.

Example 3 A situation where you might find a strong but not perfect positive correlation would be if you examined the number of hours people practiced at golf and their scores. This won't be a perfect correlation because two people could spend the same amount of time practicing and get different scores. Generally, the rule will hold true that as the amount of time practicing increases, the lower (better) the scores.

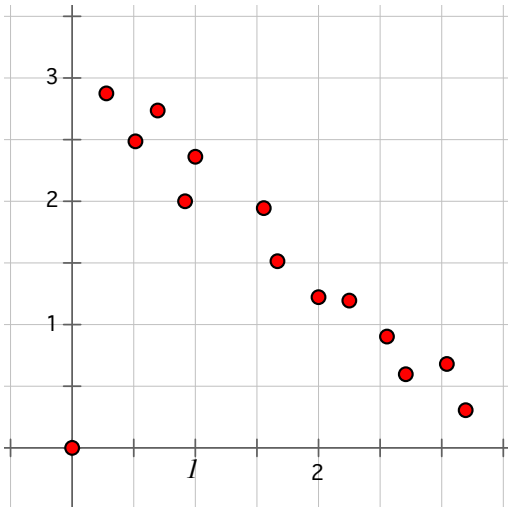
If we look at these examples, the graphs that were shown above each had a perfect correlation, so their values were 1 and –1. The graphs below do not have perfect correlations. Which graph would have a correlation of 0? What about 0.7? -0.8? 0.3? -0.3?



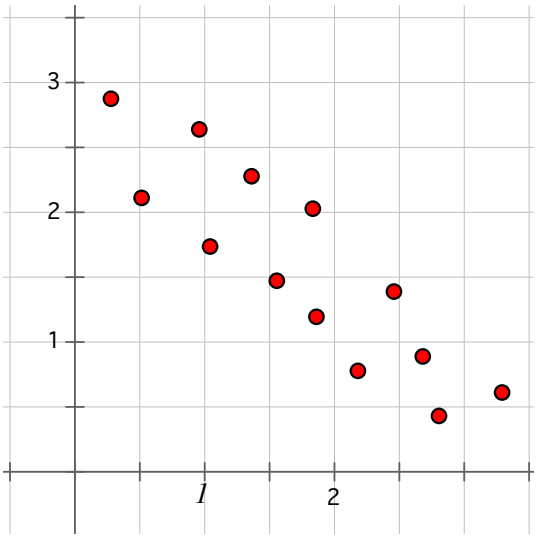
High Positive



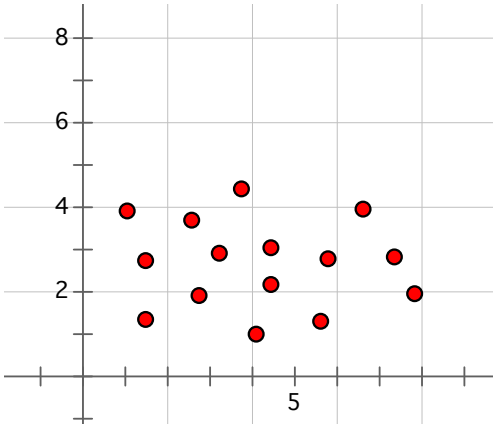
Low Positive



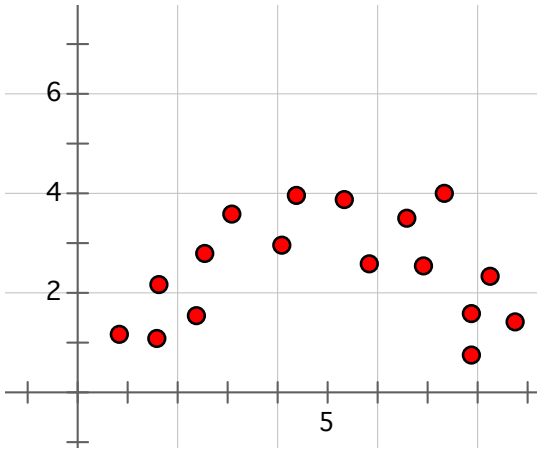
High Negative



Low Negative



No Correlation



Curvelinear

Line of Best Fit – Trend Line

There are different ways of finding lines of best fit. The one we are going to use is eyeballing the information. Another method is Grouped Averages, and a third method is Least Squares.

A **line of best fit** is a straight **line** drawn, as best we can, through the center of a group of data points plotted on a scatter plot. Scatter plots depict the results of gathering data on two variables. The "trend" line or "line of best fit" allows us to make predictions based on the information.

To find the equation describing the trend line, we use the Slope Intercept form of a Line; $y = mx + b$

To find the trend line

1. Draw a line through the center of the data.
2. Using $y = mx + b$, find the y_{int} (b)
3. Find the slope of the line.
4. Write the equation $y = mx + b$ substituting the values for b and m

To be sure, people could have picked drawn 2 different trend lines, thereby creating two different equations. Using this method, both equations would qualify as the line of best fit as long as the math was done correctly. For our purposes that is ok. Later, we can look at more formal ways of finding the line of best fit so we get the same equation.

This line is called the ***line of best fit***, it can be used as a predictor when more values are used. It's not a perfect fit, it's the best fit we can find. We will look at the difference between what our line will suggest will occur and what actually happens, that difference is called a residual.

Residuals

In order to determine how closely our ***line of best fit*** fits the data, we tabulate for each of the given values of x the difference between the observed value "y" and the value computed "y_c" by using our new equation.

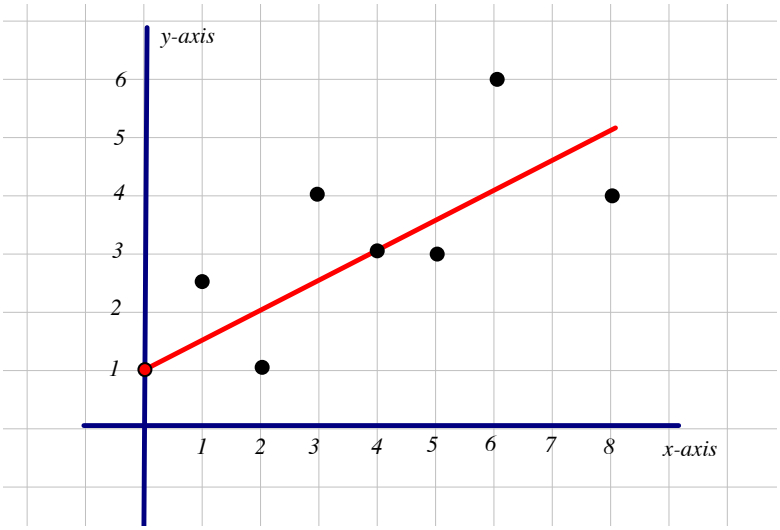
That difference $y - y_c = r$ is known as the residual. Next, we find the sum of the squares for each residual.

To be clear, we will look at actual values observed, which are our y-values. We will then look at computed values of $\boldsymbol{y}$, denoted by $\boldsymbol{y_c}$, that come from the equation of best fit that we found.

Let's look at an over simplified scatter plot, find a line of best fit (trend line) and find the residuals for different values of x.

Example Given the following data: (0, 1), (1, 2.5), (2, 1), (3, 4), (4, 3), (5, 3), (6, 6), and (8, 4). Use the red line as the line of best fit.

a. Find the equation of the line of best fit
b. Find the residual values when x has the following values; 1, 2, 4, 6 and 8.



a. The line of best fit, $y = mx + b$, the y_{int} (b) is 1. The slope, m, is ½. Therefore, the equation is $y_c = ½ x + 1$.

x	Observed y	Computed y_c	$r (y - y_c)$
1	2.5	1.5	2.5 – 1.5 = 1
2	1	2	1 – 2 = –1
4	3	3	3 – 3 = 0
6	6	4	6 – 4 = 2
8	4	5	4 – 5 = – 1

(GRAPHING)

1. The Jones family has a budget. Each month it uses its income in the following manner: 30% for food, 25% for rent, 20% for transportation, 10% for savings, 5% for entertainment, and 10% for unexpected expenses. Construct a pie graph representing this information.

2. Each dollar that the government obtains in taxes is spent in the following manner: 25 cents goes to defense, 30 cents goes to social security, 10 cents goes to farm subsidies, 15 cents goes to government salaries, and 20 cents is spent on miscellaneous social programs. Construct a circle graph representing this information.

3. In 1988 a university received the indicated amount of revenue from the following sources:

Federal Aid:	$600,000
State Aid:	700,000
Private Donations:	100,000
Corporate Donations:	200,000
Student Tuition:	300,000
Other:	100,000

   Construct a bar graph to represent the data.

4. There are 20,000 students attending a certain college. The classes are distributed in the following manner: 4,000 seniors, 3,000 juniors, 5,000 sophomores, 6,000 freshmen, and 2,000 graduate students. Construct a bar graph representing this information.

5. A statistics experiment consists of tossing a group of 8 fair coins and recording the number of heads. Construct a histogram for the thirty tosses listed below.

   6, 1, 8, 3, 6, 7, 5, 4, 5, 3, 3, 3, 7, 8, 2, 5, 2, 8, 4, 5, 4, 6, 5, 4, 1, 2, 2, 4, 6, 1.

6. A student in a math class recorded the number of doughnuts purchased by the first 30 customers in Al's doughnut shop. Construct a histogram for this data.

   2, 3, 10, 1, 4, 5, 6, 7, 9, 8, 3, 6, 3, 2, 4, 2, 5, 10, 2, 6, 2, 8, 1, 8, 8, 7, 7, 6, 5, 6.

7. The following were test scores for 33 students in a math class.

   58, 92, 85, 66, 72, 81, 60, 90, 70, 71, 77, 84, 75, 58, 89, 67, 98, 96, 70, 87, 74, 64, 64, 59, 87, 73, 91, 63, 86, 81, 72, 72, 73.

   a. Construct a grouped frequency distribution for these scores using the intervals 95-99, 90-94, 85-89, and so on.

   b. Use the frequency distribution from part (a) to construct a histogram, a frequency polygon, and a cumulative frequency graph.

   c. Construct a stem & leaf graph.

8. A survey of 32 college students was made to determine the number of books purchased for their classes in the fall semester. Construct a histogram.

   8, 7, 14, 7, 8, 10, 16, 8, 9, 15, 14, 16, 10, 14, 8, 14, 13, 8, 13, 8, 13, 8, 12, 11, 9, 12, 13, 12, 12, 7, 15, 14.

9. The scores on a math test of 40 grade school students are as follows:

   62, 65, 94, 85, 90, 43, 73, 87, 74, 42, 62, 61, 83, 68, 84, 90, 66, 71, 63, 84, 84, 76, 96, 47, 53, 78, 53, 64, 68, 58, 46, 58 58, 86, 84, 53, 87, 77, 75, 62.

   a. Construct a histogram for these grades using the intervals 95-99, 90-94, 85-89, and so on.

   b. Construct a stem & leaf plot

   c. Construct a box & whisker plot

10. The heights of 40 high school students (in inches) are given as follows:

62, 65, 54, 55, 50, 73, 73, 57, 64, 52, 62, 61, 53, 68, 64, 70, 66, 71, 63, 54, 64, 66, 56, 57, 63, 68, 53, 64, 68, 58, 66, 58, 58, 56, 64, 53, 67, 67, 70, 62.

a. Construct a grouped frequency distribution for these heights using the intervals 72-75, 69-71, 66-68, and so on.

b. Using the frequency distribution from part (a), construct a frequency polygon and a cumulative frequency graph.