Geometry and Trigonometry Module 2 Data Analysis

Section 2.1 Median-Median Line

Looking Back 2.1

In a model linear function there is a line of best fit (a straight line on a scatterplot that best represents the data). Sometimes, these can be very difficult to find. Two people may get different equations for a line of best fit given the same data.

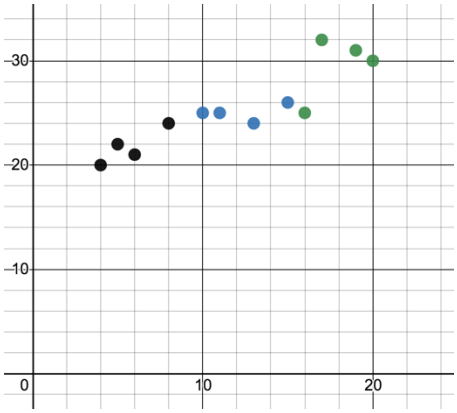
One method that can be used to find the line of best fit is called the median-median line. With this method, we divide the points into thirds and find the median-median point in each group.

Looking Ahead 2.1

Let three points called M_1, M_2, and M_3 represent an entire set of data. The equation that fits these three points the best is the line of best fit for all the data. Point M_1 represents the median of the x-coordinates and y-coordinates in the lower part of the data. Point M_2 represents the median of the x-coordinates and y-coordinates in the middle part of the data. Point M_3 represents the median of the x-coordinates and y-coordinates in the upper part of the data.

Below is a sample set of the data along with its graph.

	x	y
I	4	20
	5	22
	6	21
	8	24
II	10	25
	11	25
	13	24
	15	26
III	16	25
	17	32
	19	31
	20	30



1) Break the table into thirds. Break the graph into thirds as well to match the table. If the total number of points is not divisible by 3, then make the first and last group the same size and let the middle group have one more or one less point.

2) Find the x median and y median (the middle points) of each set of data. (You might have to order them to do this.) Mark the following points on the graph:

I: $(5.5, 21.5) = M_1$

II: $(12, 25) = M_2$

III: $(18, 30.5) = M_3$

3) The median line should go between M_1, M_2, and M_3, but be closer to M_1 and M_3 because it contains two-thirds of the data while M_2 only contains one-third of the data. Find the slope of M_1 and M_3.

This is the slope of the median-median line.

4) Find the equations and y-intercepts of M_1 and M_3. They will be the same since they are on the same line.

5) Find the equation of the line through the point M_2 using the slope above from M_1 and M_3.

6) Find the median-median line by calculating the mean of the three y-intercepts. Remember, the y-intercepts of M_1 and M_3 are the same. Use the slope from M_1 and M_3.

Section 2.2 Average-Mean Line

Looking Back 2.2

Eudoxus of Cnidus was born in 400 B.C. and has become known as one of the greatest of the ancient mathematicians. He was born near the Black Sea, went on to study medicine in Sicily, and at 23 years of age, attended Plato's academy in Athens to study philosophy and rhetoric. He would go on to be a respected mathematician as well as an astronomer. His original writings do not exist, but much of what he contributed to mathematics can be found in Euclid's Elements, as shown below:

- Definition 5 of Book V: The Theory of Proportion
- Proposition 1 of Book V: The Method of Exhaustion
- Proposition 1 of Book XII: Circles are to one another as the squares on their diameters.
- Proposition 2 of Book XII: Similar polygons inscribed in circles are to one another as the squares on their diameters.
- Proposition 5 of Book XII: Pyramids of the same height with triangular bases are to one another as their bases.

Eudoxus' contributions being restored in Euclid's writings is reminiscent of the transmission of the Bible (the copying of the Hebrew transcripts). God passed information down to prophets, such as Isaiah and Micah, who then wrote physical books of the Bible. Some prophets did not write books of the Bible; however, other prophets were inspired by God's word and wrote their stories. Elijah and Elisha were two prophets who did not write books of the Bible but their stories are told in I Kings and II Kings.

When Euclid wrote his books, Geometry rested on axioms, which are propositions regarded as being established and accepted. They do not have to be proven as they are self-evident. You will learn more about these in the Geometry section of this book.

In Statistics, the math is often self-evident. These truths are used to support or reject claims regarding data. Much data gathering and calculating is done to determine a coefficient of correlation. You studied these concepts in depth in Algebra 2 and they will be investigated further for the remainder of this module.

Looking Ahead 2.2

In the previous section, the median-median line method was used to determine the line of best fit. The final equation for the line of best fit for the data on total fat and total calories in fast-food sandwiches was $y = 12.41x + 174.21$. There are other calculations that can be made to determine how "good" a fit this line really is, and if it is the line of best fit. These calculations can be done rather quickly on a properly programmed calculator, but doing these by hand can help you understand what they mean and where they come from.

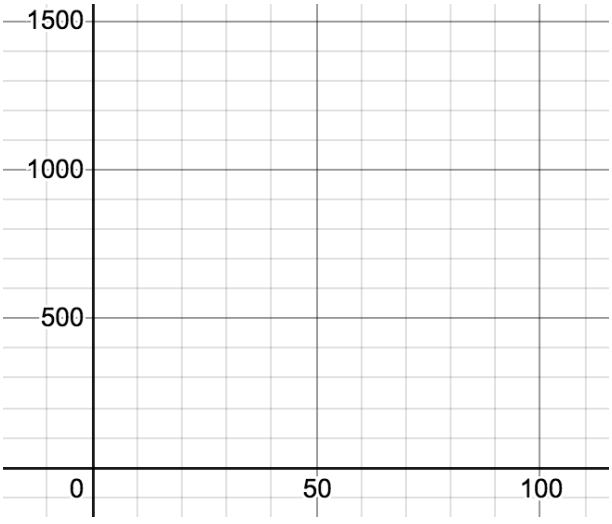
The average-mean line is one method that can be used to determine deviations. Deviation is how far something is from the norm. In statistics and mathematics, deviation is the difference between an observed or experimental value and the mean (average) of the population (not people but data).

Going back to the table on total fat and total calories in sandwiches (from the previous section), let x be total fat and y be total calories. The mean (average) can be calculated and the distances of the data from that line are the deviations from the mean.

Example 1: Find the mean of the y-values. This is called $\bar{y}$ (pronounced, "y-bar") and is the best estimate of the average calories.

Total Fat (x)	**Total Calories (y)**
3.5	310
14	360
18	340
20	780
29	550
34	520
35	630
43	700
100	1,354

Example 2: Make a graph showing the deviations of the data from the average-mean line. This can be done by drawing a line from each data point to the average-mean line.



Section 2.3 Standard Deviation

Looking Back 2.3

Some of the methods we use in Statistics are used to make sense of and compare data. Many of the calculations are measures of variability to see how much the data differs. It is important to know if the data is close to what it should be. If it is not, the "should be" might have to change (that is, we must find what the data should be) or the outlier might have to be discounted. The outliers for Problem 6 of the previous section may be the points (1, 3.5) and (20, 54) because they have the largest deviations from the norm, which we called the average-mean line.

The standard deviation can be calculated using arithmetic formulas. The average is the central location of all the data. The deviation is how much the data deviates or varies from what is considered the norm, or in this case, the average. Using mathematical notation, the equation for the average is as follows:

$$\bar{y} = \frac{\Sigma\, y}{n}$$

The symbol Σ means "sum of" so "Σ y" means "the sum of all the y data points." The symbol n represents the number of pieces of data.

Looking Ahead 2.3

The most common measure of variability is the standard deviation or is sometimes called the residual standard deviation. Standard deviation measures the spread of the data set and the relationship of the mean to the rest of the data. If the data points are close to the mean there is little variance and the standard deviation will be small. If the data points are far from the mean there is a wide variance and the standard deviation will be large. If all the data points are equal the standard deviation the variance will be 0. Using mathematical notation, the equation for sample standard deviation, s, is as follows:

$$s^2 = \frac{\Sigma(y-\bar{y})^2}{n-1} \qquad \text{Or} \qquad s = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}}$$

The variance is s^2 and is the average of the squared differences from the mean. The sample standard deviation, s, is the "sum of" the squares of the deviations divided by one less than the number of data points. The $n-1$ is used as a correction for the mean.

The sandwiches in our example is a sample from the population of sandwiches at fast food restaurants. If these sandwiches were the only sandwiches we were interested in, then they represent the population. The equation for population standard deviation is as follows:

$$\sigma^2 = \frac{\Sigma(y-\bar{y})^2}{n} \qquad \text{Or} \qquad \sigma = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}$$

The variance is σ^2 and is the average of the squared differences from the mean. The population standard deviation, σ (the Greek letter sigma), is the "sum of" the squares of the deviations divided by the number of data points.

Example 1: Find the sample mean deviation and mean deviation squared of the total fat and total calories in the table below.

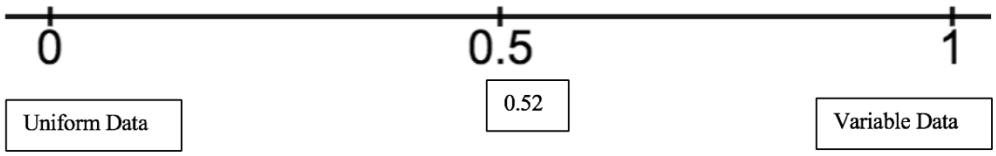
Total Fat (x)	**Total Calories (y)**	**Mean Deviation** $(y - \bar{y})$	**Mean Deviation Squared** $(y - \bar{y})^2$
3.5	310		
14	360		
18	340		
20	780		
29	550		
34	520		
35	630		
43	700		
100	1,354		

Now, this information lets us know the standard deviation and how many standard deviations our data is from the norm, but not whether the data has a great deal of variability or uniformity. The coefficient of variation, C, tells us about variability or uniformity. The formula for that is as follows:

$$C = \frac{s}{\bar{y}}$$

Earlier, $\bar{y}$ (the mean of the calories of the sandwiches) was calculated and found to be 616. If we substitute the values for s and $\bar{y}$ into the equation for the coefficient of variation we get the following solution:

The closer this value is to zero the more uniform the data is. The closer this value is to 1 the more the data varies. What do you think this tells us?



The coefficient of variation (CV) is often multiplied by 100 to get a percent. This is sometimes called the relative standard deviation (RSD) because it is the ratio of the sample or population standard deviation to the mean.

Section 2.4 Residual Deviation

Looking Back 2.4

As you now know, there are many methods that can be used to evaluate data, but some are more reliable than others. These methods include: the median-median line (Section 2.1), the average-mean line (Section 2.2) and calculating the squares of the deviations to find the coefficient of variance (Section 2.3). In this section, the method we will use will be calculating the squares of the residual deviation, which is often called residuals. This is a very efficient method to find a line of best fit as will be shown in the Practice Problems.

Looking Ahead 2.4

The residuals are calculated from the line that represents the linear equation trend line, or line of best fit, rather than the horizontal average-mean line. This measures the spread of the data and the relationship of the predicted data to the actual data. The larger the residual, the farther the actual data point is from the predicted value. The smaller the residual, the closer the actual data point is to the predicted value.

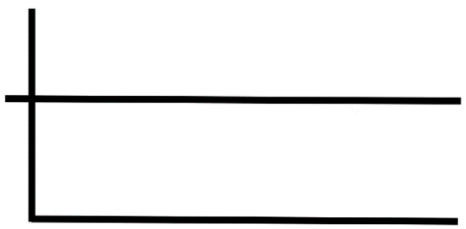
Remember, functions often model real-world data but not perfectly. Experimental data is just that-experimental, not perfect. We can predict what we think will or should happen in an experiment (which is called a hypothesis) but the conclusion may or may not support our prediction.

Using the data in Example 1, one standard form equation for a line of best fit line that compares saturated fat to total fat is $2.5x - y = -4$. The y-intercept or slope-intercept form of the equation is $y = 2.5x + 4$. If the x-values representing saturated fat are substituted in for x, the y-values representing total fat can be determined. These are predicted values from the equation, not actual values that were measured. The predicted value given by the equation is called $\hat{y}$ (pronounced, "y-hat"). The difference between the actual y-value and the predicted y-value is $y - \hat{y}$ and is called the residual.

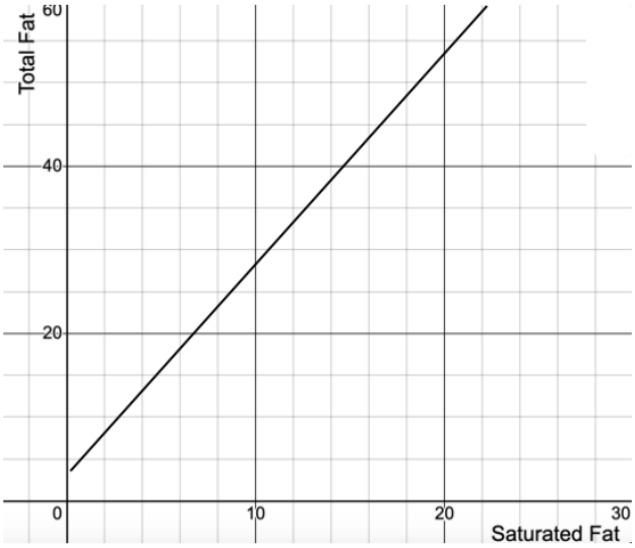
Example 1: Find the predicted value of each data point and calculate the residuals. The fats are measured in grams.

Saturated Fat (x)	**Total Fat (y)**	**Predicted Values ($\hat{y}$)**	**Residuals ($y - \hat{y}$)**
5	14		
10	35		
1	3.5		
19.5	43		
5	18		
13	34		
10	29		
20	54		

A residual plot shows the residual values on the vertical axis with a horizontal axis of 0 and the independent variables on the *x*-axis. A random plot demonstrates a linear relationship and a non-random plot demonstrates a non-linear relationship.



Example 2: Draw the graph of the residuals from Example 1 by connecting the actual data point to the predicted data point on the line of best fit.



Section 2.5 Coefficient of Correlation

Looking Back 2.5

In this section, we will be performing an experiment relating the density of a beverage to its sugar content. The coefficient of determination, r^2, is the proportion of variation of one variable that is predictable from another and $0 \leq r^2 \leq 1$. This shows how well a regression line represents the data. The linear coefficient of correlation, r, shows the strength and direction of a linear relationship between two variables and $-1 \leq r \leq 1$.

Density Analysis to Determine Sugar Content in Beverages

Safety Considerations: The samples you will use are nonhazardous; however, follow all normal safety guidelines. Do not taste or ingest any materials in the lab.

Materials:

- Sucrose solution (at room temperature): either 1%, 5%, 10%, 15%, 20% or 25%
- Graduated cylinder
- Balance
- Beverage sample

Procedure for Standard Solution

1. Tare the balance (get it to 0) and measure the mass of a clean graduated cylinder. Record the mass in grams in Data Table 1.
2. Fill the cylinder to the 10.0 mL line (exactly) with the 1% sucrose solution.
3. Tare the balance and measure the mass of the sucrose solution and graduated cylinder. Record this mass in grams in Data Table 1.
4. Add the sucrose solution to the graduated cylinder until it is at the 20.0 mL line.
5. Tare the balance and measure the mass of the solution and graduated cylinder. Record this mass in grams in Data Table 1.
6. Repeat Step 4 and Step 5, adding 10.0 mL solution each time until it reaches the 90.0 mL line.
7. Graph the mass on the y-axis and volume on the x-axis using the graphing calculator. Find the line of best fit and record the slope, equation of the line, and the value of r value under Data Table 1.
8. Using the data from your lab and from other groups, fill in Data Table 2. Record the slope values for other concentrations of sucrose solution from other lab groups.
9. From this data, plot the Density (slope) on the y-axis and Percent Sucrose on the x-axis using the graphing calculator. Find the line of best fit and record the equation for the line along with the value of r below Data Table 2.

Procedure for Beverage Sample

10. Tare the balance and measure the mass of a clean graduated cylinder. Record this mass in grams in Data Table 3.
11. Fill the cylinder to the 10.0 mL line (exactly) with the chosen beverage sample.
12. Tare the balance and measure the mass of the beverage sample and graduated cylinder. Record this mass in grams in Data Table 3.
13. Add the beverage sample to the graduated cylinder until it is at the 20.0 mL line.
14. Tare the balance and measure the mass of the beverage sample and graduated cylinder. Record this mass in grams in Data Table 3.
15. Repeat Step 13 and Step 14, adding 10.0 mL of beverage each time until it reaches the 90.0 mL line.
16. Graph Mass on the y-axis and Volume on the x-axis using the graphing calculator. Find the line of best fit and record the slope, equation of the line, and the value of r below Data Table 3.

Data:

Table 1: Standard Solution 1%

Mass of Graduated Cylinder (grams)	Mass of Graduated Cylinder & Solution (grams)	Mass of Solution (grams)	Volume of Standard Solution (mL)
			10.0
			20.0
			30.0
			40.0
			50.0
			60.0
			70.0
			80.0
			90.0

Standard Solution Graph:

Slope:

Equation for the Line of Best Fit:

r:

Table 2: Data Collected from Other Lab Groups

Percent Sucrose	Density (Slope)
1%	
5%	
10%	
15%	
20%	
25%	

Equation for the Line of Best Fit:

r:

Table 3: Beverage Sample

Mass of Graduated Cylinder (grams)	Mass of Cylinder & Solution (grams)	Mass of Beverage (grams)	Volume of Standard Solution (mL)
			10.0
			20.0
			30.0
			40.0
			50.0
			60.0
			70.0
			80.0
			90.0

Beverage Sample Graph:

Slope:

Equation for the Line of Best Fit:

r:

Results:

1. Using the line equation for the line of best fit from Step 9 (from density vs. percent sucrose of standard solutions), use the slope of the graph from your beverage to determine the amount of sucrose in your beverage. What do the values of x and y represent? What does the slope represent?

2. The Percent Sucrose formulated into the beverage for this experiment is 12%. Compare this value to the one you determined above and calculate the percentage of error.

$$\text{Percent of error} = \frac{|\text{experimental} - \text{accepted}|}{\text{accepted}} \cdot 100$$

3. This lab examines the relationship between the density of a beverage and its sugar content. What assumption is made concerning the other ingredients in the beverage and their effect on its density? Is this a valid assumption? Why or why not? What assumption can be made concerning density and sugar content?

Section 2.6 Permutations, Combinations, and Binomial Probability Distributions

Looking Back 2.6

Let us suppose that eight runners are competing in a finals event and only the first three finishers receive medals: gold for first, silver for second, and bronze for third. The medals are different for each place so order matters. This is an example of permutations of eight people taken three at a time. To find ${}_8\mathrm{P}_3$ there are only $3! = 6$ possible permutations for any one combination:

There are $3! = 6$ possible permutations. Therefore, ${}_8P_3 = 3!{}_8C_3$ or ${}_8C_3 = \frac{{}_8P_3}{3!}$. Remember, the number of combinations for n things taken r at a time is written ${}_n\mathrm{C}_r$ or $\binom{n}{r}$ and is read, "n choose r." Each combination of r objects is arranged $r!$ ways, yielding the following equation:

The equation for the combination of n things taken r at a time for all whole numbers n and r, is ${}_n\mathrm{C}_r = \frac{\mathrm{n!}}{(\mathrm{n-r})!\mathrm{r!}}$.

Example 1: Find the number of combinations for three of eight runners in a race receiving first, second, and third place medals.

Looking Ahead 2.6

In a previous course, we investigated a problem where a basketball player, Ethan, made 70% of his free-throw attempts. This represents a binomial experiment because there are only two outcomes, a success (making a free-throw) or a failure (missing a free-throw). In a binomial experiment there is a fixed number of trials, and in this one, each trial (each free-throw attempt) is an independent event, so each trial has the same probability of success.

Suppose Ethan shoots four free-throw attempts:

Outcomes in Four Trials for Free-Throws				
0 Successes	**1 Success**	**2 Successes**	**3 Successes**	**4 Successes**
F F F F	S F F F F S F F F F S F F F F S	S S F F F F S S F S F S F S S F S F F S S F S F	S S S F F S S S S F S S S S F S	S S S S

Here we see Pascal's Triangle again. There are sixteen total outcomes for four trials.

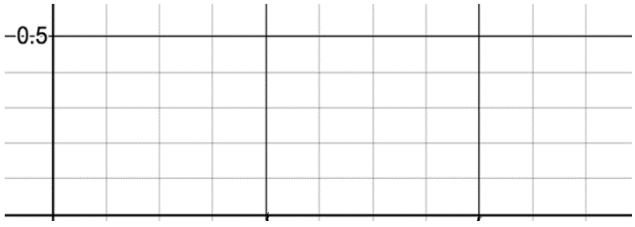
If the probability of success in each trial is p, then the probability of failure is $q = 1 - p$. The probability of having a number of exactly k successes is shown as follows:

$$P(k) = {}_nC_k \cdot p^k q^{n-k}$$

The probability of k free-throws being made in n attempts by a player with a success percentage p is a binomial probability distribution.

Example 2: Graph the probability distribution for the number of free-throws Ethan makes in four attempts. Draw the graph with k on the horizontal axis and P(k) on the vertical axis.

k	${}_nC_k$	p^k	q^{n-k}	$p(k)$
0	1	$(0.70)^0$	$(0.30)^{4-0}$	0.081
1	4	$(0.70)^1$	$(0.30)^{4-1}$	0.0756
2	6	$(0.70)^2$	$(0.30)^{4-2}$	0.2646
3	4	$(0.70)^3$	$(0.30)^{4-3}$	0.4116
4	1	$(0.70)^4$	$(0.30)^{4-4}$	0.2401



Section 2.7 Normal Distributions

Looking Back 2.7

Let us revisit the problem of Ethan's free-throw attempts. The expression ${}_{n}C_{k}p^{k}(1-p)^{n-k}$ has three variables: n represents the number of trials; p represents the probability of successes with exactly k successes. This expression is called the binomial distribution function. Therefore, $B(k,n,p)$ is the probability of k successes in n binomial trials with a probability of p successes in a single trial. The numbers representing n and p are characteristics of the binomial distribution. If two of the three variables remain constant and only one is changed, the function can be graphed on a 2-dimensional plane. Graphing all three values of the function would require a 4-dimensional plane.

Looking Ahead 2.7

Example 1: Draw the histogram for Ethan's four free-throw attempts when the probabilities are $p = 0.05$, $p = 0.5$, and $p = 0.95$.

	$\boldsymbol{p = 0.05}$	$\boldsymbol{p = 0.5}$	$\boldsymbol{p = 0.95}$
$\boldsymbol{k}$	$B(k,n,p)$	$B(k,n,p)$	$B(k,n,p)$
0	0.814506	0.0625	0.0006
1	0.171475	0.25	0.000475
2	0.013538	0.375	0.013538
3	0.000475	0.25	0.171475
4	0.00006	0.0625	0.814506

On the Graphing Calculator:

$\boldsymbol{k}$	$\boldsymbol{c}$	$\boldsymbol{c \cdot 0.05^{k} \cdot 0.95^{4-k}}$
0	1	0.814506
1	4	0.171475
2	6	0.013538
3	4	0.000475
4	1	0.00006

Example 2: Draw the histograms for free-throw attempts with a probability of 0.50 of making a free-throw when the number of trials is 2, 3 and 5?

$\boldsymbol{p = 0.5}$

$\boldsymbol{k}$	$\boldsymbol{c}$	$B(k,n,p)$
0		
1		
2		

$\boldsymbol{p = 0.5}$

$\boldsymbol{k}$	$\boldsymbol{c}$	$B(k,n,p)$
0		
1		
2		
3		

$\boldsymbol{p = 0.5}$

$\boldsymbol{k}$	$\boldsymbol{c}$	$B(k,n,p)$
0		
1		
2		
3		
4		

Example 3: What is the probability of getting exactly 4 heads when a coin is tossed 8 times?

Example 4: Women represent 45% of the people who drive sports utility vehicles. If 12 sports utility vehicle owners are called about a warranty survey, what is the probability that exactly 9 of the people contacted for the survey will be women?

Section 2.8 Mean or Expected Value of a Binomial Experiment

Looking Back 2.8

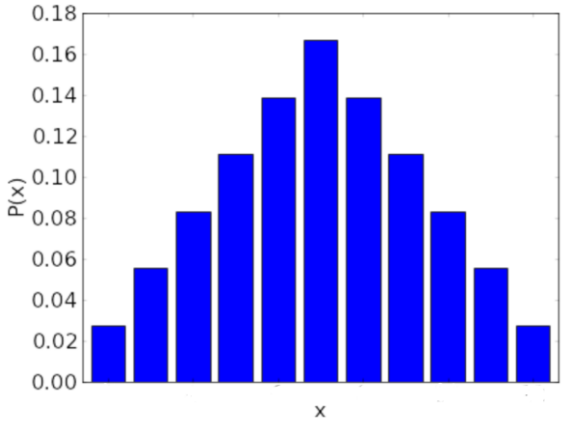
In a previous course, we played a game involving sums when a pair of dice was rolled. When a pair of fair dice are rolled, the smallest possible sum is 2 and the largest possible sum is 12.

	1	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

The probability of each sum is given in the table and the sum of the probabilities is equal to 1.

Sum (x)	2	3	4	5	6	7	8	9	10	11	12
Probability of Sum ($P(x)$)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Notice the patterns in both the chart and the table. The random variable x is the sum of the numbers on the faces of the two dice rolled. It is determined by the outcome in the experiment. The ordered pair $(x, P(x))$ represents a probability distribution. The function maps the random variable onto its probability.



The mean of the probability distribution is 7. We will see how this is calculated in Example 1.

Example 1: Draw the graph of the histogram for $(x, F(x))$ where x is the sum of rolling a pair of dice and $F(x)$ is the probability of getting that same result. For this experiment, the dice are rolled 50 times.

Sum (x)	2	3	4	5	6	7	8	9	10	11	12
Frequency of Sum ($F(x)$)	$\frac{2}{50}$	$\frac{2}{50}$	$\frac{5}{50}$	$\frac{6}{50}$	$\frac{7}{50}$	$\frac{8}{50}$	$\frac{8}{50}$	$\frac{5}{50}$	$\frac{4}{50}$	$\frac{3}{50}$	$\frac{1}{50}$

Looking Ahead 2.8

To find the mean of the frequencies in the experiment, add all 50 outcomes together and then divide that sum by 50. There was one 12, two 2's, two 3's, three 11's, four 10's, five 4's, five 9's, six 5's, seven 6's, eight 7's, and eight 8's. The sum of the x's is shown as follows:

$$2 \cdot 2 + 3 \cdot 2 + 4 \cdot 5 + 5 \cdot 6 + 6 \cdot 7 + 7 \cdot 8 + 8 \cdot 8 + 9 \cdot 5 + 10 \cdot 4 + 11 \cdot 3 + 12 \cdot 1 =$$

This means that the mean (or expected value) is $\bar{x} = \frac{352}{50} = 7.04$.

Another way to write this using the frequency of each sum is shown as follows:

$$\bar{x} = 2 \cdot \frac{2}{50} + 3 \cdot \frac{2}{50} + 4 \cdot \frac{5}{50} + 5 \cdot \frac{6}{50} + 6 \cdot \frac{7}{50} + 7 \cdot \frac{8}{50} + 8 \cdot \frac{8}{50} + 9 \cdot \frac{5}{50} + 10 \cdot \frac{4}{50} + 11\frac{3}{50} + 12 \cdot \frac{1}{50}$$

$$\bar{x} = \frac{2}{25} + \frac{3}{25} + \frac{2}{5} + \frac{6}{10} + \frac{21}{25} + \frac{28}{25} + \frac{32}{25} + \frac{9}{10} + \frac{4}{5} + \frac{33}{50} + \frac{6}{25}$$

$$\bar{x} = \frac{12}{25} + \frac{6}{5} + \frac{15}{10} + \frac{33}{50} = 7.04$$

Both methods yield the same answer. The formula for the mean of the frequencies of the sum can generally be written as follows:

$$\bar{x} = x_1 F(x_1) + x_2 F(x_2) + \ldots x_n F(x_n)$$

This equation can be written using summation notation, which will be used in Pre-Calculus and Calculus. It is shown in summation notation as follows:

$$\bar{x} = \sum_{i=1}^{n} (x_i \cdot F(x_i))$$

The summation for mean is $\frac{1}{n} \sum_{i=1}^{n} x_i$ ($\bar{x}$ for n data values).

For example: $\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$

$= 2 + 3 + 4 + 5 + 6$

$= 20$

This summation leads us to a formula with the variable μ for the mean or expected value of any given probability distribution $\{(x_1, P(x_1)), (x_2(P(x_2)), \ldots (x_n, P(x_n))\}$. The mean (or expected value) of the distribution is shown as follows:

$$\mu = \sum_{i=1}^{n} (x_i \cdot P(x_i))$$

Example 2: Ethan has the probability p of making a free-throw with a single shot, and all free-throw attempts are independent. Find the expected value of Ethan making a free-throw shot in three attempts.

The mean of a binomial distribution with n trials and p successes can be generalized as follows: $\mu = np$.

Section 2.9 Variance and Standard Deviation of a Binomial Distribution

Looking Back 2.9

Median is a measure of center for interquartile range. Variance and standard deviation are two other measures that describe how data spreads out in relation to the mean. Both variance and standard deviation are calculated from the deviation, which is the difference of each data value from the mean. The variance is the average of the squared deviations and the standard deviation is the square root of the variance. You worked with some of these concepts in Section 2.5 when calculating the coefficient of correlation.

After calculating the mean of a set of data, we find the difference (deviation) of each value from the mean. Then we find the sum of all the squares of each deviation. We divide this sum by $n - 1$ and the number we get is the variance. We take the square root of the variance and we get the standard deviation.

Example 1: Find the variance and standard deviation of the test scores given below.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{800}{10} = 80$$

Test Score (x)	**Deviation $(x - \bar{x})$**	**Square of the Deviation $(x - \bar{x})^2$**
68		
72		
73		
74		
74		
76		
78		
80		
82		
83		

Sum of Squares of the Deviation:

Sometimes this is called the sample variance, and σ^2 (which is the lowercase Greek letter sigma) is used to represent population variance. The denominator is n, rather than $n - 1$. Therefore, σ, which is the square root of sigma squared, represents the standard deviation of the population.

Variance: $\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$ Standard Deviation: $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$

These are all statistics symbols that are used by calculators: s^2 and s; σ^2 and σ.

Looking Ahead 2.9

The variance of a binomial probability distribution where $P(x) = B(x, n, p)$ can be calculated when n and p are known. The population variance that follows…

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

… can be written as follows:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i{}^2)}{n} - \bar{x}^2$$

(This was demonstrated in Example 1 of Section 2.8 where $\bar{x}$ represented a mean of numbers.)
Therefore, the variance formula for a probability distribution with a mean μ and n outcomes can be written as follows:

$$\sigma^2 = \left(\sum_{i=1}^{n} \left(x_i{}^2 \cdot P(x_i) \right) \right) - \mu^2$$

Example 2: Ethan has the probability p of making a free-throw with a single shot, and all free-throw attempts are independent events. Find the variance of Ethan making a free-throw shot in three attempts. Use the table below from Example 1 of Section 2.8.

Number of Shots (x)	**Probability of Made Shots ($P(x)$)**
3	p^3
2	$3p^2q$
1	$3pq^2$
0	q^3

The general formula for variance of a binomial distribution with n trials each, and with a probability of p successes and q failures is $\sigma^2 = npq$.
The standard deviation of the same binomial distribution is $\sigma = \sqrt{npq}$.

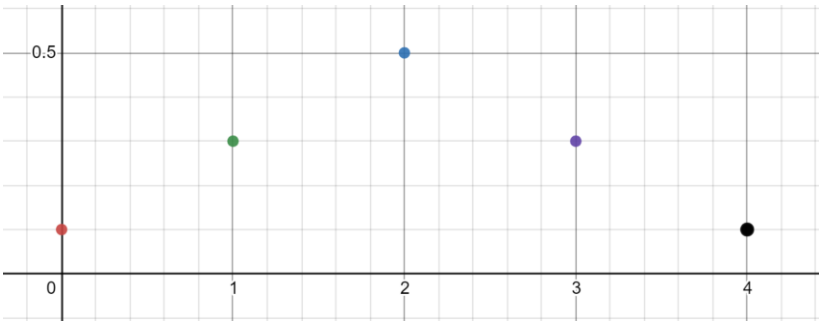
Section 2.10 Standard Normal Distributions and z-scores

Looking Back 2.10

Probabilities are the areas under the curve of a binomial distribution.

The number of free-throws made in four attempts is given by the binomial probability distribution graph. The width of each bar of the histogram is one and the height is $p(k)$. Therefore, the area is $A = 1 \cdot p(k) = p(k)$ where k is a continuous random variable. The area under the bar is equal to the probability of the random number k.

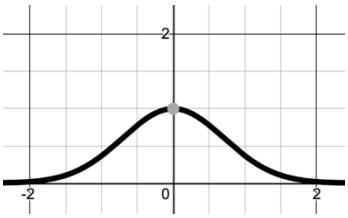
Example 1:	Given the graph below, find $p(x < 1)$ and $p(x < 2)$.



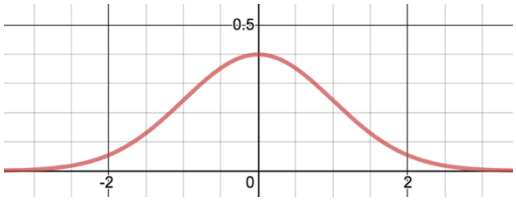
Looking Ahead 2.10

As the number of trials of a binomial probability distribution with a fixed success probability (p) increases, the binomial probability distribution approaches a normal bell curve. This is called a normal distribution and the graph is called a normal curve. This is also called the Gaussian probability distribution and is most used in science. Named after the brilliant mathematician and physicist, Carl Friedrich Gauss, who contributed greatly to mathematics and science, and pondered infinitely greater problems such as our relation to God which he found to be beyond us and outside of the realms of science to solve.

The parent function of a normal curve is $f(x) = e^{-x^2}$ and e is the base of natural logarithms (that naturally occur in nature) and will be discussed further in Calculus. Any transformation of this is also a bell-shaped curve.



This parent function has several properties that make it usable in data analysis. A child of this parent function is $f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$, which is known as the standard normal curve and the probability distribution for it is called the standard normal distribution.



The area under the curve but above the z-axis is 1. It is an even function symmetric to the line $z = 0$. The $p(z < 0) = 0.5$, and because it is symmetric, $p(z > 0) = 0.5$ as well. Probabilities derived from this function are important and are recorded in tables and on graphing calculators.

For example, $p(z < 0.88) \approx 0.8106$. This means that $p(z > 0.88)$ is $1 - p(z < 0.88) = 1 - 0.8106 = 0.1894$.

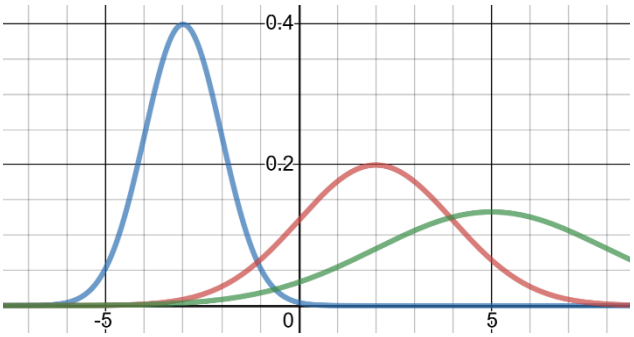
The $p(0 < z < 0.88)$ is the area under the normal curve between 0 and 0.88.

$$P(0 < z < 0.88) = p(z < 0.88) - p(z \leq 0)$$

$$= 0.8106 - 0.50$$

$$= 0.3106$$

Example 2: Below are three normal curves. What does m appear to represent? What does s appear to represent?



Blue: $m = -3$; $s = 1$
Red: $m = 2$; $s = 2$
Green: $m = 5$; $s = 3$

Example 3: The $p(0.88 < z < 1.7) = p(z < 1.7) - p(z < 0.88) = 0.9554 - 0.8106 = 0.1448$ is the probability that a tool will not be manufactured within the standard specifications. What is the probability (as a percentage) that the tool will meet the standard specifications?

The z-scores will be looked at more closely in the next section, as they are extremely beneficial in statistical reasoning.

Section 2.11 Null and Alternative Hypothesis Testing

Looking Back 2.11

Judgements and inferences about information can be made using binomial or normal distributions. A judgement and/or an inference can help us decide whether to support a given conclusion. This is called hypothesis testing.

Let us suppose a coin is tossed 20 times and you want to know the probability of getting tails such that the number of times the tail appears is five away from the expected value of the mean. The probability for getting a tail is $p = 0.5$. The expected number of tails is $20(0.5) = 10$.

Example 1: When a fair coin is tossed 20 times, what is the probability of getting tails such that the number of times that tails appears is five away from the expected value of the mean of 10 tails?

To use a z-table for this problem, find where the tenths place in the z-row and the hundredths place in the z-column meet. A positive z-score is to the right of the mean and are greater than the mean. A negative z-score is to the left of the mean and are less than the mean. The normal bell curve is symmetric.

When a fair coin is tossed 20 times, the probability of getting tails such that the number of times that a tails appears is five away from the expected value of the mean of 10 tails is about 0.0258 (about 1 chance in 50).

Looking Ahead 2.11

When researchers conduct experiments, the z-score is compared to the p-value set beforehand. This is to determine if there is evidence to support the hypothesis being tested and keep the findings ethical. This value is the calculated probability, of finding the observed, or more extreme values of the hypothesis being tested. This type of analysis is called hypothesis testing. It is often viewed in terms of a null hypothesis, H_0, (the hypothesis being tested) or an alternative hypothesis, H_a(which may be anything else).

The null hypothesis (H_0 (pronounced "H-naught")) predicts the experiment will go as expected. It is what is accepted and being tested. There is no remarkable difference between the specified variables or populations. Any observed differences are explained by experimental errors or sampling methods. The alternative hypothesis (H_a) predicts a contrasting conclusion for the same event.

Let us investigate an experiment with rats that have cancer. The average baby rat, called a pup, has a birth weight from 6 to 8 grams. Pups whose mothers have cancer have an average birth weight from 5 to 6 grams. The mothers with cancer are given medicine to determine if their pups have any significant change in weight.

H_0: the medicine will not affect birth weight $\mu = 5.5$
H_a: the medicine will affect birth weight $\mu \neq 5.5$

The method of choosing H_0 and H_a is subjective and declared at the beginning of an experiment. Since we are dealing with probabilities, statisticians try to determine the more reasonable of the two hypotheses. Sometimes the p-value or significance level is subjective also. It can be based on results of sampling and testing.

At the beginning of testing, statisticians choose a value for an experiment that is called the significance level, α (the Greek letter alpha) to determine the minimal probabilities that are acceptable. For an experiment, significance levels are often at 0.10, 0.05, or 0.01. If the significance level is $\alpha = 0.10$, then there is a 1 in 10 probability of rejecting the null hypothesis when it is true. The confidence level is $c = 1 - \alpha$, which means there is a 9 in 10 probability of not rejecting the null hypothesis when it is true.

Tests of significance begin with a statement of the null hypothesis (H_0) and the alternative hypothesis (H_a). Calculations of the test statistic, z, are performed. These will be based on how far the data are from H_0. Calculations of the p-value will be performed and then statisticians will base a conclusion on the significance level less than or equal to α.

Section 2.12 Central Limit Theorem and Confidence Intervals

Looking Back 2.12

In a survey, when the population is a group of men and women, the distribution is normal (bell-shaped) if the population distribution is normal. A sample of the group, such as just men or just women, would also be normal.

However, sometimes data is skewed to one side or the other. Many population distributions are not normal. Samples can also be skewed to a side and not normal.

In probability, as the sample size gets larger, the distribution of $\bar{x}$ gets closer to a normal distribution. The original shape of the graph of the population does not matter. This is called the Central Limit Theorem.

The sample of size n drawn from any population with mean μ and finite standard deviation σ is approximately normal and has a mean of $\mu_{\bar{x}} = \mu$ and a standard deviation of $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Example 1: Phone calls are made to survey consumers about their newspaper readership. The standard deviation of the population is $\sigma = 169.72$ seconds. The length of one phone call can often be far from the population mean of μ. Find the standard deviation for the mean length for samples of 20, 40, and 80 calls.

Looking Ahead 2.12

We can estimate with confidence using confidence intervals. The 68-95-99.7 rule says that 68% of data falls within 1 standard deviation of the mean ($\bar{x}$) of a set of data, 95% of data falls within 2 standard deviations of the mean, and 99.7% falls within 3 standard deviations of the mean.

Let us suppose 50 students of a population of 450 students in a school take an ACT exam and that the standard deviation is 3 points for the population.

$$\sigma_{\bar{x}} = \frac{3}{\sqrt{50}}$$

$$\sigma_{\bar{x}} \approx 0.43$$

The 68-95-99.7 rule tells us that the probability $\bar{x}$ will be within 0.43 points (1 standard deviation) of the population mean score μ is about 68% (0.68). The probability that $\bar{x}$ will be within 0.86 points (2 standard deviations) of the population mean score μ is about 95% (0.95).

We can be 95% confident the mean score for the students lies between $\bar{x} - 0.86$ and $\bar{x} + 0.86$ of the mean μ. If $\bar{x}$ is 27, then $27 - 0.86 = 26.14$ and $27 + 0.86 = 27.86$.

H_0: the interval between 26.14 and 27.86 contains the true μ
H_a: the interval between 26.14 and 27.86 does not contain the true μ

A confidence interval C for a value is computed from sample data by a method that has probability C of providing an interval containing the true value.

The z_c (z-score) corresponds to an interval. The table and graph below show data for the most common confidence interval.

z_c	1.645	1.960	2.576
C	90%	95%	99%

The sample mean $\bar{x}$ has the normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, so the probability that $\bar{x}$ lies between $\mu - z_c \frac{\sigma}{\sqrt{n}}$ and $\mu + z_c \frac{\sigma}{\sqrt{n}}$ is C. This means that the population mean μ of the random sample mean $\bar{x}$ lies between $\bar{x} - z_c \frac{\sigma}{\sqrt{n}}$ and $\bar{x} + z_c \frac{\sigma}{\sqrt{n}}$. The margin of error for confidence interval C added to or subtracted from the mean is $m = z_c \frac{\sigma}{\sqrt{n}}$.

The value of z_c on the normal curve with area C between $-z_c$ and z_c has a level C confidence interval for μ of $\bar{x} \pm m$.

Example 2: A sample group of 1,500 high school seniors passed the ACT and went on to college. The mean of their college debt was $24,500 with a standard deviation of approximately $36,000.00. Compute a 95% confidence interval for the mean college debt of the proportion of students in debt for college.

Section 2.13 Probability and Two-Way Tables

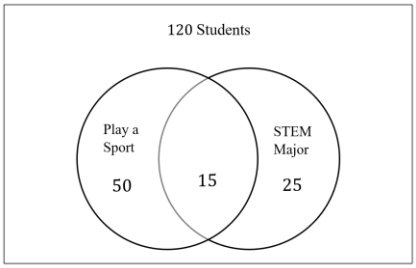
Looking Back 2.13

At the beginning of this module, the data being investigated was related to scatter plots. Standard deviation, residual deviation, and coefficient of correlation were explored. Bivariate data was described as the relation between two variables in which one is the independent variable and the other is the dependent variable. Univariate data includes a single variable and the data is analyzed in terms of spread and center.

In these last few sections, we looked at the probabilities that involve categorical data such as heights and weights of people in given groups. We also looked at the probability of given values appearing randomly in games of chance, results of surveys, and random selections.

There is one final method we will investigate that can be used to evaluate probabilities of categorical data: two-way tables. A two-way table displays the same information as a Venn diagram, but instead of two overlapping circles that represent the two categories, one category is represented by a row and the other is represented by a column.

Example 1: A survey of 120 students at a community college asked whether they play a sport and/or have a STEM (Science, Technology, Engineering, and Math) major. The Venn diagram below represents the data. Complete the two-way table.



	Have a STEM Major	Do not Have a STEM Major	Total
Play a Sport			65
Do not Play a Sport			55
Total	40	80	120

Looking Ahead 2.13

Joint relative frequency is the ratio of a frequency (how often a point appears in data) in a specific collection and the whole collection. These points are on the interior of the two-way table. To find a joint relative frequency, we divide each frequency of points by the total number in the collection.

Example 2: Find the joint relative frequencies for the table.

	Have a STEM Major	Do not Have a STEM Major	Total
Play a Sport			
Do not Play a Sport			
Total			

Analyzing a two-way table:

a) About _____ of the students do not play a sport that do not have a STEM Major.

b) About _____ of the students do not play a sport that do have a STEM Major. However, about _____ of students do play a sport that have a STEM Major.

c) About _____ of the students surveyed do not play a sport. About _____ of the students surveyed do have a STEM Major.

Marginal relative frequency is the sum of the joint relative frequencies in the rows and columns of the two-way table.

Conditional relative frequency is the ratio of the joint relative frequency to the marginal relative frequency. These can both be found using a row or column total of a two-way table.

Example 3: Use the marginal relative frequency of each row to find the conditional relative frequency of the table below.

	Have a STEM Major	**Do not Have a STEM Major**	**Total**
Play a Sport			
Do not Play a Sport			
Total			

	Have a STEM Major	**Do not Have a STEM Major**	**Total**
Play a Sport			
Do not Play a Sport			
Total			

Example 4: An employee maps out three different routes to work. He drives to work using each route for one month (30 days) and records whether he is late or on time. The table below shows the joint relative frequencies and marginal relative frequencies for the data.

	On Time	**Late**	**Total**
Route A	(7) 0.23	(5) 0.17	0.4
Route B	(4) 0.13	(4) 0.13	0.26
Route C	(8) 0.27	(2) 0.07	0.34
	0.63	0.37	≈ 1

a) What is the probability that the employee is on time using Route A?

b) What is the probability that the employee is on time using Route B?

c) What is the probability that the employee is on time using Route C?

d) Which route should the employee choose to drive to work to be late as little as possible?