Algebra: Modeling Our World

Michael Catalano
Dakota Wesleyan University
Spring, 2009 edition

Contents

Algebra: Modeling Our World

1. Introduction

Example is the school of mankind. He will learn at no other.
— Edmund Burke

Example is not the main thing in influencing others. It is the only thing.
— Albert Schweitzer (1875 - 1965)

You have very likely taken an algebra course or two in your time, and this text is designed for use in a college algebra course or what might be termed quantitative literacy courses at a similar level. You will likely find a good deal which will be familiar.

However, courses using this text are also likely to be different than other algebra courses you may have taken. In particular,

1. This course will emphasize learning through example. Information will often be presented in the context of real-world situations. Many of our examples will be related to social and economic issues, including, hunger, poverty, energy, and the environment both within the United States and worldwide.

2. A large part of the work you do in this course will be done in groups, including much of the daily homework.

3. This course is designed to help you become a better learner, especially of mathematics. In our society, knowledge is expanding at an ever increasing rate, and the knowledge which is relevant and useful today may not be tomorrow. This means that a college education is no longer *only* about acquiring "all" the knowledge you need to know to be an educated citizen. Rather, it is about acquiring a base of useful knowledge and examples, and the ability to build on this base as you progress through life. We will explicitly address this issue as part of the course.

4. The text is in more of a narrative style than most other mathematics textbooks. Although there will be examples and problems worked out in the text, the book is designed to be read 'cover to cover' rather than used only as a reference. I encourage to read thoroughly and read ahead. There are approximately 160 pages of text to read, not counting exercises and group activities. In a 15 week semester, you will be roughly reading 10 to 12 pages each week.

5. This course will emphasize the usefulness of mathematics. While mathematics is a worthy (and beautiful!) discipline in its own right, we will concentrate on what mathematics can say about the real-world, and help you develop skills in applying mathematics to understanding real-world situations.

The overall goal is for you to not only develop your mathematical knowledge and skills, but to become more confident in those skills, gain a greater appreciation for the usefulness of these skills in understanding aspects of our modern society, and enhance your ability to apply those skills in whatever your chosen field happens to be.

In addition, this course is designed to provide more relevant content and a different pedagogical experience for future elementary school teachers.

> *As technology has "mathematicized" the workplace and as mathematics has permeated society, a complacent America has tolerated underachievement as the norm for mathematics education. We have inherited a mathematics curriculum, conforming to the past, blind to the future, and bound by a tradition of minimum expectations*
> —- from Everybody Counts: A report to the nation on the future of mathematics education (Washington, DC: National Research Council/Mathematical Sciences Education Board, 1989)[1] (from Summary)

Mathematics in the Real World

> *Providing nourishment, health, and happiness to all of God's children will cost each American only a few dollars. That's a small price to pay. In the long run it will cost us a lot more if we let 300 million children remain hungry and poorly educated. How many Albert Einsteins, Marie Curies, Martin Luther Kings, and Michael Jordans will never develop if we do nothing to feed the hungry children of the world? Who can evaluate that loss?*
> — Ambassador George P. McGovern

As noted earlier, this course will emphasize examples from the real world, including examples related to issues of hunger and poverty. Why this particular focus? There are a number of reasons.

The first is that hunger and poverty are prevalent throughout the world and directly affect billions of people. Even in the United States, one of the wealthiest nations on earth, a significant proportion of the population has to cope with not having sufficient resources for basic necessities on a monthly, if not daily, basis. The author would not be surprised if there are people reading this introduction who feel they have experienced this. The second is that these issues are related to many other aspects of our world wide society, including agriculture, energy, political democracy, the environment, education, and even violent conflicts including terrorism. While we may not think about these issues on a day to day basis, they greatly influence the larger context of our lives, from what types of jobs and other opportunities might be available to us, to where we live and the quality of life that we experience there.

Finally, this focus partially arose out of the mission of the McGovern Center for Leadership and Public Service at Dakota Wesleyan University.[2] The center is named after former South Dakota Senator George McGovern, a Dakota Wesleyan University graduate, and probably most famous for his 1972 campaign for President of the United States against Richard Nixon. McGovern has spent much of his career working to alleviate world hunger, including recent service as United Nations Global Ambassador on World Hunger.

Although hunger and poverty issues will be emphasized throughout the course, this will certainly not be the sole focus. We will consider quite a variety of examples and applications of mathematics. Here are a few of the issues and situations we will be considering.

1. **Baby boom:** The so-called "baby boom" is the result of 75 million babies being born in the 18 years from 1946 to 1964, as compared to 44 million from 1926 to 1946.[3] This population surge has had, and will continue to have profound effects on American society at least through 2050. For example, in 1950, there were 3 persons 85 years and up for every 100 persons age 50 to 64. In 2050, this ratio is projected to increase to 27 to 100. In the year 2030, there will be as many 70-year-olds as 30-year-olds (up from 1 to 3 in 1990). The number of working age citizens for every person over 65 is projected to go from 3.2 to 2 by 2030.

[1] http://www.napanet.net/ jlege/allcount.htm
[2] http://www.mcgoverncenter.com/
[3] http://www.hometechsystems.com/demographics2.htm

How are such projections made, and how confident can we be of these projections? Assuming these projections prove true, what sort of impact will this have on our society? Are there other countries where “baby booms” have occurred or are occurring? How differently will you have to plan for retirement than your parents or grandparents?

2. **Ozone Ouch:** The ozone hole over the Antarctic grew from about 180,000 square miles in 1980 to about 9 million square miles in 2001. How can we use the known data to predict what might happen in the future? What assumptions are we making in our projections? How reasonable are these assumptions?

3. **It’s Never Too Early to Think About Retirement:** What would you like to be doing when you are 65 or 70? Whether you have ambitious plans (a series of world cruises?) or not, you will need to set aside a certain amount of resources to live on, unless you are counting on working until the day you die. There is the Social Security System, but as noted in item one above, there will be fewer and fewer workers to support retirees as time goes on, so Social Security benefits may be diminished by the time you are of retirement age. How much should you be saving for your retirement? When should you start?

4. **Demographic Details:** As part of our consideration of hunger and poverty issues, we will be looking at a variety of world demographic data. For example, the chart below shows the annual population growth rate worldwide since 1950, and projected out until 2050.

   How are these future predictions made? What assumptions are part of making these projections? What will be the effect on the total population? What happened around 1960?

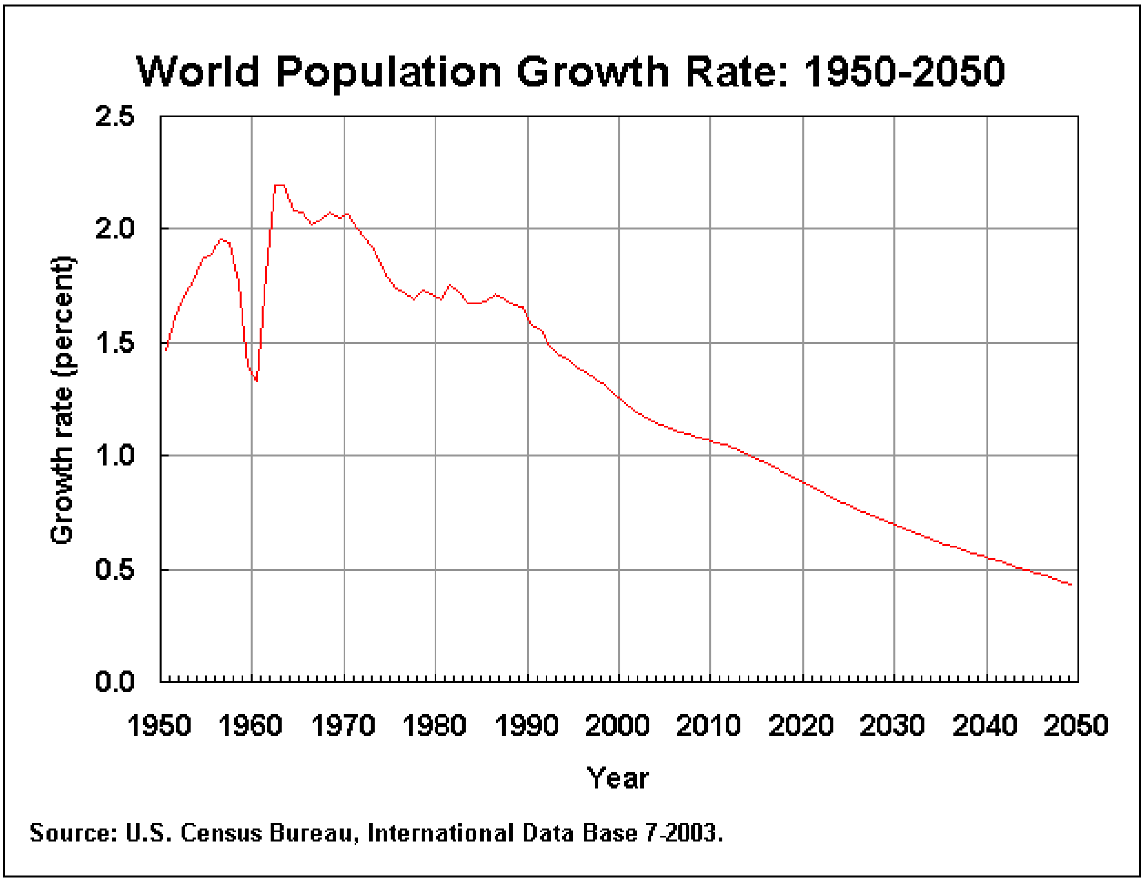


Figure 1: World population growth rates

5. **Medical Testing:** You have tested positive for a dangerous, though not too common disease. How concerned should you be? (In other words, what are the chances you actually have the disease?)

6. **The Great California Recall:** In October of 2003, action hero Arnold Schwarzenegger was elected Governor of California, replacing Governor Grey Davis who was removed using the recall provision of California state law. In one poll of 801 registered California voters leading up to the election, 50% indicated they would vote to remove Governor Gray Davis from office, and Democratic Lieutenant Governor Cruz Bustamante led the pack of replacement candidates with 35%, with Schwarzenegger favored by 22% of the voters. The margin of error for the poll was reported as 3%.

   How do we find such "margins of error," and what do they really tell us? More generally, how can we use sampling appropriately to make predictions? How did Schwarzenegger end up getting elected?

Hunger and Poverty

We all have some awareness of the problem of world hunger, and you may have some ideas about why it occurs, and what can or cannot be done about it. This course will provide you with the opportunity to study some aspects of hunger and poverty in more detail, and to see how much of the "conventional wisdom" is really true. Here are some questions we might consider as we think about world hunger. These are difficult questions, and even with the help of mathematics, it may only be possible to provide incomplete answers.

> Is hunger really just a matter of there not being enough food in particular locations at particular times, or is the problem more complicated?
>
> Does the world currently produce enough food to feed everyone? If so, how many more people can world resources reasonably support? How long before the world population reaches this level?
>
> Could we eradicate hunger by simply giving enough food to those who are hungry? If we tried this, how much food would donor nations have to provide on an annual basis?
>
> Is hunger simply a consequence of some being rich and others being poor? Could we alleviate hunger by making the world society more equitable, from an income or wealth standpoint?
>
> Is "free trade" a good way to alleviate hunger? Would hunger disappear if we were able to establish a completely free global market system?
>
> Can we alleviate hunger and still maintain reasonable environmental standards worldwide, or will we have to sacrifice environmental concerns in order to feed everyone?

As we consider questions like these, we will see how *quantitative information* and *quantitative methods*, in other words mathematics, can help us better understand what is happening, and what may happen in the future.

George McGovern

The Biographical Directory of the United States Congress provides the following information on George McGovern.[4] McGovern served in the U.S. Senate from 1963 to 1981, and the U.S. House of Representatives from 1957 to 1961.

[4] http://bioguide.congress.gov/scripts/biodisplay.pl?index=M000452

> McGOVERN, George Stanley, a Representative and a Senator from South Dakota; born in Avon, Bon Homme County, S.Dak., July 19, 1922; attended the public schools of Mitchell, S.Dak., and Dakota Wesleyan University, 1940-1942; enlisted in the United States Army Air Corps in June 1942, flew combat missions in the European Theater, and was discharged from the service in July 1945; returned to Dakota Wesleyan University and graduated in 1946; held teaching assistantship and fellowship at Northwestern University, Evanston, Ill., 1948-1950, receiving his Ph.D. from that university in 1953; professor of history and government at Dakota Wesleyan University 1950-1953; executive secretary of South Dakota Democratic Party 1953-1956; member of Advisory Committee on Political Organization of Democratic National Committee 1954-1956; elected as a Democrat to the Eighty-fifth and Eighty-sixth Congresses (January 3, 1957-January 3, 1961); was not a candidate for renomination in 1960, but was unsuccessful for election to the United States Senate; appointed special assistant to the President January 20, 1961, as director of the Food for Peace Program, and served until his resignation July 18, 1962, to become a candidate for the United States Senate; elected to the United States Senate in 1962; reelected in 1968 and again in 1974, and served from January 3, 1963, to January 3, 1981; chairman, Select Committee on Unmet Basic Needs (Ninetieth Congress), Select Committee on Nutrition and Human Needs (Ninety-first through Ninety-fifth Congresses); unsuccessful candidate for reelection to the U.S. Senate in 1980; unsuccessful candidate for Democratic presidential nomination in 1968 and 1984; unsuccessful Democratic nominee for President of the United States in 1972; lecturer and teacher; U.S. Ambassador to the United Nations Food and Agricultural Agencies in Rome, Italy, 1998-2001; United Nations Global Ambassador on World Hunger, 2001-.

Acknowledgements

This book owes a great deal to *Understanding Our Quantitative World*, published by the Mathematical Association of America, and its authors Janet Andersen and Todd Swanson of Hope College in Holland, Michigan. The incorporation of social issues, collaborative learning, and an emphasis on reading in this text are all informed by their work. I am very grateful to Janet and Todd for their help and encouragement in developing this text.

I also would especially like to acknowledge Janet for our many years of friendship, the many conversations we had during the 2003-2004 academic year while I was on sabbatical at Hope College, and for her help in arranging for my sabbatical appointment at Hope.

Very tragically, Janet was killed on Thanksgiving Day of 2005 in an automobile accident. I am dedicating this work to her. She has been greatly missed by many people, including myself, and leaves not only the mathematical community, but also our larger human family, poorer in her absence. She was a truly outstanding person.

I am grateful for the extensive support I have received from the National Science Foundation. Firstly, NSF support through the AIRE program allowed work to begin on *Algebra: Modeling Our World* during my 2003-2004 sabbatical year. Secondly, the NSF has supported continuing work on this project through a grant under the Course Curriculum and Laboratory Improvement - Educational Materials Development program (CCLI-EMD award number 0442979) to Dakota Wesleyan University, for which the author is the Principal Investigator. I would like to acknowledge the help that Todd Swanson, former students Nikki Hobbie and Kristi Kogel, and also my colleagues Dr. David Mitchell and Mr. Kevin Lein have provided this project. In addition, I would like to acknowledge the suggestions and input of many students who participated in pilot sections using this text at Dakota Wesleyan.

Finally, I would like to acknowledge Dr. Rocky VonEye and Dr. Mike Farney for their incredible support and wonderful friendship during my years at Dakota Wesleyan. I don't think I could have better mathematical colleagues.

2. Representing Quantitative Information

Numeracy is a proficiency which is developed mainly in mathematics but also in other subjects. It is more than an ability to do basic arithmetic. It involves developing confidence and competence with numbers and measures. It requires understanding of the number system, a repertoire of mathematical techniques, and an inclination and ability to solve quantitative or spatial problems in a range of contexts. Numeracy also demands understanding of the ways in which data are gathered by counting and measuring, and presented in graphs, diagrams, charts and tables.
— from *National Numeracy Strategy* published by the United Kingdom's Department for Education and Skills[5]

We see quantitative information in a variety of forms and contexts nearly every day. In this chapter, we will consider two ways in which quantitative information is commonly presented, namely as graphs and as tables. Each of these has its own advantages and disadvantages. Employing more than one form at a time often greatly enhances our understanding of a particular situation and our ability to solve problems. In Chapter Four, we will consider two additional ways to represent quantitative information, in symbols or equations, and in verbal statements.

As will be the case throughout the book, we will often introduce concepts via real-world examples.

Energy Supplies

Our ignorance is not so vast as our failure to use what we know.
— M. King Hubbert[6]

Energy issues are almost constantly in the news, and are now even becoming the subject of Hollywood films (e.g. *Syriana*, released in 2005 and starring George Clooney). The price at the pump for gasoline (nearing a national average of $4 per gallon in May of 2008), the cost of heating during a hard winter, the diminishing supply of fossil fuels, and battles between environmentalists and the energy industry over drilling in environmentally sensitive areas are all recurrent themes in political discussions.

Energy supply and cost is increasingly an issue with respect to hunger. During the 20th century, agriculture evolved from an activity predominantly dependent upon nature and human labor into a largely technological enterprise dependent on academic research, chemical inputs, and a high use of fossil fuels. As developing countries modernize their agricultural practices, they become dependent on fossil fuels and put greater demands on the world energy supply. As known reserves of oil and other fossil fuels are used up, prices are likely to increase, making it harder for farmers, especially smaller farmers, to remain economically viable. How much longer can we expect the earth's fossil fuel supplies to last?

In 1956, Dr. M. King Hubbert predicted that the volume of oil produced annually in the U.S. would increase more slowly over the coming 15 years, and peak around 1970. After this point, annual oil production would begin decrease. At the time, his prediction was greeted with deep skepticism, even ridicule. Figure 1 shows a graph of the actual data, from 1954 to 2002.[7]

In the graph, we can visually estimate the total production for any given year. For example, in 1970 the production was approximately 9,900 thousand barrels, or equivalently 9.9 million barrels, per day. We would say that the **value** of U.S. Oil Production in 1970 was about 9,900 thousand barrels per day.

From the graph, we can not only see that Hubbert's prediction was dead on, but also get an idea of how large that maximal production was, and how much production has dropped off since then.

[5]http://www.standards.dfes.gov.uk/primary/publications/inclusion/1152065/

[6]This quote and other information on Hubbert's work is from www.hubbertpeak.com

[7]The data on which this graph is based is from the Energy Information Administration of the Department of Energy at http://www.eia.doe.gov/emeu/aer/txt/ptb0502.html.

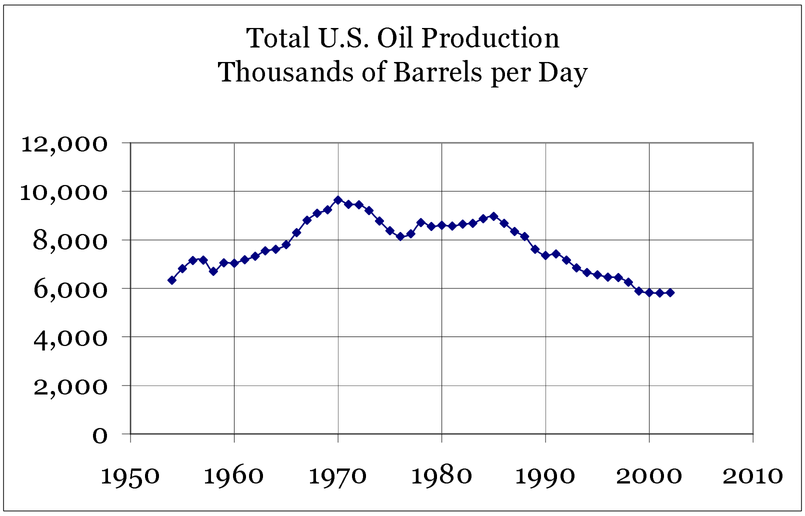


Figure 1: Total U.S. Production of Crude Oil

Hubbert made his forecast assuming that oil is a finite resource, and that production starts at zero and increases over time until a peak production level is reached. Once the peak has been passed, production declines until the resource is depleted. Since the 1950's, so-called Hubbert curves have been used to study production in every oil-producing country, the world at large, as well as for other finite resources (e.g. coal). Here is another production graph, which predicts that world oil production will peak around 2010.[8]

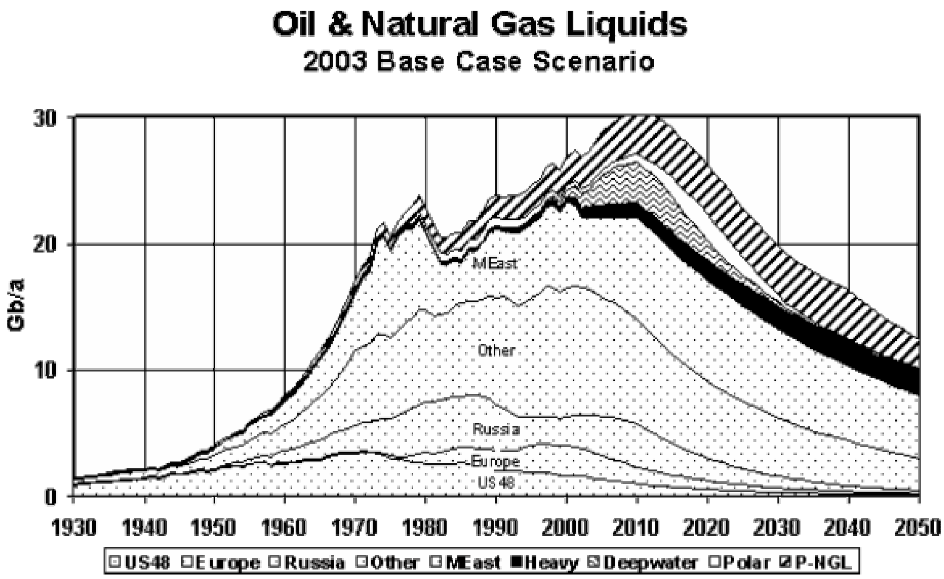


Figure 2: World Oil and Gas Production

[8]Graph from the The Association for the Study of Peak Oil and Gas (ASP), http://www.asponews.org/.

One of the implications of Hubbert's work is that, once we reach the peak, we will have used approximately half of the earth's total oil reserves. If the peak occurs in 2010, it will have taken about 150 years to go from zero production to depleting half of the resource.[9] By comparing production estimates with current trends in energy consumption, we will be able to estimate how long the remaining half of the reserves might last. Even without any data, we can reasonably assume that we are using energy at a much greater rate today than in years past, and so there is some reason to believe reserves will not last even close to 150 years.

These graphs provide us with an easily understandable picture of U.S. and world oil production. One could certainly use these to make the case that a global energy crisis looms in the not too distant future. The importance and usefulness of graphs like these are that they greatly assist us in understanding quantitative information, and reaching conclusions based on that information.

> *A whole essay might be written on the danger of thinking without images.*
> — Samuel Taylor Coleridge

What Makes a Good Graph?

Not all graphs are created equal. Good graphs are constructed with several considerations in mind.

A graph should indicate the variables which are being displayed. In Figure 1, Total U.S. Oil Production is indicated as the **dependent variable**, and one can infer from the graph that the **independent variable** is time measured in years. Unless otherwise noted, the dependent variable values are indicated along the vertical axis and the independent variable is indicated along the horizontal axis. This is standard for two-variable graphs. We could have also titled the graph "U.S. Oil Production Versus Time" or "U.S. Oil Production over Time" to indicate time as the independent variable and production as the dependent variable. Sometimes it is useful to label each of the axes with the appropriate variable. In Figure 2, the vertical axis is labelled Gb/a, which is short for "gigabarrels annually" or "billions of barrels per year."

The scale on both the vertical and horizontal axes is given and is uniform. In other words, marks are made at regular, evenly-spaced intervals along each axis and each mark represents the same number of units. Note that the scale and units on one axis could certainly be different than that on the other axis. In Figure 2, the scale on the horizontal axis is ten years, and the scale on the vertical axis is 10 Gb/a.

Choosing an appropriate **window** for the graph is important. The window is given by the minimum and maximum values for both the horizontal and vertical scale. In Figure 1, we would say the window is $1950 \leq \text{Year} \leq 2010, 0 \leq \text{Oil Production} \leq 12,000$. The **horizontal baseline** of the graph is 1950, and the **vertical baseline** is 0. The main goal in choosing a window is to allow the general characteristics of the data (or equation) depicted to be easily seen. All the data values should be included without including a huge amount of empty space. In Figure 1, we could have used a vertical baseline of 4,000 instead of 0, as this still would have easily encompassed all of our data.

Types of graphs

The graphs in Figures 1 and 2 are called a **line graphs**. Such graphs consist of a (usually unbroken) "line" or curve. Although "line" usually implies straightness, in this context it does not. Line graphs can also include designated points which are connected by the line, as in Figure 1. A graph can certainly include several line graphs at the same time, as in Figure 2, where we have separate line graphs for oil, gas, coal, total production, etc.

Figure 3 shows an example of a **bar graph**. The independent variable is time, and the dependent variable is the percentage of U.S. citizens who report ever having used marijuana. Bar graphs are often used when there are not too many values for the dependent variable.

[9] According to data provided by Jean Laherrre at the ASPO website, the first oil wells in the U.S were drilled around 1860.

For example, if we had used a bar graph instead of a line graph to represent U.S. Oil Production depicted in Figure 1, we would have had to choose between having nearly 50 bars, one for each year, or skipping some years so as to reduce the number of bars.

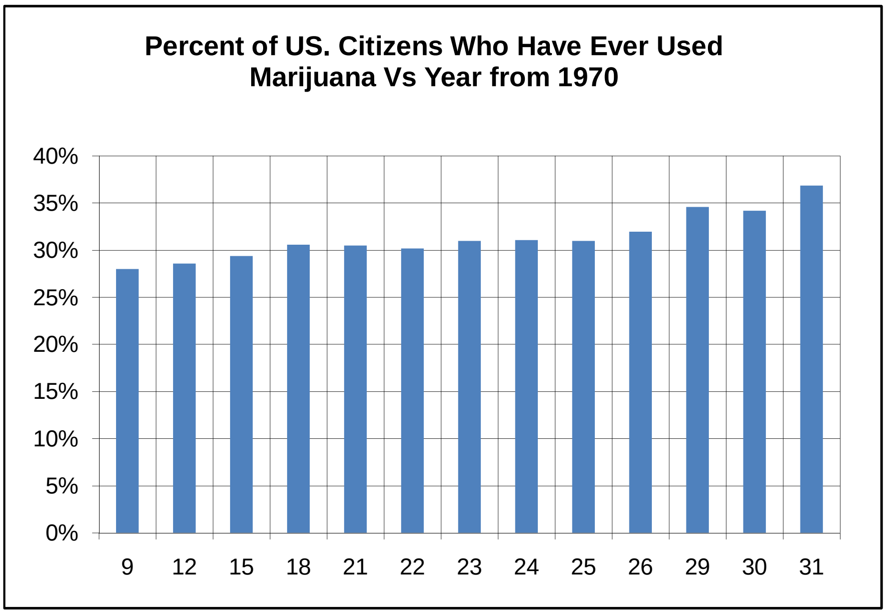


Figure 3: A bar graph of Marijuana Use in the U.S.

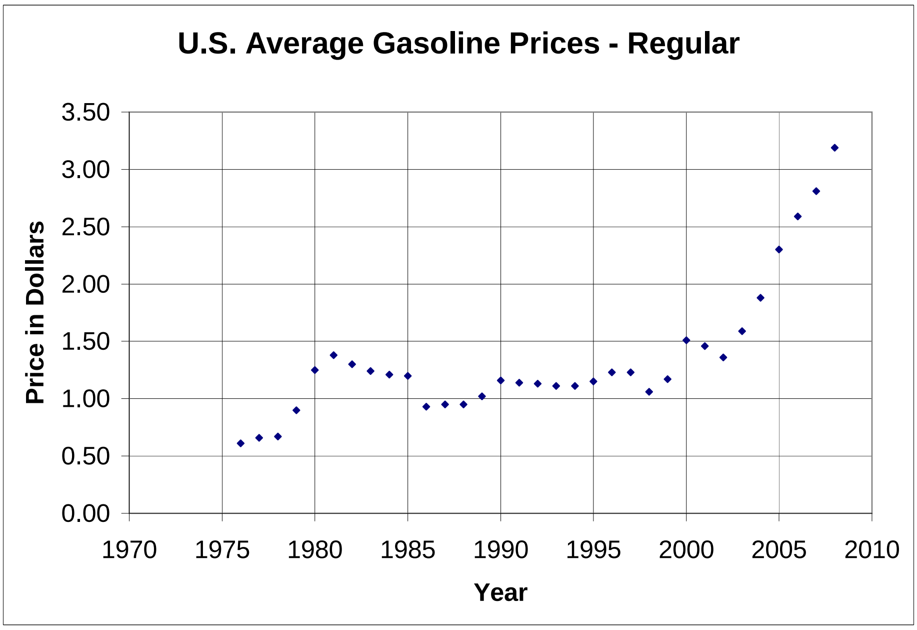


Figure 4: Graph of U.S. gasoline prices

Finally, Figure 4 depicts what is called a **scatterplot**. As in Figure 1, the independent variable is years. Here, instead of "connecting the dots" with a line, we simply plot one point for each year. The lack of a line does not keep us from seeing any general trends in what is happening. Note that the window is $1970 \leq \text{Year} \leq 2010, 0 \leq \text{Gas Prices} \leq 3.50$.

Representing Quantitative Information Numerically

When a set of data is presented in the form of a table, we refer to this as a **numerical** representation. For example, Table 1 provides the same information on gasoline prices as in Figure 4, but in numerical form.[10]

Table 1: Unleaded Gasoline Prices in the U.S.

Year	1976	1977	1978	1979	1980	1981	1982	1983	1984
Price	0.61	0.66	0.67	0.90	1.25	1.38	1.30	1.24	1.21
Year	1985	1986	1987	1988	1989	1990	1991	1992	1993
Price	1.20	0.93	0.95	0.95	1.02	1.16	1.14	1.13	1.11
Year	1994	1995	1996	1997	1998	1999	2000	2001	2002
Price	1.11	1.15	1.23	1.23	1.06	1.17	1.51	1.46	1.36
Year	2003	2004	2005	2006	2007	2008			
Price	1.59	1.88	2.30	2.59	2.81	3.19			

While the associated table gives the precise information, the graph gives you a better picture of how prices have changed over time. A main advantage of describing data in graphical form is that it allows you to observe trends and see overall behavior quickly and easily, much more so than tables or equations. On the other hand, tables often take up less space. Finally, in many cases quantitative information starts as data in numerical form. In this form, it can be efficiently stored or transmitted electronically, transported between different software packages, and easily manipulated or statistically summarized, as we will see in the next chapter. For example, the gas prices data can be stored numerically using 4 kilobytes (kb) of computer space, while the graph shown takes up over 28 kb.

Interpreting Quantitative Information: An Example

Displaying data in tables or graphs is not an end in itself. Rather, we display data in order to help us answer questions, reach conclusions, or solve problems. Often, we collect and display data in order to answer a particular questions. Here is one example.

> To what extent do countries with a large percentage of poor people also have a large percentage of hungry people?

We might reasonably guess that the answer to this question is "to a (very) large extent." This question concerns the relationship between two variables. Table 2 shows these two variables, along with the country labels for the data. The first variable, denoted Poverty Rate in the table, is more precisely the National Poverty Rate given as the percent of the total population defined as living in poverty in that country.[11] The second variable is the percentage of children under five who are defined as being severely underweight for their age. This percentage is one of several common ways to measure the extent of malnutrition in a country, or among children in a country. [12]

[10]Data from the Energy Information Administration at http://www.eia.doe.gov. Data for 2008 is the average of monthly averages through April.

[11]Poverty Data is available from the World Bank.

[12]Malnutrition Data is from UNICEF, http://www.childinfo.org/eddb/malnutrition/index.htm

Table 2: National Poverty Rates and Percentage of Severely Underweight Children Under Five.

	Pov. Rate	Underwt.		Pov. Rate	Underwt.
Albania	25.4	4.3	Kazakhstan	34.6	0.4
Algeria	12.2	1.2	Kenya	42	6.5
Armenia	53.7	0.1	Kyrgyz Republic	64.1	1.7
Azerbaijan	49.6	4.3	Lao PDR	38.6	12.9
Bangladesh	49.8	13.1	Lesotho	49.2	4
Benin	33	7.4	Madagascar	71.3	11.1
Bolivia	62.7	1.7	Malawi	65.3	5.9
Bosnia and Herzegovina	19.5	0.6	Malaysia	15.5	1.2
Brazil	17.4	0.6	Mauritania	46.3	9.2
Burkina Faso	45.3	11.8	Mauritius	10.6	2.2
Burundi	36.2	13.3	Mexico	10.1	1.2
Cambodia	36.1	13.4	Mongolia	36.3	2.8
Cameroon	40.2	4.2	Morocco	19	1.8
Chad	64	9.8	Mozambique	69.4	9.1
Colombia	64	0.8	Nepal	42	12
Costa Rica	22	0.4	Nicaragua	47.9	1.9
Cote d'Ivoire	36.8	4	Niger	63	14.3
Djibouti	45.1	5.9	Nigeria	34.1	10.7
Dominican Republic	28.6	1	Pakistan	32.6	12.8
Ecuador	35	1.9	Peru	49	1.1
Egypt	16.7	2.8	Romania	21.5	0.6
El Salvador	48.3	0.8	Rwanda	51.2	7.1
Eritrea	53	17	Senegal	33.4	4.1
Ethiopia	44.2	16	Sierra Leone	68	8.7
Georgia	11.1	0.2	Sri Lanka	25	4.8
Ghana	39.5	5.2	Tanzania	35.7	6.5
Guatemala	56.2	4.7	Togo	32.3	6.7
Guinea	40	5.1	Tunisia	7.6	0.6
Guinea-Bissau	48.7	5.2	Uganda	55	6.7
Honduras	53	4	Ukraine	31.7	0.5
Hungary	17.3	0.2	Uzbekistan	27.5	5
India	28.6	18	Venezuela, RB	31.3	0.7
Indonesia	27.1	8.1	Vietnam	50.9	5.8
Jordan	11.7	0.5	Yemen, Rep.	41.8	14.5
			Zimbabwe	34.9	1.5

Perhaps the first thing one notices about this table is that this is a lot of data! In fact, there are 69 countries listed, with the data given in alphabetical order by country. Both the amount of data and the way it is organized in the table make it a little hard to make any general conclusions. We could certainly pick out of few countries and see if those with a low(high) Poverty Rate also have a low(high) Underweight Rate. However, this is a little awkward and we could also easily reach the wrong conclusion depending on which countries we happened to pick. What we need is a quick and easy way to see if our assumption about the effect of poverty on malnutrition really seems to be true. Here we can use a scatterplot to good advantage.

Recall that standard practice is to put the independent variable on the horizontal axis and the dependent variable on the vertical axis. In this example, we would probably think of malnutrition as depending on the poverty rate. Given this, we would have Poverty Rate as our independent variable and put it on the horizontal axis and Underweight Rate as the dependent variable and put it on the vertical axis.[13]

A scatterplot of the data from Table 2 is shown in Figure 5. From this scatterplot, you can see that, although the points are widely scattered, it does seem the points are generally getting higher the further to the right you go. This is an example of a positive association. Two variables have a **positive association** when an *increase* in the value of the independent variable leads to an *increase* in the value of the dependent variable. In our situation, this means that countries with higher poverty rates tend to have more malnutrition. Since there is a lot of scatter, the positive association is **weak**.

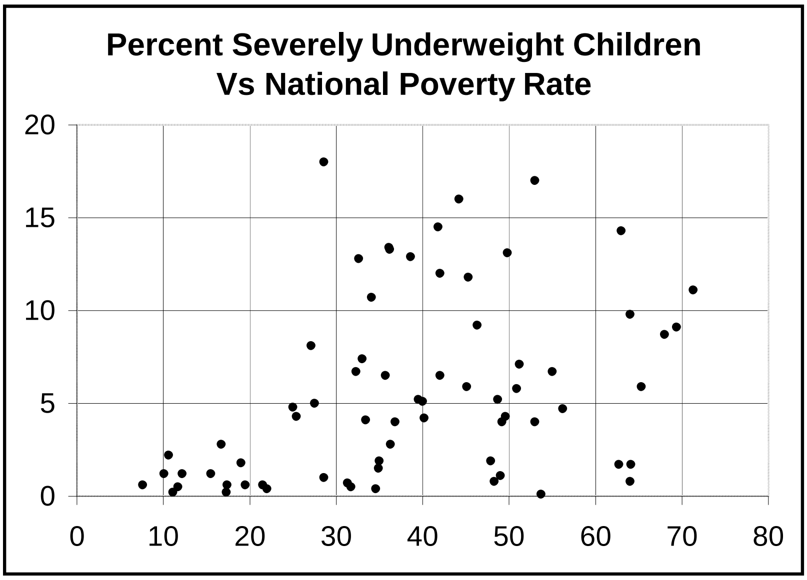


Figure 5: A scatterplot of Percent of Severely Underweight Children Versus Poverty Rate.

What more could we say about our poverty and malnutrition data from the scatter plot in Figure 5? Spend a couple minutes considering the plot. You may want to look at particular points or groups of points and see if you can find the corresponding countries in the table. What observations do you make?

(You may want to refer back to this example when doing the activities for this chapter)

Here are a few examples of observations that one might make. This is not meant to be a complete list, or even a list of "what you should have observed." The observations are simply meant to give

[13]It is not always clear in all cases which of two variable should be the dependent and which the independent. Often it is a judgement call, but at other times there may be a fairly obvious choice.

you examples of the types of things you might look for when considering a scatterplot.

One observation is that the data almost seems to fit into sort of a triangular or pie-piece pattern, except for a single point (the highest point in the plot). There are points clustered close to the horizontal axis all through the plot, and there is also a very linear increasing pattern along the top of the plot. Perhaps we could say the following.

> A low poverty rate implies a low malnutrition rate, but a high poverty rate does not necessarily mean a high malnutrition rate, as many countries with higher poverty rates also have low malnutrition rates.

Going back to the high point of the graph, we might wonder what is special about this country. The point seems "far away" in some sense from the rest of the points, and we might classify it as an **outlier**. With a little searching, you could see from the table that this country is India. Why would India have such a high malnutrition rate, especially as its poverty rate is not that high? You might know that India is the second most populous country on earth (after China). Could the large population be having an effect? We could look for other populous countries in our table. Unfortunately, the data we have does not include all countries, and in particular, China, the United States, and the Russian Federation are all populous countries that are not included. Another speculation might be that the data for India is not accurate. We might want to know more about how the data was collected. Did UNICEF collect the data, or did some other organization or the individual countries themselves supply the data? We would be jumping to a conclusion if we claimed that the data must be inaccurate just because the given point seems to be an outlier, however, it is sometimes the case that outliers are the result of errors of one sort or another.

One might also notice that the countries with poverty rates below about 25% seem to form their own group. All of these countries have low malnutrition rates, and there is sort of a "gap" between these points and the rest of the points. If you ignore these countries (say, by covering them up when you look at the scatterplot), and look only at the higher poverty countries, then there seems to be almost no association between the two variables. As a hypothesis, we might conjecture that 25% is a sort of "cutoff point". Countries that are below this point in poverty have low malnutrition, and for countries above this point, the poverty rate does not seem to have much effect on how much malnutrition there is.

There are certainly other observations one might make. At this point, we are only making some initial observations and also bringing up further questions for consideration. As we proceed through the course, we will discuss how our observations can be made more precise, more "quantifiable", and how we can better decide which conclusions we can have high confidence in and which are more speculative.

Negative and Positive Associations

It was previously stated that the graph in Figure 5 had a positive association since the points in the graph tended to get higher as we go to the right. In other words, the dependent variable tended to increase as the independent variable increased.

Two variables have a **negative association** when an *increase* in the value of the independent variable leads to a *decrease* in the value of the dependent variable. An example of this is the relationship between latitude and average January temperature for cities in the United States. We would think of temperature as depending on the latitude, so we graph the independent variable, latitude, along the horizontal axis. As the latitude increases (moving farther north), the average January temperature decreases. Figure 6 graphs the latitude and average January temperature for twelve cities in the eastern United States.[14] Since there is a lot less scatter in this graph than in Figure 5, this graph shows a much **stronger** association.

To reiterate, one difference between Figure 5 and Figure 6 is that the points in Figure 6 seem to closely follow a "linear" pattern, while the points in Figure 5 have much more "scatter" to them. We would say that the relationship between temperature and latitude is **strong** while the relationship between poverty and malnutrition is **weak**, or at least much weaker than the temperature and latitude relationship. In Chapter Five, we will make this idea of strong and weak more precise. For

[14]The data was gathered from *The World Almanac and Book of Facts 1996*, pp 180, 594, 595.

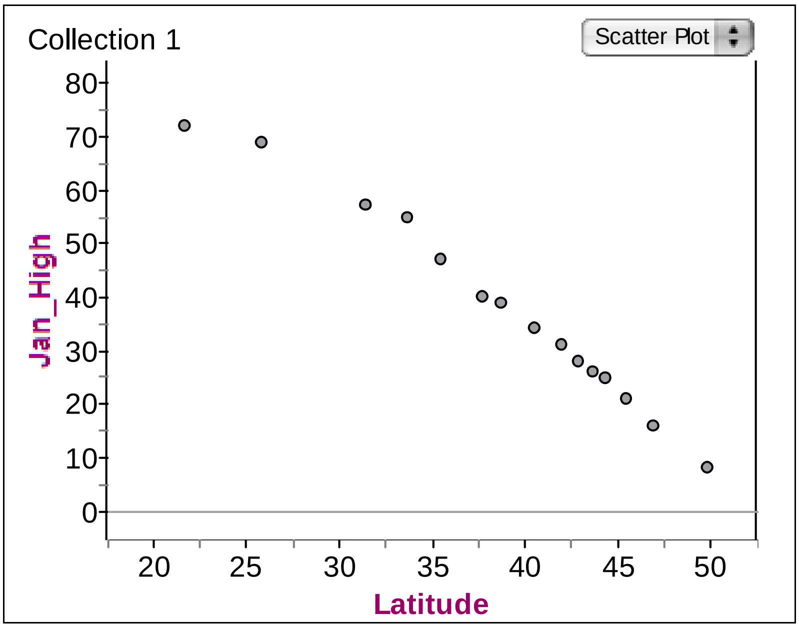


Figure 6: A scatterplot of latitude and average January temperatures for cities in the eastern United States.

now, the idea is that the more closely the points in a scatter plot cluster around a straight line, the stronger the (linear) relationship between those variables.

Displaying and Interpreting One-variable Data

Scatterplots, like the one for malnutrition versus poverty in Figure 5, help us investigate the relationship between two variables. Often, we wish to consider one variable at a time. As with two-variable data, the first step is often deciding on the best way or ways to present or display the data so that the information contained can be easily understood. In this section, we will look at what types of graphs and tables are most appropriate for displaying various types of one-variable data sets.

The data set listed in Table 3 gives the percentage of females of the appropriate age group who completed primary school (elementary school) in 23 countries. Each individual number given is called a **data point**.

This data involve just one *numerical* variable, the percentage completion rate. The data are listed from the lowest percentage country to the highest. There is no reason the data have to be ordered in this way. Another reasonable choice would be to list the countries alphabetically. The names of the countries are examples of what are called **(data) labels**.

Spend a minute or two looking over the data in Table 3. Make notes of any particularly noteworthy aspects of the data. What do the data seem to tell you about the state of education throughout the world? Are there any questions that arise in your mind about these data?

There are many things we could say about this data set.[15] Here are a few examples.

The percentage of girls who complete primary school goes from a low of under 10% to over 100%. This is a very wide **range**, especially given the relatively small number of countries.

One question we might have is "how is it possible to get over 100%?" It would be beneficial

[15]Data is from the World Bank data base

Table 3: Completion Rates for Primary (Elementary) School, Females, Year 2000

Country	Primary Comp. Rate	Country	Primary Comp. Rate	Country	Primary Comp. Rate
Chad	8.53	Nepal	58.35	Mexico	93
Guinea-Bissau	23.5	Congo, Rep.	59.68	Guyana	95.15
Guinea	24.27	Nigeria	60.66	Korea, Rep.	97.78
Sierra Leone	30	Laos PDR	64.23	Jamaica	98.3
Senegal	34.11	Nicaragua	70.04	Uruguay	101.12
Congo, Dem. Rep.	34.21	Bangladesh	71.67	Jordan	105.59
Yemen, Rep.	38.28	Trinidad & Tob.	84	Dominica	107
Cambodia	51.47	Indonesia	91.71		

to have a more complete description of how these percentages were computed. One possible explanation is that the number of girls who completed primary school in a given year is more than the number of girls who are at the age where they should be completing school because some girls have taken longer to complete school and are older than the usual completion age.

We might also notice that there are only 23 countries listed. What about all the other countries? Most of the countries are from Africa and Asia. There are no European countries, and very few industrialized countries (like the U.S., Great Britain, Japan, Russia, etc.). Were these countries intentionally left out by the World Bank, or was data simply not available for these countries? Because this group of countries does not seem to have been randomly selected, and is not representative, we cannot use this data to produce (say be averaging) a valid measure of completion rates for the world as a whole.

Notice that it looks like more than half of the countries have completion rates over 50%, so there are fewer countries with very low completion rates.

As with the example discussing poverty and malnutrition, the discussion above is not meant as a complete "right answer" to what you might notice about this data set. Rather, it is meant to illustrate a way of thinking about data, as well as an attitude of questioning. If you did not notice these aspects of the data, or did not ask the same questions, that is OK for now. As you progress through the course, you should gain skill in becoming more thorough at this kind of analysis and develop a more questioning attitude regarding data. The important thing for now is to begin to reflect on how you can think about, and learn from, numerical data.

Frequency Distributions

In our examples with two variable data, we found the visualization provided by a scatterplot very helpful in understanding the data. Similarly, a graph or some other way of summarizing the data in Table 3 might be helpful.

One common method of organizing data is a **frequency distribution**, which is a table in which the data is grouped into categories (often called **classes**) and the number of data points in each of these categories is given. The numbers in Table 3 are percentages. We might choose to divide our data up in categories in 10% intervals, 0%, 10%, 20%, . . . up to say 110%.[16] Then, we count how many data points are in each category. For example, there are 3 countries that would

[16] There are other ways to group these data. For example, we could go in intervals of 15% or 20%

fit between 20% and 30%. Standard practice is for us to account values that are *strictly greater than* 20%, and less than or *equal to* 30%. Counting the number of percentages in each class and listing that number as its frequency, we obtain the **frequency distribution** shown in Table 4.

Completion Rate	Frequency
0 to 10	1
10 to 20	0
20 to 30	3
30 to 40	3
40 to 50	0
50 to 60	3
60 to 70	2
70 to 80	2
80 to 90	1
90 to 100	5
100	3

Table 4: A frequency distribution for Primary School Completion Rates.

The frequency distribution gives us a better understanding of how the completion rates are distributed. For example, we can easily see that seven of the 23 countries have rates below 50%.

Histograms

A **histogram** is a bar graph of a frequency distribution. In a histogram, one axis (usually the horizontal) is scaled according to the classes in the frequency distribution. The other axis gives the frequency in any appropriate units. For each class, there is one bar with height equal to the *frequency* for that interval. This gives us a visual representation of the number of data values in each class. The advantage of a histogram is that, with just a quick glance, we see how the data is distributed. A histogram of our primary school completion data is shown in Figure 7. Notice that each class is labelled with the lower bound for that class. So, for example, the bar with the 20 underneath represents the three countries between 20% and 30%, as noted above in the frequency distribution.[17]

Compared to the frequency chart, the histogram allows us to see patterns in the data more easily and quickly. For example, we easily see that more than half of the countries have completion rates more than 50%. In fact, by using the vertical scale, we can see there are exactly 7 that have completion rates below 50%, as we noted from the table. We also notice that there are no countries with rates between 10% and 20%, and 40% and 50%.

There are some standard terms that are used to describe histograms, and the examples below will illustrate these.

Figure 8 shows two histograms that are **skewed right**. This means that the peak is towards the left and the data trail off towards the right. As we will see later, data that are skewed right will have an average (mean) larger than the median.

The first histogram shows the distribution of child mortality rates for children under 5. The data are given in number of deaths per 1000 people in the population. We see that many countries have mortality rates on the low end, less than 50 deaths per 1000. The number of countries drops off drastically at first, and then shows an overall slow decease as the mortality rate increases.

The second histogram shows the distribution of faculty salaries, for tenured and tenure-track faculty, at the University of Maryland, a large research university. There are about 1200 faculty

[17] How histograms appear can vary with the computer software used. One unfortunate aspect of Microsoft Excel is that the scale numbers on the horizontal axis are sometimes shown in the middle of the interval, rather than at the left end, which would be more standard

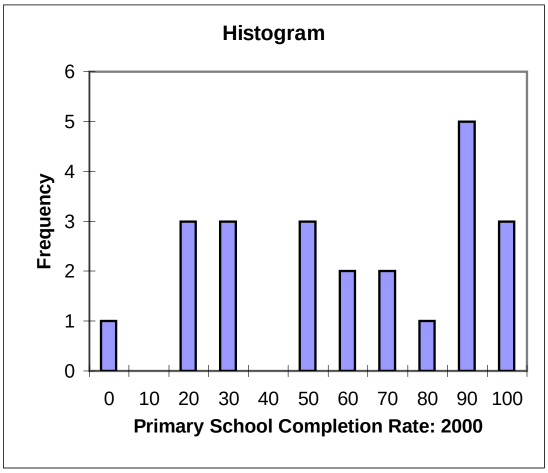


Figure 7: A histogram for Primary School Completion Rates (see footnotes).

members in the data set. The frequency has a **peak** at salaries between \$50,000 and \$60,000. There is a **gap** in the histogram towards the right, showing that there are no salaries between \$140,000 and \$160,000. The few salaries over \$160,000 would be considered **outliers**. Outliers are data points that are, in some sense, "far away" from the rest of the data. We will define more precisely what "far away" means in Chapter Three.

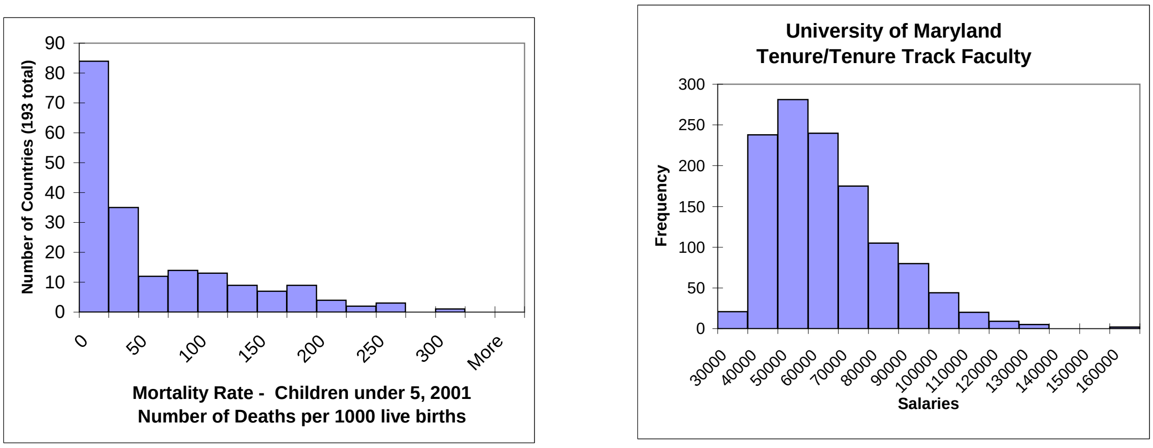


Figure 8: Two histograms that are skewed right

Figure 9 shows a histogram that is **skewed** (slightly) to the **left**. It shows the frequency for the percent of children who were ever breast fed for 62 countries. We see, for example, that there is one country where more than 99% of infants were breast fed at least once. Notice that the **(horizontal) range** for this data is only from 87% up to 100%. This is a fairly narrow range, much narrower than the range in percentages in our primary completion data. As noted in the graph, only 62 countries are represented, and so we might not consider the data representative of all countries, especially without knowing which countries are included and which are not.

Figure 10 shows two histograms which we would call roughly **symmetric**. The first depicts national poverty rates, and the second shows per capita incomes for the 50 states for 2002. The "tails" on each side of the peak are roughly the same length and overall, the histograms have about the same shape on either side of the peak.

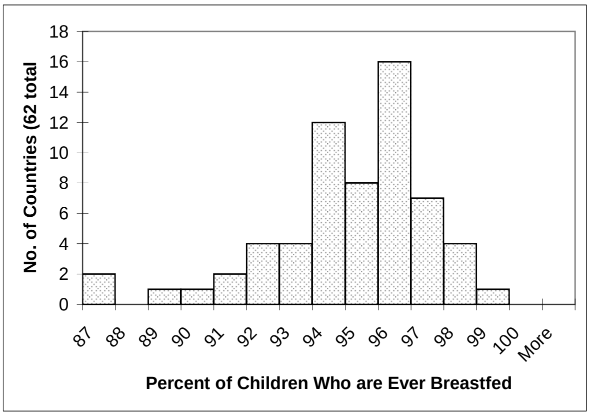


Figure 9: A histogram that is skewed slightly left

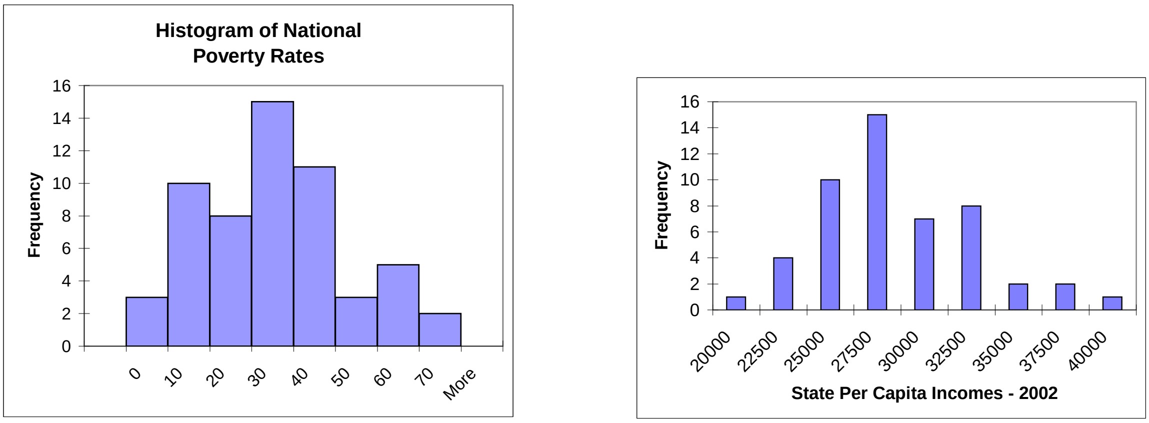


Figure 10: Two histograms that are roughly symmetric

Deciding whether a histogram is skewed or symmetric is somewhat subjective. Notice also that some of the histograms have more "up and down" from bar to bar. We might say that these histograms have some irregularity. Our first histogram on primary school completion rates might be called irregular.

Finally, Figure 11 shows a histogram that is **bimodal**, which means it has two well-defined peaks. Not all bimodal histograms have the peaks at the extremes. The data shown here are from the Polity IV Project, which has compiled data on the nature of governments for all countries over time, and given each country a Democracy Score based on the openness of the political institutions in that country.[18] At least according to this measure, countries tend towards complete democracy or complete lack of democracy.

Figure 12 shows a second bimodal histogram. This graph gives the frequency of bear weights in pounds for a sample for 143 bears.[19] There is a smaller, secondary peak to the right of the first peak. Since the sample consisted of 44 females and 99 males, the different peaks might represent the average for the two different genders (can you tell which is which?).

So far, we have discussed skewness merely as a description of a histograms shape. As we will see in the next chapter, however, the main importance of skewness is that it tells you something about the relation between **mean** and **median** of a set of data, and between these and peak of a

[18]Polity IV website is www.cidcm.umd.edu/inscr/polity/index.htm

[19]The data was collected by Dr. Gary Alt and is available in Fathom as a sample document.

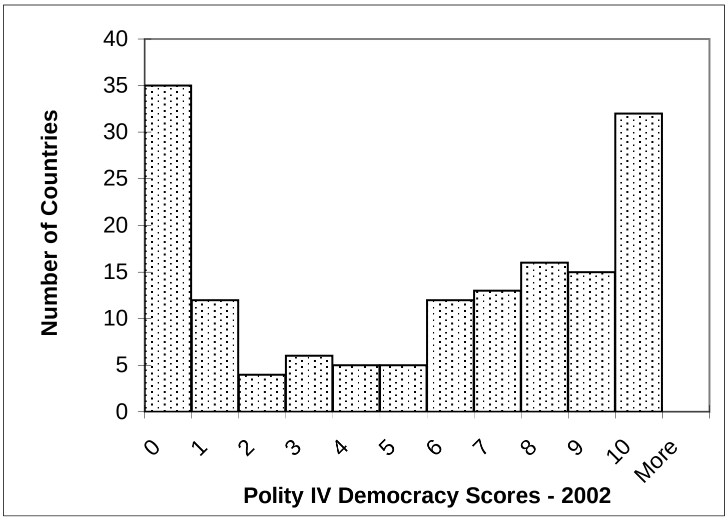


Figure 11: Democracy Scores: A bimodal histogram

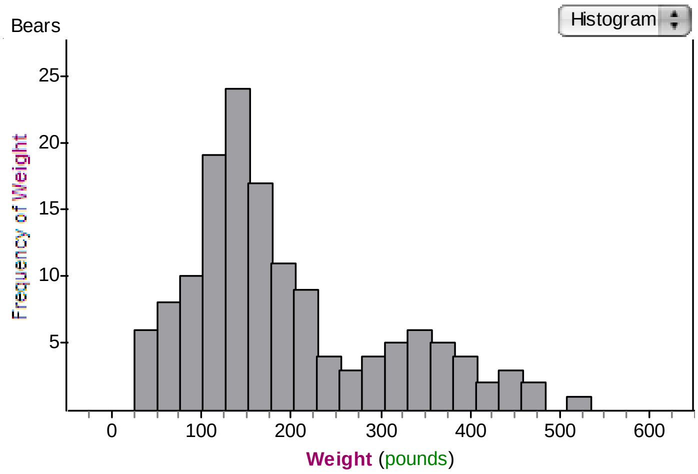


Figure 12: Bear Weights: Another bimodal histogram

histogram. For a histogram which is skewed to the right, the mean will in general be larger than the median, and the tendency will be for both the mean and median to be to the right of the peak. The opposite will occur if the histogram is skewed to the left.

The Gini Scale for Income Inequality

> *Today, there is a huge disparity between the technology, education, health care and agricultural methods that are available in the developed and developing world. The principal challenge we face is to close that gap...The countries, businesses and individuals that are on the right side of the divide have to think hard about what kind of world they want us all to live in 20 years from now. Narrowing the gap benefits everyone, and we have*

the means to do it. If we don't, we will have missed an amazing opportunity.[20]
— Bill Gates

We conclude this chapter by introducing the Gini Coefficient, which is designed to measure inequality. In particular, we will use Gini coefficients to measure income inequality in particular countries, regions, or populations. Here's how it works. Consider Table 5.

Cumulative % of Pop.	10	20	30	40	50	60	70	80	90	100
Total % of Income	1	2	4	7	14	25	49	58	78	100

Table 5: Percentage of cumulative income versus cumulative percentage of population

Table 5 shows, for example, that if you took the bottom 10% of the population by household income, this group earns 1% of all the income earned in this country. The bottom 20% of the population earns 2% of the income, etc. Note that the bottom 50% of the population earns 14% of all the income earned in the country, which means the top half of the population earns 86% of all the income. This seems to be a country with a lot of poor people and very unequal income distribution. The data above is graphed below in Figure 13. This type of a graph of Cumulative Income vs Cumulative Population is called a **Lorenz Graph** or Lorenz Curve.

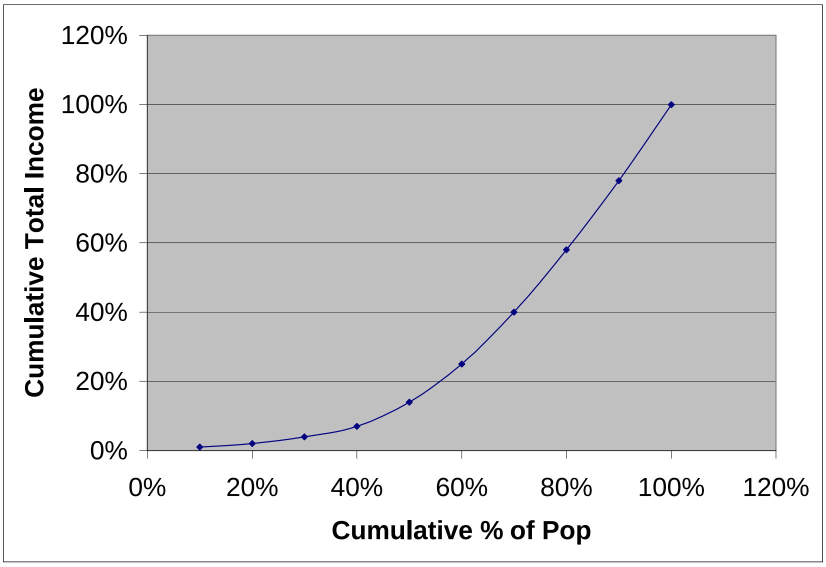


Figure 13: Lorenz graph for a society with significant inequality in incomes

Table 6 shows another example of an income distribution table.

Table 6 depicts the ideal egalitarian society. Every household earns exactly the same income!! We can tell this since the bottom 10% of society earns exactly 10% of all the income, the bottom 20% earns 20% of the income, etc. This means, for example, that the "second 10%" (those in the bottom 20% but not in the bottom 10%) must earn 10% of all the income, just like the bottom 10%. The Lorenz curve for this line would be a straight line, going from $(0\%, 0\%)$ to $(100\%, 100\%)$.

In Lorenz graphs, the closer the curve gets to the bottom right corner, the more inequality exists in the population under consideration.

Figure 14 shows graphs for both Tables 5 and 6 together on the same coordinate axes.

[20]from P. Hofheinz (2001, January 29). Gates on technology, AIDS and why Malthus was wrong. Retrieved October 3, 2007 from ZDNet Web site: $http://news.zdnet.com/2100-9595_2 2-527652.html$

Cumulative % of Population	Total Income Earned by Pop
10	10
20	20
30	30
40	40
50	50
60	60
70	70
80	80
90	90
100	100

Table 6: Table for a completely egalitarian society

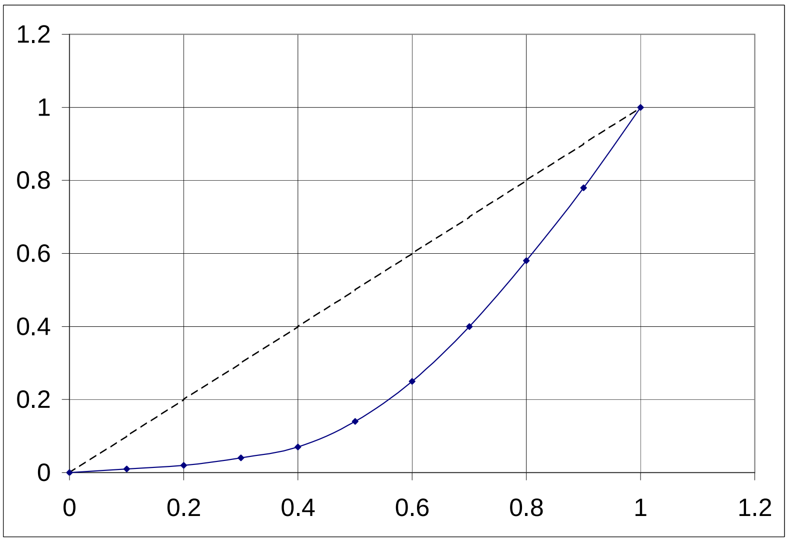


Figure 14:

It is reasonable to think that the bigger the area between the dashed line, representing total equality, and the given Lorenz graph, the more unequal the society. We define the Gini Coefficient to be twice this area between the two graphs.

The smallest this area can be is zero. What is the largest it can be? Well, imagine a country where one person earns all the income and everyone else earns nothing at all. What would the Lorenz graph look like?

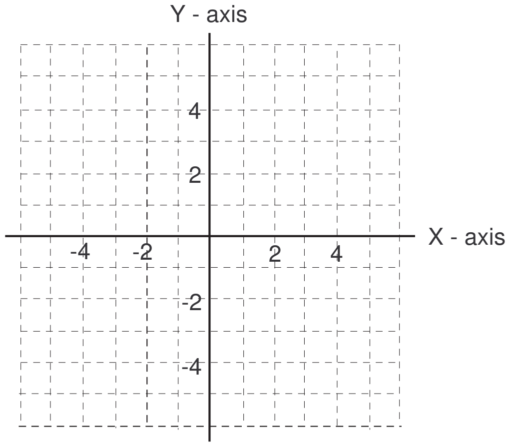
Note that the total area under the dashed line is the area of a triangle with base equal to 1 and height equal to 1 (for 100%). So, the area of this triangle is 1/2 which would mean a Gini Coefficient of 1 (remember the coefficient is defined as twice the area). This is the largest a Gini Coefficient can be. You can think of the 1 as representing 100% inequality.

Using Figure 14, we can make a rough estimate of the Gini Coefficient for this population. A little less than half of the triangular area under the line is between the line and the curve. Note that the area of each grid square is $0.2^2 = 0.04$. By adding up squares and estimating areas of partial squares, we get an estimate of the area between the line and the curve of about 0.23. This would give a Gini Coefficient of about 0.46 or 46%. In fact, it happens the the Gini Coefficient for the United States in recent years has been around 0.45.

Reading Questions for Representing Quantitative Information ______

1. Plot each of the given points on the grid provided. Label the point with its corresponding letter.

A.(−1,4) B.(0,−3) C.(1,−5)
D.(−2,1) E.(5,2) F. (−2,0)



2. Consider Figure 2 in the reading, which depicts World Oil and Gas Production.

(a) What is the window for this graph?

(b) The graph shows not only total world production, but also breaks it down into various countries and regions. Although one has to do a little interpretation, you can estimate when these other countries and regions reach their peak productions. Based on the graph, estimate when Europe, Russia, and Other have or will reach their peaks. Explain how you know.

(c) From the graph, estimate what percentage of total oil production was produced by the MidEast in 1970 and explain how you know. Do the same for 2000 and 2050.

(d) The graph indicates that the U.S. oil supply will be essentially depleted by 2035 and Russia's by 2050. From the graph, estimate which of these two countries will have produced the greater amount of oil. Explain how you can tell?

3. Consider Figure 3 in the reading, the graph on marijuana use.

(a) What is the window for this graph?

(b) Are there any characteristics of good graphs that this graph violates?

4. If you were to graph each pair of variables listed below, which variable would most appropriately go on the horizontal axis and which one would go on the vertical axis?

(a) The amount of gasoline you put into a car *and* the cost of the gasoline.

(b) The time it takes a car to accelerate from 0 to 60 mph *and* the horsepower of the car's engine.

(c) Day of the year *and* amount of daylight (i.e. time from sunrise to sunset).

(d) A child's height *and* the child's age.

(e) The year *and* the cost of tuition for a full-time student at your institution.

(f) The number of words in the vocabulary of an eleven year old *and* the number of hours the child's parents have spent reading to him or her over his or her life.

(g) The spending per pupil in public school in a state *and* the high school graduation rate in that state.

5. Suppose you were to redraw the graph from Figure 1 of the text, changing the window to $1952 \leq \text{Year} \leq 2002, 5500 \leq \text{Oil Production} \leq 10000$. How would this change the impression created by the graph?

6. If someone wants to graphically de-emphasize a sudden increase or decrease, they will sometimes make the vertical range of the graph window extend well beyond the range of the data. For example, suppose a school superintendent used a graph to represent the change in student standardized test scores from 1990 to 2002. The average standardized test score was 85 in 1990 and was 72 in 2002.

(a) Use a vertical scale from 0 to 90 and sketch a plausible graph showing the change in average test scores.

(b) Use a vertical scale from 70 to 85 and sketch a plausible graph showing the change in standardized test scores.

(c) Which graph gives a more noticeable decrease? Why?

7. How are a frequency distribution and a histogram similar? How are they different?

8. What are the advantages of displaying data in a frequency distribution?

9. What are the advantages of displaying data in a histogram?

10. Figure 11 in the text shows the histogram for Democracy Scores, as defined and computed by the Polity IV Project. It is an example of a bimodal histogram, and we commented in the text that, based on this shape, countries tended either towards total democracy or total lack of democracy.

(a) Suppose that instead of a bimodal histogram, the histogram for Democracy Scores was skewed to the right. What would this tell you about democracy in the world? You might draw an example histogram that is skewed right. In this case, remember that all scores are between 0 and 10.

(b) What could you say if the histogram was skewed to the left?

11. By hand, construct a histogram using the following data, which gives Gini Numbers for all European countries. Use classes of width two, starting with 20. Make a sketch of the histogram.

Country	Gini No.		Country	Gini No.		Country	Gini No.
Belarus	28.53		Italy	28.30		Romania	25.33
Belgium	26.97		Latvia	26.98		Slovak Rep.	21.51
Bulgaria	27.80		Lithuania	33.64		Slovenia	27.77
Czech Rep	23.05		Moldova	34.43		Russia	30.53
Denmark	29.06		Netherlands	29.16		Sweden	28.88
Estonia	37.22		Norway	28.91		UK (Britain)	32.35
Finland	23.55		Poland	26.68		Ukraine	25.71
Germany	26.00		Portugal	34.44		Yugoslavia	31.88
Hungary	27.87						

(a) Describe the shape of your histogram.

(b) Based on your histogram, estimate that average percent inequality for European countries.

(c) Write a few sentences describing what this histogram tells you about income inequality in Europe.

12. Answer the following questions about histograms using the data given below.

1 1 2 2 4 5 5 8 9 12 12 12

(a) Construct a histogram by hand using three groups: 0 to 4 (including 4, since our standard practice is to include the upper end, but not the lower end of each class), 4 to 8, and 8 to 12.

(b) Construct a histogram using four groups: 0 to 3, 3 to 6, 6 to 9, 9 to 12. Make a sketch of the histogram.

(c) Why do your two histograms look different? What does this tell you about constructing histograms with different groupings?

13. Consider the histogram in Figure 9 of this chapter, showing the percent of children who were ever breastfed for 62 countries.

(a) In how many countries were 95% or more of children breastfed?

(b) In what percent of countries were 95% or more of children breastfed?

(c) If you picked one of the 62 countries represented in the histogram at random, what is the probability you would pick a country where 95% or more of the children are breastfed?

(d) Does the histogram support the statement "In 95% of all countries, at least 9 out of 10 infants were breastfed." Explain why or why not.

14. Why would a scatterplot of people's heights versus their total arm spans have a positive association?

15. For each of the following, determine if the two variables would have a positive association, a negative association, or no association. Briefly explain.

(a) size of a building and cost of heating or cooling

(b) automobile's age and gas mileage

(c) size of a city and number of violent crimes

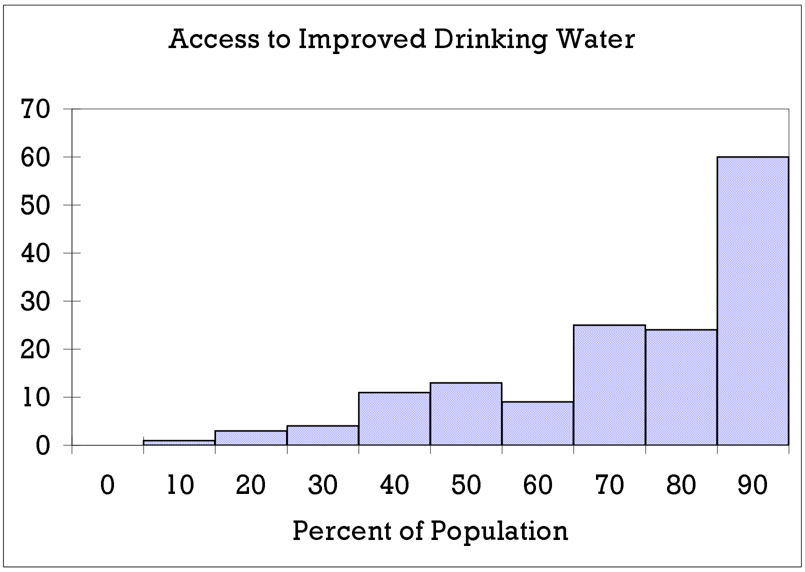
(d) number of songs on a CD and cost of the CD

(e) amount of time spent watching TV and amount of time spent studying in the library

16. If you were to graph each of the following on a scatterplot, which variable would go on the horizontal axis or does it not matter? Explain your answer.

(a) size of a building and cost of heating or cooling
(b) automobile's weight and gas mileage
(c) size of a city and number of violent crimes
(d) car's price and age
(e) number of years of education past high school and height

17. The graph below represents the percentage of the population in a number of countries who have access to improved drinking water supplies.[21]



(a) What type of graph is this?
(b) Approximately how many countries are represented in this graph?
(c) In about how many countries does at least 80% of the population have access to improved drinking water? What percent of all the countries represented is this?
(d) How would you describe the skew of this graph?
(e) How do you think this graph would have looked using data from 100 years ago? Explain why you think so.

[21] Data collected from the United Nations Conference on Trade and Development website, http://stats.unctad.org

18. Use an appropriate graph from this section to display the following data. Make a sketch of each graph.

(a) This data set represents weights (in kilograms) of 20 young women.

52	75	63	60	63	65	47	62	35	55
55	51	72	58	53	53	65	50	50	71

(b) The following table shows per capita state tax revenues, as well as the number of bachelor's degrees for each 1000 18 to 24 year olds in several midwestern states both for the year 2000.[22]

State	Per Capita Tax Revenue	Bach. Degrees per 1000 18 to 24 Year Olds
Illinois	1,835	45.7
Iowa	1,772	62.7
Kansas	1,803	53.3
Minnesota	2,711	49.2
Missouri	1,532	55.9
Montana	1,564	59.1
Nebraska	1,742	61.7
North Dakota	1,826	66.7
South Dakota	1,228	61.3
Wisconsin	2,345	52.8
Wyoming	1,952	36.0

19. Consider the histogram on child mortality rates given as part of Figure 8 in the text. Write a few sentences interpreting what this graph tells you about child morality rates. Be as thorough as you can.

[22]Data from U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System, various years; and U.S. Bureau of the Census, Population Division and the U.S. Census Bureau

Representing Quantitative Information: Activities and Class Exercises

1. **Wind Energy.** Wind energy has become more prevalent in recent years. In South Dakota, wind turbines have been built in a number of areas.

   (a) Open the Fathom file Wind Energy. You should see a collection and a case table. The variables, or attributes in the language of Fathom, are TotalMonths, Year, Month and Energy. The TotalMonths counts the number of months with January, 1998 counting as 1. The Energy variable is measured in Trillions of BTU,[23] and represents the total wind energy produced in the U.S. each month over this period of time. Create a graph of Energy versus TotalMonths. Recall that you can do this by first dragging a graph into the main window, and then dragging the appropriate columns to the vertical and horizontal axes of the graph box.

   (b) In a few sentences, describe what this graph tells you about wind energy production in the U.S.

   (c) Notice that the graph has somewhat of a wavy pattern of peaks and valleys, at least after the first year or so.

      i. Investigate when (during which part of the year) these peaks seem to be happening. What time of year do the peaks seem to be happening?

      ii. What are some possible explanations for the peaks happening at these times?

   (d) Remove the TotalMonths variable from the graph by selecting **Graph—Remove X Attribute**. The graph should change to a dot plot, which shows one dot for each month, and places that dot along the y-axis at the Energy value for that month.

      i. Use the **Object** menu to make a duplicate copy of the dotplot. You should now have two dotplots in the window.

      ii. Change one of the plots to an Energy versus Year plot. Write a couple of sentences describing what this graph tells you about wind energy production.

   (e) Change the dotplot that is still on the screen to a histogram (click on the arrows in the box at the upper right of the Graph Box).

      i. Move the mouse over a boundary between two bars until you get a double-arrow. Then click, hold, and drag. Describe what happens.

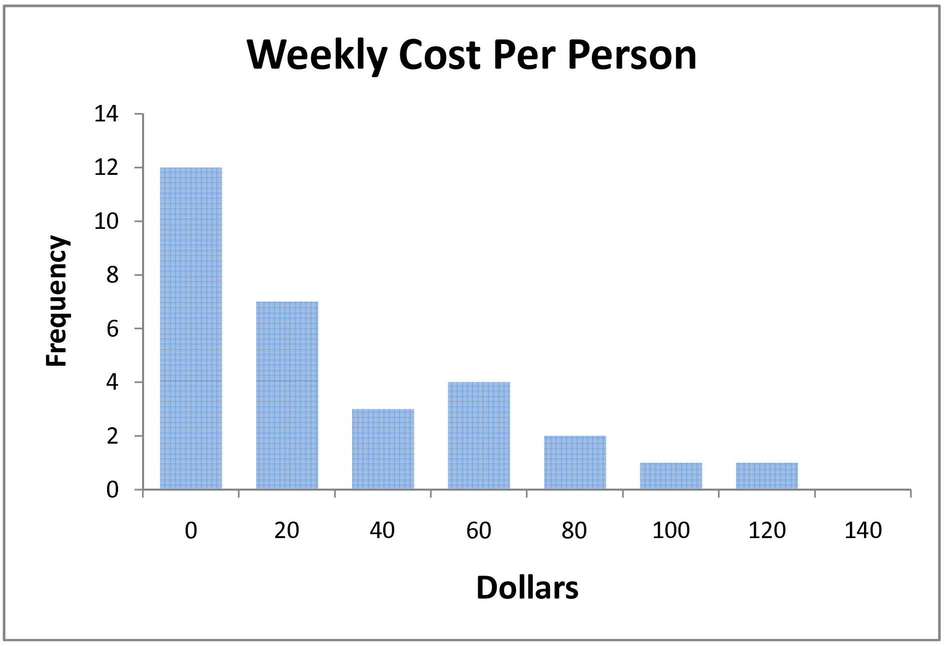
      ii. Experiment a little with changing the size of the bars. Move the boundaries around until you get each bar to cover 4 units on the vertical axis with the first bar starting at 2. You should get a total of 4 bars (one for 2 to 6, one for 6 to 10, etc.).

      iii. Once you get 4 bars in your histogram, some may go out of the window to the right. To rescale the graph so you can see everything, move the mouse over the horizontal scale values on towards the right of the axis until you get a "grabbing hand." Then, click hold and drag. Play around with this until you see the whole histogram. Try moving the hand to different parts of the the horizontal axis and also the vertical axis, and then drag to see what happens.

   (f) Be sure to save your file onto your H drive or some location where you will be able to find it later.

[23]BTU stands for British Thermal Unit, and is a standard way of measuring energy. Eight gallons of gasoline contains about 1 million BTU.

2. **Hungry Planet.** In preparing their book *Hungry Planet*, authors Peter Menzel and Faith D'Aluisio visited 30 families around the world and documented their entire food intake and its costs for a week. The histogram below shows the frequency for the cost per person for these 30 families. Note that the horizontal axis scale is offset; the first bar indicates that there were 12 families where the cost of food per person for a week was between \$0 and \$20.

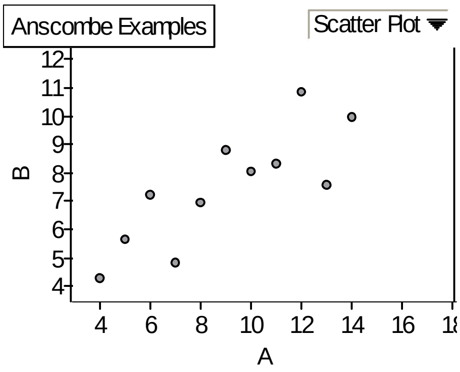


   (a) What skew does this histogram display?
   (b) In how many families was the weekly cost per person more than \$60?
   (c) In what percentage of these families was the weekly cost of food per person less than \$40?
   (d) For each member of your group, estimate as carefully as you can what you spend on food per week. If you are on a meal plan, include your weekly cost for the meal plan (even though this expense would include service, clean up, etc. that are not really part of the figures in *Hungry Planet*). Be sure to explain how you came up with the estimates for each person in the group.

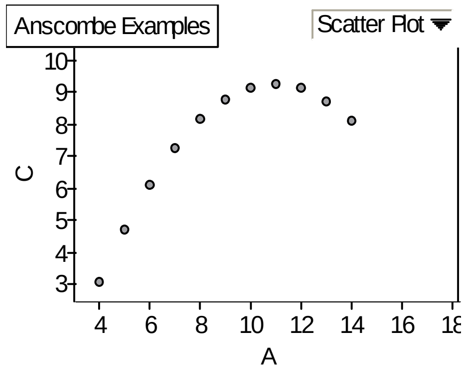
3. **Periodical Search.** Using a newspaper, magazine, or other periodical do the following:

   (a) Each group member should find at least one example of quantitative information in graphical form.
   (b) For each graph:
      i. Include a copy of the graph (scanned if possible).
      ii. Indicate the source of the graph (e.g. the July 3rd edition of the New York Times).
      iii. Indicate the type of graph (e.g. scatterplot).
      iv. Indicate what the graph represented. Include any variables named.
      v. Indicate the window used for any two-variable graphs. Indicate the scale for other types of graphs as appropriate.
      vi. Write a few sentences on what the graph is telling about the situation depicted. Consult the text for examples.
   (c) Did any of the graphs violate good graphing practice? Explain briefly.

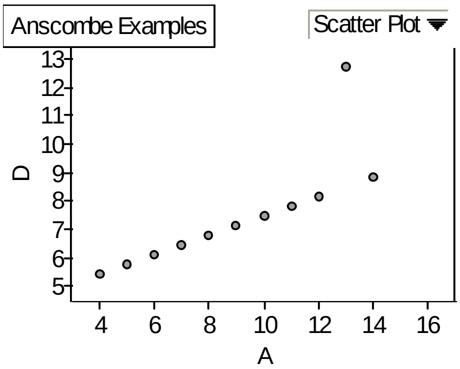
4. Anscombe I. The graphs below are scatter plots of 4 different two-variable data sets produced by Anscombe[24] as examples related to linear regression, which is a topic we will be considering later in the course. In this activity, you will "get acquainted" with the data set from a visual standpoint. We have labeled the 4 plots A, B, C, and D to the lower left of each plot.



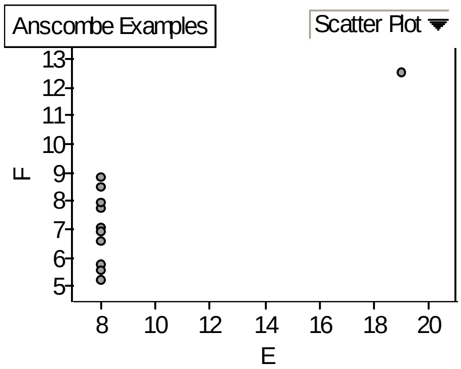
A.



B.



C.



D.

(a) For each of the 4 graphs, make your best prediction of the y-value if the x-value was 16, assuming the visual trend continues.

(b) Consider the predictions you made. Based on what you see in the graphs, you may be more or less confident about the accuracy of these predictions. In other words, if there was additional data which included y-values for $x = 16$, you might be very confident or not very confident that your prediction would be close to the actual y-value. For which graphs do you feel your prediction is highly likely to be accurate? Which predictions do you feel have a good chance of being inaccurate?

[24]Data from Anscombe, F.J., Graphs in Statistical Analysis, American Statistican, 27, 17-21

5. **Gini Coefficients and Lorenz Curves.** This activity is related to Lorenz Curves and the Gini Coefficient which were discussed in this chapter. Review the relevant sections as needed to do this activity.

   (a) Draw the Lorenz curve, and calculate the Gini coefficient for a hypothetical country where the bottom half of the population earns nothing at all, and everyone in the top half earns exactly the same amount.

   (b) Draw the Lorenz curve, and calculate the Gini coefficient for a hypothetical country where one person earns 50% of all the income and everyone else earns exactly the same amount.

   (c) Draw the Lorenz curve, and calculate an estimated Gini coefficient for a hypothetical country where the bottom 90% of the population earns a total of 10% of all the income.

   (d) Draw the Lorenz curve, and calculate the Gini coefficient for a hypothetical country where the bottom half of the population earns exactly the same amount, and the top half also each earn the same amount, but each person in the top half earns twice as much as each person in the bottom half.

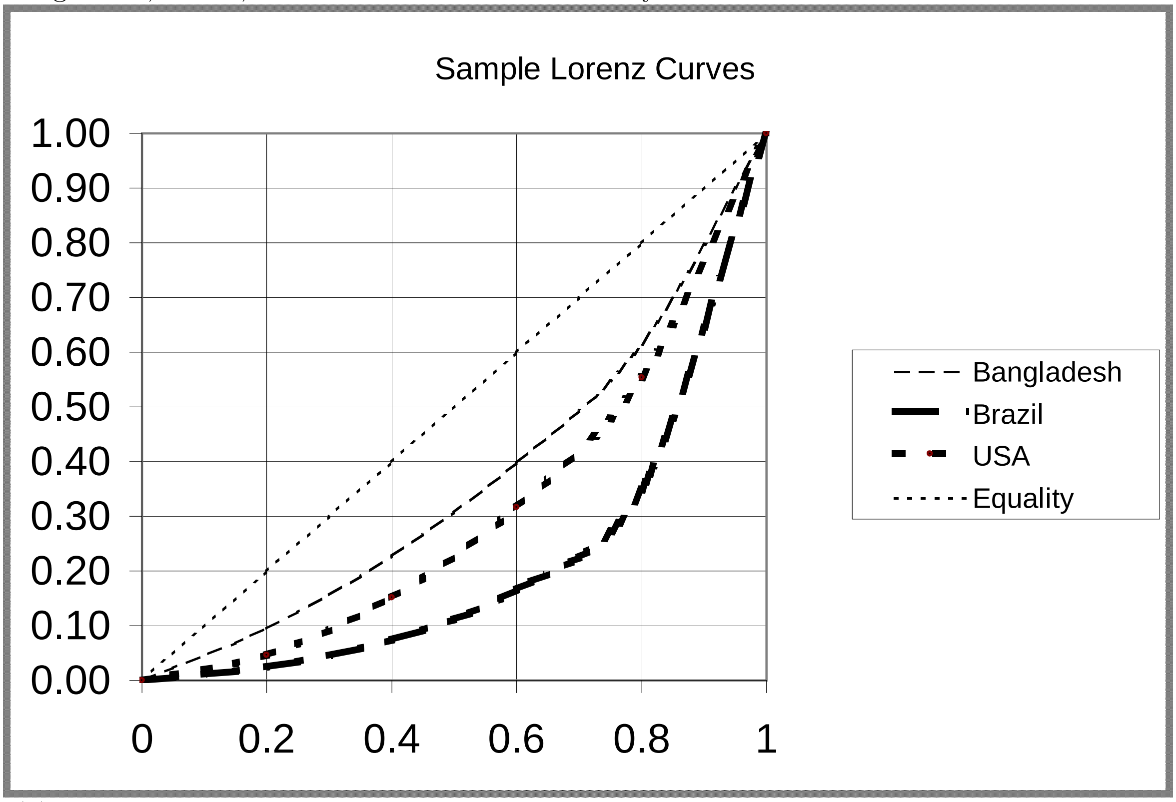
   (e) Think about a hypothetical country with a population of 10 and only two small businesses. Suppose the occupations of these 10 people are unemployed, two fast food workers, a fast food restaurant manager, a janitor, a groundskeeper, a ruler, two computer programmers, and an accountant.

      i. Assign incomes to each of the 10 people and put them in increasing order. Putting them in a table might be helpful.

      ii. Draw a Lorenz curve for this country based on your assigned incomes.

      iii. Calculate the Gini Coefficient for this hypothetical country.

6. **Income Inequality in Three Countries.** The graph below shows the Lorenz curves for Bangladesh, Brazil, and the United States for the year 1989.



   (a) Give the percentage of total income earned by the bottom half of the population in each of the three countries.

(b) Give the percentage of total income earned by the *top* 20% of the population in each of the three countries.

(c) Calculate an estimated Gini coefficient for each country, based on the graphs. As noted in the text, the Gini coefficient is twice the area between the equality line and the graph.

(d) Of the three countries for which data is given, Brazil has the highest level of inequality and Bangladesh the least. However, one drawback of the Gini coefficient is that it does not take into account the overall level of income. Per capita incomes around this time for Bangladesh, Brazil, and the U.S. were approximately $150, $2500, and $17,000 respectively. The populations for these countries (in millions) were approximately 109, 148, and 255.

   i. Based on the additional data, find the total amount of income earned in each country.

   ii. Using the Lorenz curves, find the total amount earned by the bottom 20% of the population in each country. About how much is this per person?

   iii. Find the total amount earned by the top 20% of the population in each country. About how much is this per person?

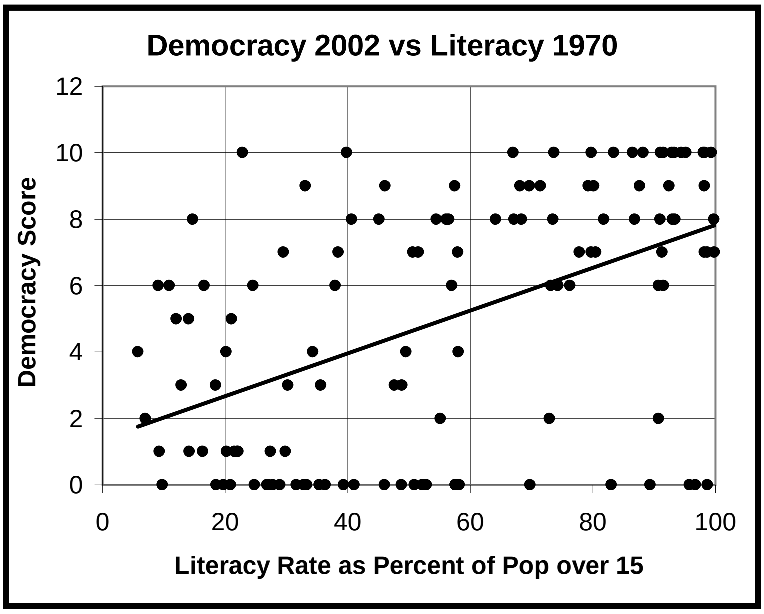
7. Looking at Democracy Vs Illiteracy.

Consider the following quote by Gwynne Dyer, a former Canadian statesman.

> *Take the year a country first reaches 50% literacy, and add one or two generations to allow the idea to sink in, and, democracy, more or less automatically, appears.*
> — Gwynne Dyer

Is Dyer correct?

Consider the scatter plot below which shows the literacy rate in 1970 as the independent variable and the Polity IV project's Democracy Score for the year 2002 as the dependent variable. The literacy rate is the percentage of the population 15 years of age or older who can read well enough to be functionally literate.



(a) Does the scatterplot seem to have a positive association, negative association, or no association? Keep in mind that areas where points are more densely clustered should be given greater weight in deciding what the overall trend is.

(b) Does this scatterplot tend to confirm or to refute the Dyer quote? Explain as completely as you can what the plot seems to imply.

(c) There are quite a number of countries with a Democracy Score of 0. Although most of these have low literacy rates, a few have higher literacy rates.

i. Speculate a little bit about which countries or regions of the world would have low Democracy Scores but high literacy rates. Include a list of at least a half dozen countries.

ii. Using the tables in the Data Samples Appendix, list which countries have a Democracy Score of 0 but literacy rates over 40%. How close to reality were your speculations?

iii. Give some possible explanations or hypotheses about why these countries have low Democracy but high literacy.

8. Income Inequality in the United States

The Gini Coefficient is one way to measure inequality. Another way is to compare incomes for different income groups within a population. The graph below shows Percent Increase in Real Average Income as a function of Time for six different income groups in the United States. Real income means income adjusted for inflation. In other words, if your income goes up at exactly the rate of inflation, your real income stays exactly the same. So, if the graph of real income versus time were a horizontal line, this would mean that *real* income has neither increased nor decreased, which in turn means that the *actual* income would have gone up at exactly the same rate as inflation.[25]

Each line in the graph represents how average real income has changed for the given group. For example, the graph for the "1st fifth" represents the percentage change in average income of those earning incomes in the bottom fifth of all incomes. These would be the poorest wage earners. Graphs are given for each fifth, as well as the top five percent, which would be the richest five percent. This richest five percent is a part of the top fifth.

The values for the dependent variable were calculated by taking the average income for the given group in the given year, and dividing by the average income in that group for 1975. For example, the average real income for those in the top five percent in 1975 was $120,364 and in 1994 it was $183,044. Taking $183,044/120,364$ we get 1.52, which is the value shown in the graph for 1994 for the top five percent group. This means that average real income went up 52% from 1975 to 1994 for this group. For comparison, the average real incomes for the bottom fifth in 1975 and 1994 were $8001 and $7762 respectively.

(a) Which group experienced the largest percentage increase in real income over this 19 year period? Which group fared the worst with respect to change in real income?

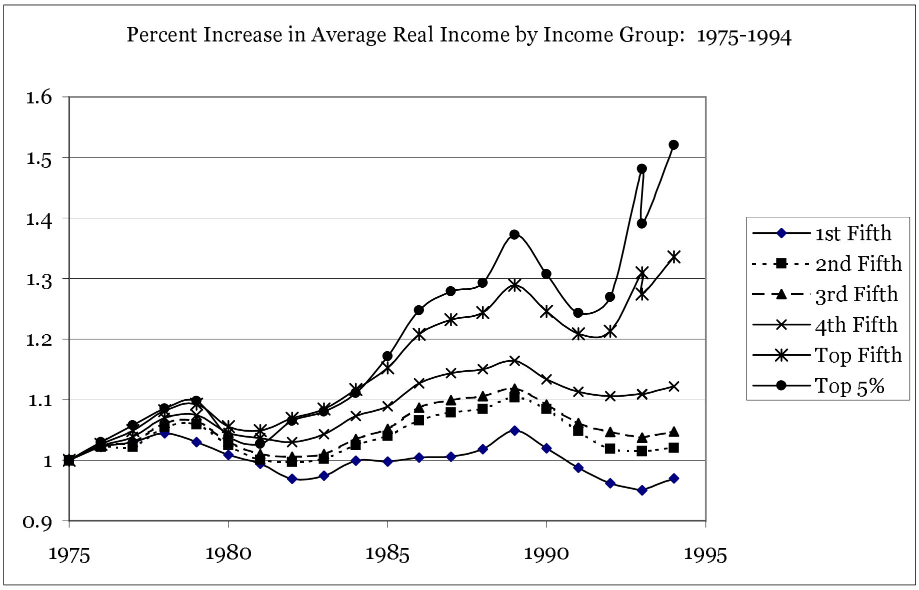
(b) Which groups, if any, had their actual income increase more slowly than inflation from 1975 to 1994?

(c) Why do all six plots have a y-value of 1 in 1975?

(d) What overall picture does the graph give regarding changes in income in the U.S. for this period?

(e) Suppose you redrew the graph, changing the range on the y-axis from .9 up to 1.6, to 0 up to 2.0. How would this change the impression of the graph?

[25] The details, which are not that hard, of calculating values which are adjusted for inflation will be covered in Chapter 7



(f) Could you change the impression of the graph by changing which years you included? For example, what would happen if you started in 1985 or 1990 instead of 1975? Do you think the person who created this graph could have manipulated the impression created by his selection of years? To help answer this question, here is a graph of percentage increases in income for the lowest fifth and second lowest fifth along with the top five percent starting in 1967.[26]

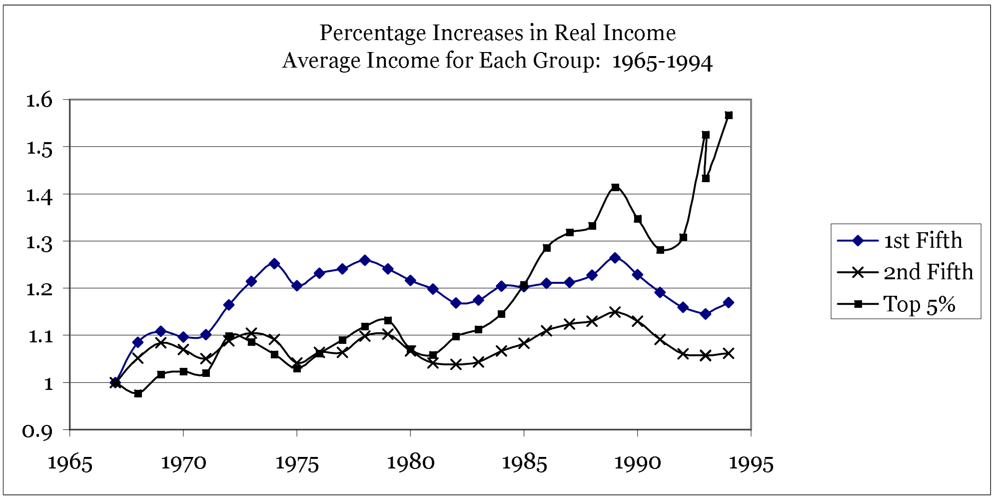


Figure 15: Another graph of percentage increase in real income

(g) Both of the graphs above depict percentage increases in average real income, rather than average real income itself. On the same set of axes, create the best plots you can for the *actual* average real income for both the top 5% group and the bottom fifth, starting in the year 1975. The dependent variable would be measured in dollars. Explain how you created your graph. How does this change the impression created by the graph?

[26]Note the significant increase for the bottom fifth from 1965 to 1974. This might be at least partially attributed to the "War on Poverty" initiated by the administration of President Johnson.

9. Bowling Alone. In his book, *Bowling Alone: The Collapse and Revival of American Community*, author Robert Putnam defines what he calls the Social Capital Index, which is a measure of how "community minded" people are. The higher the Social Capital Index for a group of people, the more likely they are to interact in constructive ways, join clubs, participate in community activities, etc. Putnam has computed the Social Capital Index for each state, except Alaska and Hawaii.

The first scatterplot in Figure A below shows the relationship between the Social Capital Index (SCI for short) and Percent Food Insecure Households (PFIH for short). PFIH represents the percent of households in a state that typically have trouble procuring sufficient food, or food of adequate quality on a monthly basis. It is considered a measure of how many people are at risk for hunger.

The second scatterplot shows the relationship between the SCI and State Poverty Rates, defined as the percentage of the population in a state that are living below the official poverty line. Data in tabular form for these and other variables are given in the Appendices.

Figure C shows the relationship between Educational Spending per Pupil and the SCI.

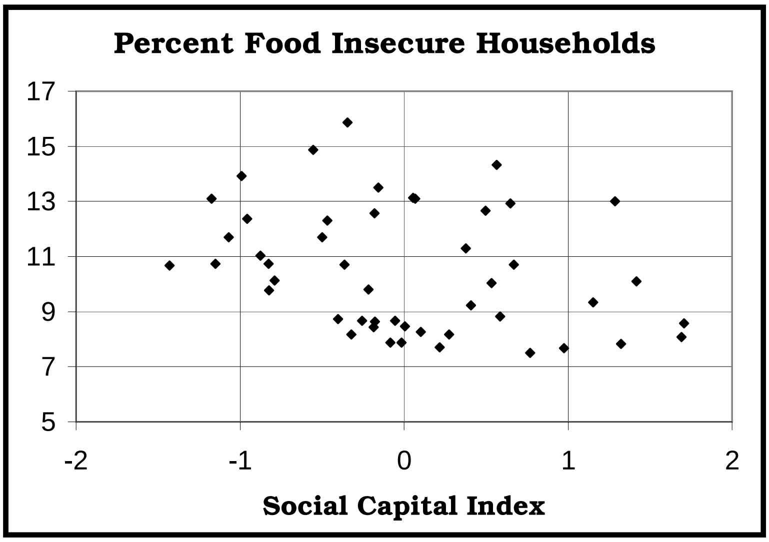


Figure A

(a) For each scatter plot, say whether the variables seem to have a positive association, negative association, or no association.

(b) Which relationship seems to be the stronger, or are they of about the same strength?

(c) In the poverty versus SCI plot, the point at the lower left could be considered an outlier. It is the lowest of all states in Social Capital, and is also has a fairly low poverty rate. Using the table in the Appendices, find out which state this is.

  i. Can you think of some reasons why this state is low in Social Capital?.

  ii. Can you think of some reasons why this state has a fairly low poverty level?

(d) In the Education Spending Versus Social Capital plot, the state that is lowest in per pupil spending has an SCI of .5. Find out which state this is.

  i. Can you think of any reasons why this state would be low in per pupil spending, but in the middle with regards to social capital?

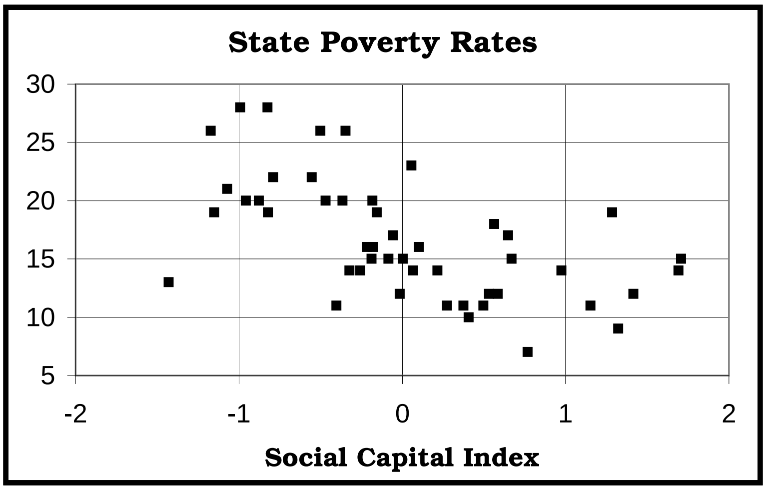


Figure B

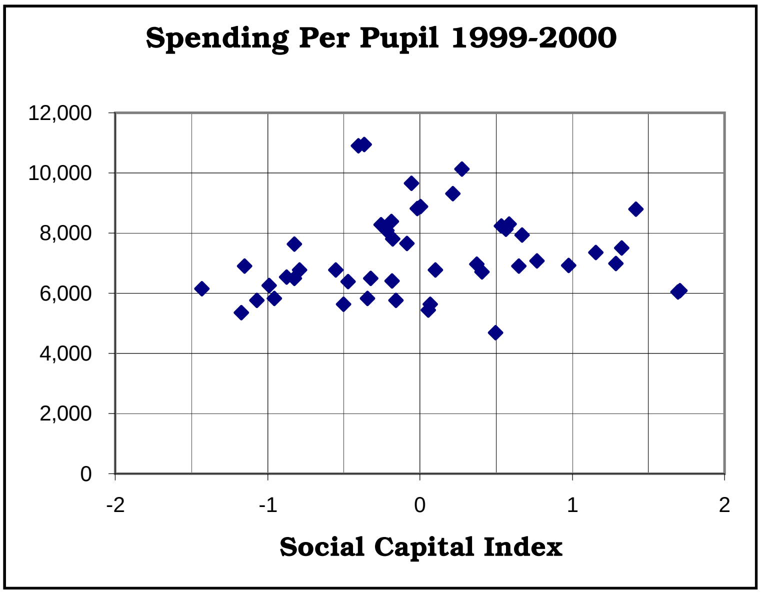


Figure C

3. Collecting and Analyzing Data

In this chapter, we make a more systematic study of some of the basic concepts that are useful when considering real world data. The basic theme will always be to ask questions of the data. What can I learn from this data? Are there any conclusions I can draw? Can I verbally describe the situation under consideration based on the data I have?

It is also important to be cautious when working with data. Can I trust that this data set is reliable? What are the shortcomings of this data set? Are there any important cases or other information that seem to be absent from these data? Overall, we want to strike a balance between analyzing and understanding the data as completely as possible, but without jumping to conclusions by reading more into the data than is actually there.

Validity and Reliability

It has been widely reported by a number of media outlets that the average age of a homeless person is nine.[27] Can this number be right? Where did it come from?

The first consideration when working with data is knowing to what extent the data are **valid** and **reliable**. Validity refers to how well the data measures what it is intended to measure. If your bathroom scale gives the true weight of the people that stand on it, the weights it produces form a valid set of data. If your bathroom scale does not produce accurate weights, or if it gives the height instead of the weight, then it is not producing valid data. Valid measurement techniques are also sometimes referred to as **accurate**.

The reason validity is important is that if we answer questions, conduct research, make predictions, or base public policy on invalid data, then our answers, research conclusions, predictions, and policies could well be wrong or counterproductive. If we develop policies to address homelessness assuming that the average age of a homeless person is nine, when the average is really 39, our policies are likely to be ineffective or wasteful.

Reliability refers to a data collection procedure producing consistent results in repeated use. If your bathroom scale consistently gives weights which are 10% heavier than the actual weight, then your scale produces *reliable* data, but not *valid* data. Another way of putting this is that if you weigh yourself on the same scale at different times or in different places (while still actually weighing the same amount), then your scale produces reliable data (even if the data is not valid!).

In many situations, validity and reliability are not a problem due to the nature of the data. For example, if we are collecting data on a certain sport, say baseball, we can easily access data that is entirely valid and reliable for each individual or each team. Major League Baseball's official records would have the exact number of wins and losses for each team, the exact number of home runs and stolen bases for each players, etc. Errors in measuring these statistics are very unlikely. The biggest possibility for error is probably in typing or copying the data incorrectly.

In other situations, validity or reliability can be a problem. For example, we might want to compare household incomes for the 50 United States, or among various countries. How would we determine how much individual households make in a given state or country? It is typically much too impractical to ask every household to provide us with their income. We typically must be satisfied with collecting data from only a **sample** of the entire **population** of all households. In general, a population is the entire group we would like to have some data on, and a sample is a subset of the population from which we actually collect data. The key to obtaining valid and reliable data in this situation is using a sampling technique that will consistently produce **representative** samples of the population.

For example, if a high school student wanted to collect data on household incomes, she might simply to decide to knock on 100 different doors in her small Nebraska town and ask one of the residents what their household's income is. If she is careful to randomly select which houses to visit, her data might be fairly representative of household incomes in her town, or in other words,

[27] see http://abclocal.go.com/wjrt/story?section=news/local&id=4975476 for one example, or simply google the last nine words of this sentence

if the population she wanted to consider was all households in her town. However, her data would not be representative if the population under consideration was all households nationwide, since no urban households and no households outside of Nebraska were included in her survey.

So, if the student were using this technique to estimate incomes nationwide, her method would not be valid. It would not accurately measure incomes nationwide. On the other hand, if she repeated her survey several times, each time randomly selecting 100 households, the method would probably be fairly reliable. The averages of her several samples would likely not vary from each other too much. How much variance she might expect we will be able to quantify when we discuss standard deviation.

It is also possible the people in an individual household do not have a very accurate idea of what their annual income is or might have some motivation to be less than truthful in reporting their income. In cases like this, even a representative sample might not produce valid data. In collecting data, or using data that has been collected for us, we always want to be aware of potential problems in the validity and reliability of the data.

Example 1. A high school student wants to determine what percentage of students in her school

have engaged in underage drinking. To this end, she randomly selects 25 of her schoolmates and schedules face-to-face interviews with them in an empty classroom. Interviews take place during school hours. One question asked during the interview is "how many times have you engaged in drinking alcohol in the last six months?" From the 25 responses she receives, she finds that the average number of times the respondents drank was 1.3.

Does this method produce a valid and reliable average?

Solution:

Given the circumstances of the interviews, the validity of the data is questionable. The students interviewed may be reluctant to be totally truthful, given the possibility that their responses might be overheard by school officials or otherwise be made available to persons other than the student doing the interviews.

As far as the reliability, this may be difficult to determine. If the student conducted several sets of 25 interviews with different groups of randomly selected students, she probably will not get the same 1.3 average every time. On the other hand, we might not expect the results to be drastically different from group to group. Later in this chapter, we will be able to quantify how much "variability" we can expect in a situation like this. In general, if the student employs appropriate randomization and asks a sufficient number of students, we would consider the data reliable.

■

Collecting a Sample

Random Samples

Data can be collected in quite a number of ways. As noted above, data are often collected through surveys or other sampling techniques. In this section, we discuss issues related to sampling.

Collecting a proper sample is key to producing good data. Without it, the conclusions that you draw from your sample are usually worthless. Typically, national polls are conducted by sampling approximately 1000 people since this gives a margin of error of no more than three percent.[28] From this relatively small number, the opinions of more than 300 million Americans can be inferred with a relatively high level of confidence. We call the percentage of all 300 million Americans who respond in a certain way to a survey question a **population parameter** p. The percent of the **sample** who respond in this way is called the **sample proportion**, and is denoted $\hat{p}$.

The most important aspect in collecting data is **randomization**. In order to create a sample which is representative of the population from which it is drawn, we seek to use methods which will insure (as much as humanly possible) that each member of the population has an equally likely chance of being in the sample. The most basic sampling technique which accomplishes this is called a **simple random sample**.

[28] We will see how margins of error are calculated and what they mean in the next chapter

For example, suppose we want to find out whether the students at a small college of 1000 students object to a new policy banning the sale of soft drinks on campus. To create a simple random sample of size $n = 20$ from this population of 1000 college students, we could assign each student a number from 1 to 1000, write these on 1000 identical pieces of paper, put them all in a hat and then draw out 20 (after mixing thoroughly). This should give each student an equally likely chance of being selected, and as a result, give each possible set of 20 students an equally likely chance of being the sample.

In practice, one uses more technologically sophisticated mechanisms for achieving randomness. Many computer programs have so-called "random number generators" which can be used instead of a hat full of paper to do sample selection.

Non-random sampling techniques can create invalid, often also called **biased** data. For example, if we were doing our soft drink survey but selected our sample of 20 by standing next to a coke machine in the basement of a residence hall and asking the first 20 students who wander by, we are not giving each student an equally likely chance of being selected. We would tend to get only students from that residence hall, and probably get students who like to drink pop, since we were standing by a coke machine (and one located in an out of the way place at that). Thus, our results are not likely to be **representative**, and we can be fairly sure the people in the sample would have more concern about the new policy than the student body as a whole. Our sample would be considered **biased**, and as a result, the data from our sample are not likely to be **valid**. We wanted an accurate measure of what percent of all students objected to the new policy, and our sampling technique is not likely to provide this.

There are other types of sampling techniques which can also produce results which effectively represent the entire population. One of these is called a stratified random sample. In a **stratified random sample**, the population is first divided into subgroups and then simple random samples are taken from each group. For example, if we conducted our soft drink policy survey by first dividing the 1000 students into residential and non-residential and then creating separate random samples from each group, this would be a stratified random sample. The advantage to such a sample is that we could then see whether the two groups differed in their opinions and by how much.

Bias and Variability

In any survey, randomization is used to reduce **bias**. In general, any sampling technique which is likely to produce data which is not valid is considered biased. Bias often occurs when the sample selected is not representative of the entire population.

Another goal in sampling is to reduce **variability**. A sampling technique has high variability if the results obtained vary significantly in repeated use of the technique. In other words, if there is significant variability, the technique does not produce reliable data. These two ideas are illustrated in Figure 1. The center of the target represents the parameter, p, for the entire population and the dots are the different $\hat{p}$ determined by several different samples. If variability is low, the dots will be clustered together. If bias is low, the dots will be around the center. Note that one can have high bias and low variability, or low bias and high variability. For example, if we conduct several soft drink surveys, but always stand near coke machines, the surveys might have low variability but will probably have high bias. We are likely to consistently overestimate the percent of students p who object to the new policy.

If a sample is selected randomly, the possibility of getting a biased sample that systematically favors a particular outcome is greatly reduced. There are, however, other things that can bias a study. For example, the way in which questions are worded can bias the results. For example, asking the question "Do you favor allowing students the freedom to choose to drink soft drinks wherever and whenever they wish?" is likely to get more " yes" responses than the question "Do you favor exposing students to risks associated with obesity by allowing them unlimited access to soft drinks?" These statements admittedly are very different in wording, but even subtle changes in wording questions, such as changing "welfare" to "aid to the poor", or "pro-choice" to "pro-abortion" can influence the responses people give to a question.

The principle way to reduce variability is to use larger sample sizes. In general, the more people

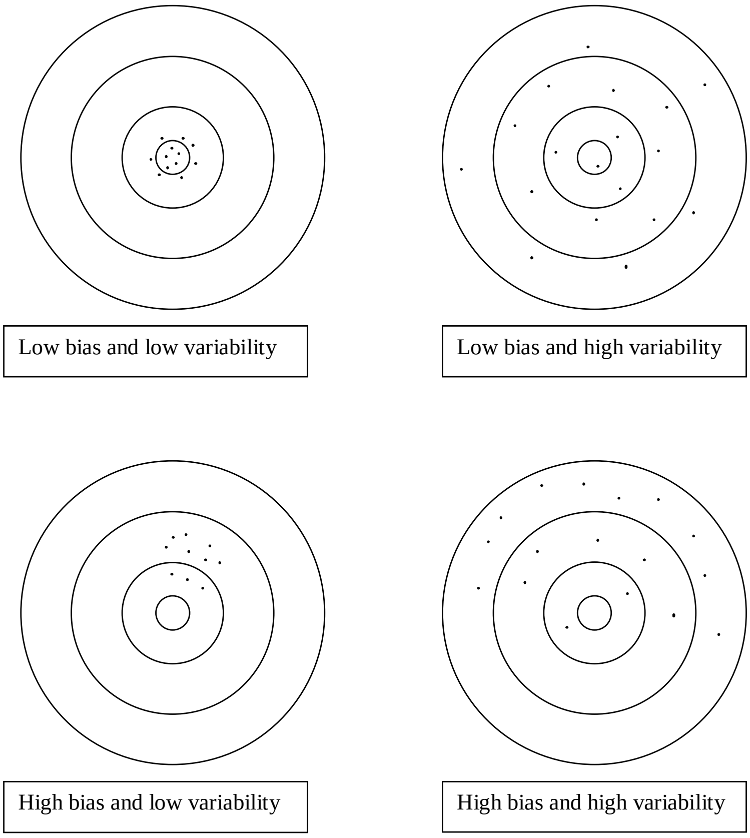


Figure 1: An illustration showing varying combinations of variability and bias.

surveyed, the less variability. Large does not mean hundreds of thousands, however. Political polls, for example, typically sample about 1000 people. The **margin of error** usually reported for a simple random sample of 1000 is about 3%. This typically means that we are very sure (where "very" usually means 95% sure) that our poll result is within 3% of the population parameter p we are trying to measure. If the sample were to be reduced to 100 instead of 1000, the margin of error would increase to plus or minus 10% which is too high to be acceptable for most studies.

To summarize, bias is generally controlled by employing randomization and careful design of questions, while variability is generally controlled via sample size.

Example 2.

The soft drink survey above is one example of a poor sampling method. Here is another.

Each day, many news websites like CNN conduct surveys. On July 20th, 2007 the question on CNN's main page was "Do you plan to read the new Harry Potter book?"[29] One could select Yes or No. This sampling technique is certainly not random. Only those who go to the CNN site will even see the poll. Of those who see it, those who have strong feelings (good or bad) about Harry Potter are more likely to respond than those who could care less. Clearly, not every person in the U.S. is equally likely to be a part of this sample. Samples like this are called **self-selected** or also **voluntary response samples**. If you are curious about the results, as of 6 p.m. Central Daylight Time there were 15458 yes votes, 30% of the total, and 36822 no votes. Also, CNN did indicate on their website that this was not a **scientific** poll. In other words, they acknowledged that randomization was not used and that the results were not necessarily representative.

[29]Harry Potter and the Deathly Hallows, the last in the Potter series, was released the following day with over 2 million advance sales in the U.S.

When you read articles that include results from sampling, be careful not to "believe everything you read." Ask questions. Did the sampling technique employed use randomness or was the sample completely self-selected? What was the sample size and the margin of error? How were the questions worded? Was the survey or study conducted by a reputable firm with experience in such work, or was it conducted by an organization likely to have a strong view on the issue at hand? This latter does not automatically mean the results are biased, but it should lead one to be careful in accepting the results at face value.

It is important to note that in this chapter, we use the word "random" in a very specific well-defined way. On the other hand, "random" also has other connotations that are less specific. For example, we speak of events happening "at random" like a "random accident." Here, "random" means something like "unexpected" or "unable to be anticipated." Or, a student might decide to do a survey by asking people he runs into "at random" her set of questions. Again, "random" here might mean "not pre-determined." However, this is not a random sample in the mathematical sense because the student has not insured that each person in the designated population is equally likely to be part of the sample.

Measuring Center

> **Average** *real family income grew by well over 15% from 1982 to 1989.*
> —Rush Limbaugh
>
> *For millions of years, on* **average**, *one species became extinct every century. ...We are now heaving more than a thousand different species of animals and plants off the planet every year.*
> — Douglas Adams
>
> *The* **average** *age of the world's greatest democratic nations has been 200 years. Each has been through the following sequence: From bondage to spiritual faith. From faith to great courage. From courage to liberty. From liberty to abundance. From abundance to complacency. From complacency to selfishness. From selfishness to apathy. From apathy to dependency. And from dependency back again into bondage.*
> — Lord Thomas MacCauley
>
> *In this era of globalization, international affairs touch the lives of* **average** *Americans in unprecedented ways. And as we wage a global campaign to purge from the world, the terrorist threat against our very way of life, the assistance we provide to friendly governments and impoverished peoples across the globe supports our ability to sustain an international coalition to fight terror and retain the popular goodwill necessary to this task.*
> — Senator John McCain

The quotes above illustrate a number of different ways that the word "average" is used. When we are considering a set of data, the average is usually meant to be the *single* number which best represents all the numbers in the data set.

Mathematicians refer to averages as "measures of center," and there are actually several ways to do this, including the common average that everyone is familiar with. In this section, we will look at two ways of measuring the center of a data set. We will also consider the concept of the **spread** of a data set and ways to numerically describe this spread.

Intuitively, the center of a set of data is a single data point or value that is meant to be most representative of the whole set. Obviously when you represent a whole set of numbers by a single number, you are going to lose some information. No one number can accurately represent the whole set.

The Mean of a Data Set

There are several measures of center that are commonly used. The first is the ordinary arithmetic average, also known as the mean. To determine the **mean** of a data set, you add up the all the data values and divide by the number of data points. For example, suppose the starters on a college basketball team have heights in inches of 72, 76, 77, 77, and 80. The sum of these heights is 382 and the mean of these numbers is $(72 + 76 + 77 + 77 + 80)/5 = 382/5 = 76.4$ inches. The standard symbol for the mean of a data set represented by x is $\overline{x}$, which is usually pronounced "x-bar." For our starting team, we have $\overline{x} = 76.4$ inches.

You can think of the mean as a way of "levelling off" a data set. For example, our basketball team above would have the same "total height" of 382 inches as a team of 5 players all of whom were exactly 76.4 inches tall. Thus, the mean is always the value of each data point if the total sum of all the data points is divided equally.

As mathematicians always like to be efficient, the following notation is often used as shorthand when dealing with means. If there are n numbers, denoted by the generic names $x_1, x_2, x_3, \ldots, x_n$, their mean is

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}.$$

This can also be expressed as

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

The Σ is the upper case Greek letter sigma, and is the standard symbol for summation. The expression, Σx_i, is just a symbolic way of saying "add up the numbers" $x_1 + x_2 + x_3 + \ldots + x_n$.

Medians

A second common way of measuring the center of a set of data is the **median**. The **median** is the middle number of an *ordered* set of data. To find the median height for our basketball team, we first order the data set as $\{72, 76, 77, 77, 80\}$ (or put in reverse order). The middle number, 77, is the median.

What if there is not middle number? In this case, the median is the average of the middle two numbers in the ordered set. For example, to determine the median for the data set $\{7, 5, 9, 11, 6, 2\}$, we first order the numbers from smallest to largest as we did before: $2, 5, 6, 7, 9, 11$. The two middle numbers are 6 and 7, and so the median is the average of these two, namely 6.5. The median of a data set has the property that the number of data points larger than the median is equal to the number of data points smaller than the median. This is why the median is a measure of center for the data set.

Example 3. The following table shows malnutrition data for a number of Latin American countries for the year 2001. The data is the percent of children under five who are severely underweight for their age. [30]

Determine the mean and median for this data.

Solution: The mean percentage of severely underweight children is

$$\overline{x} = \frac{0.4 + 0.6 + 0.7 + 0.8 + 0.8 + 1 + 1.1 + 1.2 + 1.7 + 1.9 + 1.9 + 4 + 4.7}{13} = 1.6.$$

This means that, on average, these countries have a 1.6% rate of malnutrition, using this measure. To determine the median, note that the data is already listed from lowest percentage to highest. Since this is an odd numbered data set, the median is just the middle number, which is 1.1. This tells us that approximately half of the countries have underweight rates higher than 1.1%, and half lower. ■

[30] Data is from the UNICEF end of decade database at <http://www.childinfo.org/eddb/malnutrition/index.htm >.

Country	% Severely Underwt. Chld.
Costa Rica	0.4
Brazil	0.6
Venezuela, RB	0.7
Colombia	0.8
El Salvador	0.8
Dominican Republic	1
Peru	1.1
Mexico	1.2
Bolivia	1.7
Ecuador	1.9
Nicaragua	1.9
Honduras	4
Guatemala	4.7

It is an unfortunate fact that both the mean and the median are often referred to as the average. In considering the data in Example 2, one newspaper might say "the average Latin American country has a malnutrition rate of 1.1%", while another might report this average as 1.6%. Both deliberate and unintentional confusion of mean and median is common in political and public discourse.[31] Whenever you hear an average reported, you should wonder whether you are getting the mean or the median. Is there any way you can tell without actually going back to the source of the data? Not always, but we will look at some ways where you might at least be able to tell which is more likely being reported.

Comparing the Mean and Median

Although the mean and median are different measures of center, they often give similar results. For example, we found earlier that the mean of our five basketball players heights $\{72, 76, 77, 77, 80\}$ is 76.4 while the median is 77. However, there are times when the median and mean are quite different. In Example 2, we saw that the mean percentage of severely underweight children in the given countries is 1.6 while the median number was 1.1. Here, the mean is significantly larger than the median, as compared to the range of percentages for all the countries. The reason the mean and median are noticeably different is because of the impact of Guatemala and Honduras. The data for these two countries is significantly higher than the rest, and they could be considered outliers. [32]

Intuitively, an **outlier** is a data point that falls well outside of the overall pattern of the data. When calculating the mean, every number is added to the sum. Thus, larger values like those for Honduras and Guatemala can make the sum, and thus the mean, significantly larger than it would be if these data points were only slightly higher than the others. The median, however, is not impacted by outliers because its calculation is based on just the middle number, or the middle two numbers in the list. In cases where there are outliers or the distribution is severely skewed, the median probably represents a better measure of center than the mean.

When a distribution is symmetric, such as in Figure 2, the mean and median both occur near the center or peak of the histogram and will be approximately equal. The median always occurs at the point that divides the histogram into two parts of equal area. We will represent the median in a histogram with the letter M. Equivalently, the sum of the heights of the bars to the left of M should roughly equal the sum of the heights of the bars to the right of M, as long as all the bars are of equal width.

[31] You might consider the Limbaugh quote opening this chapter regarding whether he is giving the mean or the median, and whether he is doing this intentionally or not!

[32] The key here is what we mean by "significantly." This will be quantified more precisely later.

The mean would be the "balancing point" of the histogram. In other words, if you imagine your histogram bars were made out of blocks and you set them on a see-saw, the point at which you would you put the fulcrum of the see-saw to exactly balance it would be the mean. Just as on a real see-saw, the further you sit away from the fulcrum, the more you pull down on your end of the see-saw. We all know a small person far away from the fulcrum can balance a large person who is close to the fulcrum. In a histogram, a little bit of data (short bars) far away from the fulcrum can balance a lot of data (tall bars) close to the fulcrum. For the distribution shown in Figure 2, both the mean and median are 18, and are indicated under the horizontal axis on the graph.

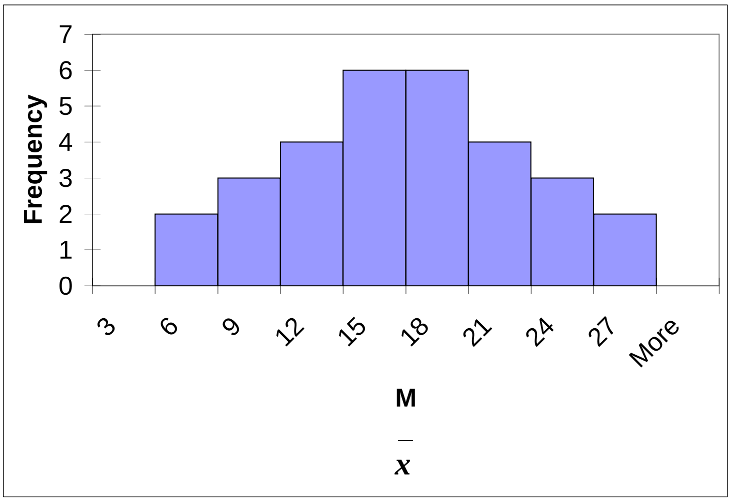


Figure 2: The mean and median are equal when the distribution is symmetric. In this distribution, they are both 18.

How do the mean and median compare if a distribution is not symmetric? Consider the data on drinking water from the Reading Questions in Chapter Two. The graph in Figure 3 represents the percentage of the population in a number of countries who have access to improved drinking water supplies.[33]

This histogram is severely skewed to the left. To estimate the median, one could first add up the estimated heights of all the bars (it turns out this is equal to 150 which is the total number of countries represented). Half of 150 is 75. To find the median, we need to find that point M where 75 or half of the countries have a percentage of the population with access to improved drinking water above M and half below. We see that around 60 countries have a percentage above 90, and about 85 countries have a percentage above 80. This means the median M must be somewhere between 80 and 90. Based on the actual data, the median turns out to be 85.

Estimating the mean is more difficult, but the mean will clearly be smaller than the median since the short bars far to the left of the peak will "pull the average down." The mean calculated from the actual data turns out to be 79.22.

Finally, Figure 4 shows a histogram of our Latin American malnutrition data. The histogram could be considered to be skewed right. Alternatively, one could consider the two countries represented by the rightmost bars as outliers. Either of these is consistent with the mean of 1.6 being larger than the median of 1.1. You might pencil in marks for these values on the histogram. Note that, in situations where you have two competing numbers, both claiming to be the "average," you might be able to tell which is the mean and which is the median if you have some idea of the (likely) skew of the distribution of the data.

In general, how do you decide if the mean or the median is a better measure of center? The answer depends on what you are looking for and how the data is distributed. If the distribution is approximately symmetric, the mean and the median will be approximately the same. If the data is

[33] Data collected from the United Nations Conference on Trade and Development website, http://stats.unctad.org

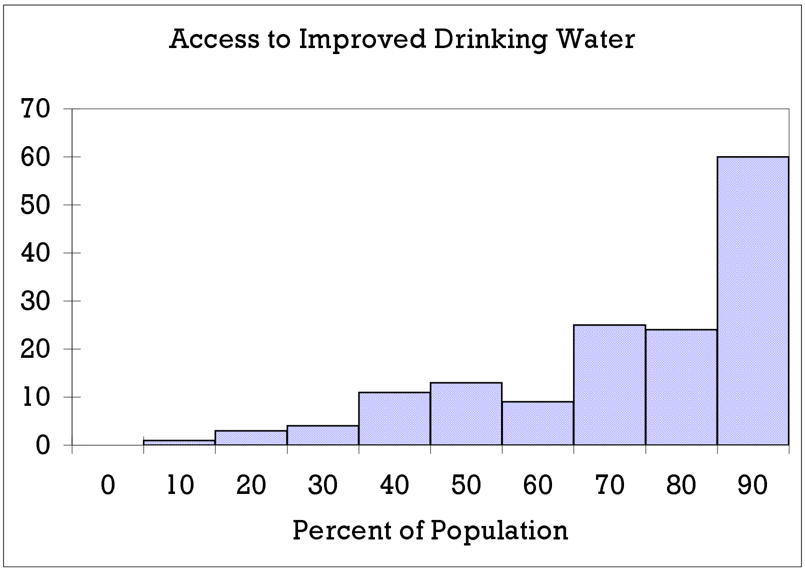


Figure 3: Histogram of data on improved drinking water

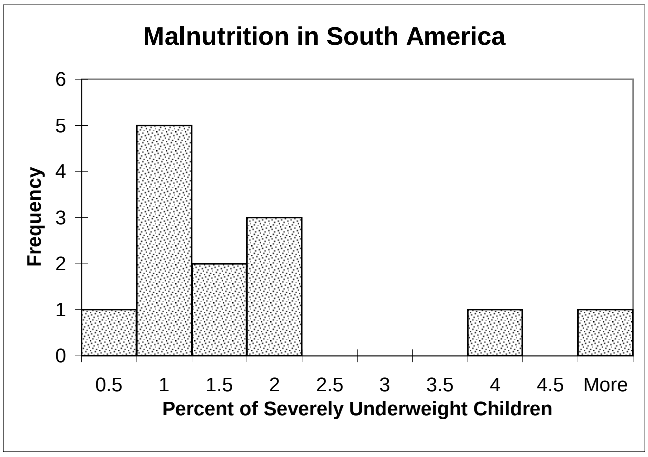


Figure 4: Histogram for childhood malnutrition in Latin America

highly skewed, then the median usually gives a better measure of center. However, the mean does have an important property not possessed by the median. As noted previously in our basketball team example, the mean "levels off" the data set. In other words, the mean gives the value of each data point assuming they were all exactly the same value. For example, suppose a company has 10 employees with a mean salary of \$31,000 and a median salary of \$27,000. The total amount of money paid out in salaries is $10 \times \$31,000 = \$310,000$. You cannot deduce this type of information using the median.

Measuring Spread

The mean and median give measures of center. They answer the question, "what is the most representative element of this data set." However, two data sets with the same mean or median can be very different, depending on how the data is distributed around the center, or its **spread**.

For example, suppose we have two companies, A and B, each with five employees. The salaries for A's employees in thousands are

$$20, 20, 30, 40, 40$$

and for B's employees

$$25, 28, 30, 32, 35.$$

Now, for both companies, the mean and median salaries are $30 thousand. But the salaries for company A have much more spread.

One measure of spread is the **range**, which is defined as the difference between the highes and lowest elements in the data set. The range for company A is $40 - 20 = 20$, and range for company B is $35 - 25 = 10$.

Standard Deviation

One problem with the range is that two data sets with the same range can still be distributed differently. Suppose we have company C with salaries

$$20, 29, 30, 31, 40.$$

Company C has the same mean, median, and range as company A. However, more of the salaries are closer to the mean. In some sense, there is less spread "on the average" among this data set.

Because of the shortcomings of the range as a measure of spread, we introduce the **standard deviation**. We will typically compute standard deviations of data sets using technology. However, it is instructive to see how the standard deviation is calculated "by hand." Here's how it works.

Calculating the Standard Deviation of a Data Set $\{x_i\}$.

1. Find the mean, $\overline{x}$, of the data set.

2. For each data point, x_i, in the data set, find the difference between it and the mean, $x_i - \overline{x}$. These are the distances from each individual data point to the mean.

3. Square the differences from step 2, $(x_i - \overline{x})^2$. [Note: Since the differences are squared, there is no distinction between data points below the mean and data points above the mean.]

4. Find the sum of the squared differences, $\Sigma(x_i - \overline{x})^2$.

5. Divide the sum of the squared differences by the number of data points in the set, $\frac{\Sigma(x_i - \overline{x})^2}{n}$. (This number is called the **variance** of the data set).

6. Find the square root of the variance, $\sqrt{\frac{\Sigma(x_i - \overline{x})^2}{n}}$. This is the standard deviation.[34]

The lower case Greek letter sigma, σ, is often used to represent standard deviation. So

$$\sigma = \sqrt{\frac{\Sigma(x_i - \overline{x})^2}{n}}.$$

Let us compute the standard deviation for two of the companies mentioned previously.

[34] The formulas that we will use for variance and standard deviation have an n in the denominator. This gives the standard deviation for a population. There are other formulas for sample variance and standard deviation that have a $n - 1$ in the denominator and give the standard deviation for a sample.

Example 4. Compute the standard deviation of the test scores for Company A,

$$20, 20, 30, 40, 40$$

and Company C

$$20, 29, 30, 31, 40.$$

Solution: The mean, $\overline{x}$, for both sets of scores is 30. We found the sum of the squared differences in the right most entry of the last row in the following table.

Company A

	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
	20	$20 - 30 = -10$	$(-10)^2 = 100$
	20	$20 - 30 = -10$	$(-10)^2 = 100$
	30	$30 - 30 = 0$	$(0)^2 = 0$
	40	$40 - 30 = 10$	$(10)^2 = 100$
	40	$40 - 30 = 10$	$(10)^2 = 100$
$\Sigma(x_i - \overline{x})^2$			400

So, $\sigma = \sqrt{\frac{\Sigma(x_i-\overline{x})^2}{n}} = \sqrt{\frac{400}{5}} \approx 8.94$ for Company A. We can think of 8.94 as being roughly the average distance of the salaries from the mean. Now for Company C:

Company C

	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
	20	$20 - 30 = -10$	$(-10)^2 = 100$
	29	$29 - 30 = -1$	$(-1)^2 = 1$
	30	$30 - 30 = 0$	$(0)^2 = 0$
	31	$31 - 30 = 1$	$(1)^2 = 1$
	40	$40 - 30 = 10$	$(10)^2 = 100$
$\Sigma(x_i - \overline{x})^2$			202

So, $\sigma = \sqrt{\frac{\Sigma(x_i-\overline{x})^2}{n}} = \sqrt{\frac{202}{5}} \approx 6.36$ for Company C. This means that the average distance of the scores from the mean, 30, is roughly 6.36 or $6,360. Since the standard deviation for Company A is greater than the standard deviation for Company C, we say that Company A has a larger spread in its salaries. ■

Percentiles and Quartiles

Another measure of the spread of a data set uses percentiles or quartiles. The **nth percentile** of a distribution is the number such that n percent of the observations fall at or below it. For example, in 2005, the 20th percentile for annual houshold income in the U.S. was $18,500. This means that 20% of all households earned $18,500 or less, and thus 80% of all households earned more than this. Here is a table of percentiles for U.S. income distribution in 2005.[35]

Percentile	20th	40th	60th	80th
Income	$18,500	$34,738	$55,331	$88,030

[35] $http\ en.wikipedia.orgwikiHousehold_income_in_the_United_states$

From this table we get an idea of the distribution of incomes in the U.S., we cannot see the range, since the maximum and minimum incomes are not shown. We can infer that 20% of all households earn between \$18,500 and \$34,738, and that 60% of the population makes between \$18,500 and \$88,030. As a side note, percentiles that are multiples of 20 like those in this table are also called **quintiles** since they divide the data into 5 equal or nearly equal groups.

By definition, the median is always the 50th percentile. Other commonly reported percentiles are the 25th and the 75th. These percentiles are also called **quartiles**. The 25th percentile is known as the first quartile, the median is the second quartile, and the 75th percentile is known as the third quartile. The 1st, 2nd, and 3rd quartiles divide a distribution into four groups with roughly the same number of data points in each group.

Percentiles, including quartiles, can often be found via technology. When finding quartiles "by hand," the 1st quartile is defined as the median of the values below the median while the 3rd quartile is defined as the median of the values above the median. So, for example, in the data set

$$\{2, 3, 7, 7, 8, 10, 12, 17, 18, 35\}$$

the median will be 9. For the five numbers below the median, the median is 7, and so 7 is the first quartile. The third quartile is 17.

The difference between the 3rd quartile and the 1st quartile is called the **interquartile range.** This is yet another measure of the spread of the data.

Previously, we had referred to data points which are far above or far below the bulk of the data as outliers. To be more precise, **outliers** are often defined as data points which are more than 1.5 times the interquartile range above the 3rd quartile or below the 1st quartile.

For example, the interquartile range in data set above is the difference between 17 and 7, or 10, and 1.5 times 10 is 15. The 3rd quartile is 17 and adding 15 to 17 gives 32. Thus, the data point 35 is an outlier since it is greater than 32.

Measuring Variability

Previously, we looked at data collection techniques and the problems of bias and variability in data collection. In this section, we will consider how to quantify variability.

Even when measures are taken to reduce bias to a minimum, any sampling technique includes the presence of variability, unless one gathers data from every single individual in the population under consideration. Variability can be measured using the **margin of error**, which is based on the idea of standard deviation discussed earlier.

In March of 2007, the McGovern Center for Leadership and Public Service and the Dakota Wesleyan University Mathematics Department conducted a statewide poll of 410 South Dakota citizens. One result was that 77% of those polled indicated that global warming was a serious or somewhat serious problem.[36] The margin of error for this poll was reported as plus or minus 5%.

What this margin of error indicates is that we can be 95% sure that the 77% poll result obtained from the sample of 410 citizens is less than 5% away from the actual percentage of all South Dakota Citizens who would indicate that global warming is serious or somewhat serious. In other words, we can be 95% sure that the population parameter this poll was attempting to estimate was between $77 - 5 = 72\%$ and $77 + 5 = 82\%$. This range, from 72% to 82% is called the (95%) **confidence interval** for this result.

One can compute margins of error and confidence intervals for all kinds of different sampling situations, giving different types of confidence intervals. All these formulas are based on the definition of standard deviation we looked at earlier. The formula used to determine a 95% confidence interval for a population proportion like the one above is

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

[36] *http www.dwu.edu* /press /2007 /jun13.htm

In this formula, $\hat{p}$ is the sample proportion and n is the sample size.[37]

This formula actually determines two numbers, namely $\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and $\hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. These two numbers determine a **confidence interval** centered at $\hat{p}$ by giving the lower and upper end-points of the interval. The number $2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the value of the **margin of error**. There is a 95% probability that the interval we obtain in this way will contain the population proportion.

For Dakota Wesleyan's global warming poll, the margin of error was calculated as

$$2\sqrt{\frac{.77(1-.77)}{410}} = 2\sqrt{\frac{.1771}{410}} \approx 0.042$$

or 4.2%. Adding and subtracting this from 77% we get a confidence interval of 72.8% to 81.2%.

Why did DWU report a 5% margin of error? As is typical for most polls, this poll considered a number of different questions and reported different percentages for each of these. For percentages different than 77%, the margin of error would be different than 0.042 or 4.2%. The worst possible margin of error for a poll is when the sample proportion $\hat{p} = .50$. In this case, the margin of error is

$$2\sqrt{\frac{.5(1-.5)}{410}} = 2\sqrt{\frac{.25}{410}} \approx 2(0.0247) = 0.049$$

which is about 5%. So, all other margins of error for a poll like this with a sample size of $n = 410$ will be smaller than 5%. Rather than provide separate margins of error for each sample proportion, pollsters typically provide the "worst case" margin of error, knowing that the actual individual margins of error will be no worse than this.

Example 5. The McGovern Center poll mentioned above included 72 individuals from Minnehaha county, which is where Sioux Falls is located, South Dakota's largest city. We could consider this a random sample of the population of Minnehaha county. Of these 72, 56 considered global warming a serious or somewhat serious problem. Find the sample proportion and the margin of error for this poll. Do Minnehaha county residents consider global warming a more serious problem than South Dakota citizens as a whole?

Solution: We know that $\hat{p} = 56/72 = 0.78$ and $n = 72$. Substituting these values into our formula, we get a margin of error of

$$2\sqrt{\frac{0.78(1-0.78)}{72}} \approx .0976$$

or about 10%.

The 78% sample proportion is only 1 percent higher than the 77% for all South Dakotans. Given the 10% margin of error for Minnehaha county and the 5% margin of error for the state, we would say there is essentially no difference between these two poll results. After all, remember that there is a 95% chance that the actual result will be within the margin of error. Since we are just as likely to overestimate the actual result as underestimate, this means, for example, there is a 47.5% chance that the Minnehaha county result will be between 68% and 78%. Thus, the Minnehaha county result could easily be below 77%, not even taking into account that the actual state result has a good chance of being higher than 77%. Thus, there will be a good chance, close to half in fact, that Minnehaha county residents are actually slightly less concerned about global warming than other South Dakotans. ■

[37]Most statistics books use the more precise value of 1.96, rather than 2, in this formula. However, the error introduced by rounding off to 2 is not important in most cases.

Weighted Averages

Welcome to Lake Wobegon, where all the women are strong, all the men are good-looking, and all the children are above average
— Garrison Kiellor

A common concern among college students, particularly as they see graduation approaching, is their GPA or Grade Point Average. Students perceive that this one number has the potential to greatly affect their futures. It can determine whether or not a student gets into medical school, law school, or graduate school. It may influence future prospective employers who are holding the keys to that "dream job." For many student-athletes, their GPA is a big factor in whether they play or not. It is used at most schools as the benchmark by which those who graduate with honors are separated from those who don't.

The GPA is what is known as a **weighted average**. A weighted average is similar to the ordinary mean, except some of the data points being averaged count more (have more weight) than others. In the case of the GPA, some classes count more than others, depending on how many credits those classes carry.

You may already be concerned enough about your GPA to have computed it yourself. Here is an example. Suppose a student has courses and letter grades as shown in Table 1.

Table 1: Grade Information for a Hypothetical Student

Course Title	Credit Hours	Letter Grade	Grade Points	Quality Points
General Biology	4	C	2.0	8.0
Foundations of Ed	2	B+	3.3	6.6
U.S. History	3	A-	3.7	11.1
Statistics I	3	B	3.0	9.0
Tennis & Golf	1	A	4.0	4.0
Human Development	3	C-	1.7	5.1
TOTAL	16			43.8

The grade points corresponding to the letter grades are given in the fourth column. The rightmost column shows what are called Quality Points in Registrar's lingo. These are the products of the course credits and the grade points for each class. The GPA is then calculated by adding up the Quality Points for each class to get the total quality points, and then dividing by the number of credits. For this student, the GPA is $43.8/16 = 2.738$. (Registrars typically round the GPA to three decimal places).

What if all the courses had been weighted the same? In other words, what if all the courses were the same number of credits, or if we just ignored the credit hours? In this case, we would just add up the grade points for each class and divide by 6. The result would be

$$\frac{2.0 + 3.3 + 3.7 + 3.0 + 4.0 + 1.7}{6} = 2.95$$

Why is this higher than the actual GPA? Looking at the grades, we see that the student's best grade was in a 1 credit course, and their worst two grades were in 3 and 4 credit courses. When computing the GPA, these lower grades were weighted 3 and 4 times as much as the A in Tennis & Golf. When we ignore the weighting, the A counts as much as the C or the C-, and so the average gets raised.

Another way to calculate the GPA is as follows.

$$\frac{4}{16} \cdot 2.0 + \frac{2}{16} \cdot 3.3 + \frac{3}{16} \cdot 3.7 + \frac{3}{16} \cdot 3.0 + \frac{1}{16} \cdot 4.0 + \frac{3}{16} \cdot 1.7 = 2.738$$

Here, we are dividing the credit hours for each course by the 16 total, instead of doing the multiplying and adding first, and dividing at the end. Each grade is "weighted" by the fraction of the 16 total credit hours represented by that class. You could even convert these fractions to percentages. In fact, many instructors use this type of weighted average approach in calculating grades for their courses. Each exam, quiz, paper, etc. is given a percentage weight, and this is multiplied by the grade each student earns on that item, and then all the items are added up.

There are many other examples of weighted averages. Although the methods for calculating these might be modified from the basic example above and be more complicated as a result, they all have in common the idea of weighting. Oftentimes, numbers that are the result of a weighted average type of calculation are called **indexes** (singular index, and sometimes the plural is indices). Examples include:

1. Stock indices, like the Dow Jones Industrial Average, the NASDAQ Composite Index, and the S&P 500 Index.

2. Football quarterback ratings.

3. The Consumer Price Index (CPI) of inflation.

4. Many, many other economic indexes, like the Consumer Confidence Index, and the Index of Leading Economic Indicators.

5. The Social Capital Index, introduced in Robert Putnam's *Bowling Alone*.

6. Other social indexes, like the United Nations Human Development Index.

7. Scoring at figure skating competitions

Some of these indexes are "just for fun." However, economic and social indexes are very important, because they are used by governments, companies, and even individual citizens in making decisions. For instance, many companies and organizations (including DWU) base salary raises on the CPI. Investment decisions usually take stock indexes into account. Economic indexes can decide who wins and who loses in presidential elections.

As is the case with the usual arithmetic mean, indexes are designed to give a single numerical measure of a particular variable, based on a larger (sometimes quite large) set of data. For example, the S&P 500 gets part of its name from the fact that it uses data from 500 companies in computing the index. The Consumer Confidence Index is based on responses to five questions on a survey collected from thousands of individuals.

The following examples will give you an idea of how indexes are designed and calculated. As you read through these, think about why the these indexes are designed the way they are. How might they have been designed differently? Does it seem reasonable that the index is actually measuring what it is intended to measure? In the activities for this chapter, you may have the opportunity to design your own index, and you can use these examples as models to follow.

Example 6. One of the most famous averages in existence, The Dow Jones Industrial Average (DJIA), was introduced on May 26, 1896 by Charles H. Dow. At the time,"stocks moved on dubious tips and scurrilous gossip because solid information was hard to come by." Originally, the DJIA was calculated as the average price of 12 stocks hand-picked by Dow. Dow hoped his average would be a good overall measure of how the stock market was performing. [38]

The DJIA has become a bit more complicated over the years. Today, the editors of The Wall Street Journal (WSJ) select a list of 30 large U.S. companies to calculate the DJIA. These companies would be called the **components** of the index. The stocks prices of these companies are averaged, but the average is adjusted to take into account changes that have occurred over time. For example the average is adjusted for "splits," which occur when a company decides to issue two shares of new stock for each old share, with the price of the new shares half of the old price. If one

[38]information on the DJIA, NASDAQ, and S&P 500 indices can be found at www.investopedia.com/ and www.cbot.com

of the 30 stocks is split, then that would lower the price of that stock, and thus lower the average. Also, there have been additions to and substitutions in which stocks have been used to compute the average over the years. For example, when one of the companies goes out of business, or is purchased by another company, this company would be replaced by another as a component. In fact, General Electric is the only company remaining from the original list. When this happens, the stock price for the substituted company is probably not the same as the company that was dropped from the index.

To compensate for these types of changes, the WSJ uses a "Dow Divisor". The sum of the 30 stock prices is divided by this Dow Divisor, rather than the number of stocks (30). If a stock splits, resulting in a lower average, then divisor is also lowered in such a way so that the DJIA does not change as a result of the split. The divisor is also adjusted when changes are made in the companies that are selected, in order that these changes *by themselves* do not produce any change in the actual DJIA index. As of March 1, 2004, the divisor was 0.13500289. The changes that have occurred over the years have substantially decreased the divisor! By adjusting the divisor as changes occur, it is hoped that the DJIA is a good measure of the stock market over time.

Example 7. Since the individual stock prices are not weighted, the DJIA is not really a weighted average. However, most other stock indexes are, including the S&P 500 and the NASDAQ composite. In these indexes, the stock prices are weighted by what is known as **(market)-capitalization**. The market-capitalization of a company is the total value of all its stock, and is one measure of what a company is worth, overall. It can be computed by taking the current stock price times the current number of shares. For example, a company with 2 million shares and a stock price of $50 per share has a capitalization of $100 million. Another company with 10 million shares and a stock price of $25 has a capitalization of $250 million.

An index that is weighted by market-capitalization would take the stock price times the capitalization for each company that is a component of the index, add these together, and then divide by the total market capitalization of all the companies. Thus, companies that have a higher overall total worth are weighted more in the index since there stock price gets multiplied by a larger number. For example, suppose we have only two companies represented in our index, company A with a capitalization of $5 million and a stock price of $20, and company B with a capitalization of $15 million and a stock price of $10. Our index value would the be

$$\frac{\$5 \cdot 20 + \$15 \cdot 10}{5 + 15} = \frac{250}{20} = 12.50$$

Notice this is smaller than the average of the share prices, which would be $(20 + 10)/2 = 15$. This is because the company with the lower stock price has the larger capitalization, and so is weighed more in the index, making the index lower. If company B loses a lot of capitalization and its total capitalization goes down to $5 million (same as company A), but it still has the same stock price of $10, the index would be

$$\frac{\$5 \cdot 20 + \$5 \cdot 10}{5 + 5} = \frac{150}{10} = 15.00$$

the same as the average of the two prices.

A capitalization weighted index is considered a much more accurate reflection of the overall stock market than the DJIA. However, because of tradition, the DJIA is still the most prevalent.

As a final caution, consider the following quote. Although indexes can be very useful, we should usually be wary of placing too much credence on any single number, no matter how well it seems to reflect what we want to measure.

> *There are difficulties with all economic statistics, but the problems do not arise because the people who designed them were stupid or lazy. The problems arise because we are trying to use a single number to summarize a phenomenon more complex than any single number can report.*
> — Robert Schenk, Professor of Economics, St. Joseph's College, Indiana

Example 8. As a less serious example of an index, let's consider the ubiquitous quarterback rating system that is used for both NFL and college football quarterbacks. We will refer to it as a rating instead of an index, although the idea is the same. The NFL would remind us that this rating is really only designed to measure the passing proficiency of the quarterback, not his overall performance. Here's how it works.

The rating has four components.

1. Percentage of completions per attempt C
2. Average yards gained per attempt Y
3. Percentage of touchdown passes per attempt T
4. Percentage of interceptions per attempt I

The rating formula is designed so that high values for C, Y, and T increase the rating, while high values of I (interceptions are bad) lower the rating. With a couple of caveats,[39] the rating is computed using the following formula.

$$\textit{Rating} = \frac{100}{6}(0.05(C-30) + 0.25(Y-3) + 0.2T + (2.375 - 0.2I))$$

The caveats are that if any of the four terms $0.05(C-30)$, $0.25(Y-3)$, $0.2T$, or $(2.375-0.2I)$ are bigger than 2.375, or less than 0, then substitute the numbers 2.375 or 0 respectively.

As an example, suppose that a quarterback has a completion percentage of $C = 65$, an average yards per attempt of $Y = 13.2$, a percentage of touchdowns per attempt of $T = 8$ and a percentage of interceptions of $I = 2$. Putting these into the formula, we get

$$\frac{100}{6}(0.05(65-30) + 0.25(13.2-3) + 0.2\cdot 8 + (2.375 - 0.2\cdot 2))$$

$$= \frac{100}{6}(1.75 + 2.55 + 1.6 + 1.975)$$

Since the 2.55 is bigger than 2.375, we replace it with 2.375 and get

$$\textit{Rating} = \frac{100}{6}(1.75 + 2.375 + 1.6 + 1.975) = 128.3$$

Our hypothetical quarterback is very, very good. In practice, it is very hard to get one of the terms to be larger than 2.375, at least over a whole season, although it does happen when considering only single games. In our case, we put in 13.2 yards per attempt to get the 2.55 term. The best quarterbacks in the league typically do not average even 10 yards per attempt over a season. In 2003, the best season long rating was 100.4 by Steve McNair. The highest possible rating is 158.

Let's look again at the formula.

$$\textit{Rating} = \frac{100}{6}(0.05(C-30) + 0.25(Y-3) + 0.2T + (2.375 - 0.2I))$$

The weighting is achieved chiefly by the numbers 0.05, 0.25, 0.2, and 0.2. In fact, these weights and the subtractions that occur are actually designed to give roughly equal weight to each of the four components. To achieve the top rating, a quarterback must make all the component terms equal to 2.375.

Notice the I is multiplied by a negative number. This reflects that a quarterback wants this number to be as low as possible, rather than as high as possible. The 100 and the 6 are present merely because the NFL wanted a rating of 100 to represent a very good quarterback (one who is "100%"). The NFL could have left these numbers out, and the result would have been that excellent passers would have ratings of 6 or more instead of around 100 or more.

[39] or exceptions

Averages of Averages

Before we leave behind discussion of averages, there is one more item that is important to discuss. What happens if the data that you are averaging are themselves averages?

Suppose you own a small business that operates in two locations, A and B. Suppose the average compensation that you provide for employees at A is $20,000, and at B it is $30,000. The average of these two numbers is $25,000. Does this mean that the average compensation for all your employees is $25,000?

Not necessarily! There is an important piece of information that is missing, and that is how many employees work at each location. If both locations had the same number of employees, than the answer is yes. Suppose, however, that there are 8 employees at location A, and only 2 at location B. Then the total amount of compensation received by all 10 employees is

$$8(\$20,000) + 2(\$30,000) = \$220,000$$

and so the average is $\$220,000/10 = \$22,000$. It is lower than the $25,000 since there are more employees at the location that has the lower average salary. If the situation were reversed and there were 8 employees at B and only 2 at A, then the average for all employees would be $28,000.

The moral of the story is that you need to beware when you are "averaging averages." It may sound silly, but remember that

the average of the averages is not necessarily the average

This idea will be relevant in many of the situations we consider. For example, we have already presented some data in graphical form based on the 50 states. If our data is in the form of averages, say average household income in each state, then the average of these 50 numbers is not the same as the average household income of all households in the U.S.

If a large state (like California, which has about one fifth of all the households in the U.S.) has a lower income than most states, than it will cause the overall U.S. average to be lower than the average of the 50 state averages, in the same way that the employees at location A (which had four fifths of the employees) caused the overall average for the 10 employees to be lower than the average of the two location averages. The only difference is we now have 50 averages that we are averaging, instead of just two.

In fact, the same idea of "weighting" that we introduced with the GPA example applies here. When averaging the 50 state income averages, the California data point is weighted equally with all other 49 states. However, if we actually count the households in each state, than the California income number would be weighted more than all the other states, about one fifth of the total weight. A small state like South Dakota would have a weight of about 1/375, based on the fact that about 1 out every 375 households in the U.S. is in South Dakota.[40]

[40] So, unfortunately, in some sense we don't count for much. On the other hand, there are advantages, since we don't see nearly the number of negative political ads as they do in California

Reading Questions for Collecting and Analyzing Data

1. What does it mean for a sample to be a simple random sample?

2. What does it mean for a sampling technique to be biased?

3. A very conservative talk radio host decides to conduct a survey in his city. He hires a local survey firm to conduct a random sample of 500 residents of the city. One of the survey results is that 48% of those surveyed support a proposed constitutional amendment outlawing gay marriage. Is this sample biased? Why are why not?

4. A survey firm conducts surveys on customer satisfaction with local cable service every two weeks for an entire year. Assume that the actual level of customer satisfaction is approximately constant 64% over the whole year. The 26 survey results were as follows:

$$52, 66, 63, 75, 59, 84, 68, 56, 49, 67, 60, 72, 71, 66, 54, 61, 58, 63, 79, 64, 68, 45, 58, 73, 66, 70$$

(a) Does this situation show high bias, low bias, or is it impossible to tell? Explain.

(b) Does this situation show high variability, low variability, or is it impossible to tell? Explain.

(c) Does this survey firm seem to be producing reliable data? Why or why not?

5. An elementary teacher has her class measure the height of each student in "hands." She pairs off the students. Each student measures the other student in their pair using their hand width, counting the number of hand width they need to go from foot to head.

(a) If this measuring technique is used repeatedly, mixing up the pairs each time, would the technique produce reliable data? Why or why not?

(b) Does this technique produce valid data? Why or why not?

(c) Suppose that, instead of pairing students up, the teacher picks one student to measure all the other students. Would this technique produce valid and reliable data?

6. If you repeat a sample several times and get very similar answers, you have (low, high, can't tell) variability and (low, high, can't tell) bias.

7. A local environmental organization likes to include polls in its fund-raising mailings because it seems to increase response rates and the amount raised. Often, they ask pretty much the same questions from one poll to the next. The polls get sent to everyone on the organizations mailing list. One poll question that has been asked several times is whether congress should pass laws to mandate a reduction in greenhouse gas emissions. The results from the polls are consistently that 75% to 80% of the membership agree with this statement.

(a) Does this situation shows high bias or low bias? Why?

(b) Does this situation show high variability or low variability? Why?

(c) Does the sampling technique used in this situation produce simple random samples?

8. Identify the population and the sample in each of the following situations.

(a) A biologist wants to know the mercury level present in a certain species of fish in Lake Catalano. Over the course of two days, she traps 37 fish live, collects a blood and tissue sample from each, and returns the fish to the lake. She finds that 17% of the fish sampled have mercury levels higher than recommended by the EPA.

(b) On July 31st, 2006, CNN conducted a web poll asking if people felt the term ''tar baby,' used by Massachusetts Governor Mitt Romney to describe Boston's 'Big Dig' construction project, was a racist term. Of 63,240 votes registered as of 3:30 p.m. CDT, 45% responded 'yes.'

9. Find the median for each of the following.

(a) $76, 52, 46, 55, 49, 80$

(b) $78, 83, 96, 58, 95, 87, 91$

(c) What did you do differently for part (b) as compared to part (a)?

10. A fund-raiser for an elementary school organization brings in the following amounts of money:

$$\$60, \$125, \$80, \$25, \$130, \$90, \$50, \$75, \$40, \$140$$

All the money is to be used for a trip and the adult leader decides that each of the 10 children will get the same amount of money. How much does each child receive?

11. List 8 numbers that have exactly one outlier.

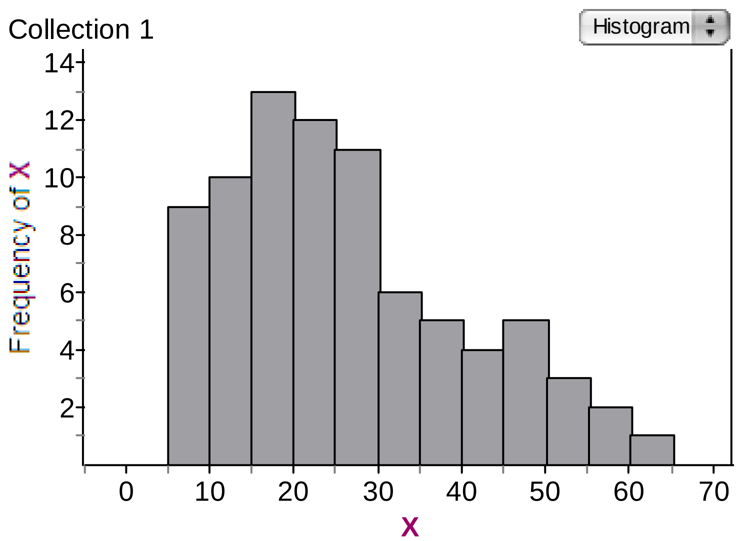
12. List 6 numbers whose mean is smaller than the median.

13. List 10 numbers whose histogram will be skewed to the right. Sketch the histogram for your data.

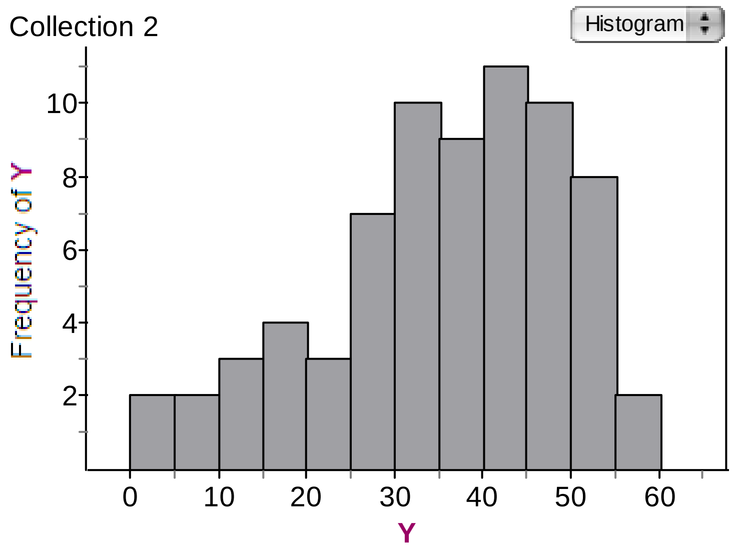
14. Suppose there are nine people working in a dentist's office. The three people that keep track of the patients and the books each earn \$36,000 per year, the four dental hygienists each earn \$42,000 per year, and the two dentists each earn \$106,000 per year. What is the mean and the median salary of people working in this office? Which is the better measure of center in this case? Explain.

15. When is it better to use the median instead of the mean as a measure of center?

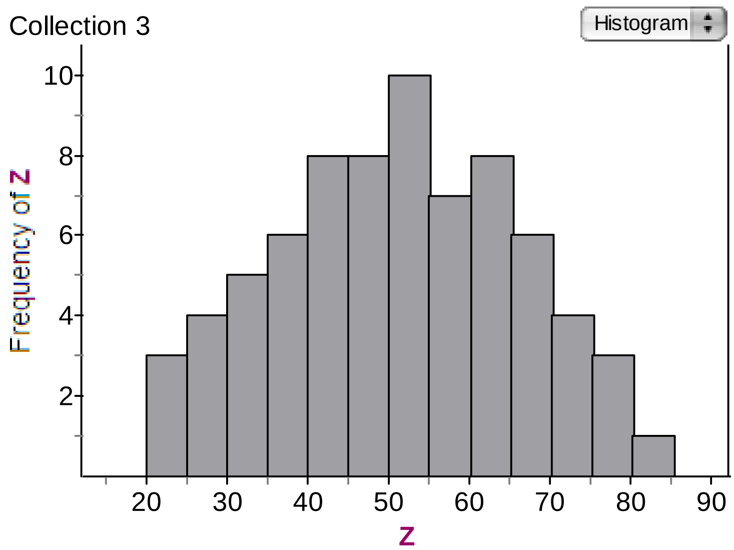
16. Points X, Y, and Z are labelled on the following histograms. Determine if the labelled number represents the mean, median, both, or neither.



(a)



(b)



(c)

17. What does it mean to be at the 80th percentile for ACT scores?

18. Find the standard deviation and the interquartile range for the following sets of data (you may use a calculator or software):

(a) $20, 20, 21, 23, 26, 27, 27, 30, 31, 31, 32$

(b) 10.1 10.8 12.6 13.1 14.3 14.4 14.4 14.6 14.7 15.1 16.0 16.3 16.8 18.1

19. Consider the histogram on University of Maryland faculty salaries given in Figure 8 of Chapter Two.

(a) Looking at the histogram, estimate as accurately as you can the minimum, first quartile, median, third quartile, and maximum.

(b) Looking at the histogram, and considering its skew, estimate the mean salary for University of Maryland faculty.

20. Consider the histogram on National Poverty Rates given in Figure 10 of Chapter Two.

(a) Looking at the histogram, estimate as accurately as you can the minimum, first quartile, median, third quartile, and maximum.

(b) Looking at the histogram, and considering its skew, estimate the mean national poverty rate.

21. A student has scored 78, 89, 86, and 72 on the first 4 tests in her accounting class. There will be one more test, and all tests are out of 100 points. Her grade will be based on the average of all 5 tests.

(a) What is the highest average this student can possibly achieve?

(b) What is the lowest average this student can possibly achieve?

(c) What score does she need to get on the last test to get an average of 80?

22. From Figure 10 in Chapter Two, carefully estimate both the mean and the median National Poverty Rate for the countries represented.

23. The following table shows childhood mortality data for a number of Eastern European countries for the year 2001. The data is the number of deaths of children five and under per 1000 live births. [41] Use this to answer the following questions.

Country	Deaths per 1000 births	Country	Deaths per 1000 births
Czech Rep.	5	Bosnia and Herzegovina	18
Germany	5	Yugoslavia	19
Slovenia	5	Belarus	20
Croatia	8	Ukraine	20
Hungary	9	Latvia	21
Lithuania	9	Romania	21
Poland	9	Russian Federation	21
Slovakia	9	TFYR Macedonia	26
Estonia	12	Albania	30
Bulgaria	16	Moldova, Rep. of	32

(a) Without doing any calculations, which of the following is most likely to be true? Briefly justify your answer.

- The mean will be greater than the median.
- The mean will be less than the median.
- The mean will be almost equal to the median.

(b) Compute the mean and the median for this data.

(c) Compute the standard deviation for this data.

(d) Draw a histogram for this data, and then mark the mean and median along the horizontal axis.

(e) Compare your histogram with the mortality histogram in Figure 8 of Chapter 2. How do mortality rates in Eastern Europe differ from mortality rates for the world overall?

[41] Data is from the UNICEF end of decade database at http://www.childinfo.org/eddb/malnutrition/index.htm.

24. As noted in the text, one must be careful to interpret what an average means when the numbers being averaged are based on different sized populations.

(a) Find the average of the mortality rates for the four countries Estonia, Latvia, Lithuania, and Russia.

(b) Now, use the table below, which gives the total number of births (in thousands) in each of these countries, and the information in the previous question to find the total number of child deaths in each country.

Country	Live Births (in thousands)
Estonia	30
Lithuania	12
Latvia	18
Russia	1230

(c) Now, find the total number of deaths in the four countries combined and the total number of births (in thousands) combined. Use this to find the overall child mortality rate in these four countries.

(d) Why is the overall rate from part (c) so much bigger than the average of the four rates you found in part (a).

25. Using your calculator or software, find the median, quartiles, mean, and standard deviation of the following frequency distribution of student heights. Write a brief sentence describing each of the results.

Height	66	67	68	69	70	71	72
Frequency	1	3	5	8	7	3	2

26. Suppose we added two individuals to the height data in the previous question, one who is 67 inches tall, and one who is 81 inches tall. How will these additions change the mean and median of the data? Explain how you can answer this question without necessarily recalculating both values.

27. Consider the quote from Rush Limbaugh a the beginning of this chapter. This quote occurred within a section of his show where he was arguing that middle class Americans had not suffered during the so-called "decade of greed", the 1980's, during which Republicans Ronald Reagan and George Bush (senior) were President.

(a) Explain why it is reasonable to assume that the distribution of family incomes in the U.S. is skewed to the right.

(b) Given that incomes are skewed to the right, which is larger, the median family income or the mean family income?

(c) Explain why it possible for the mean income to increase over time, while the median income remains the same.[42]

[42] In fact, according to The Statistical Abstract of the United States, from 1982 to 1989 real median income for all U.S. households went up 10.4%. Mean Per Capita Income went up about 20%. This number is a mean, since it is per capita, and the denominator is the total population. The median computation is done per household. By way of comparison, households at the 95th percentile (those earning more than 95% of all households) had their income go up by 17%, while those at the 20th percentile went up 8%, over the same period.

28. According to Bread for the World, a hunger advocacy organization, the average U.S. farm family earned $64,117 in 2002, almost $6000 more than the average American family. However, the same publication also notes that half of all farmers earn less than $6000 per year in farm income, and many farmers make most of their family income from non-farm work.[43]

(a) Do you think the $64,117 figure is a mean or a median? Explain why you think so.

(b) Suppose that, based on the information above, someone made the claim that "The average American farmer earns $6000 per year from their farming operation and $58,117 in non-farm related income." Explain carefully why this statement is probably not correct.

29. In the poll mentioned in Example 4 in the text, there were 20 respondents from the 7 county area around Mitchell, South Dakota. Of these, 16 indicated they thought global warming was a serious or somewhat serious problem.

(a) Give a 95% confidence interval for the proportion of all the people in the Mitchell area that think global warming is a problem.

(b) State carefully what this confidence interval means in terms of probabilities.

30. Suppose a student scores 82% on quizzes, 76% on tests, 89% on his project, and 78% on his final. Suppose these are weighted 25%, 40%, 15% and 20% respectively. His final grade percentage is the weighted average. Find the student's final grade percentage.

[43]Statements are from Bread for the World's 2003 annual report, *Agriculture in the Global Economy*

Collecting and Analyzing Data: Activities and Class Exercises

1. **The Curse Reversers.** Below is a table showing the salary for members of the 2004 Red Sox, the only team to come back from a 3 games to none deficit in a championship series in baseball history.[44] They are also given credit for reversing the 'curse of the Bambino,' placed on the Red Sox after then owner Harry Frazee sold Babe Ruth to the New York Yankees in 1920 for $100,000. The 2004 team beat the curse by being the first Red Sox team since 1918 to win the World Series.[45]

Player	Salary	Player	Salary
Ramirez, Manny	22,500,000	Embree, Alan	3,000,000
Martinez, Pedro	17,500,000	Timlin, Mike	2,500,000
Schilling, Curt	12,000,000	Mueller, Bill	2,100,000
Garciaparra, Nomar	11,500,000	Reese, Pokey	1,000,000
Damon, Johnny	8,000,000	Mirabelli, Doug	825,000
Varitek, Jason	6,900,000	Burks, Ellis	750,000
Ortiz, David	4,587,500	Kapler, Gabe	750,000
Lowe, Derek	4,500,000	Daubach, Brian	500,000
Nixon, Trot	4,500,000	McCarty, David	500,000
Wakefield, Tim	4,350,000	Bellhorn, Mark	490,000
Mendoza, Ramiro	3,600,000	Arroyo, Bronson	332,500
Foulke, Keith	3,500,000	Crespo, Cesar	309,000
Kim, Byung-hyun	3,425,000	Shiell , Jason	303,000
Millar, Kevin	3,300,000	Garcia, Reynaldo	301,500
Williamson, Scott	3,175,000	DiNardo, Lenny	300,000

   (a) Without doing any calculations, decide if the median salary or the mean salary will be higher. Explain how you determined your answer.

   (b) Calculate the mean, median, and standard deviation of the salaries for the 2004 Red Sox using a calculator or software.

   (c) Assume players on the Red Sox are paid according to their abilities. What should a player of average ability expect to get paid? Explain your answer.

   (d) What percent of Red Sox players made below the mean salary? What percent made below the median salary?

   (e) Suppose each player on the Red Sox was given a $50,000 pay raise. Without recalculating, determine the new mean, median, and standard deviation. Explain how you determined your answers.

   (f) What percent of the payroll was earned by Manny Ramirez? Starting at the bottom of the pay scale, how many players does it take to get a total salary equal or higher than Manny's?

[44] The salary data is from <http://asp.usatoday.com/sports/baseball/salaries/playerdetail.aspx?player=1684.>

[45] For more information on 'the curse', see <http://www.bambinoscurse.com/whatis/.>

2. **Hungry Planet II.** The table below gives some of the data for the 30 families surveyed in *Hungry Planet*. For each family, the table gives the country, number of adults, children, and total persons, the total weekly food costs for the family, and the cost per person.

Country	Adults	Children	Persons	Weekly Cost	Per Person
Australia	3	4	7	376.45	53.78
Australia	2	2	4	303.75	75.94
Bhutan	6	7	13	34.09	2.62
Bosnia	2	3	5	157.43	31.49
Chad	1	5	6	25.59	4.27
Chad	2	7	9	43.77	4.86
China	3	1	4	155.06	38.77
China	5	1	6	59.23	9.87
Cuba	2	2	4	69.92	17.48
Ecuador	2	7	9	34.75	3.86
Egypt	5	6	11	68.53	6.23
France	3	1	4	419.95	104.99
Germany	2	2	4	500.07	125.02
Great Britain	2	2	4	253.15	63.29
Greenland	2	3	5	498.1	99.62
Guatemala	2	6	8	79.82	9.98
India	2	2	4	39.27	9.82
Italy	2	3	5	260.11	52.02
Japan	2	2	4	317.25	79.31
Okinawa	3	0	3	221.51	73.84
Kuwait	4	4	8	221.45	27.68
Mali	4	11	15	26.39	1.76
Mexico	2	3	5	189.09	37.82
Mongolia	2	2	4	40.02	10.01
Philippines	5	3	8	49.42	6.18
Poland	4	1	5	151.27	30.25
Turkey	3	3	6	145.88	24.31
United States	2	2	4	159.18	39.80
United States	2	2	4	341.98	85.50
United States	3	2	5	242.48	48.50

(a) Find the five number summary for the weekly per person costs. Write a few sentences describing what these data tell you about what it costs to feed a person for a week.

(b) Note that the average of the per person costs represents an average of the family averages. By summing up the number of persons in each family, we get that there are 183 people in these 30 families. What is the average cost per person for these 183 people, and why is it different than the average you found in part (a)?

(c) Find both the average of the family sizes and the average number of people per family and explain why these are the same.

(d) Does this data support the statement that large families tend to spend more *total* on food than small families? Use graphs or appropriate statistics to support your answer.

(e) Explain as thoroughly as possible why the data represented by these 30 families is probably not representative of average food costs world wide.[46]

[46] As a hint, note that China and India are the most populous countries in the world, each with over 1 billion people. The United States currently has around 300 million people.

3. Homelessness. We noted in the text that reports that the average age of a homeless person is nine have been widespread. Adam Molnar[47], a mathematics professor at Bellarmine University, found that this claim originated with ads created by the Coalition for the Homeless, which bills itself as the oldest homeless advocacy and direct service organization in the U.S.[48]

(a) Create a set of five ages where the average age is nine, and then discuss how likely you think it would be to find a group of five homeless people with these ages.

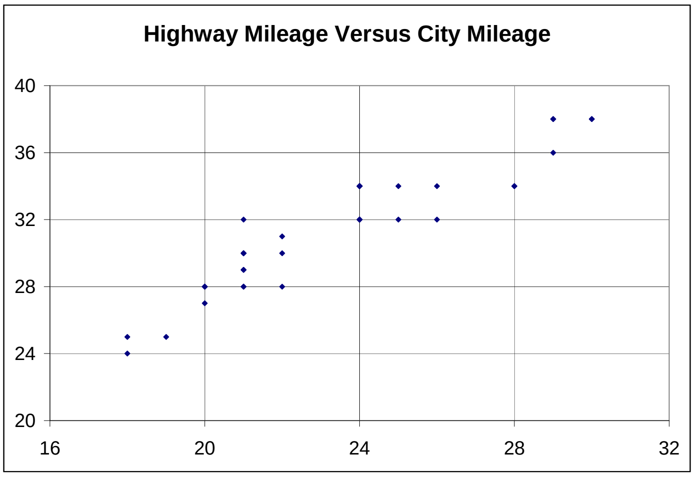
(b) Create a set of five ages where the average age of the five people is nine, and one of the people is 40 years old. Would you be likely to find a group of five homeless people with these ages?

(c) Using Fathom, create a data set of 25 ages and an associated histogram so that the average of the 25 ages is nine. Use at least 10 different ages, and include at least 5 adults in your histogram.

(d) Describe the skew and other characteristics of your histogram.

(e) How likely do you think it is that this claim is true? Describe how you might go about determining whether or not it is true.

4. City Versus Highway. The graph below represents the relationship between Highway Mileage and City Mileage for 40 different car models.



(a) What type of graph is this?

(b) Even without the graph heading, how can you tell that highway mileage is represented on the vertical axis?

(c) What are the highway gas mileages for cars which get 21 miles per gallon in the city?

(d) Every car gets better mileage on the highway than in the city. What are the highway and city mileage ratings of the models that have the smallest difference between city and highway?

(e) What are the highway and city mileage ratings of the models that have the largest difference between city and highway?

(f) There are only 22 points shown on the graph, even though the graph represents 40 different models. How is this possible?

[47]see http://mathematicsafterthefall.twelvefruits.com/ for Adam's blog entry on this claim.
[48]http://www.coalitionforthehomeless.org/home.asp

5. Gross Domestic Product. Gross Domestic Product. Below is a table showing the Per Capita Gross Domestic Product (GDP) for a number of countries for the year 2002. The Gross Domestic Product is an estimate of the total value of all the goods and services produced in a country, and is considered one measure of the total wealth of a country. Per Capita GDP (PC GDP) is the GDP divided by the population, and is considered a measure of the productivity of a country.

Country	PC GDP
United States	36123
United Kingdom	26376
Trinidad and Tobago	7109
Venezuela	3760
Uruguay	3645
Turkey	2626
Tunisia	2163
Yugoslavia	1459
Turkmenistan	1383
Tonga	1345
Ukraine	849
Zimbabwe	640
Yemen	559
Vietnam	436
Uzbekistan	383
Zambia	352
Uganda	251

(a) Without doing any calculations, use the table to decide if the median PC GDP or the mean PC GDP will be higher. Explain how you determined your answer.

(b) Now, calculate the mean, median, and standard deviation of the PC GDP. How many countries lie within one standard deviation of the mean?

(c) Consider how you would set up a histogram for this data. Without drawing it out, explain how the histogram would look (you could create a frequency chart to help). Why is a histogram probably not a very good way to display this data?

(d) List the minimum (smallest number), 1st quartile, 2nd quartile, 3rd quartile, and maximum (largest number) for this data.

(e) Determine which countries are outliers with respect to the PC GDP, and explain why. Without doing any calculations, discuss how the mean and median would change if these countries were omitted from the data.

(f) Recalculate the mean, median, and standard deviation without the countries that you identified as outliers. How do these calculations compare with your conjectures from part e?

(g) Based on what you have done so far, explain in a few sentences what this PC GDP data indicates. Mention particular countries as you feel appropriate.

6. **Textbook Prices.** College textbooks can be both fat and expensive. The table below gives data on 13 different calculus texts. [49]

Author	Pages	Price
Stewart	1127	180.95
Larson	1138	123.35
Larson	760	119.31
Stewart	1336	147.62
Hughes-Hallett	528	97.35
Hughes-Hallett	688	117.75
Lial	864	126.67
Bittinger	673	121.33
Hass	922	106.67
Foerster	778	100.57
Berresford	672	111.49
Varberg	864	102.01
Goldstein	736	104.14

(a) Sketch a scatterplot of the data. How would you describe the relationship between price and the number of pages in a calculus textbook?

(b) How do you think the relationship would be different if we picked a different subject, like literature for example?

(c) How would you expect the relationship to change if picked textbooks from several different subjects? Explain.

(d) What is the standard deviation for the price data?

(e) Based on this data, how much would you budget for a calculus text if you were planning on taking the class next fall, not knowing which textbook would be required? Explain your reasoning.

(f) Estimate how much a page of calculus text costs, based on these data. Be sure to explain how you arrived at your answer.

[49] Data acquired from amazon.com, August 14th, 2007

7. Basketball Salaries. The table below lists the teams in the National Basketball Association along with their total team payrolls in millions of dollars and the winning percentage for the 2004-2005 season.[50] We want to determine if the team owners are getting their money's worth from their highly paid players.

Team	Payroll (in millions)	04-05 Winning Percentage
New York	102.44	0.402
Dallas	91.55	0.707
Portland	83.67	0.329
Philadelphia	71.95	0.524
Minnesota	70.06	0.537
Memphis	67.01	0.549
Orlando	66.45	0.439
Indiana	65.79	0.537
L.A. Lakers	65.06	0.415
Boston	64.58	0.549
Sacramento	61.81	0.61
Toronto	61.70	0.402
Houston	60.22	0.622
Miami	58.95	0.72
Chicago	57.28	0.573
Milwaukee	57.14	0.366
New Orleans	56.57	0.22
Golden State	54.94	0.415
New Jersey	54.73	0.512
Detroit	54.57	0.659
Seattle	53.82	0.634
Washington	49.55	0.549
Cleveland	49.18	0.512
San Antonio	47.15	0.72
Denver	45.62	0.598
L.A. Clippers	45.17	0.451
Phoenix	44.26	0.756
Utah	43.16	0.317
Atlanta	40.68	0.159
Charlotte	23.38	0.22

(a) Looking through the table, does it seem as though the higher paid teams had higher winning percentages? If so, were there exceptions to this rule?

(b) Make a scatterplot of this data with the team payroll on the horizontal axis and the winning percentage on the vertical axis.

(c) Does your scatterplot have a positive or a negative association?

(d) Which team would you say is the best value for the money? Why?

(e) Which team would you say is the worst value for the money? Why?

(f) Make a histogram of the salary data. What skew does this histogram have? What does this tell you about team payrolls in the NBA?

[50]From Patricia Bender's Various Basketball Stuff at <http://www.dfw.net/ patricia>.

8. **Standard Deviation.** Given the set of numbers {6, 7, 8, 9, 10}, you are to choose four of them with repetitions allowed (in other words, you can choose a particular number more than once).
   (a) Give a list of four of these numbers so that the standard deviation is as small as possible. Is your list unique or is there another list of four that will have the same standard deviation? Explain.
   (b) Give a list of four of these numbers so that the standard deviation is as large as possible. Is your list unique or is there another list of four that will have the same standard deviation? Explain.
   (c) How would your answers change if the set of numbers is $\{106, 107, 108, 109, 110\}$ instead of the original set?

9. **Exam Grades.** A college professor states that he will curve exams 10 percentage points if the class average is below 70. Suppose the test scores are:

   75, 51, 60, 97, 90, 30, 62, 89, 83, 78, 25, 77, 80, 70, 82, 59, 83, 85, 61, 92, 78, 82, 90, 77, 78

   (a) What is the median test score?
   (b) What is the mean test score? Will the professor curve the exam?
   (c) What are the percentages for the first and third quartiles? What is the interquartile range?
   (d) What is the standard deviation? What does this tell you about the spread?
   (e) Are there any outliers? If so, list them.
   (f) Do you think the professor should change his syllabus and use the median test score instead of the mean for curving exams? Explain.

10. **Creating an Index I.** In this activity, you will have the opportunity to create your own "car rating index" following the ideas outlined in the text. You will use the Fathom file **Car Index Data**. When you open this file you should see the collection and a short description of the data at the top. Read through these notes to get an idea of the variables/attributes represented in the data.

    On the left are several "sliders" as they are called in Fathom. These can represent either attributes (variables) or **parameters** that can be incorporated into formulas, and can be dynamically changed by grabbing and dragging. Try playing around with the slider labelled A a bit, dragging the grey slider icon back and forth. Notice how the value given in the upper left of the slider box changes.

    (a) Click on the formula in the *Index* attribute (click in the box directly under the attribute name). The **Formula** dialog box should pop up. Write down the formula for future reference. Which attributes and sliders are involved in this formula?
    (b) Notice that most of the attributes are multiplied by a slider value. However, in the formula, we divide by the *AvgPrice* attribute. This is because, everything else being equal, we would rather pay less for a car than pay more. When we divide by the price, the higher the price, the smaller the number we get, and the less our rating goes up because of the price.
    (c) Sort the data from highest to lowest Index value. Which model has the highest rating? Which has the lowest? Which has the median rating?
       i. The highest rated model should be the Toyota Prius. Consider each of the four components that are part of the index. How much does each component contribute to the Index value? To do this, calculate each of the four summands of the formula separately. Which component is giving the biggest contribution towards the Prius' rating?

ii. Now, look at the lowest rated model and explain why it is rated so low, again considering each of the four components and the data for this model.

iii. We could change the values of *Index* by adjusting the sliders. Try changing the values of the sliders so that the model that was rated second moves ahead of the Prius, and becomes the highest rated model. Which sliders did you increase in making this happen, and which attributes are these sliders associated with? Which did you decrease?

(d) The Index formula as it was originally set up involved only four attributes. We could certainly get different ratings if we used more attributes, or different attributes, or changed the values of the sliders.

i. Pick your favorite model. Then, pick the four attributes, of those that are given, that you think you can use to create an Index which will give your car the highest rating. For example, if your model is one of the least expensive models, you probably want to include *AvgPrice* as a component in your formula.

ii. Create a new Index formula, using the four attributes you selected in the previous question as components. You can do this by clicking on the Index formula and editing the formula to include only the attributes you selected. You may also want to adjust parts of the formula to weight certain components more or less (although you can also do this with the sliders).

iii. Now adjust the sliders, trying to get your favorite model to be the highest rated model. Note that it may not be possible to make your model the highest rated model, so just try to get it as high as possible. Once you have done the best you can, record the formula you used and the values of the sliders as well.

(e) Finally, discuss any problems or shortcomings with this method of rating cars, including any deficiencies in the data set. You might consider what other attributes would have been good to include in the data set.

11. Creating an Index II. In this activity, you will pick your own variable to measure and create an index for that variable. You can pick any variable you want subject to the following conditions.

1. Do not use any of the examples outlined in the text.
2. Your index should involve at least four components.

(a) Explain what variable you are measuring, and why you picked it.

(b) Say which components you are including in your index and why you decided on these.

(c) Decide how you are going to measure each component. For a component to be included in a numerical index, you need to be able to assign a numerical measure to it.

(d) Decide how you want to initially weight each component. One way to do this is to use percents as weights, and have all the weights add up to 100%. If, for example, you think a particular component should count as much as all the rest put together, then you could weight it 50%.

(e) Decide if there are any other adjustments to be made to the computation of your index, other than the weights. (See the NFL quarterback rating scheme as an example).

(f) Find at least three actual data points to try out your index. For example, if your variable was "level of democracy in a country", you might pick Sweden, Somalia, and Senegal as three sample countries, and use information for each of your components for these countries.

(g) Consider how your index rated the data points from part e). Were these the results you expected? Do you feel that the data points have been ranked appropriately? Based on the results, would you alter how your index is computed in any way?

(h) Summarize what you feel are the advantages and disadvantages of your index.

12. The 2008 Election. In 2008, the U.S. elected its first President of African-American background.[51] Going into election day, there were a number of states designated as "battleground states" because polls indicated those states were too close to call with any confidence. Most pundits included Florida, Indiana, Missouri, North Carolina, Ohio, and Virginia in this category. The table below gives poll results from shortly before the election, where (O) gives the percentage of prospective voters supporting Obama in the poll, and (M) McCain. These poll results are from Zogby International, have sample sizes of about 600, and can be accessed at `http://www.zogby.com/news/ReadNews.cfm?ID=1632`.

State	Last Poll (O)	Last Poll (M)	Elec. Result (O)	Elec. Result (M)	Elec. Votes
Florida	49.2%	48.0%	51.0%	48.2%	27
Indiana	45.1%	50.4%	50.0%	48.9%	11
Missouri	48.8%	48.8%	49.3%	49.4%	11
North Carolina	49.1%	49.5%	49.7%	49.4%	15
Ohio	49.4%	47.4%	51.5%	46.9%	20
Virginia	51.7%	45.3%	52.6%	46.3%	13

(a) What is the margin of error for these Zogby poll results?

(b) How many of the actual state election results were within the margin of error of the poll results?

(c) Write a short paragraph outlining several reasons why the poll results might have been different than the election results. You should refer to relevant concepts discussed in the text for at least some of these reasons.

(d) How many people would Zogby have to poll to reduce their margin of error to 2%?

(e) Obama won the electoral vote tally 365 to 173. What would the tally have been if McCain had won all six of these battleground states?

[51] Actual election results are available at www.cnn.com; the results we provide in the table are from wikipedia

4. Functions

If you think your teacher is tough, wait until you get a boss. He doesn't have tenure.
— Charles Sykes, author of *Dumbing Down Our Kids*

Every fall, students at thousands of other colleges and universities across the country pay for tuition. The amount each student is charged typically depends on the number of credit hours taken. At Dakota Wesleyan University, in the fall of 2007, this was determined as follows[52]:

1-5 hours $375 per hour
6 hours $2,525 flat rate
7 hours $3,150 flat rate
8 hours $3,850 flat rate
9 hours $4,475 flat rate
10 hours $6,100 flat rate
11 hours $7,725 flat rate
12-16 hours $8,750 flat rate
Overload $375 per hour over 16 hours

If the number of credit hours are known, the tuition charge is exactly known. We say that the tuition charge is a **function** of the number of credit hours. The "number of credit hours" is called the **independent variable**, or the **input**. For a given student, it can vary anywhere from 1 up to as many as 21 or so (although there is no set upper limit, students need permission to register for more than 16 hours and 21 is the most the author has seen). The "tuition charge" is called the **dependent variable**, or the **output**, since how much a given student pays *depends on* how many credit hours that student registers for. Which variable is independent and which is dependent is sometimes given, but sometimes it will be your choice. If you are wondering how to choose, a good idea is to just ask yourself "Does A depend on B, or does B depend on A?" If the answer is neither, then the choice is somewhat arbitrary.

Alternatively, we will sometimes say that the table above describes the function "tuition charge versus number of credit hours," instead of saying "the tuition charge is a function of the number of credit hours." When we describe a function by saying **"A versus B,"** the A is always the dependent variable and the B is the independent variable.

We will see many other examples of functions as we proceed through the course. Functions are one of the most fundamental concepts of mathematics. They are very useful for describing (or what mathematicians and other scientists call **modeling**) all sorts of real-world situations and problems.

Just What Is a Function?

We have already used the same "dependent variable versus independent variable" language when discussing two-variable data, especially scatterplots and other two variable graphs. We discussed how scatterplots can represent negative or positive associations between variables. Functions are a special kind of relationship between two variables.[53]

A **function** is a rule that, for each valid input, assigns one and only one output.

A function is a special type of a more general concept called a relation.

A **relation** is *any* set of ordered pairs. In each ordered pair (x, y) we consider x to be an input and y to be an output.

[52] This is, of course, not including any financial aid
[53] Actually, more than two variables can be involved, as we will see in a few examples later

For example, the description of tuition policy at Dakota Wesleyan given above represents a function where the input is the number of credit hours and the output is the (gross) tuition charge. For any positive number of credit hours we get exactly one tuition charge. If the input is 5, the assigned output is $5 \times \$375 = \$1875$. If the input is 17, representing a 1 hour overload over the standard 12-16 hour load, the assigned output is $\$8,750 + \$375 = \$9,125$.

A number of the examples considered in Chapters Two and Three represent functions. For example, the U.S. gasoline prices given graphically in Figure 4 and numerically in Table 1 of chapter two represent a function where the input is the year and the output is the average gasoline price. For each input year, there is a *single* average price given as the output.

An example of a relation which is not a function would be considering the inputs to be the number of credit hours taken by students during the fall semester at Dakota Wesleyan, but letting the outputs be the amount paid out of pocket for each student taking a given number of credit hours. For example, two students who each take 12 hours might have different amounts paid because of differences in scholarships and other financial aid. For these two students, we might have ordered pairs like $(12, \$3,500)$ and $(12, \$4,200)$. The entire relation would consist of the set of ordered pairs for all Dakota Wesleyan students for the fall 2007 semester.

> This relation is *not* a function because we do *not* have a *single* output assigned to each input. We will surely have two students with the same number of credit hours as the input and two different amounts paid as the output.

Why is the distinction between functions and relations important?

Functions have the advantage over relations in that, for a function, knowing the input exactly determines the output. This is not true for a relation. This advantage is so useful that our standard approach to modelling data will be to create functions which will approximate the data.

Four Ways to Represent a Function

In Chapter Two, we considered two ways to represent data, namely graphically and numerically. For functions, we will consider four possible representations, including these two.

In your previous mathematical experience, you are probably most accustomed to thinking of functions as formulas given in symbols, as in $y = x^2$. In this context, unless otherwise noted, y is the dependent variable or output and x is the independent variable or input.[54] We will also consider three other ways to represent functions. The four ways are:

1. Symbolically, as in $y = x^2$.
2. Graphically.
3. Numerically.
4. Verbally, as in our tuition example above.

Symbolic Representations

As another example of a symbolic representation, consider

$$C = 2\pi r.$$

In this function, the input r represents the radius of a circle, and the output represents the circumference C of the circle. The symbol π is not a variable, but instead is referred to as a constant; $\pi \approx 3.14159$. For any input radius, the rule assigns a single circumference as the output since there

[54] The author is not sure of the origins of this practice, although it has been followed for many years.

is only one way to calculate the formula for any given input. If the input is $r = 4$, the output is $C = 8\pi$. There is no other possible output.

An example of a formula that does not represent a function is $y = \pm\sqrt{x}$. Here, if the input is 4, the output could be either 2 or -2. Since the rule does not assign a single input for each output, the rule does not represent a function.

When determining if a rule represented symbolically is a function, you need to reason whether or not any single input can lead to more than one output. There is no one procedure that you follow – how you determine if the rule is a function will depend on the particular formula given.

Graphical Representations

We have already commented that the graph of U.S. gasoline prices given in Figure 4 of Chapter Two represents a function. As another example, the World Population Growth Rates graph in Figure 1 of the Introduction also represents a function. For each input year, there is one growth rate depicted by the graph.

Typically, it is easy to determine if a rule represented graphically is a function by using the **vertical line test**. If there exists a vertical line that will cross the graph at more than one point, then there is more than one output for the input value where the vertical line crosses the horizontal axis, and so the graph does not represent a function.

An example of a graph which does not represent a function is the graph in Figure 1, depicting the relationship between Per Capita Income for each state versus the number of electoral votes for each state. There are 538 total electoral votes, and every state gets at least 3 electoral votes. Based on the most recent census in 2000, California has the most electoral votes with 55.

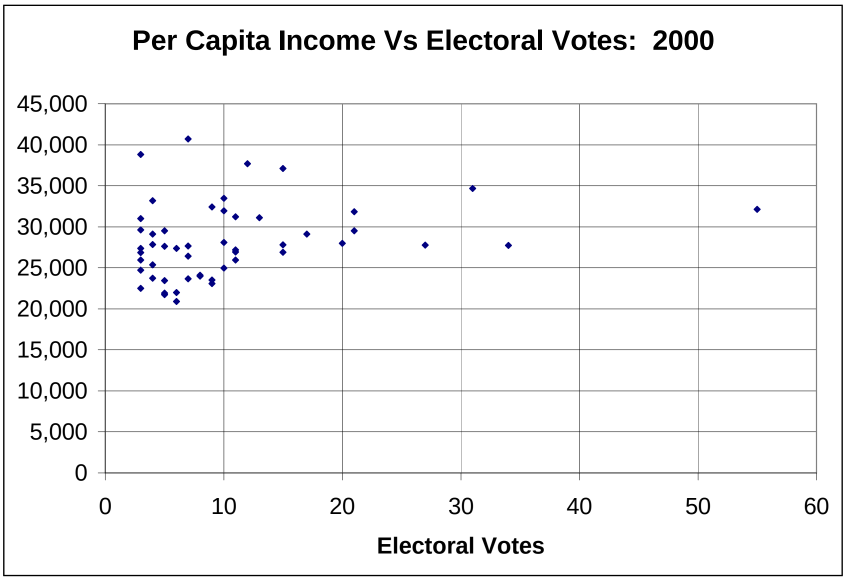


Figure 1: Do people in larger states make more money?

This graph is does not represent a function, since there are quite a number of states with 3 electoral votes (input value 3), and these states have widely varying values for their per capita incomes (output values). So, there is an input value which is *not* matched with a single output, but rather is matched with more than one output. You can probably find other input values besides 3 which are matched with more than one output value.

As long as we have this graph handy, we might ask what this graph tells us about the relationship between population and per capita income. You might recall that the number of electoral votes equals the number of representatives a state has in the House of Representatives plus two (since each state has 2 senators). The number of representatives is based upon population, with each representative representing very roughly 650,000 people.

Notice that small states seem to display a larger variation in incomes. States with fewer than 10 electoral votes include both the poorest and the richest states. However, note that all the states with under $25,000 in per capita income are smaller states. To the extent that larger population states tend to have larger cities, this might be a result of there being higher per capita incomes in urban areas than in rural areas.

Numerical Representations

Checking if a table of data represents a function is fairly easy. You need to check whether an input is paired with more than one output. If this occurs, then the table of data does not represent a function. Table 1(a) represents a function since each input has exactly one corresponding output. However, in Table 1(b), the input 3 is paired with two different outputs (6 and 8), so this does not represent a function.

Table 1: The data in Table 1(a) represents a function while the data in Table 1(b) does not.

Input	1	3	4	6	9
Output	3	6	8	6	5

(a)

Input	1	3	3	6	9
Output	3	6	8	6	5

(b)

Typically, when a function is given in numerical form as a horizontal table, the input variable is in the top row and the output variable is in the second row. If the data is in columns, the input is typically in the first or lefthand column and the output in the second or righthand column.

What if there are more than two columns (or rows) of numbers or values in a table? In this case, it may be that the table describes several functions. For example, Table 2 shows median incomes (adjusted for inflation using 1998 dollars; more on this later) for six different groups of people. So for example, we could consider each of the income columns as outputs for one of six different functions, all six functions having years as the input. One such function could be described as "Median Income of Black Women versus Year." For this function, if the value of the input is 1955, the value of the output would be $3,658.

Table 2: U.S. Median Incomes in 1998 constant dollars

Year	All Races Men	All Races Women	White Men	White Women	Black Men	Black Women
1950	15,989	5,929	16,854	6,595	9,152	2,949
1955	18,809	6,274	19,851	7,013	10,447	3,658
1960	20,653	6,383	21,747	6,844	11,440	4,237
1965	23,940	7,249	25,213	7,688	13,569	5,595
1970	26,325	8,829	27,671	8,943	16,414	8,142
1975	25,677	9,818	26,973	9,919	16,126	9,011
1980	24,816	9,744	26,397	9,798	15,862	9,071
1985	24,709	10,933	25,921	11,145	16,312	9,509
1990	25,308	12,559	26,402	12,867	16,048	10,386
1995	24,131	12,974	25,557	13,173	17,119	11,723
1998	26,492	14,430	27,646	14,617	19,321	13,137

Table 3 gives another example of a function in numerical form.

Table 3: The probability distribution for rolling two dice and recording the sum.

Outcome: Sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

In this function, the inputs are the possible numbers, 2 through 12, that can be the sum when two dice are rolled. The outputs are the *probabilities* for the given sums. If the input is 4, the output is 3/36 or 1/12. This means the probability of getting a sum of 4 when rolling two dice is 1 out of 12.

Verbal Representations

Our tuition example at the beginning of this chapter was one example of a function given in a verbal representation.

As another example, we could describe a function verbally by saying

"Give the output as the number of square miles of land area of the given country"

or

"Tell me the square miles of land of the given country"

Each country is an input, and the output is the number of square miles of land in that country. Using our "versus" language, we could say that the function is "Square miles versus country."

As a third example, "mother versus child" describes a function. Here, "child" is the independent variable, and we define the mother as the biological mother. The identity of the mother depends on the child.

Note that for this function, a given person might serve as both an input and an output. If the author is the input, Marcella Catalano is the output. However, Marcella can also be an input and the output would be Anna (Trnka) Herda. There are lots of functions where particular *values* can serve as either inputs or outputs.

On the other hand, the rule "child versus mother" is not a function. If mother is the input variable, there will very often be several outputs for a given input (anytime a mother has more than one child). For example, if the input is Marcella Catalano, there are four outputs, namely Michael, Gabriela, Aaron, and Naomi.

Sometimes, one needs to be careful in determining exactly what the input and output variables are. For example, for the rule

"Tell me the weight of a person who eats a given number of calories per day, on average."

The output is weight, but the input is not the person but the number of calories the person eats. The person is just part of the rule which associated the calories with the weight. This rule is not a function, since two different people could eat the same number of calories per day (same input) but have different weights (outputs).

If we changed the rule to say

"Tell me the weight of a given person."

then this is a function, since now the input is the actual person, and each person has only one weight (at any given time).

If we have an input and output variable combination that does not give a function, we often call it a **relation**. Any table or any graph in rectangular coordinates represents a relation. Thus, we can think of a function as a relation with the additional condition that each input corresponds to exactly one output.

One reason we use functions in verbal form is that it allows us to consider the dependence of one quantity (the output) on another (the input) before we have a symbolic representation or any numerical information.

Why All These Representations?

We will use all four representations of functions – symbolic, graphical, numerical, and verbal – throughout this book. All four representations are equally valid. There is not a "right" or "wrong" way to represent a function. Rather, there are four equally valid ways of representing the same thing. However, one of the four representations might work best for a particular situation, depending on what the function represents and what you intend to do with it. Consider the functions represented in Table 2 above, where the *input* is the year and the *outputs* are median incomes for various groups. To quickly communicate the number of people in a given year, the table probably works the best. If you want to illustrate how incomes are changing, a graphical representation would be better. If you are interested in describing these functions to someone else, you would probably give a verbal description. However, if you want to use a function to predict incomes five or ten years from now, you will probably want to find a symbolic representation. So while all four representations are equally valid, you may prefer one representation over another to convey certain information.

Function Notation

Functions in symbolic form are often denoted using the so-called **function notation**. In this notation, we let $f(x)$ (pronounced as "f of x") denote the output where x denotes the input. The $f(x)$ is a generic notation, and for a particular function, we give an explicit expression for $f(x)$. The letter f by itself can be thought of as a label representing the function, while $f(x)$ technically only represents the output of the function.

For example, $f(x) = 3x+2$ means "my function, called f, is the rule that multiplies a given input by 3 and then adds 2." There is nothing special about the letter x. It should be thought of as a placeholder or "blank" for your input. So $f(x) = 3x + 2$ should be thought of as $f(\_\_\_) = 3(\_\_\_) + 2$. For example, $f(4) = 3 \cdot 4 + 2 = 14$, $f(-1) = 3(-1) + 2 = -1$, $f(\pi) = 3\pi + 2$. In fact, $f(y) = 3y + 2$ and $f(x + 1) = 3(x + 1) + 2$, which we can simplify to $3x + 5$.

When a function is a modeling function, it is customary to use letters which represent the given quantities. For example, the function for the area of a circle is typically written as $A = \pi r^2$ rather than $f(x) = \pi x^2$ to remind us that the input is the $r$adius of the circle and the output is the Area of the circle.

Notational Issues

A couple of notes on function notation and terminology. When we write $f(x)$, remember $f(x)$ does not mean f times x. The f by itself is not even a *number* but rather is the generic designation for a *function.* The expression $f(x)$ will sometimes be referred to as the *value* of the function f at the input value x. So, for example, if $f(x) = 3x + 2$, the value of f at 6 is equal to $f(6)$ or, using the formula, 20.

Here is an analogy that might be helpful. Let's let f stand for the function which represents the instructions for filling out a 2007 tax return to get the amount of tax owed for the 2007 tax year, and let x represent any given tax payer. Then $f(x)$ represents the amount of tax owed by taxpayer x for 2007. Notice that f by itself does not have any monetary value. You cannot find the tax owed unless you know what particular tax payer you are talking about. The letter f is merely the name for the rule (a very elaborate rule with dozens of pages of instructions) which tells you how to find the value of the tax owed for each taxpayer x. If you are a taxpayer, $f(you)$ is the amount *you* owe for 2007.

Also, we will often write the equation $y = f(x)$. In this case, we are simply using a single letter, namely y, to denote the output values, rather than the expression $f(x)$. This is sort of like calling the President "Mr. Bush" instead of "President George W. Bush."

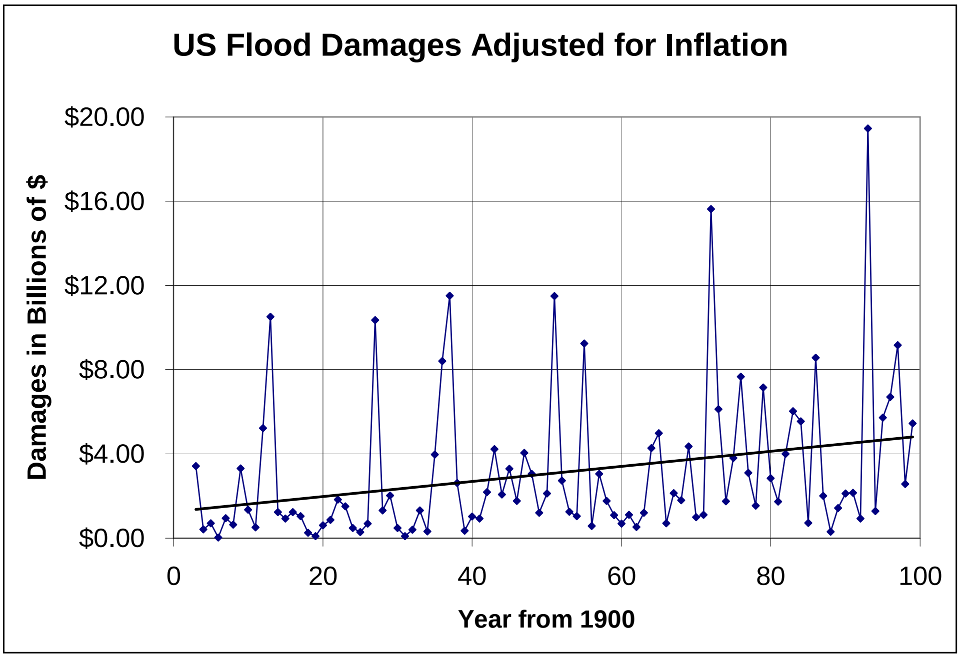


Figure 2: Two functional representations of US Flood Damages.

Modeling Functions

Many of the functions we will consider throughout this text will be examples of modeling functions. A **modeling function** is a function that describes a real world phenomenon by representing the relationship between two real world variables. For example, $y = x^2$ is a function which gives the square of the input. However, if this function is used to find the area of a square, then $y = x^2$ (or $A = s^2$) is now a *modeling function* because it represents the area of a square as a function of the length of the side. We might also specify the units for s, say in feet. The units for A would then be square feet.

Modeling functions can be represented in all four ways – symbolic, graphical, numerical, verbal or by combinations of these. However, if we wish to use a modeling function to predict something which has not yet occurred, to calculate an output value which is difficult to estimate in another way, or to expand the list of input output pairs in a table, a symbolic representation is often the most useful.

For example, global warming is a topic that has been in the news a great deal of late, and is likely to remain so into the foreseeable future. Global warming has many potential effects, including more frequent and more intense hurricanes with resulting damage caused by wind, rain, and flooding. Suppose we wish to use a function to model flood damage in the U.S. The graph in Figure 2 actually shows two functions. The first, represented by the "squiggly line graph," shows actual flood damages in billions of dollars adjusted for inflation. The second "straight line" is a linear approximation of the first.[55]

Both graphs give us a picture of how flood damages have been changing over time. In general, flood damages have been increasing, albeit with a lot of variability from year to year represented by the squiggles.

In order to predict what might happen in the future, a symbolic representation might be useful. In fact, the linear approximation graphed above has the symbolic representation

$$D(t) = 0.036t + 1.25$$

[55]Data from *U.S. National Oceanic and Atomspheric Administration (NOAA)* $http : //www.nws.noaa.gov/oh/hic/flood_stats/Flood_loss_time_series.htm.$

where $D(t)$ is the approximate damages in billions of dollars and t is the time in number of years since 1900.[56] Notice we are using the letter D to denote our Damages function rather than the more generic f. This is not necessary, but can be helpful, especially when one is considering several different modeling functions simultaneously.

Using this symbolic representation, we can estimate flood damages in future years. For example, for the year 2010, t would be equal to 110. The estimated damages would be

$$D(110) = 0.036 \cdot 110 + 1.25 \approx 5.21.$$

Going further into the future, the estimated damages for 2030 would be $D(130) = 0.036 \cdot 130 + 1.25 \approx 5.93$.

Now, we should make a couple of comments about these predictions. The first is that, because our symbolic representation is only an approximation and we know that the actual damages shows a high degree of variability, we should expect that our estimate could be wrong by several billions of dollars. Secondly, the model assumes that no changes in the behavior of the system that might significantly effect the dependent variable occurs in the future. For example, if ocean temperatures begin to increase more rapidly in the future, this may lead to even more frequent and stronger hurricanes, with increased resulting flood damages.

Example 1. For the function D modelling U.S. flood damages, explain what $D(60) = 3.41$ means in the given context.

Solution: Since D is giving an estimate of U.S. flood damages adjusted for inflation where the input is years from 1900, $D(60) = 3.41$ means we estimate that U.S. flood damages were about 3.41 billion dollars in 1960. We will note that actual damages were about 0.69 billion. Again, this is mostly because our symbolic calculation does not give actual flood damages but only estimated damages. ■

Example 2. We noted earlier that the graph of fertility versus contraceptive use in Figure 1 of this chapter probably does not represent a function. However, the line that is shown along with the scatterplot does define a function. One way to see this is that any vertical line will intersect any other line (except another vertical line) in exactly one point. While we do not have a symbolic representation for this function, like the function D given above which models U.S. flood damages, this line was created to be the best *linear* model for this data. Having only the graph of the line, we could still use the line to make "eyeball" estimates of "average" fertility rates as a function of contraceptive use. For example, following the line, a country with a contraceptive use rate of 60% would average about 3 births per woman.

Example 3. We are certainly allowed to consider different functions as models of the same data or phenomenon. Another model for U.S. flood damages is given by

$$D_1(t) = 0.741(1.0168^t)$$

where we have used D_1 to distinguish this model from the previous one. What does this second model predict for flood damages in 2010 and 2030?

Solution: The estimates provided by D_1 are $D_1(110) = 0.741(1.0168^{110}) \approx 4.63$ and $D_1(130) = 0.741(1.0168^{130}) \approx 6.46$. Notice that D_1 gives a lower estimate than D for 2010 but a higher estimate for 2030.

Whether D_1 is a better model than D is a question we will consider further in later chapters. ■

Composition of Functions

Previously, in our "mother versus child" function, we noted how a given value, like Marcella Catalano, can serve as either an input or an output. Similarly, using function notation, a given

[56] This equation was derived using linear regression – something we will learn about later in this book. The given points will not necessarily lie on the line and, therefore, will not necessarily be solutions to the equation.

output $g(x)$ can serve as the input for another function f. For example, if g is the function which assigns to each person his or her mother, and f is the function which assigns to each person his or her father, then for a given person x we can first find the person's mother $g(x)$ and then use this as the input in the function f, finding $f(g(x))$. The expression $f(g(x))$ represents both the father of $g(x)$ and also the maternal grandfather of x.

Example 4. Let g be the "mother function" and f be the "father function" as described above. Explain what $g(f(x))$, $f(f(x))$, and $g(g(x))$ mean.

Solution: The expression $g(f(x)$ signifies that we first find the father $f(x)$ of person x, and then use $f(x)$ as the input in the function $g(x)$ to find the mother of $f(x)$. The whole expression $g(f(x))$ thus represents the mother of x's father, or in other words, x's paternal grandmother.

Similarly, $f(f(x))$ represents x's paternal grandfather and $g(g(x))$ represents x's maternal grandmother. ■

The process of using the output $g(x)$ of the function g as the input for the function f is called the **composition** of f and g. We use the notation $f \circ g$ to denote the composition of f and g. Thus, $(f \circ g)(x)$ or $f(g(x))$ denotes the value of the composition of f and g.

Expressing the composition $f \circ g$ symbolically using the symbolic expressions for f and g is fairly straightforward. For example, if $f(x) = 2x + 5$ and $g(x) = 3x - 1$ then

$$(f \circ g)(x) = f(g(x)) = f(3x - 1) = 2(3x - 1) + 5.$$

If we wish, we can simplify the last expression using the distributive law, so that

$$(f \circ g)(x) = 2(3x - 1) + 5 = 6x - 2 + 5 = 6x + 3.$$

Note that what we are doing is merely substituting the expression $3x-1$ for the x in the expression for $f(x)$. Alternatively, we could first substitute the explicit expression for f, then substitute for g, and still get the same expression.

$$(f \circ g)(x) = f(g(x)) = 2(g(x)) + 5 = 2(3x - 1) + 5$$

Example 5.
Let $f(x) = x^2 + 3$ and $g(x) = x - 2$. Find $(f \circ g)(x)$ and $(g \circ f)(x)$.

Solution: We have

$$(f \circ g)(x) = f(g(x)) = f(x - 2) = (x - 2)^2 + 3 = x^2 - 4x + 4 + 3 = x^2 - 4x + 7$$

and

$$(g \circ f)(x) = g(f(x)) = g(x^2 + 3) = x^2 + 3 - 2 = x^2 + 1.$$

■

Note well from the previous example that $(g \circ f)(x)$ and $(f \circ g)(x)$ give you different formulas, and thus are different functions. This will be the case for the vast majority of pairs of functions f and g.

Domains and Ranges

Sometimes, we may wish to talk about the whole *set* of possible input values, or the whole *set* of possible output values, rather than individual inputs or outputs.

- The set of all possible inputs for a given function f is called the **domain** of f.
- The set of all possible outputs for a given function f is called the **range** of f.

For example, if we are considering the function *"Tell me the weight of the given person,"* then the domain is the set of all people, and the range is the set of all weights of people. If we are measuring weights in pounds, we might say the range is all number bigger than 0 (premature babies can weigh only a few ounces) up to the largest weight ever recorded (say 1200 pounds as a guess, or consult the Guiness Book of World Records for a more precise upper limit).

For a function given in numerical form, the domain would be simply the list of inputs, unless it is understood that the function includes more inputs and outputs than are listed. For example, for the functions given in Table 2 on median incomes in the U.S., we might say the domain is the list of years 1950, 1955, 1960, up to 1995 and 1998. Similarly, the range of a function represented by one of the income columns would simply be the list of 11 incomes. For example, the function "Median Income for Black Women versus Year" has range

$$\{\$2949, \$3658, \$4237, \$5595, \$8142, \$9011, \$9071, \$9509, \$10386, \$11723, \$13137\}.$$

For a function given in symbolic form, we usually take the domain to be the largest set of real numbers which are valid when substituted into the formula for the function. For example, if the function output is given as $f(x) = 1/x$, then the only input that would not work of x would be $x = 0$, since division by 0 is not defined. Thus, we would say the domain is "the set of all real numbers except 0."

For now, the two main considerations when considering domains of symbolic functions are to:

- Exclude all input values which result in division by 0 from the domain.
- Exclude from the domain all input values which result in taking the square root of a negative number.

For example, if $f(x) = \frac{\sqrt{x-2}}{x-5}$, then we would exclude all numbers smaller than 2 from the domain, since these numbers would result in taking the square root of a negative number. We would also exclude 5, since that would result in division by 0. So, the domain would be "all real numbers bigger than or equal to 2 and not equal to 5." In mathematical notation, we might write this as $\{x|x \geq 2, x \neq 5\}$.

If a function given in symbolic form is also a modeling function, we might further restrict the domain and range based on the context. For example, for the function $y = x^2$, we could say the domain is all real numbers, since there is not restriction on what x might be for this equation. The function $A = s^2$, where s is the side length of a square and A is the area, is given by essentially the same equation. However, it does not make sense for s to be 0 or negative in this context, so we would say the domain is all positive real numbers. In general, for modeling functions we simply pick a reasonable set of values for the domain, based on the context, and allow some flexibility on what counts as "reasonable."

Finally, for functions in graphical form, we can often infer the domain and range off of the graph. For example, for the line shown in Figure 5, we might say the domain is "all percentage values from 0% to about 90%," based on where the line starts and ends. We could say the range is approximately "all real numbers from about 1 up to about 6.2."

Functions and solving equations

Previously, we used the function $D(t) = 0.036t + 1.25$ to model U.S. flood damages as a function of the year since 1900, denoted by t. We estimated the flood damages in 2010 by evaluating $D(t)$. What if we want to know when flood damages will reach \$7 billion or some other amount?

Here, we are asking what input t would produce an output $D(t)$ equal to 7. Since $D(t) = 0.036t + 1.25$, this results in the equation

$$7 = 0.036t + 1.25.$$

You very likely have solved equations like this before, even if they were not in the context of working with a function. The basic rules are:

- Equals can be added to both sides of any equation.
- Equals can be subtracted from both sides of any equation.
- Equals, *except for zero*, can be multiplied on both sides of any equation.
- Equals, *except for zero*, can be divided by on both sides of any equation.

Following any of these rules changes a given equation into a different equation, but the second equation will still have the same solutions(s) as the first.

The "except for zero" is important in the last two items. For example, the equation $x - 2 = 5$ has one solution for x, namely 7. However, if we multiply both sides by zero, we have

$$0(x - 2) = 0 \cdot 7$$

$$0 = 0$$

This last equation is trivially true, and does not even have the variable x appearing. Thus, it is true no matter what the value of x is. So, the second equation does not have the same solutions as the first. This is a result of multiplying by zero, and thus, is why multiplying both sides of an equation by zero is not a legitimate step in solving equations.

Now, to solve $7 = 0.036t + 1.25$, we first subtract 1.25 from each side of the equation.

$$7 - 1.25 = 0.036t + 1.25 - 1.25$$

$$5.75 = 0.036t.$$

Then we divide both sides by 0.036 (which is not zero).

$$\frac{5.75}{0.036} = \frac{0.036t}{0.036}$$

$$159.7 \approx t$$

If we round the value $t = 159.7$ to the nearest year, we get 160, which corresponds to the year 2160. So, our equation would predict U.S. flood losses should reach \$7 billion (on the average) around 2160.

Example 6. For the function $f(x) = 2x + 5$, find the input x when the output $f(x)$ is 3.

Solution: We have

$$\begin{aligned} 3 &= 2x + 5 \\ -2 &= 2x \\ -1 &= x \end{aligned}$$

■

Example 7. Consider the function $C(r) = 2\pi r$, where $C(r)$ represents the circumference of a circle as a function of the radius r.

- Find the circumference of a circle of radius 2.5.
- Find the radius of a circle which has a circumference of 220.

Solution: If the radius is 2.5, then the circumference is equal to $C(2.5) = 2\pi 2.5 = 5\pi \approx 15.71$.

Now, if the circumference is 220, this means $C(r) = 220$. So, we have

$$\begin{aligned} 220 &= 2\pi r \\ 220/(2/pi) &= r \\ r = 110/pi &\approx 35 \end{aligned}$$

So, if the circumference is 220, the radius is approximately 35.
■

Characteristics of Graphs of Functions

In this section, we introduce some terminology we can use to describe the shape of the graph of a function. These terms generally apply to graphs which are in the form of a line graph, versus a scatterplot.

Increasing and Decreasing

A function is said to be **increasing** over an interval if, for each input, a, in the interval, every input to the right of a has a greater output.[57] Symbolically, we say a function is **increasing on an interval** if for any two numbers a and b in the interval with $b > a$, then $f(b) > f(a)$. In words, this says the larger the input, the larger the output. Increasing graphs are going "uphill" as x increases toward the right. This does not have to occur over the entire graph since increasing is a property on an interval. In our world oil production example (Figure 2), the graph is increasing on the interval from 1930 to about 2010, except for a short interval in the early 1970's, and a somewhat longer interval from 1979 to about 1982.

A graph is **decreasing** over an interval if for each input, a, in that interval, every input to the right of a has a smaller output. Symbolically, we say a function is **decreasing on an interval** if for any two numbers a and b on the interval with $b > a$, then $f(b) < f(a)$. In words, this says the larger the input, the smaller the output. Decreasing graphs are going "downhill" as x increases toward the right. The graph in Figure 2 is decreasing from 2010 on. To determine if a graph is increasing or decreasing, we always compare outputs as we move from *left to right.*

Steepness is another word we can use to describe a graph or a section of a graph. The more quickly a graph increases or decreases, the steeper it is. We can think of steepness as indicating the **rate of change** of the graph, and will say more about this later.

One-to-one Functions

Previously, we discussed the vertical line test for functions. Any graph that is intersected at most once by any vertical line is a function. We can also consider a horizontal line test for graphs that we already know are functions. If the graph of a function is intersected at most once by any horizontal line, then we say the function is **one-to-one**. Verbally, this means that each output is matched with only one input. This is the converse of the definition of a function which states that each input is matched with exactly one output.

Symbolically, a function f is one-to-one if the only way that $f(a) = f(b)$ is if $a = b$. So, for example, the function $g(x) = x^2$ is not one-to-one since $g(3) = g(-3)$, but $3 \neq -3$. Functions that are one-to-one include those that would be increasing through their entire domain, or those that would be decreasing through their entire domain. Both of these types of functions would satisfy the horizontal line test.

Concavity

In Chapter Two, we discussed line graphs as one type of graph. For line graphs that are the graphs of functions, we can categorize the graph or sections of the graph according to their **concavity**. Informally, a graph is **concave up** if it bends upward, resembling a bowl that can

[57] The interval must be contained in the domain of the function.

hold water. A graph is **concave down** if it bends downward, resembling an umbrella when it sheds water. The concavity of a graph and whether it is increasing or decreasing over an interval are independent. An increasing function can either be concave up or concave down. Similarly, a decreasing function can be either concave up or concave down. (See Figure 3.)

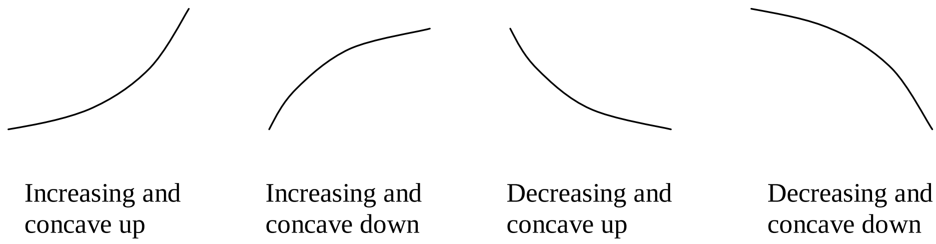


Figure 3: Examples of concavity

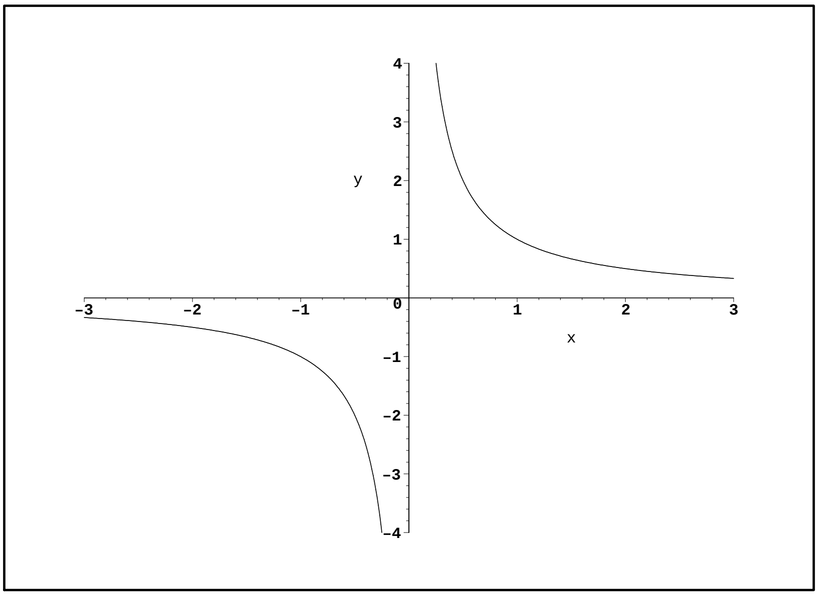


Figure 4: The graph of $y = 1/x$ is always decreasing, is concave down when $x < 0$ and concave up when $x > 0$.

One needs to be careful in distinguishing between the concepts of increasing/decreasing, concave up and concave down, and slope and steepness. Increasing and decreasing describe how the *output* of the function is changing over an interval while the concavity of a function tells how the *rate* of increase or decrease is changing on the interval. This is connected to the definition of concavity. A graph is **concave up** on intervals where the rate of change or **slope** of the graph increases. A graph is **concave down** on intervals where the rate of change decreases. Note that when the graph is increasing, we say the slope is positive, and when it is decreasing the slope is negative.

Also, we need to be careful that slope is not the same as steepness. If a graph is increasing and concave up, the slope of the graph is increasing and we also would say it is getting steeper. On the other hand, if a graph is decreasing and concave up, then its slopes are negative but are still increasing. The slope becomes "less negative." However, the graph is also getting less steep or "flatter." One needs to carefully consider this difference between steepness and slope and how it is affected by whether the graph is increasing or decreasing.

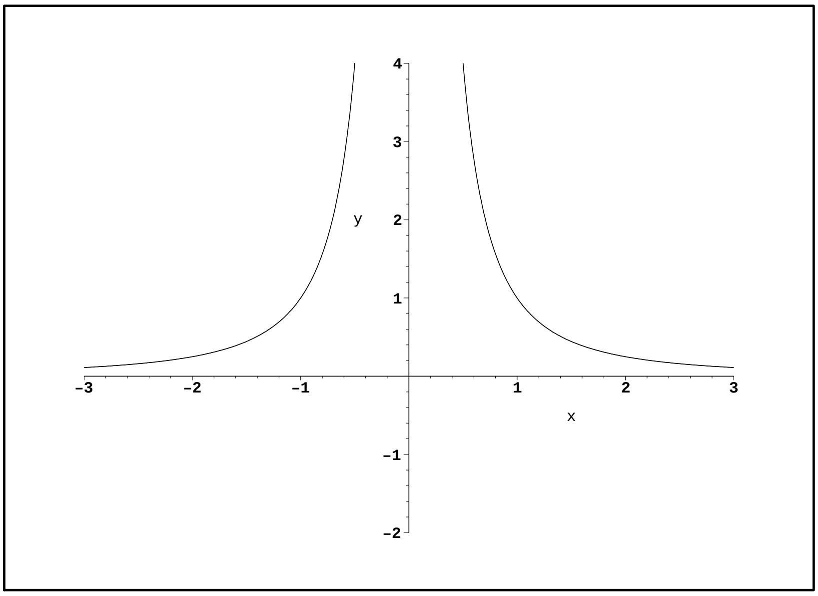
For example, the graph in Figure 4 is always decreasing. Where it is concave down (when

$x < 0$), its rate of change is decreasing, and in this case it is becoming steeper. The slope is becoming more and more negative.

Where it is concave up (when $x > 0$), the slope is still increasing (becoming less negative), but it is becoming less steep. In other words, its rate of change is increasing, even though its steepness is decreasing.

Example 8. The graph of $y = 1/x^2$ is shown in Figure 5. Describe where this function is decreasing, increasing, concave up, and concave down.

Figure 5



Solution: The function is increasing when $x < 0$ and decreasing when $x > 0$. It is always concave up and never concave down, so the rate of change of the function is always increasing. ■

Functions Related to Hunger

Every gun that is made, every warship launched, every rocket fired, signifies in the final sense a theft from those who hunger and are not fed, those who are cold and are not clothed.
—Dwight D. Eisenhower

How can one "measure" hunger?

Hunger is not a numerical quantity. However, we could define and try to measure it in a number of ways. For example, we might define hunger to mean not getting the minimum number of calories needed to function normally on a day to day basis. As long as we can measure a person's calorie intake we can measure hunger, under this definition. We might consider a person hungry if they often have to skip meals, or go days at a time without eating. We can measure this by counting meals skipped, assuming 3 meals a day as normal. Alternatively, we might be more concerned with the long term effects on people who are chronically hungry, including mortality that can be attributed to hunger.

These measures all can be applied to individual people, but we may also want to develop measures which apply to countries, social groups, or other populations of people. For example, UNICEF has three different standard ways of measuring malnutrition within countries. These are the percentage of children underweight for their age, the percentage of children suffering from stunting, and the percentage of children suffering from wasting. Stunting is a measure of height deficiency, and can result from chronic hunger over months or years during a child's growth period. Wasting results when a child is underweight for their height, in other words, chronically thin. The UNICEF data set including these malnutrition measures can be found in the appendices, and we will use these data in a number of examples, exercises, and activities throughout this book.

Of course, we want to do more than just measure hunger. Ideally, we want to identify, if possible, what causes hunger, so that we can address the problem constructively. This might include identifying **variables** which we can numerically measure, and seeing how closely these are **correlated** with our measures of hunger.[58] It would be unrealistic to expect that we could find variables X for which "hunger is a function of X." However, we should certainly be able to create functions, in symbolic or other representations, which serve as modeling functions for the relationship between hunger and other variables.

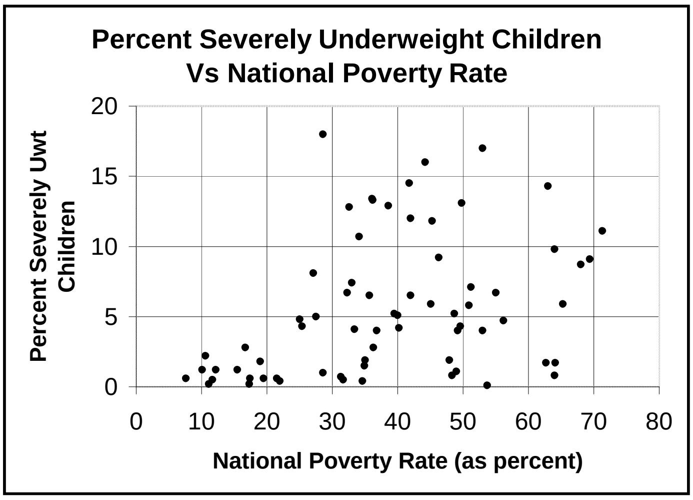


Figure 6: Percentage of Severely Underweight Children versus National Poverty Rate

For example, Figure 6 shows a scatterplot of malnutrition versus poverty, where malnutrition is measured using the underweight measurement. Poverty is measured for each country by its national poverty rate, which is the percentage of the population considered below the nation-adjusted poverty line. What does this graph seem to say about the relationship between poverty and hunger?

In the activities for this chapter, you may be asked to identify other variables which might be related to hunger.

[58]Correlation will be defined in chapter 6. For now, we will note that two variables are said to be correlated if a numerical change in one tends to produce a change in the second.

Reading Questions for Functions

1. Name the four different ways to represent a function, **and give examples of each**. [Note: The examples should be different from those given in the reading.]

2. For each of the following, determine if the rule represents a function. **Write a sentence justifying your answer.**

(a) Tell me the date of birth of a given person.

(b) Tell me the total amount owed in student loans of a given graduating senior.

(c) Tell me the number of calories you consumed on a given day.

(d) Tell me a color on the flag of a given country.

(e) Total yearly income versus number of years of education attained by given person.

(f) Number of malnourished children in a country versus number of personal vehicles in the country.

(g) Tell me the grade of a given person in College Algebra this semester.

(h) Tell me the grade of a given person in College Algebra where the input is how many hours they study per week for the course, on average.

3. For each of the following, determine if the rule represents a function. **Write a sentence justifying your answer.**

(a)

Input	-4	-1	0	2	5
Output	3	4	6	6	9

(b) $y = x + 3$, where x represents the input.

(c) $y^2 = x^2$, where x represents the input.

(d) $y^2 = x + 2$ where x represents the input.

(e) $y^2 = x + 2$ where y represents the input.

(f)

Input	3	-1	0	3	5
Output	7	2	5	5	3

4. For each of the following modeling functions, give reasonable values for the domain and also for the range.

(a) $w(h) = 1.4h - 20$ where h is the height in inches of a child under 3 years of age, and $w(h)$ is the estimated weight in pounds of the child.

(b) $C = 2.959g$ where C is total cost of gasoline and g is the number of gallons purchased filling up a single gas tank.

(c) The cost of tuition at your college in a given year.

(d) The number of children a particular woman has as a function of her age (as their age varies over their lifetime).

(e) The total yearly income of a particular person as a function of the age of that person.

5. In the previous problem, what does the 2.959 in part (b) represent?

6. What is typically the best way to represent a modeling function if you want to use it to predict something that has not yet occurred?

7. Use the symbolic representation $D(t) = 0.036t + 1.25$ of the function from this chapter that modeled U.S. flood damages to estimate flood damages for the year 2000. Remember that t is the number of years since 1900. Use the symbolic representation of the function $D(t)$ to estimate flood damages in 1850. Does your answer seem reasonable? Why or why not?

8. The following table gives the estimated percentage of all Americans who had ever used marijuana in the given year.

Year	1979	1982	1985	1988	1991	1993	1994	1996	2001
Marijuana Use	28.0%	28.60%	29.4%	30.6%	30.5%	31.0%	31.1%	32.0%	36.9%

(a) What row is associated with the range?
(b) How would you describe the range as a set?

9. When finding the domain for a function given symbolically, what two things should you check?

10. (a) What are the domain and the range of the function $f(x) = \pi x^2$?
(b) What are the domain and the range of the function $A = \pi r^2$ used for finding the area A of a circle, given the radius r?

11. Give an appropriate domain for each of the following functions. You may use symbols, words, or both when giving your answer.

(a) $f(x) = 5x^2 - 2$
(b) $f(x) = \sqrt{x+6}$
(c) $f(x) = \sqrt{4x-3}$
(d) $f(x) = \frac{1}{x-8}$
(e)

Input	0	2	5	8
Output	-1	1	2	1

(f) Your summer job pays an hourly wage of \$7.25 an hour. Given the number of hours you worked in one week, determine your gross pay for that week.

12. The cost C in cents of first class postage is a function of the weight w (in ounces) of the item mailed. We write $C = f(w)$. Explain the meaning of $f(2) = 0.60$.

13. Suppose that, in a particular state, the amount of the fine you pay for speeding F is a function of the number of miles per hour over the speed limit you were clocked at, denoted m. We write $F = f(m)$ (mathematics is generally case sensitive). Explain the meaning of $f(12) = \$75$.

14. Let $f(x) = 3x^2 + 4$. Find each of the following:

(a) $f(0)$
(b) $f(2)$
(c) $f(a)$
(d) $f(x+1)$

15. Let $f(x) = \frac{\sqrt{x-3}}{2x+1}$. Find (where possible) each of the following:

(a) $f(0)$
(b) $f(4)$
(c) $f(7)$
(d) $f(x+1)$, and simplify

16. Let $f(x) = 3x + 4$. Solve each of the following equations:

(a) $f(x) = 7$
(b) $f(x) = 0$
(c) $f(x) = 11$
(d) $f(x + 1) = 11$
(e) $f(x) = x + 10$

17. Let $f(x) = 3x^2 + 4$. Solve each of the following equations. You may need to use the 'guess and check' method.

(a) $f(x) = 7$
(b) $f(x) = 0$
(c) $f(x) = 16$
(d) $f(x) = 22$

18. Let $f(x) = 3x + 4$ and $g(x) = 5x - 2$. Find $(f \circ g)(x)$ and $(g \circ f)(x)$ and simplify.

19. Let $f(x) = 2x + 3$ and $g(x) = \frac{1}{2}x - 6$. Find $(f \circ g)(x)$ and $(g \circ f)(x)$ and simplify.

20. Let $f(x) = x^2 + 4$ and $g(x) = 2x - 5$. Find $(f \circ g)(x)$ and $(g \circ f)(x)$ and simplify.

21. Let $f(x) = 3x^2 + 4$ and $g(x) = 2x - 5$. Find $(f \circ g)(x)$ and $(g \circ f)(x)$ and simplify.

22. Suppose the function $m(n) = 2\sqrt{\frac{.25}{n}}$ gives the upper limit on the margin of error for a randomly conducted poll of n people

(a) Find $m(400)$ and $m(1600)$.
(b) Carefully explain what $m(1100) = .03$ in terms of the number of people polled and the margin of error.
(c) What sample size would be required to achieve a margin of error of .02?

23. A function approximating the data on marijuana use given previously in the Reading Questions is $y = 0.0033x + .2414$, where x is the year from 1970 and y is the estimated percentage of all Americans who have used marijuana, as a decimal.

(a) What is the estimated percentage use in 1970?
(b) According to the equation, in what year did marijuana use reach 26%?
(c) According to the equation, in what year will marijuana use reach 40%?
(d) Would this equation eventually predict 100% marijuana use? What does this imply about the applicability of this equation?

24. Graph the function $f(x) = x^2 - 2x + 1$ on a computer or graphing calculator.

(a) Where is the function increasing?
(b) Where is the function decreasing?
(c) Where is the function concave up?
(d) Is this function one-to-one or not?

Functions: Activities and Class Exercises

1. **Midges.**[59] In this activity, you will be introduced to the Fathom Dynamic Data software package, and gain some experience working with functions in numerical and graphical form.

   (a) A good way to look at all of the data at once is to use a *case table*. To open a case table in Fathom, first select the collection by clicking on it. Then, you can either drag the **Table** icon into the main window, or choose **New—Case Table** from the **Object** menu. Do this.

   i. How many Af and how many Apf midges are represented in our data set? (You can record your answer in the same text box. Resize or move the box as needed).

   ii. To find the longest and shortest wing span for this set of midges, you can select the table and then the **WL** column, and then **Table—Sort Ascending**. Record the longest and shortest wing spans in your text box. Note that Apf midges tend to have longer wings than the Af midges.

   iii. Do the same for Antenna length. Which variety of midges seems to have longer antennae?

   iv. You can also use functions or formulas in symbolic form in Fathom. Let's add another attribute and create it as a formula. To do this, click on the **new** at the top of the rightmost column of the table and then type in "Difference" which will be our name for the new attribute. Then, click on **Table—Show Formulas**. A new blank row should appear above the first case. Now, click on the box in this row just under your Difference heading. You should get a "formula dialog box." Let's enter in the formula WL - AL. You can do this by typing or by selecting the WL and AL from the attribute menu in the formula box. Click OK when done. The differences between WL and AL are calculated and entered into the table for each midge.

   v. sort your table using the Difference column. What do you notice? What does the difference tell you about the variety of midge?

   (b) Create a graph of AL versus WL.

   (c) Looking at either the table or the graph, decide whether or not AL is a function of WL, and explain why, using the definition of function.

   (d) (Optional for those using tablet PC's) You can include Fathom Documents in other windows documents, including Microsoft Journal notes. To do this, open the journal writer from the START menu. Then, select **File—Import**. Then, select the Fathom file. You should get a copy of your Fathom file to appear on the journal page. However, it may be that some of the Fathom screen got 'cut off'. If so, try again after deleting the note (use **File** menu) and moving the objects in the Fathom window to the left, so they will all fit in the journal window.

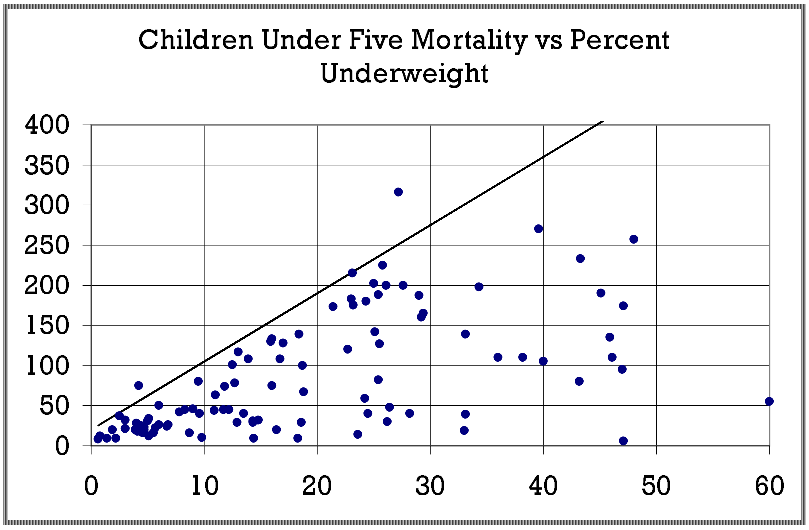
2. **Hunger.** The graph below shows a scatterplot of Child Mortality versus Percent Underweight Children. Child mortality is measured in number deaths of children under five per 1000 live births. There are 80 data points, representing 80 different countries. The data are from the year 2000 and was compiled by UNICEF.[60]

   (a) Why is the scatterplot probably not the graph of a function?

   (b) Explain why the line graph shown with the scatterplot is a function.

[59] This activity uses data provided by Key Curriculum Press with the Fathom software

[60] see /www.childinfo.org



(c) The line shown is meant to represent the upper limit on child mortality given the percent of underweight children. Note that only two countries have points over this line. The equation for this function is $f(x) = 8.5x + 20$, where x represents percent underweight children and $f(x)$ represents the approximate upper limit on the child mortality rate for a country with the given underweight percentage x.

i. Find $f(12)$, and explain briefly what this value means.

ii. Find $f(18)$, and $f(24)$.

iii. Suppose a country with a child mortality rate of 120 deaths per 1000 live births wants to reduce its child mortality rate to no greater than 80 deaths per 1000 live births. Use the symbolic representation for $f(x)$ to estimate what the percent of underweight children should be for the country to achieve this goal.

(d) So far, we have been considering the variable "Childhood Mortality" as it is related to hunger, where the latter is measured using "Percent of Underweight Children." What other variables do you think might be related to hunger? Identify as many possibilities as you can. Variables might include those which you believe are causes of hunger, but you could also include variables which you believe might be correlated with either hunger a "lack of hunger." Some of your variables might be numerical, but do not necessarily restrict yourself to variables which are easily measurable. If you possibly can, however, identify how the variable might be measured.

3. **Education Spending.** The following table gives the average amount of money spent per student in the United States for a number of years, going back to 1960, adjusted for inflation. The year represents the year in which that school year was completed (so 1960 represents the 1959-60 school year).

Year	1960	1970	1980	1981	1986	1990	1991
Spending	\$2,161	\$3,657	\$4,954	\$4,890	\$5,843	\$6,639	\$6,647
Year	1994	1995	1996	1997	1998	1999	2000
Spending	\$6,678	\$6,741	\$6,735	\$6,811	\$6,987	\$7,216	\$7,392

(a) Does this table represent a function? Why or why not?

(b) One can quickly see from the table that spending, even after adjusting for inflation, has increased over the years. We might wonder during which period of years education spending was increasing the fastest. To do this, we might consider the percentage increase in spending from one year to the next. One problem with our data would be that we do not have spending values for every year.

i. Describe how would you calculate the percentage increase from one year to the next.
ii. Estimate the percentage increase from 1990 to 1991.
iii. How could you calculate yearly percentage increases for periods where you do not have data for every year?
iv. Use your idea from the previous question to estimate the yearly percentage increase from 1960 to 1961. Using your method, would you get the same answer for all the yearly increases for the 1960's?

4. **How much does your car contribute to global warming?** Based on a sample of 38 car models, the function $y = G(m) = -.267m + 13.48$ approximates the tons of carbon dioxide emissions $G(m)$ produced by a car in a year as a function of the EPA's city mileage rating m for that car in miles per gallon.

(a) Find the value for $G(20)$ and describe what it means in terms of gas mileage and carbon dioxide emissions.
(b) Find the estimated carbon dioxide emissions for a car model that gets 30 miles per gallon.
(c) Find the estimated gas mileage for a car model which produces 7 tons of carbon dioxide annually.
(d) According to the given function, what is the estimated gas mileage for a car which produces no carbon dioxide emissions?
(e) Suppose that $m(w)$ denotes a function which models the city mileage $m(w)$ as a function of the weight of the vehicle w in pounds.

i. Suppose $m(1300) = 27$. Explain what this means in terms of weight and mileage.
ii. Suppose $m(w) = \frac{9000}{.82^w}$. Find an expression for the composition $(G \circ m)(w)$.
iii. Identify the input and output variables of the function $G \circ m$ and explain what this function means in terms of these variable.

5. **World Population Growth Rates.** World overpopulation is frequently cited as a contributing factor to world hunger. The world population has grown from about 4,434,700,000 in 1980 to about 6,336,200,000 in the latter part of 2003. However, the annual growth rate of the population has been decreasing for a number of years, and in 2003 was about 1.16% per year. This means that in 2004, there would be about 1.16% more people than there were in 2003. The graph below shows the percent growth rate for world population, going back to 1950 and projected out to 2050.

(a) Given the estimated population in 2003 noted above, and the estimated 1.16% growth rate, about what would the population be in 2004?
(b) We could define a function verbally by letting the input be the year, starting with 2003, and the output be estimated world population assuming that the 1.16% holds forever and the 2003 population is 6,336,200,000. You have already calculated the value for this function when the input is 2004. Begin to develop a numerical representation for this function by filling in the second column of the table below, giving the values for the function in the given years.

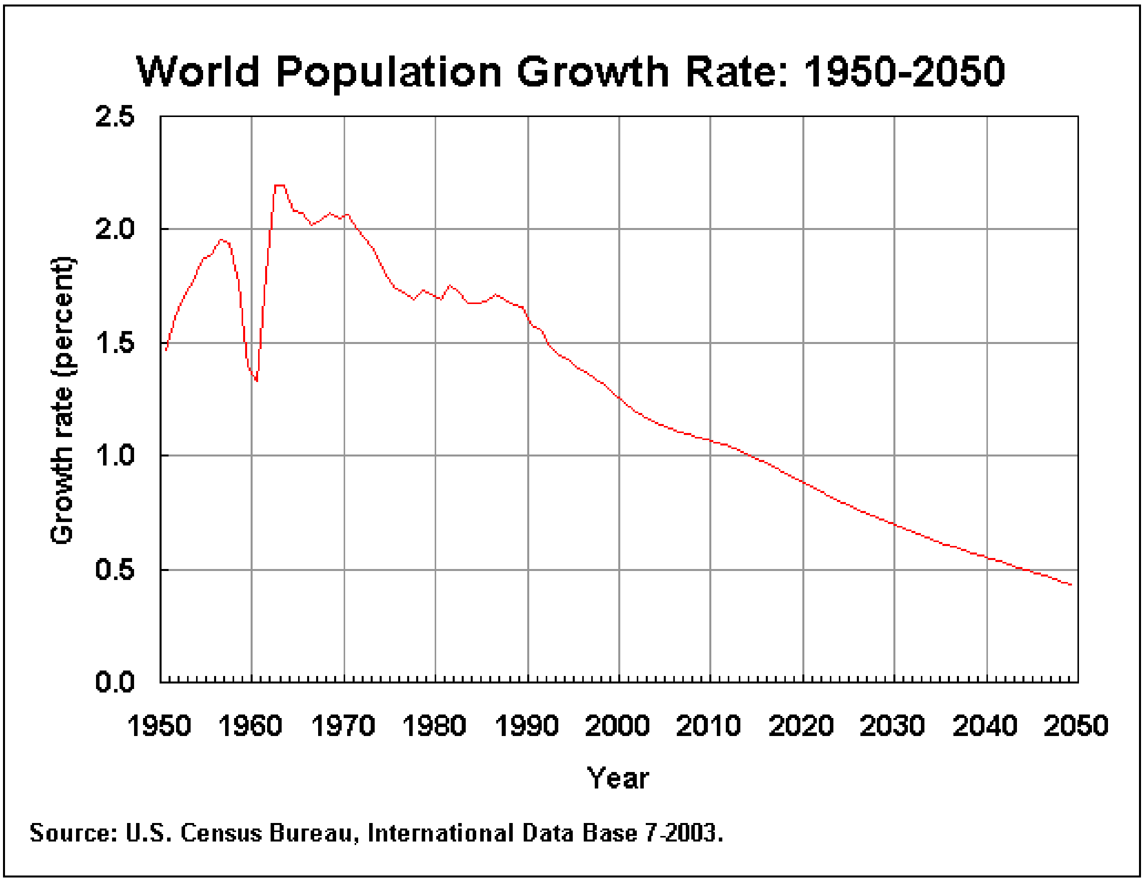


Figure 7: World population growth rates

Year	Est. Pop. at 1.16% annual growth	Est. growth rates	Est. Pop. at Est. growth rates
2003	6,336,200,000	1.16	6,336,200,000
2004		1.14	
2005		1.12	
2006		1.11	
2007		1.10	
2008		1.08	
2009		1.07	
2010			

(c) Since the world population growth rates are decreasing, the estimated population function from part (b) probably overestimates what the actual population will be (although we won't know this for awhile!). To get a better estimate, we could use the function where the output is calculated using the estimated growth rates, as they change over time. Fill in the last column of the table above using the percentages in the third column. Use the growth rate in one year to calculate the estimated population in the next year.

(d) How much of a difference is there between the values in 2010 of these two functions?

(e) We might wonder whether this difference might be significant enough to affect decisions that governments might make concerning the availability of food, energy, and other resources, etc. Make a rough estimate of how much "the world" would save in food costs if the population in 2010 is the lower of the two estimates instead of the higher. Make the most reasonable assumptions you can in developing your estimate, and explicitly say what these assumptions are and how you arrived at them.

6. **Composition of Functions.** In this activity, you will gain practice in working with composition of functions.

   (a) Let $f(x) = 2x + 1$.

      i. Find $(f \circ f)(x)$, $(f \circ f \circ f)(x)$, and $(f \circ f \circ f \circ f)(x)$. Do you notice any pattern?

      ii. In order not to have to write lots of $f's$ and $\circ$'s, let's define $f^{(n)}(x)$ to denote the composition of f with itself n times. So, for example, $f^{(3)}(x) = (f \circ f \circ f)(x)$. Continue with the work you started in part (a) i., computing formulas for $f^{(n)}(x)$ for larger values of n until you see a general pattern. Describe this pattern as best you can in words or symbols.

      iii. If you did not already do so in part ii., write a general formula for $f^{(n)}(x)$. This formula can include the variable x and the letters f and n. For example, your answer might be something like $f^{(n)}(x) = (n - 1)x + 2n$.

   (b) Repeat part (a) using the function $g(x) = 3x + 2$.

7. **Weight versus Height.**For the first two years of a child's life, the function $y = w(h) = 1.4h - 20$ can be used to approximate the average weight y, in pounds, of a child whose height h (or length) is given in inches.

   (a) Find the value for $w(25)$ and describe what it means in terms of height and weight of a child.

   (b) Find the estimated weight of a child who is 30 inches in length.

   (c) Find the estimated *height* h of a child who weighs 30 pounds.

   (d) Suppose we wanted to create a new function which calculated the weight in kilograms instead of pounds as the output, and still let the input height be measured in inches.

      i. There are approximately 0.45 kilograms in 1 pound. What is the average weight, in kilograms, of a child that is 20 inches tall?

      ii. The new function could be described as $y = 0.45w(h)$. Find an equation for this new function, which gives the output y as a function of h.

      iii. What is the average height, in inches, of a child who weighs 12 kilograms?

8. **Do people who read live longer?** Based on data collected for countries in North and South America, the function $y = S(r) = -.142r + 98.3$ approximates the percent of females who survive to age 10 in a given country $S(r)$ as a function of the percent of women in the given country who do not know how to read r.

   (a) Find the value for $S(10)$ and describe what it means in terms of illiteracy and survival.

   (b) Find the estimated survival to age 10 for a country where the illiteracy rate is 2%.

   (c) Find the estimated illiteracy rate for women in a country where 90% of females survive to age 10.

   (d) What is the estimated survival rate for a country where none of the women know how to read?

5. Linear Functions

One should not increase, beyond what is necessary, the number of entities required to explain anything.
— Willam of Occam (Ockham)

A problem should be stated in its basic and simplest terms. In science, the simplest theory that fits the facts of a problem is the one that should be selected.
— Dr. James Schombert

We have looked at various ways to present data as well as various ways to represent functions. We will now look at different types of functions, what we will call "families" of functions. All the functions in a given family will share some general characteristics that define that family, much like all the animals in a given species share the characteristics of that species.

We will begin with the family of linear functions, the simplest and most common type of function. Common examples of linear functions include the conversion of feet to meters, dollars to euros, and temperature from the Fahrenheit scale to the Celsius scale.

Linear functions are important partly because they are the simplest types of functions to work with, and partly because linear functions supply good models for many, many different types of real-world situations. When looking for a mathematical model, look linear first. In this chapter, we will explore some of the important characteristics of linear functions and how to determine if a given function is linear, as well as some examples of linear models.

Definitions

What is a linear function? What are the characteristics that define the family of linear functions? The easiest answer is that linear functions have graphical representations which are straight lines. However, we want to be able to tell if a particular function is linear whether it is represented graphically, symbolically, numerically, or even verbally.

Numerical and Graphical Representations of Linear Functions

Consider the the graph in Figure 1 of a linear function.

Two points on the graph are labelled. We can see that for each one unit increase in x, the y-value increase by 3. If the graph is linear, this property must hold for every 1 unit increase in x, otherwise the angle or steepness of the line would change. This leads us to the numerical representation given in Table 1.

Table 1: Tabular representation of a linear function.

Input	0	1	2	3	4
Output	2	5	8	11	14

The table reflects that for each 1 unit increase in x, the y-value increase by 3. What if the x-value increases by 2 units? In this case, the y-value increases by 6 units. Doubling the change in x also doubled the change in y.

In general, if we pick any two points either on the graph or in the table, the change in outputs divided by the change in inputs will always be the same, namely 3. This happens regardless of whether the input is changing by 1 unit, by 4 units, or by any other number of units. For this function, the change in output is always three times the change in input. We say that 3 is the **rate of change** of y versus x. One way to define linear functions is that they are continuous functions

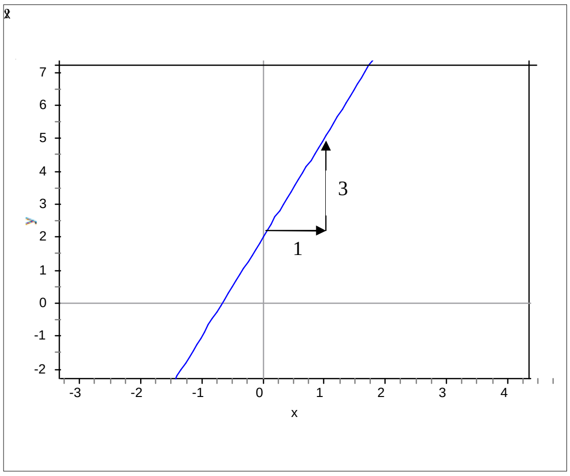


Figure 1: Graph of a linear function that has a starting point of $(0, 2)$ and whose change in output is 3 times the change in input.

with a constant rate of change. Numerically, a **linear function** is a function where the ratio, $\frac{\text{change in output}}{\text{change in input}}$, is constant over the entire domain of the function.

The symbol Δ (the upper-case Greek letter delta) is commonly used to denote change in a variable. We can therefore refer to the change in x as Δx (or "delta x")and the change in y as Δy (or "delta y"). Then,

$$\frac{\Delta y}{\Delta x} = \frac{\text{change in output}}{\text{change in input}}$$

denotes the rate of change for any linear function.

This constant rate of change can be used to create a graphical representation of a linear function. One way to specify a linear function (or, equivalently, to draw a straight line), is to pick a starting point (somewhere to put your pencil) and a rate of change (how much the output changes when the input changes by 1 unit).

For example, suppose you are asked to draw a line starting on the y-axis at the point $(0, 1)$ with $\frac{\Delta y}{\Delta x} = 2$, or in other words, the change in y will be 2 times the change in x. You just need to plot the starting point and a second point and connect them to draw your line. You are given the starting point, $(0, 1)$, so you need to locate a second point. Since the change in output is 2 times the change in input, you can begin at $(0, 1)$, move to the right one unit and then move up 2 times one unit or two units to locate the second point. This produces $(0+1, 1+2) = (1, 3)$ as your second point. Connecting these two points with a line gives you the graph shown in Figure 2.

Note that these two pieces of information, where to start and the constant rate of change, are sufficient for drawing the graph of our linear function. This is what makes linear functions simple. You only need two pieces of information to completely define the function.

We define these two pieces of information in general as follows. The ***y*-intercept** is the point where the line crosses the y-axis (or, equivalently, the output when the input is zero). This can be thought of as our starting point. The **slope** is the constant rate of change or

$$\text{slope} = \frac{\text{change in output}}{\text{change in input}} = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

where (x_1, y_1) and (x_2, y_2) are any two points on the line. For the linear function given in Table 1, the y-intercept is 2 (since this is the input when $x = 0$) and the slope is 3. For the linear function given in Figure 2, the y-intercept is 1 and the slope is 2.

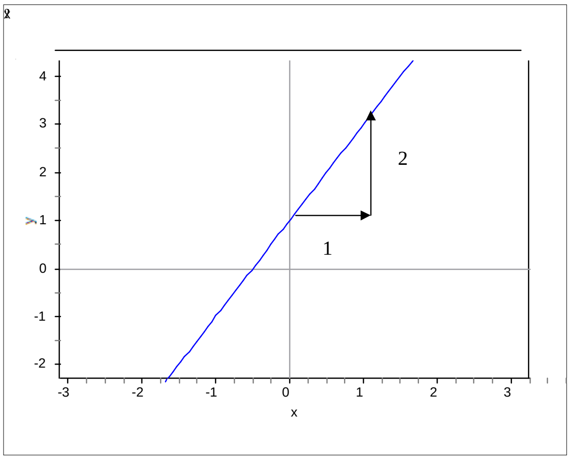


Figure 2: Graph of a linear function that has a starting point of $(0, 1)$ and whose change in output is 2 times the change in input.

Linear functions can have a negative slope. For example, a slope of negative 3 means that, for every *increase* of one unit in the input, the output *decreases* by 3 units. To draw a graph with slope -3 and y-intercept 2, we start at $(0, 2)$, and for each one unit we go to the right (increase the input by one), we go *down* three units (decrease the output by three). So the second point is $(0+1, 2-3) = (1, -1)$. Connecting these two points creates the graph of this linear function. (See Figure 3.)

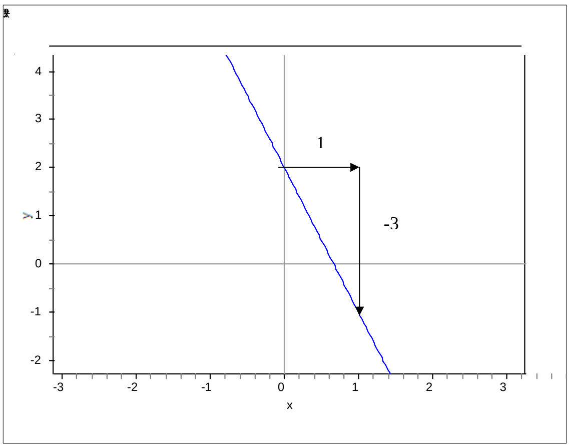


Figure 3: Graph of a linear function with a y-intercept of 2 and a slope of -3.

Example 1. Consider the functions represented in Table 2(a) and Table 2(b). One is a linear function and one is not. Can you tell which is which?

Solution: To determine if a function given numerically is linear, we need to determine if there is a constant slope or rate of change, $\Delta y/\Delta x$. The function given by Table 2(a) has a constant rate of change (or slope) of 2 over the entire domain, while the function given by Table 2(b) does not. In Table 2(a), as x increases by 1, y always increases by 2, so the rate of change is 2. Note that Table 2(b) does follow a discernible pattern. This pattern is not *linear*, however. ■

Table 2: Deciding if a table of numbers represents a linear function.

x	-2		-1		0		1		2
y	-5		-3		-1		1		3
$\Delta y/\Delta x$		2/1		2/1		2/1		2/1	

(a)

x	-2		-1		0		1		2
y	8		2		0		2		8
$\Delta y/\Delta x$		$-6/1$		$-2/1$		2/1		6/1	

(b)

Symbolic Representations of Linear Functions

In this section, we will see that linear functions have symbolic representations of the form $y = mx + b$ or equivalently $f(x) = mx + b$ where m will equal the slope and $(0, b)$ will be the y-intercept. Let's see why this is.

Consider the linear function from Table 1 above, given again here.

Table 3: Table 1 reproduced.

Input	0	1	2	3	4
Output	2	5	8	11	14

The slope is 3 and the y-intercept is 2. Is $f(x) = 3x+2$ a symbolic representation of this function?

By definition, the y-intercept of the function f is $f(0)$, which in this case equals $3 \cdot 0 + 2 = 2$. This is the same value we see in the table. By hand, we can also check that $f(1) = 3 \cdot 1 + 2 = 5$, $f(2) = 3 \cdot 2 + 2 = 8$, $f(3) = 3 \cdot 3 + 2 = 11$, and $f(4) = 3 \cdot 4 + 2 = 14$. So, this function f does correspond to the table, at least for the values we have.

We noted previously that a linear function in numerical form will always give the same change in y for each 1 unit change in x. We can add additional entries to the table using this property, and then check that the symbolic form gives the same y values. Instead, let's check that the symbolic form always produces a $\frac{\Delta y}{\Delta x}$ equal to 3.

Let's designate an arbitrary input x_0. The output under the function f would be $f(x_0) = 3x_0+2$. If we increase the input by 1, we get x_0+1 and a corresponding output of $f(x_0+1) = 3(x_0+1)+2$. Calculating the change in y divided by the change in x, we have

$$\frac{\Delta y}{\Delta x} = \frac{f(x_0+1) - f(x_0)}{x_0 + 1 - x_0} = \frac{3(x_0+1) + 2 - (3x_0+2)}{1} = 3x_0 + 3 + 2 - 3x_0 - 2 = 3$$

So the formula $f(x) = 3x + 2$ always results in an increase in y of 3 whenever x is increased by 1 unit. Since it has the same slope of 3 and y-intercept of 2 as the function given by Figure 1 and Table 1, it must be the symbolic representation of that function.

We can use this same logic to show that the general form $f(x) = mx + b$ will give a linear function with slope m and y-intercept b. If $x = 0$, we have $f(0) = m \cdot 0 + b = b$. So, the y-intercept of $f(x) = mx + b$ is b.

For the slope, we have

$$\frac{\Delta y}{\Delta x} = \frac{f(x_0+1) - f(x_0)}{x_0 + 1 - x_0} = \frac{m(x_0+1) + b - (mx_0+b)}{1} = mx_0 + m + b - mx_0 - b = m.$$

This shows the slope of the function f is constant and equal to m.

The conclusion is that the function $f(x) = mx + b$ gives us the two pieces of information we need, the slope and the y-intercept, for the linear function with slope m and y-intercept b.

Finding the equation of a line through any two points

Because the function $f(x) = mx + b$ is linear with slope m and y-intercept b, we can find the linear function in symbolic form given any two points. If we let (x_1, y_1) and (x_2, y_2) denote any two points, we can express the slope as

$$\frac{\Delta y}{\Delta x} = \frac{\text{change in } y}{\text{change in } x} = \frac{y_2 - y_1}{x_2 - x_1}$$

Using this fact, we can use the following procedure to find the equation.

- Find the slope using $\frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$
- Substitute the slope and the x and y values from either known point into the equation $y = mx + b$
- Solve for b
- Write the equation using the calculated m and b.

Example 2.
Find the equation of the line through the points $(2, 9)$ and $(6, 1)$.

Solution:
Following our procedure, we first find the slope. We have

$$\frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{1 - 9}{6 - 2} = -2$$

so the slope is -2.

Using the point $(6, 1)$ and the slope of -2, we substitute into $y = mx + b$ to get

$$1 = -2 \cdot 6 + b$$

$$1 = -12 + b$$

$$1 + 12 = 13 = b.$$

So, $b = 13$. This gives us the equation $y = -2x + 13$, or in function notation, $f(x) = -2x + 13$. We can double-check that this is correct by checking that the two points we started with are valid input-output pairs. We have $f(6) = -2 \cdot 6 + 13 = -12 + 13 = 1$. Also, $f(2) = -2 \cdot 2 + 13 = 9$. So, this function does give us a linear function through both given points. ■

Verbal Representations of Linear Functions

Linear functions provide good models in a wide variety of contexts. Principally, any two variables that show an approximately *constant* rate of change can be described by a linear function. In some cases, the rate of change might be precisely or nearly precisely constant.

For example, if you travel at a *constant* velocity of 75 mph, then the distance, d (in miles), that you travel is represented by the linear function

$$d = 75t,$$

where t is time (in hours). When we have driven for 0 hours ($t = 0$), we have gone 0 miles ($d = 0$). So the y-intercept is zero. The slope is 75 since the constant rate of change (i.e. velocity) is 75 mph. For every 1 hour change in the input (for every hour you drive), your output (distance) increases by 75 miles.

In a situation where two variables do show a constant rate (or approximately constant) of change, we can find the equation by determining two pieces of information, the slope and the y-intercept. In a physical (or verbal) situation, the y-intercept will be your starting point (i.e. your output when your input is zero) and the slope will be your constant rate of change.

Example 3. Suppose you call a professional to take care of a furnace problem. He charges \$20 per visit plus \$30 per hour. Let C be the function whose input is n, the number of hours worked, and whose output is $C(n)$, the *total* cost.

1. Explain why this is a linear function.
2. Find an equation for C.

Solution:

1. This is a linear function since the rate of change (in this case, the charge per hour) is constant.
2. To find the equation, we need to find the slope and the y-intercept. The slope, or rate of change, is the \$30 charge per hour. The y-intercept is the cost for the furnace person to show up or the starting cost which is \$20. So the equation for this function is

$$C(n) = 30n + 20$$

where $C(n)$ is the total cost (in dollars) and n is the number of hours the plumber works.

■

Note in the previous example that the units for the slope is dollars per hour, and that the units for the output is dollars, and the units for the input is hours. This general pattern works for all linear modeling functions. We always have:

The units for the slope m are unit of output per unit of input.

Going back to our distance function $d = 75t$, the units for the output d were miles and the units for the input t were hours. The units for the slope is miles per hour.

Example 4. Suppose you are currently spending about \$80 per month on electricity, and that over the past 12 months, your cost has been increasing about \$1.20 per month. What linear function would model your electricity costs?

Solution: If we let x represent months with $x = 0$ the current month, then \$80 represents your y-intercept or starting point. The \$1.20 represents your constant rate of change. Assuming this rate remains constant at about this level into the future, your electric costs y as a function of x will be $y = 1.20x + 80$. So for example, 20 months from now your electric bill would be approximately $1.20 \cdot 20 + 80 = \$104$. ■

Linear functions are often the basis for converting units of measurement. This is because the rate of change for units of measurement is nearly always *constant.* For example, there are always 12 inches per foot, 5280 feet per mile, or 4 quarts in a gallon. So, to convert from feet to inches, we multiply the number of feet by 12 to get the number of inches. Symbolically, if i is the number of inches and f is the number of feet, we have $i = 12f$. This is a linear equation with slope 12 and y-intercept 0. If we put in the units, we have

$$i \text{ inches} = \frac{12 \text{ inches}}{1 \text{ foot}} \times f \text{ feet}.$$

We can find similar conversion equations for feet to miles (slope is $\frac{1}{5280}\frac{\text{miles}}{\text{foot}}$), or gallons to quarts (slope is $\frac{4}{1}\frac{\text{quarts}}{\text{gallon}}$).

How long is a billion seconds in years? (Go ahead and take a guess.)

Let's look at linear functions used for converting one unit of time into another. To start with, let's convert seconds to years.

There are 60 seconds in a minute, 60 minutes in an hour, 24 hours in a day, and 365 days in a year. So,

$$\text{number of seconds} \times \frac{1 \text{ minute}}{60 \text{ seconds}} \times \frac{1 \text{ hour}}{60 \text{ mins}} \times \frac{1 \text{ day}}{24 \text{ hours}} \times \frac{1 \text{ year}}{365 \text{ days}} = \text{number of years}$$

or (multiplying all the fractions together)

$$\frac{1}{31,536,000}s = y$$

where s represents number of seconds and y number of years. Notice, again, the cancellation of units to produce the correct unit for the output.[61]

If $s = 1,000,000,000 = 10^9$, then

$$y = \frac{1}{31,536,000}10^9 \approx 31.71 \text{ years}$$

or 31 years and about 8 and 1/2 months.[62] Also, we should mention that Appendix 4 contains a list of some other "large numbers" with comparisons for you to get a sense of "how large is large?"

Also notice that the y-intercept is again zero. This is because zero in the first unit of measurement corresponds to zero in the second unit of measurement. Zero seconds means zero years. Zero feet means zero inches. Will the y-intercept for a linear conversion function always be zero?

The answer is "no." One example is the conversion from temperature in degrees Celsius to temperature in degrees Fahrenheit. The Celsius scale was created so that 0 degrees Celsius is the freezing point of water and 100 degrees Celsius is the boiling point of water. So, 0° C equals 32° F, and 100° C equals 212° F. This means that the linear function that converts from temperature in the Celsius scale to the Fahrenheit scale must contain the points $(0, 32)$ and $(100, 212)$. The y-intercept will be 32. We find that the slope is $\frac{212-32}{100-0} = 1.8$. So the conversion equation is the linear equation $F = 1.8C + 32$. The slope of 1.8 means that a change of 1 degree Celsius is equivalent to a change of 1.8 degrees Fahrenheit.

Solving Literal Equations

In Chapter Four, we discussed how to solve equations like $2x + 5 = 21$. This equation had one variable, and we were able to find the single value for x which solved the equation.

Many times, we will consider equations that have more than one variable or "letter," like $y = mx + b$. When an equation has more than one variable or undetermined parameter, we call it a **literal equation**. Literal equations are solved using the same methods as single-variable equations. One difference is that the solution will not be a number but an expression. Another difference is that one must specify which of the several variables one is going to solve for.

[61] Using cancellation of units is a good way of deciding if you want $\frac{1 \text{ minutes}}{60 \text{ seconds}}$ or $\frac{60 \text{ seconds}}{1 \text{ minute}}$.

[62] If you are 20 now and want to be a billionaire by the time you are 50, you need to earn more than $1 per second, on average, for 30 years.

Example 5. Solve the following literal equations.

1. Solve $y = mx + b$ for the variable m.

2. Solve $B = P(1 + r)^t$ for P.

Solution:

1. To solve for m we can add, subtract, multiply or divide by equals on both sides, as long as we don't multiply or divide by 0. It often helps to think of using the reverse of the usual order of operations. To start, we will subtract b from both sides.

$$y = mx + b$$

$$y - b = mx + b - b = mx$$

Then, we divide by x.

$$\frac{y-b}{x} = \frac{mx}{x} = m$$

our solution is that $m = (y - b)/x$.

2. The equation $B = P(1 + r)^t$ looks a little more complicated than the first, but is actually easier to solve. Simply divide by $(1 + r)^t$.

$$B = P(1 + r)^t$$

$$\frac{B}{(1+r)^t} = \frac{P(1+r)^t}{(1+r)^t} = P$$

So, $P = \frac{B}{(1+r)^t}$.

■

Variations in symbolic form

Sometimes a function that actually is linear may be "in disguise", that is, written symbolically in a format other than $y = mx + b$. Given any literal equation with variables x and y, if we can rewrite it in the form $y = mx + b$, then we know we have a linear function.

One way to do this is to algebraically solve for y (or whatever letter represents the output) before deciding if your function is linear. Consider the equation

$$\frac{y-5}{x} = 3.$$

This does not "look" like the equation of a line, but if we solve for y (by multiplying both sides by x and then adding 5), we get

$$y = 3x + 5$$

which is the equation for a linear function with slope 3 and y-intercept 5.[63]

[63]The original equation $\frac{y-5}{x} = 3$, excludes the input $x = 0$ because of division by zero. So, technically, this is a linear function which is not defined for $x = 0$.

What about the equation

$$\frac{y-5}{x} = 3 + x?$$

To solve for y we multiply both sides by x and then add 5. This gives us the equation

$$y = 3x + x^2 + 5.$$

This is not a linear function since the equation has an x^2 term. To be safe, always solve the equation for y (or whatever letter represents the output) and see if it looks like $y = mx + b$ before deciding if the function is linear.

Regression and Correlation

When we looked at scatterplots in Chapter Two, we introduced the idea of strong and weak associations. In Chapter Four, we saw a couple of examples of scatterplots where we included what we called "the linear function that best fits the data." In this section, we will make both of these concepts more precise by quantifying what we mean by "best fit" and "strength of association."

The scatterplot in Figure 4 shows how U.S. energy consumption has been increasing since 1982. Although the data is not precisely linear, it is approximately linear and so we can *fit a line to this data* by drawing a line that is close to all the points. We can then use the equation of this line as a symbolic description of the data. The line plotted in the graph is called the **linear regression** line (also know as the **least-squares** line) for this data. The equation is $E(t) = 0.19t + 10.6$, where t represents years from 1980 and $E(t)$ represents total energy consumption in quadrillions of BTU's.[64] Let's look at where this equation came from.

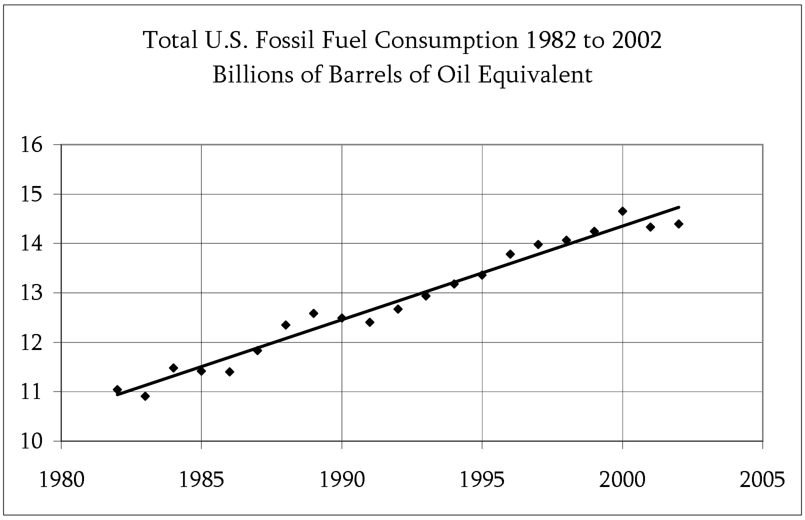


Figure 4: U.S. Energy Consumption

The Method of Least Squares

The most common method of determining the equation of the line that best fits a set of data is called the **method of least-squares**. The resulting line is referred to as the least-squares regression line or simply the (linear) regression line. Formally, a **least-squares regression line** is

[64] BTU is short for British Thermal Unit. Eight gallons of gasoline contains the equivalent of about 1 million BTU's of energy.

a line that best fits two-variable data by minimizing the sum of the squares of the vertical distances between the points in the scatterplot and the line.

To illustrate this, let's consider the data shown in Table 4, which gives heights and arm spans of a small sample of college students.

Table 4: Heights and arm spans (in centimeters) of sample college students.

Height	152	160	165	168	173	173	180	183
Actual Arm Span	159	160	163	164	170	176	175	188

A graph of the data from Table 4 is shown in Figure 5.[65]

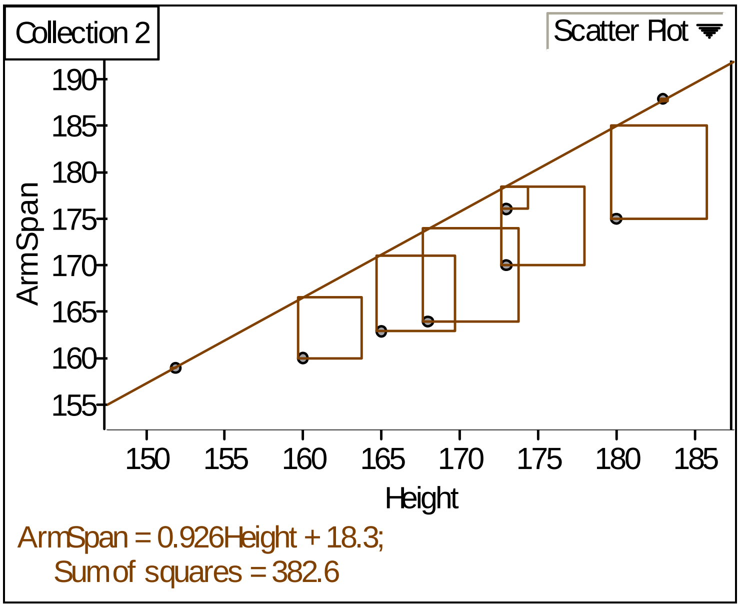


Figure 5: Scatterplot and sample line for student arm span versus height data

We have also fitted a line to this data, but this is *not* the best fit line. Notice, for example, that while the line goes through two of the points, all the rest of the points are below the line. Finally, we have also included the "squared errors" in the graph. For each point, a vertical line from that point is drawn to the graph; the length of this vertical line is called the **error** for this data point. Then, a square is created using this line as one of the sides. The closer a point is to the line, the smaller the error is, and so the smaller the square. Adding up the area of all these squares we get what is called the **sum of squared errors**, or just sum of squares for short. For our line, the sum of squared errors is equal 382.6. This number is a measure of the overall error between the line and the data points.

The regression line is the line which makes this sum of squares which measures the overall error as small as it can possibly be. Now, of all the infinitely many lines that we could draw, how can we find this best fit line?

It is possible to find this line "by hand," using ideas from calculus. However, we will use technology which is programmed to do linear regression. For example, you can find linear regression

[65]This graph was created using a statistical software package called Fathom

lines using most graphing calculators, the Excel spreadsheet program, or most any statistical software package. We found that the least square regression line for the arm span data is

$$\text{arm span} = 0.875 \times \text{height} + 21.4.$$

Figure 6 shows the data again, but now with the linear regression line, instead of the line we used above. Notice that the sum of squares is now only 125.3.

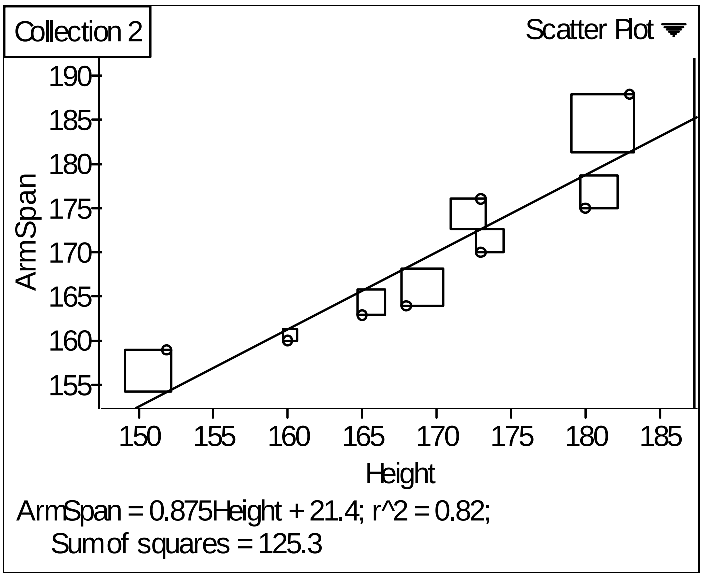


Figure 6: Scatterplot with Regression Line

Table 5 is the same as Table 4, except that it gives the values predicted by the linear regression line, and the errors and squared errors of our line, i.e. the difference between the actual output (using the data points) and the predicted output (using the equation of the line). The sum of the squares of all these errors is 125.42, essentially the same as the 125.3 given in the graph (the discrepancy is due to round off error).

Table 5: Heights and arm spans with regression predictions and errors.

Height	152	160	165	168	173	173	180	183
Actual Arm Span	159	160	163	164	170	176	175	188
Predicted Arm Span using the line	154.4	161.4	165.8	168.4	172.8	172.8	178.9	181.5
Error	4.6	-1.4	-2.8	-4.4	-2.8	3.2	-3.9	6.5
Squares	21.16	1.96	7.70	19.36	7.70	10.40	15.21	41.93

One more note on the regression line. Notice that the slope of the line is 0.875. Recalling that the slope is the rate of change between the two variables, this means that the linear regression line predicts that each one centimeter increase in height results in an increase in arm span of .875 centimeters. Of course, since the line does not follow the data exactly, this is only an approximate rate of change, but we can consider it to be the best estimate of the rate of change between these two variables.

Correlation

The two examples of regression lines that we have looked at so far do a reasonably good job of predicting values for inputs close to our data points. Both lines visually fit the data reasonably well, and give predictions reasonably close to the actual data values. However, data cannot always be expected to behave so nicely, and regression lines may or may not fit the data well. For example, in Chapter Two, we considered a scatterplot showing malnutrition versus poverty, and we have reproduced it here in Figure 7 along with the linear regression line.

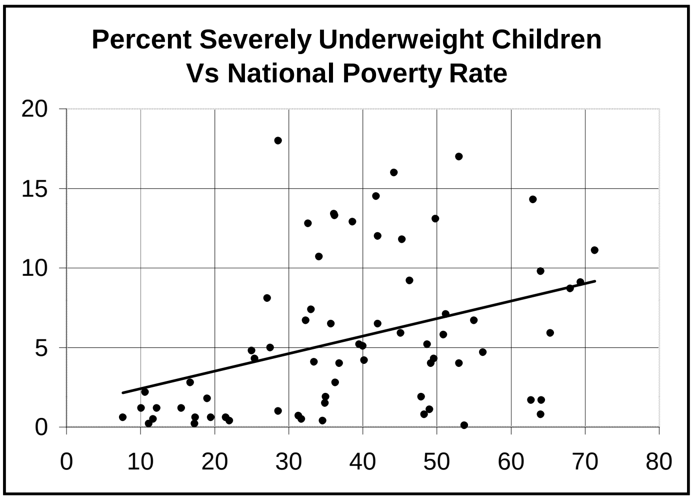


Figure 7: Percent of Severely Underweight Children Versus National Poverty Rate.

As noted in Chapter Four, there does seem to be a general increasing trend, but the association is certainly not as strong as in our previous two examples. We can see that the regression equation will not give very accurate predictions since many of the data points are not very close to the line.

The standard measure of how well a regression line fits a set of data is called the **correlation coefficient**, or correlation for short. This coefficient measures both the strength and direction of a linear relationship between two numerical variables. The height and arm span data in Figure 6 is said to be highly correlated while the malnutrition and poverty data in Figure 7 has a lower correlation.

Correlation is usually denoted by the letter r. Correlation is always a number between -1 and 1. If the correlation is close to -1 or 1, then the data points are highly correlated, i.e. the plotted data points lie very close to a line indicating a very strong association. In fact, if $r = 1$ or $r = -1$, the data all lie exactly on a straight line and the regression line perfectly predicts the outputs. If the correlation is close to 0, then the data points have little correlation, i.e. the plotted data points do not follow a linear pattern and there is no association or a very weak association. If r is sufficiently larger than 0, then there is a positive association between the two variables and the regression line will have a positive slope. If the correlation is sufficiently negative, then there is a negative association between the two variables and the regression line will have a negative slope.

The correlation of the height and arm span data shown in Table 5 and Figure 6 is $r = 0.904$ while the correlation of malnutrition and poverty data is $r = .369$.[66] This means that the height and arm span data is highly correlated and has a positive association, i.e. people who are taller have longer arm spans. The malnutrition data has a much lower correlation, but there is still a positive association.

You can often compare correlations simply by looking at scatterplots of the data. Figure 8 shows a number of different scatterplots with corresponding correlations.

[66]You might notice in Figure 6, it is noted that $r^2 = .82$. Sometimes r^2 is used instead of r because it can be interpreted to mean "the percentage of the variation in the dependent variable which is explained by the independent variable."

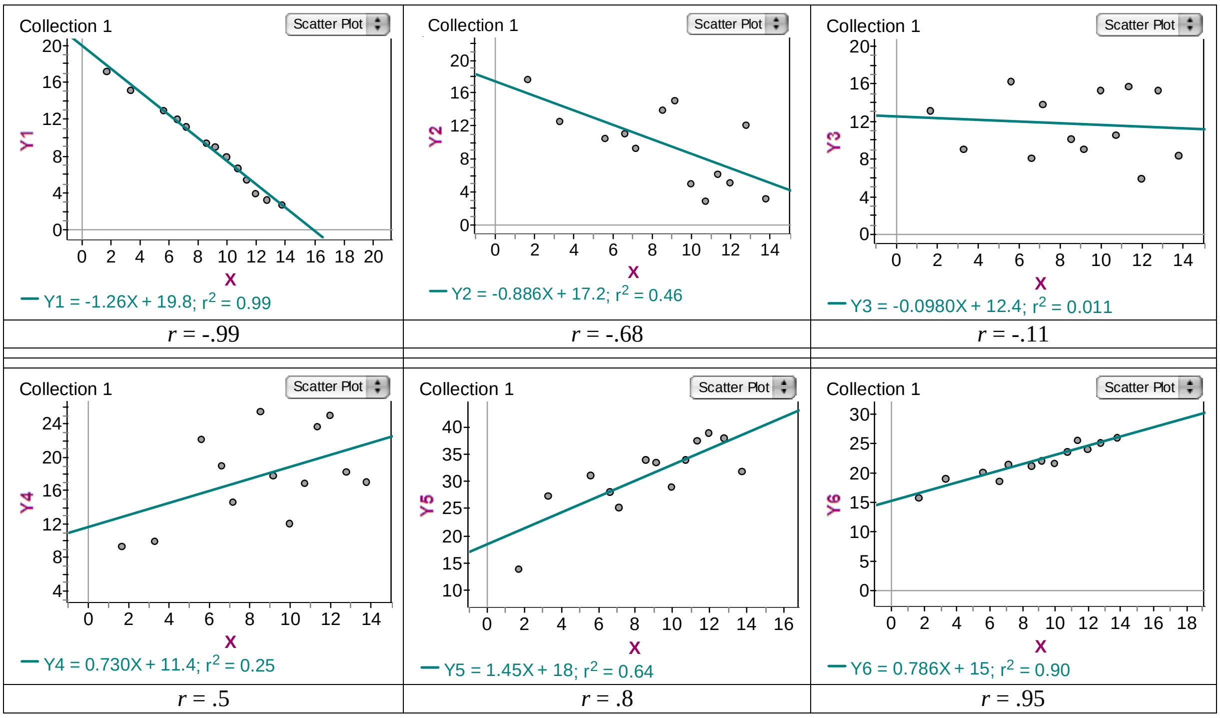


Figure 8: Examples of various scatterplots and their corresponding correlation.

As with the least-squares regression equation, there are rather complicated formulas to determine the correlation.[67] However, as with regression lines, we will rely on graphing calculators or software packages such as Excel to compute these numbers.

Correlation is often confused with **causation**. An example of two variables that have a causal relationship would be Amount Earned versus Hours Worked. An increase in the number of Hours Worked will in general cause an increase in the Amount Earned. If you measured these variables in a particular month for a number of people, the results would vary, but you would have a definite strong positive correlation, and this correlation is the result of the causal relationship between the variables.

However, two variables can be highly correlated without there being any causal relationship. For example, if you were studying the verbal ability of elementary school students, you would probably find a high positive correlation between a child's height and his or her vocabulary, measured in words. This does not mean that the growth of a child causes the child to learn more words. The correlation is a result of the child growing and learning more words as they age. Correlation is merely a numerical measure of the relationship between the data.

To establish causation, one would need to provide proof, or at least a plausible explanation, why change in one variable *causes* a corresponding change in the other variable. For example, an increase in vocabulary probably has a causal relationship with standardized verbal test scores for children.

Correlation and changing the data set

In Chapter Two, we also commented that one might divide the points in our malnutrition versus poverty scatterplot into two groups, one consisting of countries with a poverty less than 25% and the second containing all the other countries. We noted that, looking at only the second group, there seems to be almost no pattern to the data. We can quantify this using correlation. If we look

[67]The formula for correlation involves the sum of squared errors, as well as the means and standard deviations of the two variables.

at only the data points in the second group, the correlation is $r = 0.045$, instead of the $r = 0.369$ we obtained using all the data points, showing there is almost no association within the second group of higher poverty countries.[68]

By excluding some data points, we have drastically altered the correlation and so possibly the conclusions we might reach based on the data. This is an important point to keep in mind. When considering correlation, it is a good idea to consider whether the selection of which data points to include has been biased, either intentionally or unintentionally.

One further point with respect to correlation is that it is dependent not only on the strength of the association between the two variables, but also on the *number* of data points. To give an extreme example, if you do linear regression on a data set with only two points, you will always[69] get a correlation of 1 or -1. This does not mean that there is really any relationship between the two variables. It just means that, given any two points, there is a line that will go exactly through those points. If we had picked any two points out of our malnutrition and poverty data, we would have gotten a correlation of either 1 or -1.

The tendency will be that the more data points you have, the lower the r-value will be. This means that the larger the data set, the smaller the r-value needs to be in order to indicate a strong correlation. Table 6 shows a few of what are called **critical** values of r based on the number of data points n.[70] For a data set with n points, an r-value above the critical value for that n indicates that you can be confident there is a definite positive association. For now, it is enough to be aware that a large data set could have a stronger association than a smaller data set with a higher correlation. For example, Table 6 indicates that a data set with 10 data points and $r = 0.632$ would have the same strength of association as a data set with 47 points and $r = 0.288$.

n	r	n	r	n	r
3	.997	13	.553	27	.381
5	.878	15	.514	37	.325
10	.632	20	.444	47	.288

Table 6: Critical Values for r

As a visual comparison with the previous graphs, Figure 9 shows graphs with 47 points, instead of the 13 points in the graphs in Figure 8.

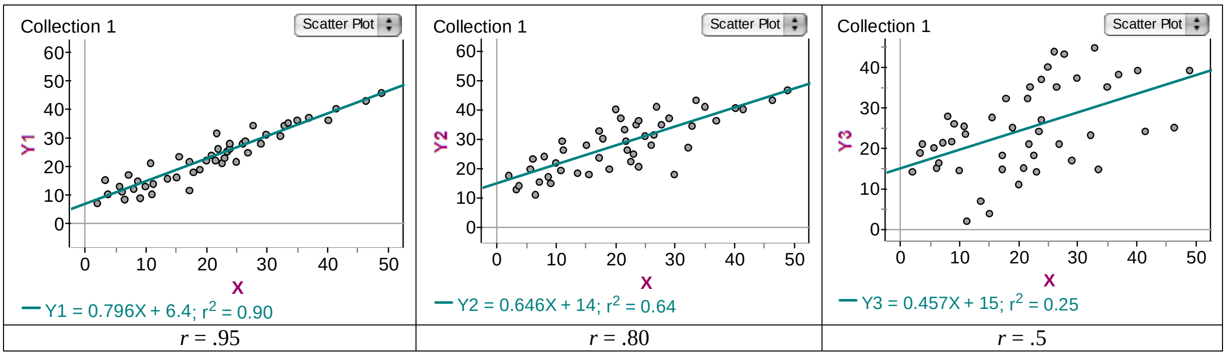


Figure 9: Examples of various scatterplots with their corresponding correlation coefficients.

[68] Again, we are just using Excel to compute this.
[69] unless the two points have the same x-coordinates or the same y-coordinates
[70] Values are from a larger table found in Gordon, et. al.

Application: Carbon Emissions from Fossil Energy Consumption

Table 7 gives carbon dioxide emissions from energy consumption for the commercial sector in the United States.[71] There are two variables, one for non-electric energy usage, and one for total energy usage from all sources. Non-electric includes direct use of petroleum products, coal, and other fuels besides electricity.

Table 7: Carbon emissions from the U.S. commercial sector from 1990 to 2003.

Year	1990	1995	1996	1997	1998	1999	2000	2001	2002	2003
Non-electric	224	228	237	237	220	222	234	228	226	231
Total	777	837	869	912	930	944	1,004	1,021	1,020	1,026

A scatterplot Total Emissions versus Year is shown in Figure 10. We measured time as years since 1990. So 1990 is represented by 0, 1995 is represented by 5, etc. These points seem to portray a linear relationship. The least squares regression line for this data is $y = 21.9x + 757$, with correlation $r = 0.975$.

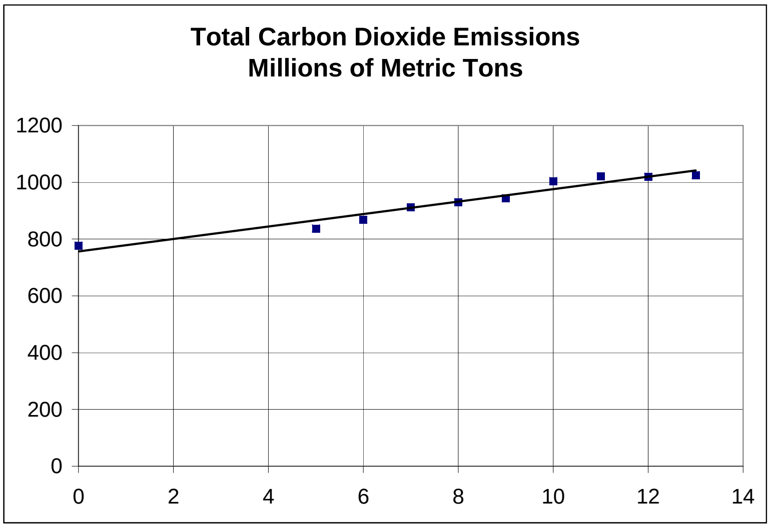


Figure 10: Total Carbon Dioxide Emissions for the U.S. Commercial Sector

Example 6. What is the physical interpretation of the slope and y-intercept of the least squares regression line?

Solution: The y-intercept is 757, which means that at time 0 (the year 1990), approximately 257 million metric tons of carbon dioxide were emitted from energy consumption by the U.S. commercial sector. This is a little below but reasonably close to the actual number million metric tons emitted in 1990 given in Table 7. The line has a slope of 21.9 which means the equation predicts the amount of emissions increases an additional 21.9 million metric tons each year. ■

Looking at this trend, it is easy to see why environmental groups are becoming increasingly concerned with the need for the government to set higher standards for the commercial sector. If this pattern continues, carbon emissions just from the commercial sector will be 1414 million metric tons in the year 2020, almost double what they were in 1990.

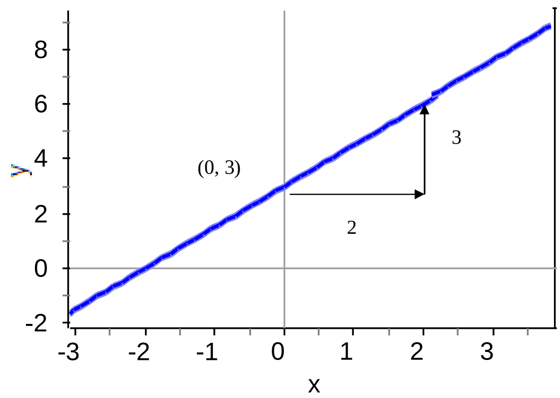
[71] *Data from the Energy Information Administration*

Reading Questions for Linear Functions

1. What is the definition of a linear function?

2. What 2 pieces of information determine a line?

3. The graph below gives the y-intercept, rise, and run for a line. Find the equation of this line.



4. Suppose $y = f(x)$ is a linear function whose slope is 2. If $f(1) = 5$, what is $f(2)$? $f(3)$? $f(10)$?

5. Let $y = f(x)$ be a linear function whose y-intercept is -1 and whose output changes by 3 when the input changes by 1. What is $f(2)$?

6. Suppose $f(x) = 5x + 4$.

(a) How do you know f is a linear function?

(b) What is the slope?

(c) What is the y-intercept?

(d) What is $f(0)$?

(e) What is $f(1)$?

7. Sketch each of the following. Label at least 3 points on your graph.

(a) A linear function with y-intercept 0 and slope $\frac{1}{3}$.

(b) A linear function with y-intercept 1 and slope -2.

8. Let A,B, and C be three points such that the slope between A and B equals the slope between B and C. What, if anything, do you know about the slope between A and C? Explain.

9. Find the equation of the line through the points $(5, 3)$ and $(13, 9)$.

10. Determine if each of the following tables represent a linear function. If so, find an equation for the line. If not, **explain** why.

(a)

Input	-2	-1	0	1	2
Output	3	0	-3	-6	-9

(b)

Input	-2	-1	1	2
Output	-4	0	8	12

(c)

Input	-2	-1	0	1	2
Output	12	10	8	10	12

11. Find the equation of the line for each of the following.

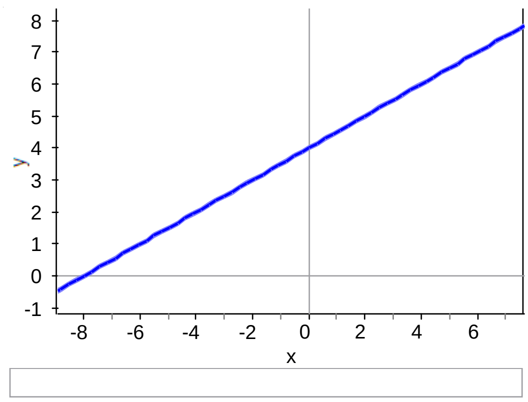
(a) The y-intercept is 5 and the slope is $-\frac{1}{4}$.

(b) It contains the two points $(0, 10)$ and $(1, 11)$.

(c)

Input	-2	-1	0	1	2
Output	0	2	4	6	8

(d)



12. Determine whether or not each of the following represents a linear function.

(a) $\dfrac{y-1}{x+2} = 4$

(b) $y - 4 = 1.2(x + 3)$

(c) $\dfrac{x+3}{y} = 16$

13. A calling card offers a price of \$0.23 cents per minute at any time during the day with a monthly service charge of \$1.00.

(a) Write a linear equation (including the \$1.00 monthly service charge) where the input is minutes and the output is total monthly cost.

(b) Describe the practical interpretation of the y-intercept.

(c) Using your equation from part(a), calculate the total cost for a month if you talk, on average, for 20 minutes a day. Assume the month is thirty days long.

14. Give the linear equation that will convert from the first unit to the second.

(a) centimeters to inches (2.54 centimeters $\approx$ 1 inch)

(b) yard to inches (1 yards = 36 inches)

(c) degrees Fahrenheit to degrees Celsius. [Hint: Use the points (32, 0) and (212,100).]

15. We are interested in converting from miles per hour to feet per second.

(a) Convert 70 mph to feet per second.

(b) Write a linear conversion function whose input is the velocity in miles per hour and whose output is the velocity in feet per second.

16. Suppose the linear function $y = 2,500x - 15,000$ represents the average salary y of a U.S. citizen who has complete x years of school. Explain carefully, using the specific parameters given in the equation and appropriate units, what the slope and y-intercept of this equation mean.

17. Suppose a baby boy's birth weight was 8.25 pounds and he gained 2.1 pounds each month for the first few months of his life. The function describing his weight is linear and could be written as

$$y = 2.1x + 8.25,$$

where x is his age in months and y is his weight in pounds.

(a) Give the slope for the linear equation, and explain what the slope means in terms of weight and age.

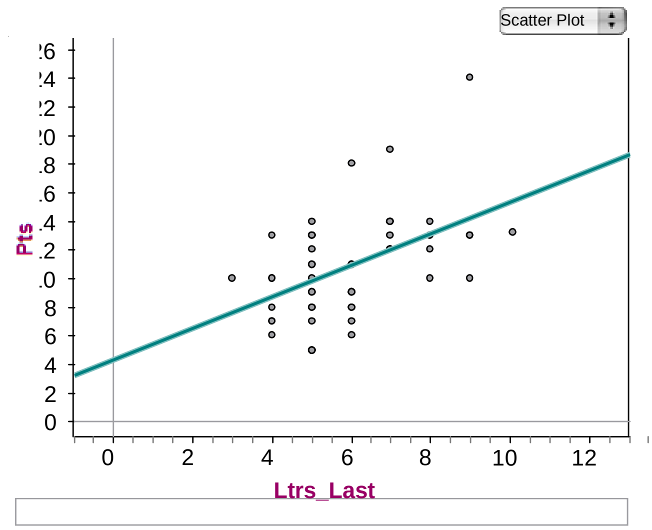
(b) Determine the baby's weight after 3 months.

(c) Determine the child's weight after 36 months. Is your answer reasonable? Explain.

(d) If the baby's birth weight was 8.5 pounds instead of 8.25, how would the function change?

(e) If the slope of the function were 2.0, what would this mean?

18. The following is the graph of a scatterplot and its regression line. The data is from students in two algebra sections in the fall of 2007. The relationship is scrabble points versus number of letters in the last name for each student.



(a) One student claimed that the regression equation gave a value of 3 scrabble points for a student with 2 letters in their last name. Is this correct? Why or why not?

(b) Another student claimed that the regression equation gave a value of 15 scrabble points for a student with 10 letters in their last name. Is this correct? Why or why not?

(c) Which of the following is the equation of the regression line for this data? For each answer you do *not* choose, explain why.

i. $y = 4.2x + 1.1$

ii. $y = 1.1x + 4.2$

iii. $y = 10.1x + 7.5$

iv. $y = -1.1x + 4.2$

(d) Using your answer from the previous question, give the slope and y-intercept of the regression equation and explain what each of these mean in terms of scrabble points and number of letters in a person's last name.

19. The table below gives the data for U.S. per capita energy consumption but in quadrillions of BTU's, rather than billions of barrels oil equivalent as in Figure 4 in the text.

Year	82	83	84	85	86	87	88
Consumption	73.2	73.1	76.7	76.4	76.7	79.2	82.8
Year	89	90	91	92	93	94	95
Consumption	84.9	84.6	84.5	85.9	87.6	89.3	91.2
Year	96	97	98	99	100	101	102
Consumption	94.2	94.7	95.2	96.8	98.9	96.3	97.4

(a) Find the linear regression equation for this data, letting $x = 0$ represent the year 1980.

(b) What is the correlation coefficient for this data?

(c) Explain the meaning of the slope using appropriate units.

(d) Explain the meaning of the y-intercept using appropriate units.

20. Can you ever have a least-square regression line where all the data points are below the line? Explain.

21. What does correlation measure?

22. Give an example of two variables not mentioned in the reading that have a negative correlation.

23. If a person's arm span was *always* 3 inches shorter than his or her height, what would be the correlation between arm span and height? Explain.

24. Explain what is wrong with the following statements about correlation and regression.

(a) The regression equation that describes the relationship between the age of a used Ford Mustang and its value is $y = -1000x + 10,000$, where the input is the age of the automobile (in years) and the output is its value (in dollars). The correlation describing this relationship is 0.83.

(b) The correlation between the number of years of education and yearly income is 1.23.

25. Just because two variables are highly correlated does not mean that one necessarily causes the other. There may be something else that is causing one or both events.

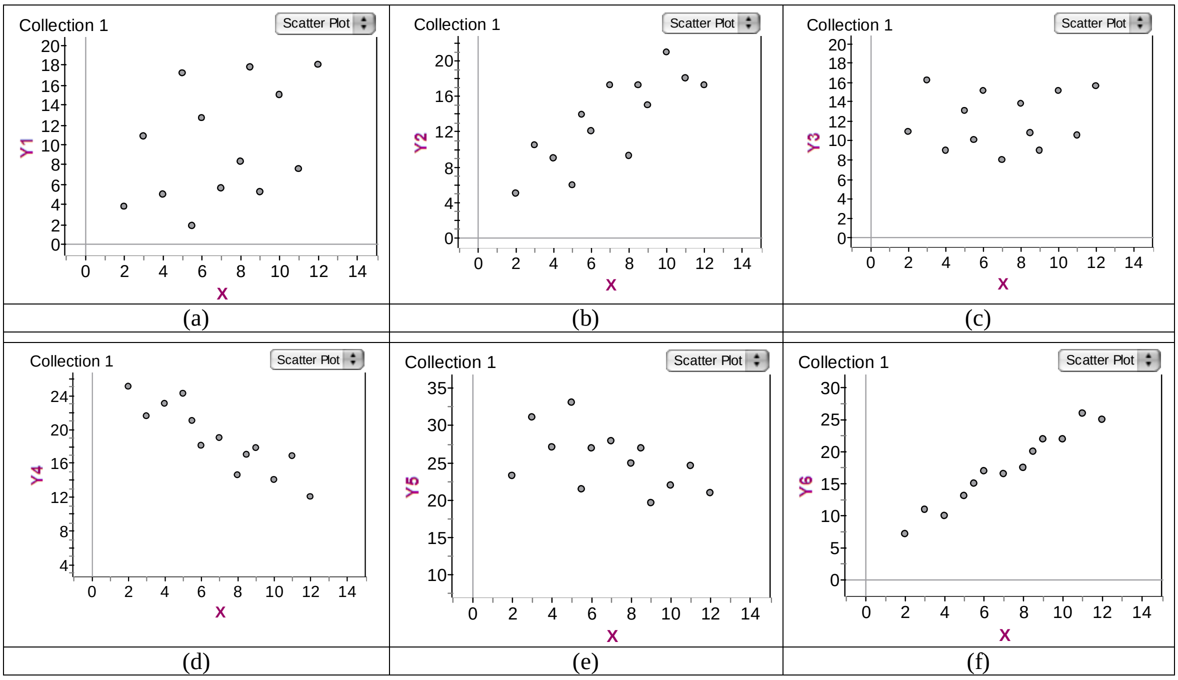
(a) There is a high correlation between a child's height and his or her ability in mathematics. Does this mean that a child's physical growth is causing the increased ability in math? Explain.

(b) There is a high correlation between the outside temperature and the number of colds people get. Does this mean that the number of colds people get causes the temperature to decrease? Explain.

(c) There is a high correlation between the number of drinks someone has at a bar and incidents of lung cancer. Does this mean that drinking causes lung cancer? Explain.

26. Match the following correlations with the corresponding scatterplot.

(i) $r = 0.98$ (ii) $r = 0.8$ (iii) $r = 0.4$ (iv) $r = 0.1$ (v) $r = -0.5$ (vi) $r = -0.9$



27. Input the following data into your calculator.

x	1	2	3	4	5	6	7	8
y	5	7	9	11	13	15	17	19

(a) Determine the regression equation and correlation for the data.

(b) What does the correlation tell you about the data and the regression equation?

28. In the reading, we found that the least squares regression line for the Carbon Emissions data shown in the following table was $y = 21.9x + 757$, where x was the number of years since 1990 and y was carbon emissions in millions of metric tons for that year.

(a) The correlation for the regression line was $r = 0.975$. What information does this give you about the carbon emissions data?

(b) Use the regression line to estimate the level of carbon emissions in the commercial sector in 2008. How confident are you about the accuracy of your answer? Explain.

29. The following table gives three variables for 26 countries in North and South America.[72] All the variables measure the female population in the given country. The variables are:

- Illiteracy Rate, measured as the percent of females over 15 years of age who are functionally illiterate.

[72]Illiteracy data is from the World Bank. Life expectancy data is from the World Health Organization. All variables are from circa 2000

- Additional life expectancy at age 10, which represents how many additional years a female aged 10 can expect to live, on the average.
- Survival to 10, the percentage of newborn females expected to live to age 10.

Country	Illit. Rate	Additional Life Exp.	Percent Surviving to Age 10	Country	Illit. Rate	Additional Life Exp.	Percent Surviving to Age 10
Barbados	0.3	68.5	0.99	Panama	8.8	68.2	0.98
Guyana	1.9	61.9	0.93	Jamaica	9.3	67.8	0.98
Uruguay	2.0	69.1	0.98	Ecuador	10.1	66.9	0.96
Trin. & Tob.	2.4	64.8	0.99	Mexico	10.9	68.3	0.97
Argentina	3.2	69.5	0.98	Brazil	13.2	65.1	0.96
Cuba	3.4	68.2	0.99	Peru	14.8	65.4	0.95
Bahamas	3.7	65.9	0.98	Dom. Rep.	16.3	65.3	0.95
Costa Rica	4.4	70.0	0.98	Bolivia	20.8	59.5	0.91
Chile	4.4	70.3	0.99	El Salvador	24.0	66.0	0.96
Belize	6.9	67.3	0.97	Honduras	25.0	64.4	0.95
Paraguay	7.8	66.7	0.97	Nicaragua	33.3	64.3	0.96
Venezuela	8.0	68.3	0.98	Guatemala	38.9	62.6	0.94
Colombia	8.4	66.8	0.98	Haiti	52.2	52.6	0.89

(a) Consider the relationship Additional Life Expectancy Versus Illiteracy Rate. What is the correlation? What does this number tell you about the relationship between illiteracy and life expectancy?

(b) Consider the relationship Percent Surviving to Age 10 Versus Illiteracy Rate. What is the correlation? What does this number tell you about the relationship between illiteracy and life expectancy?

(c) How does the strength of the first relationship compare with the second? Is this surprising?

(d) Make separate scatterplots of both relationships. In each case, which countries seem to be outliers? If these outliers were removed from the data, what would happen to the correlation coefficients? (You could do this by recalculating the coefficients, but you should be able to answer correctly without doing this!)

(e) Give the slopes for the linear regression equations for both relationships and explain what these slopes mean in the given context.

Linear Functions: Activities and Class Exercises

1. **Protein Requirements Revisited.** In 1985 the World Health Organization (WHO) published revised daily protein requirements which translate into 56g of protein a day for a 75kg man, and 48 grams for a 64 kg woman

   (a) Assume that the amount of protein a person requires is proportional to the person's weight. Using this assumption and the information given above, find the linear function expressing the protein requirement for women as a function of weight.

   (b) Explain what the slope of the line you found in part a) means. Do you get the same slope whether you use the data given for women as for men?

   (c) Calculate how many grams of protein each of the members of your team requires.

   (d) The table below gives the number of calories for each gram of protein for a variety of plant foods and lean ground beef. [73]

Food	No. of cals per gram protein	Number of cals total for protein RDI
Spinach	5	
Broccoli	10	
Collards	12.5	
Lettuce(1)	14	
Celery	20	
Corn (3)	28	
Potato	44	
Soybeans	12	
Lentils	13.5	
Peas (4)	15	
Peanuts	21	
Wheat(2)	25	
Oatmeal	24	
Sunfl. sds.	27	
Almonds	28	
Grd. beef	11	

   (1) Iceberg variety. (2) Hard Spring Wheat. (3) fresh. (4) frozen

   Given the information in the second column, one can calculate how many calories from any particular food item one would need to eat to get their entire RDI of protein from that food item. Calculate this value for each of the food items in the table for a 64kg woman, putting the results in the third column of the table.

   (e) Which of the food items in the table could a typical 64kg women use to fulfill her protein needs without going over the calorie limit of 2100 recommended for such an individual?

   (f) What additional important information about these food items would be helpful in planning a daily menu which fulfills basic nutritional needs?

[73] Data from the USDA, was accessed at http://www.caloriecountercharts.com/chart3a.htm

2. **World Population Growth Rates.** As noted in the Activities for Chapter Four, the annual growth rate of world population has been decreasing for a number of years, and the graph below shows the percent growth rate for world population, going back to 1950 and projected out to 2050.

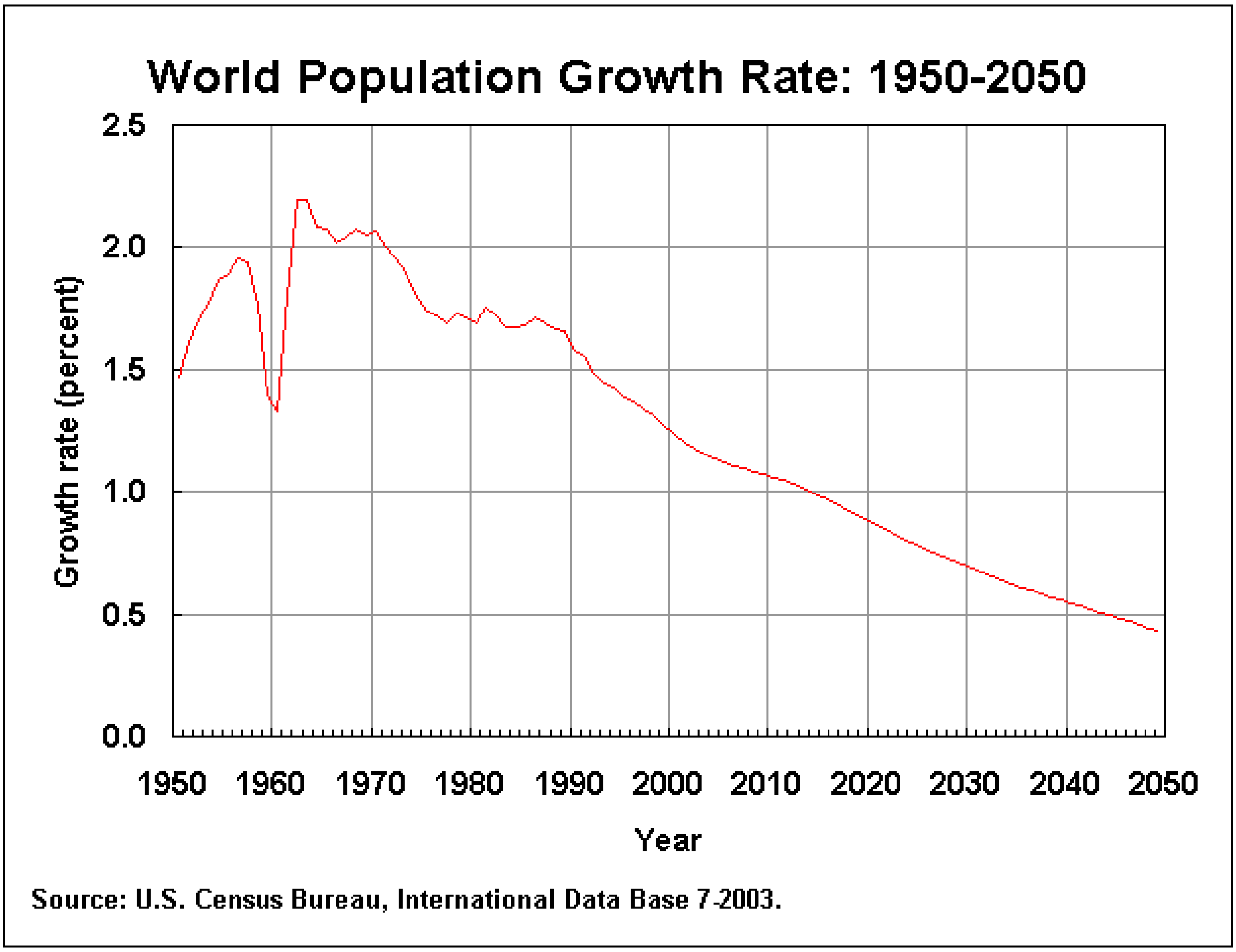


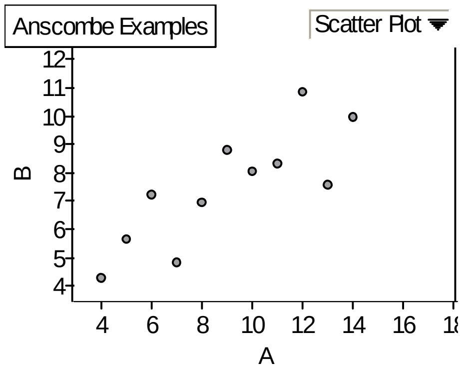
Figure 11: World population growth rates

(a) After about 2000, and certainly after 2020, the graph is very nearly linear. Looking at just the portion of the graph from 2020 on, estimate the slope of the graph. In words, describe what the slope is telling you about world population growth rates.

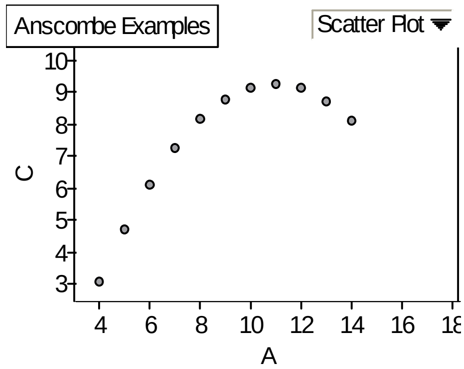
(b) Using the slope from part (a), and assuming the graph continues in the same way, estimate when the growth rate will reach 0%.

(c) What would it mean for the population growth rate to be negative? Do you think this is possible? Why or why not?

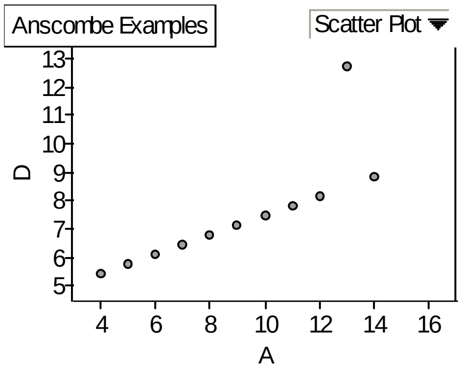
3. Anscombe II. The graphs below are scatter plots of 4 different two-variable data sets produced by Anscombe[74] as examples related to linear regression. You may have done an activity using these graphs in chapter three. We have labeled the 4 plots A, B, C, and D to the lower left of each plot.



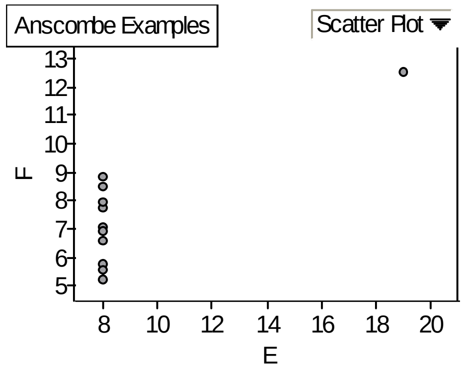
A.



B.



C.



D.

(a) For each of the 4 graphs, make your best prediction of the correlation coefficient r for the data. Using your estimates, rank the 4 data sets from highest to lowest with respect to your estimated correlation coefficients.

(b) On each of the 4 graphs, draw in your best intuitive guess for the linear regression line. Then, for each of the 4 lines, estimate the slope of the line.

(c) In fact, all four data sets have exactly the same correlation coefficient of $r = .82$, and they all have exactly the same linear regression line of $y = .5x + 3.0$. Compare your estimates for r and for the slope of your estimated linear regression lines with the actual values.

(d) Even though the correlation coefficients and regression equations are identical, the linear regression model is not equally appropriate in all cases. For which of the 4 data sets is the linear regression model a good model?

(e) Consider the C graph (with variable D on the vertical axis). If this graph was based on real-world data, what explanation might there be for the outlier in this graph?

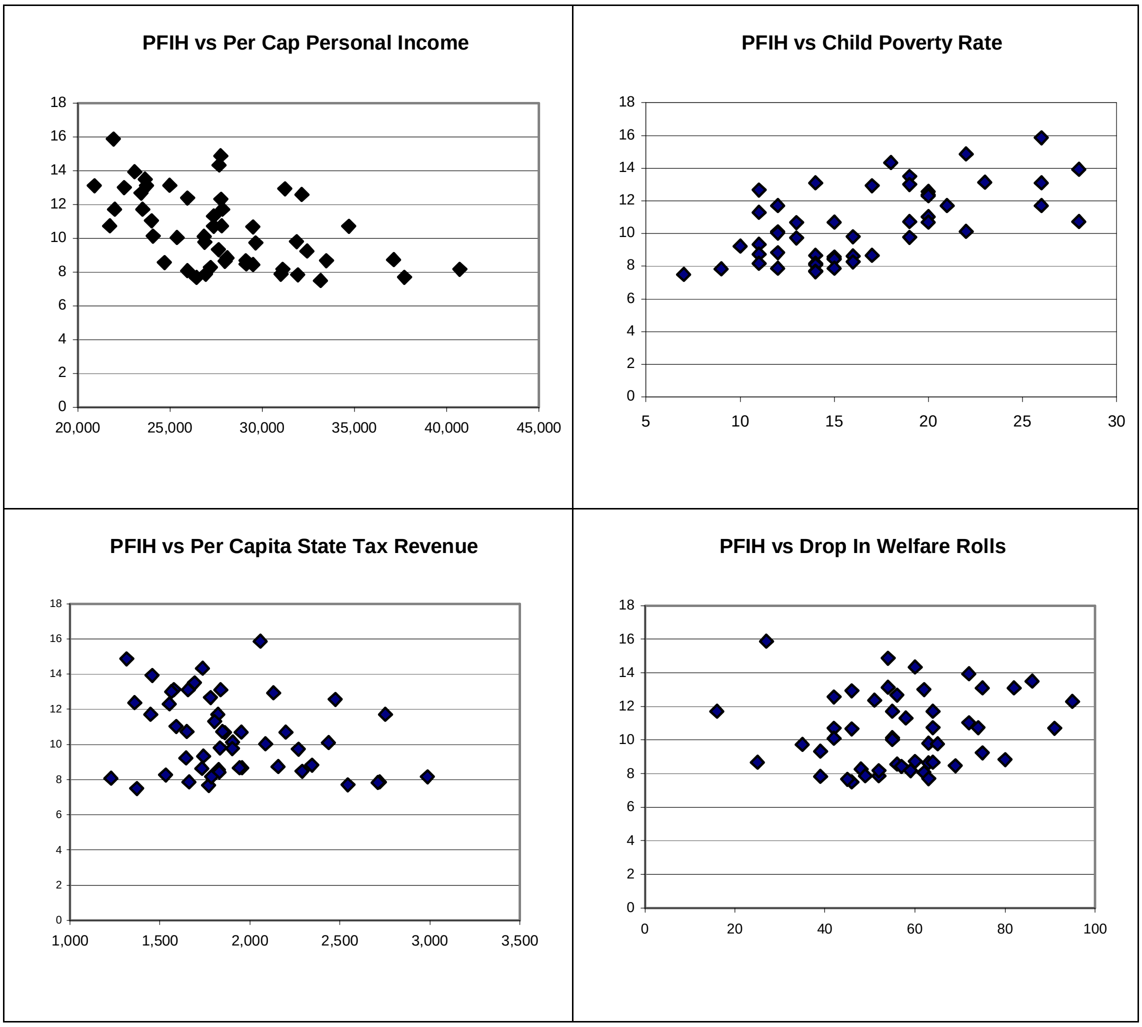
[74]Data from Anscombe, F.J., Graphs in Statistical Analysis, American Statistican, 27, 17-21

4. Food Insecurity. The USDA uses the following definition for food insecurity.

> Food insecurity is limited or uncertain availability of nutritionally adequate and safe foods or limited or uncertain ability to acquire acceptable foods in socially acceptable ways.

One way to measure this is using Percent Food Insecure Households (PFIH), which gives the percent of all households which are experiencing food insecurity. In this activity, we'll look at PFIH and its relationship with four other variables. These variables, and scatter-plots of PFIH as the dependent variable versus them are given below. Each point represents one of the 50 states. South Dakota's PFIH was 8.07% in 2000. The other four variables are:

- Per Capita State Tax Revenue
- Percent Drop in Welfare Rolls 1995-2002
- Per Capita Personal Income
- Child Poverty Rate



A few words about the variables.

- Per Capita State Tax Revenue is equal to the total amount of state tax revenues divided by the population. For South Dakota, this was $1228 in 2000, lowest among the fifty states.[75]
- Percentage Drop in Welfare Rolls measures the percent decrease in the number of families on welfare from 1995 to 2002. The reduction in welfare rolls in each state is largely a result of the Welfare Reform legislation that was passed during the Clinton Administration. South Dakota's drop was 62%. On average, more than 6 out of every 10 families on welfare in 1995 were dropped from welfare rolls. However, one must keep in mind that this was not a fixed population, and some families went on welfare while others moved into or out of the state, etc. [76]
- Personal Income measures how much personal income was generated per person, on the average. It equals the total amount of income generated in that state divided by the mid-year population of that state. For South Dakota, this was $25,958 in 2000.
- The Child Poverty Rate measures what percent of children in each state are classified as living in poverty. The rate for South Dakota was 14% in 2000.[77]

(a) The four scatter-plots above each show a relationship between one of the given variables and PFIH. For each of these relationships, consider the plots, and say whether there seems to be a positive association, a negative association, or no association, and indicate why you decided as you did.

(b) The table below shows the correlation coefficients for each of the four relationships, along with the squared coefficients r^2.

Variable	r^2	r
Per Capita State Tax Revenue	0.076428	-0.27646
Percent Drop in Welfare Rolls	0.012054	0.109792
Per Capita Income	0.277138	-0.52644
Child Poverty Rate	0.392424	0.626437

i. According to these coefficients, which relationships show a positive association? Which relationships show a negative association?

ii. Which of the 4 relationships shows the weakest association (one that seems to be closest to having no association)?

iii. Which relationship shows the strongest positive association? Do you think it is surprising that this variable has the strongest positive association, or would you have expected this? Explain briefly.

iv. Which relationship shows the strongest negative association? Do you think it is surprising that this variable has the strongest negative association, or would you have expected this? Explain briefly.

(c) When the welfare reform legislation was passed in the mid-1990's, many critics predicted that there would be adverse consequences, including an increase in hunger and food insecurity. The argument was that as individuals and families were dropped from welfare programs, many would have trouble making ends meet, including procuring enough food. According to the information above, does this prediction seem to have proven true or not? Explain as completely as you can why you think so.

[75] Data is from the U.S. government Census Bureau, http://www.census.gov/govs/statetax/00staxss.xls

[76] Welfare data is from http://www.infoplease.com/ipa/A0774474.html

[77] Income and poverty data found at http://www.hungerfreeamerica.com/facts/statistics/index.cfm

5. **Ozone Ouch.**[78] The following table gives the maximum area of the ozone hole over Antarctica for most years since 1980.[79] The area is in millions of square kilometers.

Year	Area (Millions sq. km)	Year	Area (Millions of sq. km)
1980	3.27	1993	24.017
1981	3.15	1994	23.429
1982	10.8	1995	NA
1983	12.24	1996	26.96
1984	14.65	1997	25.13
1985	18.79	1998	28.21
1986	14.37	1999	26.09
1987	22.45	2000	30.31
1988	13.76	2001	26.52
1989	21.73	2002	21.74
1990	21.05	2003	28.51
1991	22.60	2004	22.76
1992	24.90	2005	26.77

For purposes of comparison, the United States is a little over 9 million square kilometers, and North America is a little over 25 million square kilometers.

(a) Let time be measured in years since 1980. So 1980 is represented by $t = 0$, 1990 by $t = 10$, etc. Find the equation for the regression line where years since 1980 is the input and the area of the ozone hole is the output.

(b) Using your equation from part (a), compute the area of the ozone hole for 1980, 1987, 1994, and 2001. For each year, find the difference between your result and the actual area of the hole given in the table. What does this tell you about your regression line?

(c) What is the correlation for your regression line? What information does this give you about how well the line fits the data?

(d) Using your equation from part (a), predict the area of the hole in the year 2015. How confident are you about the accuracy of your answer? Why?

(e) Notice that the trend in the graph seems to be steeper before 1990 than after 1990. Suppose that you were doing this research in 1990 and only had the data through that year. Find the regression line and correlation coefficient for this smaller data set. (One way to do this in Fathom is to create a copy of the case table, and delete the years you do not want to include). Let's call this second linear model $f_2(x)$

i. Graph $y = f_2(x)$ together with the full data set. This can be done in Fathom by adding a movable line, and adjusting it until the equation approximates the one you have for $y = f_2(x)$. How well would f_2 have predicted the area of the ozone hole for the years 1990 through 1995? 1996 through 2001?

ii. Does this new information change your answer to part (d)?

[78]adapted from *Understanding Our Quantitative World* by Janet Andersen and Todd Swanson

[79]Data from http://www.theozonehole.com/ozoneholehistory.htm

6. **Inequality and Poverty.** The table below gives the Poverty Rate[80] as well as the Gini Number (or Coefficient) for 57 countries. The Poverty Rate is the percentage of people in that country considered to be living in poverty. As noted in chapter three, the Gini Number is a measure of income inequality, and can be considered in some sense the "percentage of inequality" in a country.[81]

   The inequality data were collected by Deininger and Squire in the early 1990's. The countries listed are those for which Deininger and Squire had both poverty data and Gini Coefficients. If you are curious, the Gini Number for the USA is estimated at 37.2. One question one might consider is "Do poorer countries have higher levels of inequality?"

   The data is available in the Fathom File **Inequality and Poverty Data**

Country	Pov Rt.	Gini No	Cont.	Country	Pov Rt.	Gini No	Cont.
Bangladesh	49.8	28.27	Asia	Malawi	65.3	62	Africa
Belarus	41.9	28.526	Europe	Mali	63.8	54	Africa
Bulgaria	12.8	27.7967	Europe	Mauritania	46.3	37.8	Africa
Burkina Faso	45.3	39	Africa	Mauritius	10.6	36.69	Africa
Chile	17	52.35	SA	Mexico	10.1	50.31	NA
China	4.6	29.9662	Asia	Moldova	23.3	34.43	Europe
Colombia	64	50.0233	SA	Morocco	19	36.2167	Africa
Djibouti	45.1	38.1	Africa	Nicaragua	47.9	50.32	NA
Dom. Rep.	28.6	49	NA	Niger	63	36.1	Africa
Ecuador	35	43	SA	Nigeria	34.1	39.31	Africa
Egypt	16.7	32	Africa	Pakistan	32.6	31.15	Asia
El Salvador	48.3	44.77	NA	Paraguay	21.8	39.8	SA
Estonia	8.9	37.224	Europe	Peru	49	38.42	SA
Ethiopia	44.2	44.2	Africa	Philippines	36.8	45	Asia
Ghana	39.5	33.94	Africa	Poland	23.8	26.6845	Europe
Guinea	40	40.4	Africa	Romania	21.5	25.33	Europe
Guin. Bissau	48.7	56.12	Africa	Russian Fed.	30.9	30.53	Asia
Guyana	35	40.22	SA	Senegal	33.4	54.12	Africa
Honduras	53	53.726	NA	Sri Lanka	25	30.1	Asia
Hungary	17.3	27.8743	Europe	Tanzania	35.7	51.37	Africa
India	28.6	31.992	Asia	Thailand	13.1	50.264	Asia
Indonesia	27.1	33.1371	Asia	Tunisia	7.6	40.62	Africa
Jamaica	18.7	39.825	NA	Uganda	55	39.6933	Africa
Jordan	11.7	40.66	Asia	Ukraine	31.7	25.71	Europe
Kazakhstan	34.6	32.67	Asia	Venezuela	31.3	49.12	SA
Kenya	42	54.39	Africa	Vietnam	50.9	35.71	Asia
Kyrgyz Rep.	64.1	35.32	Asia	Zambia	72.9	49.1033	Africa
Laos	38.6	30.4	Asia	Zimbabwe	34.9	56.83	Africa
Madagascar	71.3	43.44	Africa				

   (a) Create a scatterplot of *Gini No* versus *Pov Rate*. Visually, how would you describe the association between these variables? Is it positive or negative, weak or strong?

   (b) Find the correlation coefficient for this relationship. What does this additional information tell you about the association?

[80] The data were downloaded from the World Bank website at http://www.worldbank.org/research/growth/dddeisqu.htm.

[81] Recall, for example, that if everyone in the country made exactly the same amount of money, the Gini number would be 0. If one person in the country had all the income, and everyone else had none at all, the Gini number would be 100 (for 100% inequality).

(c) Find the linear regression line for this data. Graph this line on the scatter-plot. Explain what the slope and y-intercept of this line tell you in terms of the two variables.

(d) Use the linear regression equation to predict the Gini Coefficient for countries with 0% poverty, 50% poverty, and 100% poverty. Do these predictions seem reasonable? (Hint: Do your predictions correspond to the approximate values you'd get from the graph?)

(e) It is possible that certain regions or continents might show a weaker or stronger relationship between these variables than the data as a whole.

i. Construct a scatter-plot just for African countries. You can do this by creating a **Filter** in Fathom, or by creating a duplicate table and deleting the unwanted cases. Does the African data seem to show a stronger or weaker association than the data as a whole?

ii. Find the linear regression line and correlation coefficient for the African data. How do these compare with those from the data as a whole?

iii. Find the linear regression line and correlation coefficient for the Asian countries. How do these compare with those from Africa? Do you find anything surprising?

(f) Look over the list of countries shown. Are there any particular types of countries or regions that seem to be left out?

(g) Finally, let's investigate which continent, Africa or Asia, has more inequality, as measured by the Gini Coefficient. Create one histogram for Gini Coefficients, including separate bars for Africa and Asia. An easy was to do this in Fathom is to create a graph with *Continent* as the Y attribute and *Gini No* as the X attribute. This will give you all 5 continents, and you can select Africa and Asia using a Filter. What do you conclude?

7. Comparing Car Models. The data in the table below represent 38 car types of 7 different makes and 19 different models and variations for the 2005 model year. The data is from www.autos.msn.com. All models have automatic transmissions. The average price is based on the Kelley Blue Book price for the 2004 model year. City Mileage, Highway Mileage, and Greenhouse Gas Emissions ratings are as determined by the EPA. Gas emissions is given in tons of carbon dioxide per year under normal driving conditions. Engine Size is in liters. Passenger Room and Luggage Room are given in cubic feet.

Make	Model	Price	City mpg	H'way mpg	Carbon Emiss.	Engine Size	Passenger Room	Lugg. Room
Honda	Civic	15325	29	38	5.9	1.7	91	10
Honda	Accord (2dr)	22622.5	24	34	7	2.4	103	14
Honda	Accord (4dr)	22622.5	24	34	7	2.4	103	14
Toyota	Avalon	23500	22	31	7.6	2.2	107	14
Honda	Accord (2dr)	22622.5	21	30	7.8	3	103	14
Honda	Accord (4dr)	22622.5	21	30	7.8	3	103	14
Toyota	Corolla	15667.5	30	38	5.7	1.8	84	14
Honda	Civic	15325	29	38	5.9	1.7	91	10
Ford	Crown Victoria	27475	18	25	9.2	4.6	111	21
Toyota	Matrix 4x2	16855	28	34	6.2	1.8	94	22
Toyota	Matrix 4x4	16855	28	34	6.2	1.8	94	22
Toyota	Celica	18600	29	36	6	1.8	78	17
Toyota	Camry	21875	24	34	6.9	2.4	102	17
Chevrolet	Cobalt (4dr)	17527.5	24	32	6.7	2.2	87	14
Ford	Taurus SW	13875	19	25	8.8	3	109	39
Toyota	Camry	21875	21	29	8	3.3	102	17
Ford	Taurus	22165	20	27	8.4	3	105	17
Toyota	Camry	21875	20	28	8.2	3	102	17
Chevrolet	Monte Carlo	25317.5	21	32	7.6	3.4	96	16
Toyota	Corolla	15667.5	30	38	5.7	1.8	84	14
Chevrolet	Aveo	11160	26	34	6.5	1.6	91	12
Saturn	Ion (2dr)	11900	24	32	6.7	2.2	91	15
Saturn	Ion (4dr)	11900	24	32	6.7	2.2	91	15
Ford	Focus	15842.5	26	32	6.8	2	94	15
Chevrolet	Malibu Classic	21630	25	34	6.7	2.2	99	16
Toyota	Camry	21875	24	34	6.9	2.4	102	17
Honda	Accord (4r)	22622.5	24	34	7	2.4	103	14
Chevrolet	Cobalt (4dr)	17527.5	24	32	6.7	2.2	87	14
Dodge	Stratus (4dr)	21845	22	30	7.6	2.4	94	16
Toyota	Camry	21875	21	29	8	3.3	102	17
Honda	Accord (4r)	22622.5	21	30	7.8	3	103	14
Dodge	Stratus (4dr)	21845	21	28	8.1	2.7	94	16
Toyota	Camry	21875	20	28	8.2	3	102	17
Dodge	Stratus (2dr)	21845	21	28	8	2.4	86	16
Dodge	Stratus (2dr)	21845	20	28	8.3	3	86	16
Dodge	Neon	13750	25	32	6.8	2	90	13
Chevrolet	Cavalier (2dr)	14017.5	24	34	6.7	2.2	92	14
Chevrolet	Cavalier (4dr)	14017.5	24	34	6.7	2.2	92	14

(a) How closely related is city mileage and highway mileage for these car models? To answer this question, create a scatterplot and find the linear regression line and correlation coefficient for these two variables.

  i. What is the slope of your linear regression equation and what does this slope mean in terms of city and highway mileage?

  ii. What is the y-intercept of your linear regression equation and what does this slope mean in terms of city and highway mileage?

(b) Let's consider greenhouse gas emissions.

i. What is the average amount of carbon dioxide emissions for these models?

ii. Of the variables city mileage, highway mileage, average price, engine size, and passenger room, which is most highly correlated with carbon dioxide emissions? Explain how you know.

iii. Based on these data, in order to reduce greenhouse gas emissions by one ton per year, how much would you need to increase your city gas mileage? Explain how you know.

(c) Do these data support the statement "consumers who wish to reduce their contribution to greenhouse gas emissions are going to have to pay more for their vehicles"? Consider how you can use the data to answer this question and justify your answer.

(d) This past fall, the "big three" U.S. automakers, Ford, Chrysler (Dodge), and General Motors (makers of Chevrolet and Saturn) lobbied congress and the President for "bailout loan funds." These car makers were criticized for not producing the fuel efficient, "green" vehicles that the times demand. Based on these data, is it fair to say that U.S. car models are less fuel efficient and less environmentally friendly than cars produced by foreign car companies like Honda and Toyota? Be sure to justify your answer.

8. Miles Per Gallon.[82] With the significant increases in the cost of gasoline that have occurred since 2005, owning a vehicle is becoming even more expensive. The purchase price is the biggest one time cost. Continuing costs include not only gasoline, but maintenance, and insurance. The cost of gasoline is related to the miles per gallons averaged by the car. More efficient cars will reduce this cost, sometimes dramatically (assuming you continue to drive the same number of miles).

(a) Suppose you were considering buying a car in the fall of 2008 or 2009, and were choosing between a Honda Accord and a Ford Taurus. A 2-door Honda with a 2.4 liter engine and an automatic transmission costs about $23,725 MSRP (manufacturers suggested retail price), gets 21 miles per gallon in the city and 30 on the highway. A 4-door Taurus with a 3.5 liter engine has a MSRP of $25,670, gets 18 mpg in the city and 28 on the highway. Assume most of the miles you drive are highway miles, and the cost of gasoline is $1.69 per gallon. We'll also assume maintenance and insurance costs are roughly the same for both models.[83]

i. Write linear equations for the Accord and the Taurus where the input is miles and the output is the cost for owning and operating the vehicle. Include the initial cost of buying the vehicle and gasoline costs but ignore maintenance and insurance costs.

ii. Graph the equations on the same axis system, using a window that covers how long you expect the cars to last, and shows the intersection point of the two graphs. Indicate the window used and the intersection point.

iii. What is the physical meaning of the intersection point?

(b) Some energy experts consider $5 per gallon gasoline a realistic possibility in the near future. Redo part (a) under the alternate assumption that gas is $5 per gallon.

(c) Some metropolitan areas are looking at light rail transit and other forms of mass transportation as a way to reduce traffic, pollution, and gasoline usage.

i. How much would it cost to run a car that is driven 15,000 each year and gets 24 mpg on the average. Assume gas is $5 per gallon.

ii. How much would it cost to run 200,000 such cars for a year?

iii. How much money would be saved (say, to help pay for mass transit) if these 200,000 cars cut their mileage by two-thirds, to 5,000 miles from 15,000?

[82]adapted from Understanding Our Quantitative World by Janet Andersen and Todd Swanson

[83]Information on these two models is from www.kbb.com, the Kelley Blue Book site

6. Exponential Functions

The single biggest problem of the human race is its inability to understand the exponential growth function
—Albert Bartlett

In the last chapter, we considered linear functions, our first "family." We saw that linear functions provide good models for relationships where we know there is a constant rate of change, or where the scatterplot of the data seems to follow a roughly linear pattern. Of course, lots of relationships do not have a constant rate of change, and lots of of scatterplots or line graphs are not very linear. Recall, for example, Figures 1 and 2 of Chapter Two which depicted Hubbert Curves of energy productions data, or the Lorenz curves we used in introducing the Gini Coefficient in Chapter Two.

In this chapter and the next four, we will discuss several more families of functions, consider some of the basic characteristics that define these families, and see how they can be useful in modeling nonlinear patterns of data. There are certainly many families of nonlinear functions, but, as with linear functions, we will consider the simplest first and progress to some more complicated examples. In this chapter, we will consider the **exponential** family of functions.

Exponential Functions Symbolically

Symbolically, linear functions are those which can be written in the form $y = mx + b$. We called m the slope, and b is the y-intercept. To define a particular linear function, one need only specify particular values for m and b. We call m and b **parameters**. Although designated by letters, parameters are not considered variables (x and y are the variables here), but rather are considered unspecified constants.

In linear functions, the independent variable x does not involve an exponent, unless you want to write it as $x = x^1$. Exponential functions, and also power functions which will be considered in detail in the next chapter, will involve exponents that are not equal to 1. Recall that, in an expression of the form b^r, b is called the **base** and r is called the **exponent**. The appendices contain some additional background and examples related to exponents and roots that you can refer to as needed throughout this section.

- **Exponential functions** are those which can be written in the form $y = a \cdot b^x$. The variable is an *exponent* and the *base* is a constant.
- **Power functions**, which we define here but which will be discussed in detail in Chapter Eight, are those which can be written in the form $y = ax^b$. The variable is a *base* and the exponent is a constant.

In both cases, a and b are **parameters**,[84] just as m and b are parameters for the family of linear functions $y = mx + b$. The families are defined by which operations are performed on parameters, and which on the independent variable. In exponential functions, the independent variable is the exponent, and we denote the base by the parameter b.

Here are some examples of exponential functions:

$$y = 3 \cdot 2^x \qquad f(x) = 2.15(0.87)^x \qquad p(t) = 6.2(1.016)^t$$

Here are some examples of power functions:

$$y = 3 \cdot x^2 \qquad f(x) = 2.15(x^{0.87}) \qquad p(l) = 10.5(l)^{1.5}$$

[84]Which letters are used to denote the parameters is a somewhat arbitrary choice. We could have written $y = ba^x$ so that a was the base. The important thing is to substitute consistently into the general equation given for a family.

Example 1. Here are a few more functions in symbolic form. Identify which of the three families, linear, exponential, or power, they belong to, or if they belong to none of the three (there is one of these!).

1. $y = 0.0345x - 5$

2. $y = 10500(x^{2.3})$

3. $y = 2x(2.3 + x)$

4. $y = 2^x$

5. $y = 3.4(1.032^x)$

Solution: Item 1 is linear, item 2 is a power function, and items 4 and 5 are exponential functions. Item 3 is none of these.[85] ■

Example 2. Sometimes, functions might appear "in disguise," so that you might have to do a little algebraic manipulation to recognize them. Which do you think are linear? Which exponential or power?

1. $y = 2^{x-0.15}$

2. $y = 0.02(x - 0.15) + 2$

3. $y = (3x)^2$

4. $y = 7(10^{0.023x})$

5. $y = 4\sqrt[5]{x^3}$

Solution: Let's go through each of these five examples in turn.

1. Let's try simplifying this, using properties of exponents.

$$y = 2^{x-0.15} = 2^{x+-0.15} = 2^x 2^{-0.15} \approx 2^x(0.901) = 0.901 \cdot 2^x$$

This is an exponential function with parameters $a \approx 0.901$ and $b = 2$

2. This is a linear function.

$$y = .02(x - 0.15) + 2 = 0.02x - 0.02(0.15) + 2 = 0.02x - 0.003 + 2 = 0.02x + 1.997$$

The slope is 0.02 and the y-intercept is 1.997.

3. This is a power function. $y = (3x)^2 = 3^2x^2 = 9x^2$, and so is of the form $y = ax^b$ with $a = 9$ and $b = 2$.

4. You might guess this is also exponential.

$$y = 7(10^{0.023x}) = 7((10^{0.023})^x) \approx 7(1.0544^x)$$

In fact, it is an exponential functions with base $10^{.023}$ or about 1.0544.

[85] In fact, this function is a **quadratic** function, which is another family we will eventually consider

5. We have not considered root functions as yet. However, you may know that roots can be changed to exponents.[86] For example,

$$\sqrt[3]{x^7} = x^{7/3}$$

In our case, we have

$$y = 4\sqrt[5]{x^3} = 4x^{3/5}$$

which is a power function.

In considering these five functions from Example 2, we used basic properties of exponents, the distributive law, and in the last, the definition of fractional exponents. We also used a calculator to get decimal approximations in two cases. Be sure you understand the steps involved in these examples, in case we run into any more functions "in disguise!" ■

Exponential Functions in Graphical Form

In the previous chapter, we saw that any linear function had a symbolic representation of the form $y = mx + b$ and this corresponded to the graph of the function being a straight line. We saw that if m was positive, the graph was increasing, and if m was negative, the graph was decreasing. We also know that b is the value of the y-intercept. How can we characterize the graphs of exponential functions, and how are these characteristics related to the parameters in the symbolic representation?

In these examples, and the ones that follow, we will be interested not only in the general characteristics that are true for all functions in the family of exponential functions, but also how these characteristics are affected by the parameters. We will consider these questions for the other families of functions we consider, as well as the similarities and differences between functions in *different* families.

Figure 1 shows the graphs of exponentials functions $f(x) = 3 \cdot 1.5^x$, $g(x) = 3 \cdot 2^x$, and $h(x) = 3 \cdot 2.5^x$.

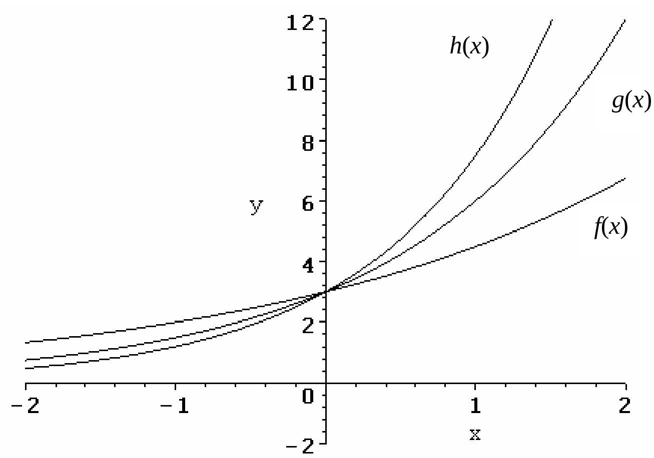


Figure 1: Graphs of $f(x) = 3 \cdot 1.5^x$, $g(x) = 3 \cdot 2^x$, and $h(x) = 3 \cdot 2.5^x$

[86] As noted above, see appendices for more information on exponents and roots

The graphs are fairly simple, so in some sense there may not seem to be much to notice. We can say that all the graphs are increasing, and they are all concave up. We might also notice that the graphs all have the same y-intercept of 3, and that this is the the same as the parameter a in the symbolic form $y = ab^x$! Why is this? Recall that the y-intercept is the y-value of the functions when $x = 0$. In the case of an exponential function $y = ab^x$, if $x = 0$, we have

$$y = ab^0 = a \cdot 1 = 1$$

since any non-zero number to the zero power is equal to 1.

Do the graphs have any x-intercepts? At least in the window we see, there are none. In fact, there are no x-intercepts because if b is any positive number, b^x will also be positive for any value of x.

How does the base b affect the graphs? If we look on the right side of the graph (where $x > 0$), we see that the larger the base, the steeper the graph. For the function h, $b = 2.5$ which is the largest base and h has the steepest graph. On the other hand, to the left of the origin ($x < 0$), all the graphs are very flat. As you go further to the left (in other words, as x becomes more and more negative) the graphs seem to become flatter and the y-values become closer to 0. We say that y approaches 0, as x goes to negative infinity, and that $y = 0$ is a **horizontal asymptote** for the graph. An asymptote is a straight line which a graph gets closer and closer to, oftentimes without ever touching. Exponential functions get closer and closer to the x-axis without touching.

Why does this happen? Recall that, if n is positive, b^{-n} means the same as $1/b^n$. In other words, negative exponents mean "take the reciprocal." When we let x be a negative number, then b^x is really the reciprocal of b to the corresponding positive power. For example, if $x = -3$ and $b = 2$, $b^x = 2^{-3} = 1/2^3 = 1/8$, which is fairly close to zero. If the base b is bigger than 1, then the "more negative" the exponent is, the larger the denominator in the resulting fraction, and the smaller the resulting fraction.

Example 3. In Figure 1, we looked at three functions of the form $y = ab^x$, but in all cases, b was bigger than 1. What happens if we use b where $0 < b < 1$? On your graphing calculator (or a computer software program), graph several examples of exponential functions where the parameter b is between 0 and 1. Write a short paragraph or two, similar to what we did above, describing the characteristics of these exponential functions and how the parameters a and b affect the graphs. You might also experiment with different graphing windows. The main idea is for you to "play around" with examples of these types of exponential functions until you feel like you understand how they work.

Solution: Figure 2 shows some possible examples for graphs of $y = ab^x$ with $0 < b < 1$. You may or may not have graphed any of these particular functions.

These graphs are also concave up, but instead of increasing, they are all decreasing. They are steeper on the left and flatter on the right, and these functions seem to level off at $y = 0$ as x becomes larger and larger (in other words, as x goes to infinity). As in the previous case, $y = 0$ seems to be a horizontal asymptote. Also as before, the parameter a represents the y-intercept. ■

Finally, note that since exponential functions are either increasing throughout their domain, or decreasing throughout their domain, every exponential function is a one-to-one function. Recall from Chapter Four that this means that if $f(a) = f(b)$, then a must equal b. So, for example, if $2^a = 2^b$, this can only happen if $a = b$.

Numerical Representation of Exponential Functions

If we have a function given in numerical form, we can determine if it is linear by seeing if it has a constant rate of change. In other words, for each one unit increase in x, we can check to see if we always have the same increase in y. Alternatively, we can see if any two points selected from the table give us the same slope.

Is there a way we can tell if a function given in numerical form is exponential?

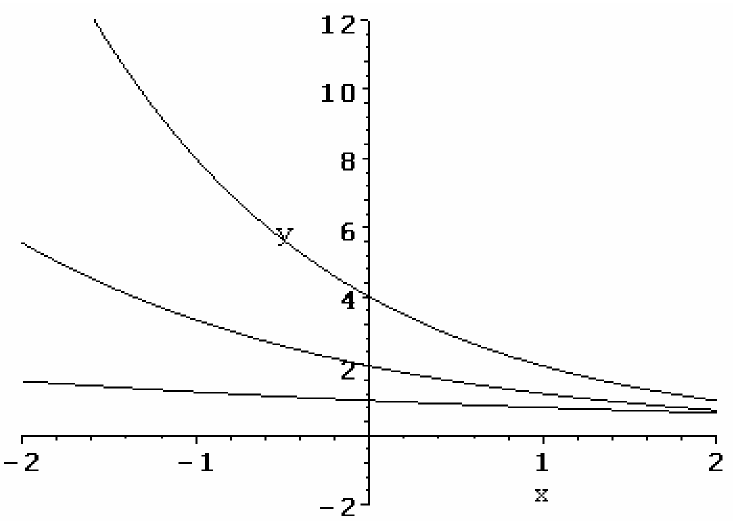


Figure 2: Graphs of $f(x) = 0.8^x$, $g(x) = 2(0.6)^x$, and $h(x) = 4(0.5)^x$

YES!!

While a linear function $g(x)$ has a constant rate of change, equal to $g(x+1) - g(x)$, we will see that an exponential function $f(x)$, has a constant **growth factor**, which is equal to $\frac{f(x+1)}{f(x)}$. For example, let E denote the exponential function $E(x) = 2^x$ (parameters $a = 1$ and $b = 2$). Some values for $E(x)$ are given in Table 1. Notice that if we take any y-value divided by the previous y-value in the table, the result is 2. In other words $\frac{E(x+1)}{E(x)} = 2$ for any x. The growth factor of this function is 2. Alternatively, we could write $E(x+1) = 2 \cdot E(x)$, so that taking any y-value times the growth factor of two gives the next y-value in the table. Thus, the result of increasing x by 1 is that y gets *multiplied* by the growth factor 2.

> In other words, addition is to linear functions as multiplication is to exponential functions. Table 1 shows an example of an exponential function with growth factor 2 and an linear function with slope 2 for comparison.

Table 1: Exponential function $E(x) = 2^x$, and linear function $L(x) = 2x + 1$.

x	0	1	2	3	4	5	6	7	8	9	10
$E(x)$	1	2	4	8	16	32	64	128	256	512	1024
$L(x)$	1	3	5	7	9	11	13	15	17	19	21

Notice how the exponential function begins to increase very, very rapidly as compared to the linear function's constant growth rate.

Let's generalize this. If $E(x) = a \cdot b^x$, then we can create a table for the function as always, except that the values will involve the parameters a and b.

Notice that if we multiply any y-value by the growth factor b, we get the next y-value ($ab^2 \cdot b = ab^3$, $ab^3 \cdot b = ab^4$, etc.). So, the base b is the growth factor.

Table 2: Exponential function where $E(x) = ab^x$.

x	0	1	2	3	4	5	6	$\cdots$	n
$E(x)$	a	ab	ab^2	ab^3	ab^4	ab^5	ab^6	$\cdots$	ab^n

Example 4.
For each of the following tables, say which are tables for exponential functions, which for linear functions, and which for neither.

1.

x	-2	-1	0	1	2
$f(x)$	1	4	16	64	81

2.

x	-2	-1	0	1	2
$f(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$

3.

x	-2	-1	0	1	2
$f(x)$	0.8	0.4	0	-0.4	-0.8

4.

x	-2	-1	0	1	2
$f(x)$	0.3	0.45	0.675	1.0125	1.51875

Solution:
Notice that, in each case, the x-values in the table increase by 1 each time, so we can check whether the functions are exponential by calculating the ratios $\frac{f(x+1)}{f(x)}$.

1. This function looks like it may be exponential since $\frac{4}{1} = \frac{16}{4} = \frac{64}{16} = 4$, but since the ratio $\frac{81}{64} = 1.265625 \neq 4$ the function is not exponential. If we changed the value 81 to $4 \cdot 64 = 256$, then we would have an exponential function.

2. This function is exponential with common ratio $1/2$.

3. This function is not exponential since $\frac{0.4}{0.8} = \frac{1}{2} \neq \frac{0}{0.4}$. Instead, this is actually a linear function where the slope is $m = -0.4$. In addition, note that 0 is an output for this function, which can never be true for an exponential function.

4. Calculating the ratios, we get a common ratio of $\frac{0.45}{0.3} = 1.5$ (you can check the others via calculator). So, this is an exponential function.

■

The Number e

Any positive number $b \neq 1$ can be used as the base for an exponential function $f(x) = ab^x$. Some bases are used more often than others, however. The most frequently used are probably $b = 10$ and $b = e$. The latter is particularly used by many software packages.

The number e is approximately 2.71828. Like the symbol "π" which stands for the number used in formulas for the area and circumference of circles ($\pi \approx 3.14159$), the symbol "e" is used simply as a short hand for a rather complicated number that we do not want to have to write out constantly. Like π, the number e is an irrational number. You may recall that this means it is represented by an infinite non-repeating decimal. In practice, calculations with e are done via calculators; nearly any calculator these days has an e^x button.

Additional background on e as well as some history on why it has come to be important are provided in the appendices. The main importance for us will be the use of e in software packages, and the interchangeability between base e and other bases. Also, a useful fact is that

$$(1 + r)^t \approx e^{rt}$$

for "small" values of r. Recall that in the expression $(1 + r)^t$, $(1 + r)$ is the growth factor and so r represents the percentage growth rate. Similarly, in the expression e^{rt}, the r is also referred to as the percentage growth rate. Table 3 gives some values of these two expressions for various r for purposes of comparison.

Table 3: Comparison of $(1 + r)^t$ and e^{rt}.

r	0.01	0.03	0.05	0.1	0.15	0.2
$(1+r)^t$	1.0510	1.1593	1.2763	1.6105	2.0114	2.4883
e^{rt}	1.0513	1.1618	1.2840	1.6487	2.1170	2.7183

If you have a function of the form $f(x) = a \cdot e^{rt}$, you can convert it to the form $a \cdot b^t$ simply by calculating e^r. For example, $2e^{0.035t} = 2(e^{0.035})^t \approx 2(1.03562)^t$.

Non-linear Regression Models

Many spreadsheets, statistical software packages, and graphing calculators, including TI graphing calculators, have built in procedures for finding exponential and other **non-linear** regression models. Typically, exponential and logarithmic regression are the most popular, but you may also find power regression, quadratic regression, and other regressions available, depending on what technology you are using. These non-linear regression features give you the function from the particular family of functions (eg. exponential) that "best" fits your data. We will discuss the details of how the technology is finding these models later, as well as how the definition of "best" may be somewhat different than the definition for linear regression, but for now, let's just do a few examples using these capabilities. More detailed explanations on how to find non-linear regressions using a TI graphing calculator, as well as how to find exponential and logarithmic regressions with EXCEL, are give in the Appendices.

For example, in Chapter Four[87], we considered two different models for damages due to floods in the U.S. One was a linear model, and the second was an exponential model (although we did not use the term "exponential" at that point). The linear model was found using linear regression, as in Chapter Five. The exponential model, $D_1(t) = 0.741(1.0168^t)$, was found using **exponential regression**.

Notice that the model graphed with the data is slightly concave up and is increasing, consistent with an exponential function. The exponential model is given by Excel in the graph, and is $y = .7414(e^{.0167t})$. Calculating $e^{.0167}$, we get 1.0168, so this equation is equivalent to the model $D_1(t) = 0.741(1.0168^t)$ originally given in Chapter Four.

Most mathematical technologies will also give you a correlation coefficient for non-linear regressions. As with linear regression, this is a measure of how well the function fits the data. Unfortunately, the correlation coefficients for the exponential regression and other non-linear regressions are not based on the sum of squared errors between the given equations and the data, as they are for linear regression. This means, for example, that if linear regression gives you a correlation coefficient of $r = .9$ and exponential regression gives a value of $r = .95$, you cannot conclude on this basis alone that the exponential regression function fits the data better than the

[87]See Example 3 and preceding.

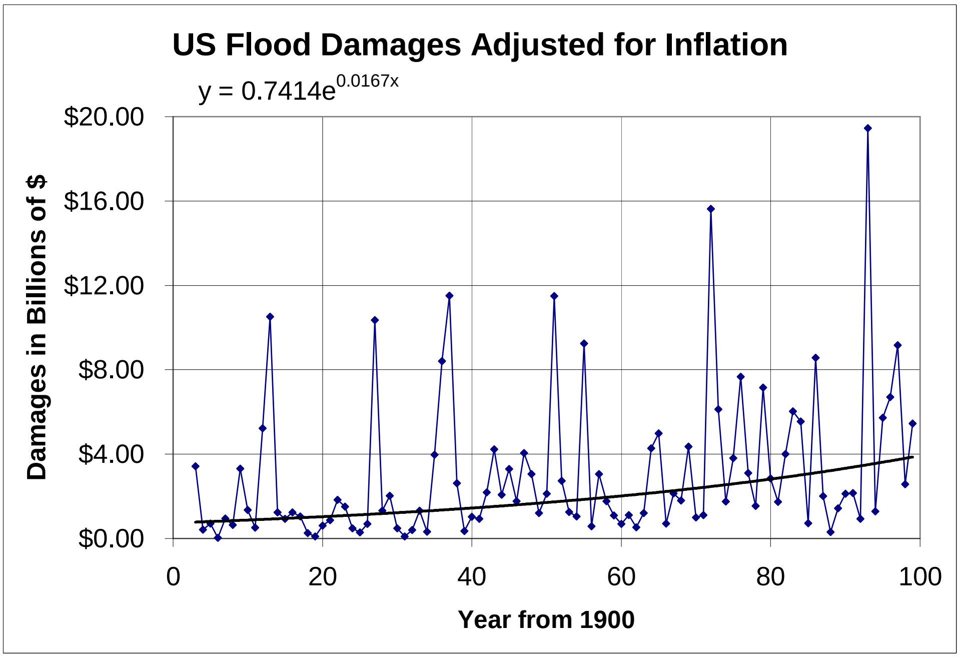


Figure 3: U.S. Flood Damages Data in Real Billions of Dollars with Exponential Regression Model

linear. However, if you have two different exponential models, you can compare them to each other using the correlation coefficients.

Application: Compound Interest

Nearly all fixed rate investments today use compound interest. Compound interest simply means that you are payed interest at fixed intervals, typically every month, and once the interest is paid, it becomes part of the principal that is invested and so earns interest along with the rest of the principal.

For example, suppose way back when you were born, your grandmother placed \$2000 in an account which paid a 6% annual interest rate compounded monthly. Since the 6% is an annual rate, the monthly rate would by $6\%/12 = 0.5\%$, and each month her account would be credited with interest equal to 0.005 times the previous month's balance. The following table shows how the account would grow over the first few months. We have shown some of the expressions unsimplified to illustrate the pattern of how compound interest works.

Month	Beginning Balance	Interest Earned	Ending Balance
1	2000	$0.005 \cdot 2000 = 10$	$2000 + 0.005 \cdot 2000 = 2000(1 + 0.005) = 2010$
2	2010	$0.005 \cdot 2010 = 10.05$	$2010 + 0.005 \cdot 2010 = 2010(1 + 0.005) = 2020.05$
3	2020.05	$0.005 \cdot 2020.05 = 10.10$	$2020.05 + 0.005 \cdot 2020.05 = 2020.05(1 + 0.005) = 2030.15$
4	2030.15	$0.005 \cdot 2030.15 = 10.15$	$2030.15 + 0.005 \cdot 2030.15 = 2030.15(1 + 0.005) = 2040.30$

Table 4:

Notice that we can write the ending balance each month as 1.005 times the beginning balance. If we consider months as the input, increasing the input by 1 results in the output being multiplied by 1.005. This is an exponential function with base 1.005. In fact, using t as the input variable, the function $B(t)$ which gives the balance at the end of t months is given symbolically by

$$B(t) = 2000(1.005)^t$$

which fits the form of our exponential function family with $a = 2000$ and $b = 1.005$.

Knowing the symbolic form for this function, we can find out how much money she will have to give you (if that's what she decides to do) on your 18th birthday. There are $12 \cdot 18 = 216$ months in 18 years. So, the account would be worth $B(216) = 2000(1.005^{216}) \approx \$5873.53$.

Now, what if we wanted to consider another compound interest situation, with an interest rate different than 6% compounded monthly?

Well, we could use the same logic as in our original example for any interest rate. We could even allow for other **compounding periods**, besides monthly compounding, as well as allowing for a change in the amount invested. In our original example, 0.005 was the monthly interest rate, \$2000 was the amount invested, and t denoted the number of months. To generalize to any rate, any amount invested, and any compounding period, we will let i denote the **periodic interest rate**, P the amount invested, and t the number of periods. Using this notation, our general compound interest formula is

$$B(t) = P(1 + i)^t$$

For example, with an account with an *annual* interest rate of 8% compounded quarterly (four times per year), we would have $i = 0.08/4 = 0.02$ and our formula would be

$$B(t) = P(1.02)^t$$

where t would be the number of quarters we leave the money in the account. If we wanted to leave the money in for 18 years, this would be $4 \cdot 18 = 72$ quarters, so we would let $t = 72$.

Example 5. Suppose you invest \$5000 at 6% compounded monthly. How much will you have in 8 years?

Solution: Since you are compounding monthly, your periodic interest rate is $i = 0.06/12 = 0.005$. In eight years, there are $8 \cdot 12 = 96$ months. So your balance will be $B = 5000(1+0.005)^{96} \approx \$8070.71$. ■

Application: Inflation and Real Dollars

You have probably heard your grandparents saying something like, "In my day, I could go to the movies for 50 cents!" However, what they usually fail to tell you is that 50 cents was worth a lot more in their day than it is now. The concept of money changing in value through time is known as inflation. Table 5 shows the **inflation factor** for every 5 years from 1935 to 2000, and then 2001, 2002, and 2003. The inflation factor represents how much it would cost in that year to buy an "average" set of items that cost \$1 in 2003.[88]

Year	1935	1945	1950	1955	1960	1965	1970	1975
Inflation Factor	0.074	0.098	0.131	0.146	0.161	0.171	0.211	0.293
Year	1980	1985	1990	1995	2000	2001	2002	2003
Inflation Factor	0.448	0.585	0.711	0.829	0.936	0.963	0.978	1.000

Table 5: Inflation factors from 1935 to 2003.

So, for example, say it cost \$1 in 2003 to buy an apple, a can of soup, and a pencil. Then, because of inflation, we can see from the table that the same set of items should have cost only 13

[88] Inflation factors from www.infoplease.com.

cents in 1950, 16 cents in 1960, etc. For a set of items that cost several dollars in 2003, to find the cost in any other year N you can multiply the 2003 cost by the inflation factor in year N to get the estimated total cost for the items in that year. So, if you bought a bag of apples, a pair of jeans, and a box of laundry soap for a total of $35 in 2003, the estimated cost for those same items in 1935 would be $0.074 \cdot 35 = \$2.59$. The idea is that, on average, things cost more today than they did years ago. The inflation factor keeps track of this average rise in the price of consumer goods.

Where did these inflation factors come from? Recall in Chapter Three that we introduced the idea of an **index**, and even mentioned the Consumer Price Index as one example. Each month, the government surveys retail outlets all over the country and finds out how much more a "typical basket" of consumer goods costs this year than it did last year. The goods that are selected to be a part of this "basket" are the components of the index, and you can imagine that there might have been some changes in these components over time.[89] The percentage increase in the Consumer Price Index (CPI for short) is the standard measure for inflation. When you hear on the news that the inflation rate was 3.2% last year, this means that the CPI is 3.2% higher this year than it was last year.

The inflation factors are calculated so that the percent increase in the inflation factor from one year to the next is the same as the inflation rate. So for example, if we compare the inflation factors for 2002 and 2001 in Table 5, we have $0.978/0.963 = 1.015576$. This represents a 1.5576% increase, and this increase is the same as the percentage increase in the CPI from 2001 to 2002, which is the inflation rate.[90]

The difference between the CPI and inflation factors is that we pick a particular year to set the inflation factor to 1, representing $1. This is called the **reference** year (sometimes also called the **base** year). We could have picked any year we wanted to be the reference year. All you need to do to compute the inflation factors from the CPI is pick your reference year, and then divide all the CPI values for the years you want by the CPI for the reference year.

In Table 5 above, we picked 2003 as the reference year. In Table 6 we have changed the reference year to 1980. Table 6 shows the CPI values for the given years, and the inflation factors are all calculated by dividing the CPI in the given year by the CPI in 1980, namely 82.4.

Year	1935	1945	1950	1955	1960	1965	1970	1975
Inflation Factor	0.165	0.219	0.292	0.326	0.359	0.382	0.471	0.654
CPI	13.7	18.0	24.1	26.8	29.6	31.5	38.8	53.8
Year	1980	1985	1990	1995	2000	2001	2002	2003
Inflation Factor	1.000	1.306	1.587	1.850	2.089	2.149	2.183	2.232
CPI	82.4	107.6	130.7	152.4	172.2	177.1	179.9	184.0

Table 6: CPI and Inflation factors from 1935 to 2003, with reference year 1980.

Currently, the CPI is calculated so that the average for the years 1982 through 1984 is 100. This is a somewhat arbitrary choice, just like our decision on which year to use as the reference year for inflation factors. The important thing is that the percent increase in the CPI reflect the average increase in prices of all items as accurately as possible. If we wanted, we could use inflation factors with any reference year we wish to keep track of inflation instead of the CPI, since they will all have the same percentage increases from any given year to the next year.

Example 6. Find an exponential model for the inflation factors in Table 6. Does your model seem to fit the data well? What does the percentage growth rate represented by this model tell you?

Solution: Using exponential regression on the data in Table 6 and letting $t = 0$ represent 1935, we get an exponential model of $y = .142(1.0421^t)$. The correlation coefficient is $r = 0.99$. A

[89]in 1900, you might have included supplies for your horse, and today you might include items like DVD's which did not exist even a few years ago.

[90]The rate would probably be reported as 1.6% or 1.56%.

scatterplot of the data along with the exponential model is shown in Figure 4.

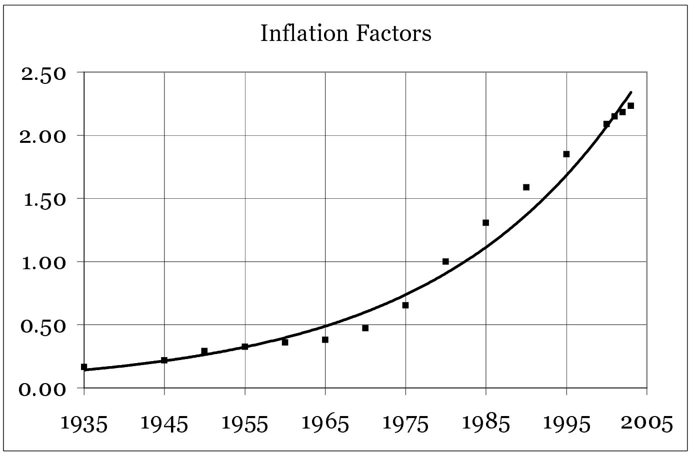


Figure 4: Graph of inflation factors, reference 1980, with exponential model

The model seems to fit the data reasonably well, although some of the errors might be considered significant. The largest error seems to occur in 1985 and to be around 0.15. In fact, the model would predict an inflation factor of $0.142(1.0421^{50}) \approx 1.16$, versus the actual value of 1.306. This means the model would underestimate what you would have needed to spend in 1985 by 15 cents for every \$1.30 spent.

The trend for the last 3 years seems to show a slower increase than does the model, so the model may overestimate the inflation factors for years following 2003.

The percentage growth rate represented by the exponential model is 4.21%. This can be considered the average percent increase in the inflation factors over these years. In other words, the average rate of inflation over this period was 4.21%. ■

Real Dollars versus Current Dollars

The CPI, and the inflation rates that represent the annual percentage increase in the CPI, keep track of how prices increase over time.[91] The flip side of this is that money is worth less over time. If you had stuffed a dollar bill in the sofa in 2000, it would not be worth as much today because it could not buy as much as it could in the year 2000. To keep track of how the value of money changes due to inflation, and to adjust prices and other monetary amounts for inflation, we use what are known as constant dollars.

The terms **current dollars** or **nominal dollars** refer to an amount in actual dollars not adjusted for inflation. For example, suppose that in 1990, \$1 would buy you 5 pieces of candy. You go into the store in 1990, lay down your dollar bill, and leave with 5 candies. The price of one piece of candy in 1990 current dollars is 20 cents. Now, suppose you go back 13 years later, in 2003, and you lay down your dollar bill and they only give you 2 candies. This means the price in 2003 current dollars for one piece is 50 cents. The price of a piece of candy in current dollars, or the nominal price, has gone up from 20 cents to 50 cents in 13 years. If you want to buy 5 candies in 2003, you will have to pay \$2.50.

Now, the 2 pieces of candy you got in 2003 for your dollar would have cost you only 40 cents in 1990. So, at least with respect to its power to purchase candy, your dollar in 2003 is only worth 40 cents, compared to what the dollar was worth in 1990. The **real** value or purchasing power of your dollar has decreased from \$1 to \$.40 over that 13 year period (at least for this type of candy).

[91]It should be noted that prices sometimes *decrease* over time, and this is called deflation. Deflation can be taken into account simply by letting the inflation factors decrease by the appropriate percentage. This in fact did happen during the Great Depression, and at several other points in U.S. history

Constant dollars work the same way, except that instead of looking at just a piece of candy, we use the CPI, because it represents how prices of all consumer items have risen on average over time. As with inflation factors, we pick a reference or base year. Then we find the value of a dollar in any other year in reference to that year.

For example, let's say we pick 1980 to be our reference year. This means we are going to compare the worth of a dollar in any other year to the worth of $1 in 1980. To find the value of a dollar in any year N in **constant dollars** relative 1980, take the CPI in 1980 divided by the CPI for the year N. For example, in Table 6 we see that the CPI for 1980 is 82.4. If we want the value of a dollar in the year $N = 2000$ in real 1980 dollars, we take 82.4 divided by the CPI for 2000, which is 172.2. We get,

$$\frac{82.4}{172.2} = 0.478513357 \approx 0.48$$

or 48 cents. So, a dollar in 2000 can only buy 48 cents worth of stuff compared to what a dollar bought in 1980.

Now, you may notice that to calculate the value of a dollar in real dollars, we are doing the same thing we did for inflation factors except we are switching the numerator and denominator. If R is our reference year, then

$$\text{Inflation Factor for Year } N \text{ with Reference Year } R = \frac{\text{CPI in Year } N}{\text{CPI in Year } R}$$

$$\text{Value of 1 Dollar in Year } N \text{ in Real Dollars with Reference Year } R = \frac{\text{CPI in Year } R}{\text{CPI in Year } N}$$

Here is another way to think about real versus constant dollars. If you had time travel capability, you could use it to make a lot of money by using the changing value of money. Suppose a particular item sells for $1000 in 2000. If you travelled back from the year to 2000 to 1980, you could buy that same item for only $480 (48 cents on the dollar), assuming that the price of that item had increased at the same rate as inflation from 1980 to 2000. If you could bring it back to the year 2000 and sell it, you'd make $520. You made money by spending it in 1980, when your money was worth more.

Adjusting for Inflation

Of course, our hypothetical $1000 item may or may not have increased in price at exactly the rate of inflation. The CPI, after all, is an average based on prices of many items, and some items will go up in price more than others. For example, you may have heard that health care costs (or college tuition) are rising faster than inflation. To **adjust for inflation**, simply take whatever nominal prices you are looking at and divide by the inflation factors. For example, Table 7 shows gasoline prices, based on national average, for a number of years starting in 1976. The nominal prices are shown in the second column, and inflation factors referenced to 1980 are in the third column. The last column is the average price of gasoline in constant 1980 dollars. These values are found by taking the column two nominal prices divided by the column three inflation factors.

Notice that even though the nominal price in 2001 is more than twice the nominal price in 1976, the real price, adjusted for inflation, is lower. We would say that gasoline prices have gone up more slowly than inflation from 1976 to 2001. Another way to say this is that gasoline prices have fallen **in real terms**.[92] The chart in Figure 5 shows both the nominal and the real prices for gasoline for all the years from 1976 through 2001. We have actually given the real prices two different ways, once in 1980 dollars and once in 1996 dollars. Notice that the shapes of the graphs of the real prices are exactly the same! The only difference that changing the reference makes is where the plot of real prices crosses the plot of nominal prices. For the plot of real prices in 1980

[92] This is the same use of the word "real" made by Rush Limbaugh as quoted in Chapter Four

Year	Nominal Price	Inflation Factors	Price in Real $ (1980)
1976	0.61	0.691	0.88
1978	0.67	0.791	0.85
1980	1.25	1.000	1.25
1985	1.20	1.306	0.92
1990	1.16	1.586	0.73
1995	1.15	1.850	0.62
1997	1.23	1.948	0.63
1998	1.06	1.978	0.54
2000	1.51	2.090	0.72
2001	1.46	2.149	0.68

Table 7: Nominal and Real Gasoline Prices

dollars, the graph will meet the nominal graph in 1980, since that is where the nominal price equals the real price in 1980 dollars. Similarly, the nominal price will equal the real price in 1996 dollars in the year 1996, and that is where these two plot intersect.

Wherever the graph of real prices is increasing, the nominal price is going up faster than inflation. Wherever the real price is decreasing, the nominal price is going up more slowly than inflation. Since the real prices in 2001 are lower than the real price in 1976, as noted above, nominal gasoline prices have gone up more slowly than inflation overall during that period.

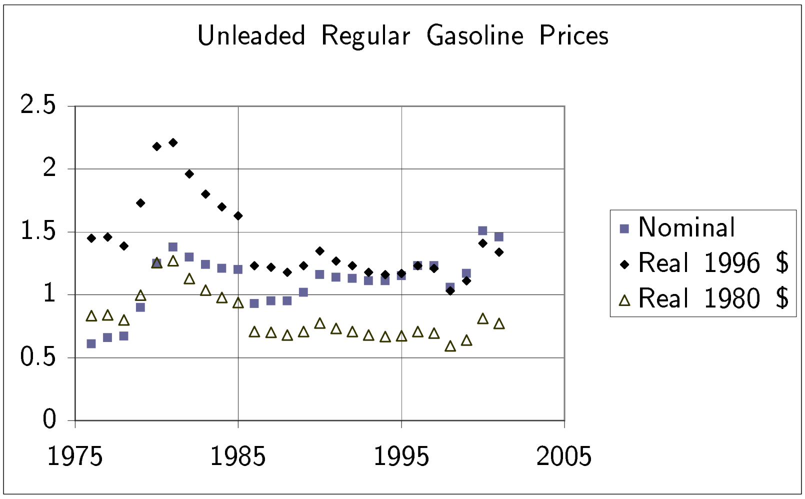


Figure 5: Nominal and Real Gasoline Prices

Based on the fact that real gasoline prices have been going down (and they were going down even more steeply before the 1970's), many politicians and others have said that complaints about expensive gasoline are unwarranted. Americans actually spend much less for gasoline than almost any other people in the world, in real terms. If Americans in 2001 paid, in real terms, the same amount for gasoline as in 1976, the nominal price of gasoline would have been $1.90 per gallon.

Conclusion

The above example illustrates one example of the use of exponential functions, as well as the relationship between exponential functions and *any* quantity that grows at a fixed percentage growth rate. Money in a savings account, the cost of college tuition, the size of a population, and the decay of a radioactive substance are all examples of situations which typically increase (or decrease) by a fixed percentage. All of these can be modelled by exponential functions, just as quantities which grow at a constant rate can be modelled by linear functions.

Note that our base b, which is also referred to as the growth factor, for our exponential function modelling grandmother's money was 1.005, while the monthly interest rate was 0.005. The 0.005 is what we will call the **percentage growth rate** and it will be true in general that

$$\text{The Growth Factor} = 1 + \text{The Percentage Growth Rate}$$

This will be true regardless of what the percentage growth rate is, even if the percentage growth rate is negative (in which case we might call it the percentage rate of decay).

Reading Questions for Exponential Functions

1. For each of the following functions given in symbolic form, decide whether the functions is linear, exponential, power, or none of these.

(a) $y = 13(10^{0.015x})$

(b) $y = 3(\sqrt[5]{x^2})$

(c) $y = 2^{x^2}$

(d) $y = 4(x - 7) + 3$

(e) $y = 6.23(1.016^x)$

(f) $y = \frac{23}{x^{0.75}}$

2. How is a linear function similar to an exponential function? How is a linear function different from an exponential function?

3. Each of the following is an exponential function. In each case, give the growth factor, the y-intercept, and the percentage growth rate.

(a) $f(x) = 4^x$

(b) $f(x) = 3(\frac{1}{3})^x$

(c) $f(0) = 5,\ f(1) = 15$

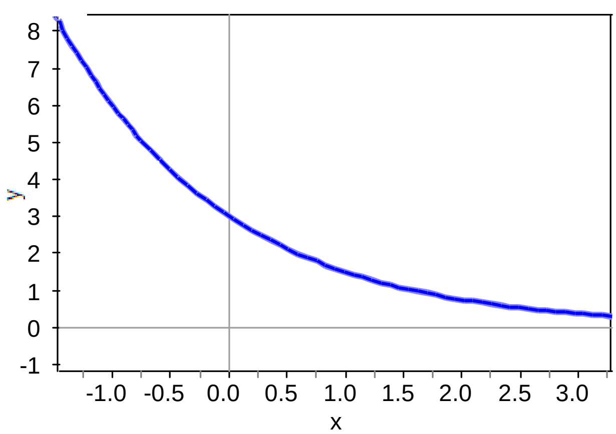
(d) $f(0) = 1,\ f(2) = 36$

4. Let $f(x) = ab^t$ be an exponential function which models the value of a particular car over time as it ages. What do you know about the value of b? Justify your answer.

5. Give the equation for the exponential function whose y-intercept is 3.2 and whose growth factor is 1.4.

6. Let $f(x) = ab^x$ be the exponential function whose graph is given below.

?



(a) What is the value of a? How do you know?

(b) Is the value of b bigger or less than 1? How do you know?

7. Let $h(x) = 5(4)^x$. Find each of the following.

(a) $h(-3)$.
(b) $h(-1)$
(c) $h(0)$
(d) $h(8)$

8. Determine if each of the following is an exponential function. If so, give its equation. If not, briefly explain why.

(a) $2y = 4^x$

(b)

x	-2	-1	0	1	2
$f(x)$	4	2	1	0.5	0.25

(c) The function describing the amount of money received by returning pop bottles if the deposit is $0.10 per bottle.

9. Determine if each of the tables given below describes an exponential function. If so, give the equation for the function. If not, briefly explain why.

(a)

x	-2	-1	0	1	2
$f(x)$	1	3	9	27	81

(b)

x	-2	-1	0	1	2
$f(x)$	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	1	$\sqrt{2}$	2

(c)

x	-2	-1	0	1	2
$f(x)$	0.08	0.4	0	0.4	0.08

(d)

x	-2	-1	0	1	2
$f(x)$	0.309	0.555	1	1.8	3.24

(e)

x	0	2	4	6	8
$f(x)$	3	12	48	192	768

10. One of the following involves an exponential function and one involves a linear function. Identify which is which.

(a) Find the cost of gasoline if the price is $1.23 per gallon.
(b) Find the population of a city that is increasing at a rate of 3.5% per year.

11. Let $f(x) = x^b$. Explain why it is always true that the graph passes through the point $(1, 1)$.

12. Evaluate each of the following:

(a) $16^{1/2}$
(b) 4^7
(c) $125^{4/3}$
(d) $10^{3/4}$
(e) $e^{1.7}$
(f) $4e^{1.7}$

13. Write each of the following using radicals (i.e. write $x^{1/2}$ as $\sqrt{x}$).

(a) $3x^{9/5}$

(b) $x^{4/3}$

(c) $6x^{54/5}$

(d) $8x^{7/4}$

14. Suppose you have $5000, and you want to save up to buy a new pickup truck to replace your aging 1998 Ford F-150. You estimate you can get at least 4 more years out of your current vehicle, and so tentatively plan to invest the money for 4 years and use the balance as a down payment on a new truck.

(a) Assume you can invest your money at fixed annual rate of 6% compounded quarterly. How much would you have at the end of 4 years?

(b) Instead of investing your money at fixed annual rate of 6% compounded quarterly, you decide to take a somewhat riskier approach and invest in a variable rate account. The upside is you can currently get an 8% annual rate in such an account. The downside is the interest rate may go down, possibly significantly, before the 4 years is up.

i. If you are lucky, the interest rate will stay at 8%, compounded quarterly. How much would you have in the account after 4 years at this interest rate?

ii. If you are not lucky, the interest rate will go down. To keep things simple, suppose the account stays at 8% for the entire first year, goes to 6% during the second year, 5% during the third, and 4% during the fourth.[93] How much will you have after 4 years under this scenario?

15. The table below gives child mortality data for Libya, Cuba, and the Bahamas.

Country	1990	1995	2000	2001
Bahamas	29	23	17	16
Cuba	13	10	9	9
Mexico	46	36	30	29

(a) Find the exponential regression equation for the Bahamas's child mortality rate. It might be convenient to let 0 represent the year 1990.

(b) Find the exponential regression equation for Cuba.

(c) Find the exponential regression equation for Mexico.

(d) Using your exponential regression equations, predict child mortality for each of these countries for they years 2010 and 2050.

[93] In reality, interest rates in such accounts may vary from quarter to quarter.

16. The table below is similar to Table 5 in the text, except that (nominal) median household incomes for the U.S. are given instead of the nominal price of gasoline. Fill in the household incomes in real dollars in the last column. Have median incomes increased faster than inflation or not?

Year	Nominal Salary	Inflation Factors	Income in Real $ (1980)
1980	21,023	1.000	
1985	27,735	1.306	
1990	35,353	1.586	
1995	40,611	1.850	
1997	44,568	1.948	

Table 8: Nominal Median Incomes and Inflation Factors, U.S.

17. Explain briefly the difference between real dollars and constant dollars.

18. Suppose the nominal price of a certain item increases at exactly the same rate as inflation from year to year. When the price is adjusted for inflation, giving the price in real dollars, what will be true about the price in real dollars from year to year?

Exponential Functions: Activities and Class Exercises

1. **Moore's Law.**[94]

   In 1965, Gordon Moore, cofounder of Intel Corporation, predicted that computing power would double roughly every 2 years or so.[95] This prediction has come to be known as "Moore's Law."

   The data for Moore's Law can be found in the Fathom file **Moore's Law Data**. Upon opening the file, you should see a case table, two sliders, and a graph with a scatterplot of the data and an exponential model $y = A \cdot B^x$. The dependent variable is the number of transistors in an average CPU, a measure of computing power. The independent variable is the year from 1970. The values of A and B are given by the sliders.

   (a) Play with the sliders until the exponential model gives a reasonable approximation to the data. What values of A and B did you use to give the best fit?

   (b) Do these values seem to confirm Moore's law? Explain carefully why or why not based on the value of the growth factor B. You might want to find the value of B which would result in a doubling every two years.

   (c) As we will see in Chapter 8, a two-variable data set Y versus X will follow an exponential model if $Log(Y)$ versus X follows a linear pattern. Add a new attribute *Log Transistors* by calculating the logarithm of the *Transistors* variable as a formula.

      i. Make a graph of *Log Transistors* versus *Since1970*. Is this a fairly linear graph?

      ii. Find the linear regression $y = mx + b$ equation and the correlation coefficient for *Log Transistors* versus *Since1970*.

      iii. As we will also see in Chapter 8, the **exponential regression model** for a data set Y versus X has parameters A and B with $A = 10^b$ and $B = 10^m$, where m and b are the slope and y-intercept of the linear regression equation for $Log(Y)$ versus X. Find these values of A and B for your regression equation.

      iv. Use the sliders to change the values of A and B to those you found in the previous part. How does this model fit the data? Does it seem to fit better than your original model?

      v. Finally, does this model seem to confirm Moore's Law? Why or why not?

2. **Stock Market Growth.** The following table gives the Dow Jones Industrial Average (DJIA) on the last day of January for each year from 1970 through 2004.[96]

Year	DJIA	Year	DJIA	Year	DJIA	Year	DJIA
1970	800.4	1979	805	1988	1938.8	1997	6448.3
1971	838.9	1980	838.7	1989	2168.6	1998	7908.3
1972	890.2	1981	964	1990	2753.2	1999	9181.4
1973	1020	1982	875	1991	2633.7	2000	11497.1
1974	850.9	1983	1046.5	1992	3168.8	2001	10786.85
1975	616.2	1984	1258.6	1993	3301.1	2002	10021.56
1976	852.4	1985	1211.6	1994	3754.1	2003	8341.63
1977	1004.7	1986	1546.7	1995	3834.4	2004	10453.92
1978	831.2	1987	1896	1996	5117.1		

[94] This activity is adapted from *Fathom Dynamic Data Software: Workshop Guide*, by Finzer and Erickson, p. 54

[95] see Wikipedia at http://en.wikipedia.org/wiki/Moore's_law.

[96] Data from www.djindexes.com

(a) Graph a scatterplot of this data, letting $x = 0$ represent 1970. Which family of functions, linear or exponential, do you think is most appropriate to model this data? Explain.

(b) Find the exponential regression equation for this data.

(c) Graph your exponential regression equation along with the data. How well does the equation seem to fit the data?

(d) What percentage growth rate does your exponential equation represent? This percentage can be considered the (effective) average annual growth rate in the DJIA.

(e) One type of investment that has become popular in recent years is an **index fund**. An index fund is a mutual fund which invests in the same stocks that are components of one of the stock indexes. The idea is that the index funds performance should then mirror that of the index. If you bought a Dow Jones based index fund, your investment should grow (and fall!!) along with the DJIA.

i. Using your average percentage growth rate from part (d), estimate what an investment of $10,000 made in 1980 would be worth today.

ii. Assuming this average growth rate continues into the future, how much would $10,000 invested today in a Dow Jones Index Fund be worth 40 years from now?

(f) From the scatterplot, you can see that the DJIA has a fair amount of **volatility**. In other words, although the general trend is upwards, there is a lot of up and down in the graph. So, the average percentage growth rate is not a very good predictor of the growth rate in any particular year.

i. Create a new table which gives the actual percentage growth rate (or decline) in the Dow for each year. You can do this by dividing the value in one year by the value the previous year, and then subtracting one.

ii. What is the largest percentage increase seen in these years?

iii. What is the largest percentage drop? How many years out of these 35 years did the Dow experience a drop?

iv. Find the mean and the median of the percentage increases. How close are these to the percentage growth rate represented by the exponential growth equation? What do the mean and the median tell you about the skew of the distribution of the percentage growth and decline rates?

v. (Optional bonus for those who follow the business news). Starting in 1996, the DJIA experienced a large spike, peaking in 2000, and then declining until 2003. Can you think of any explanation for this spike?

3. **Comparing Car Models II.** The data in the table below represent 38 car types of 7 different makes and 19 different models and variations for the 2005 model year. The data are from www.autos.msn.com and also appear in one of the activities in Chapter Five. As noted there, City Mileage, Highway Mileage, and Greenhouse Gas Emissions ratings are as determined by the EPA. Gas emissions is given in tons of carbon dioxide per year under normal driving conditions. Engine Size is in liters. Passenger Room and Luggage Room are given in cubic feet.

Make	Model	Price	City mpg	H'way mpg	Carbon Emiss.	Engine Size	Passenger Room	Lugg. Room
Honda	Civic	15325	29	38	5.9	1.7	91	10
Honda	Accord (2dr)	22622.5	24	34	7	2.4	103	14
Honda	Accord (4dr)	22622.5	24	34	7	2.4	103	14
Toyota	Avalon	23500	22	31	7.6	2.2	107	14
Honda	Accord (2dr)	22622.5	21	30	7.8	3	103	14
Honda	Accord (4dr)	22622.5	21	30	7.8	3	103	14
Toyota	Corolla	15667.5	30	38	5.7	1.8	84	14
Honda	Civic	15325	29	38	5.9	1.7	91	10
Ford	Crown Victoria	27475	18	25	9.2	4.6	111	21
Toyota	Matrix 4x2	16855	28	34	6.2	1.8	94	22
Toyota	Matrix 4x4	16855	28	34	6.2	1.8	94	22
Toyota	Celica	18600	29	36	6	1.8	78	17
Toyota	Camry	21875	24	34	6.9	2.4	102	17
Chevrolet	Cobalt (4dr)	17527.5	24	32	6.7	2.2	87	14
Ford	Taurus SW	13875	19	25	8.8	3	109	39
Toyota	Camry	21875	21	29	8	3.3	102	17
Ford	Taurus	22165	20	27	8.4	3	105	17
Toyota	Camry	21875	20	28	8.2	3	102	17
Chevrolet	Monte Carlo	25317.5	21	32	7.6	3.4	96	16
Toyota	Corolla	15667.5	30	38	5.7	1.8	84	14
Chevrolet	Aveo	11160	26	34	6.5	1.6	91	12
Saturn	Ion (2dr)	11900	24	32	6.7	2.2	91	15
Saturn	Ion (4dr)	11900	24	32	6.7	2.2	91	15
Ford	Focus	15842.5	26	32	6.8	2	94	15
Chevrolet	Malibu Classic	21630	25	34	6.7	2.2	99	16
Toyota	Camry	21875	24	34	6.9	2.4	102	17
Honda	Accord (4r)	22622.5	24	34	7	2.4	103	14
Chevrolet	Cobalt (4dr)	17527.5	24	32	6.7	2.2	87	14
Dodge	Stratus (4dr)	21845	22	30	7.6	2.4	94	16
Toyota	Camry	21875	21	29	8	3.3	102	17
Honda	Accord (4r)	22622.5	21	30	7.8	3	103	14
Dodge	Stratus (4dr)	21845	21	28	8.1	2.7	94	16
Toyota	Camry	21875	20	28	8.2	3	102	17
Dodge	Stratus (2dr)	21845	21	28	8	2.4	86	16
Dodge	Stratus (2dr)	21845	20	28	8.3	3	86	16
Dodge	Neon	13750	25	32	6.8	2	90	13
Chevrolet	Cavalier (2dr)	14017.5	24	34	6.7	2.2	92	14
Chevrolet	Cavalier (4dr)	14017.5	24	34	6.7	2.2	92	14

(a) Make a scatterplot of greenhouse gas emissions versus city mileage, and graph the linear regression line along with the scatterplot.

i. Notice that data points at the left and right end of the scatterplot tend to be above the line and those in the middle tend to be below the line. This suggests that a decreasing exponential function might provide a better fit to the data than a linear function. Find the exponential regression equation for this data and graph it along with the scatterplot.

ii. Which do you think provides a better fit, the linear regression or the exponential regression equation. Recall that, as discussed in the text, it is *not valid* to use the correlation coefficients to answer this question.
iii. Predict the greenhouse gas emissions for a car that gets 40 miles per gallon in the city using both the linear and the exponential models. How much difference is there in these predictions?
iv. Repeat the previous part for a car that gets 10 miles per gallon in the city.

(b) Make a scatterplot of greenhouse gas emissions versus average price in thousands.
i. Looking at the scatterplot, which family of functions do you think is more likely to produce a better fit to these data, linear or exponential?
ii. Find the exponential regression model for these data.
iii. What is the growth factor and percentage growth rate for your exponential model? What does the percentage growth rate mean in terms of emissions and price?
iv. What does your exponential model predict for a car that costs $50,000?

4. **Airline Deregulation.** The graph below is from a report by the Heritage Foundation on airline deregulation. Congress passed deregulation legislation in 1978, cutting back drastically on price, route, and other restrictions that airline companies had been required to follow before then.

The foundation supported deregulation of the airline industry, and the report tries to make the case that, based on 20 years of experience, deregulation has been good both for the industry as a whole and for consumers. The foundation cited this chart to support the claim that, after deregulation, the rate of increase in miles flown has been greater than the rate of increase prior to deregulation. This claim was meant to respond to critics of deregulation who have claimed that deregulation led to fewer options and a lower rate of service for consumers in many areas of the country. Here is a short excerpt from the report.[97]

> Airlines also are logging more miles than before deregulation. Whereas carriers flew roughly 2.5 billion miles in 1978, they logged more than double that number last year alone, flying approximately 5.7 billion miles in 1997. Furthermore, as Chart 5 illustrates, this trend increased much more rapidly after deregulation than it did before market liberalization.

(a) The bars represent the actual number of miles flown by U.S. domestic air carriers. What function family do you think would be most appropriate to model this data?

(b) The black line, up to the year 1978, is meant to represent what is called the three-year moving average. This is typically calculated by taking the values of three consecutive years, including the year in question, and averaging them. What seems to be wrong with the three-year averages as plotted here?

(c) In the excerpt from the report given above, it was noted that the trend of miles flown increased much more rapidly after deregulation than before. The implication is that more miles were flown after deregulation than would have been flown if deregulation had not occurred.
i. What function family is being used in creating the extrapolated trend line?
ii. In order to create the extrapolated line, the author probably calculated the linear trend line for years up to 1978. Why do you think the author would do this? What is wrong with using a linear model in this situation?

[97] The full report can be found at http://www.heritage.org/Research/Regulation/BG1173.cfm. The author is Adam D. Thierer, former Alex C. Walker Fellow in Economic Policy at The Heritage Foundation, and now at the Cato Institute. Mr. Thierer has graciously allowed the use of his report for this activity.

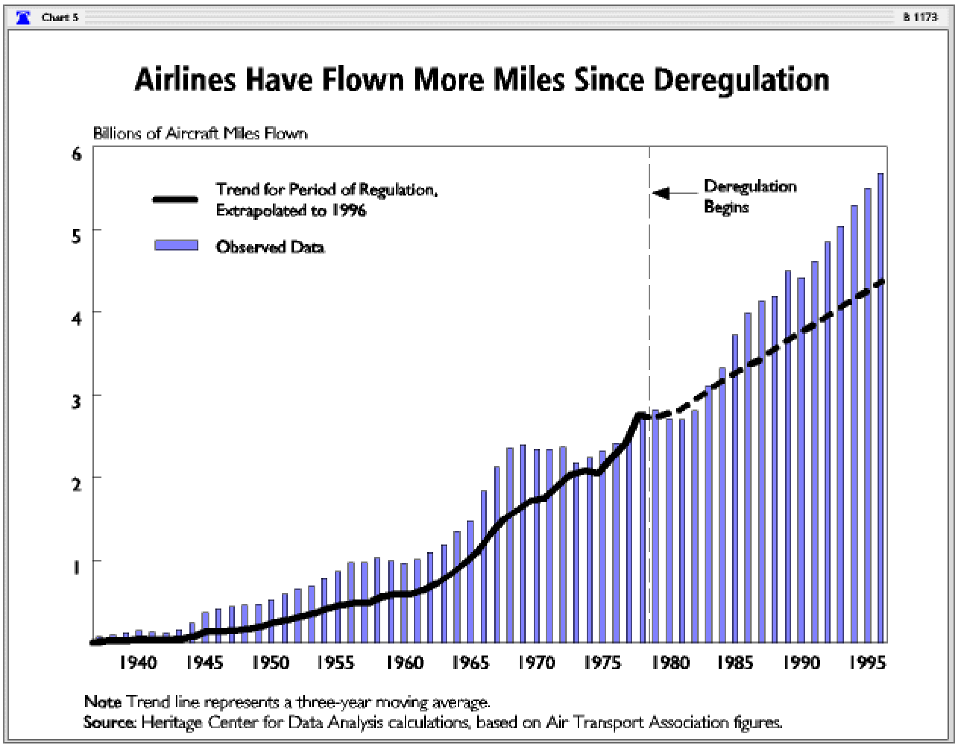


Figure 6: Billions of Aircraft Miles

iii. The actual data on which the graph is based were given later in the report. The number of miles flown in 1950 was about .48 billion, and the number flown in 1970 was about 2.39 billion. Letting $t = 0$ represent 1950, use these two points to create an exponential model $M = a(b^t)$ where M represents estimated number of miles flown and t represents time in years from 1950.

iv. Use your exponential model to estimate the number of miles flown in 1980, 1990, and 1996. How do these values compare with those indicated by the graph?[98]

v. One might argue that the years 1950 and 1970 used to create the previous model would lead to an overestimate, since the graph has a peak near 1970. Picking a lower point from a later year would give a different model with lower predictions for future years. In the year 1974, only 2.18 billion miles were flown. Use this point together with 1950 to create an exponential model, and then use this model to predict the number of miles flown in 1980, 1990, and 1996.

vi. What are the percentage increases in numbers of miles flown represented by your two exponential models?

(d) Comparing your two exponential models to the actual data, and using the other information you have discovered, what would you say about the Heritage Foundation's claim that the rate of increase in the number of miles flown was higher after deregulation than it was before deregulation?

5. Inflation and investing. We saw in the text that one effect of inflation is that money is worth less over time. This includes not only cash in your wallet, but also other monetary amounts like the value of your stock portfolio.

The table in activity 2 gives data on the Dow Jones Industrial Average (DJIA) over time, starting in 1970. Although the DJIA is not measured in dollars, we could consider the

[98]The actual values in these years were 2.82, 4.49, and 5.5 billion miles

amounts as representing the value of an investment as it grows over time. Thus, we could say that an investment valued at $800.40 in 1970 would grow to $10453.92 in 2004 if it grew at the same rate as the DJIA.

However, these values would be in nominal dollars. Inflation would make the value of the investment grow more slowly in real terms.

(a) Using Table 6 in the text, which gives the CPI and inflation factors every five years starting in 1935, create a table showing the value of an investment that mimics the DJIA, but adjusted for inflation to real 1970 dollars. Since you only have inflation factors for every fifth year, you need only adjust the DJIA values for those same years.

(b) Using your table from the previous part of this activity, find an exponential model for the DJIA adjusted for inflation, starting in 1970.

   i. What is the percentage growth rate of the DJIA in real terms?

   ii. If you invested $1000 today and it grew exponentially according to this percentage growth rate, what would it be worth in 40 years in real dollars, indexed to today?

6. Population Growth. Each of you should collect data on the population of a particular country. You should find yearly data going back to at least 1970. For your country, answer the following questions.

(a) Find the exponential model that best fits your data. Then discuss how well the exponential model fits the data. Are there periods of time for which the model fits better or worse than at other times?

(b) What is the average percentage growth rate of the population over the period you've modeled? (This is the percentage growth rate represented by your exponential model)

(c) Use your exponential model to predict the population for your country 5 years from now and 10 years from now.

(d) Find each individual year's percentage growth rate. To find the percentage growth rate of the population in a given year N, take the year N population and divide it into the population for the next year N+1, and then subtract 1. (You can do this for all years at once on a calculator or on a spreadsheet).

(e) Graph the percentage growth rates you found in part (d) over time. How are they changing over time? How have they been changing over the last 10 years? Select a family of functions to model the percentage growth rate over the last several years (anywhere from 5 to 20, depending on how "nice" the data seems to be), and find the member of this family which best fits the data. Discuss how well you think your function fits the data.

(f) Use the function you found in part (e) to project what the percentage growth rate should be for each of the next 10 years. Give the results in a table.

(g) Now, use the results from part (e) to make new predictions for the population of your country 5 years from now and 10 years from now. Compare these to the predictions you made in part (c) Discuss the differences between the two sets of predictions. Also decide which set of predictions you have more confidence in and why.

7. Logarithmic Functions

So far, we have looked in detail at linear and exponential functions. While different in many ways, these functions are similar in that, for the most part, they only involve the basic arithmetic operations of addition and multiplication.[99] Logarithmic functions, while they "look" different and do not explicitly involve basic arithmetic operations, are actually closely related to exponential functions. In fact, the most succinct way to describe logarithms is that:

Logarithms are exponents.

Let's see exactly what this means.

Definitions

Any exponential function $y = ab^x$ involves two parameters. The parameter b is called the base, and the exponent is the independent variable. Logarithmic functions will also involve a parameter b called the base, and an exponent as one of the variables, but in this case the exponent will be the *dependent variable*. For any base b, we define the **logarithm base** b as follows.

$$y = \log_b x \qquad \text{means that} \qquad b^y = x$$

In other words, $\log_b x$ represents the exponent y such that $b^y = x$. This is what we mean when we say "logarithms are exponents." The first equation above is called the **logarithmic form** and the second **exponential form**.

Example 1. Using the definition above, find the following logarithms.

1. $\log_3 9$
2. $\log_3 1$
3. $\log_2 8$
4. $\log_2(1/4)$

Solution:

1. $\log_3 9$ is the exponent y such that $3^y = 9$. Since $3^y = 9 = 3^2$ and the function $f(x) = 3^x$ is a one-to-one function, we must have that $y = 2$. Therefore, we have $\log_3 9 = 2$.
2. We are looking for the exponent y such that $3^y = 1$. Since any nonzero number to the 0 power is 1, we have $\log_3 1 = 0$. In fact, for any positive base b, $\log_b 1 = 0$, since $b^0 = 1$.
3. Since $2^3 = 8$, $\log_2 8 = 3$.
4. We are looking for the exponent y so that $2^y = \frac{1}{4}$. Recalling that negative exponents give reciprocals, we can deduce that -2 is the correct exponent here. $2^{-2} = 1/2^2 = 1/4$. So, $\log_2(1/4) = -2$.

■

Now, what if someone asks you to calculate $\log_2 10$? They want to know the exponent y such that $2^y = 10$. This is not as easy as the examples above! We know that $2^3 = 8$ and $2^4 = 16$, so we can deduce that this exponent is somewhere between 3 and 4. We could start guessing at numbers

[99] Exponents are really just repeated multiplication. To be fair, exponents that are not integers are a bit more complicated, but in this course, we will not consider the theoretical mathematical underpinnings of non-integral exponents.

for y, and then calculating 2^y until we find one that gets us close to 10, but this does not sound like much fun. Fortunately, technology will come to our aid.

Your calculator very likely has *two* buttons to calculate logarithms. The one which we will typically use is the **LOG** button. The **LOG** button calculates logarithms with base 10. For example, if you press **LOG 100** your calculator should give you a result of 2, which makes sense since $10^2 = 100$. If you press **LOG 35** you should get 1.54407, or something close to this, depending on how many digits your calculator displays. If you check, you will see that $10^{1.54407}$ gives you an answer pretty close to 35. It will not be exact, since we did some rounding off. If you see a log with no base given, this will typically mean the logarithm base 10.

The second logarithm button on your calculator is probably labelled **LN**. This calculates logarithms with base e, where $e \approx 2.718281828$. The logarithm base e is called the natural logarithm, and is used extensively in scientific fields. The appendix contains more background on the natural logarithm, and where the number e comes from.

What about our problem with $\log_2 10$? We can use the following property of logarithms to calculate this.[100]

$$\log_b x = \frac{\log x}{\log b}$$

Let's apply this property to our example. We get,

$$\log_2 10 = \frac{\log 10}{\log 2} \approx \frac{1}{0.30103} \approx 3.322$$

To check, calculate $2^{3.322}$ on your calculator.

The Graph of the Logarithm Function

Although there is a whole family of logarithmic functions $y = \log_b x$, one for each b, we will predominantly use $y = \log x$, the base 10 logarithm, or $\ln x$ the logarithm base e.[101] Because logarithms are exponents, we will be able to relate the characteristics of the function $y = \log x$ to the function $y = 10^x$. Table 1 shows some values for both of these functions. [102]

Table 1: The logarithm function and the exponential function base 10.

x	0.01	0.1	.5	1	2	5	10	20	30	40	50	100
$\log x$	-2	-1	-0.30	0	0.30	0.70	1	1.30	1.477	1.602	1.699	2
x	-2	-1	-0.30	0	0.30	0.70	1	1.30	1.477	1.602	1.699	2
10^x	0.01	0.1	.5	1	2	5	10	20	30	40	50	100

Notice that the tables are exactly the same except that the dependent and independent variables are reversed!! In other words, if a point (u, v) is in the table (and thus on the graph) for $y = \log x$, then the point (v, u) is in the table (and thus on the graph) for $y = b^x$. The effect of this is that the graph of $y = \log x$ will have the same shape as the graph of $y = 10^x$, but with the x and y axes switched. In other words, if you have the graph of $y = 10^x$ on a transparency, flipping the transparency over diagonally so that it is now face down and with the x-axis running vertically will give you a picture of the graph of $y = \log x$.

Two versions of the graph are shown in Figure 1. On the left, the window is $0 \leq x \leq 110$, $-1 \leq y \leq 2.5$. In this window, we see the very slow growth of the logarithmic function as x increases, and also the vertical asymptote at $x = 0$. The version on the right is an extreme close-up

[100] Proofs for this and the other properties of logarithms we give later are found in the Appendices.
[101] As you can see in the appendices, b must be positive and not equal to 1.
[102] We have rounded many of the values, but they are all accurate to within a hundredth.

of a section of the graph very close to the y-axis (notice the horizontal scale). This section of the graph would be indistinguishable from the y-axis in any window including x values 2 or larger, but here we see that the graph is indeed not touching the y-axis.

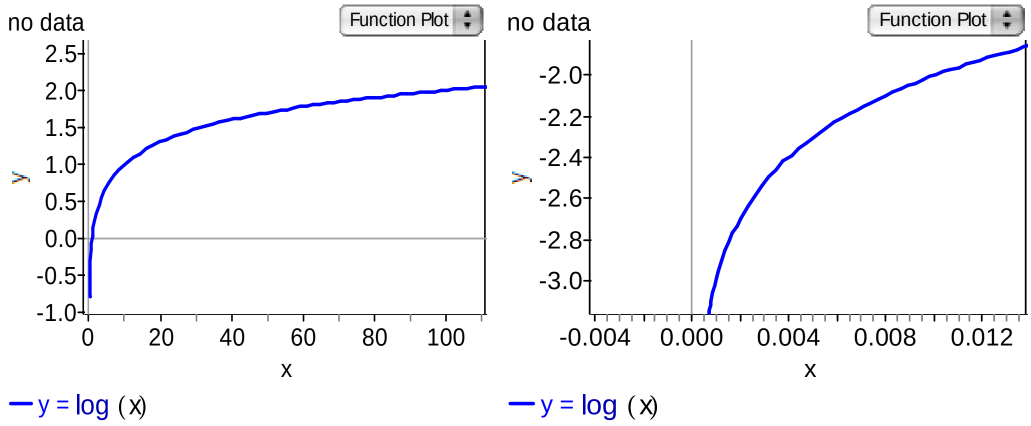


Figure 1: Two views of the logarithmic function base 10.

Since the graph of the logarithmic function $y = \log x$ has the same shape as the exponential function $y = 10^x$ except that it is "flipped diagonally," we have the following comparisons between the two graphs, and these hold for logarithmic and exponential functions of bases b other than 10 as long as $b > 1$.

Exponential Functions	Logarithmic Functions
Always concave up	Always concave down
y-values always positive	x-values always positive
y-intercept equal to 1	x-intercept equal to 1
Horizontal asymptote $y = 0$ as x goes to negative infinity	Vertical asymptote $x = 0$ as y goes to negative infinity
Graph is very steep for large x	Graph is very flat for large x.

Finally, recall that for the exponential function with base b, the result of increasing the input by 1 is that the output gets multiplied by b. For $y = \log x$, this means that if we multiply the input by 10 (the base b), the output will increase by 1. You can see this in the table, for example, by comparing the outputs for $\log x$ for the inputs 5 and 50, or for the inputs 2 and 20, or for the inputs 0.1, 1, 10, and 100. We could write this property in equation form as

$$\log 10x = 1 + \log x.$$

Properties of Logarithms

To summarize what we know so far about logarithms, we have the following properties of the logarithm function.

- $\log 10^x = x$.
- $\log(1) = 0$.
- $\log x$ is negative for $0 < x < 1$ and positive for $x > 1$.
- $y = \log x$ has a domain of $x > 0$ and a range of all real numbers.

- The graph of $y = \log x$ is increasing and concave down.
- $\log_b x = \frac{\log x}{\log b}$ for any positive $b \neq 1$.

There are four more properties that we will also find useful. These are:

1. $\log(uv) = \log u + \log v$.

2. $\log(u/v) = \log u - \log v$.

3. $\log x^n = n \log x$.

4. $10^{\log x} = x$.

The first property is a generalization of the equation $\log 10x = 1 + \log x$ given above. Note that $\log 10 = 1$. So,

$$\log 10x = \log 10 + \log x = 1 + \log x$$

The more general property 1 says we can substitute any other number u for the 10 in $\log 10x$ by using $\log u$ instead of $\log 10$. Why does this property work in general?

Let's illustrate with an example. Let a and b be numbers such that $\log(a) = 3$ and $\log(b) = 4$. This means $a = 10^3$ and $b = 10^4$. Notice that $\log(a) + \log(b) = 3 + 4 = 7$. What input would give an output of 7 under the log function? It would be 10^7, the *product* of the original two inputs 10^3 and 10^4, since $\log 10^7 = 7$. Since $ab = 10^7$, we get

$$\log ab = \log 10^7 = 7 = 3 + 4 = \log a + \log b$$

In general, property 1 is a consequence of the fact that, whenever you multiply two numbers with exponents that have the same base, you add the exponents. Symbolically, if $\log a = p$ and $\log b = q$, then $10^p = a$ and $10^q = b$. So $ab = 10^p 10^q = 10^{p+q}$ (we're adding the exponents). This means $\log(ab) = p + q = \log(a) + \log(b)$.

Discussion of why properties 2 and 3 work are given in the appendices.

Let's consider property 4. Suppose we start with $10^{\log x} = y$. If we change this equation to logarithmic form, we get

$$\log_{10} y = \log x$$

or just $\log x = \log y$. Since the logarithm function is increasing, the only way this can happen is if $x = y$. Combining this with our original equation $10^{\log x} = y$, we get $10^{\log x} = x$.

In interpreting these properties, it is important to keep in mind how logarithms fit into the usual order of operations. In general, unless there are parentheses involved, logarithms are done after any multiplications, exponents, or divisions that come immediately after the log sign, but before any additions or subtractions and before any multiplications immediately preceding the log sign. Refer to the appendices as needed for more examples and discussions on this and other properties of logarithms.

Why are these properties important? One reason is that knowing these properties gives you a better understanding of how logarithms work. A second is that they will be useful at times in solving problems that involve either logarithms *or* exponential functions.

The Natural Logarithm

So far, we have concentrated on the logarithm base 10. The other commonly used logarithm is the logarithm base e, also referred to as the **natural logarithm**, and denoted in symbols as ln. Recall e is an irrational number approximately equal to 2.71828. By definition

$y = \ln x$ means that $e^y = x$.

Example 2. Using the definition above, find the following logarithms.

1. $\ln e^2$

2. $\ln 1$

3. $\ln(1/e)$

4. $\ln 10$

Solution:

1. $\ln e^2$ is the exponent y such that $e^y = e^2$. The only possible solution for y is 2. So $\ln e^2 = 2$. In fact, the same logic implies that $\ln e^x = x$ for any x.

2. We are looking for the exponent y such that $e^y = 1$. Since any nonzero number to the 0 power is 1, we have $\ln 1 = 0$.

3. Since $\frac{1}{e} = e^{-1}$, we get $\ln(1/e) = \ln e^{-1} = -1$.

4. We are looking for the exponent y so that $e^y = 10$. Unless we want to use the "guess and check" method, we would evaluate $\ln 10$ using a calculator. We get $\ln 10 \approx 2.3026$. To check, we can evaluate $e^{2/3026}$ to get 10.00015 or very close to 10.

■

Properties of the Natural Logarithm

Previously, we gave a number of properties satisfied by the common base 10 logarithm. The natural logarithm satisfies essentially these same properties, as do all logarithms, and for the same reasons. Explicitly, we have the following properties of the natural logarithm.

- $\ln e^x = x$.
- $\ln(1) = 0$.
- $\ln x$ is negative for $0 < x < 1$ and positive for $x > 1$.
- $y = \ln x$ has a domain of $x > 0$ and a range of all real numbers.
- The graph of $y = \ln x$ is increasing and concave down.

1. $\ln(uv) = \ln u + \ln v$.

2. $\ln(u/v) = \ln u - \ln v$.

3. $\ln x^n = n \ln x$.

4. $e^{\ln x} = x$.

Example 3.
Use the properties of the natural logarithm to express the following expressions using only $\ln x$, $\ln y$, and $\ln z$.

1. $\ln x^4 y$

2. $\ln \frac{\sqrt{x}}{yz}$

Solution:

1. We have

$$\begin{aligned} \ln x^4 y &= \ln x^4 + \ln y && \text{by property 1} \\ &= 4 \ln x + \ln y && \text{by property 3} \end{aligned}$$

2. First, recall that $x^{1/2} = \sqrt{x}$. We have,

$$\begin{aligned} \ln \frac{\sqrt{x}}{yz} &= \ln \sqrt{x} - \ln yz && \text{by property 2} \\ &= \ln x^{1/2} - (\ln y + \ln z) && \text{by property 1} \\ &= \frac{1}{2} \ln x - \ln y - \ln z && \text{by property 3 and the distributive law} \end{aligned}$$

■

Logarithmic Regression

In Chapter Six, we used exponential regression as a standard way to find a "best fit" exponential model for a given set of data. We would tend to apply exponential regression to data where the graph is increasing (or decreasing) and concave up.

Given the basic shape of the graph of the logarithmic function, we would consider using a logarithmic model when the graph is increasing and concave down, and contains only positive x values. As with exponential regression, a number of software packages and graphing calculators have logarithmic regression features. Typically, these give equations of the form $y = a + b \ln x$ or $y = a + b \log x$.

Example 4. Table 2 shows how life expectancies changed in the United States during the 20th century.[103] The value given for each year represents how long a person born in that year could expect to live, on the average. Find a logarithmic model for this data, and discuss how good the fit is. Use the model to predict the life expectancy for U.S. children born in 2010 and 2050. How reasonable do these predictions seem to be?

Year	1900	1915	1930	1945	1960	1975	1990	2000
Life Expectancy	47.3	54.5	59.7	65.9	69.7	72.6	75.4	76.4

Table 2: U.S. Life Expectancies

Solution: Figure 2 shows a scatterplot of the data. The pattern seems to be slightly concave down. This indicates that a logarithmic function might be an appropriate model.

If we let the independent variable be years from 1900, however, we will have to take into account that we cannot use 0 as a value when taking logarithms (since log(0) is not defined). To adjust for this, we leave off the first data point and construct the plots using only the years starting with 1915, which corresponds to an x-value of 15.

Using logarithmic regression on a TI calculator, we get the logarithmic model is $y = 12.05 \ln x + 20.521$, which uses the natural logarithm. The original data with the model is graphed in Figure 3. It seems to fit the data well.

If you prefer, you can change the $\ln x$ in the model to $\log x$ using the change of base formula.

$$\ln x = \frac{\log x}{\log e} \approx \frac{\log x}{.434494} \approx 2.3026 \log x$$

[103]This example is adapted from *Functioning in the Real World* by Sheldon Gordon, et. al., p. 189

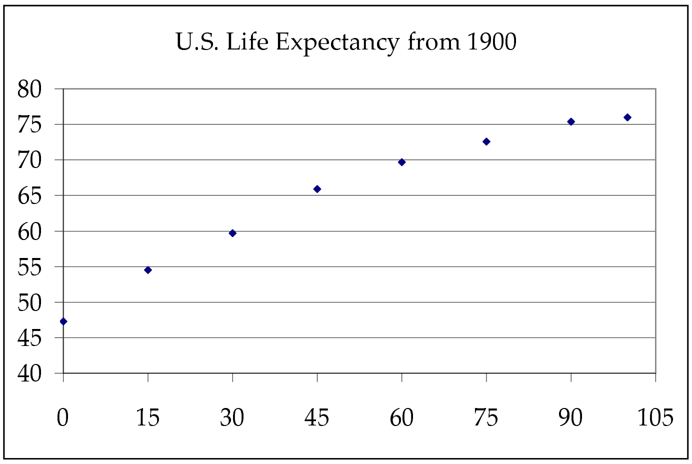


Figure 2: U.S. Life Expectancy

This changes the form of the equation to $y = 12.05(2.3036 \log x) + 20.521$ or $y = 27.746 \log x + 20.521$. This is really the same equation, and will give the same estimates as the equation using the natural logarithm $\ln x$.

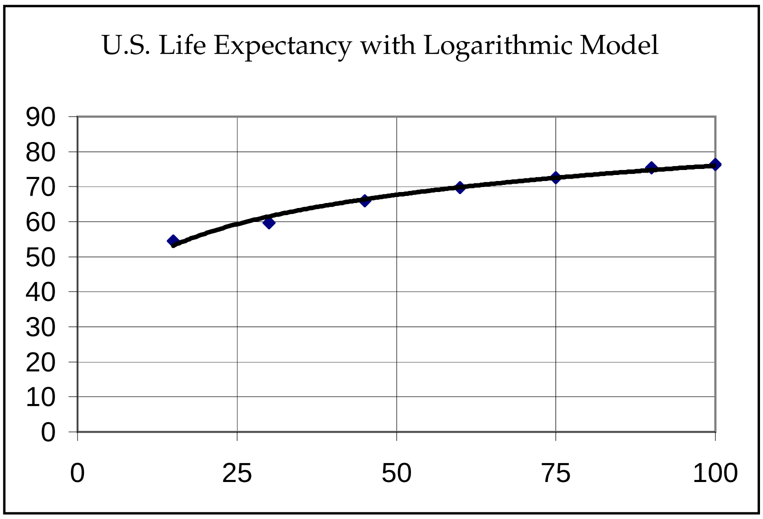


Figure 3:

For the year 2010, the logarithmic model predicts a life expectancy 77.02. ■

Solving Exponential Equations

One very important use of logarithms is in solving exponential equations. An exponential equation is any equation where the variable appears as an exponent. Let's begin with an example.

More on Compound Interest

In Chapter Six, we considered a scenario in which your grandmother invested $2,000 when you were born in order to help you with your college tuition. At a 6% annual interest rate compounded monthly for 18 years you had a total of \$5873.53 to apply towards your first year.

Suppose, however, that you have been successful in obtaining your tuition dollars by other means and don't need grandmother's money now. You decide to keep saving the money until you have \$10,000, at which point you intend to use it as a down payment on a house or perhaps for a new car. How long will it take until you have your \$10,000? To answer this question, we will use logarithms, principally property 3 in the list of properties of logarithms.

In this situation, we want to know for which value of t will $B(t)$ be equal to \$10,000, where $B(t) = P(1+i)^t$. We can set this up as an equation, using $P = \$2000$ and $i = 0.06/12 = 0.005$. Since we are compounding monthly, t represents the number of months. We obtain the equation

$$10,000 = 2000(1.005)^t.$$

We want to solve this for t. In general, whenever you want to solve an equation where the variable is an exponent, use logarithms. First, however, we divide both sides by 2000 to get

$$5 = (1.005)^t$$

Then we take the logarithm of both sides of the equation, and use property 3.

$$\log 5 = \log(1.005)^t = t \cdot \log 1.005$$

$$t = \frac{\log 5}{\log 1.005} \approx 322.7$$

You will have your \$10,000 in 323 months (from your birth) which means you can have your house or your car before your 27th birthday.

This example illustrates the general use of logarithms in solving exponential equations. Here are a few more examples.

Example 5.
Solve the exponential equations.

- $12^x = 456$
- $4(5^{x+1}) = 76$

Solution:

- We start solving the equation $12^x = 456$ by taking the logarithm of both sides. We can use any base, but base 10 and base e are the most used.

$$\log 12^x = \log 456$$

$$x \log 12 = \log 456$$

$$x = \frac{\log 456}{\log 12} \approx \frac{2.659}{1.079} \approx 2.464$$

We can check this by computing $12^{2.464}$.

- To solve $4(5^{x+1}) = 76$ we first divide by 4.

$$4(5^{x+1}) = 76$$

$$\frac{4(5^{x+1})}{4} = \frac{76}{4} = 19$$

$$\log 5^{x+1} = \log 19$$

$$(x+1)\log 5 = \log 19$$

$$x + 1 = \frac{\log 19}{\log 5}$$

$$x = \frac{\log 19}{\log 5} - 1 \approx .8295$$

■

Reading Questions for Logarithmic Functions

1. Express the following without using logarithms. You should not need a calculator.

(a) $\log_2 32$
(b) $\log_a a^2$
(c) $\log 100,000,000$
(d) $\log_b b$
(e) $\log_2 \frac{1}{8}$
(f) $\ln e$
(g) $\ln e^6$
(h) $\ln e^{x+1}$

2. Solve the following equations.

(a) $3^x = 81$
(b) $3^x = 110$
(c) $2 \cdot 3^x = 110$
(d) $3 \cdot 2^x = 110$
(e) $5 \cdot 1.3^{x+2} = 294$
(f) $2365 = 12(1.016)^x$
(g) $\frac{2^{x-3}}{5} = 11.7$

3. Suppose you have $5000, and you want to save up to buy a new pickup truck to replace your aging 1998 Ford F-150. You estimate you can get at least 4 more years out of your current vehicle, and so tentatively plan to invest the money for 4 years and use the balance as a down payment on a new truck.

(a) Assume you can invest your money at fixed annual rate of 6% compounded quarterly. How much would you have at the end of 4 years?
(b) As in part a), assume you can get at annual rate of 6% compounded quarterly. However, at the end of the 4 years, you decide your current truck is doing fine, and so decide to wait to buy the new truck until you have $8000 in the account. How much longer until this happens?
(c) Instead of investing your money at fixed annual rate of 6% compounded quarterly, you decide to take a somewhat riskier approach and invest in a variable rate account. The upside is you can currently get an 8% annual rate in such an account. The downside is the interest rate may go down, possibly significantly, before the 4 years is up.
   i. If you are lucky, the interest rate will stay at 8%, compounded quarterly. How long would it take for this account to grow to the same balance as you had in part a), where you earned 6% and left the money in the account for 4 years?
   ii. If you are not lucky, the interest rate will go down. To keep things simple, suppose the account stays at 8% for the entire first year, goes to 6% during the second year, 5% during the third, and 4% during the fourth.[104] How much will you have after 4 years under this scenario?

[104] In reality, interest rates in such accounts may vary from quarter to quarter

4. Match each of the following functions in symbolic form with the appropriate graph.

(a) $y = 3^x$

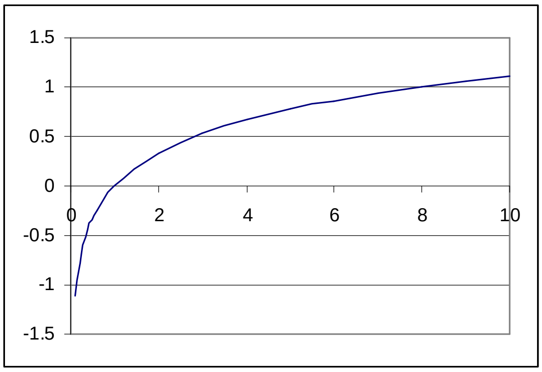
(b) $y = \log_2 x$

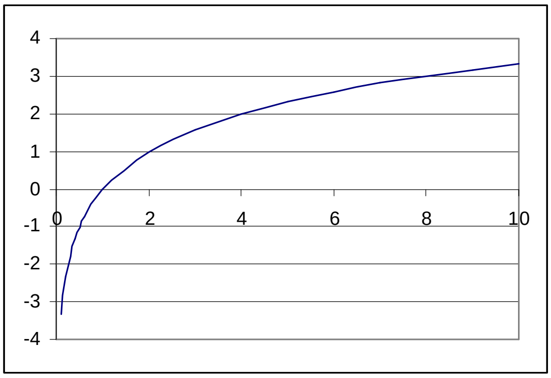
(c) $y = 10 \log_2 x$

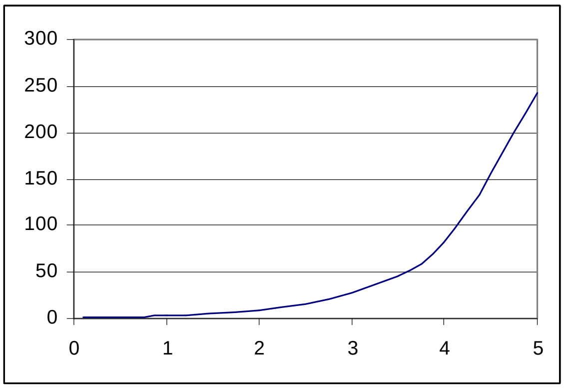
(d) $y = \log_8 x$

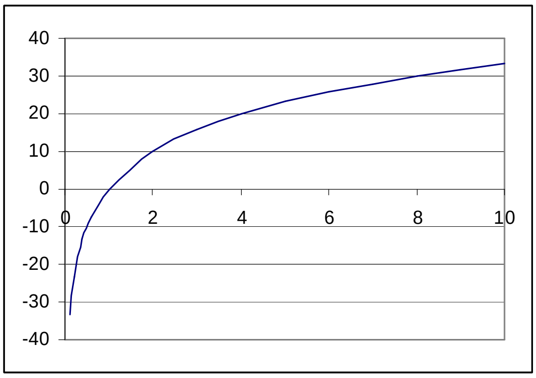
(e) $y = 2 \cdot 1.5^x$

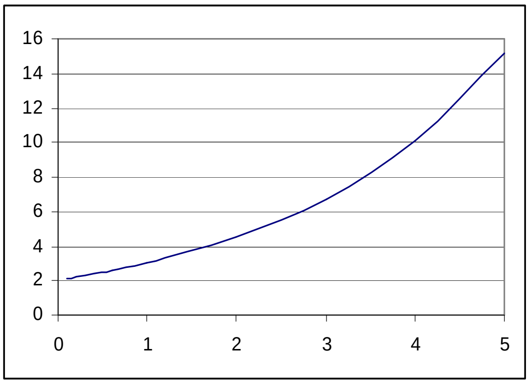
(f) $y = \log_2 4^x$











5. Solve the following equations. Changing to exponential form may help.

(a) $\log_3 x = 4$

(b) $\log_x 100 = 2$

(c) $\ln x = 2.5$

6. For each step, say which of the properties 1 through 4 of the natural logarithm are being used.

$$\begin{aligned} \ln a \cdot b^x &= \ln a + \ln b^x \\ &= \ln a + x \ln b \end{aligned}$$

7. For each step, say which of the properties 1 through 4 of the natural logarithm are being used.

$$\begin{aligned} \ln a \cdot x^b &= \ln a + \ln x^b \\ &= \ln a + b \ln x \end{aligned}$$

8. Use properties 1 through 4 of the natural logarithm to write the given expression in so that any logarithms are only $\ln x$, $\ln y$, or $\ln z$.

(a) $\ln xy^2$
(b) $\ln x^2y^3$
(c) $\ln \frac{x}{z}$
(d) $\ln \frac{zx^2}{y}$
(e) $\ln \sqrt{xyz}$

9. What does it mean when we say "logarithms are exponents?"

10. Say whether the following are true or false.

(a) $\ln(1) = 0$.
(b) $\ln(uv) = \ln u \cdot \ln v$.
(c) $\ln e^x = e$.
(d) The graph of $y = \ln x$ is increasing and concave down.
(e) $\ln(u/v) = \ln u - \ln v$.
(f) $\ln x^n = n \ln x$.
(g) $e^{\ln x} = x$.
(h) $y = \ln x$ has a domain of $x > 0$.

Logarithmic Functions: Activities and Class Exercises

1. **Airline Revenues.**

Table 3: Airline Yields for Domestic U.S. Flights

Year	Nominal Yield	Real (78) Yield(78$)	Log of Real Yield	Year	Nominal Yield	Real Yield(78$)	Log of Real Yield
1926	12.03	44.31	1.647	1964	6.12	12.87	1.110
1927	10.6	39.72	1.599	1965	6.06	12.54	1.098
1928	11	41.94	1.623	1966	5.83	11.73	1.069
1929	12	45.75	1.660	1967	5.64	11.01	1.042
1930	8.3	32.4	1.511	1968	5.61	10.51	1.022
1931	6.7	28.74	1.458	1969	5.79	10.29	1.012
1932	6.1	29.03	1.463	1970	6	10.08	1.003
1933	6.1	30.59	1.486	1971	6.33	10.19	1.008
1934	5.9	28.71	1.458	1972	6.4	9.98	0.999
1935	5.7	27.13	1.433	1973	6.63	9.74	0.989
1936	5.7	26.74	1.427	1974	7.52	9.95	0.998
1937	5.6	25.36	1.404	1975	7.69	9.32	0.969
1938	5.18	23.45	1.370	1976	8.16	9.35	0.971
1939	5.1	23.92	1.379	1977	8.61	9.26	0.967
1940	5.07	23.61	1.373	1978	8.49	8.49	0.929
1941	5.04	22.35	1.349	1979	8.96	8.05	0.906
1942	5.34	21.36	1.330	1980	11.49	9.09	0.959
1943	5.35	20.16	1.304	1981	12.74	9.14	0.961
1944	5.34	19.78	1.296	1982	12.02	8.12	0.910
1945	4.95	17.93	1.254	1983	12.05	7.89	0.897
1946	4.63	15.48	1.190	1984	12.8	8.03	0.905
1947	5.05	14.77	1.169	1985	12.21	7.4	0.869
1948	5.76	15.58	1.193	1986	11.08	6.59	0.819
1949	5.78	15.83	1.199	1987	11.45	6.57	0.818
1950	5.56	15.04	1.177	1988	12.31	6.78	0.831
1951	5.61	14.07	1.148	1989	13.08	6.88	0.838
1952	5.57	13.7	1.137	1990	13.43	6.7	0.826
1953	5.46	13.33	1.125	1991	13.24	6.34	0.802
1954	5.41	13.11	1.118	1992	12.85	5.97	0.776
1955	5.36	13.04	1.115	1993	13.74	6.2	0.792
1956	5.33	12.78	1.107	1994	13.12	5.77	0.761
1957	5.31	12.32	1.091	1995	13.52	5.78	0.762
1958	5.64	12.72	1.104	1996	13.76	5.72	0.757
1959	5.88	13.17	1.120	1997	13.97	5.68	0.754
1960	6.09	13.41	1.127	1998	14.08	5.63	0.751
1961	6.28	13.69	1.136	1999	13.96	5.46	0.737
1962	6.45	13.93	1.144	2000	14.57	5.52	0.742
1963	6.17	13.15	1.119	2001	13.25	4.88	0.688
				2002	11.98	4.34	0.637

December, 2003 marked the 100th anniversary of the world's first powered flight, conducted by the now famous Orville and Wilbur Wright. Since then, air travel has mushroomed into a huge industry with revenues of over 100 billion dollars annually. This activity considers how airline revenues have changed over the first century of flight.

One standard measure of airline revenues is called the yield. In short, the yield is the amount of revenue taken in per person per mile flown. For a single flight, this can be calculated by taking the total revenue from all paying passengers, dividing by the number of passengers (giving the average fare per passenger), and then dividing this by the number of miles. This is typically converted to cents per person per mile. The accompanying graph shows the average airline (Nominal) yields for all U.S. domestic flights for the years 1926-2002, along with real yields. The real yields are the nominal yields adjusted for inflation. In this chart, real yields are calculated in 1978 real dollars. The table of the data is given above.

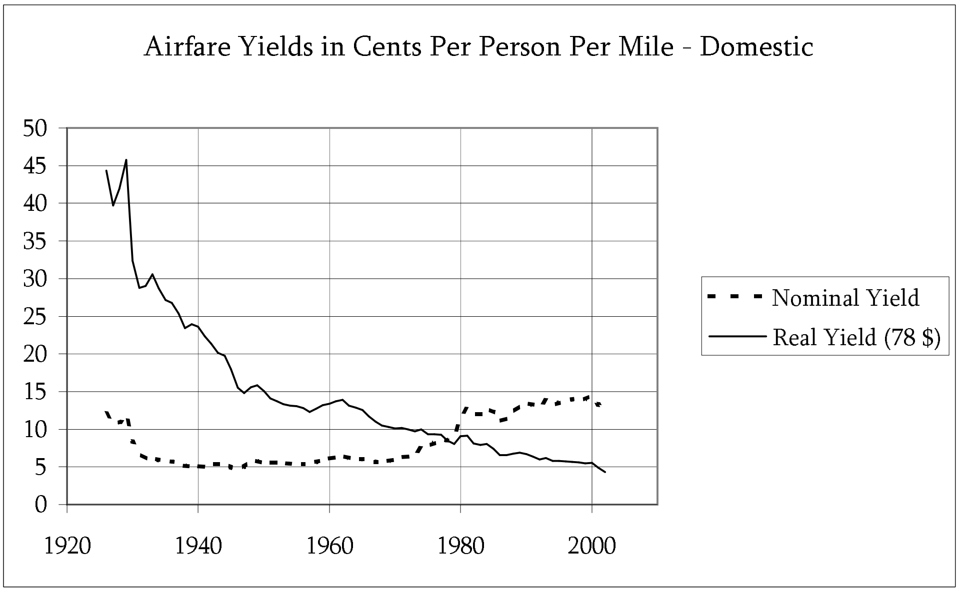


Figure 4: Graph of Airline Yields - Nominal and Adjusted for Inflation

(a) Using nominal yields, about how much did it cost a person to fly a mile in 1940? About how much did it cost in 2000?

(b) The nominal yields decreased sharply during the early 1930's. Why do you think this happened? You could discuss general sorts of reasons which might occur during any period, or particular historical events that were occurring during that period.

(c) During the 1970's, nominal yields increased sharply, and yet real yields continued to decline at roughly the same rate. Explain how this is possible using the idea of inflation.

(d) The graph of real yields decreases steadily throughout this period, and looks as if it might follow an exponential curve with a negative percentage growth rate. Use exponential regression to find an exponential model for real yields as a function of time. Instead of using all of the data, you could use data from just the years which are multiples of five (1930, 1935, etc.). Graph a scatter plot of the data you used along with your regression equation. Visually, how well does the curve seem to fit the data?

(e) We have seen that, when using linear regression, it is possible that the correlation coefficient may imply a fairly strong correlation, and yet the scatter plot clearly shows a linear model is not appropriate. The same thing might occur with exponential regression, as discussed in the text, and we can test visually whether an exponential model is appropriate by taking the Log of the dependent variable data, and seeing whether this results in a linear model. Take the Log of the yield data you used in part (d), and then create a scatter-plot of the Log of the real yields versus time. Does this relationship appear to be linear? If so, estimate the slope of a line that would model this data (you may do this from the graph or use linear regression on the Log of real yields data). If m is the slope, calculate 10 to the m. How close is this to the growth factor for your regression equation?

(f) The graph of nominal yields seems to have taken a sharp decrease after the year 2000. Why do you think this happened?

(g) How can you tell from the table that the real yields are calculated using 1978 dollars?

2. **Flood Damages.** In both chapter four and chapter six, we considered data on U.S. Flood Damages. The scatterplot from chapter six is recreated below along with the exponential model given by the formula $D_1(t) = 0.741(1.0168^t)$. Here, t gives the year from 1900, and $D(t)$ is total flood damages in billions of dollars. Note that the equation looks different than the one given in the graph, but is essentially the same as $e^{0.0167} \approx 1.0168$.

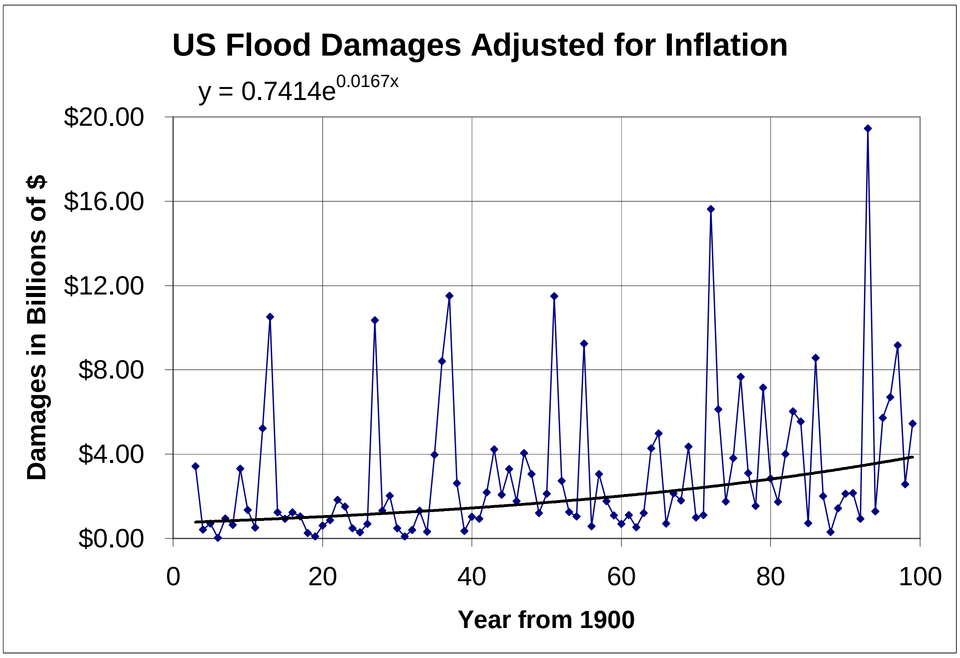


Figure 5: Two functional representations of US Flood Damages.

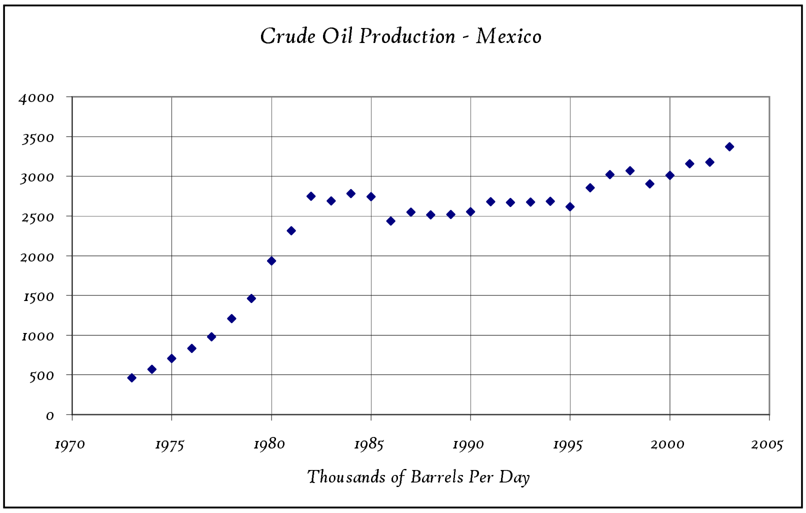
(a) What is the percentage growth rate represented by the exponential model D_1?

(b) What does the exponential model predict for the years 2050 and 2100?

(c) According to the model, when will flood damages reach 25 billion, and also 50 billion dollars?

(d) The data shows a lot of variability, and so the exponential model is likely to give estimates very different than the actual damages, both in past and future years. Create an additional model which will give estimated upper limits on flood damages. To create the "upper limit model," you could estimate the values for several of the " peak points" including 1913 and 1993.

3. Mexican Oil Production. The scatterplot below shows total crude oil production for Mexico from 1973 to 2003.



(a) If you looked *only* at the scatterplot, and did not consider any other information, which family of functions, linear, exponential, or logarithmic, would you select as the most appropriate to model this data? Explain why you picked as you did.

(b) The data below corresponds to the scatterplot above.

Year	Production	Year	Production	Year	Production
1973	465	1984	2780	1994	2685
1974	571	1985	2745	1995	2618
1975	705	1986	2435	1996	2855
1976	831	1987	2548	1997	3023
1977	981	1988	2512	1998	3070
1978	1209	1989	2520	1999	2906
1979	1461	1990	2553	2000	3012
1980	1936	1991	2680	2001	3157
1981	2313	1992	2669	2002	3177
1982	2748	1993	2673	2003	3371
1983	2689				

i. Let $t = 0$ represent 1970. Find the linear regression model for this data.

ii. Again, let $t = 0$ represent 1970. Find the logarithmic regression model for this data.

(c) Graph the two models from part (b) together with the data. Which model seems to give the best fit.

(d) Why would a logarithmic model not be appropriate for modeling crude oil production in the long run?

4. Benford's Law. In 1938 Dr. Frank Benford, a physicist at the General Electric Company, noticed that, given a list of numbers, more numbers started with the digit 1 than any other digit. Random numbers are equally likely to start with a 1, 2, 3, ..., 9 yet that did not appear to be what really happened in lists of data. Dr. Benford analyzed 20,222 sets of numbers including areas of rivers, baseball statistics, numbers in magazine articles, and street addresses. He concluded that the probability of the numeral d appearing as the first digit was

$$\log\left(1+\frac{1}{d}\right)$$

In other words, according to Benford's Law, the probability that the first digit is a 1 is $\log(1+1) \approx 0.30$, the probability that the first digit is a 2 is $\log(1+\frac{1}{2}) \approx 0.18$, etc. So about three-tenths of the numbers in a list typically start with the digit 1 (even though a random list of numbers should have only one-ninth of the numbers starting with the digit 1.)[105]

(a) Complete the following table.

d	1	2	3	4	5	6	7	8	9
$\log(1+1/d)$									

This table, according to Benford's law, gives the proportion of first digits that are ones, twos, etc.

(b) We looked at 60 retired baseball jersey numbers from the American League.[106] The following table indicates how many numbers started with a 1, 2, 3, ..., 9.

First Digit of Jersey Number	1	2	3	4	5	6	7	8	9
Frequency	17	11	12	5	5	2	2	3	3

Do these numbers seem to conform to Benford's Law? Justify your answer by comparing a relative frequency distribution of the first digit of the jersey number to the distribution from part (a).

(c) We also looked at the populations in 1990 of counties in Michigan.[107] The following table indicates how many numbers started with a 1, 2, 3, ..., 9. There are 83 numbers in all.

First Digit of Population Number	1	2	3	4	5	6	7	8	9
Frequency	26	17	10	4	9	2	7	5	3

Do these numbers seem to conform to Benford's Law? Justify your answer by comparing a relative frequency distribution of the first digit of the population number to the distribution from part (a).

(d) Using one page from a newspaper, list all of the numbers that appear. Do they seem to conform to Benford's Law? Justify your answer.

(e) Benford's Law is currently used in software designed to detect fraud and tax evaders. Explain how you think it does this and why this is likely to work.

5. **Child Growth.** The table below gives average weights for a sample of children in the U.S.

(a) Make a scatterplot of this data. Which family of functions, linear, exponential, or logarithmic seems most appropriate to model this data and why?

(b) Find a logarithmic regression model for this data, and graph it along with the scatterplot. How well does the model seem to fit the data?

(c) Fill in the third column of the table, using your equation to predict the weights for each of the given inputs.

[105] Browne, Malcom W. "Looking Out for No. 1." *New York Times* 4 August 1998, p F4.
[106] *Baseball Almanac- The "Official" Baseball History Site.* http://www.baseball-almanac.com. 22 August 2002.
[107] Time Series of Michigan Population Estimates by County. http://eire.census.gov/popest/data/counties/tables/CO-EST2001-07/CO-EST2001-07-26.php. 23 August 2002.

Age in months	Weight in pounds	Log Model
3	12.3	
6	16.2	
9	19.1	
12	21.1	
15	22.8	
18	24	
21	25	
24	25.8	

(d) Use your model to predict weights for children who are 5 years old, 10 years old, and 15 years old. Do these predictions seem reasonable?

(e) Use your model to predict the weight of someone who is 50 years old. Does this prediction seem reasonable?

(f) Why would this model not be appropriate to model the weights of infants less than a month old?

6. **Comparing Car Models III.** This activity uses the same data on car models that appeared in chapters five and six. You can refer there for a table giving the complete data set.

(a) Make a scatterplot of greenhouse gas emissions versus city mileage.

i. The exponential regression model for this data is $y = 17.3(.963^x)$ where y represents the tons of carbon dioxide emitted per year by a car which gets x miles per gallon in city driving. Graph this equation along with your scatterplot. How well does it seem to fit?

ii. What does your exponential model predict for a car which gets 36 miles per gallon in the city?

iii. Using your exponential model, estimate the city miles per gallon for a car which emits 4 tons of carbon dioxide per year.

iv. Repeat the previous part for a car that emits 3 tons of carbon dioxide per year.

(b) Make a scatterplot of greenhouse gas emissions versus engine size. Note that engine size is measured in liters.

i. Explain why an exponential model would probably not provide a good fit to this data.

ii. Recall that the concavity of a graph indicates how the rate of increase or decrease in the dependent variable is changing with respect to the independent variable. This scatterplot is slightly concave down, and increasing. What does the concavity say about the *rate* of increase?

iii. Find the logarithmic regression model for this data and graph it along with the scatterplot.

iv. What does your logarithmic model predict for a car that has an engine size of 1.3 liters? What about 5 liters?

v. If you want to buy a car that only emits 4 tons of carbon dioxide per year, about what size of engine should you be considering?

(c) It is estimated that there are about 600 million passenger cars in the world.

i. Assuming that these cars emit an average that is the same as the 38 cars in our sample, how many tons of carbon dioxide are being emitted each year by the estimated 600 million cars?

ii. How much could this total be reduced if the average car emitted the same amount of carbon dioxide as a Toyota Celica?

8. Power and Quadratic Functions

It is the mark of an educated mind to be able to entertain a thought without accepting it.
— Aristotle

In the last two chapters, we considered the properties of exponential and logarithmic functions, and mentioned power functions briefly. Recall that:

- **Power functions** are those which can be written in the form $y = ax^b$. The variable is a *base* and the exponent is a constant.
- **Exponential functions** are those which can be written in the form $y = ab^x$. The variable is an *exponent* and the base is a constant.

In this chapter, we will examine power functions in greater detail, as well as a quadratic functions, which is a family related to the power function family with exponent $b = 2$.

Power Functions

As in the previous chapter, when considering examples of functions from a particular family, it is a good idea to experiment with your own particular examples so that you get a feel for how that family of functions work. We have divided the examples of power functions into cases (or we might say subfamilies), so that you need not consider all possible power functions all at once.

Example 1. Graph several examples of power functions $y = ax^b$, where b is a positive integer. Use a variety of values for the parameter a, although it is probably a good idea to try several different values of b for the same a as you begin. You might start with $a = 1$, varying b, and then try a different a with the same set of b values. In each case, determine any x or y intercepts, decide whether the function is increasing or decreasing, and describe the concavity. Then, write a few sentences summarizing your results. Try to characterize the graphs of power functions with positive integer exponents as completely as possible. How does the exponent effect the graphical behavior?

Solution: Figures 1 shows two examples for graphs of $y = ax^b$ where b is a positive integer. For $x > 0$, both graphs are concave up and increasing. What if $x < 0$? If b is an even integer, then the graph is decreasing and concave up. If b is an odd integer, the graph is increasing and concave down. Both graphs go through the origin, and this is the only point at which the graphs cross either of the axes. Also, we might notice that all the graphs go through the point $(1, a)$, whatever the parameter a happens to be (this fact is easy to miss, unless you try several examples with the same a). You can see why this happens by substituting $x = 1$ into the general equation $y = ax^b$.

Finally, you might notice that the larger the value of b, the steeper the graph is for larger x. In fact, since all the graphs of $y = ax^b$ go through $(1, a)$, we can see that larger b give a steeper graph from this point on.

One final note. If you graph $y = x^b$ using a small window from $x = 0$ to $x = 1$, you will see that the larger the b, the closer the graph is to the x-axis. Why is this? Think about what happens if you raise a number between 0 and 1 to larger and larger powers (you could even try some examples on your calculator). You could try $x = 0.5$, for example, and then raise this number to the second power, the third power, etc. The larger the exponent b, the smaller the resulting value of x^b. ■

In the next example, we will consider *negative* values for b in $y = ax^b$. If you are using a graphing calculator, you should change your graphing mode from **Connected** to **Dot** when graphing functions with negative b. To see the reason for this, consider $f(x) = 1/x$. This function is not defined for $x = 0$ (in other words, 0 is not in the domain of the function). We can evaluate f for x-values which are close to 0 (like $.1, -.1, .01$, etc.), and Table 1 shows the results from some of these evaluations.

The closer x is to zero, the larger (in absolute value) the value of $1/x$. The graph of $y = 1/x$ is shown below (Figure 2), and we see that the y-axis is a **vertical asymptote** for the graph. The

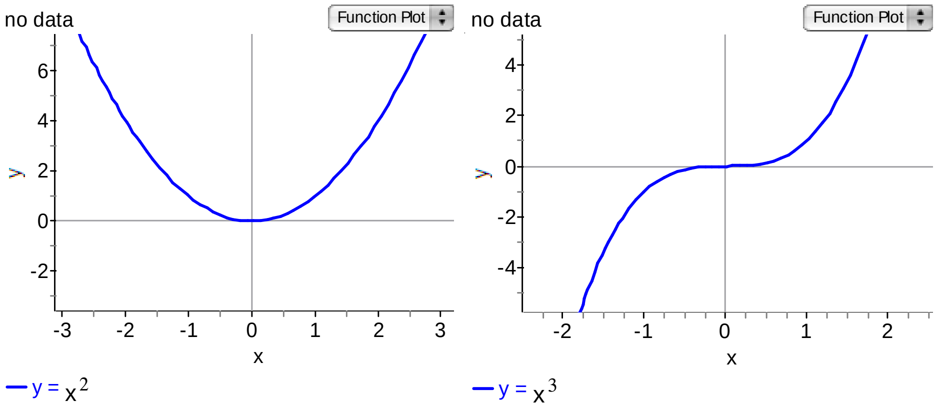


Figure 1: Two basic power function shapes.

x	5	0.1	0.01	0.001
$1/x$	2	10	100	1,000
x	-0.5	-0.1	-0.01	-0.001
$1/x$	-2	-10	-100	$-1,000$

Table 1:

graph gets closer and closer to this line as x gets closer and closer to 0, but the graph never touches the axis.

When a graphing calculator (or other technology that is not designed to handle vertical asymptotes) tries to graph a function like $y = 1/x$, it does not know there is a vertical asymptote, so it simply plots a number of points and then connects the dots. If you set your calculator to graph in **Dot** mode, you will see just the points that the calculator is graphing, without the connecting lines, and this will give you a more accurate picture of the graph.

Example 2. Now, experiment with power functions $y = ax^b$ with b a negative integer. Try several examples, until you feel comfortable that you understand the general characteristics of these functions, and then write a few sentences describing what you know.

Solution:
Figure 2 shows two examples you may have considered, $y = x^{-1} = 1/x$ and $y = x^{-2} = 1/x^2$. ■

Example 3. Finally, let's consider what happens if b is *not* an integer. Graph several examples of $y = ax^b$, and use a variety of exponents b, some between 0 and 1 and some larger than 1, as well as a variety of values for the parameter a. As before, it is a good idea to only vary one of these at at a time (for example, you might graph several functions with different values of b but all with $a = 1$). In each case, determine any x or y intercepts, decide whether the function is increasing or decreasing, and describe the concavity. Then, write a few sentences summarizing your results. Try to characterize the graphs of power functions as completely as possible. How does the exponent effect the graphical behavior?

Solution: Several examples are shown in Figure 3. Notice that, if the exponent b is not an integer, the graph often exists only where $x \geq 0$. We will see why this happens later.

For any $b > 1$, the graphs are concave up, and increasing for $x > 0$. If $0 < b < 1$, then the graph is increasing and concave down. In both cases, the larger the value of b, the steeper the graph is for $x > 1$. As in the previous cases of power functions, the graph of $y = ax^b$ goes though $(1, a)$. ■

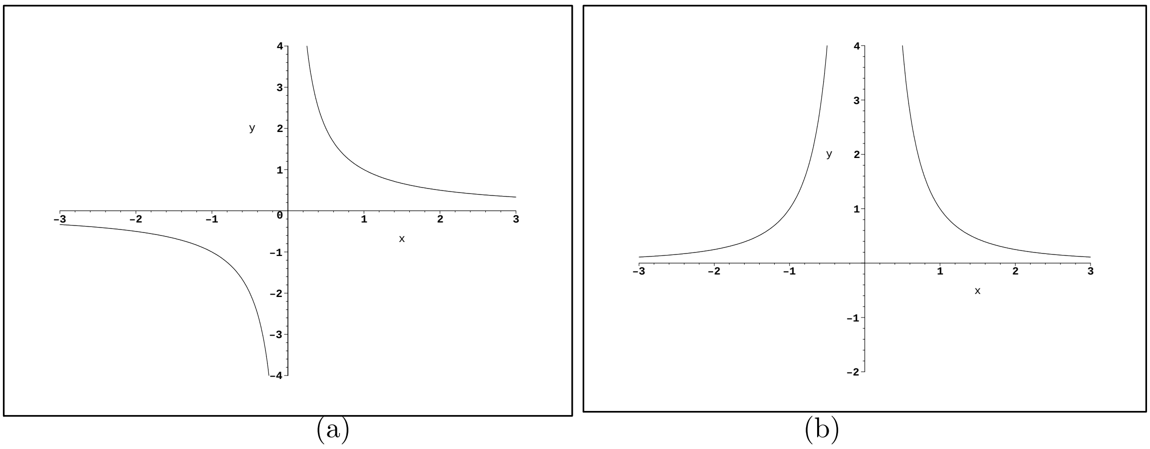


Figure 2: Two basic power function shapes for negative integer powers. Figure 2(a) is $f(x) = \frac{1}{x}$ and Figure 2(b) is $g(x) = \frac{1}{x^2}$.

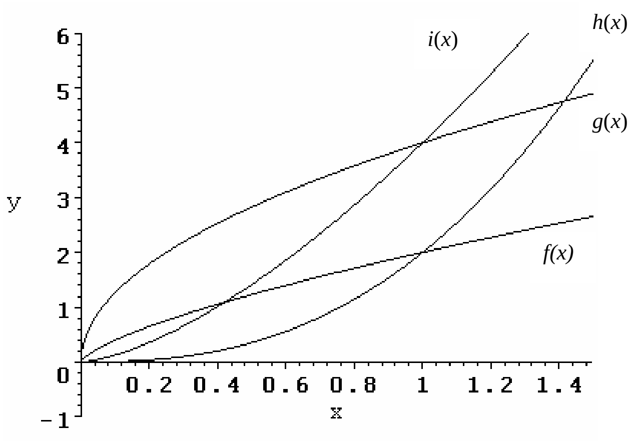


Figure 3: Graphs of $f(x) = 2x^{0.7}$, $g(x) = 4x^{0.5}$, $h(x) = 2x^{2.5}$ and $i(x) = 4x^{1.5}$

Domains of Power Functions

> Why does the graph of $y = ax^b$ exist only for non-negative x for many values of b which are not integers?

Let's consider $f(x) = x^{0.5}$ as an example. Recall that

$$f(x) = x^{.5} = x^{1/2} = \sqrt{x}$$

Since $\sqrt{x}$ is not defined for negative x, the domain of this function is $x \geq 0$.[108] So, the graph of f exists only for $x \geq 0$. In general, if b is not an integer, the evaluation of $f(x) = ax^b$ involves some type of root, and many of these roots will not be defined for negative values of x. Thus, some graphing technologies will not graph $y = ax^b$ for negative x if b is not an integer.

[108] We can define $\sqrt{x}$ for negative x if we use imaginary or complex numbers, but in this course, we will only use real numbers for domains and ranges of functions, unless otherwise noted.

Quadratic Functions

A **quadratic function** is a function of the form

$$f(x) = ax^2 + bx + c.$$

Here, a, b, and c are parameters. We have already met one member of the family of quadratic functions, namely $y = x^2$, in Figure 1. Another example is shown below in Figure 4. The shape of the graph of a quadratic function is called a **parabola**. All quadratic functions have this shape, although if the parameter a is negative, the shape will be turned "upside-down." The low point (or high point) of a parabola is called the **vertex**. We will soon see how the three parameters of a quadratic function determine the vertex, as well as the steepness or flatness of the corresponding parabola. Quadratic functions are used extensively in physics, and also have applications in business, economics, and other fields.

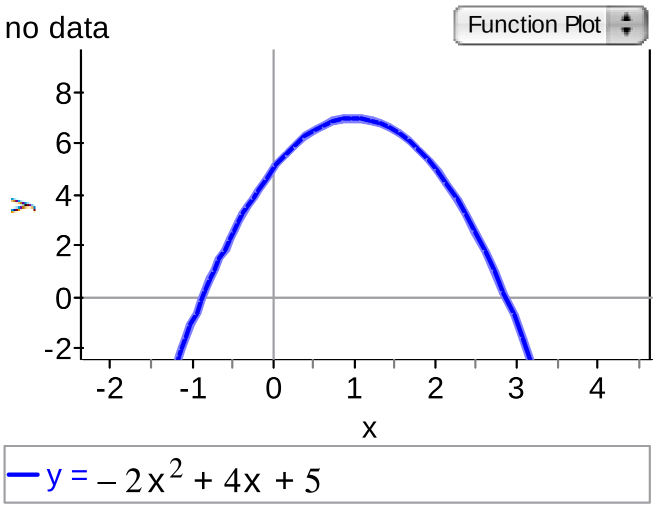


Figure 4: The parabola $y = -2x^2 + 4x + 5$. Here $a = -2$, $b = 4$, and $c = 5$.

The Quadratic Formula

At some point in your mathematical studies, you have probably seen the so-called **quadratic formula**. This formula can be used to solve *any* equation of the form $ax^2 + bx + c = 0$. The formula says that solutions for x are given in terms of the parameters a, b, and c by

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

For a quadratic function $f(x) = ax^2 + bx + c$, the solutions given by the quadratic formula tell you the x-intercepts of the parabola, as long as the quantity under the square root sign is not negative.[109]

One convenient property of parabolas is that they are symmetric about the vertical line through the vertex. So, whenever there are two intercepts, the x-coordinate of the vertex will be half way in

[109] If the quantity $b^2 - 4ac$ under the root sign is negative, then the solutions to the equation are complex numbers, and will not correspond to any points on the graph of the parabola. The expression $b^2 - 4ac$ is called the **discriminant** of the equation.

between the x-intercepts. Because of this, one can deduce that the vertex of a quadratic function will be the point

$$\left(\frac{-b}{2a}, f\left(\frac{-b}{2a}\right)\right)$$

To see why this is, note that the quadratic formula can be written

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{-b}{2a} \pm \frac{\sqrt{b^2 - 4ac}}{2a}$$

This means that one x-intercept is equal to $-b/2a$ plus a particular number, given by the expression after the $\pm$ in the equation above. The other x-intercept is $-b/2a$ minus this same number. So, the two x-intercepts are the same distance away from $-b/2a$ which means $-b/2a$ must be half way in-between the two intercepts.

You can use the quadratic formula and the vertex formula to easily relate the parameters of any quadratic function to its x-intercepts and vertex.

Example 4. Find the vertex and x-intercepts of the graph of $y = 2x^2 - 8x + 3$.

Solution: The parameters are $a = 2$, $b = -8$, and $c = 3$. So, $\frac{-b}{2a} = \frac{8}{4} = 2$. We have $f\left(\frac{-b}{2a}\right) = f(2) = 2 \cdot 2^2 - 8 \cdot 2 + 3 = -5$. So, the vertex is $(2, -5)$.

The x-intercepts can be found using the quadratic formula and are

$$x = \frac{-b}{2a} \pm \frac{\sqrt{b^2 - 4ac}}{2a} = 2 \pm \frac{\sqrt{(-8)^2 - 4 \cdot 2 \cdot 3}}{4} = 2 \pm \frac{\sqrt{40}}{4} \approx 2 \pm 1.58$$

or about 0.42 and 3.58. ■

From this chapter and the last, we now have a fairly good idea of the general characteristics of graphs for exponential, logarithmic, power, and quadratic functions, based on observations of graphs. We also have some connections between the symbolic and the graphical representations. For exponential functions $y = ab^x$, the base tells us whether the graph is increasing or decreasing and how steeply it does so. The parameter a tells us the y-intercept. For power functions $y = ax^b$, the exponent b determines where the function is concave up and concave down as well as the steepness.

Knowing these general characteristics will help us decide which family is more likely to give us a good mathematical model when we are presented with a scatterplot of data.

Power Regression Models

In Chapters Six and Seven, we used exponential and logarithmic regression to find models. Similarly, we can use power or quadratic regression to find models when either of these families of functions seems appropriate. This can be done on a TI calculator or in Excel.

> It is important to remember when using the power regression feature in any technology to exclude any zero or negative values in either the independent or dependent variable lists! The reason for this is that the domain of $\log x$ is all positive real numbers. We will discuss the details of why this is in the next section of this chapter.

Example 5. In Chapter Two, we introduced Lorenz curves and the Gini coefficient. In the Activities, you may have looked at the Lorenz curves for several countries, including the U.S. Figure 5 shows the Lorenz curve, and the data on which it is based is in Table 2. The data is for the U.S. for the year 1989. Recall that the Gini coefficient is twice the area between the curve and the "equality line."

Table 2: Distribution of Income by Percentile for the U.S., 1989

Population Percentile	0	0.20	0.40	0.60	0.80	1.00
Percent of Income	0.000	0.046	0.152	0.317	0.554	1.000

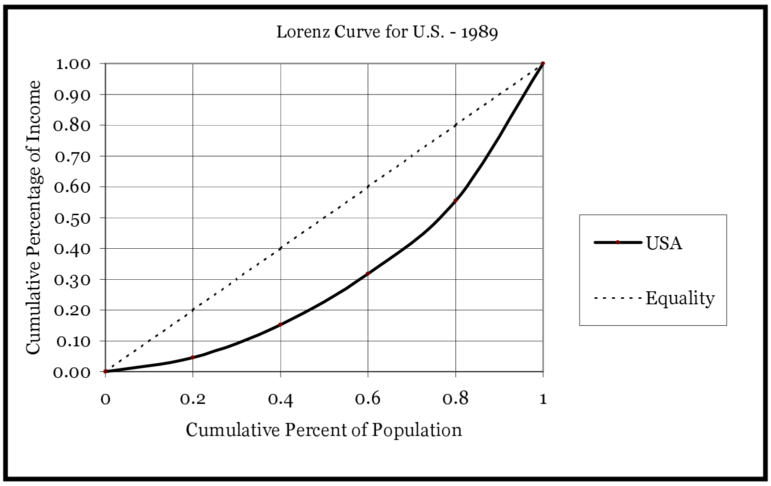


Figure 5: Lorenz Curve for U.S., 1989

In the Chapter Seven Activities, you may have found power functions of the form $L(x) = x^b$ to model this data by finding b so that the graph of $y = L(x)$ went through one of the points for which x was not 0 or 1, in addition to the two points $(0, 0)$ and $(1, 1)$. The problem with these functions was that, even though they went exactly through three of the points, they were somewhat far away from some of the other points. Can we find a power function which is "on average", closer to all the points?

Solution: To do this, we will find the **power regression** function modeling this data and using technology, for example a TI-83 calculator.[110] We simply input the x-values .2, .4, .6, .8, and 1.0 into one list, and the y-values (as shown in the table above) in another list (but not $y = 0$). Remember that we cannot use 0 as an x-value or a y-value when doing power regression. The power regression feature can be found in the **STAT** menu, under **CALC**. Then scroll down to **PwrReg**. The calculator returns the two parameters which define the power function. The resulting function is $y = 0.89x^{1.88}$. Figure 6 shows the scatterplot for the data in Table 2, values for a Lorenz curve, along with our regression function and the function $L(x) = x^{2.25}$. The regression equation seems visually to be a better fit for the data, except near the point $(1, 1)$.

■

Re-expression and Nonlinear Regressions

Let's look at exactly how your TI calculator or whatever technology you have available is calculating the exponential and power regression functions. As we do this, we will explain some of the problems with non-linear regressions. Understanding the mathematical details of how non-linear regression works will give you a better understanding of how to use these models, and also

[110]Excel does not have a built in power regression feature. However, we will eventually see how we can use Excel's linear regression feature to find the power regression function.

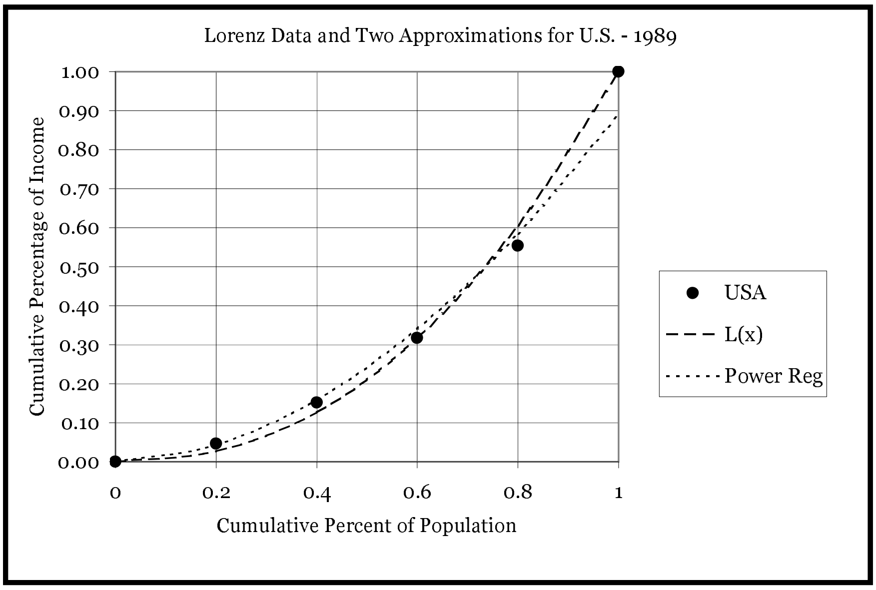


Figure 6: Scatter plot of Lorenz information with two power models

help you make judgements on which non-linear model is the most appropriate. In addition, it will give us an example of the use and importance of logarithms and their properties.

In Chapter Five, when we introduced linear regression, we noted that the linear regression line is the line which minimizes the sum of the squared errors (refer back to that chapter to review the details).[111] The calculation depends on the fact that we are using a linear function, the simplest type of function, to model the data. In cases where the data is not linear, and we want to find a non-linear model, we will first need to **re-express** the data to see if we can "make it approximately linear", and then use linear regression.

We will first explain the steps a calculator or other piece of technology would do internally to find the exponential regression, and then explain why this process works to give a good exponential model. We will then do the same for power regression.

Exponential Regression and Re-expression

Figure 7 shows the Child Mortality Rate for Albania over time. Obviously the Child Mortality Rate will always be at least 0. The scatterplot is decreasing, and the trend seems to be leveling off towards zero over time. The general shape is concave up. All these indicate that an exponential model with base between 0 and 1 might be appropriate.

To find the exponential regression function for this data, your technology would go through the following steps (but of course, automatically and very quickly).

1. The technology takes the logarithm of the given dependent variable values, in this case the (Child) Mortality values. The Mortality Rates along with Log(Mortality) are shown in Table 3.

2. The technology finds the linear regression line for Log(Mortality) versus Year data. Figure 8 shows a scatterplot of Log(Mortality) versus Year along with the graph of the linear regression line for this relationship. Notice that the line does seem to fit the data fairly well. The equation for the regression line $y = Mx + B$ shown has slope $M = -0.016455$ and y-intercept $B = 2.12646$. We are using capitals M and B to denote the parameters for this linear equation to avoid confusion with the parameters we will get for the exponential model below.

[111]The details of how the linear regression calculation is done can be found in most elementary statistics books.

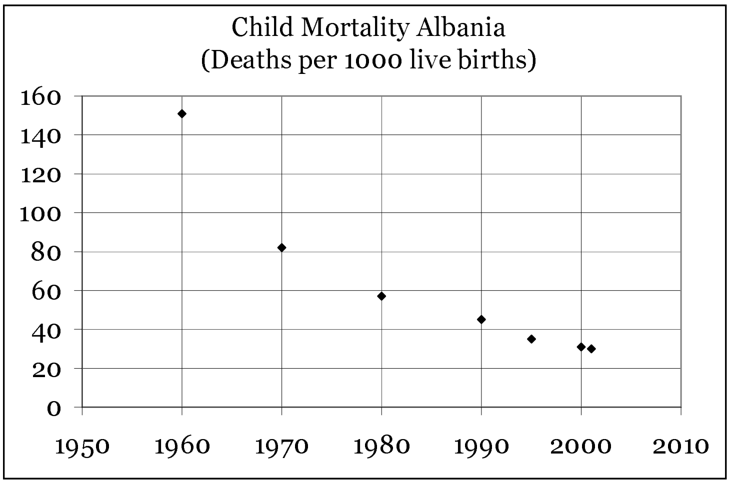


Figure 7: Child Mortality Rates for Albania over time

Year from 1960	Child Mortality	Log(Mortality)
0	151	2.178977
10	82	1.913814
20	57	1.755875
30	45	1.653213
35	35	1.544068
40	31	1.491362

Table 3: Mortality and Log(Mortality) as a function of Year since 1960

3. The technology finds the parameters a and b for an exponential equation $y = ab^x$ modeling the original Mortality versus Year data. It does this by converting the linear regression equation for Log(Mortality) versus Year into an exponential model for the the original Mortality versus Year relationship. This is done using basic properties of exponents and logarithms.

- The parameter a, which is the y-intercept for the exponential model, equals 10^B, where B is the y-intercept for the linear model for the logarithm data.
- The parameter b, which is the base for the exponential model, is equal to 10^M, where M is the slope for the linear model of the re-expressed (logarithmic) data.

In our case, we have $a = 10^B \approx 10^{2.12646} \approx 133.8$ and $b = 10^M \approx 10^{-0.016455} \approx .9628$. The exponential equation $y = 133.8(.9628^x)$ is shown below along with the original scatterplot in Figure 9. The correlation coefficient for the exponential regression given by the technology is $r = 0.988$.

Why does this process work?

When we find the linear regression model for Log(Mortality) versus Year, we are really getting a model of the form $\log y = Mx + B$, where y will give the predicted values for the original Mortality data. Using basic algebra, including the properties of logarithms we considered in the previous chapter, we can transform this equation into an exponential model for y versus x. Recall that

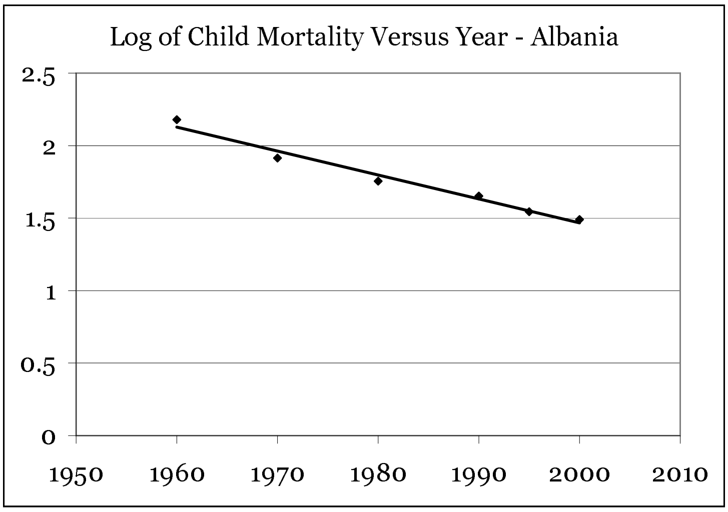


Figure 8: A plot of the logarithms of child mortality versus year, with linear regression line

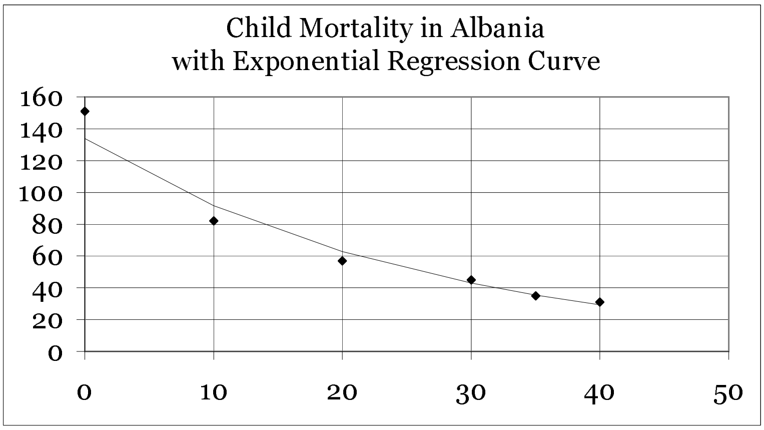


Figure 9: Child Mortality in Albania with Exponential Regression Model

$$\log y = x \qquad \text{means the same as} \qquad 10^x = y$$

So, applying this to the equation $\log y = Mx + B$, we get

$$\log y = Mx + B \qquad \text{means the same as} \qquad 10^{Mx+B} = y$$

Using properties of exponents we get

$$y = 10^{Mx+B} = (10^{Mx})(10^B) = (10^M)^x(10^B) = 10^B(10^M)^x$$

If we let $a = 10^B$ and $b = 10^M$ as above, this becomes $y = ab^x$.

Power Regression and Re-expression

For power regression, your technology also uses re-expression. In this case, if we let Independent and Dependent represent your data variables, then your technology lets X = Log(Independent) and $Y = \text{Log(Dependent)}$ and then finds the linear regression model for X and Y. This results in a model of the form.

$$\log y = M \log x + B$$

Where x and y are the variables modeling the original Dependent versus Independent data relationship. Changing to exponential form as before, we get

$$y = 10^{M \log x + B} = (10^{M \log x})(10^B)$$

Now, from our properties of logarithms, $M \log x = \log x^M$. Also, recall from our properties of logarithms that $10^{\log x^M} = x^M$. From these, we get that

$$y = (10^{M \log x})(10^B) = (10^{\log x^M})(10^B) = x^M \cdot 10^B = (10^B)x^M$$

so, the parameters a and b for the power regression for y versus x are given by $a = 10^B$ and $b = M$.

Also, we can now say why you cannot use 0 in your data set for either the dependent or independent variable when doing power regression. In fact, you also cannot use negative numbers. Recall the $\log x$ is not defined unless x is positive. If your data has 0 or negative values for the independent variable (or x), then your technology will give an error when it tries to take the logarithm of these values. The same thing happens if you have any dependent variable values that are zero or negative.

With exponential regression, the technology is not taking the logarithm of the independent variable values, but it is taking the logarithm of the dependent variable values so you will typically not be able to do exponential regression on data where there are any dependent variable (or y) values which are negative or 0.

Correlation Coefficients for Non-linear Regressions

Above, we cautioned that the linear regression correlation coefficient cannot be compared to the exponential (or other non-linear) correlation coefficient. We will illustrate why using the Albanian Child Mortality Data above.

Recall that the linear regression line is the line that minimizes the sum of the squared errors, where the errors are the differences between the actual data values for the dependent variable for each x, and the y-values predicted by the linear regression line for each x. We might hope that the exponential regression equation would give the smallest sum of squared errors over all possible exponential models. However, this is *not* the case. Table 4 gives the original data along with the values predicted by the exponential regression equation $y = 133.8(.96282^x)$, the values predicted by another exponential model $f(x) = 151(.9591^x)$, as well as the errors between these models and the data, the squared errors, and the sum of the squared errors. Note that the sum of the squared errors for the exponential regression equation is 427.308, which is larger than the sum of the squared errors for $f(x)$, 387.018. If we measure "best" by which function gives us the smallest sum of squared errors, the exponential regression function is *not* the best exponential model.

If the exponential regression model is not the best in this sense, how is it the best and why do we use it?

To answer the second question first, the technology uses the regression procedure because it reduces to the known linear regression procedure. Finding the nonlinear function that minimizes the sum of squared errors is possible, but more involved. The non-linear regressions are the best linear fit to the corresponding re-expressed data. Thus, they provide a reasonably good model in most cases, even if it is not necessarily the best we can do.

Logarithmic Regression

Some technologies will also do logarithmic regression. This type of regression gives you the "best" logarithmic model for a given two-variable data set. These models would be of the form

Table 4: Comparison of Exponential Regression Errors with a Second Exponential Model

Year	Child Mort.	Exponential Regression	$151(.9591^x)$	Regression Errors	Errors $151(.9591^x)$	Squrd. Reg. Errors	Squrd. Err. $151(.9591^x)$
0	151	133.800	151.000	-17.200	0.000	295.840	0.000
10	82	91.583	99.453	9.583	17.453	91.842	304.591
20	57	62.687	65.502	5.687	8.502	32.342	72.285
30	45	42.908	43.141	-2.092	-1.859	4.376	3.455
35	35	35.499	35.012	0.499	0.012	0.249	0.000
40	31	29.370	28.414	-1.630	-2.586	2.658	6.687
					Sum of Err^2	427.308	387.018

$$y = a \log x + b \qquad \text{or} \qquad y = a \ln x + b$$

depending on whether the technology is using the common base 10 logarithm or the base e natural logarithm. As with exponential and power regressions, re-expression is involved. In this case, the re-expression takes the logarithm of the x or independent variable values.

Applications of Power Models

Example 6. Table 5 shows how life expectancies changed in the United States during the 20th century. These same data were also considered in Chapter Seven, where we found a logarithmic model.[112] The value given for each year represents how long a person born in that year could expect to live, on the average. Find an appropriate power model for this data, and discuss how good the fit is. Use the model to predict the life expectancy for U.S. children born in 2010 and 2050. How reasonable do these predictions seem to be?

Year	1900	1915	1930	1945	1960	1975	1990	2000
Life Expectancy	47.3	54.5	59.7	65.9	69.7	72.6	75.4	76.4

Table 5: Life Expectancy for U.S. children born from 1900 to 2000.

Solution: Figure 10 shows a scatterplot of the data. The pattern seems to be slightly concave down. This indicates that a power function with exponent between 0 and 1, as well as the logarithmic model we had previously found, might be appropriate.

To test if either is appropriate, and which one might be better, we do what is called "re-expressing the data." If the plot of Log(Life Expectancy) versus Log(Year) is approximately linear, then a power function would be a good model. If the plot of Life Expectancy versus Log(Year) is fairly linear, then a logarithmic function would make a good model. If we let the independent variable be years from 1900, however, we will have to take into account that we cannot use 0 as a value when taking logarithms. To adjust for this, we leave off the first data point and construct the plots using only the years starting with 1915, which corresponds to an x-value of 15. Both plots are shown in Figure 11, along with the linear regression line for these relationships.

Both plots seem to give a very strong linear pattern, which means either type of function should make a good model. The Log(y) versus Log(x) seems to give a slightly stronger linear relationship, so the power model might be the best. We can do both a power regression and a logarithmic regression, either using technology that does this automatically, or by re-expressing the data in the

[112] This example is adapted from *Functioning in the Real World* by Sheldon Gordon, et. al., p. 189

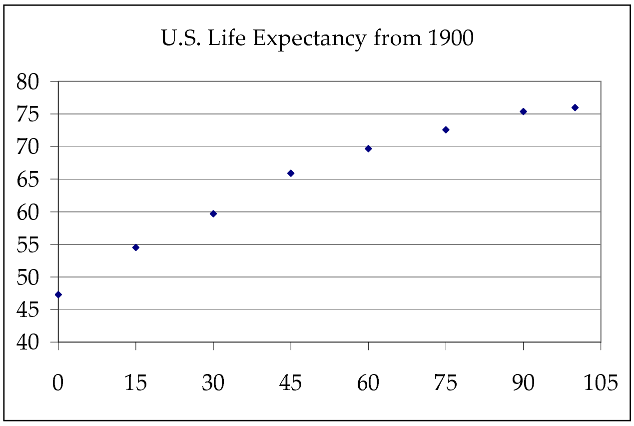


Figure 10: U.S. Life Expectancy

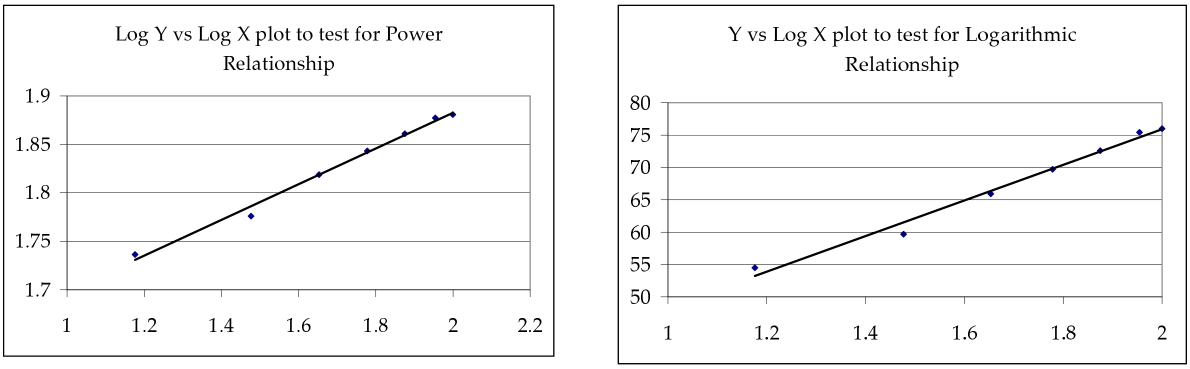


Figure 11: Plots of re-expressed life expectancy data

table and using linear regression on the re-expressed data (again, leaving out the year 1900). We get the following models:

- The power model is $y = 32.66(x^{0.1842})$
- The logarithmic model is $y = 12.05 \ln x + 20.521$

The original data with each of these models is graphed in Figure 12. Both seem to fit the data well.

For the year 2010, the power model predicts a life expectancy of 77.64, while the logarithmic model predicts 77.02. These are quite close to each other. For the year 2050, the power projection is 82.20, while the logarithmic projection is 80.73. The power function is providing a higher estimate as time goes on. This should not be surprising, as power functions, even when concave down, grow more quickly than logarithmic functions.■

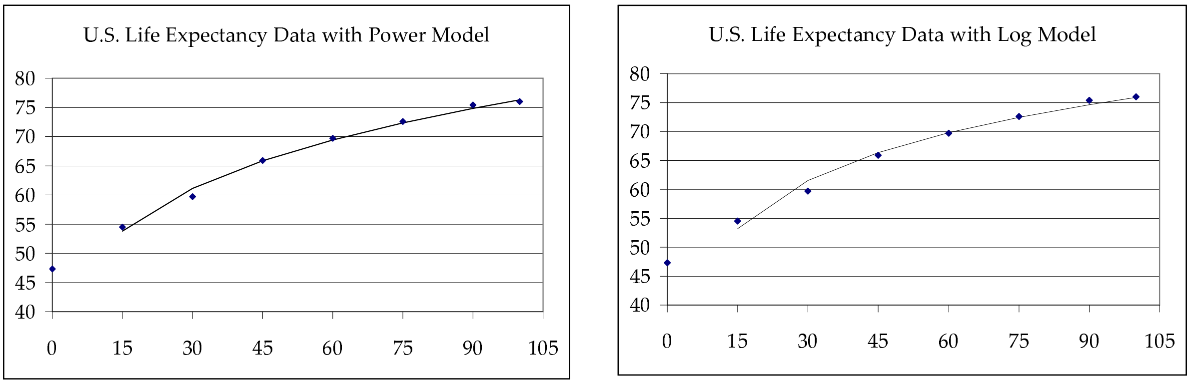


Figure 12:

Modeling Revisited

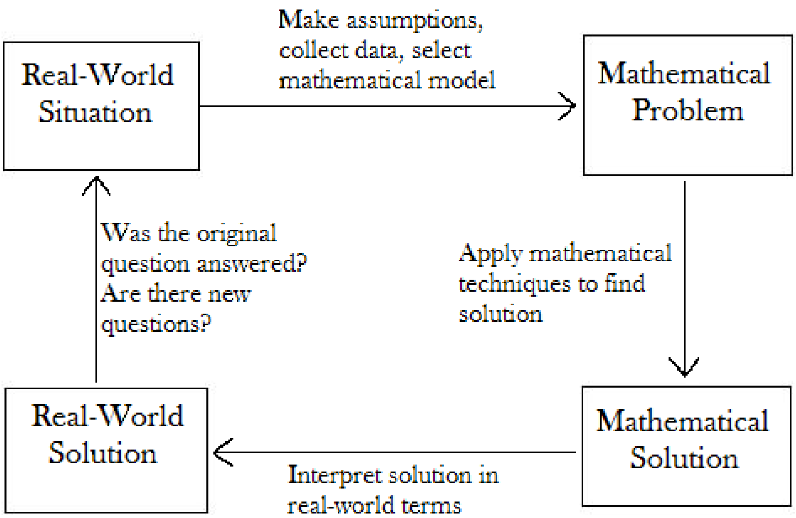


Figure 13: The Mathematical Modeling Process

In Chapter Four, *Functions*, we defined a modeling function as one that describes a real world phenomenon. However, a modeling function often only gives an approximation of the real world situation. In general, a mathematical model is a representation of a particular phenomenon which attempts to take into account the most important characteristics of that phenomenon, while ignoring irrelevant details. This means that models are not often (perhaps almost never) a completely accurate description of the phenomenon. They are simplified versions of reality.

In fact, in Chapter Five, *Linear Functions*, we commented that one reason linear functions are so important is that they are the simplest types of functions. The two quotes that began that chapter highlight the notion that one should always pick the simplest model that adequately explains the given situation. Oftentimes, one can create a better model than the current model, but this often means making the model more complicated. Inherent in all modeling is this trade-off between quick and easy, but perhaps over-simplified models, and complicated, hard, but perhaps more accurate and complete models.

In practice, models are often created using an iterative process like the one in Figure 13.

Two Kinds of Models

When we considered linear models, there were actually two different approaches we took to creating models. In Example 3 of Chapter Five, we considered a furnace repair person who charges \$20 to visit your house, and then \$30 per hour for his work. The function which models the total charge for a visit is $C(n) = 30n + 20$, where n represents the number of hours and C represents the cost. This type of model is called a **theoretical** model. It is based on known facts or assumptions about the relationship between the hours worked and the cost. We end up with a linear model because we are told the rate of change measured as cost per hour is constant.

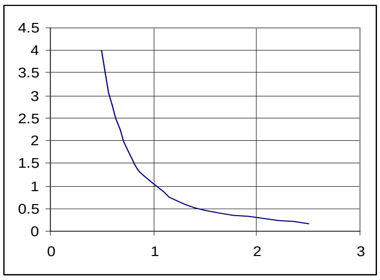
The second approach we took involved creating linear models using linear regression. For example, in Table 5 and Figure 6 in Chapter Five, we looked at data relating the Arm Span and Height of a number of college students. In this case, we started with a set of data, and did not know ahead of time whether or not the relationship between the two variables involved a constant rate of change. We simply took the data we had and found the linear equation that "best fit" the data, where best is defined by minimizing the sum of the squared errors. Regression models are one kind of what are called **empirical** model. In this chapter and the last, we used regression to find "best fit" non-linear models from a given function family for a given set of data. This is sometimes called **curve-fitting**.

As we proceed, we will have even more options for creating models for a given set of data. The choice of which family (or combination of families) to use is often more of an art than a science. We will continue to try and create "the best" models we can, but will always need to acknowledge the inherent limitations of any model.

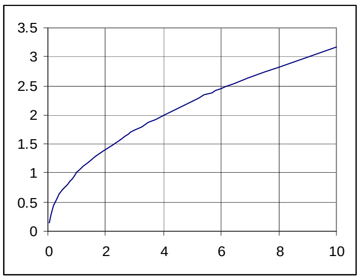
Reading Questions for Power and Quadratic Functions

The function of education is to teach one to think intensively and to think critically ... Intelligence plus character - that is the goal of true education.
— Martin Luther King, Jr.

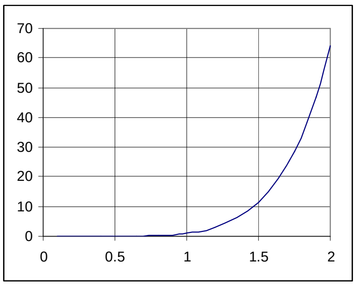
1. Below are the graphs of $y = x^{-2}$, $y = x^{-1/2}$, $y = x^{1/6}$, $y = x^{1/2}$, $y = x^2$ and $y = x^6$. Identify each, briefly justifying your answer.



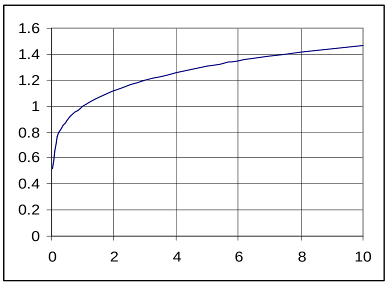
(a)



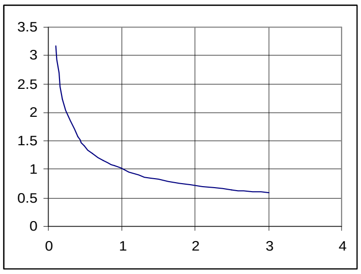
(b)



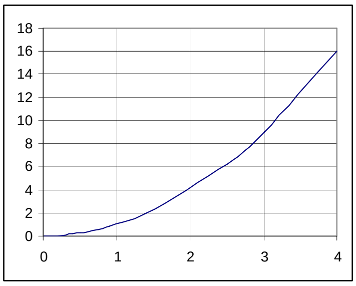
(c)



(d)



(e)



(f)

2. If you are given a power function whose graph is concave up in the first quadrant, what do you know about the exponent?

3. If you are given a power function whose graph is concave down in the first quadrant, what do you know about the exponent?

4. Let f be a power function of the form $f(x) = ax^b$.

(a) Suppose $f(1) = 5$. What is the value of a?

(b) Suppose $f(1) = 2$ and $f(2) = 8$. What is the value of b?

(c) Suppose $f(1) = 5$ and $f(2) = 190.27$ (rounded to two decimal places). What is the value of b?

5. Let f be a function whose y-intercept is 5. How do you know that this cannot be a power function?

6. Solve the following equations. Some of these are exponential equations, and some are power equations.

(a) $23 = 3.6(1.2^x)$

(b) $23 = 3.6(x^{1.2})$

(c) $22,500 = 3,000(1.005)^{12x}$

(d) $17 = 2.5(1 + x)^3$

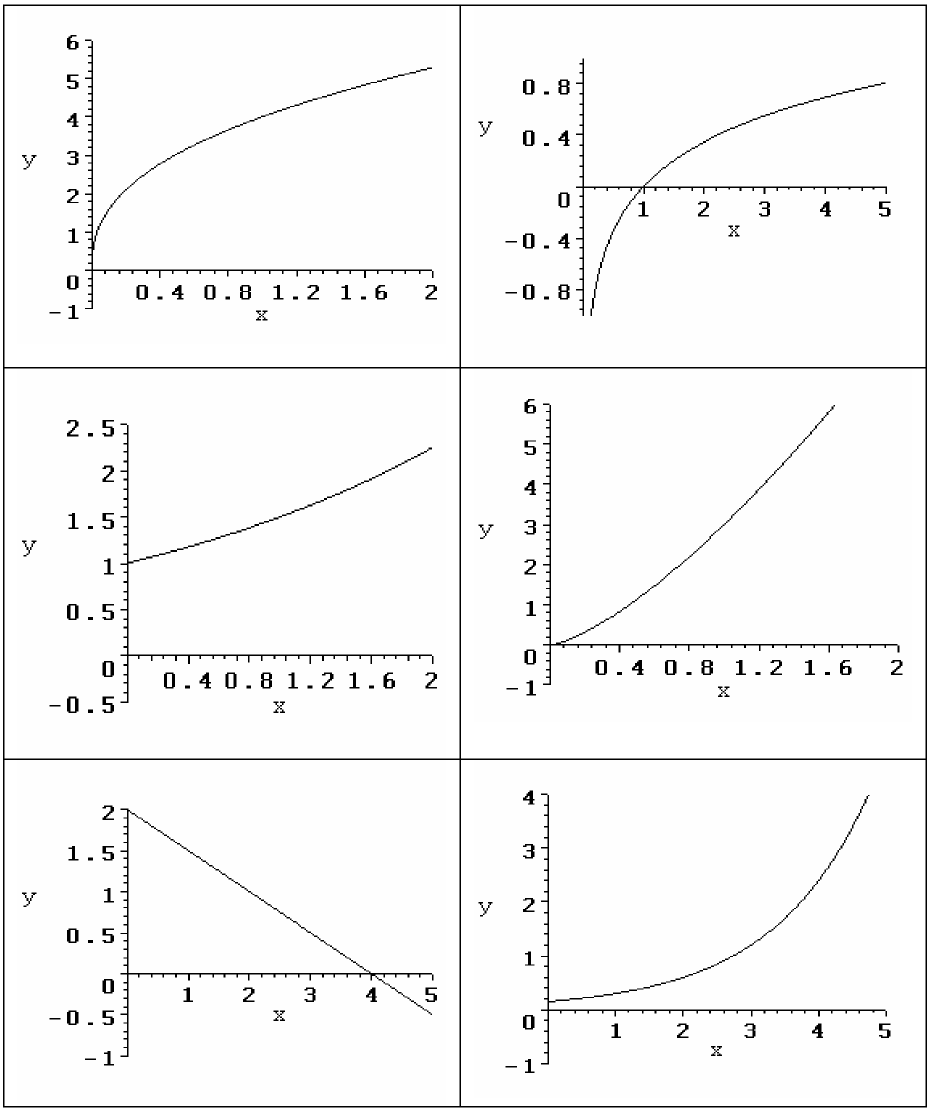


Figure 14:

7. For each of the functions given in graphical form in Figure 14, decide whether the function is linear, exponential, power, logarithmic, or none of these.

8. Solve the equations.

(a) $x^2 - 4x + 3 = 0$

(b) $x^2 - 8x + 4 = 0$

9. For each verbal description, decide which type of function, linear, logarithmic, exponential, or power, is most likely to fit the given information.

(a) The U.S. population has been growing at a rate of about .7% per year.

(b) World energy consumption has been growing at a constant rate of about 6 quadrillion BTU's per year since 1980.

(c) f is concave down and increasing, with a y-intercept of about 0.
(d) f is concave down and increasing, with an x-intercept of about 1.
(e) f is concave up and decreasing, with a y-intercept of about 5.
(f) f is concave up and decreasing, with no y-intercept.

10. In Chapter Eleven, one of the activities includes creating a model for the total amount of energy consumed worldwide starting in the year 1980. One possible model, letting $n = 0$ represent the year 1980, is

$$C(n) = \frac{n(6n + 560)}{2} = 3n^2 + 280n$$

where $C(n)$ represents an estimate of total world energy consumption from 1980 through the year represented by n in quadrillions of BTU's.

(a) Estimate the year in which the total consumption from 1980 through that year will be 10,000 quadrillion BTU's.
(b) Estimate the year in which the total consumption from 1980 through that year will be 40,000 quadrillion BTU's. (40,000 quadrillion BTU's was the U.S. Energy Department's estimate of total known energy reserves, including coal, oil, and gas, as of 2002.)

11. The table below shows total U.S. energy consumption, in quadrillions of BTU, for the years 1949 through 2002. Years are counted from 1900.

Yr	Cons.		Yr	Cons.		Yr	Cons.		Yr	Cons.
49	32.0		63	49.7		77	78.0		91	84.5
50	34.6		64	51.8		78	80.0		92	85.9
51	37.0		65	54.0		79	80.9		93	87.6
52	36.8		66	57.0		80	78.3		94	89.3
53	37.7		67	58.9		81	76.3		95	91.2
54	36.6		68	62.4		82	73.2		96	94.2
55	40.2		69	65.6		83	73.1		97	94.7
56	41.8		70	67.8		84	76.7		98	95.2
57	41.8		71	69.3		85	76.4		99	96.8
58	41.7		72	72.7		86	76.7		100	98.9
59	43.5		73	75.7		87	79.2		101	96.3
60	45.1		74	74.0		88	82.8		102	97.4
61	45.7		75	72.0		89	84.9			
62	47.8		76	76.0		90	84.6			

(a) Use logarithmic regression to find a logarithmic model for this data. Logarithmic regression gives you a model of the form $y = m \log x + b$, with parameters m and b.
(b) Graph the data along with your logarithmic model. How well does the model seem to fit the data?
(c) Is there a different family that you think would provide a better model? Why do you think so?
(d) Using the family you selected in part (c), find a model from this family that you think best fits the data. How does the fit of this model seem to compare to the fit of the logarithmic model?

12. Solve the following equations, where $f(x) = 2.76x^{1.4}$

(a) $f(x) = 0$
(b) $f(x) = 1$
(c) $f(x) = 2.76$
(d) $f(x) = 10$

13. Solve the following equations.

(a) $2x^2 - 8x = 7$
(b) $2x^3 - 4 = 9$
(c) $2(3^x) - 4 = 9$

14. Solve the following quadratic equations.

(a) $x^2 + 3x + 2 = 0$
(b) $-16t^2 + 85t + 5 = 0$
(c) $0.35x^2 - 2.4x + 4.7 = 0$
(d) $3x^2 - 2x = 7$

15. The function $h(t) = -16t^2 + 85t + 5$ represents the approximate height of a baseball thrown up in the air at 85 feet per second from a height of 5 feet.

(a) Evaluate $h(0)$, $h(1)$ and $h(2)$.
(b) Use the formula for the vertex of a parabola in the text to find the vertex of this parabola.
(c) Interpret the vertex coordinates in terms of the flight of the ball. What does the x-coordinate of the vertex tell you in this situation? What does the y-coordinate tell you?
(d) According to this model, when will the ball hit the ground?
(e) Suppose the ball was thrown from a height of 10 feet at 70 feet per second. How do you think this would change the symbolic form for $h(t)$?

16. $f(x) = ax^2 + bx + c$ is the general form for a quadratic function.

(a) What does the parameter a tell you about the graph of the function? Experiment with a few examples as necessary.
(b) What does the parameter c tell you about the graph of the function?

Power and Quadratic Functions: Activities and Class Exercises

1. **Baby Weights.** The following table gives the average weight for babies according to the National Center for Health Statistics.

Age (months)	3	6	9	12	15	18	21	24
Weight (lb)	12.3	16.2	19.1	21.1	22.8	24.0	25.0	25.8

   (a) Make a graph of weight versus age. (Your graph should have age on the horizontal axis.)

   (b) By looking at your graph, explain why a power function should do a better job of fitting these data than an exponential function.

   (c) Using power regression, find a power function that fits the data where age is the input and weight is the output.

   (d) Plot your power function along with the data. How well does your function fit the data?

   (e) The average weight of a one month old baby is 9.7 pounds. Using your power equation, what is the average weight of a one month old baby?

   (f) The average weight of a newborn baby is 7.9 pounds. What does this imply about the appropriateness of a power function for modeling baby weights over time (at least for very young infants)?

   (g) The following table gives the average weight of children ages 3 through 10.

Age (years)	3	4	5	6	7	8	9	10
Weight (lb)	32	36	40	46	51	57	63	70

   Plot these data along with your power function and the original data from part (d). In doing so, remember to convert the years to months. For example, the first new point should be 36 months with a weight of 32 pounds.

   (h) How well does your function fit the new data points from part (g)? How does this compare with the fit of the original data points?

2. **Lorenz Curves and Power Functions.** In Chapter Two, we introduced Lorenz curves and the Gini coefficient. In one of the activities in that chapter, you may have looked at the Lorenz curves for several countries, including the U.S. The graph below shows the Lorenz curve, based on the data in the table which is for the U.S. for the year 1989.

Table 6: Distribution of Income by Percentile for the U.S., 1989

Population Percentile	0	0.20	0.40	0.60	0.80	1.00
Percent of Income	0.000	0.046	0.152	0.317	0.554	1.000

The Lorenz curve is concave up, and goes through the points $(0, 0)$ and $(1, 1)$. This information we have is clearly consistent with a power function. In fact, since $(1, 1)$ is on the graph, we might use a power function of the form $L(x) = x^b$.

   (a) As a first model, find a b which gives a power function $L_1(x) = x^b$ going through the point $(0.40, 0.152)$. If this point is on the curve of $L_1(x) = x^b$, then $0.152 = 0.4^b$. Use logarithms to find b.

   (b) Find power functions L_2 and L_3 which go through the points corresponding to $x = 0.6$ and $x = 0.8$.

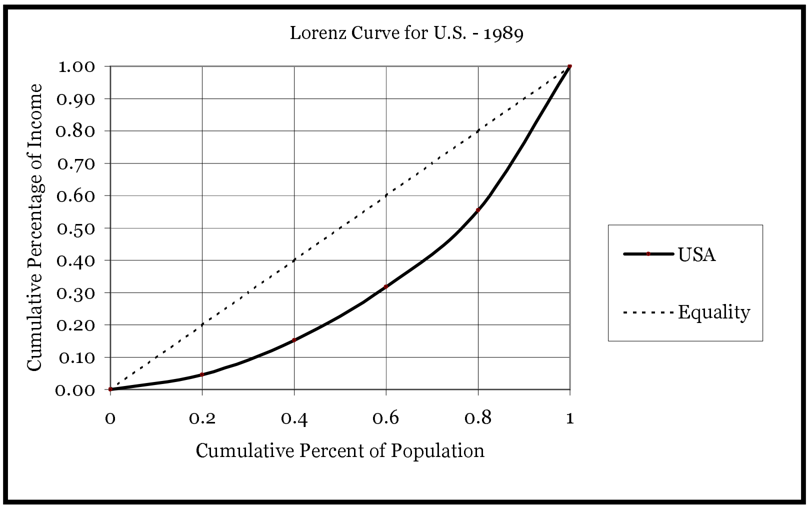


Figure 15: Lorenz Curve for U.S., 1989

(c) Graph a scatterplot of the data in the table along with all three power functions L_1, L_2, and L_3. Which function seems to be the best model?

3. Power Models for Lorenz Curves.

The table below shows Income Inequality Data for the Soviet Union, similar to the data shown in Example 5 for the United States where we were modeling the Lorenz curve for the U.S.

Table 7: Distribution of Income by Percentile for the Soviet Union, 1989

Income Percentile	0	0.2	0.4	0.6	0.8	1
Percent of Population	0.00	0.0891	0.2260	0.4050	0.6340	1.0000

(a) Find the power regression model for this data, as in Example 1.

(b) Graph your power regression model along with a scatterplot of the data. How well does your model fit the data?

(c) Graph a scatterplot of Log(Percent of Population) versus Log(Income Percentile). Does this graph seem to have a strong linear relationship? What does this tell you about the appropriateness of a power model for this data?

(d) Find a second power model, one of the form $y = x^b$. Models of this form will go through both the origin and the point $(1, 1)$. Experiment with various values of b until you feel you have the best fit you can. Do you think this model fits the data better than your first model? Why or why not?

4. Child Mortality Models.

The table and graph below show Child Mortality Rates for children under 5, in deaths per 1000 live births, for four countries from 1960 to 2001.

Year	Haiti	Kenya	Congo	Laos
1960	253	205	220	235
1970	221	156	160	218
1980	195	115	125	200
1990	150	97	110	163
1995	137	111	108	134
2000	125	120	108	105
2001	123	122	108	100

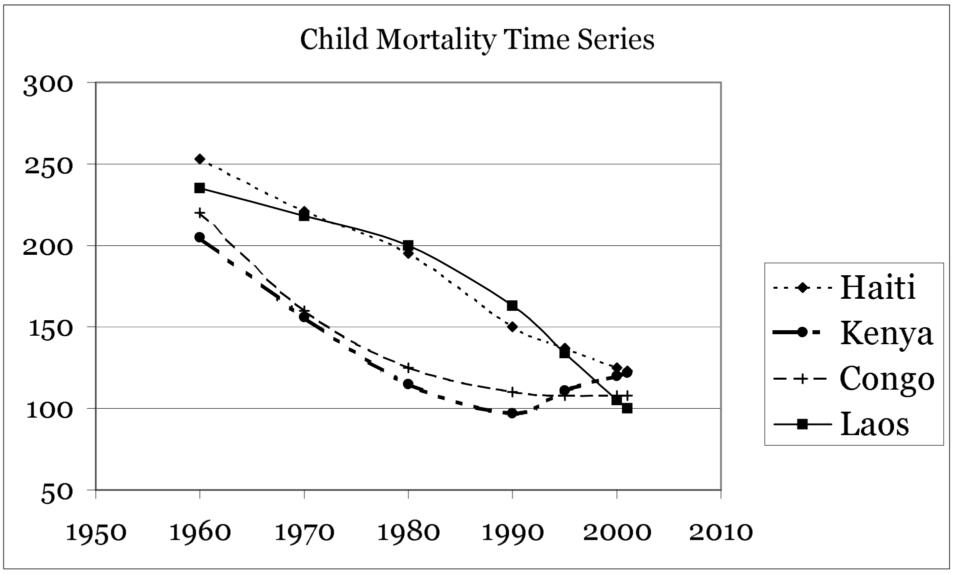


Figure 16: Childhood Mortality Time Series for Four Countries

(a) For the plots for Congo and Kenya, discuss why *none* of the four families linear, exponential, power, or logarithmic would give an ideal model. In other words, for each plot and for each of the four families, give at least one characteristic of the line plot (intercepts, asymptotes, etc.) that would not be a characteristic of the family.

(b) Which family of functions do you think would be most appropriate to model the graph for Haiti and why?

(c) For each of the four countries, give your best estimate for what the child mortality rate will be in the year 2020. One way to do this is to create models for each country, selecting an appropriate family given the scatterplot. Whatever method you use, carefully justify your estimate for each country.

5. Crude Oil Production Models. As noted in Chapter Two, fossil fuel supplies are finite. The U.S. has already passed its peak production years, and is now on the decline. Many other countries are past their peak as well. With the quadratic family of functions, we finally have a family that can model phenomenon which have this kind of peak.

The table below gives primary energy production from 1980 for three countries. This production includes crude oil, natural gas, and coal. The units are in quadrillions of BTU.

Year from 1980	Egypt	Norway	Sweden
0	1.453	3.012	0.873
1	1.481	3.065	0.988
2	1.691	3.062	0.954
3	1.841	3.422	1.071
4	2.080	3.638	1.210
5	2.233	3.787	1.309
6	2.093	3.947	1.318
7	2.299	4.392	1.404
8	2.238	4.764	1.390
9	2.320	5.718	1.391
10	2.361	5.942	1.414
11	2.396	6.230	1.408
12	2.444	7.079	1.392
13	2.547	7.268	1.373
14	2.594	7.643	1.328
15	2.673	8.351	1.390
16	2.729	9.280	1.264
17	2.600	9.588	1.397
18	2.569	9.347	1.494
19	2.673	9.531	1.458
20	2.628	10.212	1.389
21	2.782	10.217	1.543
22	2.719	10.610	1.362
23	2.705	10.402	1.211

(a) Create scatterplots of production versus year for each of the three countries.
   i. For which two of the three countries does it appear a quadratic model would fit the data well? Why?
   ii. For the third country, why does it appear a quadratic model would not fit well?

(b) For each of the two countries for which you believe a quadratic model would fit well, find the quadratic regression model for the production in this country.
   i. Graph your models together with the corresponding scatterplots. Which models seem to fit well?
   ii. For each country, use the model for that country to predict the energy production in the years 2010, 2020, and 2050. Do these estimates seem to make sense? Why or why not?

6. **Kepler's Laws:** The table below shows the period in earth days, and the average distance from the sun in millions of miles for all nine "traditional" planets.

Planet	Period	Distance
Mercury	88	36.0
Venus	225	67.2
Earth	365	92.9
Mars	687	141.5
Jupiter	4,329	483.3
Saturn	10,753	886.2
Uranus	30,660	1782.3
Neptune	60,150	2792.6
Pluto	90,670	3668.2

(a) Notice the earth is 92.9 million miles from the sun and takes 365 days to orbit the sun.

   i. Find the distance the earth travels around the sun (in miles) and then use this to find how fast the earth is traveling in miles per hour.

   ii. Find out how fast Jupiter is traveling in miles per hour as it orbits the sun.

(b) Graph a scatterplot of these data. Which family of functions do you think would be most appropriate to model these data? Explain why you think so.

(c) Find the power regression model for the data. How well does the model fit the data?

(d) Astronomers have recently been discovering large objects that may qualify as planets past Pluto. One of these has been named Quaoar (pronounced kwah-o-wahr). Quaoar's period is estimated to be about 285 earth years. Use your power model to estimate how far Quaoar is from the sun, on the average.

(e) In the 1950's, Dutch astronomer Jan Oort conjectured that there was a massive sphere of comets orbiting the sun at distances around 4 trillion miles. The existence of this so-called "Oort cloud" has since been confirmed. Using your power model, estimate the period of objects in the Oort cloud. Give your answer in Earth years.

(f) Using your answer from part (e), about how fast are objects in the Oort cloud travelling in miles per hour?

9. Transformations of Functions

Nothing has afforded me so convincing a proof of the unity of the Deity as these purely mental conceptions of numerical and mathematical science which have been by slow degrees vouchsafed to man, and are still granted in these latter times by the Differential Calculus, now superseded by the Higher Algebra, all of which must have existed in that sublimely omniscient Mind from eternity.
— Mary Somerville

The graph in Figure 1 shows the relationship between survival of female children to age 10 and the illiteracy rate for a number of countries in North and South America. Both variables are measured in percent. The trend is definitely decreasing, but it is not clear from the plot whether a linear or non-linear model would be more appropriate. There are three points which could be considered outliers.

If we were to model this data **empirically**, we might try both a linear model and a decreasing exponential model and use our own "visual judgement," correlation, and perhaps an analysis of the sum of squared errors to decide which was a better fit. Leaving out the outliers would allow us to get a much better model for the remaining points.

On the other hand, we might instead think about making some assumptions about this relationship, and use these to create a **theoretical** model, or at least a model that is partially based on theoretical assumptions.[113]

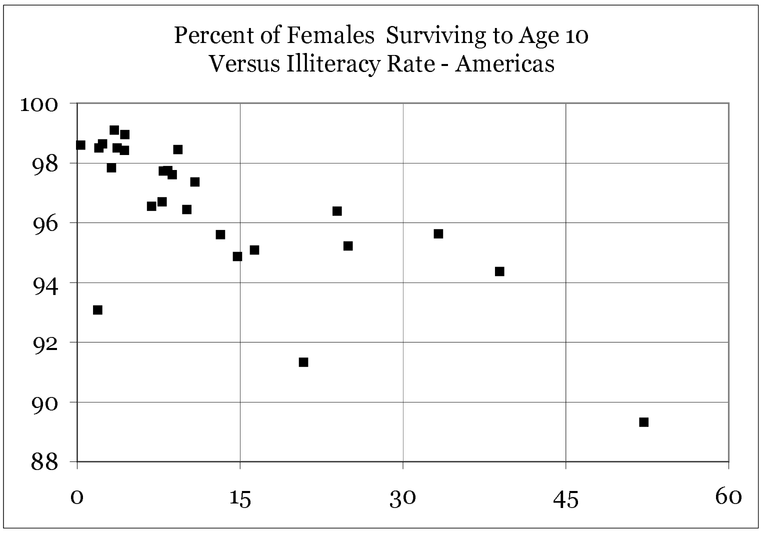


Figure 1: Survival of Female Children to Age 10 Versus Illiteracy Rate

What are some reasonable assumptions might we make?

Well, we obviously can't have a survival rate higher than 100% or lower than 0%, and the same goes for the illiteracy rate. From the scatterplot, we might assume that a 0% illiteracy rate corresponds to a 100% survival rate (or perhaps 99.9% or something close to 100%), and create our model accordingly. Also, although it is conceivable you might have countries with illiteracy rates ranging from near 0% to near 100%, it is very doubtful that the survival rate would ever get too close to 0%. We might assume that there is some **lower bound** on how low the survival rate can be, and that it will never go lower than this bound. Of the countries in our scatterplot, none has a female survival rate to age 10 lower than 89%.

[113]We discussed empirical and theoretical models briefly at the beginning of the last chapter.

Given these assumptions, we might want a model which is decreasing and concave up, and levels off at whatever lower bound seems reasonable. Of the models we have looked at so far (linear, exponential, power, and logarithmic) none would satisfy all these criteria. An exponential might work, except that decreasing exponential functions level off at a lower bound of $y = 0$. What might work is to have a model which has the same shape and characteristics of an exponential function, but levels off somewhere besides 0. We can in fact create such a model using exponential functions and what are called **transformations** of functions.

We will consider four basic types of transformations to start with. We will introduce these using graphical representations, after which we will show how to transform a function using its symbolic representation. The four types of transformations are:

1. **Vertical shifts:** The graph is moved up or down, without distorting the shape.
2. **Horizontal shifts:** The graph is moved left or right, without distorting the shape.
3. **Vertical stretches and compressions:** The graph is stretched or compressed (squashed) in the vertical direction. When this is done, points on the x-axis do not move, and points off the x-axis are moved away from (for a stretch) or towards (for a compression) the x-axis.
4. **Horizontal stretches and compressions:** The graph is stretched or compressed (squashed) in the horizontal direction. When this is done, points on the y-axis do not move, and points off the y-axis are moved away from (for a stretch) or towards (for a compression) the y-axis.

The following "fun house mirror" pictures illustrate a couple examples of these transformations.[114]

(a) (b) (c)

Figure 2: Image (a) represents an undistorted graph. Image (b) represents a horizontal compression. Image (c) represents a vertical compression.

Now, let's relate the visual effects of these transformations to what is happening symbolically.

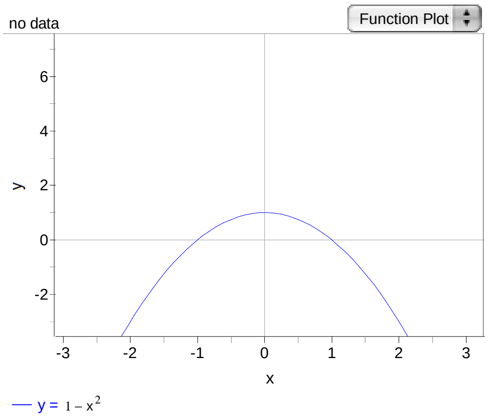
Transformations of Functions from the Symbolic Point of View

Vertical shifts: Adding a Constant to the Output

Consider $f(x) = 1-x^2$ and let $g(x) = f(x)+5$. Note that we can write $g(x)$ as $g(x) = (1-x^2)+5$. For each input x, the output or y-value for g is 5 more than the output for x. Graphically, this means that the graph of g will be the same shape as the graph for f but 5 units higher over the

[114]Original photo of *De Zwaan* taken by the author in Holland, Michigan.

whole domain. For instance, if $x = 1$, then $f(1) = 1 - 1^2 = 0$, so the point $(1, 0)$ is on the graph of f. We have, $g(1) = (1 - 1^2) + 5 = 0 + 5 = 5$, so the point $(1, 5)$ is on the graph of g. The point $(1, 5)$ lies 5 units directly above the point $(1, 0)$. Figure 3 shows graphs for both $y = f(x)$ and $y = g(x) = f(x) + 5$.



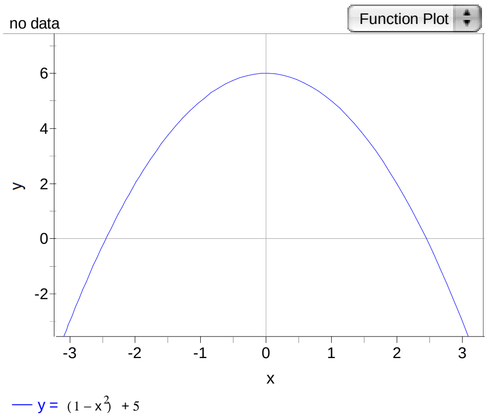


Figure 3: Graph of $y = 1 - x^2$ and $y = (1 - x^2) + 5$. Notice that the second graph has the same shape, but is shifted 5 units higher.

What would happen if we subtracted 5 instead of adding 5? The graph of $h(x) = (1 - x^2) - 5$ would have the same shape as the graph of $f(x) = 1 - x^2$ but would be shifted 5 units down. We are subtracting 5 from the outputs of the function f to get the outputs of the function h. We could write $h(x) = f(x) - 5$. Note that we could think of this as adding a negative constant instead of subtracting.

Horizontal shifts: Adding a Constant to the Input

Next, let's consider the same function $f(x) = 1 - x^2$, but let $g(x) = f(x + 5)$ which equals $1 - (x + 5)^2$. Here, we are adding 5 to the *input* of the function f instead of the *output*. In this case, the graph is shifted 5 units horizontally *to the left*. The graphs are shown in this case on the same set of axes in Figure 4.

The shift to the left is probably the opposite of what you would expect, since the positive direction on the x-axis is to the right. One way to convince yourself that left is the correct direction is to consider a particular point. The point $(0, 1)$ is the vertex of the graph of $y = f(x)$. The vertex of the graph of $g(x)$ will be $(-5, 1)$, since $g(-5) = 1 - (-5 + 5)^2 = 1$, and g evaluated at any other value of x will give a smaller output. So, the vertex of the graph of f gets moved 5 units to the left to get the vertex the graph of g.

Vertical stretches and compressions: Multiplying the Output by a Constant

Next, consider the functions $f(x) = 1 - x^2$ with $g(x) = 3f(x) = 3(1 - x^2)$. Here, we are multiplying the outputs of the function f by 3. Thus, for each input x, the output $g(x)$ is three times larger (in absolute value) than the output $f(x)$. Graphically, this means the corresponding point on the graph of g is three times further from the x-axis than the point on the graph of f. Points above the x-axis are three times higher, and points below the x-axis are three times lower. Points on the graph of f which are on the x-axis have output 0, and since $3 \cdot 0 = 0$, the corresponding points on the graph of g are still on the x-axis as well. As with vertical shifts, changing the output in this case also has an effect in the vertical direction, but in a different way. The graphs of $y = f(x)$ and $y = g(x) = 3f(x)$ are shown in Figure 5.

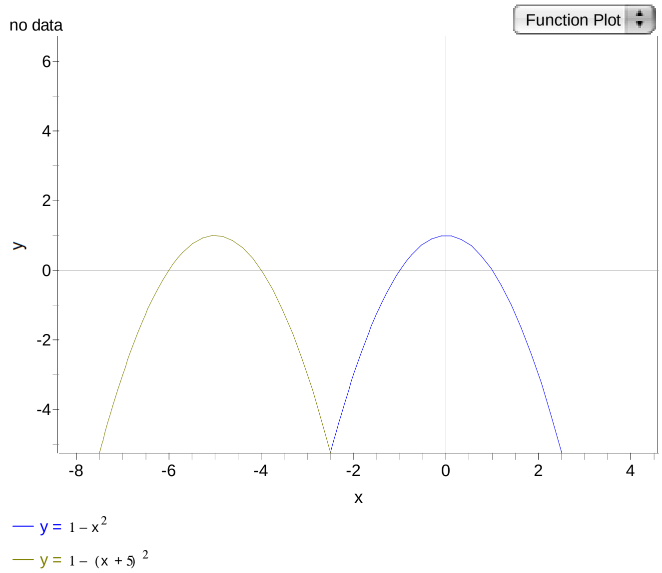


Figure 4: Graph of $y = 1 - x^2$ and $y = 1 - (x+5)^2$. Notice that the second graph is shifted horizontally 5 units to the *left*

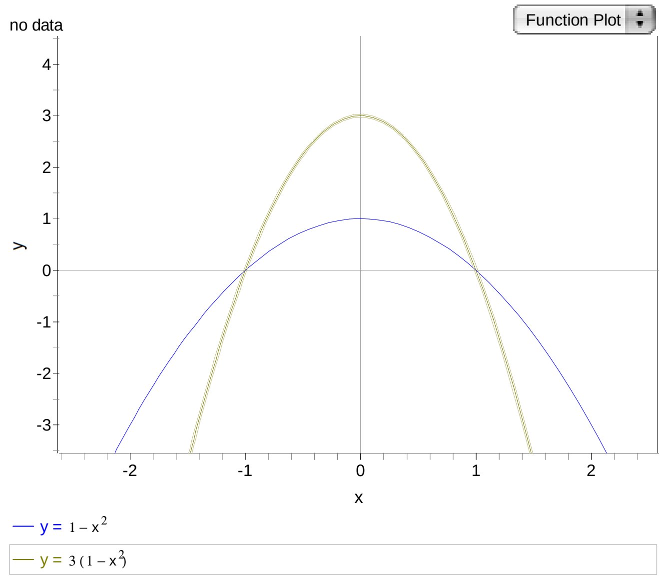


Figure 5: Graph of $y = 1 - x^2$ and $y = 3(1 - x^2)$. Notice that the second graph is vertically stretched by a factor of 3.

What do you think would happen if you multiplied the output by (1/3)?

Horizontal stretches and compressions: Multiplying the Input by a Constant

Finally, consider again $f(x) = 1 - x^2$ and let $g(x) = f(3x) = 1 - (3x)^2$. Here, we are multiplying the input of the function f by a factor of 3 to create the function g. The result is that the graph of f is compressed horizontally by a factor of 3. See Figure 6. Again, you might wonder why this does not result in a horizontal stretch.

Consider a specific point on the graph of f, say $(1, 0)$. To get the same output of 0 from the function g, we would let the input be $1/3$, since $g(1/3) = 1 - (3 \cdot 1/3)^2 = 1 - 1^2 = 0$. So the point $(1/3, 0)$ on the graph of g corresponds to the point $(1, 0)$ on the graph of f. This point is a factor of 3 closer to the the y-axis.

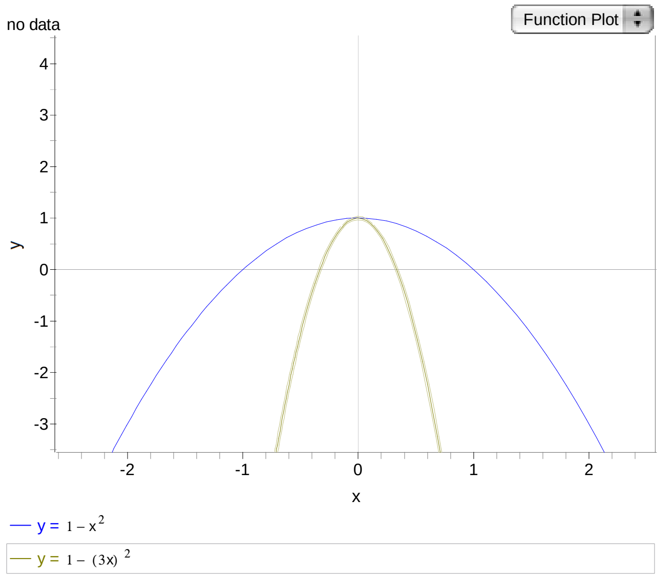


Figure 6: Graph of $y = 1 - x^2$ and $y = 1 - (3x)^2$.

To get a horizontal stretch, multiply the input by a number between 0 and 1. For example, the function $h(x) = f(.5x)$ would result in a horizontal stretch by a factor of 2 of the graph of the function $y = f(x)$.

One note on considering transformations from a visual standpoint. Remember that how the graph of a function looks depends on the window that you use. We could have created the visual impression of stretches and compressions without changing the function, but rather by changing the window in which we graph the function. As we continue to look at transformations, we just need to keep this in mind.

Summary of Shifts and Stretches

In summary, **output changes** for $y = f(x)$ include:

- $y = f(x) + c$ is the graph of $y = f(x)$ shifted up c units (assuming $c > 0$).
- $y = f(x) - c$ is the graph of $y = f(x)$ shifted down c units (assuming $c > 0$).
- $y = cf(x)$ is the graph of $y = f(x)$ vertically stretched by a factor of c if $c > 1$.
- $y = cf(x)$ is the graph of $y = f(x)$ vertically compressed by a factor of $1/c$ if $0 < c < 1$.

Input changes include:

- $y = f(x + c)$ is the graph of $y = f(x)$ shifted to the **left** c units (assuming $c > 0$).
- $y = f(x - c)$ is the graph of $y = f(x)$ shifted to the **right** c units (assuming $c > 0$).
- $y = f(cx)$ is the graph of $y = f(x)$ horizontally compressed by a factor of c if $c > 1$.
- $y = f(cx)$ is the graph of $y = f(x)$ horizontally stretched by a factor of $1/c$ if $0 < c < 1$.

Application: Child Mortality

The graph in Figure 7 shows the Under Five Child Mortality Rate for three African countries, starting in 1960. The units for the dependent variable are in deaths per 1000 live births. The good news is child mortality is down in all three countries. All the graphs are concave up and decreasing. This would suggest an exponential curve. However, the graphs seem to be leveling off at mortality rates significantly higher than zero. An exponential curve would level off at $y = 0$.

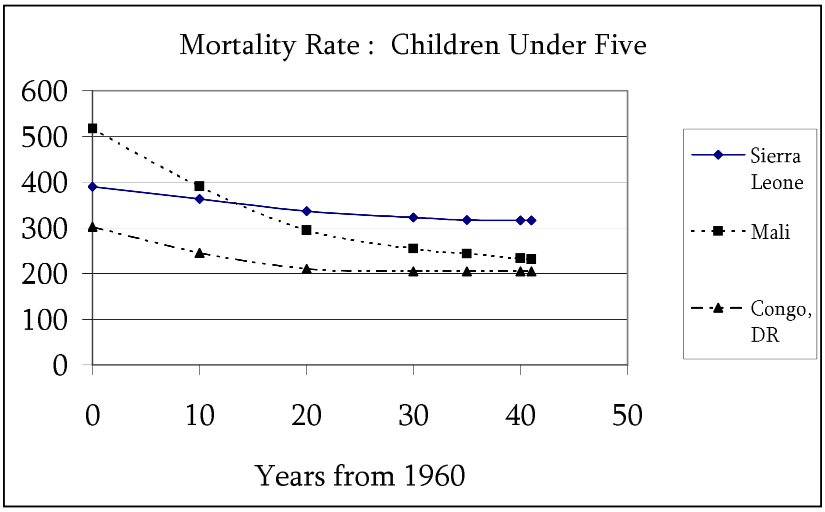


Figure 7: Mortality for Children Under Five for Three African Countries

We cannot know, of course, if the mortality rates will actually level off or not, and even if they do, we cannot know at exactly what level. So, we will make an assumption that seems reasonable and go from there. A mortality rate of 200 seems a reasonable first assumption, at least for the Congo and Mali. Sierra Leone seems to be leveling off at a higher mortality rate, or it could be that it will just take longer to get down to a rate of around 200. If this is a reasonable assumption for each of the three graphs, then exponential models shifted 200 units up should give us a reasonable fit for all three. Recall that an exponential model is of the form $y = ab^x$. So, an exponential model shifted 200 units up would be of the form $y = ab^x + 200$.

To test whether our assumption is reasonable, we can **re-express** the data, taking into account the shift we want to do. If we subtract 200 from all the mortality rates data values, this will have the effect of shifting the mortality line graphs down 200 units. This will give values which "should" level off at 0 instead of 200, if our assumption is correct. We could then investigate whether an exponential model provides a good fit for the shifted data.

To find models, we could do exponential regression on the "shifted" data and then shift the resulting function up 200. Table 2 gives the shifted data for Sierra Leone and Mali.

For Sierra Leone, the exponential regression on the shifted data is $y = 183(0.988)^x$ with $r = 0.981$. The final model for the original data would then by $y = 183(0.988)^x + 200$. For Mali, the final model, using exponential regression and then shifting up by 200 is $y = 316(0.944)^x + 200$. For

Years from 1960	0	10	20	30	35	40	41
Sierra Leone	190	163	136	123	117	116	116
Mali	317	191	95	54	43	33	31

Table 1: Shifted data for Sierra Leone and Mali (Mortality minus 200)

Mali, $r = 0.999$. These models are graphed along with the original data in Figure 9. Both models seem to give a very good fit, with the model for Mali even better than that for Sierra Leone.

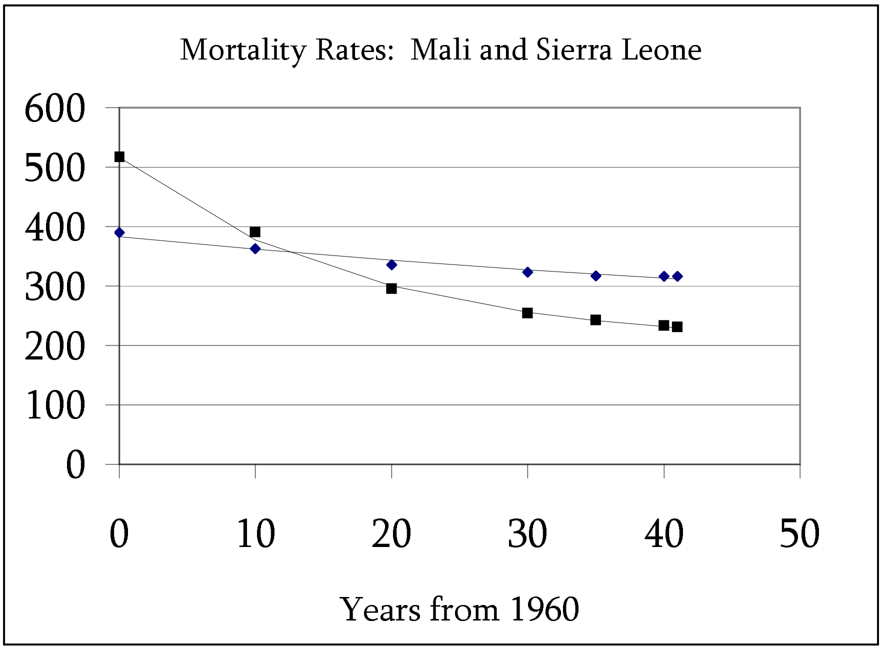


Figure 8: Mortality for Children Under Five with Shifted Exponential Models: Mali and Sierra Leone

Could we have found even better models by making a different assumption about where the mortality rates will level off? In the case of Mali, probably not. In the case of Sierra Leone, we probably could get a better fit of the existing data by assuming a higher mortality rate as the leveling off point. In fact, by shifting the Sierra Leone data down 300 instead of 200, and following the same steps above, we get a model $y = 91(0.956^x) + 300$ with correlation on the regression of 0.994, versus the 0.981 we achieved previously. This may not seem like a huge improvement, but if you compare the scatterplot of the data with both models on your own, you should see that the second model does seem to give a better fit.

One final word of caution. As with all empirical models, we must be aware that the predictions the model provides are subject to error, and that we should be less confident of the predictions the further into the future we go. The model for Mali may accurately predict what happens in the early 21st century or it may not. However, if mortality rates decrease more rapidly than the model predicts (or if they reverse the trend and start to increase), we will at least have a basis for comparison. We might then ask "what, if anything, has changed in Mali that might explain the difference between our predictions and reality?"

Reading Questions for Transformations of Functions

1. Match each of the following function transformation operations with the correct graphical transformation.

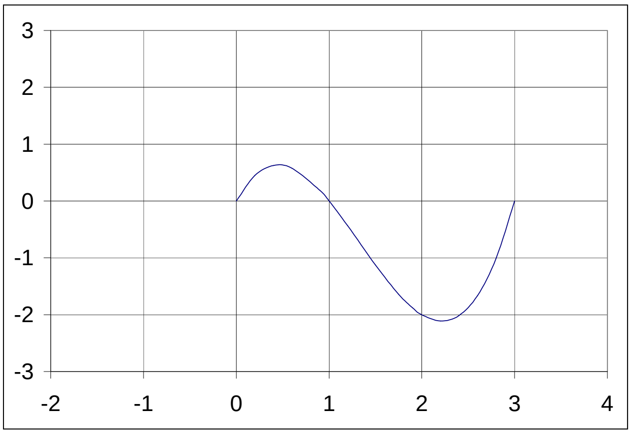
A. $f(x+c)$ (i) vertical stretch

B. $cf(x)$ (ii) horizontal shift

C. $f(x)+c$ (iii) horizontal stretch

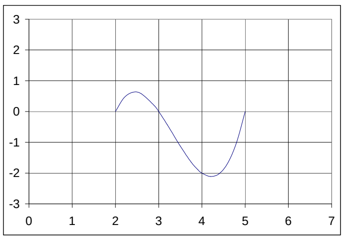
D. $f(cx)$ (iv) vertical shift

2. Given the accompanying graph of $y = f(x)$, match the following function transformations to the following graphs.[115]

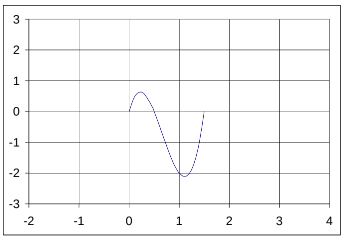


(a) $y = f(2x)$ (b) $y = 2f(x)$ (c) $y = f(\frac{1}{2}x)$

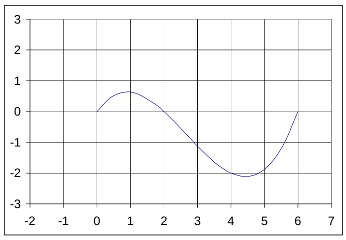
(d) $y = f(x) + 2$ (e) $y = f(x-2)$ (f) $y = f(x+2)$



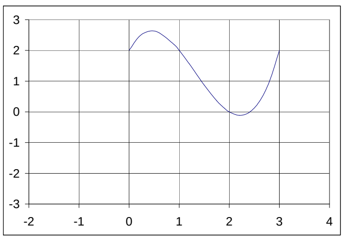
(i)



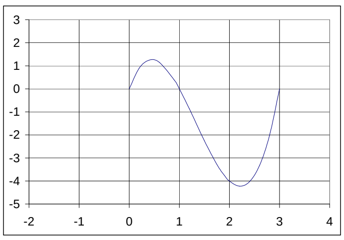
(ii)



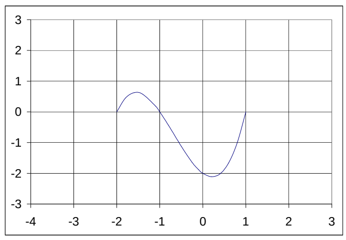
(iii)



(iv)



(v)



(vi)

[115] Exercise adapted from Andersen and Swanson

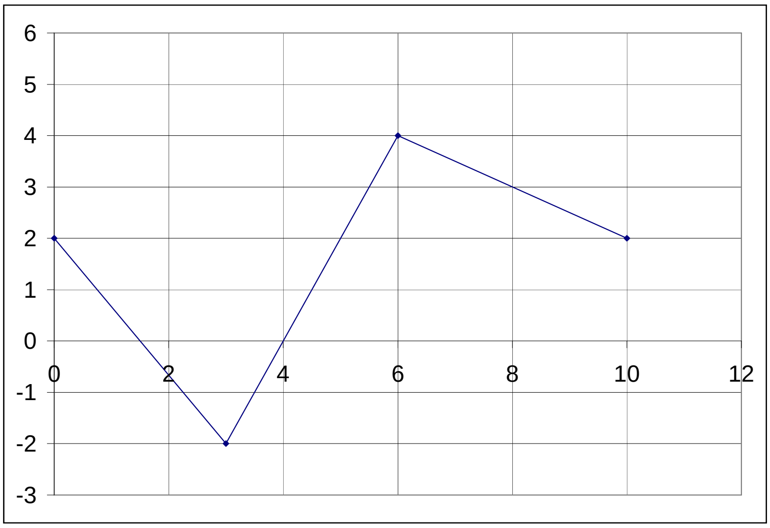
3. Suppose that $t(c)$ represents the function in which c is the number of credits you are taking this semester and $t(c)$ is the cost of tuition. Suppose a change was made in the cost of tuition such that $100 + t(c)$ describes the new tuition function. This new function describes which of the following situations?

(a) The number of credit hours has increased by 100.
(b) The cost per credit has increased by \$100.
(c) Tuition has increased by 100%.
(d) The total cost of tuition has increased by \$100 (e.g. a technology fee was added).

4. Which of the following type(s) of transformations will *always* result in all the points on the graph moving. (Circle all that apply.)

(a) adding a constant to the output
(b) multiplying the output by a constant
(c) adding a constant to the input
(d) multiplying the input by a constant

5. Let $y = f(x)$ be given below. Sketch a graph and give the **domain and range** for each of the following functions. Be sure to label the coordinates of the two endpoints and the two "corner" points on each graph.



(a) $y = f(x) - 1$
(b) $y = f(x - 1)$
(c) $y = 2f(x)$
(d) $y = f(x) + 2$
(e) $y = \frac{1}{2}f(x)$
(f) $y = f\left(\frac{1}{2}x\right)$
(g) $y = f(2x)$

6. Suppose that $y = t(c)$ represents the function in which c is the number of credits you are taking this semester and $y = t(c)$ is the cost of tuition. If the cost per credit increases by 10% (which means the cost of tuition is 110% times last year's cost), what function represents the new cost of tuition?

(a) $y = t(1.1c)$ (b) $y = t(c + 1.1)$ (c) $y = t(c) + 1.1$ (d) $y = 1.1t(c)$

7. Suppose a navy submarine was exercising tactics by beginning at sea level, diving to -300 feet in $\frac{1}{2}$ hour, then climbing to -100 feet in $\frac{1}{2}$ hour. It then dove to -500 feet in 1 hour where it rested for 2 hours. Let $y = s(t)$ be the function where t is time and $s(t)$ is the distance of the submarine from the surface.

(a) Sketch a graph of $y = s(t)$. Let the horizontal axis represent sea level.

(b) Sketch a graph of $y = s(\frac{1}{2}t)$. What did the submarine do differently for this graph?

8. Suppose a person left an office, walked slowly down a hall at a constant rate, stopped to talk to someone for a few seconds, and ran back to the office when the phone rang. Sketch a graph of this event where time is the input and the distance from the person to the office door is the output.

Transformations of Functions: Activities and Class Exercises

1. **Changing units and Transformations.**

   (a) In the Chapter 2 activities, we considered the function $w(h) = 1.4h - 20$ which is a linear model for the average weight, in pounds, of a child whose height (or length) is given in inches. As part of that activity, we also considered the functions:

   i. $y = 0.45w(h)$ which gives the weight in kilograms as a function of the height in inches. Let's denote this function $k(h)$. What kind of transformation, applied to $w(h)$, gives the function $k(h)$?

   ii. $y = w(\frac{c}{2.54})$ which gives the weight in pounds as a function of the height c in centimeters. Let's denote this function $m(c)$. What kind of transformation, applied to $w(h)$, gives the function $m(c)$.

   iii. $y = .45w(\frac{c}{2.54})$ which gives the weight in kilograms as a function of the height in centimeters. Let's denote this function $G(c)$. Describe how $G(c)$ is created from $w(h)$ via transformations.

   (b) In Chapter Eight, one of the activities involved creating a power function to model baby weights, using the data below, as a function of the time in months.

Age (months)	3	6	9	12	15	18	21	24
Weight (lb)	12.3	16.2	19.1	21.1	22.8	24.0	25.0	25.8

   The power regression function for these data is $w(x) = 8.48x^{.36}$, where w is the weight in pounds and x is the age in months.

   i. Suppose you want a function to model baby weights as a function of the time in *years* instead of months. Use a function transformation of the power regression function w to create this new function.

   ii. Suppose you want a function to model baby weights in *kilograms* instead of pounds. Use a function transformation on w to create this function.

   iii. Suppose you want to model baby weights as a function of the time in months, but starting at *conception* instead of at birth. Assuming a nine month pregnancy, use the function w to create this new function.

   (c) Earlier in this chapter, we created shifted exponential models for child mortality rates for Mali and Sierra Leone where the independent variable was the time in years from 1960. The model for Mali was $y = 316(.944)^x + 200$.

   i. Use a transformation to transform this model into a model which gives the child mortality rate for Mali as a function of the *actual* year instead of the year from 1960.

   ii. To check your model from part (i), graph your model together with a scatterplot of the data, but using the *actual* year as the independent variable in your scatterplot instead of the year from 1960.

   iii. We could have used the actual year as our independent variable instead of the year from 1960 when we created our child mortality model for Mali. Discuss what difference this would have made. Discuss any advantages or disadvantages of using the actual year versus using the year from 1960?

2. Child Mortality and Literacy.

In Figure 1 of this chapter, the relationship between illiteracy and childhood mortality for females as measured by survival rates to age 10 was depicted in a scatterplot. The data on which this graph was based is given below.

Country	Illiteracy Rate	Pct. Newborn Fem. Surviving to Age 10	Country	Illiteracy Rate	Pct. Newborn Fem. Surviving to Age 10
Barbados	0.3	98.6	Panama	8.8	97.6
Guyana	1.9	93.1	Jamaica	9.3	98.4
Uruguay	2.0	98.5	Ecuador	10.1	96.4
Trinidad & Tobago	2.4	98.6	Mexico	10.9	97.4
Argentina	3.2	97.8	Brazil	13.2	95.6
Cuba	3.4	99.1	Peru	14.8	94.9
Bahamas, The	3.7	98.5	Dominican Rep.	16.3	95.1
Costa Rica	4.4	98.4	Bolivia	20.8	91.3
Chile	4.4	98.9	El Salvador	24.0	96.4
Belize	6.9	96.5	Honduras	25.0	95.2
Paraguay	7.8	96.7	Nicaragua	33.3	95.6
Venezuela	8.0	97.7	Guatemala	38.9	94.4
Colombia	8.4	97.7	Haiti	52.2	89.3

(a) As in the Childhood Mortality example, use the following steps to create a shifted exponential model for these data.

i. Assume the scatterplot will level off at $y = 88$.

ii. Input the data in a calculator or appropriate software package.

iii. Add a column re-expressing the dependent variable by subtracting 88 from each survival rate.

iv. Find the exponential regression model with the illiteracy rate as the independent variable and the re-expressed survival rates as the dependent variable.

v. Shift your exponential model up by 88 to create a model for the original data.

(b) Graph your shifted exponential model together with the original data. How well does your model fit?

3. Child Mortality in Albania. The table below gives data on child mortality in Albania. A graph is also provided. The scatterplot is decreasing, and the trend seems to be leveling off towards zero over time. The general shape is concave up. All these indicate that an exponential model with base between 0 and 1 might be appropriate.

(a) Find the exponential regression model for this data, and graph the data and the model together. How confident are you that the model will make reasonable predictions for child mortality for years after 2000?

(b) As in the application on child mortality in this chapter, create a shifted exponential model for the Albanian data, assuming that the mortality rate will level off at $y = 20$. Give the correlation coefficient that you get from the exponential regression on the shifted data, along with the equation for the shifted exponential model.

(c) Plot this shifted model together with the data. Does this model seems to fit the data better than your first model?

Year from 1960	Child Mortality
0	151
10	82
20	57
30	45
35	35
40	31

Table 2: Child Mortality as a function of Year since 1960

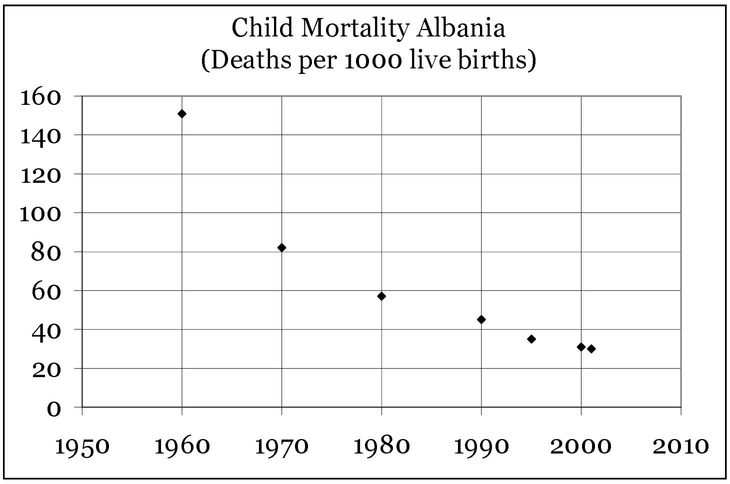
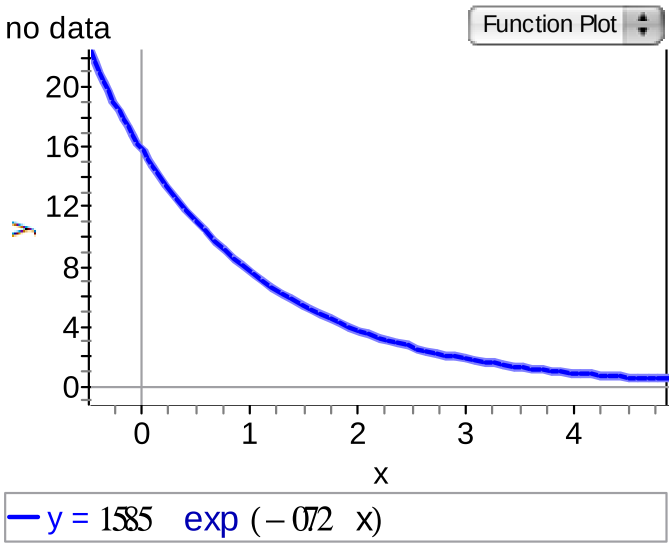


Figure 9: Child Mortality Rates for Albania over time

(d) Using both of your models, predict Child mortality in Albania for the year 2020 and 2050. Which predictions are you most confident in and why?

4. **Cost of Computer Storage.** In 1956, one IBM hard drive could store 5 megabytes of information at a cost of $50,000. In 1981, *Creative Computing* magazine predicted "the cost for 128 kilobytes of memory will fall below US $100 in the near future." [116] This would represent about $800 per megabyte. The graph below shows an exponential model, $y = 15.85(.487^x)$, approximating the cost for 1 gigabyte (1000 megabytes) of storage where x represents the year from January 1st, 2000. Note that this model predicts one gigabyte cost $15.85 on January 1st of 2000, and that eventually storage costs will become close to $0 per gigabyte. The equation for this graph is $y = 15.85(.487^x)$.

[116] http://www.alts.net/ns1625/winchest.html accessed on December 30th, 2007



(a) The graph is decreasing and concave up. Carefully explain what both of these mean for the cost of computer storage.

(b) Suppose one gigabyte started out at a price of \$15.85 in 1998 rather than 2000. Assuming the original change in price remains the same (in other words, that the exact shape of the graph is not changed), identify the transformation this implies (vertical/horizontal stretch/shift, etc.) and find the equation for this new function.

(c) Suppose computer storage started out in 2000 at the higher price of \$20. Assuming the original change in price remains the same, what transformation is this and what is the equation of this new function?

(d) Suppose both things in parts (b) and (c) occurred, so that computer storage started out in 2000 at a price of \$20 per gigabyte.

(e) Assuming the original rate of change of price remains the same, give the equation of the new function.

(f) Suppose we write the original function in function notation as $f(x) = 15.85(.487^x)$. Also, suppose that the price of computer storage had instead decreased according to the function $y = f(1.5x)$. What type of transformation is this?

(g) According to the original model, what should the price of a gigabyte of storage be today, and when will the price reach 1 cent per gigabyte?

5. Newton's Law of Cooling. Suppose it's a cold winter morning, and you have made yourself a nice cup of hot chocolate. You set it down on a table when the phone rings and subsequently forget about it. It will, of course, cool off over time, eventually getting down to room temperature.

(a) By hand, draw a reasonable graph of the temperature of the hot chocolate versus time. Include scales on both the horizontal and vertical axis. You might use minutes for the horizontal axis and degrees Fahrenheit (or Celsius) on the vertical. Consider carefully the concavity of the graph.

(b) What family of functions would seem to be most appropriate for modeling your graph from part (a)? You might want to consider a transformation of one of the basic families we have considered.

(c) Assume that your hot chocolate was at 150° Fahrenheit when you set it on the table, and that the room containing the table is at 70° Fahrenheit. If you were to use a transformed exponential function of the form $y = ab^t + c$ to model the temperature y

as a function of the time t, what could you say about the parameters a, b, and c based on this information? (Hint: You will not be able to exactly determine the value of one of these parameters, but should be able to provide partial information for it)

(d) Assume the same information as in part (c), and in addition, suppose that the temperature of the hot chocolate was 120° degrees after 10 minutes. This is enough information for you to determine all three parameters a, b, and c. Find all three parameters.

(e) Later that day, you lift some weights in the gym. Afterwards, you are hot and decide to have a cold drink. You buy a can of pop which has a temperature of 45° Fahrenheit, and open it, at which time the fire alarm goes off and you have to leave the building. You leave your pop sitting in the weight room, which is at a temperature of 78° Fahrenheit.

i. How would a graph of the temperature of the pop versus time be different than in the case of the hot chocolate? How would it be similar? Consider both the concavity and whether the graph is increasing or decreasing.

ii. You could use a transformed exponential function of the form $y = ab^t + c$ to model the temperature y of the pop over time. Use the information you have to determine the parameters a and c. (Hint: a should be negative)

iii. Do you have enough information to find b for this warming pop example?

iv. Assume $b = .8$, and graph the equation $y = ab^t + c$ using the parameters a and c you have already found for the pop. You should get an increasing and concave down graph. Graphically, you have an exponential function but "flipped upside-down", since it is concave down instead of concave up. This type of transformation is called a **reflection over the x-axis**. Symbolically, to reflect the graph of a function $f(x)$ over the x-axis, which holds the points on the axis fixed and flips everything else over the axis, you would graph $y = -f(x)$.

10. Other Families of Functions

How much space do you need?

As of January of 2004, there were about 6.2 billion people in the world. The total land area of the world is about 57.4 million square miles, with 5.1 million of that in the Antarctic. Another roughly 7 million square miles is under water of one form or another (glaciers, lakes, rivers, etc.), leaving about 45 million square miles of "exposed land."

Of course, a significant portion of the Earth's exposed land is uninhabitable or cannot be used for any agricultural purpose due to arid or mountainous terrain, or year-round frigid conditions. In fact, only around 13 million square miles is arable. Currently over half of this area is not under cultivation.

How much land does this leave for each of us to "live on?"

If we measure population density in persons per square mile, using the total amount of exposed land, we get

$$\frac{6.2 \times 10^9}{45 \times 10^6} = \frac{6.4}{45} \times 10^3 \approx 142 \text{ people per square mile}$$

This would give each person about 200,000 square feet. This is about 4 football fields. About one and one-half of these football fields would be arable land.

Is this enough land for a person to live on?

Keep in mind that not only do we need actual living space, we need land to provide many of the other necessities of life, most notably food, but also energy, water, clothing, material for shelter, etc.

Whether this 200,000 square feet is enough is a complicated question. However, it is reasonable to assume that, even with improved agricultural techniques and more advanced technology, there is some upper limit on how many people the earth can reasonably support. This implies that if we wish to model world population in the long term, we want a model that "levels off" at some point. In mathematical terms, we want a function which has a horizontal asymptote $y = M$, where M is the positive value representing the upper limit on human population. What kinds of functions can have such a horizontal asymptote?

Families of Functions So Far

Although we have considered a variety of families of functions and have used transformations to adjust the basic families to create more and more accurate models, all of our functions have been really very simple functions in two important ways. First, with only one exception, all of our symbolic functions have been either *always* increasing or *always* decreasing. We have not used a symbolic function which "changes direction," except for power functions where the power is an even integer.[117] Mathematicians call functions with no change from increasing to decreasing or vice versa **monotonic**. Second, with one exception, all of our symbolic functions have been either *always* concave up or *always* concave down. Mathematicians would say that such functions have no **inflection points**, that is, points where the concavity of the function changes from concave down to concave up, or vice versa. The only exceptions to this were the power functions where the power is an odd integer (eg. $y = x^3$). These have an inflection point where $x = 0$.

We can create increasing functions which "level off" at a positive horizontal asymptote using transformations of exponential functions or power functions with negative exponents. In each of

[117]Such functions change direction at $x = 0$, like $y = x^2$.

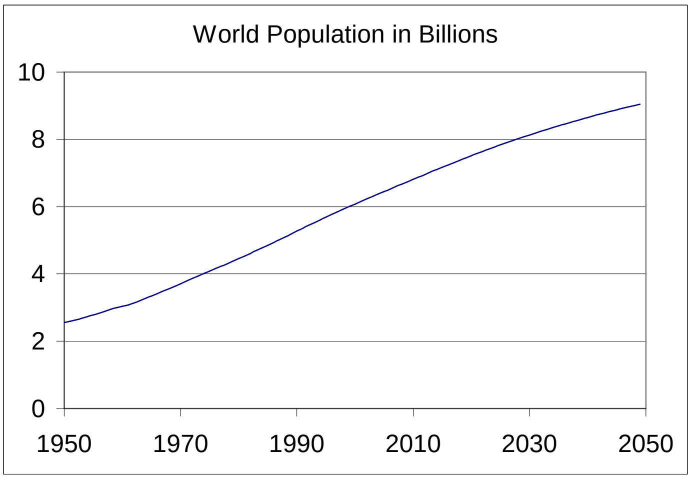


Figure 1: World Population from 1950, projected to 2050

these cases, the model would be monotonic increasing. Can we create models for situations like world population, shown in Figure 1? In both of these cases, the graphs of the data show a change in concavity as well as a leveling off over time.

What about data which shows both increases and decreases, and other more complicated behavior, like Figure 2 of Chapter Two, which shows total world energy production?

In this chapter, we will look at a few other families of functions that we can use to model data that has more complicated behavior. This will certainly not be a comprehensive treatment. However, it will give you a few more families to use as tools. The good news is that there are several standard models which we can use fairly easily. The trade off is that the more complicated the model, the fewer will be the situations where it gets used. Also, we need to be careful not to deceive ourselves that a more complicated model always means a better model.

Logistic Growth Models

The family of logistic functions can provide reasonable models for quantities that increase over time, but where there is or is assumed to be a maximum value beyond which the quantity will not go. The basic shape of logistic curves is an elongated "S" shape, similar to the graph of urban population in Botswana.

The Number e

There are several ways to express the symbolic form for logistic functions. The most usual expression involves the number e, which we introduced in Chapter Six. Also, recall that when we introduced logarithms, we said

$y = \log_b x$ means the same as $b^y = x$

and that when $b = 10$, we simply write $\log x$. If $b = e \approx 2.71828$, we write $\ln x$ instead of $\log_e x$.

Now, the symbolic form for a logistic function $L(t)$ is

$$L(t) = \frac{M}{1 + be^{-rt}}$$

We are using t to denote the independent variable since we most often use logistic functions to model data where the independent variable is time. In this expression, M, b, and r are the parameters. These parameters have the following interpretations:

(1) M represents the *limiting value* of the function. For logistic functions which are increasing,[118] M is the upper limit of growth, beyond which the function will not increase.

(2) r represents the approximate percentage growth rate in $L(t)$ when $L(t)$ is much smaller than M.

(3) $b = \frac{M}{L(0)} - 1$, where $L(0)$ represents the **initial amount** for the quantity L.

(4) Increasing logistic growth curves have an inflection point, where the concavity changes from concave up to concave down, at the point where $L(t) = \frac{1}{2}M$. This occurs when $t = \frac{ln(b)}{r}$.

Example 1. It is possible to make yogurt yourself, using commercial yogurt (as long as it contains live yogurt cultures) and milk. Simply put a few tablespoons of yogurt in a jar of warm milk (about 105 degrees Fahrenheit) and maintain the temperature for about 4 hours or so. The yogurt organisms will eat the milk and you will end up with a jarful of yogurt (and no milk), and a small amount of waste product (called *whey*, as in what Little Miss Muffet ate with her curds).

Suppose we let $L(t)$ be the amount of yogurt measured in ounces in a 32 ounce container at time t, measured in hours, during a yogurt making process. Assume the initial amount of yogurt is 1 ounce and that the percent increase per hour in the amount of yogurt for small amounts of yogurt is $r = 1.8$ which represent a 180% growth rate. Assuming L is a logistic growth function, find the symbolic representation for L. When does the model predict that there will be 31.5 ounces of yogurt?

Solution: Under the assumptions given, we have $r = 1.8$, $L(0) = 1$ ounce, and $M = 32$ ounces. We can find b using (3) above. We have $b = \frac{M}{L(0)} - 1 = \frac{32}{1} - 1 = 31$. So, our logistic model is

$$L(t) = \frac{32}{1 + 31e^{-1.8t}}$$

To find out when we get 31.5 ounces of yogurt, one way to proceed would be to graph $y = L(t)$ and then estimate the t-value when we reach $y = 31.5$. Try this on your own graphing technology and you should get about 4.2 hours. The graph is shown in Figure 2.

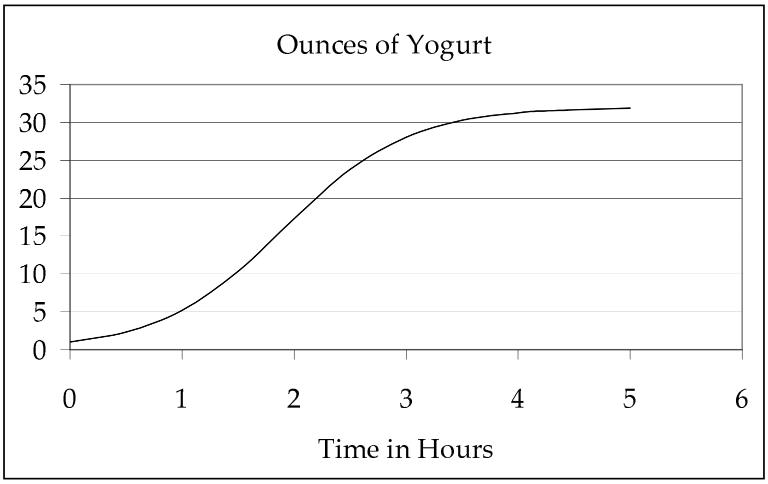


Figure 2: The Growth of Tomorrow's Breakfast

Alternatively, we could solve the equation

[118]logistic functions can be decreasing as well

$$31.5 = \frac{32}{1 + 31e^{-1.8t}}$$

for t. We can do this with some algebra and logarithms. We have,

$$31.5 = \frac{32}{1 + 31e^{-1.8t}}$$

$$1 + 31e^{-1.8t} = \frac{32}{31.5}$$

$$31e^{-1.8t} = \frac{32}{31.5} - 1 \approx .015873$$

$$e^{-1.8t} \approx \frac{.015873}{31} \approx .000512$$

Using the natural logarithm, we can change this to exponential form, getting $\ln .000512 = -1.8t$ or

$$t \approx \frac{\ln .000512}{-1.8} \approx 4.21$$

very close to the estimate given above.■

Explanation of the Parameters

Notice in Figure 2 that the growth curve does have an inflection point at about $L = 16$, which is half of the limiting value of $M = 32$, illustrating property (4) above. This is where the yogurt population is growing most rapidly. Remember that when a function is concave up, this means that the *rate of change* of the function is increasing, and when a function is concave down, the rate of change is decreasing. Logistic curves, at least those where $L(0) < \frac{1}{2}M$, will have increasing rates of change until the quantity L reaches half of the limiting value, after which the rate of growth will decrease.

Next, let's explain properties (1), (2), and (3). You will be asked to consider property (4) in the Reading Questions. Property (3), which says $b = \frac{M}{L(0)} - 1$, is perhaps the easiest. If we let $t = 0$ in our basic equation $L(t) = \frac{M}{1+be^{-rt}}$, we get

$$L(0) = \frac{M}{1 + be^{-r \cdot 0}} = \frac{M}{1 + be^{0}} = \frac{M}{1 + b}$$

$$1 + b = \frac{M}{L(0)}$$

$$b = \frac{M}{L(0)} - 1$$

Next, consider the parameter M, the limiting value for the function L, or the value at which L levels off. M will be the limiting value if the values of $L(t)$ get closer and closer to M as t gets large. We need to consider the expression e^{-rt} which is part of the function $L(t)$.

Recall that negative exponents mean "take the reciprocal." So, $e^{-rt} = 1/e^{rt}$. What happens to this expression if we evaluate it at large values of t? Well, if t is very large, than rt will still be large, since r will be positive. Even if r is only .01, for example, if we pick t larger than 10,000, then rt will be larger than 100. If we take the number e to a large power, since $e > 1$, we get an even larger number. In fact, if $rt = 100$, e^{rt} will be about 2.7×10^{23}, a huge number. So, for large values of t, $e^{-rt} = 1/e^{rt}$ will be 1 divided by a large number, and so will be very close to 0. Now, looking at the whole function $L(t) = \frac{M}{1+be^{-rt}}$, if e^{-rt} is approximately 0, then

$$L(t) = \frac{M}{1 + be^{-rt}} \approx \frac{M}{1 + b \cdot 0} = \frac{M}{1 + 0} = M$$

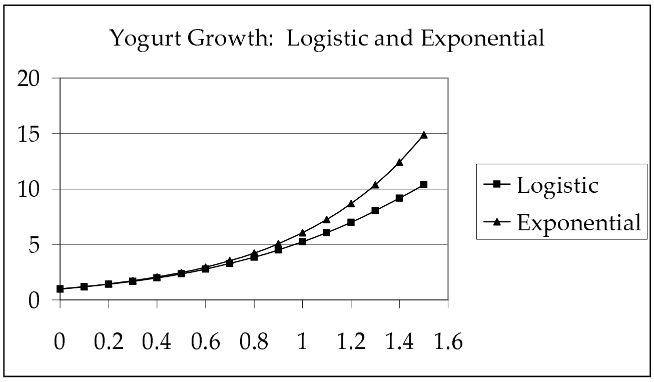


Figure 3: Logistic and exponential growth models for yogurt growth

Time	Logistic	Exponential
0.0	1.00	1.00
0.1	1.19	1.20
0.2	1.41	1.43
0.3	1.68	1.72
0.4	1.99	2.05
0.5	2.35	2.46
0.6	2.78	2.94
0.7	3.27	3.53
0.8	3.83	4.22
0.9	4.49	5.05
1.0	5.23	6.05
1.5	10.38	14.88
2.0	17.33	36.60

Table 1: Numerical comparison of logistic and exponential yogurt growth

So, M is the limiting value which L approaches as t gets large. Also, note that if $b > 0$, then so is be^{-rt}. This implies that

$$\frac{M}{1+be^{-rt}} < \frac{M}{1} = M$$

since the fractions have the same numerator and the denominator of the second fraction is smaller than the denominator of the first.

How large t must be for L to be close to M depends on r and b. If r is very small, this indicates a slow growth rate and so it will take longer for L to approach M. Similarly, the smaller b is, the larger t needs to be for L to be close to M.

Finally, let's consider the role of r in the logistic model. A complete mathematical explanation of why r is approximately equal to the percentage growth rate when L is much smaller than M requires calculus. However, we can get an idea of how the logistic and exponential models compare numerically and graphically. Table 1 shows values for the logistic model of yogurt growth given in Example 1, along with the exponential model $E(t) = e^{1.8t}$ which has the same initial amount of $E(0) = 1$ and the same percentage growth rate. At least over most of the first hour, the exponential model and the logistic model are very close. A graph of the two models over the first 1.5 hours is shown in Figure 3.

Application: World Population Growth

One of the main challenges in using logistic models is that, unlike the situation with our yogurt example, we rarely know what the maximum value M is ahead of time. Finding $L(0)$ is no problem, and we might be able to estimate r if we can be reasonably sure that the data we have is for values much less than whatever the maximum value is. However, without M we cannot directly find b and will not be able to get a model, at least in the same way we found our yogurt model in Example 1.

However, in cases where it looks like the quantity being modelled has already reached a point where the rate of growth is decreasing, then we can use property (4) to make an assumption on what M to use. In Figure 1, although there is not a great deal of "curve" to the graph, it looks as if the graph is steepest around 1990 or just before, and this is where the graph is changing from concave up to concave down. In fact, Table 2 shows the world population and the rates of growth for years around 1990. The rates of growth are given in number of people per year, and so the year which sees the biggest additional number of people is the year where the rate of growth is the largest.

Year	Population	Percent Growth	Population Increase
1980	4,454,269,203	1.69	75,864,564
1981	4,530,133,767	1.75	80,105,008
1982	4,610,238,775	1.73	80,253,764
1983	4,690,492,539	1.68	79,312,007
1984	4,769,804,546	1.68	80,596,505
1985	4,850,401,051	1.68	82,324,417
1986	4,932,725,468	1.71	85,142,812
1987	5,017,868,280	1.69	85,667,332
1988	5,103,535,612	1.66	85,671,996
1989	5,189,207,608	1.66	86,677,681
1990	5,275,885,289	1.58	83,940,351
1991	5,359,825,640	1.55	83,939,711
1992	5,443,765,351	1.48	81,404,054
1993	5,525,169,405	1.44	80,191,434
1994	5,605,360,839	1.43	80,626,257
1995	5,685,987,096	1.38	79,173,661

Table 2: World Population, 1980-1995, with percentage growth rates and growth rates in persons per year

The table confirms what we saw from the graph. The largest rate of change occurred in 1989, when the population increased by over 86 million people. If the world population is growing according to a logistic curve, then the population in 1989 should be about half of the limiting population. Under this assumption, the limiting population would be $M = 2 \cdot 5,189,207,608$ or about 10.4 billion people.

We can now create a logistic model for world population. We will use $M = 10.4$ and let $t = 0$ represent 1950. The population in 1950 was 2.56 billion so we let $L(0) = 2.56$. This gives us a value for b of $10.4/2.56 - 1 = 3.0625$. To complete our model, we need to find r. We could do this by using the parameters we have and one data point to create an equation. Which data point is a somewhat arbitrary choice, and we will use the population in 1980 which is 4.45 billion. The year 1980 corresponds to $t = 30$. From the basic model $L(t) = M/(1 + be^{-rt})$, we get

$$L(30) = 4.45 = \frac{10.4}{1 + 3.0625e^{-30r}}$$

$$1 + 3.0625e^{-30r} = \frac{10.4}{4.45} \approx 2.337$$

$$3.0625e^{-30r} \approx 1.337$$

$$e^{-30r} \approx \frac{1.337}{3.0625} \approx 0.43657$$

$$-30r \approx \ln 0.43657 \approx -0.8288$$

$$r \approx \frac{-0.8288}{-30} \approx .027627$$

This gives a logistic model for world population from 1950 of

$$L(t) = \frac{10.4}{1 + 3.0625e^{-0.027627t}}$$

This curve is plotted along with a few of the actual data points as a scatter plot in Figure 4.

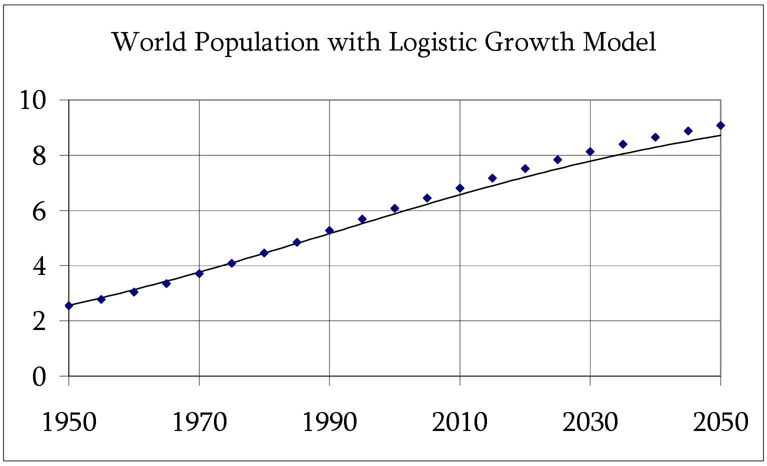


Figure 4: World Population with Logistic Model

The model follows the data fairly well over the first 40 years or so, but then underestimates the values after that. This may seem a little disappointing. However, we should keep in mind that all the values after 2003 are estimates, and in fact, would have been created by some other model.[119] In the Reading Questions for this chapter, you will be asked to create another logistic model for this data, which will hopefully give a better fit than this one.

Hubbert Curves

Figure 5 shows a graph depicting the total oil production of Norway, a significant exporter of petroleum.[120]
As noted previously, Hubbert was a geologist who used bell-shaped curves that have subsequently come to bear his name to approximate oil production data. Symbolically, **Hubbert curves** are of the form

$$H(t) = \frac{Mbre^{-rt}}{(1 + be^{-rt})^2}$$

[119] This other model may be a different logistic model, but more likely, is a more complicated model which also takes into account the actual and projected percentage growth rates

[120] Graph is from http://www.hubbertpeak.com/blanchard/

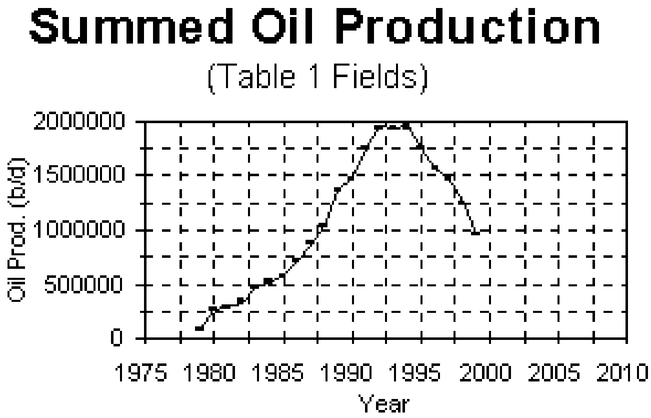


Figure 5: Norwegian Oil Production in Barrels per Day

where the parameters M, b, and r are the same as those in the logistic family of functions.[121] With respect to Hubbert curves

- M represents the total area under the curve.
- The peak occurs at $t = ln(b)/r$.
- The maximum height at the peak is equal to $Mr/4$.

If we wanted to find a Hubbert curve to model the Norwegian oil production data, we could first estimate the total area under the curve, which would be an estimate for M. To do this, let's estimate half of the area and then multiply by two. Looking at the graph, the peak year seems to be about 1993. As a rough estimate, we could say the area under the curve up to this year is about the area of the triangle with two vertices along the x-axis at the years 1980 and 1993, and the third vertex at the peak, which is at the year 1993 with a height of about 2,000,000 b/d or barrels per day. The base of this triangle has length 13 and the height is about 2,000,000. So, using the $\frac{1}{2}bh$ formula for the area of a triangle, the area is about

$$\frac{1}{2} \cdot 13 \cdot 2,000,000$$

which is half the area under the curve. So, the whole area is about $13 \cdot 2,000,000$ or about $26,000,000 = 2.6 \times 10^7$. So, we can let $M = 2.6 \times 10^7$.

The maximum height of the curve is about 2 million or $2 \cdot 10^6$. This would equal $Mr/4 = (2.6 \cdot 10^7)r/4$. So,

$$2 \cdot 10^6 = (2.6 \cdot 10^7)r/4$$

$$\frac{4 \cdot 2}{2.6 \cdot 10} = r$$

so that $r = .3077$.

Now, $ln(b)/r$ is the t-value of the peak, in this case 1993. So $\ln(b) = 1993r = 1993 \cdot .3077 = 613.25$. Solving for b, we get that $b = e^{613.25}$, which is about $2.135 \cdot 10^{266}$, a truly huge number. We could still use this for our model, although it might cause problems for a TI calculator. If we'd like, we could simplify things by letting $t = 0$ denote 1970 so that the peak would occur when $t = 23$.

[121] For those of you who do know calculus, take the derivative of the formula which gives the general form of logistic functions, and you should get the expression above for Hubbert curves

In this case, $\ln(b) = 23r = 23 \cdot (.3077) = 7.077$, and $b = e^{7.077} \approx 1184$, a much more manageable number. This gives a model

$$H(t) = \frac{2.6 \cdot 10^7 \cdot 1184 \cdot (.3077)e^{-.3077t}}{(1+1184e^{-.3077t})^2} = \frac{9.472 \cdot 10^{19}e^{-.3077t}}{(1+1184e^{-.32077t})^2}$$

where t represents the year from 1970.

Table 3 gives some values of this Hubbert curve model. You can compare these with estimated values from the graph to see how closely it fits.

Year	Hubbert
1980	141,349
1985	579,556
1990	1,628,193
1993	2,000,000
2010	42,317
2020	1,970

Table 3: Numerical values of Hubbert curve model for Norway

Piecewise Functions

Some functions given in symbolic form have different rules for different parts of their domain. For example, the function used for determining income tax involves different symbolic formulas for different income levels. Such functions are called **piecewise functions**. When determining the output for a given input, it is important to make sure you are using the correct rule. For example, let

$$f(x) = \begin{cases} -x & \text{if } x < 0 \\ x^2 & \text{otherwise.} \end{cases}$$

Then $f(-4) = -(-4) = 4$ since $-4 < 0$, while $f(0) = 0^2 = 0$ and $f(4) = 4^2 = 16$. The decision whether to use the rule "find the opposite of the input" or the rule "square the input" depends on whether your input is less than zero.

Figure 6 is the graph of a piecewise function. Notice the abrupt change at $x = 3$. Graphs of piecewise functions often have an abrupt change at the point where the rule changes.

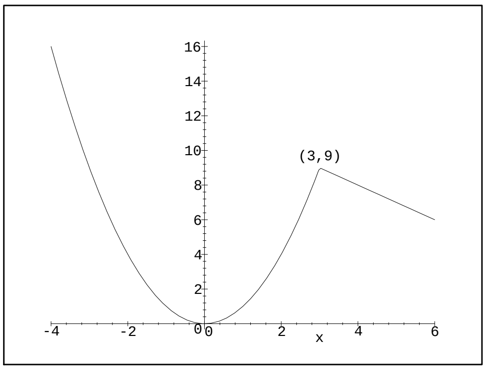


Figure 6: The graph of a piecewise function.

The symbolic representation of the function shown in Figure 6 is

$$f(x) = \begin{cases} x^2 & \text{if } x < 3 \\ -2x + 15 & \text{otherwise.} \end{cases}$$

Many modeling functions are piecewise functions. For example, phone companies often offer plans which charge customers a flat rate for so many minutes, and then a per minute charge if the customer uses more than the limit. For example, Verizon has offered a cell phone plan that costs \$39.99 per month and includes 450 anytime minutes. The cost per minute for calls going over the 450 minute limit is \$.45 per minute.[122] This verbal description can be represented in the following symbolic form, where m is the number of minutes used in a month, and $C(m)$ is the associate cost.

$$C(m) = \begin{cases} 39.99 & \text{if } 0 \leq m \leq 450 \\ 39.99 + 0.45(m - 450) & \text{if } m > 450. \end{cases}$$

So, for example, $C(300) = 39.99$, and $C(500) = 39.99 + .45(500 - 450) = 62.49$.

Example 2.

As another example, consider the Dakota Wesleyan University tuition cost function given verbally at the beginning of Chapter Four. Can we find a symbolic representation for this function?

Solution: We will give two symbolic functions, one exact and one approximate. First, we give the function in numerical form in Table 4. We also included the differences between successive cost values. Recall that if the difference in outputs for each unit increase in the input is constant, then the function is linear with the constant difference representing the slope. In this case, if portions of the table show a linear pattern, we can form a piecewise functions with those pieces represented by a linear formula.

Hrs.	Cost	Diff.	Hrs.	Cost	Diff.
1	375		11	7725	1625
2	750	375	12	8750	1025
3	1125	375	13	8750	0
4	1500	375	14	8750	0
5	1875	375	15	8750	0
6	2525	650	16	8750	0
7	3150	625	17	9125	375
8	3850	700	18	9500	375
9	4475	625	19	9875	375
10	6100	1625	20	10250	375

Table 4: DWU Tuition Costs for the Fall of 2007

From the table, we see that there is a linear pattern for credit hours between 1 and 5 inclusive, from 9 to 11 inclusive, from 12 to 16 inclusive, and over 17. The pattern from 6 to 9 is approximately but not exactly linear. A piecewise function which gives the exact values for tuition $T(c)$ for a given number of credit hours c is given by:

[122] Plan details acquired at www.pricegrabber.com on August 17th, 2007

$$T_1(c) = \begin{cases} 375c & \text{if } 0 \le m \le 5 \\ 2525 & \text{if } m = 6 \\ 3150 & \text{if } m = 7 \\ 3850 & \text{if } m = 8 \\ 4475 + 1625(c-9) & \text{if } 9 \le m \le 11 \\ 8750 & \text{if } 12 \le m \le 16 \\ 8750 + 375(c-16) & \text{if } 16 < m \end{cases}$$

Notice we have four linear pieces with slopes 375, 1625, 0, and 375 respectively. Unfortunately, we do still have a complicated function with 7 different pieces. We could make this simpler by approximating the tuition costs for the credit hours between 6 and 8. One way to do this would be to find the linear equation through the points $(6, 2525)$ and $(8, 3850)$. The slope is

$$m = \frac{3850 - 2525}{8 - 6} = \frac{1325}{2} = 662.5.$$

To find b, we solve $2525 = 662.5 \cdot 6 + b$ for b to get $b = -1450$. If we use this formula for c values between 6 and 8, we get a second function with 5 different linear pieces.

$$T_2(c) = \begin{cases} \text{375c} & \text{if } 0 \le m \le 5 \\ \text{662.5c-1450} & \text{if } 6 \le m \le 8 \\ \text{4475 + 1625(c-9)} & \text{if } 9 \le m \le 11 \\ \text{8750} & \text{if } 12 \le m \le 16 \\ \text{8750 + 375(c-16)} & \text{if } 16 < m \end{cases}$$

The only difference between these two functions will be that $T_1(7) = 3150$ and $T_2(7) = 3187.50$. Perhaps we could suggest that DWU make a slight modification of their tuition structure for next year. Also, note that the two of the linear formulas are not in $y = mx + b$ form, but could be simplified to this form. We have

$$\begin{aligned} 4475 + 1625(c-9) &= 4475 + 1625c - 14625 = 1625c - 10150 \\ 8750 + 375(c-16) &= 8750 + 375c - 6000 = 375c + 2750 \end{aligned}$$

■

Reading Questions for Other Families of Functions

1. Let

$$f(x) = \begin{cases} 4x & \text{if } x \leq 0 \\ x^2 - 1 & \text{if } x > 0 \end{cases}$$

Compute each of the following.

(a) $f(-2)$
(b) $f(0)$
(c) $f(3)$

2. Let

$$g(x) = \begin{cases} 2^x & \text{if } x \leq 0 \\ x^2 + 1 & \text{if } 0 < x < 4 \\ 2x + 9 & \text{if } x \geq 4 \end{cases}$$

Compute each of the following.

(a) $g(-2)$
(b) $g(0)$
(c) $g(3)$
(d) $g(4)$
(e) $g(6)$

3. In this chapter, we created a logistic model for world population, and commented that the model seemed to underestimate the actual population values after 1990.

(a) Create a logistic model for world population, letting $t = 0$ correspond to 1970, but use logistic regression instead of the technique in the text.
(b) Graph your model together with the data given in the appendices, starting in 1950, to see how well the model fits. You may use an electronic file with the data. Otherwise, type in the data using only the years 1950, 1955, 1960, etc.
(c) According to this model, what will the eventual upper limit on human population be? How well does this model fit the data for the years 2000 through 2050.

4. Give an example (other than the one in the reading) of a piecewise function which is a **modeling** function.

5. In the reading, we looked at the cost of a particular cell phone plan from Verizon. Alltel Wireless had a competing plan for \$49.99 per month for 1000 minutes, plus 40 cents per minute over the 1000 minute limit.

(a) Find a symbolic representation for the piecewise function giving the cost of the Alltel plan as a function of the number of minutes used.
(b) If we let m represent the number of minutes, for which values of m does the Verizon plan cost less than the Alltel plan?
(c) If you want to be sure your monthly cost is below \$80, how many minutes can you use under each of these plans?

6. In the text, we created a function $T_2(c)$ which essentially gave the tuition cost at DWU for the fall of 2007 as a function of the credit hours c.

(a) Suppose for the fall of 2008, DWU wants to increase tuition across the board by 4% over the tuition costs given by T_2. Give a piecewise function in symbolic form which does this.

(b) Suppose for the fall of 2008, DWU wants to create a more simplified tuition structure with only 3 linear pieces. The function should satisfy the following conditions:

- A student taking 0 credits is charged \$0.
- A student taking 12 to 16 credits is charged \$9000.
- Credits over the 16 hour limit are charged at \$400 per hour.

Create a piecewise function with 3 linear pieces that meet these 3 conditions.

7. Suppose that $H(t) = \frac{120{,}000e^{-0.05t}}{(1+3.5e^{-0.05t})^2}$ is a Hubbert curve model for annual U.S. oil production, where $t = 0$ represents 1950. The units for $H(t)$ are in thousands of barrels per day.

(a) Graph $y = H(t)$. Figure 1 of Chapter Three shows a scatterplot of U.S. oil production. How does the graph of H compare with this scatterplot?

(b) In which year does H show a peak?

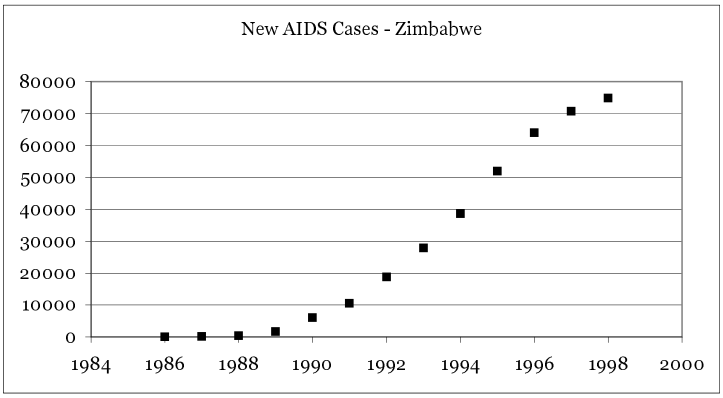
(c) Use H to estimate the total U.S. oil production in 1950, 2000, 2020, and 2050.

(d) Comparing the symbolic representation for H with the general form for Hubbert curves given in the text, we can see that for this curve, $b = 3.5$ and $r = .05$. Use these values and the general form to find M for this curve.

(e) In 2000, the U.S. consumed about 16,000 thousands of barrels of oil per day. In part (c), you estimated how much oil the U.S. produced in 2000. Based on this information, what percent of its oil needs did the U.S. have to import in 2000?

(f) The M that you found in part (d) represents the estimated total amount of oil that has been and will ever be produced by the U.S. The peak year you found in part (b) represents the year in which half of this production would have already occurred. This means that after that year, we only have half of our value of M left to use. Let's be optimistic and suppose that we still had half of M left to use in the year 2000. Assuming that the U.S. consumes 16,000 thousands of barrels per day from 2000 on, *and* is able to access this much oil every year from its own production sources, how long would it be before the U.S. used up all its oil?

8. The graph below shows the cumulative number of AIDS cases reported for Zimbabwe, a country in southern Africa, for each year. The values represent the total number of cases reported up to and including that year.



(a) Explain why a logistic model might be appropriate to model this data.

(b) The table below gives the data on which the graph above is based. Based on the graph and the table, estimate the maximum number of cases for a logistic model of these data.

Year	1986	'87	'88	'89	'90	'91	'92	'93	'94	'95	'96	'97	'98
Cases	0	119	321	1632	5994	10551	18731	27905	38552	51908	63937	70669	74782

(c) Find a logistic model for these data, and graph both the model and the data. How well does your model seem to fit? Experiment with different parameters until you get the best fit you can. You may have to try a number of different models before finding one that is satisfactory.

(d) The population of Zimbabwe is currently about 12.6 million. According to your model, about what percent of this population has AIDS? (Note that this could be significantly less than the number that are infected with the HIV virus but have not yet developed AIDS).

(e) Suppose the percentage of the entire world population that has AIDS is the same as the percentage in Zimbabwe. The world population in 2005 is estimated to be about 6.4 billion. How many people would have AIDS under this assumption?

(f) According to the Population Reference Bureau, about one quarter of the population of Zimbabwe already has the HIV virus. How does this fact affect your model?

9. In the section on logistic models, we note that a logistic graph has an inflection point when $L(t) = \frac{1}{2}M$, and that this occurs when $t = \frac{ln(b)}{r}$. You can see why this is by solving the equation

$$\frac{1}{2}M = \frac{M}{1 + be^{-rt}}$$

for t. Solve this equation for t, writing out your solution step by step. You can consult Example 1 in the text to help. As an additional hint, recall that $\ln(b^n) = n \cdot \ln(b)$, according to the properties of logarithms.

10. In the text, we estimated that currently, there is about 200,000 square feet for each person in the world. We also estimated, based on a logistic growth model, that the world's human population will eventually level off at around 10.4 billion. If and when the population reaches 10.4 billion, how many square feet will each person have? (Hint: There is an easy way and a hard way to do this!)

11. In Example 1, we considered the growth of yogurt in a 32 ounce jar, starting with 1 ounce of yogurt culture. In that example, it took 4.2 hours for the amount of yogurt to reach 31.5 ounces. Suppose you decided to speed up the process by adding more yogurt at the beginning of the process.

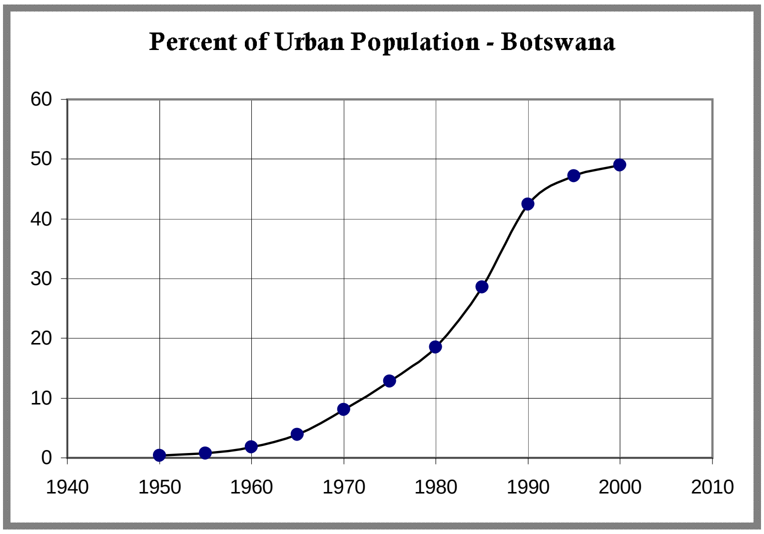
(a) How would the function L from Example 1 change if we started with 2 ounces instead of 1 ounce of yogurt?

(b) Using the new logistic model from part (a), how long would it take for the amount of yogurt to reach 31.5 ounces?

Other Families of Functions: Activities and Class Exercises

1. **The Urbanization of Botswana.** The table and graph below show the urbanization of the population of Botswana over time. The dependent variable is the percent of the population which is classified as urban.

Year	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000
Percent Urban	0.4	0.8	1.8	3.9	8.1	12.8	18.5	28.6	42.4	47.2	49



   (a) Create the best logistic model for this data that you can. You may use the logistic regression feature on your calculator, if you wish.
   (b) For each of the three parameters in your equation from part (a), explain what the parameter means in the context of urban population in Botswana.

2. **The Ozone Hole Revisited.** In the Chapter Five activities, you may have considered data related to the size of the ozone hole over the Antarctic. In this activity, we will reconsider these data, using some of the other families we have discussed since then. Refer to Chapter Five for the original data and scatterplot.

   (a) Recalculate the linear regression equation for this data, letting $x = 2$ represent the year 1982. Also, calculate the power regression equation. Based on both the correlation coefficient and the shape of the scatterplot, which model do you think is better? (NOTE: Remember for the power regression you must exclude any points with 0 as an x or y coordinate. This is why we suggested you let $x = 2$ denote the year 1982.)
   (b) Based on the scatterplot, it seems that after 1989, the data are approximately linear, but with a smaller slope than the data show up to that point. Find the linear regression model for the data but use only the years starting with 1989 ($x = 9$). Give the equation and the correlation coefficient.
   (c) Graph your data together with all three models (both linear models and the power model), over a range including all the data points. Which model do you think would give the most accurate predictions for the next 10 to 20 years? Explain why you think so.
   (d) Both linear and power models (with positive exponent) have y values which get larger and larger as x increases. However, the ozone hole can only get so big. Find a reasonable upper limit for the size of the ozone hole (in square kilometers). Knowing the surface area of the earth might be useful.

(e) One type of equation that can be used to model data with an "upper limit," is a transformed exponential function. One way to find such a model is to transform the data, and then use exponential regression. In our case, we can do this in two steps (or one step with two operations). We can subtract the upper limit from all the data points (this makes the data all negative with an upper limit of 0), and then multiply the data by -1. (On a TI-83, you could do this by creating a new list with values $-(L - U)$ where L is the list with the original data and U is the upper limit).

i. Do the transformation of the data as described, using your estimated upper limit.

ii. Find the exponential regression equation for the transformed data. Use the original x-values with the transformed y-values. Keep track of the correlation coefficient.

iii. Transform your equation to an equation that will work for the original model.

iv. Graph your transformed equation together with the data. Decide which of the four models you have created is the best model.

v. Repeat the previous steps using a couple of different upper limits. Try one that is just slightly higher than the largest area given in the actual data. How does changing the upper limit seem to effect the correlation coefficient and the graphical fit?

(f) Based on what you have done so far, which of the functions that you have developed seems to be the best model? What is your best estimate as to how large the ozone hole will eventually become? Note that this estimate is based entirely on numerical data. It would definitely be helpful if we had some scientific information or explanation of how or why the ozone hole was growing.

3. World Hunger In 1996, the United Nation's Food and Agriculture Organization convened a World Food Summit in order to begin to comprehensively address the problem of world hunger. The summit established the goal of reducing the number of chronically hungry people in the developing world from 819 million in 1990 to half that by 2015. The FAO's 2002 report, *The State of Food Insecurity in the World: 2002*, noted that progress towards the goal has been much slower than hoped for.[123]

> Worldwide, the latest estimates indicated . . . [that there were 799 million undernourished people in the developing world in 2000], . . . a decrease of just 20 million since 1990-92, the benchmark period used at the World Food Summit (WFS). This means that the average annual decrease since the Summit has been only 2.5 million, far below the level required to reach the WFS goal of halving the number of under nourished people by 2015. It also means that progress would now have to be accelerated to 24 million per year, almost 10 times the current pace, in order to reach that goal.

The report includes the following graph.

(a) The graph includes two different projections, one a "business as usual" scenario and one reaching the 2015 target.

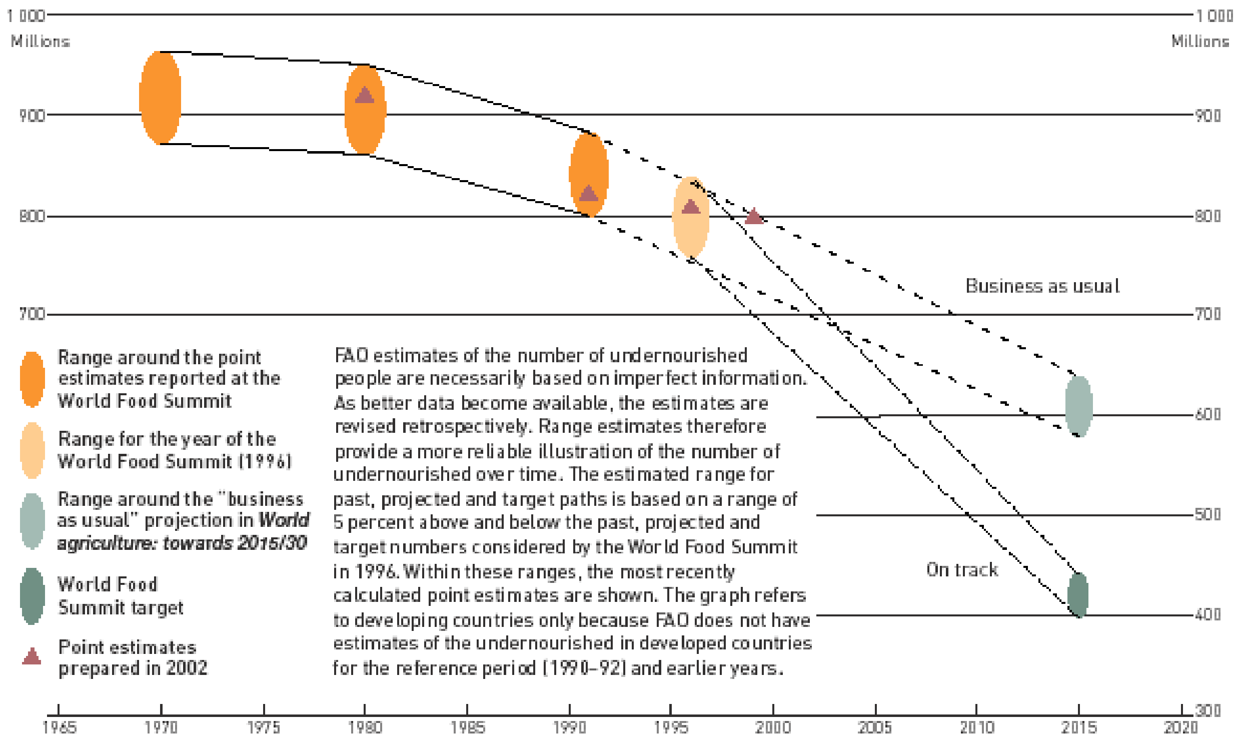
i. What family of functions do both scenarios seem to be using?

ii. Assuming the business as usual projection uses a linear model, what is the slope of the line, based on the information given in the quote from the report?

iii. Find an equation for the business as usual linear model. You can choose which year you would like to be represented by $x = 0$; any year from 1995 to 2000 might be convenient.

iv. Using this model, in what year would we reach the goal of around 410 million hungry people in the developing world?

[123] Reprinted with permission of the Food and Agriculture Organization of the United Nations. The full report can be accessed at www.fao.org



v. Consider the second scenario. What slope would be required for a linear model if we are to reach the 2015 goal by 2015, starting from a value of 799 million in 2000?

vi. Using your second model, if the trend continued indefinitely, when would hunger in the developing world be eliminated?

(b) Find a piecewise linear function which models the entire graph, from the year 1970 on, using the "On track" scenario where the goal of 410 million is met by 2015. Again, you have a choice of which year you want to correspond to $x = 0$. You can divide the graph into 3 linear 'pieces' and will need a separate equation for each. You could consider the years from 1980 to 1995 as one piece. However, you already know the slopes for two of these pieces. It will probably be most convenient to let $x = 0$ represent either the starting year (1970), or one of the years dividing the pieces.

4. **Selecting a Family.** Each of the situations below might be modeled by one of the families we have discussed so far, possibly including a function transformation. This activity will involve the first step in the modeling process given in Figure 1 of chapter nine, reproduced below. Specifically, for each situation, you should:

1. Is the situation more likely to be modeled by an empirical model, or by a theoretical model. Explain why you think so.
2. Discuss which family or families would most likely make a good model for the situation. Explain why you think so.
3. For one of the families you picked in 2, discuss whether the parameters have a particular meaning, and if so, explain what that meaning is as specifically as possible.

Feel free to be creative! Also, try to be as thorough as you can in your discussion, even though you will not have complete information.

(a) You are a member of the government in a developing country that is currently in the very beginning stages of a famine caused principally by a drought. You are very concerned

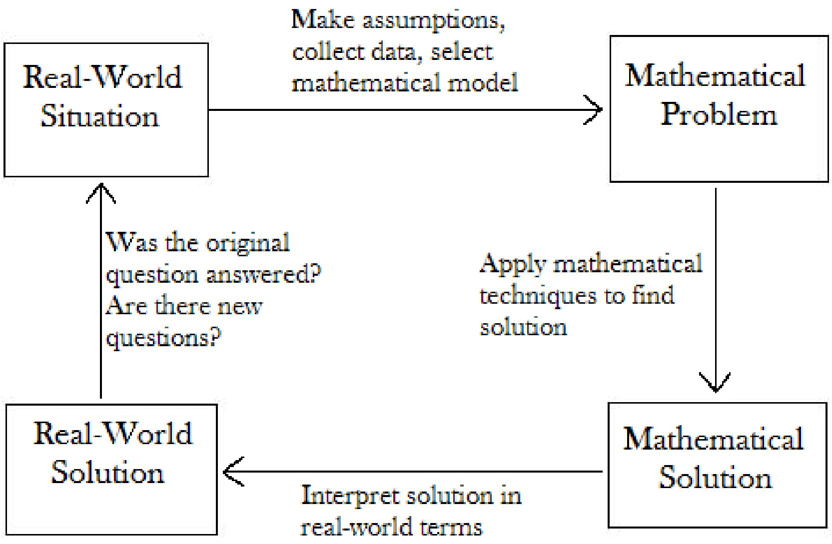


Figure 7: The Mathematical Modeling Process

about the number of people in your country at risk from starvation. As part of your planning for dealing with the situation, you decide to create a model which would estimate the total number of deaths N from the famine as a function of the time t (say, in months), assuming that your country receives little or no aid from foreign sources.

(b) You are a member of the Marketing Division of the Hewlett-Packard (HP) Corporation. The company is interested in projecting revenues from its sales of laptop computers. Part of this will involve projecting the "average *real* price" $P(t)$ of laptop computers as a function of the time t in years. Laptops (or portable computers) were first marketed in the early 1980's, and you have data on laptop prices both for HP and for several other major computer manufacturers going back to at least 1995.

(c) You work for the Social Security Administration. Your assignment is to find a model which will predict the number of people over the age of 65 as a function of time. You are principally interested in predictions for the years 2010 through 2050. You of course have complete access to all U.S. census data, and are told to assume current policies with regards to benefits and eligibility will continue.

(d) You work for the U.S. Department of Agriculture (USDA). A new President has just been elected, and one of her campaign issues involved addressing the disappearance of the family farmer. She has asked the USDA do a study which is to include the history of small family farms in the U.S. and projections on the number of small family farms should current trends continue.

5. **Canadian Oil Production.** Open the Fathom file **Canadian Oil Production Data**. You should see a case table, three sliders, and a graph which contains a scatterplot and the graph of a Hubbert curve. The scatterplot is based on data for Canadian crude oil production from 1980 through 2003. The independent variable is the years since 1980. The units for oil production is quadrillions of BTUs.

   (a) The sliders represent the three parameters M, b, and r for the Hubbert curve model, as described in the text. These sliders are linked with the Hubbert curve shown in the graph. Play with the values of the sliders until the Hubbert curve fits the scatterplot as closely as possible. Record your values for the parameters M, b, and r.

(b) Recall the M represents the total area under the curve or the total amount of oil that will ever be produced. Given your value of M, what is the total amount of oil that Canada will eventually produce?

(c) According to your Hubbert model, in which year will or did Canadian oil production reach its peak?

(d) Canadian oil production was about 10 quadrillion BTU in 1980. After the peak year, production will eventually diminish to that level again sometime in the future. Estimate which year this will be (Hint: You can use the symmetry of the Hubbert curve around the peak to help).

6. World Oil Production. Open the Fathom file **World Oil Production Data**. You should see a case table, three sliders, and a graph which contains a scatterplot and the graph of a Hubbert curve. The scatterplot is based on data for world crude oil production from 1980 through 2003. The independent variable is the years since 1980. The units for oil production is quadrillions of BTUs.

(a) The sliders represent the three parameters M, b, and r for the Hubbert curve model, as described in the text. These sliders are linked with the Hubbert curve shown in the graph. Play with the values of the sliders until the Hubbert curve fits the scatterplot as closely as possible. Record your values for the parameters M, b, and r.

(b) Recall the M represents the total area under the curve or the total amount of oil that will ever be produced. Given your value of M, what is the total amount of oil that Canada will eventually produce?

(c) According to your Hubbert model, in which year will or did world oil production reach its peak?

(d) World oil production was about 10 quadrillion BTU in 1980. After the peak year, production will eventually diminish to that level again sometime in the future. Estimate which year this will be (Hint: You can use the symmetry of the Hubbert curve around the peak to help).

11. Adding Things Up

The first lesson of economics is scarcity: There is never enough of anything to satisfy all those who want it. The first lesson of politics is to disregard the first lesson of economics.
— Thomas Sowell

Energy Consumption

In Chapter Two and again in Chapter Ten, we considered energy production, and looked at so-called Hubbert curves. We saw that oil production in the U.S. and Norway had already peaked and is on the decline, and that world production is predicted to decline starting around 2010.[124] We suggested that a global energy crisis might be in our future.

Of course, if there was not a huge demand for energy there would be no crisis. We need to consider not only the history of energy production, but also the history of energy consumption and see if we can make some reasonable predictions based on the known data. Figure 1 shows a graph of Total Annual U.S. Fossil Fuel Consumption from 1982 to 2002. The scatterplot has a strong linear trend, and the regression line is shown. The equation is $y = 0.18984x + 10.5832$, where $x = 0$ represents 1980 and the correlation coefficient is $r = 0.986$.

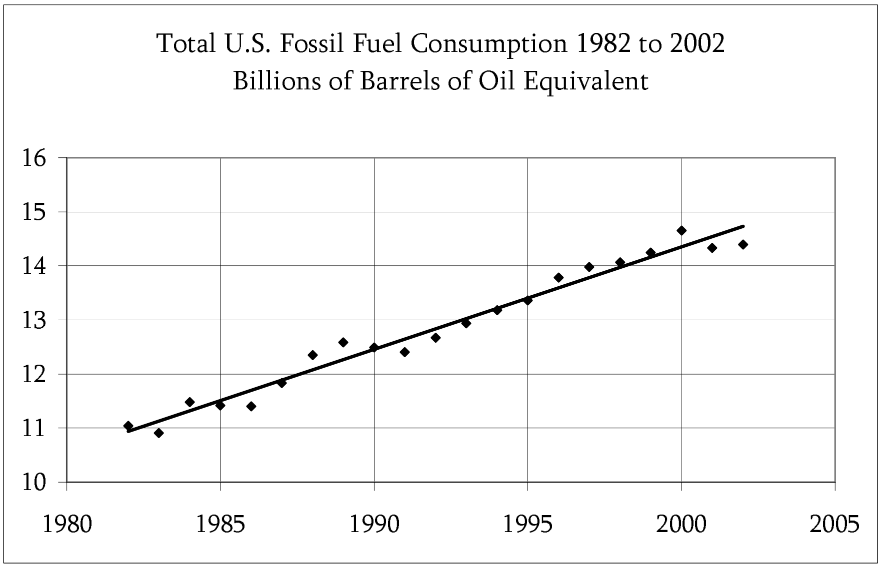


Figure 1: U.S. Energy Consumption

Total consumption includes consumption of oil, natural gas, and coal. The consumption is measured in oil equivalent. In other words, if we used only oil to satisfy our fossil fuel consumption, instead of a combination of oil, gas, and coal, this is how much oil we would need to get the equivalent amount of energy. According to the graph, in 2000, the U.S. used fossil fuels equivalent to about 14.7 billion barrels. How much energy is this?

[124]Keep in mind that this is just one prediction. There are many organizations and companies involved in predicting future oil production, and as you can imagine, their predictions vary widely. See for example *Has Global Oil Production Peaked?* at http://www.csmonitor.com/2004/0129/p14s01-wogi.html.

A barrel of oil is 42 gallons, which gives you a rough idea of its volume. However, the amount of energy is really the important characteristic. Energy is usually measured in British Thermal Units, BTU for short. The BTU is defined to be the amount of energy required to raise the temperature of 1 pound of water 1 degree Fahrenheit. To give you a more meaningful picture, eight gallons of gasoline contain about 1 million BTU. If your car gets 25 miles per gallon, then 8 gallons would get you about 200 miles. This means 1 million BTU is enough to drive your car 200 miles.

Gasoline is refined from crude oil. One barrel of crude oil contains about 6.2 million BTU.[125] So, a barrel of oil is enough for you to drive your car about $6.2 \times 200 = 1240$ miles.

In 1999, the average single-family residence in the United States consumed a little over 100 million BTU of energy, and the per capita consumption of energy was about 351 million BTU. The latter figure includes energy used for *all* purposes, not just residential use. Since $351/6.2 \approx 57$, U.S. citizens used, on the average, about 57 barrels of oil in 1999. This is enough to drive about 70,000 miles in a car getting 25 mpg.

Now, suppose we want to use our regression equation to estimate how much total energy the U.S. will use over the next 10 or 20 or 50 years. Our regression equation tells us only the consumption y in any one given year x. To get the total consumption, say for the years 2005 through 2015 we could compute the yearly consumption for each of these 11 years using the equation, and then add these up. This would be somewhat tedious, but doable in a few minutes. On the other hand, if we changed the range of years we would have to do our calculations over. If we wanted 50 years worth of production, the calculation would be a big pain (and would be highly susceptible to error).

Is there a better way?

Fortunately, the answer is yes. Based on what we know about linear functions, we will be able to easily develop a way of adding up the consecutive values of a linear function. Let's illustrate with a simple example.

Suppose you want to add up the odd numbers from 1 through 101. Let's denote this sum by S. So, we want to know the sum

$$S = 1 + 3 + 5 + 7 + \cdots + 101$$

Notice that each number is two more than the previous number. We could consider these numbers as the y-values of the linear function $y = 2x - 1$ of slope 2, for integer x-values from 1 up to 51. Now, we'll do a little arithmetic trick to simplify this sum. Let's write it down two ways, frontwards and backwards, and then add vertically.

$$\begin{array}{rcccccccccccl} S & = & 1 & + & 3 & + & 5 & +\cdots+ & 97 & + & 99 & + & 101 \\ S & = & 101 & + & 99 & + & 97 & +\cdots+ & 5 & + & 3 & + & 1 \\ \hline 2S & = & 102 & + & 102 & + & 102 & +\cdots+ & 102 & + & 102 & + & 102 = 51 \cdot 102 \end{array}$$

Solving for S we get that

$$S = \frac{51 \cdot 102}{2} = 2601$$

This was certainly easier (and less error prone) than adding up all 51 numbers by hand or even with a calculator.

Why did this work? The key is that we always got the same number, 102, when we added vertically. This happened because each number differed by 2 from the numbers before it and after it in the list. In the first line, the numbers increased two each time, and in the second line they decreased by two each time. This, in turn, occurred because the numbers we were adding were *consecutive* y-values of a *linear* function with slope 2.

Notice that, in the final calculation of our sum S, all we had to know is how many numbers were being added (51 in this case), and the sum of the first and last numbers ($1 + 101 = 102$). The

[125]Information in this paragraph is from the Department of Energy's Energy Information Adminsitration at http://www.eia.doe.gov/neic/infosheets/apples.htm.

sum S is the product of these two numbers divided by 2. The same procedure would work for any list of numbers that are *consecutive values of a linear function.* Let's use this idea to predict the total U.S. energy consumption for the years 2001 through 2050.

From our regression equation $y = 0.18984x + 10.5832$, we have that U.S. consumption in 2001 and 2050 are estimated to be

$$\begin{aligned} y &= 0.18984(21) + 10.5832 = 14.57 \\ y &= 0.18984(70) + 10.5832 = 23.87 \end{aligned}$$

Consumption for each year will be 0.18984 more than the year before, since the consumption values are given by consecutive values of the linear equation $y = 0.18984x + 10.5832$. We are adding 50 consumption values, so the sum will be

$$\frac{50(14.57 + 23.87)}{2} = \frac{50(38.44)}{2} = 961$$

In other words, if current trends continue, the U.S. will use the equivalent of 961 billion barrels of oil during the first half of the 21st century. How much is this really? It is estimated that there were 1.266 trillion barrels of proven oil reserves in 2004,[126] up from 0.936 trillion barrels 10 years before.[127] Of course, the U.S. is using coal and natural gas as well as oil. In fact, from 1990 to 2000, about 45% of U.S. fossil fuel energy consumption came from oil. This is the good news. On the other hand, the U.S. currently uses "only" about 40% of the world's energy. The rest of the countries of the world consume the remaining 60%, and many of these countries are seeing their annual consumption increase dramatically, much more quickly than the U.S. consumption.

If we assume the ratio of oil consumption to total fossil fuel consumption for the world at large is similar to the ratio for the U.S., that the ratio of U.S. to world consumption remains at about 40%, and that current trends continue, our calculations would imply that roughly 80% of the known world supply of oil will be gone in 50 years.

Obviously, if more oil is discovered, this would increase the time to depletion. Some estimates claim that actual recoverable oil reserves may be as high as 3 trillion barrels, about two and one-half times the current known reserves. Also, there are widely varying estimates for trends in future world energy consumption. Finally, under Hubbert's assumptions, annual production would decrease rapidly from the peak and then level off as production nears zero. So, our rate of production (and necessarily consumption) would have to decline, instead of continuing to grow, as we deplete the resource.

For these and other reasons, predicting the date of oil depletion is tricky. To give one other example, estimates from the Hubbert Peak of Oil Production website predict that yearly world production will drop from the current level of about 25 billion barrels[128] to below 1 billion barrels sometime between 2050 and 2075, with total depletion around the year 2100 or earlier.

Arithmetic Sequences and Series

In our discussion of energy consumption, we used a linear equation to model U.S. consumption, evaluating the function only at **integer** values representing years. When the domain of a function is restricted to the set of positive integers, or some subset of the positive integers, mathematicians often refer to the function as a **sequence**.

In Chapter Four, we introduced the "$f(x)$" notation as the standard way to denote functions in symbolic form. We could use the same notation to denote sequences, as sequences are simply functions, but mathematicians historically have used another notation. To illustrate with an example, the function $f(x) = 2x - 1$ in **sequence notation** would be denoted $a_n = 2n - 1$. The notation works the same way as the function notation, except that we usually use n as the independent variable instead of x, and n appears as a subscript instead of in parentheses.[129] However,

[126] http://www.csmonitor.com/2004/0129/p14s01-wogi.html
[127] The data is available at a variety of locations, including http://www.geocities.com/combusem/ENERCIA.HTM.
[128] This is based on data from http://www.eia.doe.gov/neic/infosheets/crudeproduction.htm.
[129] Why is n used? Positive integers have also traditionally been called Natural Numbers, thus the n

evaluating values for a sequence would still be done as with functions. For example, if $a_n = 2n-1$, we have

$$\begin{aligned} a_1 &= 2\cdot 1 - 1 = 1 \\ a_2 &= 2\cdot 2 - 1 = 3 \\ a_5 &= 2\cdot 5 - 1 = 9 \end{aligned}$$

and so on. The values of the sequence are called **terms**. We would call a_1 the first term, a_2 the second term, etc., and a_n the **general term** or the nth term.

Since the values for the independent variable in a sequence are usually the positive integers, a sequence is sometimes represented simply as a list of terms. For example, the sequence $a_n = 2n-1$ could be represented by

$$1, 3, 5, 7, \ldots$$

where " ... " simply means that the pattern continues. This is sometimes useful, especially when there is no obvious symbolic representation for the sequence. For example, for the sequence $1, 1, 2, 3, 5, 8, 13, \ldots$ we can with a little thought see that the pattern is "add two consecutive terms to get the next term." However, it is not at all obvious how we might write a symbolic representation for the general term a_n.

A sequence that is a linear function, like $a_n = 2n-1$, is called an **arithmetic sequence**. What we have called the slope is called the **common difference** when the context is a sequence.

Series

A **series** is the sum of some number of terms of a sequence. Series can be denoted as we did above, for example

$$S = 1 + 3 + 5 + 7 + \cdots + 101$$

We call this the **expanded form** of the series. The terms are written out consecutively with " $\cdots$ " used to avoid having to write out every term. Alternatively, we often use what is known as the "sigma notation" to denote series. We introduced this notation briefly in Chapter Three. Using sigma notation, the series above can be written

$$\sum_{i=1}^{50} 2i - 1$$

The variable i is called the **index**, and $2i-1$ is the general term of the series. The notation signifies that we will add up the values of the sequence given by the general term starting with the first term ($i = 1$) and going up to the fiftieth term. If we wanted to add up the terms starting with the 4th term and going to the 21st, we would write

$$\sum_{i=4}^{21} 2i - 1$$

In general, we let S_n denote the sum of the first n terms of any sequence a_n. In other words,

$$S_n = \sum_{i=1}^{n} a_n$$

Using this notation, the sum of an **arithmetic series** a_n would be

$$S_n = \frac{n(a_1 + a_n)}{2}$$

Here, n is the number of terms we are adding, and $a_1 + a_n$ is the sum of the first and the last terms.

Example 1. What is the sum of the following series.

1. $\sum_{i=1}^{70}(3i-1)$

2. $11+17+23+\cdots+119$

Solution:

1. Since the terms are given by the formula $3i-1$ which is a linear function of i, this is an arithmetic series with 70 terms. The first term is $a_1 = 3\cdot 1 - 1 = 2$ and the last term is $a_{70} = 3\cdot 70 - 1 = 209$. So the sum is

$$S_n = \frac{n(a_1+a_n)}{2} = \frac{70(2+219)}{2} = \frac{15470}{2} = 7735$$

2. Here, the first term is 11 and the last term is 119. We can see that the terms are increasing by 6 each time. We need to find the number of terms. One way to do this is to first find a general formula for a_n. Since the common difference is 6, the general formula for the sequence is of the form $a_n = 6n+k$ for some k. If $n=1$, we get $11 = a_1 = 6\cdot 1 + k$, so that k must 5. So, $a_n = 6n+5$. You can check this expression on the first few terms.

To find out how many terms n there are, set $a_n = 6n+5 = 119$. Solving this for n, we get $n = 19$. So, the sum is

$$S_n = \frac{n(a_1+a_n)}{2} = \frac{19(11+119)}{2} = 1235$$

■

Savings Plans and Investments

So far, we have looked at how to add up consecutive values of a linear function or arithmetic series, and have introduced some general notation related to sequences and series. Next, we will look at adding up the consecutive values of an *exponential* function. A sequence where the terms are given by an exponential function is called a **geometric sequence** and a sum of the terms of a geometric sequence is called a **geometric series**. We will introduce these in the context of savings plans.

Suppose you are really planning ahead and want to start saving money for your retirement. Suppose decide to put a certain amount aside each month, say \$100, into a savings account. Let's assume that interest in the account is compounded monthly and denote the monthly interest rate by i. We know that when interest is compounded periodically, the balance in the account grows exponentially according to the formula $B = P(1+i)^t$, where P is the amount invested, i denotes the *periodic* interest rate, and t the number of periods. In our context t will denote the number of *months*, and i is the monthly interest rate.

Let's assume next that you want to retire in 40 years. If you make your first payment at the end of this month, it will grow at the monthly interest rate i for 40 years, or $40\cdot 12 = 480$ months. At the end of this time, it will be worth $100(1+i)^{480}$ dollars. If the annual interest rate on your account is 6%, then $i = .06/12 = .005$ and your first payment will be worth $100(1.005)^{480} = \$1095.75$.

Now, this is not a whole lot of money relative to supporting you in your old age, although it is over ten times the \$100 you invested. Of course, this is only *one* of your monthly deposits. What would happen if you made deposits every month for 40 years?

Well, each deposit would grow according to the same formula, except that each deposit will be growing for one month less than the previous deposit. Your second deposit will grow for 479 months, your third for 478 months, etc.

What will the account be worth on the day you make your last deposit?

On the day you make that last deposit, it will not have earned any interest, so it adds only \$100 to the value of your account. The second to last deposit earns interest for one month, so it will contribute $100(1.005)^1$ to the total. Adding up all the deposits, the total in the account would be

$$S = 100 + 100(1.005)^1 + 100(1.005)^2 + \cdots + 100(1.005)^{478} + 100(1.005)^{479} + 100(1.005)^{480}$$

Again, this would be a big pain to calculate. There are actually 481 terms to add up. Is there a short cut in this case, as there was when we were adding up terms of an arithmetic sequence?

Yes, in fact, there is. In this case, the aspect that will help us is that the terms we are adding up are consecutive values of an *exponential function*, namely the function $y = 100(1.005)^t$. In the context of sequences, we will write $a_n = 100(1.005)^n$, and we call such a sequence a **geometric sequence**. Geometric sequences are simply exponential functions where the independent variable is restricted to the positive integers. What we previously called the growth factor for the exponential function will now be called the **common ratio** of the geometric sequence. The reason for this is that if you take the ratio of any two successive terms, the result is equal to the *growth factor*. In our example, for any positive integer n, we have

$$\frac{a_{n+1}}{a_n} = \frac{100(1.005)^{n+1}}{100(1.005)^n} = 1.005$$

Another way to express this is that every term in our sequence is 1.005 times the previous term. In symbols, $a_2 = 100(1.005) = 1.005a_1$, $a_3 = 100(1.005)^2 = 1.005a_2$, etc.

In general, in a geometric series with common ratio r and first term a_1, we know that $a_{n+1} = r \cdot a_n$ and that $a_n = a_1 r^{n-1}$. Notice that the exponent is one less than the number of the term.

Let's see how we can use the "exponential nature" of geometric sequences to help us find the sum. Rather than develop the formula only for our example, we will develop a formula for the sum of the first n terms of any geometric series $a_n = a_1 r^{n-1}$. We are looking for a formula for

$$S_n = \sum_{k=1}^{n} a_1 r^{k-1} = a_1 + a_1 r + a_1 r^2 + \cdots a_1 r^{n-2} + a_1 r^{n-1}$$

If we multiply both sides of this equation by r, and note that $r \cdot r^k = r^{k+1}$, we get

$$r \cdot S_n = a_1 r + a_1 r^2 + \cdot a_1 r^{n-1} + a_1 r^n$$

Notice that the right hand side of this equation is the same as in the previous equation except that the second has an additional term $a_1 r^n$ and is missing the term a_1. If we write the two equations together, lining up like terms, and subtract the second equation from the first, all the terms except these two cancel and we get the following.

$$\begin{array}{rcccccccccccc} S_n & = & a_1 & + & a_1 r & + & a_1 r^2 & + \cdots + & a_1 r^{n-2} & + & a_1 r^{n-1} & & \\ r \cdot S_n & = & & & a_1 r & + & a_1 r^2 & + \cdots + & a_1 r^{n-2} & + & a_1 r^{n-1} & + & a_1 r^n \\ \hline S_n - rS_n & = & a_1 & - & & & & & & & & & a_1 r^n \end{array}$$

Now, we can factor each side and solve this equation for S_n.

$$\begin{aligned} S_n(1-r) &= a_1(1-r^n) \\ S_n &= \frac{a_1(1-r^n)}{1-r} \end{aligned} \qquad (1)$$

This gives us our formula for the sum of a geometric series.

Your Nest Egg as a Geometric Series

Let's apply this to our retirement plan. We want to know the sum

$$S = 100 + 100(1.005)^1 + 100(1.005)^2 + \cdots + 100(1.005)^{478} + 100(1.005)^{479} + 100(1.005)^{480}$$

Our first term is $a_1 = 100$, $r = 1.005$, and $n = 481$. Substituting these values into (1), we have

$$S = \frac{100(1 - 1.005^{481})}{1 - 1.005} = \frac{100(1 - 11.01224)}{-0.005} = \$200,244.80$$

So, you would have saved $200,244.80 for your retirement. Is this enough?

As a quick estimate, suppose your account continues to pay 6% yearly interest, compounded monthly, and you want to live off the interest without taking any of the $200,244.80 principal or balance. If we take 0.005 times $200,244.80, we would get a monthly income of $1001.22 or $12,014.69 per year. This is not a huge amount to live on, especially considering that a dollar in forty years will be worth a lot less than it is today due to inflation. You might need to consider how much social security you will be entitled to, if there will be funds available to you from other sources (e.g. inheritance), and what you expect your expenses will be. For instance, perhaps you will have long since paid for a house, and so will have no rent or mortgage expense. A more comprehensive evaluation of your retirement plan would need to include these and other factors.

Owning a home is the biggest financial commitment most people make during their lives, and since it bears on a person's long term financial outlook in a number of ways, let's consider that next.

Owning Your Own Home

Sooner or later, you will probably want to leave the ranks of renters (if you haven't already) and become a home owner. There are many financial advantages to home ownership, besides the emotional and psychological benefits that most people experience. For one, a home typically appreciates in value over time so that, even if you only live in the home for a few years, you can probably sell it for more than you paid for it. It is in this sense that buying a home is considered an investment. Secondly, most people do not have enough cash on hand when they want to buy their first house, and so they must borrow the money by taking out a mortgage. While having to borrow money might seem like a bad thing to have to do, the good news is that all the interest you pay on the mortgage is tax deductible. In other words, you do not pay taxes on any funds that you use to pay mortgage interest. As we will see, this can save you a considerable amount of money. Finally, depending on which region of the country you live in, buying a home can result in a smaller monthly payment than renting a similar residence.

The first step in buying a house is to find out how much you can afford to borrow, given your household income and the funds you have available for a down payment. Mortgage lenders typically will not write mortgages where the monthly payment is more than 40% of monthly income, and usually recommend a ratio of no more than 30%. The monthly payment should include real estate taxes and home insurance, in addition to the payment on the loan. Knowing the amount you can afford for a monthly payment, the duration of the loan, and the available interest rate determines how much you can afford to borrow. Most home mortgages are paid off over 30 years or 360 months, but some may be for shorter periods, 15 years being common. We will assume payments are made monthly, which is the case in the vast majority of mortgages. Let's look at how this works mathematically.

We will develop the general formula relating the monthly payment, loan amount, loan duration, and interest rate. The basic idea is that, each month, you must pay interest on the remaining balance of the loan, using the interest rate i. You also want some of your monthly payment to go towards paying off your debt, so that you have less remaining to be paid off than the month before. After 360 months, or whatever the loan duration is, all the balance is to be paid off.

The details are a little messy, but are really just based on common sense, the basic idea of interest, and what we already know about geometric series. We will use the following notation for the variables involved so that we can develop our formula.

$$\begin{aligned} L &= \text{Loan amount} \\ M &= \text{Monthly payment} \\ i &= \text{monthly Interest rate} \\ N &= \text{Number of months duration} \end{aligned}$$

Also, we will let $B(k)$ stand for the balance of the loan remaining to be paid off after k monthly payments have been made. $B(0) = L$, the amount remaining to be paid before any payments are made. $B(N) = 0$, since the whole loan is to be paid off after N months.

Now, at the end of the first month, you will pay the bank M dollars, since M is your monthly payment. Of this, iL dollars goes to pay interest, and the remaining part of your payment, $M - iL$, goes to pay off the principle. So, after this first payment, the remaining balance is

$$B(1) = L - (M - iL) = L + iL - M = L(1 + i) - M$$

Now, for the second month, you will pay interest on the current balance $B(1)$. So, out of your second monthly payment, the amount going for interest is $i \cdot B(1) = i(L(1 + i) - M)$. The amount going to pay off the loan is thus $M - i \cdot B(1)$. Subtracting this from $B(1)$ gives the remaining balance after 2 payments, in other words $B(2)$. We have

$$\begin{aligned} B(2) = B(1) - (M - i \cdot B(1)) &= B(1) + i \cdot B(1) - M \\ &= B(1)(1 + i) - M \\ &= (L(1 + i) - M)(1 + i) - M \\ &= L(1 + i)^2 - M(1 + i) - M \end{aligned}$$

Let's do one more month. For the third payment we have

$$\begin{aligned} \text{Amount of interest paid} &= i \cdot B(2) \\ \text{Amount of principal paid} &= M - i \cdot B(2) \\ \text{Remaining balance} = B(3) &= B(2) - (M - i \cdot B(2)) \\ &= B(2) + i \cdot B(2) - M \\ &= B(2)(1 + i) - M \\ &= [L(1 + i)^2 - M(1 + i) - M](1 + i) - M \\ &= L(1 + i)^3 - M(1 + i)^2 - M(1 + i) - M \\ &= L(1 + i)^3 - M((1 + i)^2 + (1 + i) + 1) \end{aligned}$$

Now, it looks like we have a pattern that we can use to find $B(k)$ for any k.[130] After k payments, the remaining balance would be

$$B(k) = L(1 + i)^k - M[(1 + i)^{k-1} + (1 + i)^{k-2} + \cdots + (1 + i) + 1]$$

Since we want the loan to be paid off after N months, we would have $B(N) = 0$. This would be true if and only if

$$L(1 + i)^N = M((1 + i)^{N-1} + (1 + i)^{N-2} + \cdots + (1 + i) + 1) \tag{2}$$

[130] We are assuming that this pattern will continue for all possible k. Although you might be convinced that our formula is OK, we should note that we have not *proven mathematically* that our formula works. For now, we will simply note that this can be done, although we will not go into the details here

Finally, we will use what we know about geometric series to simplify. Ignoring the M, the right hand side of the equation above is a geometric series, written "backwards," with common ratio $r = (1+i)$, first term 1, and a total of N terms. Using our formula (1) for the sum of a geometric series, we have that

$$\begin{aligned}(1+i)^{N-1} + (1+i)^{N-2} + \cdots + (1+i) + 1 &= \tfrac{1(1-(1+i)^N)}{1-(1+i)} \\ &= \tfrac{1-(1+i)^N}{-i} \\ &= \tfrac{(1+i)^N - 1}{i}\end{aligned}$$

Substituting this into equation (2) above, we get

$$L(1+i)^N = M \cdot \frac{(1+i)^N - 1}{i}$$

Since, in our case, we will know the monthly payment M we can afford, we will solve this for L. We have

$$L = M \cdot \frac{(1+i)^N - 1}{i \cdot (1+i)^N} = M \cdot \frac{1-(1+i)^{-N}}{i} \qquad (3)$$

In other cases, we may wish to solve the equation for M in order to find the monthly payment for a given loan amount. We get

$$M = L \cdot \frac{i}{1-(1+i)^{-N}} \qquad (4)$$

This was a little bit of work, but now we have a formula which we can use to find out how much we can borrow knowing the interest rate, the amount we can afford to pay monthly M, and the duration of the mortgage N. Let's say we are making \$45,000 per year and that we want to use at most 30% of our income to make our monthly payments. Our monthly income is $\$45,000/12 = \$3750$ and 30% of this is \$1125. Let's suppose real estate taxes for the types of houses we might consider are around \$100 per month and that insurance will cost about \$25 per month. This leaves us \$1000 per month to pay off the loan. This will be our M.

Interest rates have fluctuated substantially over time, from recent lows of around 6% to over 12%. As an example, let's assume an annual interest rate of 9%, which translates to a monthly rate of $i = 0.0075$. If we plan on a standard 30 year mortgage, the duration will be $N = 360$ months. The loan amount we can finance would thus be

$$L = 1000 \cdot \frac{1-(1.0075)^{-360}}{0.0075} \approx \$124,282$$

We can afford a house of about \$125,000.

We can also use the formula for the geometric series $(1+i)^{N-1} + (1+i)^{N-2} + \cdots + (1+i) + 1$ to simplify our expression for $B(k)$. We have

$$\begin{aligned}B(k) &= L(1+i)^k - M[(1+i)^{k-1} + (1+i)^{k-2} + \cdots + (1+i) + 1] \\ &= L(1+i)^k - M \cdot \frac{(1+i)^k - 1}{i} \qquad (5)\end{aligned}$$

Before we go on to consider the investment aspects of owning a home, it is worth considering, and is surprising to most people, the interest paid for such a mortgage. In our example, you would be paying \$1000 per month every month for 30 years to pay off the mortgage. This means you will pay a total of \$360,000 for your house, which cost \$124,282. This is almost three times the price of the house!! In interest, you will pay $360,000 - 124,282 = 235,718$ dollars. How is this possible? Perhaps more importantly, how can this possibly be a good thing to do?

To give a partial answer to the first question, note that, out of your first \$1000 payment, you are going to pay $0.0075 \cdot 124,282 = \$932.12$ in interest, leaving only \$67.88 to go towards paying off your debt. This may seem surprising, but it is in general true that on a 30 year mortgage, only a very small fraction of your first payment (in fact, all of the payments for many years) goes towards paying off the principal of the loan. You can use the formula for $B(k)$ given above to experiment and find out how slowly the balance goes down during the initial years of the mortgage.

On the other hand, recall that you do not have to pay taxes on the interest you pay, whereas if you were renting, you have to pay taxes on all the money you would use to pay for housing. Even if you are in the lowest current tax bracket, this means you will save 15% of the amount you spend on interest. In our example, this means you save $(0.15) \cdot \$235,718 = \$35,357.70$ in taxes. If you are in the 28% tax bracket, you would save almost twice as much.

Also, keep in mind that you would pay for housing, whether you rent or buy. If you were renting, you might pay a total of \$360,000 or even more over 30 years, and not get any tax break, and you would own nothing as a result of your expense. After paying off your mortgage, you now own "free and clear" a house that was worth \$125,000 when you bought it, and may well be worth several times that amount 30 years later.

Housing as an investment

How much can you expect your house to grow in value over time?

In the United States, the Office of Federal Housing Enterprise Oversight (OFHEO) publishes a Housing Price Index (HPI) which tracks how the value of a typical single-family home changes over time in various regions of the country, as well as state by state. You could use the HPI for the region where you live to get at least some idea of how your house might appreciate in value, whether you plan to live there for 30 years, or only a few years.

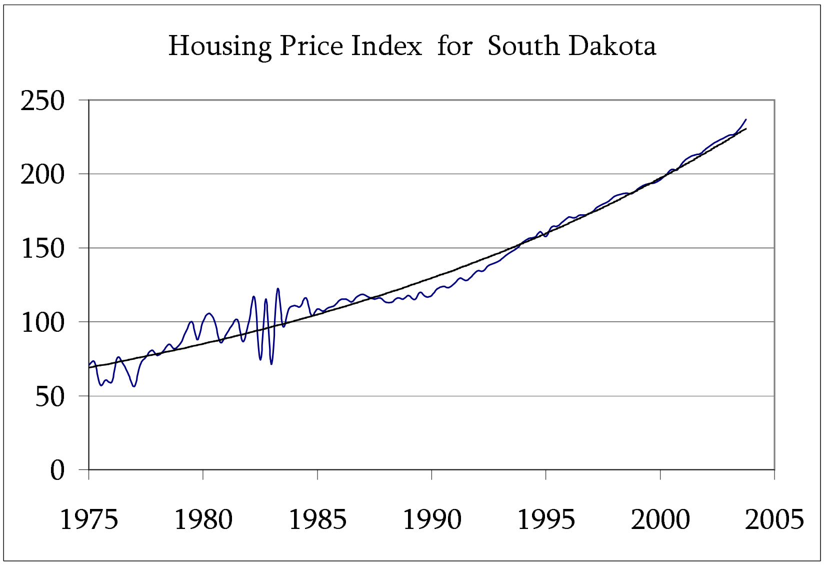


Figure 2: Housing Price Index for South Dakota

Figure 2 shows the HPI for South Dakota along with an exponential regression of the data. Although there was quite a bit of volatility in the HPI during the early 1980's, the exponential curve seems to fit the data very well. The equation for the exponential regression curve is $y =$

$68.265(1.04287^x)$, which indicates that housing prices in South Dakota have increased at about 4.287% per year.

Suppose you buy your house in South Dakota for $125,000. Assuming the same annual average growth rate of 4.287%, in 30 years, your house would be worth $\$125,000(1.04287^{30}) \approx \$440,370$. If you sold your house at this time, you would have a considerable sum of money which you could either use to buy another house, or put into savings for your retirement. If you had already raised a family, you may want to live in a smaller, less expensive home. In this case, perhaps you could buy a home using a portion of the $440,000 and use the remainder as savings. Alternatively, you may decide to go back to renting.

In any case, as a result of your "investment" of $1000 per month for 30 years, you would have close to half a million dollars *and* you would have had a place to live for that period of time. Now, you might think you could have done better by just investing $1000 every month in some sort of savings plan, and this is probably true, but you would still have to pay for a place to live. If you'd like to consider another scenario for comparison, you could calculate what would happen if you scrimped on your living conditions and rented a residence for something less than $1000 per month for the 30 years, and then saved the extra money in a savings plan. One of the activities will outline this option.

Conclusion

In this chapter, we have developed easy ways to add up values of arithmetic and geometric sequences. These correspond to linear and exponential functions respectively. Since linear and exponential functions are so important and so common, this allows us to "add things up" in lots of situations that we might run across in the real-world.

However, it is worth remembering that there are lots of (important) functions that are not linear or exponential. Could we find an easy way, for example, to add up the consecutive values of a power function like $f(x) = x^2$? What about logarithmic functions?

In general, it is unfortunately much more difficult to find sums when the sequence is neither arithmetic nor geometric. Our last resort, of course, is to just add up the specific series we have "by hand." Fortunately, spreadsheets and even graphing calculators can make this tedious task much more feasible, and there will be cases where we will have to resort to these alternate methods. The important thing is to be able to recognize when you have an arithmetic or geometric series and when you do not, so that you know which techniques will work, and which will not.

Reading Questions

1. What is a sequence? How is a sequence the same or different than a function, as functions are described in Chapter Two?

2. Describe how you could tell whether or not a sequence is an arithmetic sequence.

3. Describe how you could tell whether or not a sequence is a geometric sequence.

4. For each sequence given below, decide whether the sequence is arithmetic, geometric, or neither.

(a) $a_n = 0.6n + 1.3$
(b) $a_n = 0.2n^{1.5}$
(c) $a_n = 0.75^n$
(d) $a_n = (-1)^n(1.5^n)$
(e) $8, 6.4, 5.12, 4.096, \cdots$
(f) $3, 7, 11, 15, \cdots$
(g) $a_1 = 1000$ and each term is 1.0075 times the previous term.
(h) $a_1 = a_2 = 1$, and for $n > 2$, each term a_n is the sum of the previous two terms.
(i) $a_1 = 2.5$ and each term is 0.15 more than the previous term.

5. Write out the following sums in expanded form.

(a) $\sum_{i=1}^{5}(1.5i + 1)$
(b) $\sum_{k=3}^{6}(1.5k + 3)$
(c) $\sum_{n=0}^{7}(2^n)$
(d) $\sum_{m=4}^{8} 3(1/2)^n$

6. Find the following sums. Note that some of the series may be arithmetic, some geometric, and some neither.

(a) $1 + 4 + 7 + 10 + \cdots + 601$
(b) $1 + 4 + 9 + 16 + \cdots + 100$
(c) $\sum_{i=1}^{120} 50(1.085^i)$
(d) $1 + 1/2 + 1/4 + 1/8 + \cdots + 1/2^{12}$
(e) $\sum_{k=1}^{32}(1.5k + 1)$
(f) $\sum_{k=7}^{45}(1.5k + 1)$

7. In Chapter Five, one of the applications concerned annual carbon emissions from fossil energy consumption by the U.S. commercial sector as a function of time. The linear regression line found there was $y = 21.9x + 757$, where x was the number of years since 1990 and y was carbon emissions in millions of metric tons for that year.

(a) Explain why using the sum of an arithmetic series would be an appropriate way to estimate the total amount of carbon emissions over a period of several years.
(b) Assuming the linear trend continues, estimate the total amount of carbon emissions from 1990 through 2020.

8. Is the sequence $8, 8.8, 9.68, 10.648, \ldots$ arithmetic, geometric, or neither.

9. One of the following three sequences is arithmetic, one is geometric, and one is neither. Say which is which.

(a) $8, 12, 30, 45, \ldots$
(b) $4, 8, 14, 22, \ldots$
(c) $24, 18, 12, 6, \ldots$

10. Calculate the loan amount L in each of the following circumstances.

(a) The monthly payment $M = \$750$, the *annual* interest rate is 6%, and the term of the loan is 30 years.
(b) The monthly payment $M = \$750$, the *annual* interest rate is 12%, and the term of the loan is 30 years.
(c) The monthly payment $M = \$750$, the *annual* interest rate is 6%, and the term of the loan is 15 years.
(d) The monthly payment $M = \$750$, the *annual* interest rate is 12%, and the term of the loan is 15 years.

11. Based on your calculations in the previous question, is it true or not true that doubling the interest rate on a mortgage cuts the amount you can borrow in half?

12. Suppose you want to buy a new pick-up truck that costs \$30,000. The financing company offers you a 5-year loan at an annual interest rate of 7.78%.[131] You have \$4,000 cash on hand to put towards the truck and want to borrow the remaining cost. What would your monthly payment be?

13. Suppose you want to buy a nice home on the local lake for \$280,000. You plan to have \$60,000 cash on hand from the sale of your current home, and need to borrow the rest. Annual mortgage interest rates are at 6.2% for a 30-year mortgage and 5.5% for a 15-year mortgage.

(a) Calculate your monthly payment for each type of mortgage.
(b) In each case, calculate the total amount you would spend in interest over the duration of the loan.
(c) In each case, calculate the balance remaining to be paid off after 10 years (120 payments) using equation (4) in the text. Do these results surprise you?
(d) Assuming L, M, and i as above, graph $y = B(k)$ for both the 15 and the 30 year mortgages. Draw a sketch of your graphs on the same axes.
(e) After 15 years, the 15-year mortgage will be paid off. What is the remaining balance on the 30-year mortgage at this time?

14. Suppose you want to start saving for your retirement.

(a) How much would your total balance be if you saved \$100 per month for 40 years in an account that paid 9% annual interest?
(b) How much would you have to save each month, at a 9% annual interest rate, in order to amass 1 million dollars after 40 years?
(c) How much would you have to save each month, at a 6% annual interest rate, in order to have 1 million dollars after 40 years?

15. What is a BTU?

16. If you drive a mini-van that gets 20 miles per gallon, how many BTU's do you need to drive 200 miles?

[131] Rates quoted at http://www.bankrate.com/brm/static/rate-roundup.asp as of August 18th, 2007

Adding Things Up: Activities and Class Exercises

1. **Retirement Savings.** In our retirement planning example in the text, we calculated that saving \$100 per month at a bank or other financial institution for 40 years at a fixed 6% interest rate would result in an accumulated balance of \$200,244.80. If you lived off the interest, at the same 6%, you would have \$12,014.69 of income per year (\$1001.22 per month), and would never use the \$200,244.80 principal.

   On the other hand, you are not going to live forever, and if you don't care about leaving any inheritance behind, you could use the interest and perhaps some of the principal each month in such a way as to have nothing left after, say, 30 years (by which time you estimate there is a high probability of being dead). How could you do this?

   You can think of this situation as a mortgage, where *you* are the one loaning the bank \$200,244.80, and the bank will be paying you back over 30 years. After 30 years of monthly payments, the bank will have paid you off; in other words, your principal would be gone. The monthly payment will be what the bank pays you each month for your monthly income.

   (a) Calculate your monthly income under this scenario. How much more do you get per month than in the previous case where you did not touch the principal.

   (b) What if you "plan" on living longer, and want the income to be paid out over 40 years? Calculate your monthly income in this case.

   (c) For a long term investment plan, like the 40 year savings plan in our example, you may decide to invest in mutual funds, which do not have a fixed rate of return, but which have historically grown at higher *average* rates, typically around 10%, rather than the 6% we originally used. Calculate the accumulated amount of saving \$100 per month for 40 years at a 10% annual rate.

   (d) When you retire (or even before) you may want to reduce the risk level of your investments and so may decide to put your accumulated balance in an account that has a guaranteed, albeit lower, fixed interest rate. Calculate how much income your new, larger balance would generate at 6% interest assuming

      i. You never touch the principal
      ii. You have the principal paid out over 30 years
      iii. You have the principal paid out over 40 years

2. **Energy Production.** Figure 1 of Chapter Two showed a graph of U.S. Oil Production. Production peaked in 1970 and has been generally falling since. The graph below shows production from 1985 to 2002. The dependent variable here is given in billions of barrels per year.

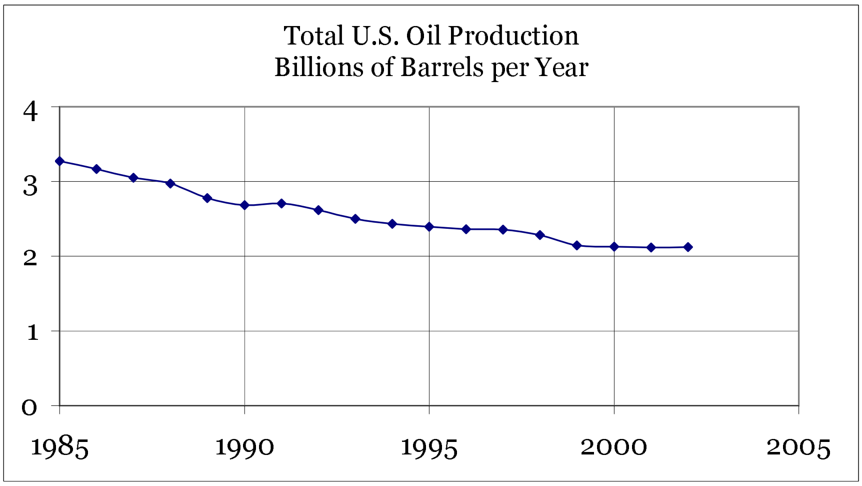
   For this activity, we will assume that production declines exponentially. The table below gives the data for the years 1985 to 2002.

Year	1985	1986	1987	1988	1989	1990	1991	1992	1993
Production	3.27	3.17	3.05	2.97	2.78	2.68	2.71	2.62	2.50
Year	1994	1995	1996	1997	1998	1999	2000	2001	2002
Production	2.43	2.39	2.36	2.35	2.28	2.15	2.13	2.12	2.12

   (a) Find the exponential regression for this data. How well does the exponential model seem to fit the data?

   (b) What does your model predict for the year 2000? Also, what does your model predict for the year 2050?

   (c) By adding the yearly productions, we can find the total production over any number of years. What type of series would we have if we were summing values of our exponential model?



(d) Estimate the total U.S. oil production for 2000 through 2050.

(e) Geometric series which have common ratio between −1 and 1 have the property that you can add an *infinite* number of terms and still get only a *finite* sum. In our context, if U.S. production continued forever, there would still only be a finite amount of oil produced. To see why this is, recall that the sum of geometric series is given by

$$S_n = \frac{a_1(1 - r^n)}{1 - r}$$

i. If r is between 0 and 1 and n is a large exponent (say, 100), what happens to r^n? Try this for the r in your exponential model.

ii. If $r^n \approx 0$, then the sum is approximately $S = \frac{a_1}{1-r}$. We call this the sum of the *infinite* geometric series. Estimate the total U.S. production from now on by finding the infinite sum of the geometric series, starting in the year 2000. How much bigger is this than the total production from 2000 to 2050?

3. Housing versus Saving as an Investment. In the text, we calculated that at a 9% annual rate of interest, we could buy a house worth about $125,000 with a monthly payment (not including taxes and insurance) of $1000. We also estimated that this house, located in South Dakota, would be worth about $440,000 in 30 years. In this activity, you will investigate what will happen if you decide to spend less on the house and invest the difference. Let's suppose you decide to buy housing at $800 per month and invest $200 per month in a savings plan.

(a) What 30-year loan amount can you afford at a monthly payment of $800 assuming a 6% annual interest rate?

(b) Assuming you buy a house for the loan amount you found in (a), and assuming the same 4.287% percent growth rate in the value of your house as in the text, how much will your house be worth in 30 years? (Hint: You should not have to recalculate 1.04287^{30} to do this).

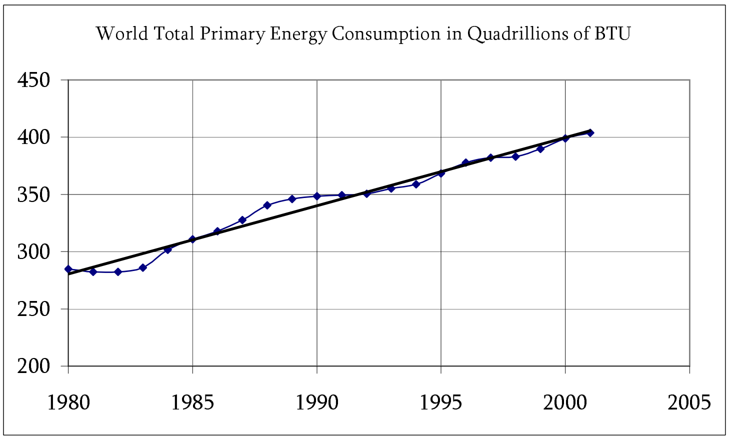
(c) Now, if you invest $200 per month at 9% compounded monthly for 30 years, how much will the total balance be at the end of the 30 years?

(d) Combining parts (b) and (c), what are your total assets between the house and the savings plan after the 30 years?

(e) Now suppose you had spent the entire $1000 per month on housing. Calculate the value of the house you could buy at 6% interest and how much this house would be worth after 30 years at the same 4.287% growth rate. How does this compare with your total in part d).

(f) Can you think of any other information or assumptions that we haven't included here that might be included to make our comparison more realistic?

4. World Energy Consumption. The graph below shows estimated *annual* total world energy consumption in quadrillions of BTU.



(a) A linear regression line is included with the scatterplot. From the graph, estimate the equation of this line. You could use $x = 0$ to represent the year 1980.

(b) Estimate the *total* amount of energy consumed for all the years 1980 through 2000.(Hint: This will involve computing the sum of an arithmetic series)

(c) According to figures from the Department of Energy, total world known reserves of crude oil, natural gas, and coal as of 2002 would give an energy equivalent of roughly 40,000 quadrillion BTU.[132] Assuming continued linear growth in annual consumption according to your equation from part (a), how long will it be before the world consumes a *total* of an additional 40,000 quadrillion BTU's?. What year will this be?

i. Since we are concerned with the total consumption over a number of years, and not just a single year's consumption, we will want to use the formula for the sum of an arithmetic series

$$S_n = \frac{n(a_1 + a_n)}{2}$$

However, in this case, we are given the sum or total which is $S_n = 40,000$. What is our unknown in this situation?

ii. We can consider a_1 to be the annual production in 2002. Use your linear regression equation to calculate this value.

iii. We do not know a_n since we do not know n, but we can use our linear regression equation to express a_n in terms of n. Do this, and substitute the expression into the sum formula for a_n. Then, solve the equation for n (Hint: the equation is a quadratic equation, so you may need to do some algebra and then use the quadratic formula). What year corresponds to the solution you found for n.

iv. In the year from the previous question, what will be the *annual* consumption, according to your linear model?

[132] Accessed from www.eia.doe.gov. However, these estimates are compiled from a number of sources, including *Oil & Gas Journal*, and are very tentative in nature, even though they are referred to as proved reserves

v. Now, suppose that by the time we reach the year from part (i), we have discovered another 40,000 quadrillion BTU's of recoverable energy. If we assume the same linear growth in consumption, how long before this additional 40,000 BTU's is used up?

(d) Assuming your projections are reasonable, what recommendations would you make with regards to future world energy policy?

12. Probability

When you have eliminated the impossible, what ever remains, however improbable, must be the truth.
—Sir Arthur Conan Doyle, in *The Sign of Four*

I will never believe that God plays dice with the universe.
—Albert Einstein

Probability, and the related field of statistics, have become increasingly important and relevant in modern society. Here are a few examples of situations where the idea of probability plays a role.

1. **Medical Testing:** You have tested positive for a dangerous, though not too common disease (HIV for example). How concerned should you be? (In other words, what are the chances you actually have the disease?)

2. **DNA Testing:** The term "DNA fingerprinting" first appeared in the public lexicon around 1985, and has been attributed to British geneticist Alec Jeffreys. Since that time, the technique has been used in the U.S. more than 50,000 times. Typically, one might read statements like "the probability that a randomly selected person would have a DNA profile matching the defendant's is one in a million" or some such very small probability. The implication often made is that the defendant is almost surely guilty based on the DNA evidence alone. Is this really a valid conclusion? In other words, if the lab says there is a DNA match between the defendant and the sample provided by the police, how likely is it that the defendant is really guilty?

3. **Lotteries:** As of 2001, 37 states had some form of government sanctioned lottery, and about $34 billion was spent by Americans on tickets that year.[133] What are your chances of winning the jackpot in a particular lottery? How much do lottery players win in prizes, on the average, for each dollar that is spent? How much revenue do states generate from lotteries?

In this chapter, we will consider the basic concepts of probability, as well as how probability can help us understand a variety of real-world situations, including those mentioned above.

The most basic examples of probability involve a **random trial** or **experiment**, a set of **equally likely outcomes**, and a particular **event** consisting of some subset of these outcomes. If the experiment is rolling one (unbiased!) die, there are six equally likely outcomes which we can list $\{1, 2, 3, 4, 5, 6\}$. Mathematicians would call this set of all possible outcomes the **sample space** of the experiment. Now, if the event is described as "rolling a prime number," then the outcomes related to this event are $\{2, 3, 5\}$ and the probability of this event is $3/6 = 1/2$. If the event is "rolling a 4," then the probability is $1/6$.

The "equally likely outcomes" condition is very important. For example, if you are considering what tomorrow might bring with respect to precipitation, the list of outcomes might be

$$\{\text{rain}, \text{snow}, \text{sleet}, \text{hail}, \text{freezing rain}, \text{none}\},$$

but these are not equally likely outcomes. It would certainly not be correct to say, on the basis of this list, that the chance of rain for tomorrow is $\frac{1}{6}$ or that the chance of some form of precipitation is $\frac{5}{6}$.

In the examples so far, we have given the probabilities as fractions, and this is a very common way to represent probabilities. We can also represent probabilities as either **percentages** or **relative frequencies**.

As a percentage, the probability of rolling a prime number when rolling a single fair die is 50% or .50. As a relative frequency, we could have said that we expect to roll a prime number 3 out of every 6 times.

[133] Data from *The Statistical Abstract of the United States: 2001.*

It is important to remember that probability can never tell you for sure what will happen in any given situation, unless you know for sure that the probability is 100%. So, when we say "we expect to roll a prime number 3 out of every 6 times," we can really only expect this to happen in some "on the average" sense. In our next six rolls, we could get anywhere from 0 up to 6 prime number rolls.

Three Types of Probabilities

We will classify a probability as one of three different types, depending on what information is used to calculate its numerical value. These types are **empirical**, **theoretical**, and **subjective**.

An empirical probability is one that is calculated on the basis of collected data. For example, suppose that medical researchers tested a new pain relief medicine on 2000 individuals and found that 1827 experienced at least some pain relief. We have $\frac{1827}{2000} = .9135$. The researchers might report that the probability that the drug will provide pain relief is over 91%.

Our dice probabilities are examples of theoretical probabilities. Here, we knew (or assumed) we had a fair die so that all outcomes were equally likely and simply counted the total number of outcomes and the total number of outcomes in the event. We did not conduct any empirical tests or collect any data. If each item in the sample space has an equally likely chance of occurring, the **theoretical probability** of an event is defined as

$$\frac{\text{the number of outcomes in an event}}{\text{the total number of outcomes in the sample space}}.$$

Note that the numerator of this fraction must be at least 0, but is no larger than the denominator since the outcomes in any event must also be contained in the sample space. So, the theoretical probability of an event will always be greater than or equal to zero and less than or equal to one. A probability of zero means that the event will never happen and a probability of one means that it will always happen.

If you did not know the die was fair, you could calculate probabilities for this particular die empirically by rolling it a whole bunch of times and counting how many times it came up in each of the six outcomes. If you rolled it a 200 times, you might put your results in a table like that in Table 1. For example, based on these data, we would say the empirical probability of rolling a 4 with this die is 20.5%.

Outcome	1	2	3	4	5	6
Frequency	25	22	6	41	32	74
Probability	.125	.11	.03	.205	.16	.37

Table 1: Empirical probabilities for an evidently weighted die

When a list of outcomes together with their probabilities are given in table form, as in Table 1, we say we have a **probability distribution**. In any such distribution, all possible outcomes should be listed and the probabilities must add up to one.

Empirical and theoretical probabilities are both legitimate ways to calculate probabilities. However, you often run across examples where people are reporting probability numbers simply based on their best judgement. For example, a political pundit might say that the probability of the Republican presidential candidate winning the next election is better than 60%. A sports writer might say "the Lakers are a slam dunk to win the NBA title," implying that the probability that they will win the championship is 100% or nearly so. These are examples of subjective probabilities.

Subjective probabilities are generally not reliable. How reliable depends more on the knowledge and objectivity of the person involved than anything else. You might have more confidence in the 60% number given in the preceding paragraph if the person reporting this number had a great deal of experience in presidential politics (and was not a Republican), then if he or she was a conservative talk radio host.

Counting and Listing All Outcomes

Odometer Method

For theoretical probabilities, if we can easily count all the possible outcomes and the number of outcomes in the event of interest, then everything is great. On the other hand, this is often not the case. For example, if a lottery involves drawing 5 different numbers from 1 to 45, we cannot possibly hope to count the number of possible combinations "by hand." There are just too many. We need an easier way to count all the possible outcomes, without having to actually list them all.

Let's illustrate with an example. Suppose a student is going to a fast food restaurant for lunch. The student likes cheeseburgers, chicken sandwiches, and fish sandwiches. He likes french fries and onion rings. He likes coke, cherry coke, sierra mist and root beer. How many different lunch combinations are acceptable to this student?

A lunch combination has one sandwich, either french fries or onion rings, and one drink. We will first "list" all the possible food combinations and then we will "count" all the possible combinations. We will list all the possible food combinations by using the so-called the "odometer method." This method is analogous to how a car odometer keeps track or "counts" miles traveled. For simplicity, let's consider an odometer with 3 positions where each position is a number between 0 and 9 (10 possible numbers).

Leftmost (hundreds) position	**Middle (tens) position**	**Rightmost (units) position**
0,1,2,3,4,5,6,7,8,9	0,1,2,3,4,5,6,7,8,9	0,1,2,3,4,5,6,7,8,9
choose one	*choose one*	*choose one*

The initial reading of the odometer is 000. When the car has traveled 1 mile, the reading is 001. When the car has traveled 9 miles, the reading is 009. When the car reaches 10 miles, the odometer reads 010. As the car traveled the first 10 miles, the rightmost position on the odometer went from 0 to 9 then back to 0. As the care continues, the rightmost position will continue to cycle through these 10 digits.

Similarly, the middle position on the odometer will cycle through the digits 0 to 9 repeatedly, changing once every 10 miles or once for each complete cycle of the rightmost position. The same pattern occurs with the leftmost position. It will cycle through all 10 digits, 0 to 9, changing digits for each complete cycle of the middle position. In this way, the odometer would count through a total of 1000 miles, starting at 000 and ending at 999. After mile 999, it would presumabley cycle back to 000.

If we were to try and list the odometer reading after each mile of travel we could generate a list that would look like this:

000	001	002	003	004	005	006	007	008	009
010	011	012	013	014	015	016	017	018	019
020	021	022	etc..........						
etc									
570	571	572	573	etc.					
etc.									
990	991	992	...............etc...........						999

Note that if we look at the three numbers showing on the odometer, no "reading" is repeated in our list. For example, the reading 572 appears only once in our list. Thus, each of the 1000 different readings occurs once and only once. By listing the numbers in this organized fashion, we can see there are 1000 different numbers. We might also realize this total by noting there are 10 different options for each of the three positions, so there are $10 \times 10 \times 10 = 1000$ total odometer readings.

We can apply this same idea to our food combinations. Just as our car odometer has 3 "slots" or positions that are to be filled with numbers, we can think of food combinations as having three slots, namely the sandwich slot, the side order slot and the drink slot. For example, one "food odometer" reading would be as follows.

fish	french fries	sierra mist
(sandwich)	(side order)	(drink)

We could say that the "reading" (fish, french fries, sierra mist) means the student ordered this combination for his lunch. Since there are 3 options for the sandwich slot, two for the side order slot, and 4 for the drink slot, our food odometer will have a total of $3 \times 2 \times 4 = 24$ possible readings, and this will correspond to 24 possible food combinations. We can list these in an organized fashion as in Table 2.

Table 2: The students twenty-four lunch possibilities

(cheeseburger, french fries, coke)	(cheeseburger, french fries, cherry coke)
(cheeseburger, french fries, sierra mist)	(cheeseburger, french fries, root beer)
(cheeseburger, onion rings, coke)	(cheeseburger, onion rings, cherry coke)
(cheeseburger, onion rings, sierra mist)	(cheeseburger, onion rings, root beer)
(chicken, french fries, coke)	(chicken, french fries, cherry coke)
(chicken, french fries, sierra mist)	(chicken, french fries, root beer)
(chicken, onion rings, coke)	(chicken, onion rings, cherry coke)
(chicken, onion rings, sierra mist)	(chicken, onion rings, root beer)
(fish, french fries, coke)	(fish, french fries, cherry coke)
(fish, french fries, sierra mist)	(fish, french fries, root beer)
(fish, onion rings, coke)	(fish, onion rings, cherry coke)
(fish, onion rings, sierra mist)	(fish, onion rings, root beer)

In practice, we forego writing out the list. We simply use the logic of the odometer method to count how many possibilities there would be in a given situation without listing them. This method will work as long as we can identify our "slots," can count how many options there are for each slot, and can fill in each slot **independently** of how the other slots are filled in. Independently means, for example, that we can order any drink with chicken and fries, or any side order with cheeseburger and cherry coke, etc. There are no restrictions like "the student will not order onion rings with a fish sandwich." Independence implies that every possible combination is a valid one, and would appear in the list if we created the list. The odometer method of counting is also know as the **general counting method** or the **multiplication principle**.

To summarize, any time you have several **independent** choices to make (slots), the number of ways to make all of these choices is the product of the number of ways of making each individual choice. The choices are independent if how you make one choice does not effect your options in the other choices. In our example above, the choices are independent since you can choose and side item with any sandwich with any drink. If the restaurant had said you can only have fries if you do not choose the fish sandwich, then these choices would not be independent, and you might be stuck having to try and list out all the options by hand.

More on Theoretical Probabilities

We can apply the multiplication principle and other counting techniques whenever we are calculating theoretical probabilities. Recall that if each item in the sample space has an equally likely chance of occurring, the (theoretical) probability of an event is defined as

$$\frac{\text{the number of outcomes in an event}}{\text{the total number of outcomes in the sample space}}.$$

If our sample space is all the possible lunch combinations listed above, and the event is "eating a cheeseburger," we need to count the outcomes containing a cheeseburger and divide that by 24 (the total number of outcomes for our lunch combinations). Referring back to Table 2, we see that there are 8 combinations containing a cheeseburger. So the probability of eating a cheeseburger is 8 divided by 24 or 1/3.

Often, we will use function notation as a shorthand for denoting probabilities. In this context, we let the inputs E be events, and use P for Probability to denote the function. So $P(E)$ denotes the probability that E will occur. This will be a function, since given any sample space S and any event E in this sample space, there is only one fraction which will give the probability $P(E)$.

Dice Probabilities

Many classic probability problems deal with dice or cards. In fact, the foundations of probability were the result of determining how to fairly divide the winnings in a game of chance. Suppose you were determining probabilities for rolling a pair of dice. To answer this question, we need to start by finding the number of equally likely outcomes for the sum. You might guess that, since you could have a sum of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, or 12 showing on the two dice, there are eleven outcomes. However, these outcomes are not equally likely. There is only one way to get a sum of two (rolling two ones) while there are many ways to get a sum of seven (rolling a three and a four, two and a five, etc.). To predict what will happen if the process is repeated many times, the process must be broken down into equally likely outcomes. For rolling two dice, one gray and one white for example, there are 36 equally likely outcomes. We could calculate this using our general counting principle by noting that there are six equally likely outcomes for each die, so the total number of outcomes when rolling both dice is $6 \cdot 6 = 36$. This is not a huge number, and we have listed all 36 outcomes in Figure 1.

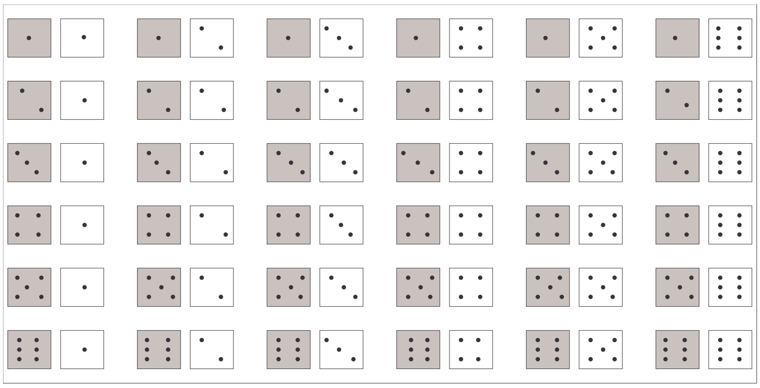


Figure 1: The sample space for rolling two dice.

From Figure 1, you can see that since there is one way to get a sum of two and 36 equally likely outcomes, the probability of rolling a two is $\frac{1}{36}$. You can also see that there are six ways to roll a sum of seven (these show up on the diagonal running from the bottom left to the upper right). Therefore, the probability of rolling a seven is $\frac{6}{36} = \frac{1}{6}$. (Notice that getting a total of 7 when the gray die comes up 3 and the white die comes up 4 is a different outcome from the gray die coming up 4 and the white die coming up 3.)

An event can be more complicated than rolling a given sum. For example, another possible event is rolling a sum that is less than 4. To find the probability of this event, we use Figure 1 to count the number of outcomes that give a sum which is less than 4. Since there are 3 of these (rolling two ones, rolling a one and a two, or rolling a two and a one), the probability of the event is $\frac{3}{36} = \frac{1}{12}$.

Oftentimes we can simplify probability calculations by using what is called the **complement** of an event. Given any event, the complement is defined to be the set of all outcomes in the sample space that are not in the event. For example, the complement of rolling a sum that is less than 4 is rolling a sum that is 4 or greater. Since the sample space has 36 elements and the event "rolling a sum less than 4" has 3 elements, the complement of that event will have $36 - 3 = 33$ elements. So, the probability of rolling a sum that is 4 or larger is $33/36$. In general, if we let E denote an event, E' will denote the complement of that event.

Since an event E and its complement E' include every outcome in the entire sample space with no overlap, we have $P(E) + P(E') = 1$ or equivalently $P(E') = 1 - P(E)$. For example, if E is the event " rolling a sum less than 4," then E' is "rolling a sum 4 or greater." We have $P(E') = 1 - P(E) = 1 - 3/36 = 33/36$ or $11/12$, just as we calculated above.

To summarize, we have two basic properties of probability:

- The probability of any event, $P(E)$, is a number between 0 and 1 (i.e. $0 \leq P(E) \leq 1$).
- For any event, E, and its complement, E', we have $P(E) + P(E') = 1$. This means that $P(E') = 1 - P(E)$.

Permutations and Combinations

The odometer method or multiplication principle works in general for any set of independent choices. Two special cases based on the multiplication principle are the counting of **permutations** and **combinations**. In both of these, we are looking at choosing a set of r items from a larger set of n items. The basic difference is that in permutations, there is an order or some other way of distinguishing between the r items, whereas in combinations, there is no order; we only care which r items we get.

Permutations

Let's say the campus Amnesty International chapter has 5 members, who we will list as $\{A, B, C, D, E\}$ for short. The club is required by student organization by-laws to elect a President, Secretary, and Treasurer. No person can hold more than one position. How many ways is there to do this?

What the club needs to do is an example of selecting a permutation. They are selecting $r = 3$ members from the set of all $n = 5$ club members, and there is an "order" to the selection. One of the three selected will by President, one will be Secretary, and one will be Treasurer. If A is President, B Secretary, and C Treasurer this is a different permutation than if B is President, A is Secretary, and C is Treasurer.

To calculate the number of possible permutations, we can use the multiplication principle (or equivalently, the odometer method). We have three choices to make, President, Secretary, and Treasurer. If we choose them in that order, we have 5 ways to choose the President, then 4 ways to choose the Secretary (since the President cannot also be the Secretary according to the by-laws), then 3 ways to choose the Treasurer. This gives us a total of $5 \cdot 4 \cdot 3 = 60$ ways to choose the club officers.

What if we did allow a person to hold more than one office?

Well, we would no longer call this a permutation, since we might be choosing 1, or 2, or 3 different people to be officers. However, we could still use the multiplication principle. The number of ways of doing this would be $5 \cdot 5 \cdot 5 = 125$.

If we listed all 125 ways of choosing the 3 officers allowing a person to hold more than one office, we might start AAA, AAB, AAC, etc. This list would include all 60 of the permutations we had previously calculated.

Combinations

Now, let us suppose the 5 member club wants to select a homecoming committee of 3 people. Here, there is to be no distinguishing between the 3 people chose. There is no President, Secretary, and Treasurer, just 3 people with the same role as committee members. This is an example of a **combination**.

We can find the number of these combinations by using our previous calculation of the number of permutations. One combination would be the set $\{A, B, C\}$. These 3 people could also have been chosen as President, Secretary and Treasurer. How many different times would this set of three people appear in our list of 60 ways to choose President, Secretary and Treasurer?

Well, if we know A, B, and C are going to be the people chosen, then any 3 of them could be the President. Again, assuming no doubling up on offices, any of the other 2 could be Secretary, and then the one remaining would have to be Treasurer. This means there are $3 \cdot 2 \cdot 1 = 6$ ways these 3 people could have been chosen for the 3 offices. Thus, these 3 people appear 6 times in the list of 60.

Now, the same would be true for any other list of 3 people ($\{A, C, D\}$ or $\{B, D, E\}$, etc.). In each case, any set of 3 people appears 6 times in the list of 60. So, if we want to count how many different sets of 3 people there are, there must be $60/6 = 10$ such sets. We can even list them as follows:

{A, B, C}	{A, D, E}
{A, B, D}	{B, C, D}
{A, B, E}	{B, C, E}
{A, C, D}	{B, D, E}
{A, C, E}	{C, D, E}

Calculational Shortcuts

Although we have illustrated the thinking behind calculating the number of permutations and combinations using the multiplication principle in the previous examples, in practice we will use our calculators to find these values. First, for any nonnegative integer n, we define $n!$, called ***n* factorial**, by

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1$$

So, for example, $3! = 3 \cdot 2 \cdot 1 = 6$ and $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$. We can then use factorials to calculate the number of permutations and combinations as follows:

- The number of permutations of r items selected from a set of n items, denoted $P(n,r)$, is

$$P(n,r) = \frac{n!}{(n-r)!}.$$

- The number of combinations of r items selected from a set of n items, denoted $C(n,r)$, is

$$C(n,r) = \frac{n!}{r!(n-r)!} = \frac{P(n,r)}{r!}.$$

Going back to our club example, the number of ways of picking the 3 officers from the club of 5 is

$$P(5,3) = \frac{5!}{(5-3)!} = \frac{5!}{2!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} = 5 \cdot 4 \cdot 3 = 60.$$

The number of ways of choosing the 3-person homecoming committee is

$$C(5,3) = \frac{5!}{3!(5-3)!} = \frac{5!}{3!2!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 2 \cdot 1} = \frac{120}{12} = 10.$$

Even better news, most calculators have buttons for $P(n,r)$ and $C(n,r)$, or have these built into a menu. For example, to find these on a TI graphing calculator, hit the MATH menu, and then choose PRB (short for probability), and you will find $P(n,r)$, $C(n,r)$ and $n!$. To calculate $P(5,3)$, go to the main screen and enter the 5, then select the $P(n,r)$ from the menu, then type 3 and ENTER. You should get 60.

Thus, your main task in using permutations and combinations is not computational. The important point will be to identify if the situation your are considering is *either* a permutation or a combination or *neither*. Remember that permutations and combinations are only *special cases* of all the possible counting problems that one might run across. To decide if the situation you are looking at is one of these, you might ask yourself the following:

1. Does this situation involve choosing a set of r items from a larger set of n items?
2. Is there or is there not an "order" or some other distinction being made between the r items being chosen. If so, you are looking at a permutation. If not, you are looking at a combination.

Probability Distributions

A finite **probability distribution** is the collection of all the outcomes of a random phenomenon together with their associated probabilities. For example, a probability distribution for rolling two dice and recording the sum is shown in Table 3. Recall that these probabilities are theoretical probabilities, assuming fair dice and equally likely outcomes.

Table 3: The probability distribution for rolling two dice and recording the sum.

Outcome: Sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

The probability distribution shown in Table 4 lists the percentages for various religious preferences among 1,384,486 U.S. Armed Forces personnel, as of March 2005.[134] In one sense, these are simply percentages. However, if we consider the random experiment of selecting a U.S. service person and determining their religious preference, these percentages serve as empirical probabilities.

Table 4: The probability distribution for Religious Preferences of U.S. Armed Forces Personnel.

Denomination Name	Total Pct.
Roman Catholic Church	21.5%
No Religious Preference	19.6%
Baptist churches	14.7%
Christian, no denominational preference	14.0%
Unknown	8.1%
Protestant, no denominational preference	3.3%
Methodist churches	2.6%
Lutheran churches	2.4%
Southern Baptist Convention	1.3%
Church of Jesus Christ of Latter Day Saints (Mormon)	1.3%
All Other	11.0%

Note that the outcomes in Table 3 are represented by numbers, but the outcomes in Table 4 are religious preferences that have no numerical value. For the dice probabilities, it would make

[134]These data are compiled by the United States Department of Defense, and were accessed at http://speakingoffaith.publicradio.org/programs/servingallah/index.shtml

sense to ask what the “average roll” would be. For the religious preferences, the best we could do for an “average” would be to pick the denomination that had the highest percentage (the so called mode).

For probability distributions which have numerical outcomes, we calculate the **mean of a probability distribution** as follows. Multiply each outcome by the respective probability and add the products. For example, the mean of the probability distribution in Table 3 is

$$2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7.$$

This means that, although we cannot say what will happen on any particular roll, in the long run the average of our rolls should be 7 or at least close to 7.

Coin Flipping Distributions

Suppose three coins are flipped and we are interested in the event of getting a certain number of tails. Using the odometer method, we can think of each coin as a slot, with two choices (H or T) for each slot. Thus, there are $2 \cdot 2 \cdot 2 = 8$ possible outcomes, which we can easily list: HHH, HHT, HTH, THH, HTT, THT, TTH, TTT. If it helps, we can think of the coins as a penny, a nickel, and a dime, so heads on the penny and the nickel and tails on the dime – HHT – is different from heads on the penny and dime and tails on the nickel – HTH.

From the list of elements in this sample space, we can see that there is 1 way to get 0 tails, 3 ways to get 1 tail, 3 ways to get 2 tails and 1 way to get 3 tails. The probability distribution for flipping 3 coins and recording the number of tails is shown in Table 5.

Table 5: The probability distribution for the number of tails observed when tossing 3 coins.

Number of Tails	0	1	2	3
Probability	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

To find the mean of this probability distribution, multiply each outcome times its corresponding probability and then add, just as we did with the dice example.

$$\text{mean number of tails} = 0\left(\frac{1}{8}\right) + 1\left(\frac{3}{8}\right) + 2\left(\frac{3}{8}\right) + 3\left(\frac{1}{8}\right) = \frac{0}{8} + \frac{3}{8} + \frac{6}{8} + \frac{3}{8} = \frac{12}{8} = 1.5$$

Alternatively, we could calculate the mean by considering the the eight different outcomes, adding up the number of tails for each outcome, and dividing by 8.

$$\text{mean number of tails} = (0+1+1+1+2+2+2+3) \div 8 = \frac{12}{8} = 1.5$$

Notice we get the same answer either way. The first method is more efficient, and also makes the division by 8 part of the probability instead of an operation at the end.

Have we seen this idea and these types of sums before?

Yes, in fact, we have. If you go back to our discussion of GPA in Chapter Four, you will see that the way we calculated GPA is very much like what we are doing here with the mean of a probability distribution. Now, grades are not typically assigned at random (at least we hope not), but the idea of *mean* or *average* in both these situations is similar. In fact, if all the grades at your college *were* assigned at random, but according to some set percentages for each grade (10% of students in any class get an A, 25% get a B, etc.), then the mean of the probability distribution would represent the expected GPA of students at your college!

Lotteries

Lotteries of all sorts have become almost ubiquitous, and many state budgets depend heavily on lottery revenue. Let's consider one example from the state of Colorado. In one game, simply called Lotto, the player picks six numbers from the set of integers 1 - 42. The state then picks six "winning numbers." The players win prizes by matching some or all of the six winning numbers. The probability distribution where the outcomes are the number of matches with their associated probabilities is shown in Table 6.

Table 6: The probability distribution for the number of matches in the Colorado Lotto.

No. of matches	6	5	4	3	2	1	0
Probability	$\frac{1}{5{,}245{,}786}$	$\frac{216}{5{,}245{,}786}$	$\frac{9{,}450}{5{,}245{,}786}$	$\frac{142{,}800}{5{,}245{,}786}$	$\frac{883{,}575}{5{,}245{,}786}$	$\frac{2{,}261{,}952}{5{,}245{,}786}$	$\frac{1{,}947{,}792}{5{,}245{,}786}$
As Decimal	1.906×10^{-7}	4.118×10^{-5}	0.0018	0.0272	.1684	.4312	.3713

Notice that the probability of matching all six numbers is $\frac{1}{5{,}245{,}786}$. How bad are these odds? Imagine yourself in a long line of 5,245,786 people waiting to buy a single lotto ticket, and that one person in that line will be the winner. Such a line would go roughly from South Dakota to New York City! Your probability of winning is the same as the probability that you would get picked out of that line at random.

To find out how many numbers you would expect to match on average if you bought a ticket, we find the mean of the distribution using given in Table 6:

$$6\left(\frac{1}{5,245,786}\right)+5\left(\frac{216}{5,245,786}\right)+4\left(\frac{9,450}{5,245,786}\right)+3\left(\frac{142,800}{5,245,786}\right)+2\left(\frac{883,575}{5,245,786}\right)$$

$$+1\left(\frac{2,261,952}{5,245,786}\right)+0\left(\frac{1,947,792}{5,245,786}\right)\approx 0.86$$

This means that, on average, you will match 0.86 of the winning numbers. In other words, you can expect to match less than one number, on the average. Therefore, it should not surprise you if you buy a Lotto ticket and do not match a single number.

Of course, players are less interested in how many numbers they match than in how much money they win. We could calculate the money won, on average, from playing lotto by replacing the number of matches with the prizes one in Table 7. In this case, the mean is also called the expected value of the lottery. The jackpot starts at $1 million dollars, and increases each time no one matches all six numbers. The other prizes also vary somewhat. The probability distribution as of August First, 2007 was as is shown in Table 7.

Table 7: The probability distribution for prize winnings in Colorado Lotto.

No. of Matches	6	5	4	3	2	1	0
Winnings	$1,919,961	$516	$41	$3	$0	$ 0	$0
Probability	$\frac{1}{5{,}245{,}786}$	$\frac{216}{5{,}245{,}786}$	$\frac{9{,}450}{5{,}245{,}786}$	$\frac{142{,}800}{5{,}245{,}786}$	$\frac{883{,}575}{5{,}245{,}786}$	$\frac{2{,}261{,}952}{5{,}245{,}786}$	$\frac{1{,}947{,}792}{5{,}245{,}786}$

The expected value of the lottery, or in other words the mean of this probability distribution, is calculated as follows.

$$1,919,961\left(\frac{1}{5,245,786}\right)+516\left(\frac{216}{5,245,786}\right)+41\left(\frac{9,450}{5,245,786}\right)+3\left(\frac{142,800}{5,245,786}\right)+0\left(\frac{883,575}{5,245,786}\right)$$

$$+0\left(\frac{2,261,952}{5,245,786}\right)+0\left(\frac{1,947,792}{5,245,786}\right)\approx 0.543$$

So, the expected value for playing the Colorado Lotto is 54 cents. This means that you could expect to win, on average, just 54 cents per ticket. Because it costs $1 for each ticket, you lose 46 cents per ticket on average. So, if you bought 100 tickets, costing you $100 total, you should expect to win only about $54. This is not a particularly good deal for the overwhelming majority of lotto players, but it certainly helps fill the state's coffers.

Conditional Probability

DNA fingerprinting and other probability calculations used in court cases, medical testing for diseases, and many other examples of probabilities that one reads about are examples of what are known as **conditional probabilities**. Almost anytime a probability is qualified by some sort of "condition," it is a conditional probability.

For example, in 1998, the percent of all live births that were classified as low birthweight was 7.8%.[135] If we consider this number as a probability, we could say the probability that a randomly selected baby born in 1998 was of low birthweight was 7.8% (or 78 out of 1000 if you prefer relative frequencies).

However, for mothers who smoked, the percentage was 12.0%. We could say the probability that a randomly selected baby born in 1998 was of low birthweight, *given that the mother was a smoker*, was 12%. This is a conditional probability, with the condition being "given that the mother was a smoker."

Conditional probabilities can be a little cumbersome to write out in words. A shorthand way of denoting the probability that a given baby is low weight given that the mother smokes is

$$P(\text{low weight}|\text{mother smokes}).$$

So, we could write $P(\text{low weight}) = 0.078$, while $P(\text{low weight}|\text{mother smokes}) = 0.12$. The event we are interested in appears in front of the vertical line, and the condition appears after.

You might guess that both the conditional and the "unconditional" probabilities in this case are empirical probabilities. The 7.8% is the result of taking the total number of low weight births divided by the total number of births. The conditional probability of 12% was computed by dividing the number of low weight births *where the mother was a smoker* divided by the total number of births *where the mother was a smoker*. The difference is, for the conditional probability, we look at the same event, but restrict ourselves to the sample space given by the stated condition. In this first example the condition is "mother was a smoker."

Figure 2 gives a Venn diagram that might be useful in picturing this situation.

$P(\text{low weight})$ is the number of births represented by the "low weight births" circle divided by the total number of births, represented by the rectangle. The conditional probability

$$P(\text{low weight}|\text{mother smokes})$$

is calculated using only the "births to smoking moms" region. The numerator would be the number of births in the overlap of "low weight births" and '"births to smoking moms" divided by the number of births in the '"births to smoking moms" region.

In many cases, one could switch the event and the condition around and still have a probability that makes sense. For example, we might be interested in

$$P(\text{mother smokes}|\text{baby is low weight}).$$

This would be computed by dividing the total number of low weight babies born to mothers who smoke, by the total number of low weight babies. In the Venn diagram, the sample space for this probability would be represented by the "low weight births" circle.

How is $P(\text{mother smokes}|\text{baby is low weight})$ different than $P(\text{low weight}|\text{mother smokes})$? Note that, in both cases, the numerator of the probability fraction would be the number of low weight

[135]This and the following data are from *The Statistical Abstract of the United States: 2001*.

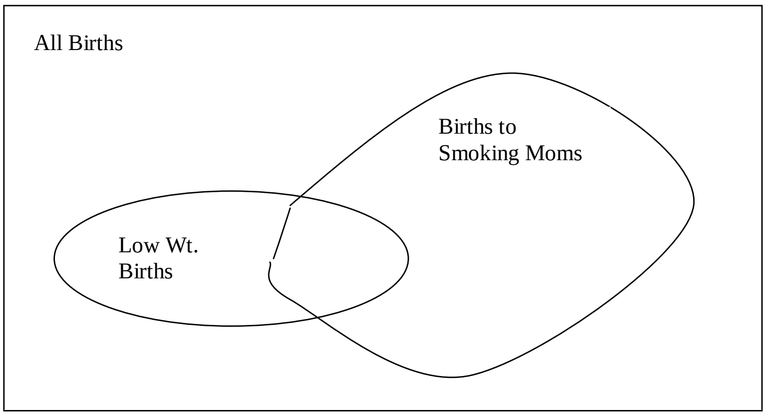


Figure 2: Venn diagram showing possible set relations for births example

babies born to mothers who smoke. What is different is the denominators of the probability fractions. In the first case, the denominator is the number of low weight babies, and in the second case the denominator is the number of babies born to mothers who smoke. These are two different groups of babies, but both of these groups contain the group counted in the numerator, namely the babies who are low weight *and* who are born to mothers who smoke.

Example 1. Suppose that, at a particular hospital, there were 25 births last month, and 6 of these were to women who smoked. Suppose that 1 of the six babies born to the smokers was low weight and 2 of the other babies was low weight. Find $P(\text{low weight}|\text{mother smokes})$ and $P(\text{mother smokes}|\text{baby is low weight})$. Also, find $P(\text{low weight}|\text{mother does not smoke})$ and $P(\text{mother does not smoke}|\text{baby is low weight})$.

Solution: For both $P(\text{low weight}|\text{mother smokes})$ and $P(\text{mother smokes}|\text{baby is low weight})$, the numerator is the number of low weight births to mothers who smoke, which is 1. We have

$$P(\text{low weight}|\text{mother smokes}) = \frac{1}{6} \approx 17\%.$$

$$P(\text{mother smokes}|\text{baby is low weight}) = \frac{1}{3} \approx 33\%.$$

For the other two probabilities, the numerator is 2, since there were two low weight babies born to mothers who do not smoke. The probabilities are

$$P(\text{low weight}|\text{mother does not smoke}) = \frac{2}{19} \approx 10.5\%.$$

$$P(\text{mother does not smoke}|\text{baby is low weight}) = \frac{2}{3} \approx 67\%.$$

■

Example 2. How do you think newspapers or television stations might report the situation with regards to low weight births at the hospital in this last example? Here are a few possible sample headlines (you might create some more of your own just for fun).

1. Local data show smoking increases chance of low weight birth 6.5%

2. Local data show smoking increases chance of low weight birth 62%

3. Smoking accounts for one third of all low weight births

4. Low weight babies twice as likely to be born to non-smokers

5. Babies and smoking: Is there anything wrong?

Hard to believe all these could be based on the same information! Explain the calculations and thinking behind each of these headlines. Which of them do you think is the most "accurate?"

Solution: Let's go through these each in turn.

1. We can get the 6.5% by subtracting the 10.5% probability of low weight birth given the mother does not smoke from the 17% probability for smokers. Is this a valid way to compare?

   One might make the case that this is not telling the whole story. After all, you would get the same 6.5% difference whether the given probabilities were 7% and 0.5%, or 56.5% and 50%, or the actual probabilities reported above. If the two probabilities are low the effects of smoking are very great, while if the two probabilities are high, the relative effect of smoking seems to be low. The main problem is that we don't know what the 6.5% compares to without having all the information.

2. This one was calculated by taking the 6.5% and dividing by the 10.5%. The advantage to this approach is that it gives you a sense of the *relative* effect of smoking versus non-smoking. The 62% calculated this way does tell you there really is a 62% greater chance of having a low weight baby if you smoke versus if you do not smoke. The headline worded this way is technically accurate.

   There is one very legitimate criticism that could be made with this approach, however. It does not take into account whether having a low weight baby is a very rare or a relatively common occurrence. In our example, there were 3 low weight births out of a total of 25 births which is a 12% overall rate. Suppose instead we had 10,000 smokers and 10,000 nonsmokers, and that there were 13 low weight babies among the smokers and 8 among the non-smokers. The difference is 5 low weight babies and 5/8 does give you 62%. We could say, accurately, that smoking increases your chances of a low weight birth by 62%. However, out of 10,000 smoking moms, you are only going to get 5 more low weight births than in the nonsmoking group. Five out of 10,000 is a very small percentage. So, even though the percentage increase in the risk is great (62%), the event is so rare that it might not be worth worrying about. The moral is that percent increases like this are more important the more common the event is.

3. The 33% is calculated by using the number of low weight births. There were 3 low weight births and one of them was to a smoking mom.

   This percentage is very misleading, and really does not provide much information by itself. The fact that one out of three low weight babies were born to smoking moms does not tell you anything unless you know what percentage of moms overall were smokers. In our case, we have 25 moms and 6 were smokers. Six out of 25 is 24%, and this is the percent of all moms who smoke. Knowing that this percentage (24%) is smaller than the 33% does tell you that smoking is correlated with an increased risk of having a low weight birth. Whenever you see a percentage like this, based on a sub-population, you should look for the percentage for the entire population. If this is not given, you really have information only about the one subgroup, in this case the smokers. You do not have any information about how the smoking groups compares with the nonsmoking group.

4. This is also based on the two low weight babies born to nonsmokers and the one born to smokers. While it is true that twice as many low weight babies were born to nonsmokers, the headline ignores that there are more than three times as many nonsmokers overall as smokers overall. As in the last item, you need to know the relative sizes of the larger populations. This headline would only be accurate if the number of smoking moms was the same as the number of nonsmoking moms.

5. We won't say too much about this one. This headline might be be the result of the same bad thinking that went into the last one.

So, what headline should we use? This author would humbly suggest something generic like "local hospital releases birth statistics." The details could then be a part of the article where they can be adequately explained. ■

Multi-stage Experiments

We have seen that determining the probability of an event by first writing down all the outcomes of the sample space and then counting the number of outcomes in the event works well when the random phenomenon is fairly simple and the number of outcomes is fairly small. When events become more complex, we used the multiplication principle to count the sample space and the number of outcomes in any event. In general, we have been calculating the numerator first, possibly by multiplying, and then dividing by the total number of possible outcomes. However, sometimes it is useful to first calculate individual probabilities, and then multiply to get the probability of an event in a multi-step process. Corresponding to our multiplication rule for counting, we have the following property of probability:

> *The probability of two or more independent events is the product of the associated probabilities.*

For example, suppose we roll a die twice and want to calculate the probability of rolling a 2 on the first roll and an even number on the second roll. We could look at the sample space in Figure 1 and count the number of outcomes for which this is true. If we think of the number on the left in each pair as the outcome for the first roll and the number on the right as the outcome for the second roll, we see that there are three outcomes in which the first roll is a 2 and the second is even. These are (2,2), (2,4), and (2,6). Therefore, the probability of this event happening is $\frac{3}{36} = \frac{1}{12}$.

On the other hand, we could compute this probability using our multiplication rule. The probability of rolling a 2 on the first roll is $\frac{1}{6}$. The probability of rolling an even number on the second roll is $\frac{3}{6}$ since there are 3 even numbers on a die. Since the outcome of the first die does not effect the outcome of the second, the two events are independent. So, we can multiply the probabilities of these two events to determine the probability of both events happening. Therefore, the probability of rolling a 2 on the first roll and an even number on the second is $\frac{1}{6} \cdot \frac{3}{6} = \frac{3}{36} = \frac{1}{12}$. Observe that $\frac{1}{6}$ of the outcomes in Figure 1 have a 2 on the first die and $\frac{3}{6} = \frac{1}{2}$ of the outcomes in Figure 1 have an even number on the second die. Therefore, $\frac{1}{6}$ of $\frac{1}{2}$ is $\frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$.

It is very important to remember that this multiplication rule for probabilities only works if the events are independent, just as with the multiplication rule for counting. Remember that two events are **independent** if the occurrence of one does not change the probability of the other occurring.

Not all events are independent. For example, the probability of you having blond hair is *dependent* on whether or not either of your parents has blond hair. You are more likely to have blond hair if one or both of your parents does. With dice, the events "rolling an odd number" and "rolling a prime number" are dependent. If you roll an odd number, 1, 3, or 5, you have a 2/3 probability of rolling a prime number. If you roll an even number, 2, 4, or 6, you only have a 1/3 chance of rolling a prime number.

> *The probability that we may fail in the struggle ought not to deter us from the support of a cause we believe to be just.*
> —Abraham Lincoln

Reading Questions for Probability

We encourage children to read for enjoyment, yet we never encourage them to "math" for enjoyment. We teach kids that math is done fast, done only one way and if you don't get the answer right, there's something wrong with you. You would never teach reading this way.
—Rachel McAnallen, from *Math? No Problem*, The Hartford Courant, October, 1998.

1. For each of the following probabilities, say whether the probability is empirical, theoretical, or subjective and why you think so. Numerical values are not given for these probabilities, and you need not do any calculations. Your answer should be based on how you think the probability is 'probably' calculated.

(a) The probability of rolling doubles when rolling two dice.
(b) The probability that a given U.S. citizen died from West Nile Virus during the last 30 days.
(c) The probability that a given U.S. citizen will die from West Nile Virus during the *next* 30 days.
(d) The probability that a Republican will win the next presidential election.
(e) The probability that you will win the jackpot in the next powerball lottery drawing.
(f) The probability that you will be selected for jury duty during the coming year.
(g) The probability that it will rain tomorrow.

2. For each of the following lists or descriptions of possible outcomes, say whether the outcomes in the list are equally likely or not, and why you think so.

(a) The outcomes are {heads, tails} when flipping one fair coin.
(b) The outcomes are all the possible combinations of numbers in the powerball lottery.
(c) The outcomes are the numbers {2,3,4,5,6,7,8,9,10,11,12} that are the sums you can get when rolling two fair dice.
(d) The three outcomes are {two heads, one head and one tail, two tails}, when flipping two fair coins.
(e) The outcomes are all the possible numbers one could get by counting how many U.S. citizens vote in the next presidential election.

3. For each pair of events, say whether you think the events are independent or not, and why you think so.

(a) You roll two fair dice. The two events are rolling a 5 on the first die, and rolling an even number on the second die.
(b) You roll two fair dice. The two events are rolling a 6 on the first dice, and rolling a sum less than 9.
(c) You select one U.S. citizen at random. The two events are the person has a college degree, and the person had an annual income of more than $40,000 during the last full calendar year.

4. Suppose a new car model comes with the following option: 3 choices of body style (2-door, 4-door or station wagon), 4 colors (black, red, green or white), and two choices of engine size (4 or 6 cylinder). List all the different ways the model could be ordered starting with (2,B,4) for a 2-door black 4-cylinder model.

5. How many different lunch choices would there be if you choose from 4 sandwiches, 2 side orders, and 5 drinks? [Note: You do not need to list all of the possible combinations.]

6. Why can a probability never be greater than 1?

7. If the probability it will rain tomorrow is 0.6, what is the probability of the complement of that event?

8. Suppose you are tossing 3 coins and want at least 2 heads.

(a) List all of the possible outcomes for the sample space.

(b) List all of the possible outcomes for the event.

(c) What is the probability of tossing 3 coins and having at least 2 heads?

9. You roll two fair dice.

(a) What is the probability that the sum is 7.

(b) What is the probability that the sum is greater than or equal to 7, but less than 10?

(c) What is the probability that the sum is 5 or 6?

(d) What is the probability that the sum is less than 6, or a prime number?

10. Suppose you have a jar with 5 white marbles and 8 black marbles.

(a) What is the probability of drawing a black marble?

(b) What is the probability of drawing a white marble?

(c) Find the sum of your answers from parts (a) and (b). Why is this true?

(d) Suppose you draw one marble, replace it, and draw a second marble.

i. What is the probability that both marble are black?

ii. What is the probability that both marbles are white?

iii. What is the probability that the first marble is black and the second one is white?

11. When rolling two dice, what is the probability that you get a three on the first die and a number greater than 3 on the second?

12. Calculate the following

(a) $C(5, 2)$

(b) $P(6, 3)$

(c) $C(6, 3)$

(d) $C(25, 10)$

13. Suppose the local College Republican group consists of 9 members.

(a) How many ways can the club choose 4 officers, President, VP, Secretary and Treasurer?

(b) How many ways can the club choose a publicity committee of 4 people?

(c) How many ways can the club choose a publicity committee of 5 people?

(d) Why is the answer to the last two questions the same!?

(e) Suppose the club has 5 men and 4 women. How many ways is there to choose a publicity committee of 4 people if they want to have equal gender representation on the committee?

14. What is a probability distribution?

15. Give an example, not mentioned in the reading, of a probability distribution where it is possible to find the mean.

16. Give an example, not mentioned in the reading, of a probability distribution where it is not possible to find the mean.

17. List the sample space for the experiment in which four coins are flipped. (Hint: There are 16 different outcomes.)

18. When flipping four coins, determine the following probabilities.

(a) Getting four heads.
(b) Getting at least 3 heads.
(c) Getting a head on the second coin.

19. Find the mean of the following probability distribution where the outcome is number of people in a family. Describe what your answer means.

Persons in Family	2	3	4	5	6	7
Probability	0.41	0.24	0.21	0.09	0.03	0.02

20. Aunt Hazel's Raisin Cookies were analyzed for their number of raisins. The following probabilities were found:

Number of Raisins, x	3	4	5	6	7	8
Probability of x Raisins	0.10	0.15	0.15	0.25	0.20	0.15

Find the mean number of raisins in a cookie.

21. Calvin is taking a test using the "Calvin Method" which involves flipping a coin and putting "true" for heads and "false" for tails. There are ten questions on the test

(a) What is the probability that Calvin gets all ten questions right?
(b) What is the probability that Calvin gets nine out of 10 correct? (Hint: there are $C(10, 9)$ ways to pick which 9 questions will be the correct ones).
(c) What is the probability that Calvin gets exactly 8 of 10 questions correct?

22. A family has three children.

(a) What is the probability that the two oldest are girls and the youngest is a boy?
(b) What is the probability that there are 2 girls and one boy (in any order)?

23. Suppose that, when throwing darts, you make a bull's eye 40% of the time. If you throw 2 darts, what is the probability that none of them is a bull's eye?

24. You and a friend decide to play "gambling dice." It costs 1 dollar to play. If you roll a 5 or 6, you get 2 dollars (a 1 dollar profit). If not, you lose your dollar. Compute the expected value for this game.

25. You have taken a survey of 235 college students. The results are:

- 67 students smoke cigarettes. Call this group A.
- 18 students use chewing tobacco. Call this group B.
- 103 students reported having one or both parents smoked. Call this group C.
- 46 students are in both A and C.
- 6 students are in both A and B.

- 9 students are in both B and C.
- 4 students are in all three groups, A, B, and C.

(a) Draw a Venn Diagram, including circles for A, B, and C, and fill in the number of students in each region of the diagram.

(b) What is the probability that a student smokes, given their parents smoked.

(c) What is the probability that a student's parents smoked, given that the student smokes.

(d) Are students who smoke more or less likely to use chewing tobacco than students who don't smoke? Explain.

(e) Are students whose parents smoked more or less likely to use chewing tobacco? Explain.

Probability: Activities and Class Exercises

1. **Medical Testing.**

   Mr. Speaker, once again, the United States has an opportunity, and the responsibility, to lead the world in confronting one of the most compelling humanitarian and moral challenges facing us today. I speak of the HIV/AIDS pandemic, a crisis unparalleled in modern times and one that threatens the entire world, embracing both developed and developing countries alike.
   —Representative Henry Hyde, U.S. House of Representatives, December 11, 2001

   One of the tests that is available to determine if someone has the human immunodeficiency virus (HIV) is called the Single Use Diagnostic System (SUDS). This test will give a positive result in approximately 97.9% of the cases where someone has the virus and 5.5% of those who do not have the virus.[136] The latter is called a false positive.

   Suppose you are living in a city of 100,000 people. It is estimated that approximately 0.3% of the United States population is infected with HIV.[137] We will assume that this percentage is true for the hypothetical city you are living in as well.

   Suppose you test positive for HIV using the SUDS test. Since there is a 5.5% chance of false positives among those without the virus, there is some chance that you do not have the virus, even though you have tested positive. **We want to find the probability that you actually have the virus assuming you have tested positive.** To do this, you need to know how many people in your city tested positive, and how many of this group actually have the virus.

   (a) Let's first consider those people who have the virus.
      i. What is the estimate for how many people in your city have HIV?
      ii. Suppose all those who have the virus in your city were tested. Approximately how many would test positive and how many would test negative?

   (b) Suppose everyone who did not have the virus were tested.
      i. How many people in your city do not have HIV?
      ii. Of these, how many would you expect to test positive, and how many would test negative?

   (c) Using parts (a) and (b), if everyone in your city were tested for HIV, both those with the virus and those without, how many would test positive? How many would test negative?

   (d) You have tested positive for HIV. What is the probability that you actually have the virus? Explain how you arrived at your answer.

   (e) Does your answer in part (d) seem surprising? Explain why the probability is so much lower than the 97.9% accuracy for those having the virus and the 94.5% chance that those without the virus test negative.

   (f) One erroneous result is a false positive. Another is a false negative. Given the accuracy of the test and the other data given, what is the probability that a person who tests negative for the virus actually has HIV?

[136] *Centers for Disease Control and Prevention,* "Rapid HIV Testing: 2003 Update," 28 October 2003, <http://www.cdc.gov/hiv/rapid_testing/materials/ USCA_Branson.pdf>.
[137] *Centers for Disease Control and Prevention,* 28 October 2003, "HIV/AIDS Update," <http://www.cdc.gov/nchstp/od/news/At-a-Glance.pdf>.

2. Relative Risks

The following passage is from a report by the U.S. State Department, cited in an article from the U.S. Embassy in Tokyo, Japan. (http://japan.usembassy.gov/e/p/tp-20030501a9.html).

> Washington – A decline in international terrorist attacks – down 44 percent in 2002 from the previous year – can be attributed to an intensively waged war on terrorism across every region of the world, the U.S. State Department's annual report on terrorism says.
>
> International terrorists conducted 199 attacks in 2002, compared with 355 attacks in 2001, according to the annual "Patterns of Global Terrorism: 2002" report, released April 30. Over the same period, the number of people killed by terrorists declined to 725, down dramatically from the 3,295 people killed in 2001. The 2001 total included those killed in the September 11 attacks in New York, Washington and Pennsylvania. The deaths in 2002 included 30 U.S. citizens.
>
> By geographic region, there were five terrorist attacks in Africa, 99 in Asia, seven in Eurasia, 50 in Latin America, 29 in the Middle East, none in the United States, and nine in Western Europe, the report said.

(a) One question we could ask is "who is more likely to be a victim of terrorism, a U.S. citizen, or someone who is not a U.S. citizen?" Note that the world population in 2002 was approximately 6.23 billion. The U.S. population was approximately 288 million, or .288 billion.

  i. What is the probability that a person who was a U.S. citizen was killed by terrorists in 2002?
  ii. What is the probability that a person who was *not* a U.S. citizen was killed by terrorists in 2002?
  iii. Based on your previous two answers, who was more at risk of being killed by terrorism in 2002, U.S. citizens or people who were not U.S. citizens?

(b) Another way to answer the previous question is to consider whether the percentage of those killed by terrorists who were U.S. citizens was larger or smaller than the percentage of the overall population that is represented by U.S. citizens.

  i. What is the probability that a person killed by terrorism in 2002 was a U.S. citizen?
  ii. What is the probability that a given person is a U.S. citizen?
  iii. Which of these two numbers are larger and what does this tell you about the risk to U.S. citizens?
  iv. Explain why "the probability that a person killed by terrorism in 2002 was a U.S. citizen" is not the same as the "probability that a person who was a U.S. citizen was killed by terrorists in 2002?"

(c) One could criticize the way we structured the question in part (a) since it disregards the facts in the third paragraph from the Embassy report. Note that not a single (international) terrorist attack occurred in the U.S. in 2002. How does this change the conclusion or the reasoning you used in part (a)? What might this imply about which U.S. citizens are at risk from terrorist attacks?

(d) One way to measure how concerned one should be about particular risks is to compare risks from different sources. From (a) and (b), you have some information about the risk of terrorism. Let's consider some other sources of risk. Just on the basis of your own gut intuition, rank the following risks of death from most likely to least likely. Assume we are only considering the U.S.

  i. Lightning
  ii. Airline crash
  iii. Automobile crash
  iv. Anthrax poisoning

v. Terrorism in general
vi. Murder in general

(e) For comparison, here are some facts on these risks.

- The number of people killed every year in car accidents in the U.S. is more than 40,000.
- On average, about 90 people a year are killed by lightning.
- According to the FBI, 15,517 people were murdered in 2000.
- Over the last 19 years, on average about 120 people have died in airline crashes each year.
- In 2001, 5 people died from Anthrax poisoning.

These are not all from 2002, but all are from recent years. Calculate the probability of dying from each of the risks listed above, ranking them from most likely to least likely. Instead of using decimals or percents, give the probabilities as a ratio "1 out of N".

(f) How many times greater is the probability of the most likely risk than the probability of being killed by terrorists?

(g) It might be worth comparing any of these to the probability of winning the jackpot in one of the popular powerball lotteries. In one such lottery, there is a 1 in 54 million chance of winning the jackpot. (Data from http://www.unitedjustice.com/stories/stats.html and http://www.worldwatch.org/pubs/mag/2000/135/mos/). Which of the above risks are more likely than winning this lottery, if any?

(h) Worldwide, the United Nations Food and Agricultural Organization estimates that about 25,000 people die *every day* from starvation. Given this, about what is the probability that a randomly selected person died from starvation in 2002? How does this compare to some of the other risks you've considered?

3. The 2008 election. In any election, politicians are most concerned with the most basic question; who wins? However, exit polls which indicate how different demographic groups voted are also of great interest. In this activity, we consider exit polling data from this past fall, accessed from `http://www.cnn.com/ELECTION/2008/results/polls.main/` on December 30th of 2008.

Table 8: Results by Race.

Total	Obama	McCain	Other
White (74%)	43%	55%	2%
African-Amer. (13%)	95%	4%	1%
Latino (9%)	67%	31%	2%
Asian (2%)	62%	35%	3%

(a) Based on the table above, what is the probability a given voter in the last election was white? African-American?

(b) Based on the table, what is the probability that a person voted for Obama given that they were white? African-American? Asian?

(c) Since a Presidential election is really a set of 51 state elections (including Washington D.C.), we cannot say who "would have" one the election under different circumstances without looking at those circumstances state by state. However, we can look at the overall popular vote totals and how they would change if certain demographic groups had voted differently.

i. According to CNN, 66,882,230 Americans voted for Obama and 58,343,671 voted for McCain. It is estimated a total of 127,003,956 Americans voted. Based on these totals, estimate how many people are represented by each of the 12 categories in the table. (Hint: You can actually do this using only the total number of voters and the percents given in the table). You may want to simply create another table, giving the total voters in each category instead of the percentages.

ii. Some Obama critics claimed that many African-Americans only voted for Obama because of his racial background. One indication of whether this is true or not is to look at how African-Americans voted in previous elections. This is not a perfect measure, since many factors can influence voters and the four years between elections can effect these factors. However, in the 2004 election, democrat John Kerry won 88% of the African-American vote to 11% for George Bush. Suppose that, in 2008, the percentages in the table above remained the same except that the percent of African-Americans who voted for Obama had been 88% instead of 95% and that McCain won 11% of the African-American vote. How would this have changed the popular vote totals?

Appendix 1: Miscellaneous Mathematics Reference

Exponents and Radicals

Exponents are shorthand for repeated multiplication. We define x^n to mean

$$x^n = x \cdot x \cdot x \cdots x$$

where there are n x's multiplied together. In particular, $x^1 = x$.

Roots are defined as follows. We let $\sqrt[n]{x}$ denote the number a so that $a^n = x$. In other words $(\sqrt[n]{x})^n = a$. This definition is subject to the following conditions.

1. If n is even, x must be greater than or equal to 0, and in this case $\sqrt[n]{x} \geq 0$.

2. If n is odd, there is no condition on x, and $\sqrt[n]{x}$ is the same sign as x.

Properties of exponents and radicals

We have the following properties of exponents. The proofs of these properties, in other words, the reason they always hold true, are all based on the definition of exponents given above.

1. $x^n \cdot x^m = x^{n+m}$

2. $\frac{x^n}{x^m} = x^{n-m}$

3. $(x^n)^m = x^{mn}$

4. $(xy)^n = x^n y^n$

5. $\left(\frac{x}{y}\right)^n = \frac{x^n}{y^n}$

6. $x^0 = 1$ for all $x \neq 0$

7. $x^{-n} = \frac{1}{x^n}$

8. $x^{m/n} = \sqrt[n]{x^m}$

Because of property 8, properties of radicals can be derived from the properties of exponents, since radicals can always be expressed as exponents.

Logarithms

In Chapter 7, we defined $\log_b x$ by

$$\log_b x = y \qquad \text{if and only if} \qquad b^y = x$$

if $b = 10$, we write $\log x$ and if $b = e$, we write $\ln x$.

As noted in Chapter Seven, it is important to keep in mind how logarithms fit into the usual order of operations. In general, unless there are parentheses involved, logarithms are done after any multiplications, exponents, or divisions that come immediately after the log sign, but before any additions or subtractions and before any multiplications immediately preceding the log sign.

Consider the following examples.

1. To calculate $\log xy$, you would first multiply x and y, and then take the logarithm of the product. This is the same as $\log(xy)$. On the other hand, to calculate $(\log x)y$, you would first take the logarithm of x and then multiply the result by y.

(a) $\log 2 \cdot 3 = \log 6 \approx .77815$

(b) $(\log 2)3 \approx .30103 \cdot 3 = .90309$

2. $x \log y$ is the same as $x(\log y)$ or $(\log y)x$. For example, $3 \log 2 \approx 3 \cdot (.30103) = .90309$.

3. $\log x^n$ is equal to $\log(x^n)$ and not $(\log x)^n$.

(a) $\log 2^3 = \log 8 \approx .90309$, but $(\log 2)^3 \approx .30103^3 \approx .02728$.

(b) Property 9 below in the properties of logarithms applies to $\log x^n$ and not $(\log x)^n$. Note that in the previous item, we calculated $\log 2^3 \approx .90309$, and that this equals $3 \log 2 \approx 3 \cdot (.30103) = .90309$.

4. Additions and subtractions come after logarithms, multiplications and exponents, unless parantheses are involved.

(a) $\log 3 + \log 2 \approx .47712 + .30103 = .97815$

(b) $\log 3 + 2 \approx .47712 + 2 = 2.47712$

(c) $\log(3 + 2) = \log 5 \approx .69897$

Properties of the Logarithm Function

In Chapter Seven, we noted the following properties of the logarithm function.

1. $\log 10^x = x$.

2. $\log(1) = 0$.

3. $\log x$ is negative for $0 < x < 1$ and positive for $x > 1$.

4. $y = \log x$ has a domain of $x > 0$ and a range of all real numbers.

5. The graph of $y = \log x$ is increasing and concave down.

6. $\log_b x = \frac{\log b}{\log x}$ for any positive $b \neq 1$.

7. $\log(uv) = \log u + \log v$.

8. $\log(u/v) = \log u - \log v$.

9. $\log x^n = n \log x$.

10. $10^{\log x} = x$.

Let's prove properties 8 and 9, as promised in the text.

First, property 8. Let $a = \log u$ and $b = \log v$. So, $u = 10^a$ and $v = 10^b$. Then $u/v = 10^a/10^b = 10^{a-b}$, using property for exponents above. Taking logarithms of both sides, we get

$$\log(u/v) = \log 10^{a-b} = a - b$$

using property 1 for logarithms above. But, $a - b = \log u - \log v$. This implies $\log(u/v) = \log u - \log v$.

Now for property 9. Let $a = \log x$. So, $x = 10^a$. This implies $x^n = (10^a)^n = 10^{an}$ using property 3 for exponents. Since $x^n = 10^{an}$, we have from the definition of logarithms that $\log x^n = an = na$, and since $a = \log x$, we get

$$\log x^n = na = n \log x$$

The number e and natural logarithms

We used the common base 10 logarithm above, but the properties and proofs work for all bases. We noted in the text that if e, which is approximately 2.71828, is the base, we write $\ln x$ instead of $\log_e x$. Where does e come from?

A rigorous definition of e is given in most calculus books. There, you might find e defined by

$$e = \lim_{x\to\infty}\left(1+\frac{1}{x}\right)^x$$

This means, that if we calculate the expression $(1 + 1/x)$ for larger and larger x, the values we get should get closer and closer to one particular number, and we use the letter e to denote this number (so we don't always have to write out "approximately 2.71828"). The table below gives a few sample calculations.

x	$(1 + 1/x)$
1	2
10	2.59374246
100	2.704813829
1000	2.716923932
1,000,000	2.718280469

Table 1: Numerical approximations to e

The number e is important because of it's relation to exponential growth. Recall the compound interest formula $B(t) = P(1+i)^t$, where i is the periodic interest rate and t is the number of years. If we let r be the annual rate, and n be the number of compounding periods per year, then $i = r/n$, and the number of periods equals n times the number of years, which we'll denote m. We get,

$$B(t) = P(1 + r/n)^{nm}$$

If we let $P = 1$, $r = 1$ (denoting 100% interest) and $m = 1$, then the expression for $B(t)$ becomes exactly the expression used to define e. Letting n get larger and larger would give you better and better approximations for e.

Letting n get larger and larger corresponds to having more and more compounding periods per year. $n = 12$ is monthly compounding, $n = 365$ is daily compounding (you get paid interest every *day*). For any P, r, and m, we get that

$$B(t) = P(1 + r/n)^{nm} \approx Pe^{rm}$$

for large n. In other words, if you compound often, you can use the formula Pe^{rm} as an approximation.

Many types of quantities, other than money in an account, grow exponentially. For example, populations might grow at an approximately fixed percentage growth rate. However, additions to the population usually happen *continuously*, not once per year or once per month (except in species that do breed at fixed times of the year). So, the Pe^{rm} is more appropriate than the compounding formula, since it represents *continuous* compounding.

Why didn't we use e when modeling populations?

Actually, we could have used e. Any model of the form $y = ab^x$ can be expressed using e by finding r so that $e^r = b$. If we find this r, then $ab^x = a(e^r)^x = ae^{rx}$. Finding r is not a problem, since if $e^r = b$, then $r = \ln b$. We used the $y = ab^x$ only to avoid the complications of discussing e.

Appendix 2: Data Samples

The effort of the economist is to "see," to picture the interplay of economic elements. The more clearly cut these elements appear in his vision, the better; the more elements he can hold and grasp in his mind, the better. The economic world is a misty region. The first explorers used unaided vision. Mathematics is the lantern by which what before was dimly visible now looms up in firm, bold outlines. The old phantasmagoria disappear. We see better. We also see further.
— Irving Fisher

World Population

The data in these tables is from the United States Census Bureau at http://www.census.gov/ipc/www/worldpop.html.

Year	Population	Percent grth. rate	Avg. ann. pop. change		Year	Population	Percent grth. rate	Avg. ann. pop. change
1950	2,555,360,972	1.47	37,785,986		1977	4,231,356,327	1.69	72,172,286
1951	2,593,146,958	1.61	42,060,389		1978	4,303,528,613	1.73	75,085,409
1952	2,635,207,347	1.71	45,337,232		1979	4,378,614,022	1.71	75,655,181
1953	2,680,544,579	1.77	47,971,823		1980	4,454,269,203	1.69	75,864,564
1954	2,728,516,402	1.87	51,451,629		1981	4,530,133,767	1.75	80,105,008
1955	2,779,968,031	1.89	52,959,308		1982	4,610,238,775	1.73	80,253,764
1956	2,832,927,339	1.95	55,827,050		1983	4,690,492,539	1.68	79,312,007
1957	2,888,754,389	1.94	56,506,563		1984	4,769,804,546	1.68	80,596,505
1958	2,945,260,952	1.76	52,335,100		1985	4,850,401,051	1.68	82,324,417
1959	2,997,596,052	1.39	42,073,278		1986	4,932,725,468	1.71	85,142,812
1960	3,039,669,330	1.33	40,792,172		1987	5,017,868,280	1.69	85,667,332
1961	3,080,461,502	1.8	56,094,590		1988	5,103,535,612	1.66	85,671,996
1962	3,136,556,092	2.19	69,516,194		1989	5,189,207,608	1.66	86,677,681
1963	3,206,072,286	2.19	71,119,813		1990	5,275,885,289	1.58	83,940,351
1964	3,277,192,099	2.08	69,031,982		1991	5,359,825,640	1.55	83,939,711
1965	3,346,224,081	2.08	70,238,858		1992	5,443,765,351	1.48	81,404,054
1966	3,416,462,939	2.02	69,755,364		1993	5,525,169,405	1.44	80,191,434
1967	3,486,218,303	2.04	71,882,406		1994	5,605,360,839	1.43	80,626,257
1968	3,558,100,709	2.08	74,679,905		1995	5,685,987,096	1.38	79,173,661
1969	3,632,780,614	2.05	75,286,491		1996	5,765,160,757	1.37	79,745,131
1970	3,708,067,105	2.07	77,587,001		1997	5,844,905,888	1.34	78,784,175
1971	3,785,654,106	2.01	76,694,660		1998	5,923,690,063	1.31	78,308,546
1972	3,862,348,766	1.95	76,183,283		1999	6,001,998,609	1.27	77,008,373
1973	3,938,532,049	1.9	75,547,134		2000	6,079,006,982	1.23	75,318,861
1974	4,014,079,183	1.81	73,265,577		2001	6,154,325,843	1.2	74,315,460
1975	4,087,344,760	1.74	71,797,582		2002	6,228,641,303	1.18	73,845,390
1976	4,159,142,342	1.72	72,213,985		2003	6,302,486,693	1.16	73,395,376

Table 1: World Population: 1950-2003, with growth rates

Year	Population	Percent grth. rate	Avg. ann. pop. change		Year	Population	Percent grth. rate	Avg. ann. pop. change
2004	6,375,882,069	1.14	72,898,133		2027	7,956,320,645	0.74	58,800,509
2005	6,448,780,202	1.12	72,714,711		2028	8,015,121,154	0.72	57,842,466
2006	6,521,494,913	1.11	72,772,754		2029	8,072,963,620	0.7	56,862,516
2007	6,594,267,667	1.1	72,776,911		2030	8,129,826,136	0.69	55,988,383
2008	6,667,044,578	1.08	72,703,236		2031	8,185,814,519	0.67	55,209,498
2009	6,739,747,814	1.07	72,485,099		2032	8,241,024,017	0.66	54,325,204
2010	6,812,232,913	1.06	72,447,829		2033	8,295,349,221	0.64	53,388,484
2011	6,884,680,742	1.05	72,509,964		2034	8,348,737,705	0.63	52,432,568
2012	6,957,190,706	1.03	72,261,423		2035	8,401,170,273	0.61	51,582,599
2013	7,029,452,129	1.02	71,803,520		2036	8,452,752,872	0.6	50,825,199
2014	7,101,255,649	1	71,144,072		2037	8,503,578,071	0.59	49,945,328
2015	7,172,399,721	0.98	70,443,548		2038	8,553,523,399	0.57	49,018,566
2016	7,242,843,269	0.96	69,755,219		2039	8,602,541,965	0.56	48,111,036
2017	7,312,598,488	0.94	68,928,253		2040	8,650,653,001	0.55	47,314,021
2018	7,381,526,741	0.92	67,997,557		2041	8,697,967,022	0.53	46,592,583
2019	7,449,524,298	0.89	66,966,195		2042	8,744,559,605	0.52	45,712,304
2020	7,516,490,493	0.87	65,973,432		2043	8,790,271,909	0.51	44,766,065
2021	7,582,463,925	0.85	65,024,404		2044	8,835,037,974	0.5	43,861,193
2022	7,647,488,329	0.83	63,958,545		2045	8,878,899,167	0.48	43,037,087
2023	7,711,446,874	0.81	62,831,316		2046	8,921,936,254	0.47	42,237,824
2024	7,774,278,190	0.79	61,670,133		2047	8,964,174,078	0.46	41,228,872
2025	7,835,948,323	0.77	60,634,107		2048	9,005,402,950	0.44	40,089,886
2026	7,896,582,430	0.75	59,738,215		2049	9,045,492,836	0.43	39,002,569
					2050	9,084,495,405		

Table 2: World Population (projected): 2004-2050

The following tables gives data on malnutrition from UNICEF.[138] UNICEF uses three different measures for malnutrition, namely percentage of children underweight (UWT), percentage of children suffering from stunting, and percentage of children suffering from wasting. For each of these three variables, UNICEF gives two percentages. The first, denoted combined in the tables, is the percentage who are classified as moderate or severe cases. The second percentage gives only the sever category.

Country	UWT combined	UWT severe	Stunting combined	Stunting severe	Wasting combined	Wasting severe
Afghanistan	48		52		25	
Albania	14.3	4.3	31.7	17.3	11.1	3.6
Algeria	6	1.2	18	5.1	2.8	0.6
Armenia	2.5	0.1	13.6	2.7	2	0.3
Azerbaijan	16.8	4.3	19.6	7.2	7.9	1.9
Bahrain	8.7	1.8	9.7	2.7	5.3	0
Bangladesh	47.8	13.1	44.8	18.4	10	1
Benin	29.2	7.4	25	7.8	14.3	2.7
Bhutan	18.7	3.1	40	14.5	2.6	0.5
Bolivia	9.5	1.7	25.6	8.9	1.8	0.5
Bosnia & Herz.	4.1	0.6	9.7	2.9	6.3	1.9
Botswana	12.5	2.4	23.1	7.9	5	1.1
Brazil	5.7	0.6	10.5	2.5	2.3	0.4
Burkina Faso	34.3	11.8	36.8	16.6	13.2	2.5
Burundi	45.1	13.3	56.8	27.7	7.5	0.5
Cambodia	45.9	13.4	46	22	15.3	4
Cameroon	21	4.2	34.6	13.3	4.5	0.8
Cape Verde	13.5	1.8	16.2	4.8	5.6	1
Central Afr. Rep.	24.3	6	38.9	19.1	8.9	2.1
Chad	27.6	9.8	28.3	13.4	11.7	2.9
Chile	0.8		1.9		0.3	
China	9.6		16.7		2.6	
Colombia	6.7	0.8	13.5	2.8	0.8	0.1
Comoros	25.4	8.5	42.3	23.3	11.5	3.7
Congo	13.9	3	18.8	6.6	3.9	0.9
Congo, Dem. Rep.	34.4	10.2	45.2	24.6	9.6	3.5
Costa Rica	5.1	0.4	6.1		2.3	
Cote d'Ivoire	21.4	4	21.9	7.8	10.3	0.9
Croatia	0.6		0.8		0.8	
Cuba	4.1	0.4	4.6	1.1	2	0.4
Czech Rep.	1	0	1.9	0.4	2.1	0.2
Djibouti	18.2	5.9	25.7	13.2	12.9	2.8

Table 3: UNICEF Malnutrition Rates

[138]Data available online at www.childinfo.org/eddb/malnutrition/index.htm

Country	UWT combined	UWT severe	Stunting combined	Stunting severe	Wasting combined	Wasting severe
Dominican Rep.	4.6	1	6.1	1.6	1.5	0.1
Ecuador	14.8	1.9	27.1	7.8		
Egypt	11.7	2.8	24.9	10.2	6.1	1.8
El Salvador	11.8	0.8	23.3	5.7	1.1	0.1
Eritrea	43.7	17	38.4	18.3	16.4	3.1
Ethiopia	47.1	16	51.2	25.9	10.7	1.4
Fiji	7.9	0.8	2.7	1.1	8.2	0.5
Gambia	17	3.5	18.7	5.9	8.6	1.2
Georgia	3.1	0.2	11.7	3.7	2.3	0.5
Ghana	24.9	5.2	25.9	9.3	9.5	1.4
Guatemala	24.2	4.7	46.4	21.2	2.5	0.9
Guinea	23.2	5.1	26.1	10.1	9.1	2.1
Guinea-Bissau	23.1	5.2	28	10.2	10.1	2.3
Guyana	11.8		10.1		11.5	
Haiti	27.5	8.1	31.9	14.9	7.8	1.5
Honduras	24.5	4	38.5	13.6	1.5	0.1
Hungary	2.2	0.2	2.9	0.4	1.6	0.1
India	47	18	45.5	23	15.5	2.8
Indonesia	26.4	8.1				
Iran	10.9	1.5	15.4	3.8	4.9	0.9
Iraq	15.9		22.1			
Jamaica	3.9		3.4		3.6	
Jordan	5.1	0.5	7.8	1.6	1.9	0.2
Kazakstan	4.2	0.4	9.7	2.5	1.8	0.2
Kenya	22.7	6.5	37.2	17.6	6.3	1.4
Korea, Dem.	60		59.5		18.7	
Kuwait	9.8	2.9	23.8	11.8	10.6	2.7
Kyrgyzstan	11	1.7	24.8	6	3.4	0.7
Lao PDR	40	12.9	40.7	20.3	15.4	3
Lebanon	3	0.3	12.2	2.9	2.9	
Lesotho	16	4	44	20	5	3
Libya	4.7	0.6	15.1	4.5	2.8	0.4
Madagascar	33.1	11.1	48.6	26	13.7	4.6
Malawi	25.4	5.9	49	24.4	5.5	1.2
Malaysia	18.3	1.2				
Maldives	43.2	10.1	26.9	6.8	16.8	3.4
Mali	43.3					
Mauritania	23	9.2	44	32.8	7.2	3.2
Mauritius	16.4	2.2	9.6	2.5	15	3.5
Mexico	7.5	1.2	17.7	5.6	2	0.6
Moldova	3.2		9.6		2.5	
Mongolia	12.7	2.8	24.6	8.5	5.5	1.2

Table 4: UNICEF Malnutrition Rates

Country	UWT combined	UWT severe	Stunting combined	Stunting severe	Wasting combined	Wasting severe
Morocco	9	1.8	22.6	8	2.3	0.4
Mozambique	26.1	9.1	35.9	15.7	7.9	2.1
Myanmar	36	8.7	37.2	14.6	9.7	1.5
Namibia	26.2	5.7	28.4	8.3	8.6	1.5
Nepal	47.1	12	54.1	22.1	6.7	0.5
Nicaragua	12.2	1.9	24.9	9.2	2.2	0.5
Niger	39.6	14.3	39.8	19.5	14.1	3.2
Nigeria	27.3	10.7	45.5	25.6	12.4	4.9
Occupied Pal. Territ.	4.4		7.2		2.7	
Oman	23.6	3.9	22.9	8	13	1.6
Pakistan	38.2	12.8				
Panama	6.8		14.4		1.1	
Paraguay	5		10.9		1	
Peru	7.8	1.1	25.8	8	1.1	0.3
Philippines	28.2		29.9		5.6	
Qatar	5.5		8.1		1.5	
Romania	5.7	0.6	7.8	1.8	2.5	0.3
Russian Fed.	3	0.5	12.7	6.7	3.9	1.6
Rwanda	29	7.1	42.7	18.6	6.7	1.3
Sao Tome & Principe	16	5	26	13	4.8	
Saudi Arabia	14.3	2.8	19.9	6.8	10.7	2.2
Senegal	18.4	4.1	19	6	8.3	
Sierra Leone	27.2	8.7	33.9	15.8	9.8	1.9
Somalia	25.8	6.9	23.3	12.1	17.2	3.5
Sri Lanka	33	4.8	17	3.5	15	
Sudan	16.7	7.2				
Syria	12.9	3.7	20.8	10.1	8.7	2.5
Tanzania	29.4	6.5	43.8	17.1	5.4	0.6
TFYR Macedonia	6	0.7	6.9	1.7	3.6	0.5
Thailand	18.6		16		5.9	
Togo	25.1	6.7	21.7	7	12.3	2.1
Tunisia	4	0.6	12.3	3.4	2.2	0.5
Turkey	8.3	1.4	16	6.1	1.9	0.4
Uganda	25.5	6.7	38.3	15	5.3	0.9
Ukraine	3	0.5	15.4	6.1	6.4	1.3
United Arab Emirates	14.4	3.2	16.7	6.8	15.2	3.8
United States	1.4	0.1	2.1	0.4	0.6	0.1
Uruguay	4.5	0.8	7.9		1	
Uzbekistan	18.8	5	31.3	14	11.6	2.8
Venezuela	4.7	0.7	13.6	4.5	3.1	0.6
Vietnam	33.1	5.8	36.4	11.9	5.6	
Yemen	46.1	14.5	51.7	26.7	12.9	2.6
Yugoslavia	1.9	0.4	5.1	1.9	3.7	0.7
Zambia	25		59		4	
Zimbabwe	13	1.5	26.5	9.4	6.4	1.6

Table 5: UNICEF Malnutrition Rates

Democracy Scores and Literacy Rates

The Democracy Scores are from the Polity IV Democracy Project. Literacy rates are the percentage of the population in the country that are considered functionally literate.

Country	Dem.	Lit.		Country	Dem.	Lit.
BELARUS	0	98.7		TUNISIA	1	27.4
UZBEKISTAN	0	96.7		TOGO	1	22.1
KAZAKHSTAN	0	95.8		HAITI	1	22.0
CUBA	0	89.3		ALGERIA	1	21.5
VIET NAM	0	83.0		BURUNDI	1	20.2
MYANMAR	0	69.8		NEPAL	1	16.4
QATAR	0	58.2		YEMEN	1	14.2
ZIMBABWE	0	57.6		CHAD	1	9.3
KUWAIT	0	57.6		TAJIKISTAN	2	90.7
CHINA	0	52.9		SINGAPORE	2	72.9
U.A.R.	0	52.2		JORDAN	2	55.1
BAHRAIN	0	50.9		BURKINA FASO	2	7.0
SWAZILAND	0	48.8		CAMBODIA	3	48.9
EQUATORIAL GUINEA	0	46.0		ZAMBIA	3	47.7
SYRIAN ARAB REP.	0	41.1		UNITED REP. OF TANZANIA	3	35.6
LAO P.D.R.	0	39.4		DJIBOUTI	3	30.2
UGANDA	0	36.4		LIBERIA	3	18.4
LIBYA	0	35.4		ETHIOPIA	3	12.9
SAUDI ARABIA	0	33.3		MALAYSIA	4	58.1
CONGO	0	32.9		COMOROS	4	49.5
EGYPT	0	31.6		IRAN ISLAMIC REP. OF	4	34.3
ERITREA	0	29.0		NIGERIA	4	20.1
RWANDA	0	27.8		NIGER	4	5.7
IRAQ	0	27.1		COTE D'IVOIRE	5	21.0
MAURITANIA	0	26.8		CENTRAL AFRICAN REP.	5	14.0
SUDAN	0	24.8		GUINEA-BISSAU	5	12.0
PAKISTAN	0	20.9		ARMENIA	6	91.6
MOROCCO	0	19.8		GUYANA	6	90.7
OMAN	0	18.5		VENEZUELA	6	76.3
GAMBIA	0	9.7		ECUADOR	6	74.3
CAMEROON	1	29.8		FIJI	6	73.2

Table 6:

COUNTRY	Dem.	Lit.		COUNTRY	Dem.	Lit.
NAMIBIA	6	57.0		GUATEMALA	8	45.1
MALAWI	6	37.9		KENYA	8	40.6
BANGLADESH	6	24.6		SENEGAL	8	14.7
MOZAMBIQUE	6	16.6		POLAND	9	98.2
BENIN	6	10.9		BULGARIA	9	92.4
MALI	6	9.1		CHILE	9	87.6
ESTONIA	7	99.8		THAILAND	9	80.2
UKRAINE	7	98.7		PANAMA	9	79.3
RUSSIAN FED.	7	98.2		PERU	9	71.5
CROATIA	7	91.3		SOUTH AFRICA	9	69.7
SRI LANKA	7	80.5		JAMAICA	9	68.1
PARAGUAY	7	79.8		BOLIVIA	9	57.5
COLOMBIA	7	77.8		BOTSWANA	9	46.1
EL SALVADOR	7	57.9		INDIA	9	33.1
ALBANIA	7	51.5		SLOVENIA	10	99.3
HONDURAS	7	50.6		LITHUANIA	10	98.3
MADAGASCAR	7	38.5		HUNGARY	10	98.1
GHANA	7	29.5		MONGOLIA	10	95.2
LATVIA	8	99.8		ITALY	10	94.5
ROMANIA	8	93.4		URUGUAY	10	93.3
ARGENTINA	8	93.0		NETH. ANTILLES	10	93.0
MOLDOVA	8	91.0		SPAIN	10	91.5
KOREA, REPUBLIC OF	8	86.8		TRINIDAD and TOBAGO	10	91.0
PHILIPPINES	8	81.8		COSTA RICA	10	88.2
MEXICO	8	73.5		GREECE	10	86.5
BRAZIL	8	68.4		CYPRUS	10	83.4
DOMINICAN REP.	8	67.2		ISRAEL	10	79.8
LESOTHO	8	64.1		PORTUGAL	10	73.7
TURKEY	8	56.5		MAURITIUS	10	67.0
INDONESIA	8	56.1		PAPUA NEW GUINEA	10	39.8
NICARAGUA	8	54.5		Czech Rep	10	22.8

Table 7:

Social Capital Indices

State	SCI	State	SCI	State	SCI
Alabama	-1.07	Maine	0.53	Ohio	-0.18
Arizona	0.06	Maryland	-0.26	Oklahoma	-0.16
Arkansas	-0.50	Massachusetts	0.22	Oregon	0.57
California	-0.18	Michigan	0.00	Pennsylvania	-0.19
Colorado	0.41	Minnesota	1.32	Rhode Island	-0.06
Connecticut	0.27	Mississippi	-1.17	South Carolina	-0.88
Delaware	-0.01	Missouri	0.10	South Dakota	1.69
Florida	-0.47	Montana	1.29	Tennessee	-0.96
Georgia	-1.15	Nebraska	1.15	Texas	-0.55
Idaho	0.07	Nevada	-1.43	Utah	0.50
Illinois	-0.22	New Hamp.	0.77	Vermont	1.42
Indiana	-0.08	New Jersey	-0.40	Virginia	-0.32
Iowa	0.98	New Mexico	-0.35	Washington	0.65
Kansas	0.38	New York	-0.36	West Virginia	-0.83
Kentucky	-0.79	North Carolina	-0.82	Wisconsin	0.59
Louisiana	-0.99	North Dakota	1.71	Wyoming	0.67

Table 1: Values for Putnam's Social Capital Index

Sources for Data Online

- The Population Reference Bureau (http://www.prb.org/). Contains reports, articles, and data related to population and demographics. Most of the data relates to recent years.
- United Nations Population Division (http://esa.un.org/unpp/). Time series data available on population and several other variables by country.
- Poverty and inequality data from the Inter-American Development Bank (http://www.iadb.org/sds/pov/)
- UNICEF site, "Monitoring the Situation of Children and Women" (http://www.childinfo.org/). Includes malnutrition data.
- The Food and Agriculture Organization of the United Nations (www.fao.org). Site includes statistical database with a large number of agricultural and nutritional variables.
- U.S. Census Bureau's poverty site (http://www.census.gov/hhes/www/poverty.html).

Appendix 3: How Large is Large?

A billion here, a billion there, and pretty soon you're talking real money
—attributed to former Illinois Senator Everett McKinley Dirksen (1896-1969)

Big numbers pervade our society. There are billions of people in the world, millions in most countries. Government budgets can be in the trillions. In some sense, all of these numbers are big numbers, and so it can seem like there is not much difference between them. Hopefully the following will help you get some intuitive sense of how big a really big number really is.

1. Millions

(a) TIME: A million seconds is about 278 hours or about 11 and a half days. A million minutes is about 694 days or almost 2 years. A million hours is about 114 years.

(b) A million dollars is enough to buy more than 30 brand new pick-up trucks. If you make $50,000 per year, it will take you 20 years to earn a total of a million dollars. A million dollars is enough to run DWU for about a month.

2. Billions

(a) TIME: A billions seconds is 278,000 hours, or about 11,500 days or about 32 years. A billion minutes is about 1900 years.

(b) MONEY:

i. If you won the million dollar jackpot on Who Wants to be a Millionaire each and every day for 2 years and 9 months, you would have won a total of a billion dollars.

ii. If DWU had a billion dollars invested at 5% interest, it could use the interest to run 3 DWU's forever, paying all the costs for everything and never ever charge a dime of tuition.

iii. On average, the U.S. Defense Department spends a billion dollars every 36 hours. The U.S. Government spends a billion about every 4 hours.

3. Trillions

(a) TIME: A trillion seconds is about 32,000 years. A trillion minutes is about 1,900,000 years. A trillion hours is about 120 million years, and that long ago, dinosaurs would have been wandering the earth.

(b) MONEY:

i. The annual U.S. budget is now over 2 trillion dollars.

ii. If you wanted to spend a trillion dollars during your adult life, say from age 20 to age 80, you would have to spend about 45 million dollars a day, every day for 60 years. If you did this by buying residential property, you could buy up all the residential property in Mitchell in a few days. Buying all the residential property in Sioux Falls should take about 2 months or less. In a year, you could buy all the residential property in South Dakota and be well into one of the neighboring states.

iii. On the other hand, if you wanted to spend a trillion dollars during your adult life, you could feed a lot of people. In many areas of the world, an adequate daily supply of food can be purchased for a few cents. Even at 50 cents per person per day, you could feed 90 million people, every day for 60 years. This is more than 10% of the total population of Africa.

(c) ECONOMICS:

i. Annual U.S. Gross National Product (GNP) is about 10 trillion. This is about 20% of the total world annual GNP.

(d) SCIENCE:

i. The total number of fish in the whole world is estimated to be about 1 trillion.

ii. The number of cells in a person is about 50 trillion.

iii. One light year is about 6 trillion miles. The nearest star (other than our sun) is about 26 trillion miles or a little over 4 light years away.

(e) FOOD, AGRICULTURE, and HUNGER:

i. Current annual worldwide production of grains is about 4 trillion pounds. This is enough for about 650 pounds per person per year or a little under two pounds per day. This is more than enough to live on.

4. Quadrillions (one quadrillion is 10^{15} or 1,000,000,000,000,000)

(a) TIME: A quadrillion seconds is about 32 million years. A trillion minutes is about 1.9 billion years. A trillion hours is about 120 billion years. The universe is estimated to be only 15 billion years old.

(b) MONEY:

i. The estimated market price of all the assets in the world is 1 quadrillion.

ii. If you wanted to spend a quadrillion dollars during your adult life, say from age 20 to age 80, you would have to spend about 45 billion dollars a day, every day for 60 years. You could buy all the residential property in South Dakota and be well into one of the neighboring states in about 7 or 8 hours. Alternatively, given the previous item on world assets, you could buy all the assets in the world in the first year. Not much left to do after that.

5. Really large numbers

(a) The number of atoms in the sun is about 10^{57}.

(b) The estimated number of atoms in the universe is about 10^{78}.

(c) The volume of the universe in cubic inches is about 10^{84}.

For a bigger list of big numbers, visit http://pages.prodigy.net/jhonig/bignum/indx.html. A few of our examples here are from this site.