1.3 Graphs and Relations between Variables

1.3.1 Overview

In physical settings, we usually consider measurements of multiple quantities at the same time. We do this because we are interested in the relationships between these different quantities. We think of each quantity of interest as a **state variable**; the collection of all such variables under consideration is called the **system**. At any instant, the variables of the system will each have a particular value and the collection of those values at that instant is called the **state** of the system. Graphs are often used to reveal relationships between different variables.

In this section, we explore graphs of data through scatter plots and graphs of equations. An equation involving multiple variables represents a relation between those variables. We consider the role of solving equations in the context of these relations.

1.3.2 Systems, States and Variables

In the course of an experiment, or even just in observation, many different quantities typically are covarying, or changing with one another. For example, an object in motion has changing position, changing velocity, and changing forces. In the course of a chemical reaction, there are changing concentrations of the different reactants and products. Other quantities might also change, such as temperature, pH, and volume. While observing a changing population, there could be changing population numbers, total biomass, birth and death rates, consumption of resources, and production of products and waste.

Mathematically, the **system** consists of all possible observable quantities associated with the experiment or observed physical system. The **state** of the system refers to the collection of instantaneous values of all such quantities at a particular instant or configuration of the system. A **state variable**, or more simply a variable, represents a single quantity that is or could be observed in the system. Quantities that can be calculated in terms of state variables are mathematically **dependent variables** and are also examples of state variables, even if they can not be directly measured.

Example 1.3.1 Consider the following data about the population, births and deaths in the United States. To conserve space, the data are given using scientific notation expressed in the standard machine form where the power of 10 follows the letter E, so that 2.521×10^8 would be written 2.521E8.

Year	Population	Births	Deaths	Year	Population	Births	Deaths
1991	2.530E8	4.111E6	2.170E6	2001	2.850E8	4.026E6	2.416E6
1992	2.565E8	4.065E6	2.176E6	2002	2.876E8	4.022E6	2.443E6
1993	2.599E8	4.000E6	2.269E6	2003	2.901E8	4.090E6	2.448E6
1994	2.631E8	3.953E6	2.279E6	2004	2.928E8	4.112E6	2.397E6
1995	2.663E8	3.900E6	2.312E6	2005	2.955E8	4.138E6	2.448E6
1996	2.694E8	3.891E6	2.315E6	2006	2.984E8	4.266E6	2.426E6
1997	2.727E8	3.881E6	2.314E6	2007	3.012E8	4.316E6	2.424E6
1998	2.758E8	3.942E6	2.337E6	2008	3.041E8	4.248E6	2.472E6
1999	2.790E8	3.959E6	2.391E6	2009	3.068E8	4.131E6	2.437E6
2000	2.822E8	4.059E6	2.403E6	2010	3.094E8	3.999E6	2.468E6

Each row (corresponding to the population in a given year) represents a distinct state of the system. The observed values in the state are the **vari-**

ables: the year, the total population at the beginning of the year, the total number of births in the year, and the total number of deaths in the year. The year is included as one of the variables—an independent variable—in order to distinguish the different states with respect to time. □

We often represent a variable by a symbol—a letter, a Greek letter, an abbreviation, or even a word. That symbol becomes a name for the variable to be used in sentences, expressions, and equations. Uppercase and lowercase letters are different symbols and should not be interchanged with one another. The choice of symbol should generally be related to the meaning of the variable. An important part of communication in modeling is in stating clearly the variables of a system and identifying the symbols that are chosen to represent them.

Example 1.3.2 In the previous example, there were four variables. A common strategy is to use the first letter of a word describing each variable. The population variable might be represented by the symbol P. Births and deaths might be represented by the symbols B and D, respectively. The year might be represented by the symbol Y.

Note that the symbols p, b, d and y are *not* the same as the symbols above, even though they have the same letter names. They should not be used for this problem. □

The next example illustrates how we might write a short explanation of a system and the variables associated with it. Note how the physical explanation of the system is described first, followed by an introduction of the measurements taken and the symbols used to represent those variables. Any time you have data and refer to the data by variables, you need a few sentences that introduce the meaning of each variable along with the units of measurement.

Example 1.3.3 In biology, scientists run electrophoresis gels to determine the size of polymers, such as proteins or DNA strands. The gel provides a porous structure for the polymers to travel through while an electric potential (voltage) creates a force that pulls the polymers through the gel. Different size polymers travel at different speeds. The experiment is setup with all polymers starting at one end of the gel, the voltage is turned on for a certain amount of time and then disconnected. Clusters of similarly sized polymers are identified visually as bands on the gel, with smaller polymers traveling a greater distance.

The image below represents an electrophoresis gel run on a standardized collection of DNA of fixed sizes. Because the image does not show a length scale, the distances traveled by the different lengths are measured in image pixels and recorded in the table below. The variables for the experiment are the length of DNA segments and the distance traveled through the gel. Let L represent the length of the segment (in nucleotides) and let D represent the distance traveled (in pixels), measured from the center of the starting well to the center of the corresponding band in the image. Each row represents a single state (L, D) of the system.

L (nts)	D (px)
100	342
200	327
300	312
400	299
500	288
600	278
700	270
800	263
900	256
1000	249

$\square$

1.3.3 Scatter Plots and Relationships Between Variables

The primary motivation for collecting data regarding different variables in the state of a system is to determine relationships between those variables. One of the ways that we look for relationships is using a **scatter plot**. A scatter plot is a graph showing the relationship between two variables. Suppose the two variables use symbols x and y. For each state of the system, there will have been observed values for both x and y. The graph will include points for each pair (x, y).

Spreadsheets (like Microsoft Excel, Apple Numbers or Google Sheets) are a common tool to generate scatter plots. The data are first put in a table. The first column of data will correspond to the variable used for the horizontal axis (x), and the second column of data will correspond to the variable for the vertical axis (y). Select the two columns at the same time and add a chart to your spreadsheet, choosing the scatter plot style of graph. You should become familiar with how to create a scatter plot. Always be sure that you label your axes, using the variables of the system rather than the generic names of x and y.

The following figure shows two different scatter plots for the electrophoresis gel data above. One plot is based on the pairs (L, D) whereas the other is based on the pairs (D, L). These graphs contains the same information but viewed from a reverse perspective. When we switch the order of the variables, we call the relationships **inverse relations**.
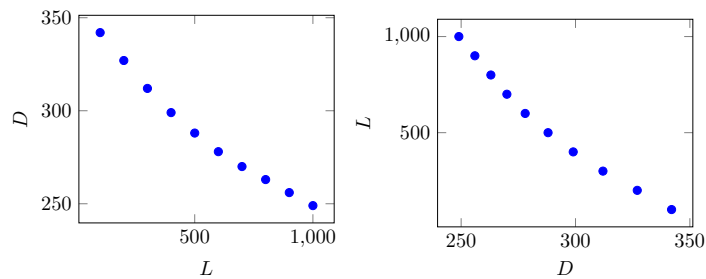

Figure 1.3.4 Scatter plot of electrophoresis data, displacement vs size and the inverse relation size vs displacement.

When a system has a state defined by more than two variables, scatter plots can be defined for each pair of state variables. For example, the population data has four state variables, (Y, P, B, D). Three scatter plots can be formed by plotting the population, the total births and the total deaths versus the year, giving graphs of points (Y, P), (Y, B), (Y, D). Because the births and deaths

are on the same scale, we can combine the plots as one. The inverse relations (P, Y), (B, Y) and (D, Y) contain the same information from a different view and are not shown.
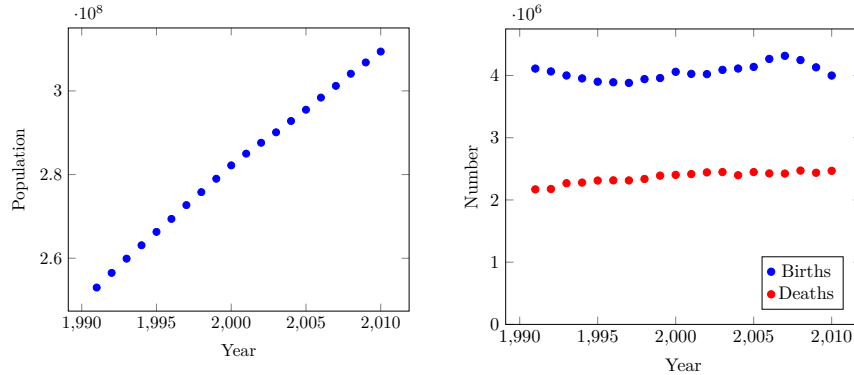

Figure 1.3.5 Scatter plot of population, births and deaths with respect to time.

We can also look at relationships between other pairs of variables. For example, we can look at how the number of births or deaths relate to the population, plotting (P, B) and (P, D), or how the number of births relate to the number of deaths with (B, D). The graph showing the relation between births and deaths to time (above) is very similar to the graph showing the relation between births and deaths to population (below). However, the relation between the births and deaths illustrates that sometimes variables do not show a clear relation.
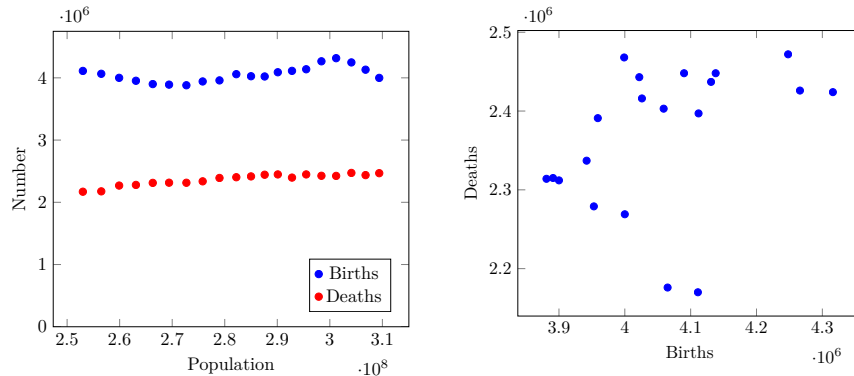

Figure 1.3.6 Graphs showing the relations (P, B), (P, D), and (B, D).

1.3.4 Graphs of Equations

An **equation** gives an abstract representation of a relationship between variables by stating that two **expressions** are equal in value. Just as the state of an experimental system is defined by the value of the variables defining the state, an equation can be considered as a mathematical way to define relationships between variables of an abstract system. A **solution** to the equation is a state for the variables such that the equation is true. The graph of an equation generalizes a scatter plot by including all solutions of the equation. If we choose an ordering for the variables (e.g., alphabetical), the values for the variables can be conveniently listed as an ordered list. When two variables are involved in an equation, the ordered list is called an ordered pair, or point, like (x, y), and the graph of the equation is typically a curve in the plane.

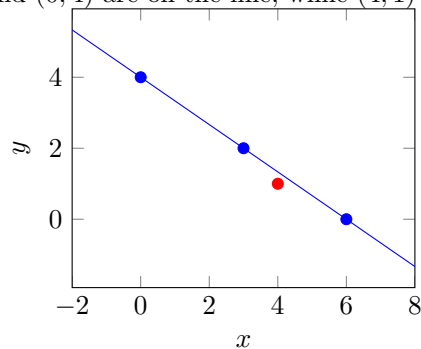**Example 1.3.7** The equation

$$2x + 3y = 12$$

involves two variables, x and y, and is the equation of a line. The expressions in the equation are $2x + 3y$ and 12. The values $x = 3$ and $y = 2$, corresponding to the ordered pair $(x, y) = (3, 2)$, provide one solution because for those values,

$$2x + 3y = 2(3) + 3(2) = 12,$$

so that the equation is true. On the other hand, $(x, y) = (4, 1)$ is not a solution because for that state,

$$2x + 3y = 2(4) + 3(1) = 11$$

and $11 \neq 12$. Some other solutions include the points $(6, 0)$ and $(0, 4)$. The line corresponding to this equation represents the set of all such solutions. The points $(3, 2)$, $(6, 0)$ and $(0, 4)$ are on the line, while $(4, 1)$ is not.

$\square$

**Example 1.3.8** The equation

$$u^2 + v^2 = 16 + 6u$$

also involves two variables, u and v. The expressions in the equation are $u^2 + v^2$ and $16 + 6u$. Using ordered pairs (u, v), the points $(3, 5)$ and $(3, -5)$ are solutions. That is, if $(u, v) = (3, 5)$, the expressions have the same value:

$$u^2 + v^2 = 3^2 + 5^2 = 9 + 25 = 34,$$
$$16 + 6u = 16 + 6(3) = 16 + 18 = 34.$$

It is possible to show that the graph of solutions for this equation is a circle centered at $(3, 0)$ with radius 5. Other points on this circle include such points as $(-2, 0)$ and $(6, -4)$. You should verify that these are also solutions, at least for one or two points to reinforce the idea that a solution makes the statement of the equation true.

□
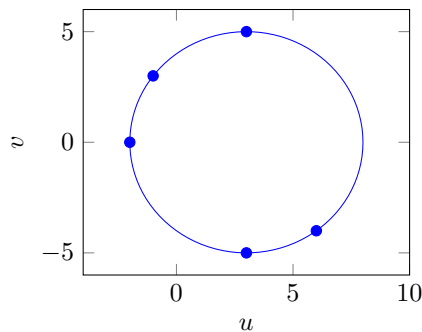
It is usually difficult to know how to sketch the graph of an arbitrary equation. Computer utilities that support implicit plots can be used. For example, the online graphing calculator at desmos.com allows you to enter an equation involving variables x and y. We also could use computational systems, such as SageMath shown below, to create an implicit plot.

```
# Declare the variables
var("u,v")
# Create a graphics object from an implicit plot
myplot = implicit_plot(u^2+v^2==16+6*u, (u,-5,10), (v,-5,5))
# Show the graph with axis labels
show(myplot, axes_labels=['u','v'])
```

When an equation is written as a dependent variable being equal to an expression involving an independent variable, we can easily generate points that are in the solution set using a table with the independent variable in the first column and the dependent variable in the second column. We choose convenient values for the independent variable, compute the value of the expression that depends on that variable, and then use that resulting value for the dependent variable. All such points will be solutions to the equation. This is precisely how a graphing calculator works internally; it computes many such points very quickly and connects the points with line segments.

Example 1.3.9 Rewrite the equation $2x + 3y = 12$ so that y is the dependent variable. Use the new equation to find four points in the solution set.

Solution. We need to isolate the variable y using balanced operations.

$$2x + 3y = 12$$
$$3y = -2x + 12$$
$$y = \frac{1}{3}(-2x + 12)$$
$$y = -\frac{2}{3}x + 4$$

The final equation $y = -\frac{2}{3}x + 4$ should be recognized as a slope-intercept equation of a line. The slope is $m = -\frac{2}{3}$ while the y-intercept value is $b = 4$. Having solved for y, we can finish the task by using four different values for x to find corresponding values for y. We do this in a table.

x	$y = -\frac{2}{3}x + 4$	(x, y)
0	$-\frac{2}{3}(0) + 4 = 4$	$(0, 4)$
1	$-\frac{2}{3}(1) + 4 = \frac{10}{3}$	$(1, \frac{10}{3})$
2	$-\frac{2}{3}(2) + 4 = \frac{8}{3}$	$(2, \frac{8}{3})$
3	$-\frac{2}{3}(3) + 4 = 2$	$(3, 2)$

□

1.3.5 Parametrized Models and Regression Curves

Suppose we have data that appear to show a relation between two variables in a scatter plot. We would like to extend the relation to data that are not in the table of known values. If we had a mathematical equation that described our relation, we could use that equation to find the solution that would match the desired values. As a practitioner, we choose a **parametrized model** and

then use a computational tool to select the best model given our data. The most common computational strategy is called **regression**.

A **parametrized model** is an equation relating state variables that includes additional variables representing **model parameters**. The model is identified by choosing particular values for each of the parameters. Once the parameters are known, the equation establishes a relation for the state variables. A given parametrized model describes an entire family of different relations, one relation for each choice of parameters.

The most common example in algebra of a parametrized model is a linear equation

$$y = mx + b.$$

The symbols m and b are the model parameters, and x and y are the state variables. The particular equation $y = 2x - 5$ is in this family of relations based on the parameter values $m = 2$ and $b = -5$.

Another example of a parametrized model,

$$y = ax^2 + bx + c,$$

which has three parameters a, b, and c, can be used to create relations whose graphs are parabolas. The simplest parabola, $y = x^2$, corresponds to the parameter values $a = 1$, $b = 0$, and $c = 0$. Curiously, linear models are contained in this family as well by choosing $a = 0$. Our earlier example $y = 2x - 5$ could have been obtained from this model using $a = 0$, $b = 2$, and $c = -5$.

Notice that the symbols used for the parameters do not have universal meaning. In the linear parametrized models, we had chosen b to represent the y-intercept value. In the quadratic models, the parameter b was used for the coefficient of x.

Regression is a strategy to select parameters for a parametrized model in such a way that it "best" matches data for a given relation. Mathematical equations are exact. Real data exhibit uncertainty and randomness. Consequently, there usually aren't parameter values that will match all of the data simultaneously. The most common regression algorithms seek to the sum of the squared errors and are called **least-squares regression**. Spreadsheets and graphing calculators that find a trend line for data use this type of regression. We revisit finding parametrized model to match exact data in a later section.

A trend line or a trend curve resulting from regression provides a model that allows us to predict values where there are not observed data. When the prediction occurs between observed data, such prediction is called **interpolation**. If the prediction is occurring beyond the extremes of the data, such prediction is called **extrapolation**. We can use the value for one variable and the regression equation to solve for the predicted value of the related variable. Often, a formula may not describe all of the data but provides a good approximation for a certain range of values. Interpolation is usually safer than extrapolation.

Example 1.3.10 Consider the population example with the scatter plot of the number of deaths plotted with respect to the total population size. Find the linear regression model for these data and predict the number of deaths in a year if the population were 300 million.

The easiest tool to find a regression model seems to be at the website desmos.com/calculator. The site desmos.com does not support scientific notation for data entry, we can make a modified model. Let $\widetilde{P} = P/10^8$ be the population in units of 100 million and let $\widetilde{D} = D/10^6$ be the annual death rate

in units of 1 million. We are going to enter the data shown in the table below.

| $\widetilde{P}$	$\widetilde{B}$	$\widetilde{P}$	$\widetilde{B}$	$\widetilde{P}$	$\widetilde{B}$	$\widetilde{P}$	$\widetilde{B}$
2.530	2.170	2.694	2.315	2.850	2.416	2.984	2.426
2.565	2.176	2.727	2.314	2.876	2.443	3.012	2.424
2.599	2.269	2.758	2.337	2.901	2.448	3.041	2.472
2.631	2.279	2.790	2.391	2.928	2.397	3.068	2.437
2.663	2.312	2.822	2.403	2.955	2.448	3.094	2.468

1. Create a table to enter the data. Either click the `+` menu and select `table` or type `table` in the formula field.

2. Enter the population values $\widetilde{P}$ in the column x_1 and the corresponding death rate values $\widetilde{D}$ in the column y_1. The data are now plotted and you should see they look roughly linear.

3. We now construct the parametrized model for the data. In Desmos, this is done by creating an equation using the tilde symbol `~` in place of an equals. If we want to use the parametrized model $\widetilde{D} = a\widetilde{P} + b$ with parameters a and b, we would type into the next formula `y1 ~ a x1 + b`.

4. Desmos will report values for the parameters a and b and draw the trend line through the scatter plot. The parameters are identified as $a = 0.486666$ and $b = 0.992711$ so that the model equation is

$$\widetilde{D} = 0.486666\widetilde{P} + 0.992711.$$

Specify static image with `@preview` attribute,
Or create and provide automatic screenshot as
`images/interactive-1-preview.png` via the `mbx` script
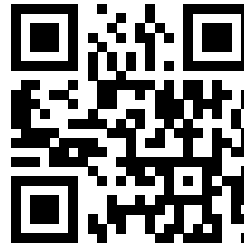
`www.desmos.com/calculator/my9xzvkfea`
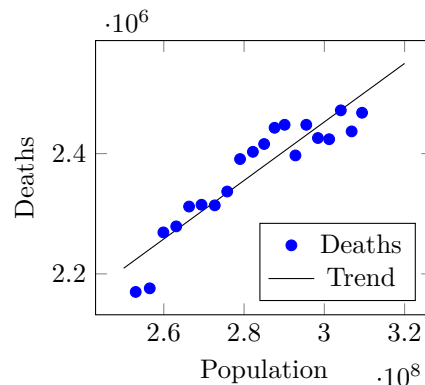
We can now use the model to predict the number of deaths per year for a population of 300 million. This corresponds to $P = 300 \times 10^6 = 3 \times 10^8$ so that $\widetilde{P} = 3$. Using the parametrized model, we find

$$\widetilde{D} = 0.486666(3) + 0.992711$$
$$= 2.452691$$

Because $\widetilde{D}$ is the number of deaths in units of millions, the model predicts 2,452,691 deaths per year for a population of 300 million. Since the original data only had four significant digits, we should not expect any more digits accuracy in the model prediction. We would predict 2.453 million deaths. □

Example 1.3.11 Consider the electrophoresis gel data. Suppose we had another DNA sample of unknown length that traveled a distance of $D = 282$ pixels. Use a model to estimate the length of the DNA sample.

Solution. Because we know the distance displaced in the gel and want to predict the length of the polymer, we treat D as the independent variable and L as the dependent variable. We will look at the scatter plot (D, L) with the length of the DNA L graphed with respect to the distance traveled in the gel D (Figure 1.3.4). The data appear smooth with a slight upward curve. A nonlinear model will be required to model the bend, such as a quadratic parametrized model,
$$L = aD^2 + bD + c.$$

We enter the data in a table and apply regression with our model. In Desmos, we would create a table for (x_1, y_1) with values of D in x_1 and values of L in y_1. We then calculate model parameters using `y1 ~ a x1^2 + b x1 + c`. The resulting model parameters are $a = 0.0573428$, $b = -43.381$, and $c = 8241.57$. Consequently, the trend curve is modeled by

$$L = 0.0573428D^2 - 43.381D + 8241.57.$$

The graph of the data with the trend curve is shown below.
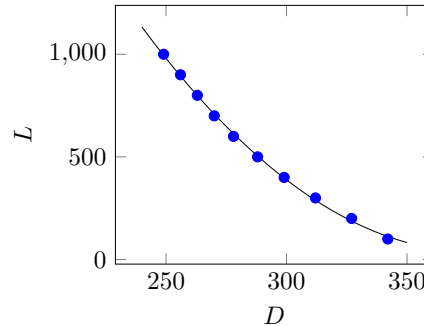
Using our value for D, we can find the value of L using the model,

$$L = 0.0573428(282)^2 - 43.381(282) + 8241.57 \approx 568.26.$$

Since our original data had 3 significant digits, we would estimate the length of the DNA in question as $L \approx 568$ nucleotides. In this way, a regression of known electrophoresis data allows us to estimate lengths of other molecules.
□

You should note that the number of significant digits reported is not the same as the uncertainty in the prediction. The degree to which the original data vary around the trend curve leads to uncertainty in the coefficients of the regression model and subsequent uncertainty to the trend curve itself. In the last example, rounding the model parameters themselves to 3 significant digits would have changed the predicted length by 11 nucleotides. Analysis of this uncertainty is a topic for statistics and is outside the scope of this text. For simplicity, we use models to make predictions and then round to comparable precision as the data.

1.3.6 Summary

- Quantities that can be measured correspond to state variables. A system is the collection of all possible variables. The state of the system is the collection of values measured for all of the variables simultaneously. An important part of communication is describing all relevant variables and introducing their names.

- A relation between two variables can often be visualized graphically using a scatterplot. An equation is the mathematical idealization of a relation. The graph of an equation involving two variables, say x and y, shows all solutions as points (x, y).

- When the equation is written as a dependent variable equal to an expression of the dependent variable, points on the graph can be quickly tabulated using the formula.

- Using regression to find a trend line or regression curve can give an approximate relation corresponding to observed data. Treating the resulting equation as a model equation can give approximate predictions of states of the system.

1.3.7 Exercises

Each problem has an equation involving two variables. Determine whether each of the given states for those variables are in the solution set.

1. $3x - 2y = 8$

     (a) $(x, y) = (0, -4)$

     (b) $(x, y) = (1, -2)$

     (c) $(x, y) = (4, 2)$

2. $2w + 5z - 3 = w^2 + z^2$

     (a) $(w, z) = (-2, 2)$

     (b) $(w, z) = (-1, 3)$

     (c) $(w, z) = (3, 2)$

Each problem has an equation involving multiple variables. Solve for the indicated dependent variable.

3. **Perimeter of Rectangle.** Given $2L + 2W = P$, solve for W.

4. **Volume of Rectangular Prism.** Given $V = LWH$, solve for L.

5. **Volume of Cylinder.** Given $V = \pi r^2 h$, solve for h.

6. **Ideal Gas Law.** Given $PV = nRT$, solve for P.

Given an equation relating two variables, solve for the indicated dependent variable. Use your resulting expression to calculate the value for that variable given the values of the indicated independent variable. Make note of any values that are not defined. Plot the corresponding points in the solution set of the equation on a graph.

7. Given the equation $4x + 5y = 20$, find y for each value $x \in \{1, 2, 3, 4, 5\}$.

8. Given the equation $np = 1000$, find n for each value $p \in \{10, 20, 25, 40, 50\}$.

Graph two dependent variables representing the expressions on each side of the equation. Use the points of intersection to identify solutions to the equation. Verify that the values you identify are solutions by testing whether make the equation true.

9. $3x - 5 = x + 2$

10. $\dfrac{20x}{x+4} = x + 3$

11. $4x^3 - 9x^2 = x - 6$

Additional DNA samples were run in the same electrophoresis gel as described in Example 1.3.11. Using the data and regression curve from that example, estimate the length of each sample. Indicate whether the approximation is appropriate.

12. Estimate the length of a DNA sample that traveled 200 pixels.

13. Estimate the length of a DNA sample that traveled 335 pixels.

14. Estimate the length of a DNA sample that traveled 350 pixels.

A Voltage–Resistance–Current Relationship   A simple electric circuit has an applied voltage V (volts) and a variable load resistance R (kilohms). When the circuit is closed, current flows through the circuit, measured as the current I (amperes). When the voltage was held constant at $V = 9$ V, the resistance and current were measured with values recorded in the table below. The following group of problems are based on these data.

V (V)	R (kΩ)	I (A)
9.0	0.84	0.0107
9.0	1.2	0.0073
9.0	1.8	0.0050
9.0	2.7	0.0033
9.0	3.4	0.0026

15. Create a scatter plot of (R, I). Would a trend line make sense for this data? Explain.

16. Conductance G is the reciprocal of resistance, $G = 1/R$. Create a scatter plot of (G, I). Would a trend line make sense for this data? Explain.

17. One of the previous scatter plots should have had a meaningful trend line. State an appropriate regression equation as a model and use it to predict the current I when the resistance is $R = 2.1$ kΩ.

Population-Growth Relationships   The number of births and of deaths in a population generally depends on the size of the population. The table below gives population data for ten of the twelve highest population cities in the state of Virginia for the year 2012. The data include the population P and the total number of births B and deaths D for the year recorded for each city. The following group of exercises are based on these data.

City	P	B	D
Virginia Beach	447021	6270	2828
Norfolk	245782	3773	1827
Chesapeake	228417	2805	1582
Richmond	210309	2939	1849
Newport News	180726	2905	1438
Alexandria	146294	2763	686
Roanoke	97469	1492	1172
Portsmouth	96470	1534	980
Suffolk	85181	1087	726
Lynchburg	77113	1062	779

18. Create a scatter plot of (P, B) and find the equation of the trend line. The cities of Hampton and Harrisonburg were left off the list with populations of $P = 136836$ and $P = 50981$, respectively. Use the trend line regression model to predict the number of births in these cities during 2012. Which calculation is an example of interpolation and which is extrapolation?

19. Create a scatter plot of (P, D) and find the equation of the trend line. Use the trend line regression model to predict the number of deaths · in Hampton and Harrisonburg during 2012. (See the previous problem for population values.) Which calculation is an example of interpolation and which is extrapolation?

Which of the calculations were examples of interpolation and which were examples of extrapolation?