2

Some Mathematical Tools

Introduction

This book is about biological modeling—the construction of mathematical abstractions intended to characterize biological phenomena and the derivation of predictions from these abstractions under real or hypothesized conditions. A model must capture the essence of an event or process but at the same time not be so complicated as to be intractable or to otherwise dilute its most important features. In this regard, differential equations have been widely invoked across the broad spectrum of biological modeling. Future values of the variables that describe a process depend on their rates of growth or decay. These in turn depend on present, or past, values of these same variables through simple linear or power relationships. These are the ingredients of a differential equation. We discuss linear and power laws between variables and their derivatives in Section 2.1 and differential equations in Section 2.4.

Sometimes a differential equation model is inappropriate because the phenomenon being studied is quantified in discrete units such as population size. If such sizes are very large, differential equations may still give correct results. Otherwise, difference equations may be more appropriate. We take up the basic principles of difference equations in Section 2.5.

Once formulated, a model contains parameters that must be specialized to the particular instance of the process being modeled. This requires gathering and treating experimental data. It requires determining values of the parameters of a model so as to agree with, or fit, the data. The universal technique for this is the method of least squares, which is the subject of Sections 2.2 and 2.3. Even though experimental data is subject to small random variations, or *noise*, and imprecision, least squares is designed to deal with this problem.

Describing noisy data and other manifestations of variation is the province of statistics. Distributions of values can be graphically portrayed as histograms or distilled to a single number, the average or mean. The most widely occurring distribution in the natural world is the normal, or Gaussian, distribution. These topics are taken up in Section 2.7.

R.W. Shonkwiler and J. Herod, *Mathematical Biology: An Introduction with Maple and Matlab*, Undergraduate Texts in Mathematics, DOI: 10.1007/978-0-387-70984-0_2,

Finally, to a greater extent in biological phenomena than in other fields of science and engineering, random processes play a significant role in shaping the course of events. This is true at all scales from diffusion at the atomic level to random combinations of genes to the behavior of whole organisms. Being in the wrong place at the wrong time can mean being a victim (or finding a meal). In Section 2.8 we discuss the basics of probabilities.

Fortunately, while an understanding of these mathematical tools is required for this book, deep knowledge of mathematical techniques is not. This is a consequence of the fruition of mathematical software. We will use the power of this software to execute calculations, invoke special functions, simplify algebra, solve differential equations, and generally perform the technical work. Above all, the software can make pictures of what is happening within the phenomenon in detail. Thereby, the curious are free to let their imaginations roam and focus on perfecting and exercising the models themselves.

As noted in the preface, you will be executing a lot of mathematical software code. As an aid to entering code, all the code in this book is posted on our webpages. Springer maintains the webpage

www.springer.com/978-0-387-70983-3,

Professor Herod's webpage is

www.math.gatech.edu/~herod,

and Professor Shonkwiler's webpage is

www.math.gatech.edu/~shenk.

In addition, as an aid to creating your own code, we provide a "code index" at the back of the book referencing the place in the text for syntax performing various mathematical and computer housekeeping tasks.

2.1 Linear Dependence

The simplest, nonconstant, relationship between two variables is a linear one. The simplest linear relationship is one of proportionality: if one of the variables doubles or triples or halves in value, the other does likewise. Proportionality between variables x and y is expressed as $y = kx$ for some constant k. Proportionality can apply to derivatives of variables as well as to variables themselves, since they are just rates of change. Historically, one of the major impacts of calculus is the improved ability to model by the use of derivatives in just this way.

Relationships among variables can be graphically visualized.

In studying almost any phenomenon, among the first observations to be made about it are its changing attributes. A tropical storm gains in wind speed as it develops;

the intensity of sound decreases with distance from its source; living things increase in weight in their early period of life. The measurable quantities associated with a given phenomenon are referred to as *constants*, *variables*, or *parameters*. Constants are unchanging quantities such as the mathematical constant $\pi = 3.14159\ldots$ or the physical constant named after Boltzmann: $k = 1.38 \times 10^{-16}$ ergs per degree. Variables are quantitative attributes of a phenomenon that can change in value, such as the wind speed of a tropical storm or the intensity of sound or the weight of an organism.

Parameters are quantities that are constant for a particular instance of a phenomenon, but can be different in another instance. For example, the strength of hair fibers is greater for thicker fibers and the same holds for spider web filaments, but the latter has a much higher strength per unit cross-section.[1] Strength per unit cross-section is a property of material that tends to be constant for a given type of material but varies over different materials.

Often two variables of a phenomenon are *linearly related*, that is, a graphical representation of their relationship is a straight line. Temperature as measured on the Fahrenheit scale, F, and on the Celsius scale, C, are related in this way; see Figure 2.1.1. Knowing that the temperatures $C = 0$ and $C = 100$ correspond to $F = 32$ and $F = 212$, respectively, allows one to derive their linear relationship, namely,

$$F = \frac{9}{5}C + 32. \tag{2.1.1}$$

In this, both C and F have *power* or *degree* one, that is, their exponent is 1. (Being understood, the 1 is not explicitly written.) When two variables are algebraically related and all terms in the equation are of degree one (or constant), then the graph of the equation will be a straight line. The multiplier, or *coefficient*, $\frac{9}{5}$ of C in (2.1.1) is the *slope* of the straight line, or the *constant of proportionality*,between the variables. The constant term 32 in the equation is the *intercept* of the straight line, or *translational term* of the equation. These parameters are shown graphically in Figure 2.1.1.

We can isolate the constant of proportionality by appropriate translation. Absolute zero on the Celsius scale is $-273.15C$, which is usually expressed in degrees *Kelvin* K. Translation from degrees K to degrees C involves subtracting the fixed amount 273.15:

$$C = K - 273.15. \tag{2.1.2}$$

From (2.1.1), we calculate absolute zero on the Fahrenheit scale as

$$F = \frac{9}{5}(-273.15) + 32 = -459.67,$$

or about -460 degrees *Rankine* R. That is,

$$F = R - 459.67. \tag{2.1.3}$$

Hence, substituting equations (2.1.2) and (2.1.3) into (2.1.1), we find that R is related to K by

[1] The strength of a material per unit cross-section is known as *Young's modulus*.

```
MAPLE
#number sign # introduces a comment
#statements must be ended by a semicolon or by a colon (suppresses printing) but can span multiple lines
> plot([C,9/5*C+32,C=0..100],-10..100,-30..220,tickmarks=[5,2]);

MATLAB
% percent sign introduces a comment in Matlab
% an end of line completes a command, or semicolon ends a command and suppresses printing results
> C=(0:1:100); % C=vector of values from 0 to 100 by ones
> F=(9/5)*C+32; % F=vector, this arithmetic to each C value
> plot(C,F); % plot the Fs vs. the Cs
> xlabel('Temperature degrees C'); %label horizontal axis
> ylabel('Temperature degrees F'); %label vertical axis
> axis([-10,110,-30,220]); % x scale from -10 to 110, y from -30 to 220
```

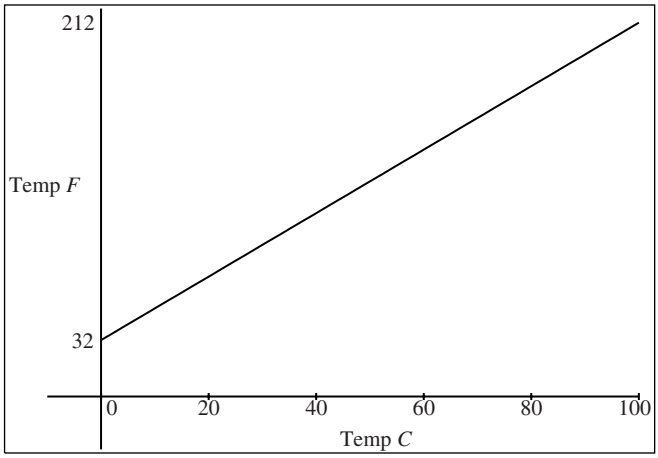


Fig. 2.1.1. Temperature conversion.

$$R = \frac{9}{5}K.$$

Thus R is proportional to K and both are zero at the same time, so there is no translational term.

One often observes that the relationship between two variables is one of proportionality but the constant is not yet known. Thus if variables x and y are linearly related (and both are zero at the same time), we write

$$y = kx$$

with the constant of proportionality k to be subsequently determined (see Section 2.2 on least squares).

Power laws can be converted to linear form.

The area of a circle does not vary linearly with radius but rather quadratically, $A = \pi r^2$; the power, or degree, of r is two. Heat radiates in proportion to the fourth power of absolute temperature, gravitational force varies in proportion to the inverse square power of distance, and diffusivity varies with the one-third power of density (see

Chapter 6). These are examples in which the relationship between variables is by a power law with the power different from one. There are many more.

In general, a power law is of the form

$$y = Ax^k \tag{2.1.4}$$

for some constants A and k. Due to the particular ease of graphing linear relationships, it would be advantageous if this equation could be put into linear form. This can be done by taking the logarithm of both sides of the equation. Two popular bases for logarithms are 10 and $e = 2.718281828459\ldots$; the former is often denoted by log and the latter by ln. (MATLAB uses log for logarithm to the base e.) Either will work:

$$\log y = k \log x + \log A; \tag{2.1.5}$$

the relationship between $\log y$ and $\log x$ is linear. Plotting pairs of (x, y) data values on special *log-log paper* will result in a straight line with slope k. Of course, on a log-log plot there is no point corresponding to $x = 0$ or $y = 0$. However, if $A = 1$ then $\log y$ is proportional to $\log x$ and the graph goes through the point $(1, 1)$. In general, A appears on the graph of (2.1.4) as the value of y when $x = 1$.

Another frequently encountered relationship between variables is an *exponential* one given by

$$y = Ca^x. \tag{2.1.6}$$

Note that the variable x is now in the exponent. Exponential functions grow (or decay) much faster than polynomial functions; that is, if $a > 1$, then as an easy consequence of L'Hopital's rule, for any power k,

$$\lim_{x\to\infty} \frac{x^k}{a^x} = 0, \tag{2.1.7}$$

or in MAPLE,

```
MAPLE
> assume(a>1); assume(k>0);
> limit(x^k/a^x,x=infinity);
```

Figure 2.1.2 demonstrates this with $k = 3$ and $a = 2$. We have drawn graphs of $y = x^3$, $y = 2^x$, and $y = 100 \cdot \frac{x^3}{2^x}$. The graphs of the first two cross twice, the last time about $x \approx 10$:

```
MAPLE
> sol:=solve(x^3=2^x,x);
> evalf({sol[1],sol[2]});

MATLAB
% make a file named fig212.m with the following two lines (without the % signs);
% MATLAB requires functions be defined in external files and finds them via the MATLAB PATH
%   function y=fig212(x);
%   y=x.^3 - 2.^x;
% resume this calculation
> fzero('fig212',10) %no semicolon to print ans.
```

$$1.3734, \qquad 9.939.$$

Taking logarithms of (2.1.6) to base e gives

MAPLE
```
> plot({[x,x^3,x=0..12],[x,2^x,x=0..12],[x,100*x^3/2^x,x=0..14]},x=0..14,y=0..4000);
```

MATLAB
```
> x=linspace(0,14); % 100 equally spaced values 0 to 14
> y=100*x.^3./2.^x; % .^means term by term power, ./ and .* mean term by term div. and mult.
> plot(x,y)
> hold on % keep axis, scale, etc., of the graph fixed
> x=linspace(0,12);
> plot(x,x.^3); % plot overlaid on the previous plot
> plot(x,2.^x); % ditto
```

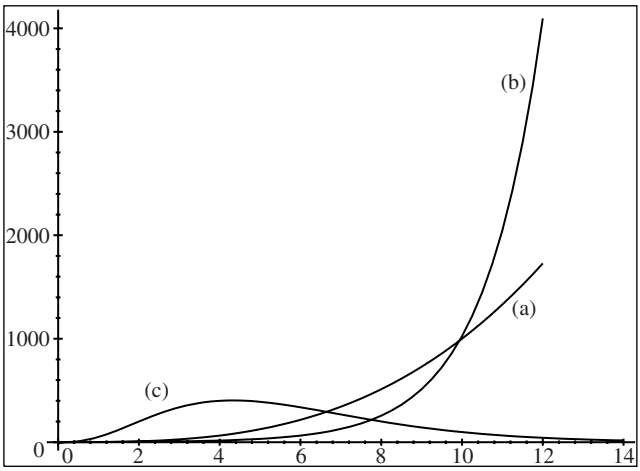


Fig. 2.1.2. Exponential vs. polynomial rate of growth graphs of (a) x^3, (b) 2^x, and (c) $100\frac{x^3}{2^x}$.

$$\ln y = x \ln a + \ln C. \tag{2.1.8}$$

If the constant a is e, then $\ln a = \ln e = 1$. Also note that any positive number can be written as some exponent of e, namely, $\ln a$. Thus $a = e^{\ln a} = e^r$ if we put $r = \ln a$. In the form of (2.1.8), it is $\ln y$ that is proportional to x. A *semilog plot* of exponentially related variables, as in (2.1.8), produces a straight line whose slope is $\ln a$.

By defining $r = \ln a$ and exponentiating both sides of (2.1.8), we get

$$y = Ce^{rx}, \quad \text{where } r = \ln a. \tag{2.1.9}$$

This is an alternative form of the relationship given in equation (2.1.6) and shows that an exponential relationship can be expressed in base e if desired.

Proportionality can pertain to derivatives, too.

A natural and simplifying assumption about the growth of a population is that *the number of offspring born at any given time is proportional to the number of adults at that time* (see Chapter 3). This expresses a linear relationship between the number of offspring and the number of adults. Let $y(t)$ (or just y in brief) denote the number

of adults at time t. In any given small interval of time Δt, the number of offspring in that time represents the change in the population Δy. The ratio $\frac{\Delta y}{\Delta t}$ is the average rate of growth of the population over the time period Δt. The derivative $\frac{dy}{dt}$ is the instantaneous rate of growth at time t, or just the rate of growth at time t, instantaneous being understood. Making the questionable, but simplifying, assumption that new offspring are immediately adults leads to a mathematical expression of the italicized statement above:

$$\frac{dy}{dt} = ky$$

for some constant of proportionality k. That is, the derivative or rate of growth is proportional to the number present.

This particular differential equation is easily solved by integration,

$$\frac{dy}{y} = kdt \quad \text{or} \quad \ln y = kt + \ln A,$$

with constant of integration $\ln A$. Exponentiating both sides gives

$$y = Ae^{kt}.$$

This situation is typical, and we will encounter similar ones throughout the book.

Exercises

1. Proportionality constants associated with changes in units are often used in making conversions after measurements have been made. Convert from the specified units to the indicated units.

(a) Convert the following: x inches to centimeters, y pounds per gallon to kilograms per liter, z miles per hour to kilometers per hour.

MAPLE

```
#Change of units is built-in
#type: ?convert.
> convert(x*inches,metric);
> convert(y*pounds/gallon,metric,US);
> convert(z*miles/hour,metric);
```

MATLAB

```
% some US to metric conversions
% Length: 1 inch = 2.54 cm (exactly), 39.3700 inch = 1 meter
% Mass: 1 lb = .45359237 kg (avoirdupois pound)
% Volume: 1 gallon = 3.785411784 liter (US gallon)
> x=0:10; y=2.54*x; plot(x,y) % plot cm vs. inch
% to plot kg/liter vs. pounds/gallon one finds the number of the former per 1 of the latter;
% use this 1 lb/gal = (1 lb/gal)*(1 gal/3.78 lit)*(.453 kg/lb)
% cancel units so that 1 lb/gal = .45359237/3.785411784 kg/lit.
```

(b) Sketch three graphs similar to Figure 2.1.1 that show the changes in units indicated above. Syntax similar to that which generated Figure 2.1.1 can be used here.

2. In this exercise, we compare graphs of exponential and power law relations with standard graphs, log graphs, and log-log graphs. For this exercise, please type

the commands manually (rather than executing pretyped commands downloaded from the Web) and view the results of each command one by one. This will help internalize the commands and aid in connecting each with its action.

(a) Sketch the graphs of πr^2 and $\frac{4}{3}\pi r^3$ on the same graph. Then sketch both of these as log-log plots.

(b) Sketch the graphs of $3x^5$ and $5x^3$ on the same graph. Then sketch both these as log plots.

```
  MAPLE
> plot({Pi*r^2,4/3*Pi*r^3},r=0..1);
> plots[loglogplot]({Pi*r^2,4/3*Pi*r^3},r=0.1..1);
> plot({3*x^5,5*x^3},x=0..1);
> plots[logplot]({3*x^5,5*x^3},x=0..1);

  MATLAB
> r=0:.1:1; % create vector of r values
> plot(r,pi*r.^2)
  % plot pi r squared vs. r, use .^(dot hat, not ^)
  % to get term by term r squared, no need for .* (dot star) since pi is a constant
> hold on % to overlay this graph
> plot(r,pi*(4/3)*r.^3);
> hold off % begin new plot
> loglog(r,pi*r.^2) % MATLAB automatically avoided r=0
> hold on
> loglog(r,(4/3)*pi*r.^3)
> hold off
> x=linspace(0,1); % divide 0 to 1 into 100 subdivisions
> plot(x,3*x.^5); hold on
> plot(x,5*x.^3)
```

3. This exercise examines limits of quotients of polynomials and exponentials. Sketch the graphs of $3x^2 + 5x + 7$ and 2^x on the same axis. Also, sketch the graph of their quotients. Evaluate the limit of this quotient.

```
  MAPLE
> plot({3*x^2+5*x+7,2^x},x=0..7);
> plot((3*x^2+5*x+7)/2^x,x=0..10,y=0..10);
> limit((3*x^2+5*x+7)/2^x,x=infinity);

  MATLAB
> x=linspace(0,7); % vector of 100 x values
> plot(x,3*x.^2+5*x+7); hold on
> plot(x,2.^x)
  % or make a matrix whose first row=polynomial and second row=exponential
> M=[3*x.^2+5*x+7; 2.^x]; % note the semicolon in M
> hold off; plot(x,M) % and plot both at once
> plot(x,M(1,:)./M(2,:))
  % quotient of first row/second row term by term
  % observe the limit is 0 graphically
```

4. This exercise solves differential equations such as we encounter in Section 2.1. Give the solution and plot the graph of the solution for each of these differential equations:

$$\frac{dy}{dt} = 3y(t), \qquad y(0) = 2,$$

$$\frac{dy}{dt} = 2y(t), \qquad y(0) = 3,$$

$$\frac{dy}{dt} = 2y(t), \qquad y(0) = -3,$$

$$\frac{dy}{dt} = -2y(t), \qquad y(0) = 3.$$

Here is syntax that will do the first problem and will undo the definition of y to prepare for the remaining problems.

MAPLE

```
> eq:=diff(y(t),t)=3*y(t);
> sol:=dsolve({eq,y(0)=2},y(t));
> y:=unapply(rhs(sol),t); plot(y(t),t=0..1);
> y:='y';
```

MATLAB

```
% for the 1st DE make an m-file, ex214a.m, say, containing
%   function yprime=ex214a(t,y); yprime=3*y;
> [t,y]=ode23('ex214a',[0 1],2);
> plot(t,y)
```

2.2 Linear Regression, the Method of Least Squares

In this section we introduce the method of least squares for fitting straight lines to experimental data. By transformation, the method can be made to work for data related by power laws and exponential laws as well as for linearly related data.

The method is illustrated with two examples.

The method of least squares calculates a linear fit to experimental data.

Imagine performing the following simple experiment: Record the temperature of a bath as shown on two different thermometers, one calibrated in Fahrenheit and the other in Celsius, as the bath is heated. We plot the temperature F against the temperature C. Surprisingly, if there are three or more data points observed to high precision, they will not fall on a single straight line because the mathematical line established by two of the points will dictate infinitely many digits of precision for the others—no measuring device is capable of infinite precision. This is one source of error, and there are others. Thus experimental data, even data for linearly related variables, are not expected to fall perfectly on a straight line.

How then can we conclude experimentally that two variables are linearly related, and if they are, how can the slope and intercept of the correspondence be determined? The answer to the latter question is by the method of least squares fit and is the subject of this section; the answer to the first involves theoretical considerations and the collective judgment of scientists familiar with the phenomenon.

Assume that the variables x and y are suspected to be linearly related and we have three experimental points for them, for example C and F in the example above. For the three data points (x_1, y_1), (x_2, y_2), and (x_3, y_3) shown in Figure 2.2.1, consider a possible straight line fit, $\ell(x)$. Let e_1, e_2, and e_3 be the errors

$$e_i = y_i - \ell(x_i), \quad i = 1, \ldots, 3,$$

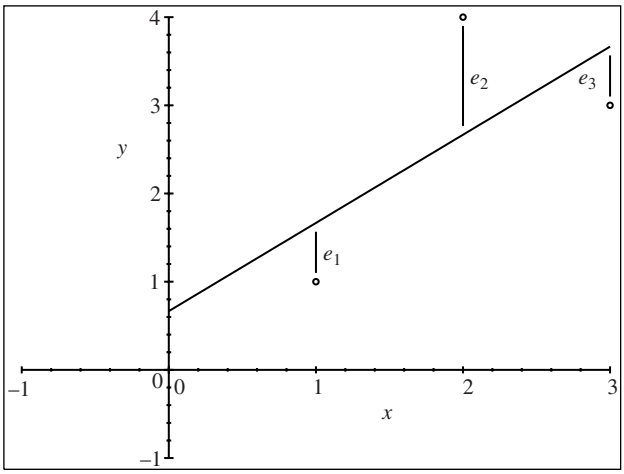


Fig. 2.2.1. The differences $e_i = y_i - \ell(x_i)$.

defined as the difference between the data value y_i and the linear value $\ell(x_i)$ for each point. Note that we assume that all x-data values are exact and that the errors are in the y-values only. This is reasonable because x is the independent variable; the x-values are the ones determined by the experimenter.

We want to choose a line ℓ that minimizes all of the errors at the same time; thus a first attempt might be to minimize the sum $e_1 + e_2 + e_3$. The difficulty with this idea is that these errors can cancel because they are signed values. Their sum could even be zero. But squaring each error eliminates this problem. And we choose the line ℓ so as to minimize

$$E = \sum_{i=1}^{3} e_i^2 = \sum_{i=1}^{3} [y_i - \ell(x_i)]^2,$$

that is, the least of the squared errors.

A line is determined by two parameters, slope m and intercept b, $\ell(x) = mx + b$. Therefore the mathematical problem becomes, find m and b to minimize

$$E(m, b) = \sum_{i=1}^{n} [y_i - (mx_i + b)]^2 \tag{2.2.1}$$

for n equal to the number of data points, three in this example. We emphasize that this error E is a function of m and b (not x and y; the x_i and y_i are specified numbers at the outset). Solving such a minimization problem is standard practice: Set the derivatives of E with respect to its variables m and b equal to zero and solve for

m and b,[2]

$$0 = \frac{\partial E}{\partial m} = -2\sum_{i=1}^{n}[y_i - (mx_i + b)]x_i,$$
$$0 = \frac{\partial E}{\partial b} = -2\sum_{i=1}^{n}[y_i - (mx_i + b)].$$

These equations simplify to

$$\begin{aligned} 0 &= \sum_{i=1}^{n} x_i y_i - m\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i, \\ 0 &= \sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i - nb, \end{aligned} \tag{2.2.2}$$

which may be easily solved.[3] The least squares solution is

$$\begin{aligned} m &= \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \\ b &= \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \end{aligned} \tag{2.2.3}$$

The expression for b simplifies to[4]

$$b = \bar{y} - m\bar{x}, \quad \text{where } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

We will illustrate the least squares method with two examples.

Example 2.2.1. Juvenile height vs. age is only approximately linear.

In Table 2.2.1, we show age and average height data for children.

With $n = 7$, age and height interpreted as x and y, respectively, in (2.2.1), and using the data of the table, parameters m and b can be evaluated from the equations in (2.2.3):

[2] Since E is a function of two independent variables m and b, it can vary with m while b is held constant or vice versa. To calculate its derivatives, we do just that: Pretend b is a constant and differentiate with respect to m as usual; this is called the *partial derivative* with respect to m and is written $\frac{\partial E}{\partial m}$ in deference to the variables held fixed. Similarly, hold m constant and differentiate with respect to b to get $\frac{\partial E}{\partial b}$. At a minimum point of E, both derivatives must be zero, since E will be momentarily stationary with respect to each variable.

[3] Verify this solution by substituting $m = \frac{nE-BF}{nA-BC}$ and $b = \frac{AF-cE}{nA-BC}$ into $mA + bB = E$ and $mC + nb = F$.

[4] Starting from $\bar{y} - m\bar{x}$ with m from (2.2.3), make a common denominator and cancel the terms $-(\sum x_i)^2\bar{y} + \bar{x}\sum x_i \sum y_i$, and the expression for b emerges.

Table 2.2.1. Average height vs. age for children. (Source: D. N. Holvey, ed., *The Merck Manual of Diagnosis and Therapy*, 15th ed., Merck, Sharp, and Dohme Research Laboratories, Rahway, NJ, 1987.)

Height (cm)	75	92	108	121	130	142	155
Age	1	3	5	7	9	11	13

MAPLE

```
> ht:=[75,92,108,121,130,142,155]; age:=[1,3,5,7,9,11,13];
> sumy:=sum(ht[n],n=1..7); sumx:=sum(age[n],n=1..7);
> sumx2:=sum(age[n]^2,n=1..7);
> sumxy:=sum(age[n]*ht[n],n=1..7);
> m:=evalf((7*sumxy-sumx*sumy)/(7*sumx2-sumx^2));
> b:=evalf((sumx2*sumy-sumx*sumxy)/(7*sumx2-sumx^2));
```

MATLAB

```
> ht=[75 92 108 121 130 142 155];
> age=[1 3 5 7 9 11 13];
> sumy=sum(ht);
> sumx=sum(age);
> age2=age.*age;
> sumx2=sum(age2);
> ageht=age.*ht;
> sumxy=sum(ageht);
> m=(7*sumxy-sumx*sumy)/(7*sumx2-sumx^2)
> b=(sumx2*sumy-sumx*sumxy)/(7*sumx2-sumx^2)
```

$$m = 6.46 \quad \text{and} \quad b = 72.3.$$

These data are plotted in Figure 2.2.2 along with the least squares fit for an assumed linear relationship $\text{ht} = m \cdot \text{age} + b$ between height and age.

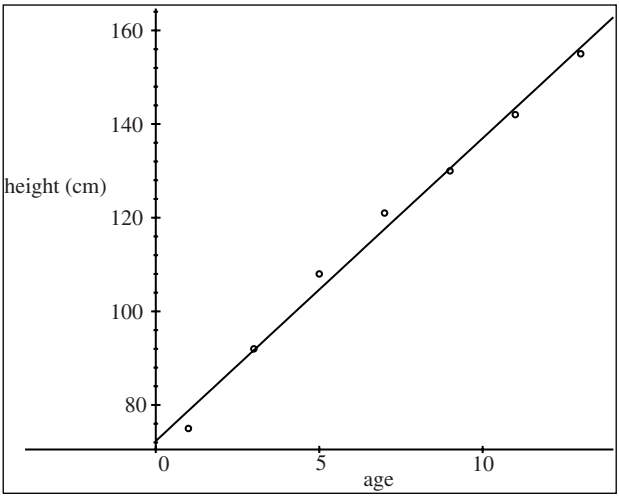


Fig. 2.2.2. Height vs. age among children.

Finding a least square fit is so important that it has its own routine in MAPLE called `fit[leastsquare]`. In MATLAB a least square fit is performed by a simple matrix statement. The mathematics of the matrix approach is the subject of the next section. Here then is the shortcut syntax for accomplishing what was done above.

```
  MAPLE
> m:='m'; b:='b'; # clears m and b (single quotes/apostrophy)
  # next create an array of (age,ht) pairs;
> pts:=[seq([age[i],ht[i]],i=1..7)]:
> with(plots): with(stats):
> Data:=plot(pts,style=POINT,symbol=CIRCLE):
> fit[leastsquare[[x,y],y=m*x+b]]([age,ht]);
  # result in y=m*x+b form, m*x is the first operand on the right-hand side
> m:=op(1,op(1,rhs(%))); # strip off x too
> b:=op(2,rhs(%%)); # use %% to get second statement back
> Fit:=plot(m*x+b,x=0..14):
> display({Data,Fit});

  MATLAB
  % Now the matrix solution
  % matrix of independent variable
  % experimental values as columns
> MT=[1 3 5 7 9 11 13; 1 1 1 1 1 1 1]; % two rows
> M=MT'; % transpose to columns
  % M = transpose of MT
  % dependent variable data next, as col. vec.
> Y=[75; 92; 108; 121; 130; 142; 155];
> s=M\Y % MATLAB syntax for leastsquare
> m=s(1); b=s(2); % plot data and fit for comparison, Figure 2.2.2
> plot(age,ht,'o') % point plot ht vs. age with circles
> hold on
> fit=m*age+b;
> plot(age,fit); xlabel('age'); ylabel('Height (cm)');
```

This demonstrates the mechanics of the least squares method. But it must be kept in mind that the method is merely statistical; it can demonstrate that data are consistent or not with a linear assumption, but it cannot prove linearity. In this example, a linear fit to the data is reasonably good, but no rationale for a linear relationship has been provided.

Example 2.2.2. The number of AIDS cases increases cubically.

As we saw in the first part of this section, when the data are obviously not linear, we can try to fit a power law of the form $y = Ax^k$. Consider the following data as reported in the HIV/AIDS Surveillance Report published by the U.S. Department of Health and Human Services concerning the reported cases of AIDS by half-year shown in Table 2.2.2. The third column is the sum of all the cases reported to that time, i.e., the Cumulative AIDS Cases (CAC).

This cumulative AIDS cases data is shown later in Figure 2.2.4. The circle symbols of the figure give the CAC data vs. year; the solid curve is the least squares fit, which we discuss next. In this figure, CAC is measured in thousands and t is decades from 1980, that is, $t = \frac{\text{year}-1980}{10}$.

We begin by first reading in the data:

```
  MAPLE
> restart:
> AIDS:=([97, 206, 406, 700, 1289, 1654, 2576, 3392, 4922, 6343, 8359, 9968, 12990, 14397, 16604,
          17124, 19585, 19707, 21392, 20846, 23690, 24610, 26228, 22768, 4903]);
```

Table 2.2.2. Total and reported cases of AIDS in the U.S.

Year	Reported cases of AIDS	Cumulative AIDS cases (thousands)
1981	97	0.097
1981.5	206	0.303
1982	406	0.709
1982.5	700	1.409
1983	1289	2.698
1983.5	1654	4.352
1984	2576	6.928
1984.5	3392	10.320
1985	4922	15.242
1985.5	6343	21.585
1986	8359	29.944
1986.5	9968	39.912
1987	12990	52.902
1987.5	14397	67.299
1988	16604	83.903
1988.5	17124	101.027
1989	19585	12.0612
1989.5	19707	140.319
1990	21392	161.711
1990.5	20846	181.557
1991	23690	206.247
1991.5	24610	230.857
1992	26228	257.085
1992.5	22768	279.853

```
> CAC:=[seq(sum(AIDS[j]/1000.0, j=1..i),i=1..24)];
> Time:=[seq(1981+(i-1)/2,i=1..24)]:
```

MATLAB

```
% year by year cases; note that ellipses continue the line
> AIDS=[97 206 406 700 1289 1654 2576 3392 4922 6343 8359 9968 12990 14397 16604 17124 19585 ...
        19707 21392 20846 23690 24610 26228 22768];
> CAC=cumsum(AIDS)/1000; % cumulative sum (scaled down 1000)
% housekeeping to get the sequence 0,0.5,1,1.5,...
> s=size(AIDS); % number of half-years
> count=[0:s(2)-1];
> time =1981+count/2;
```

To produce the fit we proceed as before using (2.2.1), but this time performing least squares on $y = \ln(\text{CAC})$ vs. $x = \ln t$:

$$\ln(\text{CAC}) = k * \ln t + \ln A. \tag{2.2.4}$$

Here we rescale time to be decades after 1980 and calculate the logarithm of the data:

MAPLE

```
> LnCAC:=map(ln,CAC);
> Lntime:=map(ln,[seq((i+1)/2/10,i=1..24)]);
```

MATLAB

```
% shifted and scaled time
```

```
> scaledTime=(time-1980)/10
  % log the data to do a log-log plot
> lnCAC=log(CAC)
> lnTime=log(scaledTime)
```

It remains to calculate the coefficients:

```
  MAPLE
> with(stats):
> fit[leastsquare[[x,y],y=k*x+LnA]]([Lntime,LnCAC]);
> k:=op(1,op(1,rhs(%))); LnA:=(op(2,rhs(%%))); A:=exp(LnA);

  MATLAB
  % form the coefficient matrix for lnCAC = k*lnTime + b fit
> MT=[lnTime; ones(1,24)] % second row is ones
> M=MT';
> params=M\(lnCAC') % do the leastsquares
> k=params(1)
> A=exp(params(2))
```

$$k = 3.29, \quad \text{and} \quad \ln A = 5.04, \quad A = 155.$$

We draw the graph of Ln(CAC) vs. Ln(time) to emphasize that their relationship is nearly a straight line. The log-log plot of best fit is shown in Figure 2.2.3 and is drawn as follows:

```
  MAPLE
> Lndata:=plot([seq([Lntime[i],LnCAC[i]],i=1..24)],style=POINT,symbol=CIRCLE):
> Lnfit:=plot(k*x+ln(A),x=-2.5..0.5):
> plots[display]({Lndata,Lnfit});

  MATLAB
  % now compare the fit to the data in log-log space
> plot(lnTime,lnCAC,'o')
> lnFit= params(1).*lnTime+params(2)
> plot(lnTime,lnFit)
```

The curve of best fit is, from (2.2.4),

$$\text{CAC} = 155t^{3.29}.$$

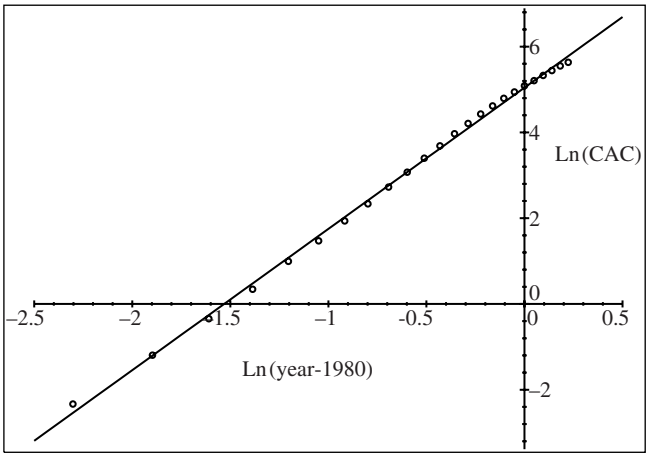


Fig. 2.2.3. Log-log plot of cumulative AIDS cases and its fit.

But we want an integer exponent; hence the exponent for the comparative graph to the data will be taken as

MAPLE
```
> n:=trunc(k);
```

$$n = 3,$$

$$\mathrm{CAC} = 155t^3 = 155\left(\frac{\text{year} - 1980}{10}\right)^3.$$

Figure 2.2.4 is drawn as an overlay of the data and this fit.

MAPLE
```
> pts:=[seq([Time[i], CAC[i]], i=1..24)];
> Fit:=plot(A*((t-1980)/10)^n,t=1980..1993):
> Data:=plot(pts,style=POINT,symbol=CIRCLE):
> plots[display](Fit,Data);
```

MATLAB
```
  % and compare in regular space
> hold off; plot(time,CAC)
> CACFit=exp(params(2)).*scaledTime.^params(1)
> plot(time,CACFit)
```

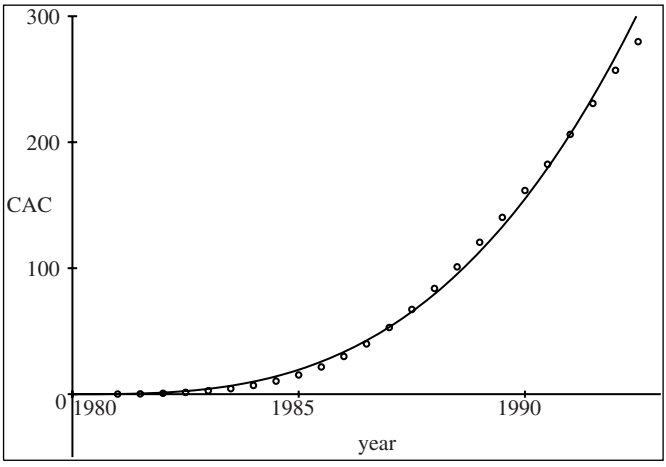


Fig. 2.2.4. Cumulative AIDS cases.

Again, we see that the fit is good. Turning from the mechanical problem of fitting the data to the scientific problem of explaining the fit, why should a cubic fit so well?

In the studies of populations and infectious diseases, it is common to ask at what rate an infected population is growing. Quite often, populations grow exponentially in their early stages, that is, according to (2.1.6). We will investigate this idea in Chapters 3 and 4.

In the first decade after the appearance of AIDS and the associated HIV, an analysis of the data for the total number of reported cases of AIDS led to the announcement

that the population was growing cubically as a function of time. This was a relief of sorts because the growth was not exponential as expected, since exponential growth is much faster than polynomial growth; see (2.1.7).

Colgate et al. [2] constructed a model for HIV infection that led to the result that the growth rate should be cubic in the early stages. A central idea in the model is the recognition that the disease spreads at different rates in different "risk groups," and that there is a statistically predictable rate at which the disease crosses risk groups.

In the exercises, we attempt an exponential fit to these data.

Exercises

1. Ideal weights for medium-build males are listed in Table 2.2.3 from [3].

Table 2.2.3. Ideal weights for medium-build males.

Height (in)	Weight (lb)
62	128
63	131
64	135
65	139
66	142
67	146
68	150
69	154
70	158
71	162
72	167
73	172

(a) Show that a linear fit for these data is

$$\text{wt} = 4.04 \cdot \text{ht} + 124.14.$$

(b) In many geometric solids, volume changes with the cube of the height. Give a cubic fit for these data.

(c) Using the techniques of Example 2.2.2, find n and A such that

$$\text{wt} = A \cdot (\text{ht} - 60)^n.$$

The following code can be used for Exercise 1(b). A modification of one line can be used for 1(a). For 1(c), modify the code for Example 2.2.2.

MAPLE

```
> ht:=[62,63,64,65,66,67,68,69,70,71,72,73,74];
> wt:=[128,131,135,139,142,146,150,154,158,162,167,172,177];
> with(stats): fit[leastsquare[[x,y], y=a*x^3+b*x^2+c*x+d]]([ht,wt]);
> y:=unapply(rhs(%),x);
```

```
> pts:=[seq([ht[i],wt[i]],i=1..13)];
> J:=plot(pts,style=POINT,symbol=CROSS):K:=plot(y(x),x=62..74):
> with(plots): display({J,K});
> errorLinear:=sum('(4.04*ht[i]-124.14- wt[i])^2','i'=1..13);
> errorcubic:=sum('(y(ht[i])-wt[i])^2','i'=1..13);
> evalf(%);
```

MATLAB

```
> ht=[62,63,64,65,66,67,68,69,70,71,72,73,74];
> wt=[128,131,135,139,142,146,150,154,158,162,167,172,177];
> MT=[ht.^3; ht.^2; ht; ones(1,13)];
> params=MT'\wt'; % MT prime, wt prime
> plot(ht,wt,'x'); hold on
> fit=params(1)*ht.^3+params(2)*ht.^2+params(3)*ht+params(4);
> plot(ht,fit)
> errorLinear=sum((4.04*ht-124.14-wt).^2)
> errorcubic=sum((fit-wt).^2)
```

2. Changes in the human life span are illustrated graphically on p. 110 of the October 1994 issue of *Scientific American*. These data appear in Table 2.2.4 in three rows: The first row indicates the age category. The next two rows indicate the percentage of people who survived to that age in the United States in the years 1900 and 1960. The last row is the percentage of people who survived to that age in ancient Rome. Get a least squares fit for these data sets. Syntax that provides such a fit is given for the 1960 data.

Table 2.2.4. Survival rates for recent U.S. and ancient Rome.

Age	0	10	20	30	40	50	60	80	100
1900	100	82	78	75	74	60	43	19	3
1960	100	98.5	98	96.5	95	92.5	79	34	4
Rome	90	73	50	40	30	22	15	5	0.5

MAPLE

```
> restart:
> age60:=[0,10,20,30,40,50,60,80,100]:
> percent60:=[100,98.5,98,96.5,95,92.5,79,34,4]:
> with(stats):
> fit[leastsquare[[x,y],y=a*x^4+b*x^3+c*x^2+d*x+e]]([age60,percent60]);
> yfit60:=unapply(rhs(%),x):
> pts60:=[seq([age60[i],percent60[i]],i=1..9)]:
> J6:=plot(pts60,style=POINT,symbol=CROSS):
> K6:=plot(yfit60(x),x=0..100):
> with(plots): display({J6,K6});
```

MATLAB

```
> age60=[0,10,20,30,40,50,60,80,100];
> percent60=[100,98.5,98,96.5,95,92.5,79,34,4];
> MT=[age60.^4; age60.^3;age60.^2; age60; ones(size(age60))];
> parms=MT'\percent60' % note the primes
> fit=parms(1)*age60.^4+parms(2)*age60.^3+parms(3)*age60.^2+parms(4)*age60+parms(5);
> plot(age60,percent60,age60,fit)
```

3. We have found a fit for the cumulative U.S. AIDS data as a cubic polynomial. We saw that, in a sense, a cubic polynomial is the appropriate choice. On first looking at the data as shown in Figure 2.2.4, one might guess that the growth is exponential. Find an exponential fit for those data. Such a fit would use (2.1.8).

Computer code to perform the calculations is only slightly different from that for the cubic fit:

MAPLE

```
> restart:
> AIDS:=([97, 206, 406, 700, 1289, 1654, 2576, 3392, 4922, 6343, 8359, 9968, 12990, 14397, 16604,
17124, 19585, 19707, 21392, 20846, 23690, 24610, 26228, 22768, 4903]);
> CAC:=[seq(sum(AIDS[j]/1000.0,j=1..i),i=1..24)];
> Time:=[seq(1981+(i-1)/2,i=1..24)]:
> pts:=[seq([Time[i],CAC[i]],i=1..24)]:
> LnCAC:=map(ln,CAC);
> Times:=[seq((i+1)/2/10,i=1..24)];
> with(stats):
> fit[leastsquare[[x,y],y=m*x+b]]([Times,LnCAC]);
> k:=op(1,op(1,rhs(%)));A:=op(2,rhs(%%));
> y:=t->exp(A)*exp(k*t);
> J:=plot(y((t-1980)/10),t=1980..1992):
> K:=plot(pts,style=POINT,symbol=CIRCLE):
> plots[display]({J,K});
```

MATLAB

```
> AIDS=[97, 206, 406, 700, 1289, 1654, 2576, 3392, 4922, 6343, 8359, 9968, 12990, 14397, 16604,…
        17124, 19585, 19707, 21392, 20846, 23690, 24610, 26228, 22768];
> CAC=cumsum(AIDS)/1000;
> s=size(AIDS); % number of half-years
> count=[0:s(2)-1];
> Time =1981+count/2;
> pts=[Time' CAC'];
> plot(pts(:,1),pts(:,2)); hold on
> Times=(Time-1980)/10; LnCAC=log(CAC);
> MT=[Times; ones(1,s(2))]; % note the space
> params=MT'\LnCAC'
> k=params(1); A=params(2);
> y=exp(A)*exp(k.*Times);
> plot(10*Times+1980,y)
```

4. Table 2.2.5 presents unpublished data that was gathered by Dr. Melinda Millard-Stafford at the Exercise Science Laboratory in the Department of Health and Performance Sciences at Georgia Tech. It relates the circumference of the forearm with grip strength. The first two columns are for a group of college women, and the following two columns are for college men. Find *regression lines* (that is, least square fits) for both sets of data:

MAPLE

```
> CW:=[24.2,22.9,27.,21.5,23.5,22.4, 23.8, 25.5, 24.5,25.5,22.,24.5];
> GSW:=[38.5,26.,34.,25.5,37.,30.,34.,43.5,30.5, 36.,29.,32];
> with(stats):
> fit[leastsquare[[x,y],y=m*x+b]]([CW,GSW]);
> pts:=[seq([CW[i],GSW[i]],i=1..12)];
> J:=plot(pts,style=POINT,symbol=CROSS):
> K:=plot(2.107*x-17.447,x=21..28):
> CM:=[28.5,24.5,26.5,28.25,28.2,29.5,24.5,26.9,28.2,25.6,28.1,27.8,29.5,29.5,29];
> GSM:=[45.8,47.5,50.8,51.5,55.0,51.,47.5,45.,56.0,49.5,57.5,51.,59.5, 58.,68.25];
> fit[leastsquare[[x,y],y=m*x+b]]([CM,GSM]);
> pts:=[seq([CM[i],GSM[i]],i=1..15)];
> L:=plot(pts,style=POINT,symbol=CIRCLE):
> M:=plot(2.153*x-6.567,x=24..30):
> with(plots): display({J,K,L,M});
```

MATLAB

```
> CW=[24.2,22.9,27.,21.5,23.5,22.4,23.8,25.5,24.5,25.5,22.,24.5];
> GSW=[38.5,26.,34.,25.5,37.,30.,34.,43.5,30.5,36.,29.,32];
> MT=[CW; ones(size(CW))];
> parmsW=MT'\GSW';
> plot(CW,GSW,'x'); hold on
```

Table 2.2.5. Forearm and grip strength, males/females.

Females		Males	
Circumference (cm)	**Grip (kg)**	**Circumference (cm)**	**Grip (kg)**
24.2	38.5	28.5	45.8
22.9	26.0	24.5	47.5
27.0	34.0	26.5	50.8
21.5	25.5	28.25	51.5
23.5	37.0	28.2	55.0
22.4	30.0	29.5	51.0
23.8	34.0	24.5	47.5
25.5	43.5	26.9	45.0
24.5	30.5	28.2	56.0
25.5	36.0	25.6	49.5
22.0	29.0	28.1	57.5
24.5	32.0	27.8	51.0
		29.5	59.5
		29.5	58.0
		29.0	68.25

```
> x=21:28; plot(x,parmsW(1)*x+parmsW(2))
  %%%
> CM=[28.5,24.5,26.5,28.25,28.2,29.5,24.5,26.9,28.2,25.6,28.1,27.8,29.5,29.5,29];
> GSM=[45.8,47.5,50.8,51.5,55.0,51.,47.5,45.,56.0,49.5,57.5,51.,59.5,58.,68.25];
> MT=[CM; ones(size(CM))];
> parmsM=MT'\GSM'
> plot(CM,GSM,'o')
> x=24:30;
> plot(x,parmsM(1)*x+parmsM(2))
```

2.3 Multiple Regression

The least squares method extends to experimental models with arbitrarily many parameters. However, the model must be linear in the parameters. The mathematical problem of their calculation can be cast in matrix form, and as such, the parameters emerge as the solution of a linear system. The method is again illustrated with two examples.

Least squares can be extended to more than two parameters

In the previous section, we learned how to perform linear regression, or least squares, on two parameters, to get the slope m and intercept b of a straight-line fit to data. We also saw that the method applies to other models for the data than just the linear model. By a *model* here we mean a mathematical formula of a given form involving unknown parameters. Thus the *exponential model* for (x, y) data is

$$y = Ae^{rx}.$$

And to apply linear regression, we transform it to the form

$$\ln y = rx + \ln A,$$

by taking the logarithm of both sides (cf. (2.1.8)). Here the transformed data is $Y = \ln y$ and $X = x$, while the transformed parameters are $M = r$ and $B = \ln A$. The key requirement of a regression model is that it be linear in the parameters.

Regression principle. *The method of least squares can be adapted to calculate the parameters of a model if there is some transformation of the model that is linear in the transformed parameters.*

Consider the Michaelis–Menten equation for the initial reaction rate v_0 of the enzyme-catalyzed reaction of a substrate having a concentration denoted by $[S]$ (see Section 8.6),

$$v_0 = \frac{v_{\max}[S]}{K_m + [S]};$$

the parameters are $v_{\max}$ and K_m. By taking the reciprocal of both sides of this equation, we get the *Lineweaver–Burk equation*:

$$\frac{1}{v_0} = \frac{K_m}{v_{\max}} \frac{1}{[S]} + \frac{1}{v_{\max}}. \tag{2.3.1}$$

Now the transformed model is linear in its parameters $M = \frac{K_m}{v_{\max}}$ and $B = \frac{1}{v_{\max}}$, and the transformed data are $Y = \frac{1}{v_0}$ and $X = \frac{1}{[S]}$. After determining the slope M and intercept B of a *double reciprocal plot* of $\frac{1}{v_0}$ vs. $\frac{1}{[S]}$ by least squares, then calculate $v_{\max} = \frac{1}{B}$ and $K_m = \frac{M}{B}$.

So far we have looked only at two-parameter models; but the principles apply to models of any number of parameters. For example, the *Merck Manual* (R. Berkow, ed., *The Merck Manual of Diagnosis and Therapy*, 14th ed., Merck, Sharp, and Dohme Research Laboratories, Rahway, NJ, 1982) gives a relationship between the outer surface area of a person as a function of height and weight as follows:

$$\text{surface area} = c \cdot \text{wt}^a \cdot \text{ht}^b,$$

with parameters a, b, and c (a and b have been determined to be 0.425 and 0.725, respectively). A transformed model, linear in parameters, for this is

$$\ln(\text{surface area}) = a \ln(\text{wt}) + b \ln(\text{ht}) + \ln c.$$

The transformed data are triples of values (X_1, X_2, Y), where $X_1 = \ln(\text{wt})$, $X_2 = \ln(\text{ht})$, and $Y = \ln(\text{surface area})$.

We now extend the method of least squares to linear models of r generalized independent variables $X_1, \ldots, X_r$ and one generalized dependent or response variable Y,

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_r X_r.$$

Note that we can recover the two variable case of Section 2.2 by taking $r = 2$ and $X_2 = 1$. Assume that there are n data points $(X_{1,i}, \ldots, X_{r,i}, Y_i)$, $i = 1, \ldots, n$. As before, let e_i denote the error between the experimental value Y_i and the predicted value,

$$e_i = Y_i - (a_1 X_{1,i} + \cdots + a_r X_{r,i}), \quad i = 1, \ldots, n.$$

And as before, we choose parameter values $a_1, \ldots, a_r$ to minimize the squared error,

$$E(a_1, \ldots, a_r) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [Y_i - (a_1 X_{1,i} + \cdots + a_r X_{r,i})]^2.$$

To minimize E, differentiate it with respect to each parameter a_j and set the derivative to zero,

$$0 = \frac{\partial E}{\partial a_j} = -2 \sum_{i=1}^{n} X_{j,i} [Y_i - (a_1 X_{1,i} + \cdots + a_r X_{r,i})], \quad j = 1, \ldots, r.$$

The resulting linear system for the unknowns $a_1, \ldots, a_r$ can be rearranged to the following form (compare with equations (2.2.2)):

$$\begin{aligned} a_1 \sum_i^n X_{1,i} X_{1,i} + \cdots + a_r \sum_i^n X_{1,i} X_{r,i} &= \sum_i^n X_{1,i} Y_i, \\ a_1 \sum_i^n X_{r,i} X_{1,i} + \cdots + a_r \sum_i^n X_{r,i} X_{r,i} &= \sum_i^n X_{r,i} Y_i. \end{aligned} \tag{2.3.2}$$

It is possible to write this system in a very compact way using matrix notation. Let M^T be the matrix of data values of the independent variables,

$$M^T = \begin{bmatrix} X_{1,1} & X_{1,2} & \ldots & X_{1,n} \\ X_{2,1} & X_{2,2} & \ldots & X_{2,n} \\ \vdots & \vdots & \ldots & \vdots \\ X_{r,1} & X_{r,2} & \ldots & X_{r,n} \end{bmatrix}.$$

The ith row of the matrix is the vector of data values of X_i. Represent the data values of the dependent variable Y as a column vector and denote the whole column by $\mathbf{Y}$,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}.$$

Denoting by M the transpose of M^T, the system of equations (2.3.2) can be written in matrix form as

$$M^T M \mathbf{a} = M^T \mathbf{Y}, \tag{2.3.3}$$

where $\mathbf{a}$ is the column vector of regression parameters.

Example 2.3.1. Can body mass and skin fold predict body fat?

Sparling et al. [4] investigate the possibility of predicting body fat from height, weight, and skin fold measurements for women. Percentage body fat can be estimated by two methods: hydrostatic weighing and bioelectric impedance analysis. As in standard practice, height and weight enter the prediction as the fixed combination of weight divided by height squared to form a factor called *body-mass index*,

$$\text{body-mass index} = \frac{\text{weight}}{\text{height}^2}.$$

The assumed relationship is taken as

$$\text{percent body fat} = a * \text{body-mass index} + b * \text{skin fold} + c$$

for some constants a, b, and c.

Table 2.3.1 gives a subset of data of Sparling [4] that we will use for this example to find these constants. The weight and height measurements were made in pounds and inches respectively; body-mass index is to be in kilograms per square meter, so the conversions 0.0254 meter = 1 inch and 2.2046 pounds = 1 kilogram have been done to calculate the body-mass index column of the table.

Table 2.3.1. Height, weight, skin fold, and % body fat for women.

Height (in)	Weight (lbs)	Body mass (kg/m^2)	Skin fold	% Body fat
63.0	109.3	19.36	86.0	19.3
65.0	115.6	19.24	94.5	22.2
61.7	112.4	20.76	105.3	24.3
65.2	129.6	21.43	91.5	17.1
66.2	116.7	18.72	75.2	19.6
65.2	114.0	18.85	93.2	23.9
70.0	152.2	21.84	156.0	29.5
63.9	115.6	19.90	75.1	24.1
63.2	121.3	21.35	119.8	26.2
68.7	167.7	24.98	169.3	33.7
68.0	160.9	24.46	170.0	36.2
66.0	149.9	24.19	148.2	31.0

We compute the third column of Table 2.3.1 from the first two:

MAPLE

```
> ht:=[63,65,61.7,65.2,66.2,65.2,70.0,63.9,63.2,68.7,68,66];
  wt:=[109.3,115.6,112.4,129.6,116.7,114.0,152.2,115.6,121.3,167.7,160.9,149.9];
> convert([seq(wt[i]*lbs/(ht[i]/12*feet)^2,i=1..12)],metric);
```

MATLAB

```
% (1 kg/2.2046 lb)/(0.0254 m/1 in)^2 = 703.076 kg-in^2/lb-m^2
ht=[63,65,61.7,65.2,66.2,65.2,70.0,63.9,63.2,68.7,68,66];
wt=[109.3,115.6,112.4,129.6,116.7,114.0,152.2,115.6,121.3,167.7,160.9,149.9];
bodymass=(wt./(ht.*ht))*703.076;
% this is the M1 in the next step
```

To apply (2.3.3), we take X_1 to be body-mass index, X_2 to be skin fold, and $X_3 = 1$ identically. From the table, M^T is

$$M^T = \begin{bmatrix} 19.36 & 19.24 & 20.76 & 21.43 & 18.72 & \dots & 24.19 \\ 86.0 & 94.5 & 105.3 & 91.5 & 75.2 & \dots & 148.2 \\ 1 & 1 & 1 & 1 & 1 & \dots & 1 \end{bmatrix},$$

and the response vector is

$$\mathbf{Y}^T = \begin{bmatrix} 19.3 & 22.2 & 24.3 & 17.1 & 19.6 & \dots & 31.0 \end{bmatrix}.$$

Solving the system of equations (2.3.3) gives the values of the parameters. We continue the present example:

MAPLE

```
> BMI:=[19.36,19.24, 20.76, 21.43, 18.72, 18.85, 21.84, 19.90, 21.35, 24.98, 24.46, 24.19];
> SF:=[86.0, 94.5,105.3, 91.5, 75.2, 93.2, 156.0, 75.1, 119.8, 69.3, 170.0, 148.2];
> PBF:=[19.3, 22.2, 24.3, 17.1, 19.6, 23.9, 29.5, 24.1, 26.2, 33.7, 36.2, 31.0];
> with(stats):
> fit[leastsquare[[bdymass,sfld,c]]]([BMI,SF,PBF]);
> bdft:=unapply(rhs(%),(bdymass,sfld));
```

MATLAB

```
  % matrix of X values (metric)
> M1=[19.36 19.24 20.76 21.43 18.72 18.85 21.84 19.9 21.35 24.98 24.46 24.19];
> M2=[86.0 94.5 105.3 91.5 75.2 93.2 156.0 75.1 119.8 169.3 170 148.2];
> MT=[M1; M2; ones(1,12)];
  % now vector of corresponding Y values
> Y=[19.3; 22.2; 24.3; 17.1; 19.6; 23.9; 29.5; 24.1; 26.2; 33.7; 36.2; 31.0];
  % do min. norm inversion (i.e., least squares)
> params=MT'\Y
```

$$a = .00656, \qquad b = .1507, \qquad c = 8.074.$$

Thus we find that

$$\begin{aligned} &\text{percent body fat} \\ &\quad \approx .00656 \times \text{body-mass index} + .1507 \times \text{skin fold} + 8.074. \end{aligned} \tag{2.3.4}$$

To test the calculations, here is a data sample not used in the calculation. The subject is 64.5 inches tall, weighs 135 pounds, and has skin fold that measures 159.9 millimeters. Her body-fat percentage is 30.8 as compared to the predicted value of 32.3:

MAPLE

```
> convert(135*lbs/((64.5/12*ft)^2),metric);
> bdft(22.815,159.9);
```

MATLAB

```
  % predict percent body fat for subject 64.5 inches tall, weight of 135 lbs, and skin fold of 159.9 mm
  % 2.2046 lbs per kilogram and 39.37 inches per meter
> bmi= (135/2.2046)/(64.5/39.37)^2
  % so percent body fat is predicted as
> pbf=params(1)*bmi+params(2)*159.9+params(3)
```

$$\text{bdft} = 32.3.$$

Example 2.3.2. Can thigh circumference and leg strength predict vertical jumping ability?

Unpublished data gathered by Dr. Millard-Stafford in the Exercise Science Laboratory at Georgia Tech relates men's ability to jump vertically to the circumference of the thigh and leg strength as measured by leg press. The correlation was to find a, b, and c such that

$$\text{jump height} = a * (\text{thigh circumference}) + b * (\text{bench press}) + c.$$

Hence the generalized variable X_1 is thigh circumference, X_2 is bench press, and $X_3 = 1$.

Data from a sample of college-age men is shown in Table 2.3.2. From the table,

$$M^T = \begin{bmatrix} 58.5 & 50 & 59.5 & 58 & \dots & 56.25 \\ 220 & 150 & 165 & 270 & \dots & 200 \\ 1 & 1 & 1 & 1 & \dots & 1 \end{bmatrix}$$

and

$$\mathbf{Y}^T = \begin{bmatrix} 19.5 & 18 & 22 & 19 & \dots & 29 \end{bmatrix}.$$

Solutions for (2.3.3) for these data are approximately found:

MAPLE

```
> thigh:=[58.5, 50, 59.5, 58, 60.5, 57.5, 49.3, 53.6, 58.3, 51, 54.2, 54, 59.5, 57.5, 56.25];
> press:=[220,150,165,270,200,250,210,130,220,165,190,165,280,190,200];
> jump:=[19.5,18,22,19,21,22,29.5,18,20,20,25,17,26.5,23,29];
```

Table 2.3.2. Leg size, strength, and jumping ability for men.

Thigh average circumference (cm)	Leg press (lbs)	Vertical jump (in)
58.5	220	19.5
50.0	150	18.0
59.5	165	22.0
58.0	270	19.0
60.5	200	21.0
57.5	250	22.0
49.3	210	29.5
53.6	130	18.0
58.3	220	20.0
51.0	165	20.0
54.2	190	25.0
54.0	165	17.0
59.5	280	26.5
57.5	190	23.0
56.25	200	29.0

```
> with(stats):
> fit[leastsquare[[x,y,z], z=a*x+b*y+c, {a,b,c}]]([thigh,press,jump]);
```

```
  MATLAB
> M1=[58.5 50.0 59.5 58.0 60.5 57.5 49.3 53.6 58.3 51.0 54.2 54.0 59.5 57.5 56.25];
> M2=[220 150 165 270 200 250 210 130 220 165 190 165 280 190 200];
> MT=[M1; M2; ones(1,15)];
  % now vector of corresponding Y values
> YT=[19.5 18.0 22.0 19.0 21.0 22.0 29.5 18.0 20.0 20.0 25.0 17.0 26.5 23.0 29.0];
  % min norm inversion
> params=MT'\(YT')
```

$$a = -.29, \qquad b = .044, \qquad c = 29.5.$$

Hence multilinear regression predicts that the height a male can jump is given by the formula

$$\begin{aligned} &\text{jump height} \\ &\quad \approx -.029 \times (\text{thigh circumference}) + 0.044 \times (\text{bench press}) + 29.5. \end{aligned} \tag{2.3.5}$$

Surprisingly, the coefficient of the thigh circumference term is negative, which suggests that thick thighs hinder vertical jumping ability.

Exercises

1. This exercise will review some of the arithmetic for matrices and vectors:

```
  MAPLE
> with(LinearAlgebra);
> A:=Matrix([[a,b],[c,d],[e,f]]); C:=Vector([c1,c2]);
```

```
  MATLAB
> a=1; b=2; c=3; d=4; e=5; f=6; c1=7; c2=8;
> A=[a,b; c,d; e,f]
> C=[c1; c2]
```

Multiplication of the matrix A and the vector c produces a vector:

```
  MAPLE
> A.C;
```

```
  MATLAB
> A*C
```

An interchange of rows and columns of A produces the *transpose* of A. A matrix can be multiplied by its transpose:

```
  MAPLE
> Transpose(A).A;
```

```
  MATLAB
> A'*A
```

2. Compute the solution for Example 2.3.1 using the matrix structure. The following syntax will accomplish this:

```
  MAPLE
> with(LinearAlgebra):
> M:=Matrix([[19.36, 86, 1], [19.24, 94.5, 1], [20.76, 105.3, 1], [21.43, 91.5, 1], [18.72, 75.2, 1],
            [18.85, 93.2, 1], [21.84, 156.0, 1], [19.9, 75.1, 1], [21.35, 119.8, 1], [24.98, 169.3, 1],
            [24.46, 170., 1], [24.19, 148.2, 1]]);
> evalm(transpose(M)); # or Transpose(M)
```

```
> A:=evalm(transpose(M).M);
> z:=vector([19.3, 22.2, 24.3, 17.1, 19.6, 23.9, 29.5, 24.1, 26.2, 33.7, 36.2, 31.0]);
> y:=evalm(transpose(M).z);
> evalm(A^(-1).y);
```

MATLAB

```
> M1=[19.36 19.24 20.76 21.43 18.72 18.85 21.84 19.9 21.35 24.98 24.46 24.19];
> M2=[86.0 94.5 105.3 91.5 75.2 93.2 156.0 75.1 119.8 169.3 170 148.2];
> MT=[M1; M2; ones(1,12)];
  % each row = multiplier of a parameter
> M=MT' % transpose of MT
> A= MT*M % square 3x3 matrix
> z=[19.3; 22.2; 24.3; 17.1; 19.6; 23.9; 29.5; 24.1; 26.2; 33.7; 36.2; 31.0]; % 12x1 vector
> y=MT*z % 3x1 vector
> params=inv(A)*y
> MT'\z % same thing
```

3. (a) In this exercise, we get a linear regression fit for some hypothetical data relating age, percentage body fat, and maximum heart rate. (See Table 2.3.3.) Maximum heart rate is determined by having an individual exercise until near complete exhaustion.

Table 2.3.3. Data for age, % body fat, and maximum heart rate.

Age (years)	% Body fat	Maximum heart rate
30	21.3	186
38	24.1	183
41	26.7	172
38	25.3	177
29	18.5	191
39	25.2	175
46	25.6	175
41	20.4	176
42	27.3	171
24	15.8	201

The syntax that follows will get a linear regression fit for these data. This syntax will also produce a plot of the regression plane. Observe that it shows a steep decline in maximum heart rate as a function of age and a lesser decline with increased percentage body fat.

(b) As an example of the use of this regression formula, compare the predicted maximum heart rate for two persons at age 40 where one has maintained 15% body fat and the other has gained weight to 25% body fat. Also, compare two people with 20% body fat where one is age 40 and the other is age 50:

MAPLE

```
> age:=[30,38,41,38,29,39,46,41,42,24];
> BF:= [21.3,24.1,26.7,25.3,18.5,25.2,25.6,20.4,27.3,15.8];
> hr:=[186,183,172,177,191,175,175,176,171,201];
> with(stats):
> fit[leastsquare[[a,b,c]]]([age,BF,hr]);
> h:=unapply(rhs(%),(a,b));
> plot3d(h(a,b),a=30..60,b=10..20,axes=NORMAL);
> h(40,15); h(40,25); h(40,20); h(50,20);
```

MATLAB
```
> age=[30,38,41,38,29,39,46,41,42,24];
> BF=[21.2,24.1,36.7,25.3,18.5,25.2,25.6,20.4,27.3,15.8];
> hrt=[186,183,172,177,191,175,175,176,171,201];
> MT=[age;BF; ones(size(age))];
> parms=MT'\hrt'
> [Xage YBF]=meshgrid(age,BF);
> R=parms(1)*Xage+parms(2)*YBF+parms(3);
> C=ones(size(R)); % for a uniform color
> surf(age,BF,R,C) % surface graph
> h=[40 15 1]*parms
> h=[40 25 1]*parms
> h=[40 20 1]*parms
> h=[50 20 1]*parms
```

4. Table 2.3.4 contains further data to relate leg size, strength, and the ability to jump. These data were gathered for college women.

Table 2.3.4. Leg size, strength, and jumping ability for women.

Thigh circumference (cm)	Leg press (lbs)	Vertical jump (in)
52.0	140	13.0
54.2	110	8.5
64.5	150	13.0
52.3	120	13.0
54.5	130	13.0
58.0	120	13.0
48.0	95	8.5
58.4	180	19.0
58.5	125	14.0
60.0	125	18.5
49.2	95	16.5
55.5	115	10.5

Find a least squares data fit for these data, which are from unpublished work by Dr. Millard-Stafford in the Health and Performance Science Department at Georgia Tech.

2.4 Modeling with Differential Equations

Understanding a natural process quantitatively often leads to a differential equation model. Consequently, a great deal of effort has gone into the study of differential equations. The theory of linear differential equations, in particular, is well known, and not without reason, since this type occurs widely.

Besides their exact solution in terms of functions, numerical and asymptotic solutions are also possible when exact solutions are not available.

In differential equations, as with organisms, there is need of a nomenclature.

In Section 2.1, we proposed a simple differential equation for mimicking the growth of a biological population, namely,

$$\frac{dy}{dt} = ky. \tag{2.4.1}$$

A *differential equation* refers to any equation involving derivatives. Other examples are

$$\frac{d^2y}{dt^2} - 4\frac{dy}{dt} + 4y = e^{-t} \tag{2.4.2}$$

and

$$\frac{dy}{dt} = y - \frac{y^2}{2 + \sin t} \tag{2.4.3}$$

and many others. If only first-order derivatives appear in a differential equation, then it is called a *first-order* equation. Both equations (2.4.1) and (2.4.3) are of first order, but (2.4.2) is a second-order equation. Every first-order differential equation can be written in the form

$$\frac{dy}{dt} = f(t, y) \tag{2.4.4}$$

for some function f of two variables. Thus $f(t, y) = ky$ in the first equation above and $f(t, y) = y - \frac{y^2}{2+\sin t}$ in the third.

A *solution* of a differential equation means a function $y = y(t)$ that satisfies the equation for all values of t (over some specified range of t values). Thus $y = Ae^{kt}$ is a solution of (2.4.1) because then $\frac{dy}{dt} = kAe^{kt}$, and substitution into (2.4.1) gives

$$kAe^{kt} = k(Ae^{kt}),$$

true for all t. Note that A is a parameter of the solution and can be any value, so it is called an *arbitrary constant*. Recalling Section 2.1, A arose as the constant of integration in the solution of (2.4.1). In general, the solution of a first-order differential equation will incorporate such a parameter. This is because a first-order differential equation is making a statement about the slope of its solution rather than the solution itself.

To fix the value of the inevitable arbitrary constant arising in the solution of a differential equation, a point in the plane through which the solution must pass is also specified, for example at $t = 0$. A differential equation along with such a side condition is called an *initial value problem*,

$$\frac{dy}{dt} = f(t, y) \quad \text{and} \quad y(0) = y_0. \tag{2.4.5}$$

It is not required to specify the point for which $t = 0$. It could be any other value of t for which $y(t)$ is known. The *domain of definition*, or simply domain, of the differential equation is the set of points (t, y) for which the right-hand side of (2.4.4) is defined. Often this is the entire (t, y)-plane.

Initial value problems can be solved analytically.

Exact solutions are known for many differential equations; cf. Kamke [5]. For the most part, solutions derive from a handful of principles. Although we will not study solution techniques here to any extent, we make two exceptions and discuss methods for linear systems below and the method of separation of variables next.

Actually we have already seen variables separable at work in Section 2.1: The idea is to algebraically modify the differential equation in such a way that all instances of the independent variable are on one side of the equation and all those of the dependent variable are on the other. Then the solution results as the integral of the two sides. For example, consider

$$\frac{dy}{dt} = ay - by^2.$$

Dividing by the terms on the right-hand side and multiplying by dt separates the variables, leaving only the integration to be done:

$$\int \frac{dy}{y(a-by)} = \int dt.$$

Instead of delving into solution methods further, our focus in this text is deciding what solutions mean and which equations should constitute a model in the first place. Happily, some of the solution techniques, such as separation of variables, are sufficiently mechanical that computers can handle the job, relieving us for higher-level tasks. Here then are (symbolic) solutions to equations (2.4.2) and (2.4.3):

MAPLE
```
> restart:
> dsolve(diff(y(t),t,t)-4*diff(y(t),t)+4*y(t)=exp(-t),y(t));
```

$$y(t) = \frac{1}{9} + C_1 e^{2t} + C_2 t e^{2t}$$

and

MAPLE
```
> dsolve(diff(y(t),t)=y(t)-y(t)^2/(2+sin(t)), y(t));
```

$$\frac{1}{y(t)} = e^{-t} \int \frac{e^t}{2+\sin(t)} dt + e^{-t} C_1.$$

Initial value problems can be solved numerically.

As mentioned above, (2.4.4) specifies the slope of the solution required by the differential equation at every point (t, y) in the domain. This may be visualized by plotting a short line segment having that slope at each point. This has been done in Figure 2.4.1 for (2.4.3). Such a plot is called a *direction field*. Solutions to the equation must follow the field and cannot cross slopes. With such a direction field it is possible to sketch solutions manually. Just start at the initial point $(0, y(0))$ and follow the direction field. Keep in mind that a figure such as Figure 2.4.1 is only a

MAPLE
```
> with(DEtools):
> dfieldplot(diff(y(t),t)=y(t)-y(t)^2/(2+sin(t)),y(t), t=0..5,y=-1..5);
```

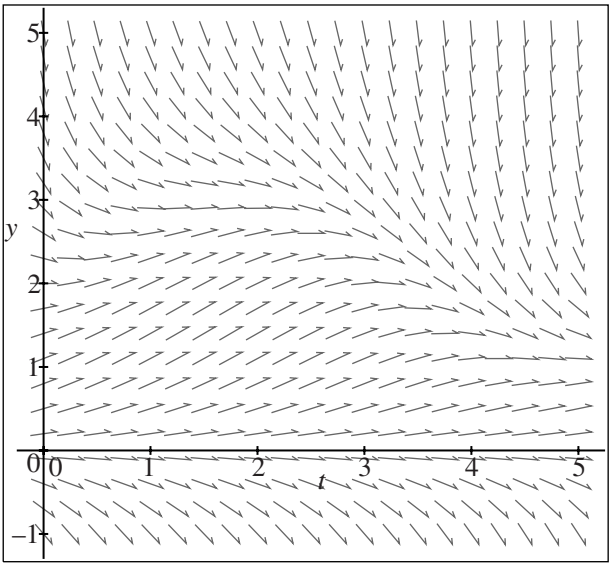


Fig. 2.4.1. Direction field for (2.4.3).

representation of the true direction field, that is to say, it shows only a small subset of the slope segments.

The mathematician Euler realized three centuries ago that the direction field could be used to numerically approximate solutions of an initial value problem in a precise way. Since Euler's time, techniques have improved—*Runge–Kutta methods* are used today—but the spirit of *Euler's method* is common to most of them; namely, the solution takes a small step Δt to the right and Δy up, where

$$\Delta y = f(t_i, y_i) \cdot \Delta t.$$

The idea is that $\frac{\Delta y}{\Delta t}$ approximates $\frac{dy}{dt}$. These increments are stepped off one after another,

$$y_{i+1} = y_i + \Delta y, \qquad t_{i+1} = t_i + \Delta t, \quad i = 0, 1, 2, \ldots,$$

with starting values $y_0 = y(0)$ and $t_0 = 0$. Figure 2.4.2 shows some numerical solutions of (2.4.3).

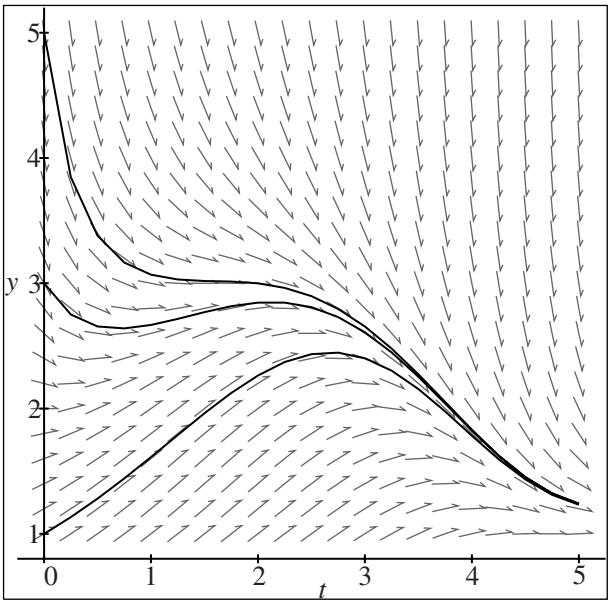


Fig. 2.4.2. Solutions and direction field for (2.4.3).

Code 2.4.1.

MAPLE

```
> with(DEtools):
> DEplot(diff(y(t),t)=y(t)-y(t)^2/(2+sin(t)),y(t), t=0..5,{[0,1],[0,3],[0,5]}, linecolor=BLACK);
```

MATLAB

```
% make up an m-file, ode243.m, as follows
% function yprim=ode243(t,y)
% yprim = y - (y.^2./(2+sin(t)));
% now for the solution with initial value=1
> tspan=[0 5];
> [t1,y1]=ode23('ode243',tspan,1);
% and for initial value=3
> [t3,y3]=ode23('ode243',tspan,3);
% and for initial value=5
> [t5,y5]=ode23('ode243',tspan,5);
% plot them all
> plot(t1,y1,t3,y3,t5,y5);
```

Linear differential equations are among the simplest kind.

A differential equation that can be put into the form

$$a_n(t)\frac{d^n y}{dt^n} + \cdots + a_2(t)\frac{d^2 y}{dt^2} + a_1(t)\frac{dy}{dt} + a_0(t)y = r(t) \tag{2.4.6}$$

is *linear*. The coefficients $a_i(t)$, $i = 0, \ldots, n$, can be functions of t, as can the *right-hand side* $r(t)$. Equations (2.4.1) and (2.4.2) are linear but (2.4.3) is not. When there are multiplications among the derivatives or the dependent variable y, such as y^2, the differential equation will not be linear. If $y_1(t)$ and $y_2(t)$ are both solutions

to a linear differential equation with right-hand side 0, then so is $Ay_1(t) + By_2(t)$ for any constants A and B. Consider the first-order linear differential equation

$$\frac{dy}{dt} = my + R(t), \tag{2.4.7}$$

where we have taken $m = -\frac{a_0}{a_1}$ and $R(t) = \frac{r(t)}{a_1}$ in (2.4.6). Its solution is

$$y = Ae^{g(t)} + \Phi(t), \quad \text{where } g(t) = \int m dt. \tag{2.4.8}$$

In this, A is the arbitrary constant and Φ is given below. To see this, first assume that R is 0, and write the differential equation as

$$\frac{dy}{y} = mdt.$$

Now integrate both sides, letting $g(t) = \int mdt$ and C be the constant of integration,

$$\ln y = g(t) + C, \quad \text{or} \quad y = Ae^{g(t)},$$

where $A = e^C$. By direct substitution, it can be seen that

$$\Phi = e^{g(t)} \int e^{-g(t)} R(t) dt \tag{2.4.9}$$

is a solution.[5] But it has no arbitrary constant, so add the two solutions, linearity allows this, to get (2.4.8). If m is a constant, then $\int mdt = mt$.

To see that finding this solution is mechanical enough that a computer can handle the job, try these commands:

MAPLE
```
> dsolve(diff(y(t),t)=m(t)*y(t)+R(t),y(t));
> dsolve(diff(y(t),t)=m*y(t)+R(t),y(t));
```

Systems of differential equations generalize their scalar counterparts.

Quite often, modeling projects involve many more variables than two. Consequently it may require several differential equations to adequately describe the phenomenon. Consider the following model for small deviations about steady-state levels of a glucose/insulin system; g denotes the concentration of glucose and i the same for insulin,

$$\begin{aligned} \frac{dg}{dt} &= -\alpha g - \beta i + p(t), \\ \frac{di}{dt} &= \gamma g - \delta i. \end{aligned} \tag{2.4.10}$$

[5] A clever idea is to try a solution of the form $y = v(t)e^{g(t)}$ with $v(t)$ unknown and substitute this into (2.4.7) to get $v'e^{g(t)} = R(t)$, since the term $vg'e^{g(t)} = vme^{g(t)}$ drops out. Now solve for v.

As discussed in Section 2.1, the second equation expresses a proportionality relationship, namely, the rate of secretion of insulin increases in proportion to the concentration of glucose but decreases in proportion to the concentration of insulin. (Modeling coefficients are assumed to be positive unless stated otherwise.) The first equation makes a similar statement about the rate of removal of glucose, except that there is an additional term, $p(t)$, which is meant to account for ingestion of glucose. Because glucose and insulin levels are interrelated, each equation involves both variables. The equations define a system; the differential equations have to be solved simultaneously.

A system of differential equations can be written in vector form by defining a vector, say $\mathbf{Y}$, whose components are the dependent variables of the system. In vector notation, (2.4.10) becomes

$$\frac{d\mathbf{Y}}{dt} = M\mathbf{Y} + \mathbf{P}, \tag{2.4.11}$$

where the matrix M and vector $\mathbf{P}$ are

$$M = \begin{bmatrix} -\alpha & -\beta \\ \gamma & -\delta \end{bmatrix}, \qquad \mathbf{P} = \begin{bmatrix} p(t) \\ 0 \end{bmatrix}.$$

Since the system (2.4.10) is linear, its vector expression takes on the simple matrix form of (2.4.11). Furthermore, this matrix system can be solved in the same way as the scalar differential equation (2.4.7). We have

$$\mathbf{Y} = e^{Mt}\mathbf{Y}_0 + e^{Mt}\int_0^t e^{-Ms}\mathbf{P}(\mathbf{s})ds. \tag{2.4.12}$$

Just as the exponential of the scalar product mt is

$$e^{mt} = 1 + mt + \frac{m^2t^2}{2!} + \frac{m^3t^3}{3!} + \cdots, \tag{2.4.13}$$

so the exponential of the matrix product Mt is

$$e^{Mt} = I + Mt + \frac{M^2t^2}{2!} + \frac{M^3t^3}{3!} + \cdots. \tag{2.4.14}$$

Since many properties of the exponential function stem from its power series expansion equation (2.4.13), the matrix exponential enjoys the same properties, in particular, the property that makes for the same form of solution,

$$\frac{d}{dt}e^{Mt}\mathbf{V}(\mathbf{t}) = e^{Mt}\frac{d}{dt}\mathbf{V}(\mathbf{t}) + e^{Mt}M\mathbf{V}(\mathbf{t}).$$

As in the case of a scalar differential equation, the system solutions can be plotted against t to help us understand how the variables behave. For example, we could plot $g(t)$ and $i(t)$ using (2.4.12) (see Figure 2.4.3). But for a system there is an alternative; we can suppress t and plot $i(t)$ against $g(t)$. This is done, conceptually, by making a table of values of t and calculating the corresponding values of g and i. But we only plot (i, g) pairs. The coordinate plane of i and g is called the *phase plane* and the graph is called a *phase portrait* of the solution (see Figure 2.4.4).

```
MAPLE
> GIdeq:= diff(g(t),t)=-g(t)-i(t), diff(i(t),t)=-i(t)+g(t);
> sol:=dsolve({GIdeq, g(0)=1, i(0)=0},{g(t),i(t)}):
> g:= unapply(subs(sol,g(t)),t); i:= unapply(subs(sol,i(t)),t);
> plot({g(t),i(t)},t=0..5);

MATLAB
  % Make up an m-file, fig243.m, as follows
  % function Yprime=fig243(t,x)
  % Yprime = [-x(1) - x(2); x(1) - x(2)];
  % for the solution with initial value g=1 and i=0
> [t,Y]=ode23('fig243',[0 5],[1;0]); % semicolon for column vector
> plot(t,Y) % plot both columns of Y vs. t
```

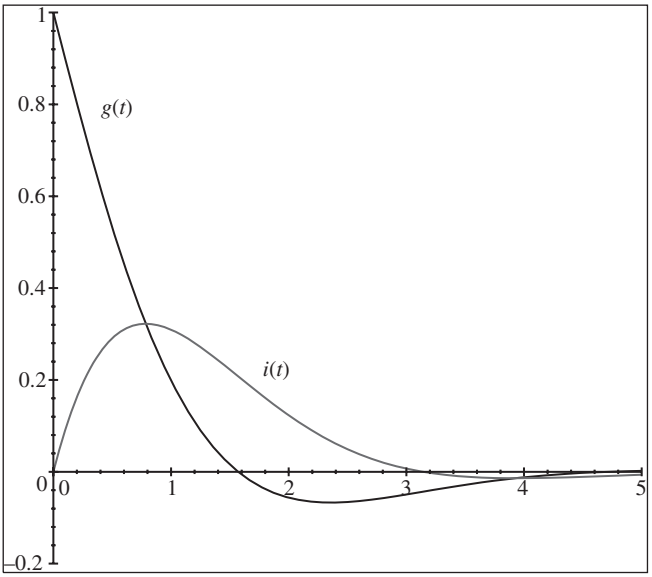


Fig. 2.4.3. Plot of solutions $g(t)$, $i(t)$ of (2.4.10).

Asymptotics predict the ultimate course of the model.

Often in science and engineering, we are interested in forecasting the future behavior of an observed process, $y(t)$. As t becomes large there are several possibilities; among them are the following: y can tend to a finite limit y_∞, known as an *asymptotic limit*,

$$\lim_{t\to\infty} y(t) = y_\infty;$$

y can tend to plus or minus infinity,

$$\lim_{t\to\infty} y(t) = \pm\infty;$$

y can oscillate periodically; y can oscillate unboundedly,

$$\lim_{t\to\infty} |y(t)| = \infty;$$

```
MAPLE
> restart:
> with(DEtools):
> GIdeq:= diff(g(t),t)=-g(t)-i(t), diff(i(t),t)=-i(t)+g(t);
> inits:={[0,1,0],[0,2,0],[0,3,0],[0,4,0]};
> phaseportrait([GIdeq],[g,i],t=0..4,inits, stepsize=.1,g=-1..4,i=-1..1.3);

MATLAB
> [t,Y4]=ode23('fig243',[0 5],[4;0]);
> [t,Y3]=ode23('fig243',[0 5],[3;0]);
> [t,Y2]=ode23('fig243',[0 5],[2;0]);
> [t,Y1]=ode23('fig243',[0 5],[1;0]);
> plot(Y4(:,1),Y4(:,2)) % plot the first component of Y4 against the second
> hold on
> plot(Y3(:,1),Y3(:,2)) %ditto for Y3
> plot(Y2(:,1),Y2(:,2)) %ditto for Y2
> plot(Y1(:,1),Y1(:,2)) %ditto for Y1
```

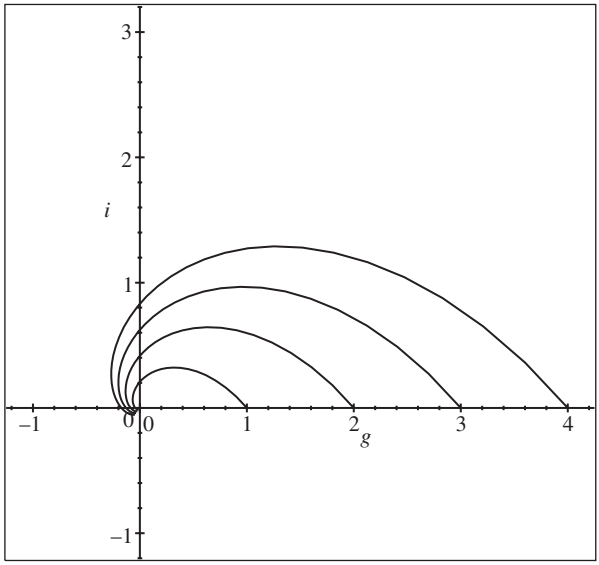


Fig. 2.4.4. Phase portrait for (2.4.10).

or y can oscillate chaotically. If y is part of a system, its fate will be linked to that of the other variables; in this case, we inquire about the vector solution $\mathbf{Y}$.

In the simplest case, $\mathbf{Y}$ has asymptotic limits. If the system is *autonomous*, meaning t appears nowhere in the system (except, of course, in the form $\frac{d}{dt}$), then to find the asymptotic limits, set all the derivatives of the system to zero. Solutions of the resulting algebraic system are called *critical points* or *stationary points*.[6] In the glucose/insulin example, suppose the glucose ingestion term, $p(t)$, were constant at p; then setting the derivatives to zero leads to the algebraic system

$$\begin{aligned} 0 &= -\alpha g - \beta i + p, \\ 0 &= \gamma g - \delta i. \end{aligned} \tag{2.4.15}$$

[6] These are also called *equilibrium points* by some authors.

MAPLE
```
> solve({-alpha*g-beta*i+p=0,gamma*g-delta*i=0},{g,i});
```

Its one critical point is $g = -\frac{\delta p}{\gamma\beta+\alpha\delta}$, $i = \frac{\gamma p}{\gamma\beta+\alpha\delta}$. If this point is taken as the initial point of the system, then for all time, g will be $\frac{\delta p}{\gamma\beta+\alpha\delta}$ and i will be $\frac{\gamma p}{\gamma\beta+\alpha\delta}$.

It is not necessarily the case that a stationary point is also an asymptotic limit. Exponential growth, $\frac{dy}{dt} = y$, is an example, since $y = 0$ is a stationary point, but if $y(0) \neq 0$, then $y \to \infty$ as $t \to \infty$. On the other hand, when it can be shown that the solution of a system tends to an asymptotic limit, a giant step has been taken in understanding the system. For example, exponential decay, $\frac{dy}{dt} = -y$, has asymptotic limit 0 for any starting point $y(0)$, for if $y > 0$, then $\frac{dy}{dt}$ is negative, so y will decrease. Similarly, if $y < 0$, then $\frac{dy}{dt} > 0$, so y will increase. Either way, 0 is the asymptotic limit.

A complication here is that the existence or the value of the asymptotic limit can often depend on the starting point $\mathbf{Y}(0)$. Given that there is an asymptotic limit, $\mathbf{Y}_\infty$, the set of all starting points for which the solution tends to $\mathbf{Y}_\infty$ is called its *basin of attraction*, $\mathbf{B}_{\mathbf{Y}_\infty}$,

$$\mathbf{B}_{\mathbf{Y}_\infty} = \left\{\mathbf{Y}_0 : \lim_{t\to\infty} \mathbf{Y}(t) = \mathbf{Y}_\infty \text{ when } \mathbf{Y}(0) = \mathbf{Y}_0\right\}.$$

If the basin of attraction of a system is essentially the entire domain of definition, the asymptotic limit is said to be *global*. By way of example, the differential equation $\frac{dy}{dt} = -y(1-y)$ has asymptotic limit $y = 0$ for solutions starting from $-\infty < y_0 < 1$; but when the starting point is beyond 1, solutions tend to infinity.

Periodicity is a more complicated asymptotic behavior. Further, just as in the asymptotic limit case, the solution can start out periodic, or can asymptotically tend to periodicity. An example of the former is $\frac{dy}{dt} = \cos t$, while the latter behavior is demonstrated by $\frac{dy}{dt} = -y + \cos t$. This second differential equation is solved by (2.4.8), $y = Ae^{-t} + \frac{1}{2}(\cos t + \sin t)$; A depends on the initial condition, but the whole term tends to zero. A well-known periodic system is the one due to Lotka and Volterra modeling predator–prey interaction. We study this in Section 4.4.

Exercises

1. Here are four differential equations with the same initial conditions:

$$\begin{aligned}
\frac{d^2y}{dt^2} + 6y(t) &= 0, \quad y(0) = 1, \quad y'(0) = 0;\\
\frac{d^2y}{dt^2} - 6y(t) &= 0, \quad y(0) = 1, \quad y'(0) = 0;\\
\frac{d^2y}{dt^2} + 2\frac{dy}{dt} + 6y(t) &= 0, \quad y(0) = 1, \quad y'(0) = 0;\\
\frac{d^2y}{dt^2} - 2\frac{dy}{dt} + 6y(t) &= 0, \quad y(0) = 1, \quad y'(0) = 0.
\end{aligned}$$

While these differential equations have a similar appearance, they have radically different behaviors. Sketch the graphs of all four equations with the same initial values. Here is syntax that will draw the graphs:

MAPLE
```
> dsolve({diff(y(t),t,t)+6*y(t)=0, y(0)=1, D(y)(0)=0},y(t));
> y1:=unapply(rhs(%),t);
> dsolve({diff(y(t),t,t)-6*y(t)=0, y(0)=1, D(y)(0)=0},y(t));
> y2:=unapply(rhs(%),t);
> dsolve({diff(y(t),t,t)+2*diff(y(t),t)+6*y(t)=0, y(0)=1,D(y)(0)=0},y(t));
> y3:=unapply(rhs(%),t);
> dsolve({diff(y(t),t,t)-2*diff(y(t),t)+6*y(t)=0, y(0) = 1,D(y)(0)= 0},y(t));
> y4:=unapply(rhs(%),t);
> plot([y1(t),y2(t),y3(t),y4(t)],t=0..4,y=-5..5, color=[black,blue,green,red]);
```

MATLAB
```
% to deal with a second-order differential equation, it has to be made into a vector-valued first-order
%   DE as follows: the first component Y1 is y and the second Y2 is dy/dt. Then dˆ2y/dtˆ2+6y=0
%   becomes the vector system dY1/dt=Y2; dY2/dt = -6Y1;
%   so make an m-file, exer241a.m, as follows:
%      function Yprime=exer241a(t,Y); Yprime = [Y(2); -6*Y(1)];
> [t,Y]=ode23('exer241a',[0 4],[1; 0]);
> plot(t,Y(:,1))
  %%%
  % DE (b) converts to first-order vector system dY1/dt=Y2; dY2/dt=6*Y1;
  % DE (c) converts to first-order vector system dY1/dt=Y2; dY2/dt=-6*Y1-2*Y2;
  % DE (d) converts to first-order vector system dY1/dt=Y2; dY2/dt=-6*Y1+2*Y2;
  % We leave it to the reader to obtain the numerical solutions and plot as above.
```

2. We illustrate four ways to visualize solutions to a single second-order differential equation in order to emphasize that different perspectives provide different insights. We use the same equation in all four visualizations:

$$\frac{d^2y}{dt^2} + \frac{y(t)}{5} = \cos(t).$$

(a) Find and graph an analytic solution that starts at $y(0) = 0$.

MAPLE
```
> dsolve({diff(y(t),t,t)+y(t)/5=cos(t), y(0)=0,D(y)(0)=0},y(t));
> y:=unapply(rhs(%),t);
> plot(y(t),t=0..4*Pi);
```

MATLAB
```
% make an m-file, exer242.m, with
%   function Yprime=exer242(t,Y); Yprime=[Y(2); -Y(1)/5+cos(t)];
% then solve and plot with
> [t,Y]=ode23('exer242',[0 4*pi],[0;0]);
> plot(t,Y(:,1))
```

(b) Give a direction field for the equation.

MAPLE
```
> restart: with(DEtools):
> dfieldplot(diff(y(t),t)+y(t)/5=cos(t),y(t),t=0..4*Pi,y=-1..5);
```

MATLAB
```
% No built-in direction field in Matlab; see DFIELD from http://math.rice.edu/~dfield.
```

(c) Give several trajectories overlaid in the direction field.

MAPLE (direction field)
```
> restart:
> with(DEtools):
> DEplot(diff(y(t),t)+y(t)/5=cos(t),y(t),t=0..4*Pi,{[0,1],[0,3],[0,5]});
```

(d) Give an animation to show the effect of the coefficient of $y(t)$ changing.

MAPLE (animation)

```
> restart: with(plots):
> for n from 1 to 8 do
  a:=n/10:
  dsolve({diff(y(t),t,t)+a*y(t)/5=cos(t),y(0)=0},y(t)):
  y:=unapply(rhs(%),t):
  P[n]:=plot([t,y(t),t=0..10*Pi],t=0..10*Pi):
  y:='y':
  od:
> display([seq(P[n],n=1..8)],insequence=true);
```

3. Find the critical points for each of the following equations. Plot a few trajectories to confirm where the basins of attractions are.

(a) $\frac{dy}{dt} = -y(t)(1 - y(t))$.

MAPLE

```
> solve(y*(1-y)=0,y);
> with(DEtools):
> de:=diff(y(t),t)=-y(t)*(1-y(t));
> DEplot(de,y(t),t=0..5,{[0,-1],[0,-1/2],[0,1/2]},y=-1..2);
```

MATLAB

```
  % make an m-file, exer243a.m, with
  %   function yprime=exer243a(t,y); yprime=-y.*(1-y);
> p=[1 -1 0]; % coefficients of p(y)=-y(1-y)
> roots(p)
> [t,y]=ode23('exer243a',[0 5],-1);
> plot(t,y); hold on
> [t,y]=ode23('exer243a',[0 5],-1/2);
> plot(t,y)
> [t,y]=ode23('exer243a',[0 5],1/2);
> plot(t,y)
```

(b) $x' = 4x(t) - x\hat{}2(t) - x(t)y(t)$; $y' = 5y(t) - 2y\hat{}2(t) - x(t)y(t)$.

MAPLE

```
> solve({4*x-x^2-x*y=0, 5*y-2*y^2-y*x=0}, {x,y});
> with(DEtools):
> deq1:=diff(x(t),t)=4*x(t)-x(t)^2-x(t)*y(t);
> deq2:=diff(y(t),t)= 5*y(t)-2*y(t)^2-y(t)*x(t);
> inits:={[0,1,1],[0,1,4],[0,4,1],[0,4,4]};
> DEplot([deq1,deq2],[x,y],t=0..4,inits,x=-1..5,y=-1..5,stepsize=.05);
```

MATLAB

```
  % contents of m-file, exer243b.m:
  % function Yprime=exer243b(t,Y);
  % Yprime=[4*Y(1)-Y(1).*Y(1)-Y(1).*Y(2); 5*Y(2)-2*Y(2).^2-Y(1).*Y(2)];
> [t,Y]=ode23('exer243b',[0 4],[1;1]);
> hold off; plot3(t,Y(:,1),Y(:,2))
> grid
> xlabel('x axis'); ylabel('y axis');
> zlabel('z axis'); hold on
> [t,Y]=ode23('exer243b',[0 4],[1;4]);
> plot3(t,Y(:,1),Y(:,2))
> [t,Y]=ode23('exer243b',[0 4],[4;1]);
> plot3(t,Y(:,1),Y(:,2))
> [t,Y]=ode23('exer243b',[0 4],[4;4]);
> plot3(t,Y(:,1),Y(:,2))
> view(30,30) % 30 deg CCW from negative y-axis, 30 deg elevation
> view(-100,30) % 100 deg CW from negative y-axis, 30 deg elevation
```

4. The solution for $Z' = AZ(t)$, $Z(0) = C$, with A a constant square matrix and C a vector is $\exp(At)C$. Compute this exponential in the case

$$A = \begin{pmatrix} -1 & -1 \\ 1 & -1 \end{pmatrix}.$$

Evaluate $\exp(At)C$, where C is the vector

$$C = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

MAPLE

```
> with(LinearAlgebra):
> A:=Matrix([[-1,-1],[1,-1]]);
> MatrixExponential(A,t);
> evalm(%.[1,0]);
```

MATLAB

```
> A=[-1, -1; 1, -1]
> t=2; At=A*t; expm(At)
> t=5; At=A*t; expm(At)
> expm(At)*[1;0] % exp(At)*C, where C is a 2x1 column vector
```

2.5 Modeling with Difference Equations

Biological systems are not always continuous. Considering population growth, individuals come in discrete units, so a differential equation model for population growth is only an approximation. When population size is large, the approximation is sufficiently accurate to describe the model's behavior and asymptotics. But there are many biological phenomena whose analysis requires a treatment in terms of discrete units.

Difference equations are similar to differential equations except that the independent variable, time or space, is taken in discrete, indivisible units. Although difference equation analysis is often more difficult than its continuous counterpart, there is a striking analogy between the two theories.

Difference equations are one example of what is more generally known as *recurrence relations*. This refers to some quantity that is defined in terms of itself.

Just as numerical and asymptotic analyses are available for differential equations, the same holds for difference equations as well.

A differential equation has a natural difference equation counterpart.

In Section 2.1 we mentioned a differential equation model for population growth,

$$\frac{dy}{dt} = ky. \tag{2.5.1}$$

This model postulates that infinitesimal units of population, dy, are added to the general population over infinitesimal units of time, dt. Of course this can only be an approximation. And indeed it is an adequate one in many cases, for example, for describing a bacterial colony.

However, for a more accurate description, an approach respecting that biological units are discrete and reproductive intervals are also discrete is called for. We are led to the discrete version of (2.5.1),

$$y_{t+1} - y_t = r y_t.$$

Here the variable t proceeds in discrete units $t = 0, 1, 2, \ldots$. As in the differential equation, a starting value y_0 is required to complete the description.

To solve the difference equation we recast it as a recurrence relation together with a starting value (denote this by y_0),

$$y_{t+1} = (1 + r)y_t, \quad y_0 = \text{starting value}.$$

The solution is easy to obtain by stepping through the generations recurrently,

$$\begin{aligned} y_1 &= (1 + r)y_0, \\ y_2 &= (1 + r)y_1 = (1 + r)^2 y_0, \\ y_3 &= (1 + r)y_2 = (1 + r)^3 y_0, \end{aligned}$$

and so on. It is easy to see that there is a closed (nonrecurrent) form for the y_t, namely,

$$y_t = (1 + r)^t y_0, \quad t = 0, 1, 2, \ldots.$$

Comparing this with the solution of the differential equation,

$$y = e^{kt} y_0 = (e^k)^t y_0,$$

shows that e^k corresponds to $1 + r$. The relationship between the per period growth rate r and the instantaneous growth rate k is

$$r = e^k - 1 \quad , \text{or} \quad k = \log(1 + r). \tag{2.5.2}$$

A second-order differential equation such as

$$\frac{d^2y}{dt^2} - 4\frac{dy}{dt} + 4y = 0 \tag{2.5.3}$$

can be written as a difference equation by noting how the second derivative converts. Since $\frac{d^2y}{dt^2} = \frac{d}{dt}(\frac{dy}{dt})$, we may write

$$\begin{aligned} \frac{d^2y}{dt^2} &\to \left.\frac{dy}{dt}\right|_{t+1} - \left.\frac{dy}{dt}\right|_t \\ &\to (y_{t+2} - y_{t+1}) - (y_{t+1} - y_t) = y_{t+2} - 2y_{t+1} + y_t. \end{aligned}$$

Then (2.5.3) becomes

$$y_{t+2} - 2y_{t+1} + y_t - 4(y_{t+1} - y_t) + 4y_t = 0.$$

This may be written as the linear recurrence relation

$$y_{t+2} - 6y_{t+1} + 9y_t = 0.$$

Just as a second-order differential equation requires two initial values for a complete solution, so also a second-order recurrence relation requires two initial values for a complete solution.

The general second-order (homogeneous) recurrence relation is

$$c_2 y_{t+2} + c_1 y_{t+1} + c_0 y_t = 0 \tag{2.5.4}$$

for some constants c_2, c_1, and c_0. On the strength of what we saw above, we expect a solution of the form $y_t = Ar^t$ for some r. Substitute this into (2.5.4):

$$c_2 Ar^{t+2} + c_1 Ar^{t+1} + c_0 Ar^t = 0.$$

Factoring out Ar^t gives

$$Ar^t(c_2 r^2 + c_1 r + c_0) = 0.$$

This is satisfied trivially if $A = 0$ or if r solves the quadratic equation

$$c_2 r^2 + c_1 r + c_0 = 0. \tag{2.5.5}$$

This is called the *auxiliary equation*.

Suppose (2.5.5) has two distinct real roots, $r = r_1$ and $r = r_2$, then the homogeneous equation has the solution

$$y_t = Ar_1^t + Br_2^t \tag{2.5.6}$$

for some constants A and B. These will be determined by the initial conditions.

Consider the equation due to Fibonacci for the growth of a rabbit population. He stated that the size of the population in terms of reproducing pairs at generation t is the sum of the sizes of the last two generations, that is,

$$y_t = y_{t-1} + y_{t-2}, \quad t = 3, 4, \ldots, \tag{2.5.7}$$

or equivalently,

$$y_{t+2} = y_{t+1} + y_t, \qquad t = 1, 2, \ldots.$$

Starting with one (juvenile) pair, after one breeding period these become adults, so there is still one pair. But in the next breeding period they produce one new juvenile pair, so now there are two pairs of rabbits. In general, the population sequence according to (2.5.7) is

$$1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots.$$

To find a closed-form solution, we use the method above. Transpose the terms on the right side of the equal sign to the left. That leads us to solve the quadratic equation

$$r^2 - r - 1 = 0.$$

From the quadratic formula, the roots are

$$r = \frac{1}{2}\left(1 \pm \sqrt{1-(-4)}\right),$$

and so the solution is

$$y_t = A\left(\frac{1+\sqrt{5}}{2}\right)^t + B\left(\frac{1-\sqrt{5}}{2}\right)^t. \tag{2.5.8}$$

Using the starting values $y_1 = y_2 = 1$ as above, substitute into (2.5.8), first with $t = 1$ and then with $t = 2$:

$$1 = A\left(\frac{1+\sqrt{5}}{2}\right) + B\left(\frac{1-\sqrt{5}}{2}\right),$$
$$1 = B\left(\frac{1+\sqrt{5}}{2}\right)^2 + B\left(\frac{1-\sqrt{5}}{2}\right)^2.$$

Finally, solve this system of two equations in two unknowns (using Code 2.5.1, for example) to get $A = \frac{1}{\sqrt{5}}$ and $B = -\frac{1}{\sqrt{5}}$.

Code 2.5.1.

MAPLE

```
> eq1:=1=A*((1+sqrt(5))/2)+B*((1-sqrt(5))/2);
> eq2:=1=A*((1+sqrt(5))/2)^2+B*((1-sqrt(5))/2)^2;
> solve({eq1,eq2},{A,B});
```

MATLAB

```
> M=[(1+sqrt(5))/2 (1-sqrt(5))/2; ((1+sqrt(5))/2)^2 ((1-sqrt(5))/2)^2]
> b=[1;1]
> sol=M\b
```

Hence

$$y_t = \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^t - \left(\frac{1-\sqrt{5}}{2}\right)^t\right). \tag{2.5.9}$$

What happens to y_t as $t \to \infty$? Since $\frac{1-\sqrt{5}}{2} = -0.618\ldots$ is less than 1 in absolute value, this quantity raised to the tth power tends to 0 as $t \to \infty$. Therefore,

$$y_t \approx \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^t$$

for large t. In fact, rounding this approximation to the nearest integer is exact for all t.

If the roots of the auxiliary equation are repeated, say $r = r_1$ with multiplicity 2, then one must use a solution of the form

$$y_t = Ar_1^t + Btr_1^t$$

instead of (2.5.6). As before, use the starting values to find the constants A and B.

Systems of equations lead to a higher-order single-variable equation.

Consider the following system of two recurrence relations:

$$x_{t+1} = c_{11}x_t + c_{12}y_t, \tag{2.5.10a}$$
$$y_{t+1} = c_{21}x_t + c_{22}y_t. \tag{2.5.10b}$$

The first may be written as

$$x_{t+2} = c_{11}x_{t+1} + c_{12}y_{t+1}.$$

Now the second may be substituted into this to give

$$x_{t+2} = c_{11}x_{t+1} + c_{12}(c_{21}x_t + c_{22}y_t).$$

Finally, use (2.5.10a) to eliminate y_t from this equation:

$$\begin{aligned} x_{t+2} &= c_{11}x_{t+1} + c_{12}c_{21}x_t + c_{22}(x_{t+1} - c_{11}x_t) \\ &= (c_{11} + c_{22})x_{t+1} - (c_{11}c_{22} - c_{12}c_{21})x_t. \end{aligned}$$

Chaos

Consider the *logistic recurrence relation*

$$y_{t+1} = \lambda y_t(1 - y_t), \tag{2.5.11}$$

where λ is a constant. This equation arises in the study of population growth. For values of λ less than 3, this equation converges to a unique asymptotic value. But if λ is greater than 3, strange behavior is exhibited. For example, if λ is 4 or greater, the y_ts are seemingly random values. More precisely, this is called *chaos* rather than random because the values are correlated; truly random values must be uncorrelated. For $3 \le \lambda < 1 + \sqrt{6}$, the y_ts asymptotically oscillate between two values, called a 2-*cycle*. For values of λ between $1 + \sqrt{6}$ and 4, cycles of various periods are encountered. The following code produces fully chaotic behavior:

MAPLE
```
> lam:=4:
> chaos:=proc() global y;
> y:= lam*y*(1-y);
> RETURN(y);
> end:
> y:=.05:
> for i from 1 to 24 do chaos();
> od;
```

MATLAB
```
> lam=4; y=0.05; for i=1:24 y=lam*y*(1-y)
> end
```

2.6 Matrix Analysis

The easiest kind of matrix to understand and with which to calculate is a diagonal matrix J, that is, one whose ikth term is zero, $j_{ik} = 0$, unless $i = k$. The product of

two diagonal matrices is again diagonal. The diagonal terms of the product are just the products of the diagonal terms of the factors. This pattern extends to all powers, J^r, as well. As a consequence, the exponential of a diagonal matrix is just the matrix of exponentials of the diagonal terms.

It might seem that diagonal matrices are rare, but the truth is quite to the contrary. For most problems involving a matrix, say A, there is a change of basis matrix *P such that PAP^{-1} is diagonal. We exploit this simplification to make predictions about the asymptotic behavior of solutions of differential equations.*

Eigenvalues predict the asymptotic behavior of matrix models.

Every $n \times n$ matrix A has associated with it a unique set of n complex numbers, $\lambda_1, \lambda_2, \ldots, \lambda_n$, called *eigenvalues*. Repetitions are possible, so the eigenvalues for A might not be distinct, but even with repetitions, there are always exactly n in number. In turn, each eigenvalue λ has associated with it a nonunique vector $\mathbf{e}$ called an *eigenvector*. An eigenvalue–eigenvector pair $\lambda, \mathbf{e}$ is defined by the matrix equation

$$A\mathbf{e} = \lambda\mathbf{e}. \tag{2.6.1}$$

An eigenvector for λ such as $\mathbf{e}$ is not unique, because for every number a, the vector $\mathbf{e}' = a\mathbf{e}$ is also an eigenvector, as is easily seen from (2.6.1).

Example 2.6.1. The matrix

$$A = \begin{bmatrix} 1 & 3 \\ 0 & -2 \end{bmatrix}$$

has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = -2$ with corresponding eigenvectors $\mathbf{e}_1 = \binom{1}{0}$ and $\mathbf{e}_2 = \binom{1}{-1}$. Before invoking the computer on this one (see Exercise 1 in this section), work through it by hand.

Eigenvalues and eigenvectors play a central role in every mathematical model embracing matrices.

This statement cannot be overemphasized. The reason is largely a consequence of the following theorem.

Theorem 1. *Let the $n \times n$ matrix A have n distinct eigenvalues; then there exists a nonsingular matrix P such that the matrix*

$$J = PAP^{-1} \tag{2.6.2}$$

is the diagonal matrix of the eigenvalues of A,

$$J = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix}.$$

The columns of P are the eigenvectors of A taken in the same order as the list of eigenvalues.

If the eigenvalues are not distinct, then we are not guaranteed that there will be a completely diagonal form; it can happen that there is not one. But even if not, there is an almost diagonal form, called the *Jordan canonical form* (or just Jordan form), which has a pattern of 1s above the main diagonal. By calculating the Jordan form of a matrix, we get the diagonal form if the matrix has one. We will not need to discuss Jordan form here, except to say that the computer algebra system can compute it.

The matrix product of this theorem, PAP^{-1}, is a change of basis modification of A; in other words, by using the eigenvectors as the reference system, the matrix A becomes the diagonal matrix J. Note that if $J = PAP^{-1}$, then the kth power of J and A are related as the k-fold product of PAP^{-1},

$$J^k = (PAP^{-1})(PAP^{-1})\cdots(PAP^{-1}) = PA^kP^{-1}, \tag{2.6.3}$$

since the interior multiplications cancel.

Diagonal matrices are especially easy to work with; for example, to raise J to a power J^k becomes raising the diagonal entries to that power:

$$J^k = \begin{bmatrix} \lambda_1^k & 0 & \dots & 0 \\ 0 & \lambda_2^k & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \lambda_n^k \end{bmatrix}.$$

As a result, the exponential of J is just the exponential of the diagonal entries. From (2.4.14),

$$\begin{aligned} e^{Jt} &= I + Jt + \frac{J^2t^2}{2!} + \frac{J^3t^3}{3!} + \cdots \\ &= \begin{bmatrix} \left(1 + \lambda_1 t + \frac{\lambda_1^2 t^2}{2!} + \cdots\right) & 0 & \dots & 0 \\ 0 & \left(1 + \lambda_2 t + \frac{\lambda_2^2 t^2}{2!} + \cdots\right) & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \left(1 + \lambda_n t + \frac{\lambda_n^2 t^2}{2!} + \cdots\right) \end{bmatrix} \\ &= \begin{bmatrix} e^{\lambda_1 t} & 0 & \dots & 0 \\ 0 & e^{\lambda_2 t} & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & e^{\lambda_n t} \end{bmatrix}. \end{aligned} \tag{2.6.4}$$

We illustrate the way in which these results are used.

The age structure of a population can be modeled so that it evolves as dictated by a matrix L, such as the following (see Chapter 5):

$$L = \begin{bmatrix} 0 & 0 & 0 & 0 & 0.08 & 0.28 & 0.42 \\ .657 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & .930 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .930 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .930 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .935 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .935 & 0 \end{bmatrix}.$$

After k generations, the pertinent matrix is the kth power of L. From the theorem, there exists a matrix P such that $J = PLP^{-1}$, and according to (2.6.3),

$$L^k = P^{-1}J^kP.$$

Letting λ_1 be the largest eigenvalue of L in absolute value, it is easy to see that

$$\frac{1}{\lambda_1^k}J^k = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & \left(\frac{\lambda_2}{\lambda_1}\right)^k & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \left(\frac{\lambda_n}{\lambda_1}\right)^k \end{bmatrix}$$

$$\longrightarrow \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \quad \text{as } k \to \infty.$$

In other words, for large k, L^k is approximately λ_1^k times a fairly simple fixed matrix related to its eigenvectors; thus it grows or decays like λ_1^k.

In another example, consider the matrix form of the linear differential equation (2.4.11) of Section 2.4. From above, the matrix exponential e^{Mt} can be written as

$$e^{Mt} = P^{-1}e^{Jt}P,$$

where e^{Jt} consists of exponential functions of the eigenvalues. If all those eigenvalues are negative, then no matter what P is, every solution will tend to 0 as $t \to \infty$. But if one or more eigenvalues are positive, then at least one component of a solution will tend to infinity.

In Chapter 9, we will consider compartment models. A *compartment* matrix C is defined as one whose terms c_{ij} satisfy the following conditions:

1. All diagonal terms c_{ii} are negative or zero.
2. All other terms are positive or zero.
3. All column sums $\sum_i c_{ij}$ are negative or zero.

Under these conditions, it can be shown that the eigenvalues of C have negative or zero real parts and so the asymptotic result above applies.

The fact that the eigenvalues have negative real parts under the conditions of a compartment matrix derives from *Gershgorin's circle theorem*.

Theorem 2. *If A is a matrix and S is the following union of circles in the complex plane,*

$$S = \bigcup_m \left\{ \text{complex } z : |a_{mm} - z| \le \sum_{j \ne m} |a_{jm}| \right\},$$

then every eigenvalue of A lies in S.

Notice that the mth circle above has center a_{mm} and radius equal to the sum of the absolute values of the other terms of the mth column.

Exercises

1. For both the following matrices A, find the eigenvalues and eigenvectors. Then find the Jordan form. Plot solutions $[Z_1, Z_2, Z_3]$ for $Z' = AZ$. Note that the Jordan structure for the two is different:

$$A_1 = \begin{pmatrix} 0 & 0 & -2 \\ 1 & 2 & 1 \\ 1 & 0 & 3 \end{pmatrix} \quad \text{and} \quad A_2 = \begin{pmatrix} 3 & 1 & -1 \\ -1 & 2 & 1 \\ 2 & 1 & 0 \end{pmatrix}.$$

Here is the syntax for A_1. Define the following matrix:

MAPLE
```
> restart;
> with(LinearAlgebra):
> A:=Matrix([[0,0,-2],[1,2,1],[1,0,3]]);
```

MATLAB
```
> A=[0 0 -2; 1 2 1; 1 0 3]
```

(a) Find the eigenvalues and eigenvectors of A. (Note that both $x_1 = (-1, 0, 1)^t$ and $x_2 = (0, 1, 0)^t$ are eigenvectors for the eigenvalue 2; therefore, so is every linear combination $ax_1 + bx_2$.)

MAPLE
```
> ev:=Eigenvectors(A);
  # first column = eigenvalues, second "column" = matrix whose columns are eigenvectors
> evals:=ev[1]; evects:=ev[2];
  # evects[1] is a row, we want the column; transpose
> whattype(Transpose(evects)[1]); # a row vector, needs to be a colmn vector
> x1:=convert(Transpose(evects)[1],Vector[column]);
> x2:=convert(Transpose(evects)[2],Vector[column]);
> x3:=convert(Transpose(evects)[3],Vector[column]);
```

MATLAB
```
> [evect, eval]=eig(A)
  % evect is P inverse and eval is J
```

(b) Find the Jordan form and verify that the associated matrix P has the property that

$$PAP^{-1} = J.$$

MAPLE (symbolic derivative)
```
> J:=JordanForm(A);
> Q:=JordanForm(A, output='Q');
> Q^(-1).A.Q;
```

MATLAB
```
> P=inv(evect)
> J=P*A*inv(P)
```
In order to get (2.6.2), take P to be Q^{-1}.

MAPLE
```
> P:=Q^(-1); P.A.P^(-1);
```

2. In a compartment matrix, one or more of the column sums may be zero. In this case, one eigenvalue can be zero and solutions for the differential equations

$$Z' = CZ(t)$$

may have a limit different from zero.

If all the column sums are negative in a compartment matrix, the eigenvalues will have negative real part. All solutions for the differential equations

$$Z' = CZ(t)$$

will have limit zero in this case.

The following matrices contrast these two cases:

$$C_1 = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -1 & 0 & \frac{1}{2} \\ \frac{1}{2} & -1 & 0 \\ 0 & \frac{1}{2} & -1 \end{pmatrix}.$$

Let C be the matrix defined below:

MAPLE
```
> with(LinearAlgebra):
  C:=Matrix([[-1,1,0],[1,-1,0],[0,0,-1]]);
```

MATLAB
```
> C=[-1 1 0; 1 -1 0; 0 0 -1]
```

(a) Find the eigenvalues and eigenvectors for C.

MAPLE
```
> Eigenvectors(C);
```

MATLAB
```
> [evects, evals] = eig(C)
```

(b) Graph each component of z with $z(0) = [1, 1, 1]$.

MAPLE
```
> exptC:=MatrixExponential(C,t);
> U:=evalm( exptC.[1,0,1]);
> u:=unapply(U[1],t); v:=unapply(U[2],t); w:=unapply(U[3],t);
> plot({u(t),v(t),w(t)},t=0..2, color=[black,blue,green]);
```

MATLAB
```
  % contents of the m-file exer252.m
  % function Zprime=exer252(t,Z);
  % Zprime=[-1*Z(1)+1*Z(2); 1*Z(1)-1*Z(2); -1*Z(3)];
> [t,Z]=ode23('exer252',[0 10],[1; 0; 1]);
> plot(t,Z)
```

2.7 Statistical Data

Variation impacts almost everything. Variation can be quantified by describing its distribution. A distribution *is the set of the fractions of observations having particular values with respect to the number of the possible values. For example, the distribution of word lengths of the previous sentence is* 3 *of length* 1, 4 *of length* 2, 2 *of length* 3, *and so on* (*all divided by* 18, *the number of words in the sentence*). *The graph of a distribution with the observations grouped or made discrete to some resolution is a histogram. Distributions are approximately described by their mean, or average, value and the degree to which the observations deviate from the mean, their standard deviation. A widely occurring distribution is the normal, or Gaussian. This bell-shaped distribution is completely determined by its mean and standard deviation.*

Histograms portray statistical data.

Given that the natural world is rife with variables, it is not surprising to find that variation is widespread. Trees have different heights, ocean temperatures change from place to place and from top to bottom, the individuals of a population have different ages, and so on. Natural selection thrives on variation. Variation is often due to chance events; thus the height of a tree depends on its genetic makeup, the soil in which it grows, rainfall, and sunlight among other things. Describing variation is a science all to its own.

Since pictures are worth many words, we start with histograms. Corresponding to the phenomenon under study, any variation observed occurs within a specific range of possibilities, a *sample space*. This range of possibilities is then partitioned or divided up into a number of subranges, or classes. A *histogram* is a graph of the fraction of observations falling within the various subranges plotted against those subranges.

Consider the recent age distribution data for the U.S. population, shown in Table 2.7.1. The possible range of ages, 0 to infinity, is partitioned into subranges or intervals of every five years from birth to age 84; a last interval, 85+, could be added if necessary for completeness. The table lists the percentage of the total population falling within the given interval; each percentage is also refined by sex. The cumulative percentage is also given, that is, the sum of the percentages up to and including the given interval. A histogram is a graph of these data; on each partition interval is placed a rectangle, or bar, whose width is that of the interval and whose height is the corresponding percentage (see Figure 2.7.1).

The resolution of a histogram is determined by the choice of subranges: Smaller and more numerous intervals mean better resolution and more accurate determination of the distribution; larger and fewer intervals entail less data storage and processing.

The cumulative values are plotted in Figure 2.7.2. Since the percentage values have a resolution of five years, a decision has to be made about where the increments should appear in the cumulative plot. For example, 7.2% of the population is in the first age interval counting those who have not yet reached their fifth birthday. Should this increment be placed at age 0, at age 5, or maybe at age 2.5 in the cumulative graph?

Table 2.7.1. Age distribution for the U.S. population.

Age	% Female	% Male	% Population	Cumulative
0–4	3.6	3.6	7.2	7.2
5–9	3.9	3.7	7.6	14.8
10–14	4.1	3.9	8.0	22.8
15–19	4.7	4.3	9.0	31.8
20–24	5.0	4.2	9.2	41.0
25–29	4.3	4.0	8.3	49.3
30–34	4.0	3.5	7.5	56.8
35–39	3.6	2.9	6.5	63.3
40–44	2.7	2.2	4.9	68.2
45–49	2.8	2.0	4.8	73.0
50–54	3.0	2.2	5.2	78.2
55–59	3.1	2.1	5.2	83.4
60–64	2.8	1.9	4.7	88.1
65–69	2.3	1.8	4.1	92.2
70–74	2.0	1.4	3.4	95.6
75–79	1.7	0.8	2.5	98.1
80–84	1.6	0.3	1.9	100

We have chosen to do something different, namely, to indicate this information as a line segment that is 0 at age 0 and is 7.2 at age 5. In like fashion, we indicate in the cumulative graph the second bar of the histogram of height 7.6% as a line segment joining the points 7.2 at age 5 with 14.8 ($= 7.2 + 7.6$) at age 10. Continuing this idea for the balance of the data produces the figure. Our rationale here is the assumption that the people within any age group are approximately evenly distributed by age in this group. A graph that consists of joined line segments is called a *polygonal graph* or a *linear spline*.

This graph of accumulated percentages is called the *cumulative distribution function*, or *cdf* for short. No matter what decision is made about placing the cumulative percentages, the cdf satisfies these properties:

1. it starts at 0,
2. it never decreases, and
3. it eventually reaches 1 (or, as a percentage, 100%).

The mean and median approximately locate the center of the distribution.

Sometimes it is convenient to summarize the information in a histogram. Of course, no single number or pair of numbers can convey all the information; such a summary is therefore a compromise, but nevertheless a useful one. First, some information about where the data lie is given by the *mean*, or *average*; it is frequently denoted by μ. Given the n values $x_1, x_2, \ldots, x_n$, their mean is

MAPLE

```
> mcent:=[3.6, 3.7, 3.9, 4.3, 4.2, 4.0, 3.5, 2.9, 2.2,2.0, 2.2, 2.1, 1.9, 1.8,1.4, 0.8, 0.3]:
  fcent:=[3.6, 3.9, 4.1, 4.7, 5.0, 4.3, 4.0, 3.6, 2.7, 2.8, 3.0, 3.1, 2.8, 2.3, 2.0, 1.7, 1.6]:
  tot:=[seq(mcent[i]+fcent[i],i=1..17)]:
> ranges:=[0..5, 5..10, 10..15, 15..20, 20..25, 25..30, 30..35, 35..40, 40..45, 45..50, 50..55, 55..60, 60..65,
          65..70, 70..75, 75..80, 80..85]:
> with(stats): with(plots):
> mpop:=[seq(Weight(ranges[i], 5*mcent[i]),i=1..17)]:
> fpop:=[seq(Weight(ranges[i], 5*fcent[i]),i=1..17)]:
> pop:=[seq(Weight(ranges[i], 5*tot[i]),i=1..17)]:
> statplots[histogram](pop);
```

MATLAB

```
> mcent=[3.6 3.7 3.9 4.3 4.2 4.0 3.5 2.9 2.2 2.0 2.2 2.1 1.9 1.8 1.4 0.8 0.3];
> fcent=[3.6 3.9 4.1 4.7 5.0 4.3 4.0 3.6 2.7 2.8 3.0 3.1 2.8 2.3 2.0 1.7 1.6];
> total=mcent+fcent;
> x=[5:5:85]; % 5, 10, 15, …, 85
> bar(x,total) % bars centered on the x values
> xlabel('Age(years)')
> ylabel('Percent in age bracket');
```

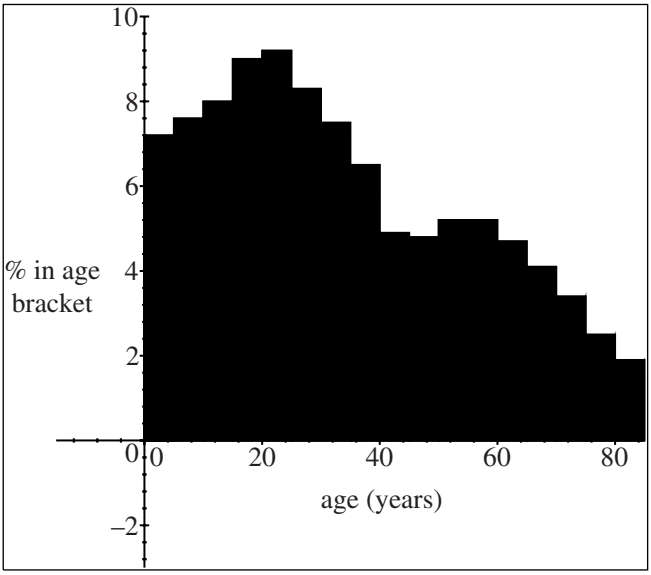


Fig. 2.7.1. Histogram for the U.S. population distributed by age.

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i . \tag{2.7.1}$$

Another popular notation for this quotient is $\bar{x}$. It is necessarily true that some values x_i are smaller than the mean and some are larger. (Either that or all xs are equal.) In fact, one understands the mean to be in the center of the x values in a sense made precise by (2.7.1). Given $\bar{x}$ and n, the sum of the xs is easily computed:

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

MAPLE

```
> age:=[2.5, 7.5, 12.5, 17.5, 22.5, 27.5, 32.5, 37.5, 42.5, 47.5, 52.5, 57.5, 62.5, 67.5, 72.5, 77.5, 82.5];
> cummale:=[seq(sum('mcent[i]','i'=1..n),n=1..17)]:
> cumfale:=[seq(sum('fcent[i]','i'=1..n),n=1..17)]:
> cumtot:=[seq(sum('tot[i]','i'=1..n),n=1..17)]:
> ptsm:=[seq([age[i],cummale[i]],i=1..17)];
> ptsf:=[seq([age[i],cumfale[i]],i=1..17)]:
> ptsT:=[seq([age[i],cumtot[i]],i=1..17)]:
> plot({ptsm,ptsf,ptsT},color=BLACK);
```

MATLAB

```
> cumM=cumsum(mcent);
> cumF=cumsum(fcent);
> cumTot=cumsum(total)
> plot(x,cumTot,x,cumM,x,cumF)
```

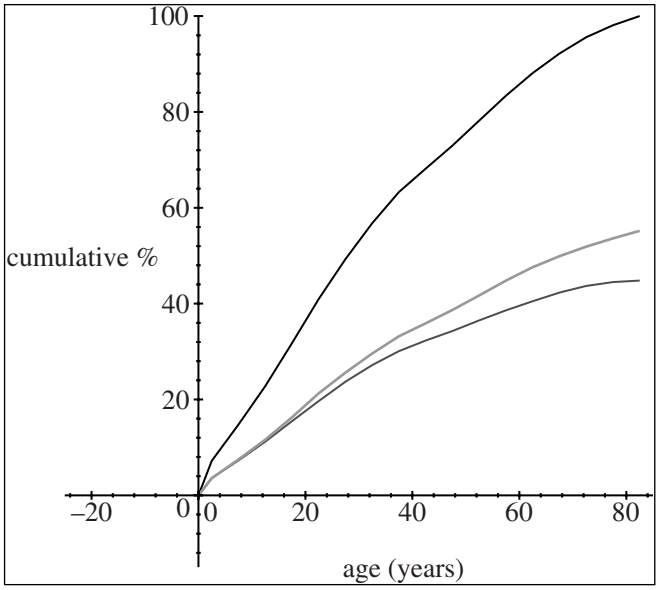


Fig. 2.7.2. Cumulative populations (% of the total vs. age).

Computing the mean of a histogram goes somewhat differently. Suppose the total number of people referred to by Table 2.7.1 to be 100 million. (It no doubt corresponds to many more than that, but it will be more convenient to calculate percentages using 100 million, and we will see that in the end, this choice is irrelevant.) Then the 7.2% in the first group translates into 7.2 million people. We do not know their individual ages, but as above, if they were evenly distributed over ages 0 to 4.999 . . . , then counting all 7.2 million as 2.5 gives the same result. Hence in (2.7.1) these people contribute a value of 2.5 for 7.2 million such people, or

$$\text{contribution of “0 to 5” group} = 2.5 \cdot 7.2 = \frac{0+5}{2} \cdot 7.2$$

in millions. Similarly the second group contributes

$$\text{contribution of "5 to 10" group} = 7.5 \cdot 7.8 = \frac{5+10}{2} \cdot 7.8.$$

Continuing in this way we get, where we are counting in millions,

$$\sum_{i=1}^{n} x_i = 2.5 \cdot 7.2 + 7.5 \cdot 7.6 + 12.5 \cdot 8.0 + \cdots + 82.5 \cdot 1.9 = 3431.0 \text{ (million)}.$$

Divide the result by 100 (million) to obtain the mean. But dividing by 100 million means a quotient such as $\frac{7.2 \text{ million}}{100 \text{ million}}$ is just the fraction .072 (or 7.2%). In other words, we do not need to know the total population size; instead, we just use the fractions, such as .072, as multipliers or weights for their corresponding interval. Completing the calculation, then, we have

$$\bar{x} = 2.5 \cdot 0.072 + 7.5 \cdot 0.076 + \cdots + 82.5 \cdot 0.019 = 34.31. \tag{2.7.2}$$

Equation (2.7.2) illustrates a general principle for calculating the mean. It applies to (2.7.1) as well:

$$\mu = \sum_{\substack{\text{over possible} \\ \text{values } x}} x \cdot \text{fraction of values equal to } x. \tag{2.7.3}$$

In (2.7.2) the possible xs are 2.5, 7.5, and so on, while the fractions are .072, .076, and so on. In (2.7.1) the possible xs are x_1, x_2, and so on, while the fraction of values that are x_1 is just 1 out of n, that is, $\frac{1}{n}$, and similarly for the other x_is.

The *median* is an alternative to the mean for characterizing the center of a distribution. The median, $\hat{x}$, of a set of values $x_1, x_2, \ldots, x_n$ is such that one-half the values are less than or equal to $\hat{x}$ and one-half are greater than or equal to it. If n is odd, $\hat{x}$ will be one of the xs. If n is even, then $\hat{x}$ should be taken as the average of the middle two x values. For example, the median of the values 1, 3, 6, 7, and 15 is $\hat{x} = 6$, while the median of 1, 3, 6, and 7 is $\frac{3+6}{2} = 4.5$.

The median is sometimes preferable to the mean because it is a more typical value. For example, for the values 3, 3, 3, 3, and 1000, the mean is 506, while the median is 3.

In the population data, the median age for men and women is between 29 and 30. This can be seen from an examination of the last column of Table (2.7.1). Contrast this median age with the average age; thus for men,

$$\begin{aligned}\text{average age for men} &= \frac{\sum_{n=1}^{17} [\text{percentage men at age } n] \cdot [\text{age } [n]]}{\text{total percentage of men}} \\ &= 32.17.\end{aligned}$$

In a similar manner, the average age for women in this data set is about 35.5, and the average age for the total population is about 33.8. The averages for these three sets of data—male population age distribution, female population age distribution, and total population age distribution—can be found with simple computer algebra commands and agree with our paper-and-pen calculations.

```
MAPLE
> Sum('age[j]'*'tot[j]',j=1..17)=sum(age[j]*tot[j],j=1..17);
> Sum('mcent[n]*age[n]','n'=1..17)/Sum('mcent[n]','n'=1..17)
  =sum('mcent[n]*age[n]','n'=1..17)/sum('mcent[n]','n'=1..17);
> with(describe): mean(pop); median(pop);

MATLAB
> xmid=[2.5:5:82.5];
> pop=xmid.*total; % term by term mult. = percentage weighted ranges
> muTotal=sum(pop)/100 % divide by 100 as data is in percent
> muM=sum(xmid.*mcent)/sum(mcent)
> muF=sum(xmid.*fcent)/sum(fcent)
```

Variance and standard deviation measure dispersion.

As mentioned above, a single number will not be able to capture all the information in a histogram. The data set 60, 60, 60, 60 has a mean of 60, as does the data set 30, 0, 120, 90. If these data referred to possible speeds in miles per hour for a trip across Nevada by bus for two different bus companies, then we might prefer our chances with the first company. The *variance* of a data set measures how widely the data is dispersed from the mean; for n values $x_1, x_2, \ldots, x_n$, their variance v, or sometimes σ^2, is defined as

$$v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{2.7.4}$$

where $\bar{x}$ is the mean as before.[7] Thus the speed variance for bus company 1 is 0 and that for bus company 2 is

$$\frac{1}{4}[(30-60)^2 + (0-60)^2 + (120-60)^2 + (90-60)^2] = 2{,}250.$$

As before, a more general equation for variance, one suitable for histograms, for example, is the following:

$$v = \sum_{\substack{\text{over possible}\\ \text{values } x}} (x - \bar{x})^2 \cdot \text{fraction of values equal to } x. \tag{2.7.5}$$

A problem with variance is that it corresponds to squared data values, making it hard to interpret its meaning in terms of the original data. If the data has units, like miles per hour, then variance is in the square of those units. Closely related to variance is *standard deviation*, denoted by σ. Standard deviation is defined as the square root of variance,

$$\text{standard deviation} = \sqrt{\text{variance}}.$$

[7] For data representing a sample drawn from some distribution, $\bar{x}$ is only an estimate of the distribution's mean, and for that reason, this definition of variance is a biased estimator of the distribution's variance. Divide by $n-1$ in place of n for an unbiased estimator. Our definition is, however, the maximum likelihood estimator of the variance for normal distributions. Furthermore, this definition is consistent with the definition of variance for probability distributions (see Section 2.8), and for that reason we prefer it.

Standard deviation is a measure of the dispersion of data on the same scale as the data itself. The standard deviation of bus speeds for company 2 is 47.4 miles per hour. This is not saying that the average (unsigned) deviation of the data from the mean is 47.4 (for that would be $\frac{1}{n}\sum_1^n |x_i - \bar{x}| = 45$), but this is, in spirit, what the standard deviation measures. For the bus companies, we make these calculations:

MAPLE
```
> bus1:=[60,60,60,60]; bus2:=[30,0,120,90];
> range(bus1), range(bus2);
> median(bus1), median(bus2);
> mean(bus1), mean(bus2);
> variance(bus1), variance(bus2);
> standarddeviation(bus1), standarddeviation(bus2);
```

MATLAB
```
> bus1=[60 60 60 60]; bus2=[30 0 120 90];
> max(bus1), min(bus1)
> max(bus2), min(bus2)
> median(bus1), median(bus2)
> mean(bus1), mean(bus2)
> cov(bus1), cov(bus2)
> std(bus1), std(bus2)
```

We can perform similar calculations for the U.S. census data of Table 2.7.1. The results are given in Table 2.7.2.

MAPLE
```
> range(mpop), range(fpop), range(pop);
> median(mpop), median(fpop), median(pop);
> mean(mpop), mean(fpop), mean(pop);
> variance(mpop), variance(fpop), variance(pop);
> standarddeviation(mpop), standarddeviation(fpop),
> standarddeviation(pop);
```

MATLAB
```
> v=(xmid-muTotal).^2 % unweighted vector of deviations squared
> var=sum(v.*total)/100 % variance of total population
> sqrt(var) % std dev of the total population
```

Table 2.7.2. Summary for the U.S. age distribution

	Range	**Median**	**Mean**	**Standard deviation**
Male	0–84	29	31.7	21.16
Female	0–84	30	35.6	22.68
Total	0–84	29	33.8	22.10

The normal distribution is everywhere.

It is well known that histograms are often bell-shaped. This is especially true in the biological sciences. The mathematician Carl Friedrich Gauss discovered the explanation for this, and it is now known as the *central limit theorem* (see Hogg and Craig [6]).

Central limit theorem. *The accumulated result of many independent random outcomes, in the limit, tends to a* Gaussian, *or* normal, *distribution given by*

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \quad -\infty < x < \infty,$$

where μ *and* σ *are the* mean *and* standard deviation *of the distribution.*

The normal distribution is a continuous distribution, meaning that its resolution is infinitely fine; its histogram, given by $G(x)$, is smooth (see Figure 2.7.3). The two parameters mean μ and standard deviation σ completely determine the normal distribution. Likewise, even though a given histogram is not Gaussian, nevertheless its description is often given in terms of just its mean and variance or standard deviation.

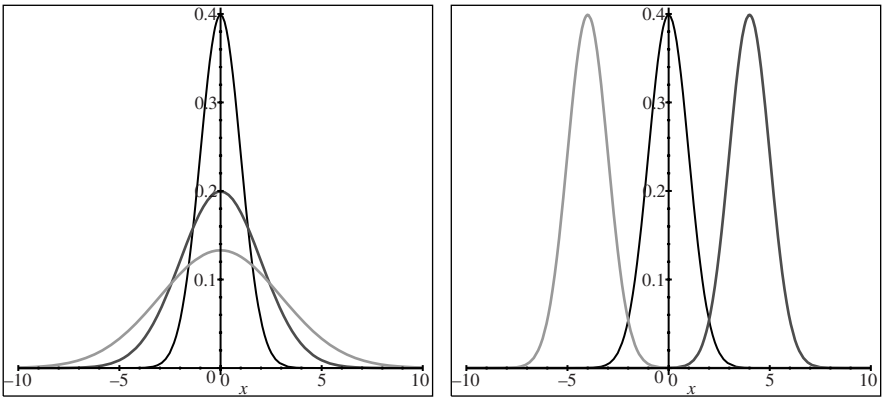
In Figure 2.7.3(a), we show three curves with the same mean but different standard deviations. In Figure 2.7.3(b), the three curves have the same standard deviation but different means.

MAPLE

```
> y:=(sigma,mu,x)->exp(-(x-mu)^2/(2*sigma^2))/(sqrt(2*Pi)*sigma);
> plot({y(1,0,x),y(2,0,x),y(3,0,x)},x=-10..10);
> plot({y(1,-4,x),y(1,0,x),y(1,4,x)},x=-10..10);
```

MATLAB

```
% make up an m-file, gaussian.m:
% function y=gaussian(x,m,s);
  %% m=mean, s=stddev
  %% note 1/sqrt(2*pi)=.3989422803
  % y=(.3989422803/s)*exp(-0.5*((x-m)./s).^2);
> x=[-10:.1:10];
> y=gaussian(x,0,1); plot(x,y);hold on;
> y=gaussian(x,0,2); plot(x,y);
> y=gaussian(x,0,4); plot(x,y);
> hold off
> y=gaussian(x,0,1); plot(x,y);hold on
> y=gaussian(x,-5,1); plot(x,y);
> y=gaussian(x,5,1); plot(x,y);
```



(a) Various σ. (b) Various μ.

Fig. 2.7.3.

Exercises

1. In the February 1994 Epidemiology Report published by the Alabama Department of Public Health, the data in Table 2.7.3 were provided as Age-Specific Mortality. Make a histogram for these data. While the data are given over age ranges, get a fit for the data so that one could predict the death rate for intermediate years. Find the median, mean, and standard deviation for the data.

Table 2.7.3.

0–1	1122.4	40–45	287.8
1–5	55.1	45–50	487.2
5–10	27.5	50–55	711.2
10–15	33.4	55–60	1116.9
15–20	118.4	60–65	1685.1
20–25	139.6	65–70	2435.5
25–30	158.0	70–75	3632.4
30–35	196.4	75–80	5300.0
35–40	231.0	80–85	8142.0
		85+	15279.0

MAPLE

```
> with(stats): with (plots): with(describe):
> Mort:=[1122.4, 55.1, 27.5, 33.4, 118.4, 139.6, 158.0, 196.4, 231.0, 287.8, 487.2, 711.2, 1116.9,
          1685.1, 2435.5, 3632.4, 5300.0, 8142.0, 15278.0]:
> MortRate:=[seq(Mort[i]/100000,i=1..19)];
> ranges:=[seq(5*i..5*(i+1),i=1..17)];
> mortdata:=[Weight(0..1,MortRate[1]), Weight(1..5,4*MortRate[2]),
              seq(Weight(ranges[i],5*MortRate[2+i]), i=1..17)]:
> statplots[histogram](mortdata);
```

MATLAB

```
> Mort=[1122.4, 55.1, 27.5, 33.4, 118.4, 139.6, 158.0, 196.4, 231.0, 287.8, 487.2, 711.2, 1116.9, …
        1685.1, 2435.5, 3632.4, 5300.0, 8142.0, 15278.0];
> MortRate=Mort/1000;
> x=[.5,2.5:5:87.5];
> bar(x,MortRate)
> x=x(2:19) % first point an outlier
> MortRate=MortRate(2:19) % ditto
```

(a) A polynomial fit:

MAPLE

```
> xcord:=[seq(3+5*(i-1),i=1..18)];
> mortrate:=[seq(MortRate[i+1], i=1..18)];
> plot([seq([xcord[i],mortrate[i]],i=1..18)], style=POINT, symbol=CROSS);
> fit[leastsquare[[x,y],y=a+b*x+c*x^2+d*x^3]]([xcord,mortrate]);
> approx:=unapply(rhs(%),x);approx(30)*100000;
> plot(approx(x),x=0..90);
```

MATLAB

```
  % cubic fit rate = d*x^3+c*x^2+b*x+a
> p=polyfit(x,MortRate,3) % use built-in polynomial fitter, third order
> y=polyval(p,x); % fit evaluated at the xs
> plot(x,MortRate,'x'); hold on
> plot(x,y)
  % or use the general leastsquares model
```

```
> MT=[x.^3; x.^2; x; ones(size(x))];
> cubic=MT'\MortRate'
> y=polyval(cubic,x); plot(x,y)
```

(b) An exponential fit:

```
  MAPLE
> Lnmortrate:=map(ln,mortrate);
> fit[leastsquare[[x,y],y=m*x+b]]([xcord,Lnmortrate]);
> k:=op(1,op(1,rhs(%))); A:=op(2,rhs(%%));
> expfit:=t->exp(A)*exp(k*t); expfit(30)*100000;
> J:=plot(expfit(t),t=0..85):
  K:=plot([seq([xcord[i],MortRate[i+1]],i=1..18)],style=POINT,symbol=CROSS):
> display({J,K});

  MATLAB
  % exponential fit log(MortRate)=a+b*x or MortRate=exp(a)*exp(bx)
> Lnmortrate=log(MortRate);
> MT=[ones(size(x)); x];
> expon=MT'\Lnmortrate'
> hold off
> plot(x,MortRate,'x'); hold on
> plot(x,exp(expon(1))*exp(expon(2)*x))
```

(c) A linear spline for the data (see the discussion in this section):

```
  MAPLE
> readlib(spline):
> linefit:=spline(xcord,mortrate,x,linear):
> y:=unapply(linefit.x): y(30)*100000;
> J:=plot(y(t), t=0..85):
> display({J,K});

  MATLAB
  % linear spline fit = straight line between points, usual MATLAB method
> hold off
> plot(x,MortRate,x,MortRate,'x')
```

Give the range, median, mean, and standard deviation of the mortality rates. Note that the first entry is applicable to humans in an age group of width one year and the second is in a group of width four years. Each of the others applies to spans of five years. Thus we set up a weighted sum:

```
  MAPLE
> summary:=[Weight(Mort[1],1),Weight(Mort[2],4),seq(Weight(Mort[i],5),i=3..19)];
> range(summary); median(summary); mean(summary);
> standarddeviation(summary);

  MATLAB
  % to interpolate any desired value, use interp1, e.g., rate=interp1(x,MortRate,70)
  % interpolated value at x=70
  % mean, median, and standard deviation (of Mortality weighted by age)
> size(Mort)
> wt=[1,4,5*ones(1,17)]
> wtSum = Mort*wt' % dot product
> mu=wtSum/sum(wt)
> median(Mort) % picks out the middle value, no duplicates here
> v=(Mort-mu).^2; % vector of squared differences
> var=sum(v.*wt)/sum(wt);
> std=sqrt(var)
```

2. What follows in Table 2.7.4 are data for the heights of a group of males. Determine a histogram for these data. Find the range, median, mean, and standard deviation for the data. Give a normal distribution with the same mean and standard deviation as the data. Plot the data and the distribution on the same graph.

Table 2.7.4.

Number of students	2	1	2	7	10	14	7	5	2	1
Height (in)	66	67	68	69	70	71	72	73	74	75

```
  MAPLE
> with(stats): with(plots): with(describe):
> htinches:=[seq(60+i,i=1..15)];
> numMales:=[0,0,0,0,0,2,1,2,7,10,14,7,5,2,1];
> ranges:=[seq(htinches[i]..htinches[i]+1, i=1..15)];
> maledata:=[seq(Weight(ranges[i],numMales[i]), i=1..15)];
> statplots[histogram](maledata);
> range(maledata); median(maledata); mean(maledata); standarddeviation(maledata);
  # note the use of back quotes in the next for formatted printing
> `The average height is`,floor(%%/12), `feet and`,floor(frac(%%%/12)*12), `inches`;
> `The standard deviation is`,floor(frac(%%%/12)*12), `inches`;

  MATLAB
> htinches=61:75;
> numMales=[0,0,0,0,0,2,1,2,7,10,14,7,5,2,1];
> bar(htinches,numMales)
> min(htinches)
> max(htinches) % range = from min to max
> unrolled=[]; % dup. each height by its #cases
> s=size(htinches);
> for k=1:s(2)
>    j=numMales(k);
>    while j>0
>       unrolled=[unrolled, htinches(k)];
>       j=j-1;
>    end
> end
> median(unrolled)
> mu=mean(unrolled+.5) % e.g., height 66 counts as 66.5
  % alternatively
> mu=dot((htinches+.5),numMales)/sum(numMales)
> v=(htinches+.5-mu).^2;
> var=sum(v.*numMales)/sum(numMales)
> std=sqrt(var)
```

In what follows, we give a normal distribution that has the same mean and standard deviation as the height data:

```
  MAPLE
> mu:=mean(maledata);
> sigma:=standarddeviation(maledata);
> ND:=x->exp(-(x-mu)^2/(2*sigma^2))/(sigma*sqrt(2*Pi));
> J:=plot(mu*ND(x),x=60..76):
> K:=statplots[histogram](maledata):
> plots[display]({J,K});

  MATLAB
> x=60:.1:76;
> y=exp(-((x-mu)/std).^2/2)/(std*sqrt(2*pi));
> bar(htinches,numMales/sum(numMales))
> hold on; plot(x,y)
```

To the extent that the graph K is an approximation for the graph J, the heights are normally distributed about the mean.

3. Table 2.7.5 contains population data estimates for the United States (in thousands) as published by the U.S. Bureau of the Census, Population Division, release PPL-21 (1995).

Table 2.7.5.

Five-year age groups	1990	1995	Five-year age groups	1990	1995
0–5	18,849	19,662	50–55	11,368	13,525
5–10	18,062	19,081	55–60	10,473	11,020
10–15	17,189	18,863	60–65	10,619	10,065
15–20	17,749	17,883	65–70	10,077	9,929
20–25	19,133	18,043	70–75	8,022	8,816
25–30	21,232	18,990	75–80	6,145	6,637
30–35	21,907	22,012	80–85	3,934	4,424
35–40	19,975	22,166	85–90	2,049	2,300
40–45	17,790	20,072	90–95	764	982
45–50	13,820	17,190	95–100	207	257
			100+	37	52

Find the median and mean ages. Estimate the number of people at ages 21, 22, 23, 24, and 25 in 1990 and in 1995. Make a histogram for the percentages of the population in each age category for both population estimates.

4. In (2.7.3), we stated that the mean μ is defined as

$$\mu = \sum_{\text{all possible } x\text{s}} x \cdot f(x),$$

where $f(x)$ is the fraction of all values that are equal to x. If these values are spread continuously over all numbers, μ can be conceived as an integral. In this sense, this integral of the normal distribution given by (2.7.3) yields

$$\mu = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx.$$

In a similar manner,

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx.$$

Here is a way to evaluate the integrals:

MAPLE

```
> sigma:='sigma': mu='mu':
> f:=x->exp(-(x-mu)^2/(2*sigma^2))/(sigma*sqrt(2*Pi));
> assume(sigma > 0);
> int(x*f(x),x=-infinity..infinity);
```

```
> int((x-mu)^2*f(x),x=-infinity..infinity);
```

```
MATLAB
% to integrate with MATLAB one can use trapz(x,y) on x and y vectors or use Simpson's rule, quad(), ...
% but this requires an m-file.
% Here we will use the trapzoidal rule.
> x=linspace(-3,3); % simulates -infinity to +infinity here
> y=exp(-x.^2/2)/sqrt(2*pi);
> trapz(x,y) % approximately 1
```

2.8 Probability

The biosphere is a complicated place. One complication is its unpredictable events, such as when a tree will fall or exactly what the genome of an offspring will be. Probability theory deals with unpredictable events by making predictions in the form of relative frequency of outcomes. Histograms portray the distribution of these relative frequencies and serve to characterize the underlying phenomenon.

Statistics deals with the construction and subsequent analysis of histograms retroactively, that is, from observed data. Probability deals with the prediction of histograms by calculation. In this regard, important properties to look for in calculating probabilities are independence, disjointness, and equal likelihood.

Probabilities and their distributions.

Probability theory applies mathematical principles to random phenomena in order to make precise statements and accurate predictions about seemingly unpredictable events. The probability of an event E, written $\Pr(E)$, is the fraction of times E occurs in an infinitely long sequence of trials. (Defining probability is difficult to do without being circular and without requiring experimentation. A definition requiring the outcome of infinitely many trials is obviously undesirable. The situation is similar to that in geometry, where the term "point" is necessarily left undefined; despite this, geometry has enjoyed great success.) For example, let an "experiment" consist in rolling a single die for which each of the six faces has equal chance of landing facing up. Take event E to mean a 3 or a 5 lands facing up. Evidently, the probability of E is then $\frac{1}{3}$, $\Pr(E) = \frac{1}{3}$, that is, rolling a 3 or 5 will happen approximately one-third of the time in a large number of rolls.

More generally, by an *event* E in a probabilistic experiment, we mean some designated set of outcomes of the experiment. The number of outcomes, or *cardinality*, of E is denoted by $|E|$. The set of all possible outcomes of an experiment is its *universe*, and is denoted by U. Here are some fundamental laws.

Principle of universality. One of the possible outcomes of an experiment will occur with certainty:

$$\Pr(U) = 1. \tag{2.8.1}$$

Principle of disjoint events. If events E and F are *disjoint*, $E \cap F = \emptyset$, that is, they have no outcomes in common, then the probability that E or F will occur (sometimes written $E \cup F$) is the sum

$$\Pr(E \text{ or } F) = \Pr(E) + \Pr(F). \tag{2.8.2}$$

Principle of equal likelihood. Suppose each outcome in U has the same chance of occurring, i.e., is *equally likely*. Then the probability of an event E is the ratio of the number of outcomes making up E to the total number of outcomes,

$$\Pr(E) = \frac{|E|}{|U|}. \tag{2.8.3}$$

To illustrate, consider the experiment of rolling a pair of dice, one red and one green. Any one of six numbers can come up on each die equally likely, so the total number of possibilities is 36; the first possibility could be 1 on red and 1 on green, the second: 1 on red and 2 on green and so on. In this scheme, the last would be 6 on red and 6 on green. So $|U| = 36$. There are two ways to roll an 11, a 5 on red and 6 on green or the other way around. So letting E be the event that an 11 is rolled, we have $\Pr(E) = \frac{2}{36} = \frac{1}{18}$. Let S be the event that a 7 is rolled; this can happen in six different ways, so $\Pr(S) = \frac{6}{36} = \frac{1}{6}$. Now the probability that a 7 or 11 is rolled is their sum

$$\Pr(S \cup E) = \Pr(S) + \Pr(E) = \frac{2+6}{36} = \frac{2}{9}.$$

Since probabilities are frequencies of occurrence, they share properties with statistical distributions. Probability distributions can be visualized by histograms and their mean and variance calculated. For example, let the variable X denote the outcome of the roll of a pair of dice. Table 2.8.1 gives the possible outcomes of X along with their probabilities. Figure 2.8.1 graphically portrays the table as a histogram. Just as in the previous section, the rectangle on x represents the fraction of times a dice roll will be x.

Table 2.8.1. Probabilities for a dice roll.

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

MAPLE

```
> with(stats): with(plots): with(describe):
> roll:=[seq(n,n=2..12)];
> prob:=[1/36,2/36,3/36,4/36,5/36,6/36,5/36,4/36,3/36,2/36,1/36];
> wtroll:=[seq(Weight(roll[i]-1/2..roll[i]+1/2, prob[i]),i=1..11)];
> statplots[histogram](wtroll);
```

MATLAB

```
> roll=ones(1,11);
> roll=cumsum(roll);
> roll=roll+1;
> prob=[1 2 3 4 5 6 5 4 3 2 1]/36;
> bar(roll,prob)
```

Equation (2.7.3) can be used to calculate the mean value $\bar{X}$ of the random variable X, also known as its *expected* value, $E(X)$,

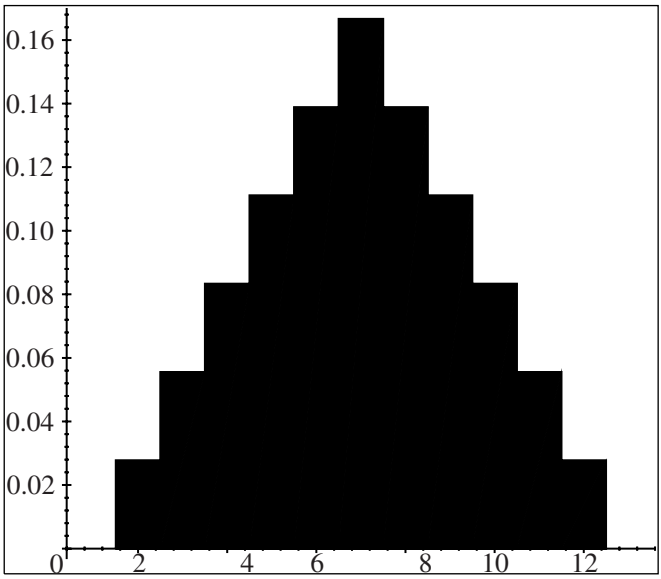


Fig. 2.8.1. Histogram for Table 2.8.1.

$$\bar{X} = \sum_{\substack{\text{over all possible} \\ \text{values } x \text{ of } X}} x \cdot \Pr(X = x). \tag{2.8.4}$$

From Table 2.8.1,

$$E(X) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7.$$

MAPLE
```
> Sum('roll[i]*prob[i]','i'=1..11)=sum('roll[i]*prob[i]',i=1..11);
```

MATLAB
```
> weightedRoll=prob.*roll;
```

Similarly, the variance is defined as

$$V(X) = E(X - \bar{X})^2 = \sum_{\substack{\text{over all possible} \\ \text{values } x \text{ of } X}} (x - \bar{X})^2 \cdot \Pr(X = x). \tag{2.8.5}$$

For a dice roll,

$$V(X) = (2-7)^2 \frac{1}{36} + (3-7)^2 \frac{2}{36} + (4-7)^2 \frac{3}{36} + (5-7)^2 \frac{4}{36} + (6-7)^2 \frac{5}{36} + (7-7)^2 \frac{6}{36} + (8-7)^2 \frac{5}{36} + (9-7)^2 \frac{4}{36}$$

$$+ (10-7)^2\frac{3}{36} + (11-7)^2\frac{2}{36} + (12-7)^2\frac{1}{36} = \frac{35}{6}.$$

MAPLE
```
> Sum('(roll[i]-7)^2*prob[i]','i'=1..11)=sum('(roll[i]-7)^2*prob[i]','i'=1..11);
> mean(wtroll);variance(wtroll);
```

MATLAB
```
> m=sum(weightedRoll)
> v=(roll-m).^2;
  % sum of squared deviations
> var=sum(v.*prob)
```

Probability calculations can be simplified by decomposition and independence.

Consider the experiment of tossing a fair coin in the air four times and observing the side landing up. Suppose we want to calculate the probability that heads will come up three of the four times. This grand experiment consists of four subexperiments, namely, the four individual coin tosses. Decomposing a probability experiment into subexperiments can often simplify making probability calculations. This is especially true if the subexperiments, and therefore their events, are *independent*. Two events E and F are independent when the fact that one of them has or has not occurred has no bearing on the other.

Principle of independence. If two events E and F are independent, then the probability that both will occur is the product of their individual probabilities,

$$\Pr(E \text{ and } F) = \Pr(E) \cdot \Pr(F).$$

One way three heads in four tosses can occur is by getting a head on the first three tosses and a tail on the last one; we will denote this by $HHHT$. Since the four tosses are independent, to calculate the probability of this outcome, we just multiply the individual probabilities of an H the first time, an H the second and also the third, and on the fourth, a T; each of these has probability $\frac{1}{2}$; hence

$$\Pr(HHHT) = \left(\frac{1}{2}\right)^4 = \frac{1}{16}.$$

There are three other ways that three of the four tosses will be H; they are $HHTH$, $HTHH$, and $THHH$. Each of these is also $\frac{1}{16}$ probable; therefore, by the principle of disjoint events,

$$\Pr(\text{three heads out of four tosses}) = 4 \cdot \frac{1}{16} = \frac{1}{4}.$$

Permutations and combinations are at the core of probability calculations.

The previous example raises a question: By direct enumeration, we found that there are four ways to get three heads (or, equivalently, one tail) in four tosses of a coin, but how can we conveniently calculate, for example, the number of ways to get eight

heads in 14 coin tosses or, in general, k heads in n coin tosses? This is the problem of counting *combinations*.

To answer, consider the following experiment: Place balls labeled 1, 2, and so on to n in a hat and select k of them at random to decide where to place the Hs. For instance, if $n = 4$ and $k = 3$, the selected balls might be 3, then 4, then 1, signifying the sequence $HTHH$.

As a subquestion, in how many ways can balls 1, 3, and 4 be selected—this is the problem of counting *permutations*, the various ways to order a set of objects. Actually, there are six permutations here; they are (1, 3, 4), (1, 4, 3), (3, 1, 4), (3, 4, 1), (4, 1, 3), and (4, 3, 1). The reasoning goes like this: There are three choices for the first ball from the possibilities 1, 3, 4. This choice having been made, there are two remaining choices for the second, and finally, only one possibility for the last. Hence the number of permutations of three objects $= 3 \cdot 2 \cdot 1 = 6$.

MAPLE
```
> with(combinat):
> permute([1,3,4]);
> numbperm(3);
```

More generally, the number of permutations of n objects is

$$\text{number of permutations of } n \text{ objects} = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1 = n!.$$

As indicated, this product is written $n!$ and called n *factorial*.

So, in similar fashion, the number of ways to select k balls from a hat holding n balls is

$$n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot (n-k+1).$$

As we said above, the labels on the selected balls signify when the heads occur in the n tosses. But each such choice has $k!$ permutations, all of which also give k heads. Therefore, the number of ways of getting k heads in n tosses is

$$\frac{n(n-1)(n-2)\cdots(n-k+1)}{k(k-1)\cdots 2 \cdot 1}. \tag{2.8.6}$$

MAPLE
```
> with(combinat):
> numbcomb(6,3);
> binomial(6,3);
```

The value calculated by (2.8.6) is known as the number of combinations of n objects taken k at a time. This ratio occurs so frequently that there is a shorthand notation for it, $\binom{n}{k}$, or sometimes $C(n, k)$, called n *choose* k. An alternative form of $\binom{n}{k}$ is

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 2 \cdot 1} = \frac{n!}{k!(n-k)!}, \tag{2.8.7}$$

where the third member follows from the second by multiplying numerator and denominator by $(n-k)!$.

Some elementary facts about n choose k follow. For consistency in these formulas, zero factorial is defined to be 1,

$$0! = 1.$$

The first three combination numbers are

$$\binom{n}{0} = 1, \quad \binom{n}{1} = n, \qquad \binom{n}{2} = \frac{n(n-1)}{2}.$$

There is a symmetry:

$$\binom{n}{k} = \binom{n}{n-k} \quad \text{for all } k = 0, 1, \ldots, n.$$

These numbers n choose k occur in the binomial theorem, which states that for any p and q,

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n. \tag{2.8.8}$$

Finally, the probability of realizing k heads in n tosses of a fair coin is, denoting it by $H_n(k)$,

$$H_n(k) = \binom{n}{k} \left(\frac{1}{2}\right)^n, \quad k = 0, 1, \ldots, n. \tag{2.8.9}$$

The distribution $H_n(k)$ is shown in Figure 2.8.2 for $n = 60$. If the coin is not fair, say the probability of a heads is p and that of a tails is $q = 1 - p$, then $H_n(k)$ becomes

$$H_n(k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \ldots, n. \tag{2.8.10}$$

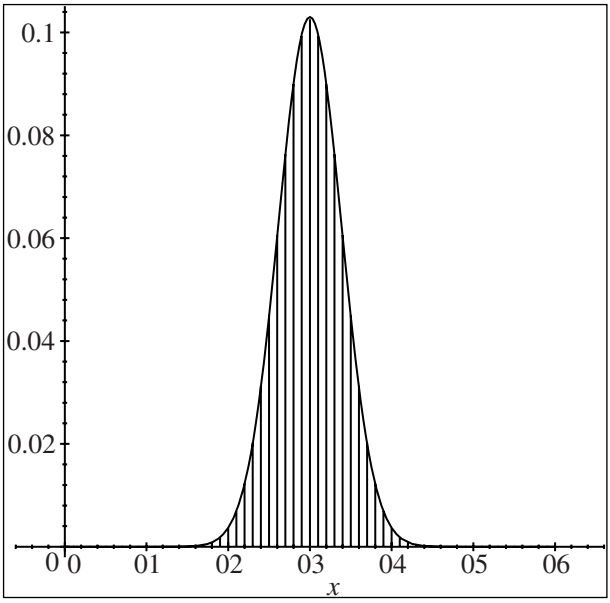


Fig. 2.8.2.

Continuous variations require continuous distributions.

In Figure 2.8.2, we show the heads distribution histogram $H_{60}(k)$ for 60 coin tosses. Notice that the distribution takes on the characteristic bell shape of the Gaussian distribution, as predicted by the central limit theorem, discussed in the previous section:

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \quad -\infty < x < \infty, \tag{2.8.11}$$

where μ and σ are the mean and standard deviation. In the figure, we have superimposed the Gaussian distribution on top of the histogram. In order to get the approximation right, we must match the means and variances of the two distributions. The mean of $H_n(k)$ for a biased coin, (2.8.10), is given by[8]

$$\mu = np. \tag{2.8.12}$$

And the variance of $H_n(k)$ is (see [8])

$$v = npq. \tag{2.8.13}$$

With $p = q = \frac{1}{2}$ and $n = 60$, we get $\mu = 30$ and $\sigma^2 = 15$.

MAPLE
```
> n:=60;
> flip:=[seq(binomial(n,i)*(1/2)^i*(1-1/2)^(n-i),i=0..n)]:
> wtflip:=[seq(Weight(i-1,flip[i]),i=1..n+1)]:
> with(stats); with(describe):
> mu:=evalf(mean(wtflip)); sigma:=standarddeviation(wtflip);
> sigma^2;
```

MATLAB
```
% use the previous m-file, gaussian.m:
%   function y=gaussian(x,m,s);
%   m=mean, s=stddev
%   note 1/sqrt(2*pi)=.3989422803
% y=(.3989422803/s)*exp(-0.5*((x-m)./s).^2);
> x=[-10:.1:10];
> y=gaussian(x,30,sqrt(15)); plot(x,y)
```

Hence Figure 2.8.2 shows the graph of

$$G(x) = \frac{1}{\sqrt{2 \cdot 15 \cdot \pi}} e^{-\frac{1}{2}\frac{(x-30)^2}{15}}.$$

MAPLE
```
> G:=x->exp(-(x-mu)^2/(2*sigma^2))/(sigma*sqrt(2*Pi));
> J:=plot(G(x),x=0..n):
> K:=statplots[histogram](wtflip):
> plots[display]({J,K});
```

[8] Using the fact that $k\binom{n}{k} = n\binom{n-1}{k-1}$ and the binomial theorem, we have

$$\mu = \sum_{k=0}^{n} k\binom{n}{k} p^k q^{n-k} = \sum_{r=0}^{n-1} n\binom{n-1}{r} p^{r+1} q^{n-r} = np.$$

The normal or Gaussian distribution is an example of a continuous distribution. Any nonnegative function $f(x) \geq 0$ with total integral 1,

$$\int_{-\infty}^{\infty} f(x)dx = 1,$$

can define a probability distribution. The condition that the total integral be 1 is dictated by the universality principle, equation (2.8.1). In this role, such a function f is called a *probability density function*. Probabilities are given as integrals of f. For example, let X denote the outcome of the probabilistic experiment governed by f; then the probability that X lies between 3 and 5, say, is exactly

$$\Pr(3 \leq X \leq 5) = \int_3^5 f(x)dx.$$

Similarly, the probability that an outcome will lie in a *very* small interval of width dx at the point x is[9]

$$\Pr(X \text{ falls in an interval of width } dx \text{ at } x) = f(x)dx. \tag{2.8.14}$$

This shows that outcomes are more likely to occur where f is large and less likely to occur where f is small.

The simplest continuous distribution is the *uniform* distribution,

$$u(x) = \text{constant}.$$

Evidently, for an experiment governed by the uniform distribution, an outcome is just as likely to be at one place as another. For example, butterflies fly in a kind of random flight path that confounds their predators. As a first approximation, we might hypothesize that a butterfly makes its new direction somewhere within 45 degrees of its present heading uniformly. Let Θ denote the butterfly's directional change; Θ is governed by the uniform probability law

$$u(\Theta) = \begin{cases} \text{constant} & \text{if } -45 \leq \Theta \leq 45, \\ 0 & \text{otherwise.} \end{cases} \tag{2.8.15}$$

By the universality principle,

$$\int_{-45}^{45} u(\Theta)d\Theta = 1;$$

therefore the constant must be $\frac{1}{90}$ in (2.8.15).

[9] This equation is interpreted in the same spirit as the concept "velocity at a point" in dynamics, which is the ratio of infinitesimals $\frac{ds}{dt}$.

Exercises

1. An undergraduate student in mathematics wants to apply to three of six graduate programs in mathematical biology. She will make a list of three programs in the order of her preferences. Since the order is important, this is a problem of permutations. How many such choices can she make?

```
  MAPLE
> restart;
> with(combinat):
  #list the permutations and count
> permute([a,b,c,d,e,f],3);nops(%);
  #calculate directly
> numbperm(6,3);
  #use the formula
> 6!/3!;

  MATLAB
  % No built-in combinatorics in MATLAB but it is easy to do factorials and hence permutations and
  %   combination calculations
  %   permutations of six things taken three at a time
> n6=1:6; n3=1:3;
> perm6t3=prod(n6)/prod(n3)
```

The student must send a list of three references to any school to which she applies. There are six professors who know her abilities well, of whom she must choose three. Since the order is not important, this is a problem of combinations. How many such lists can she make?

```
  MAPLE
> with(combinat):
> choose([a,b,c,d,e,f],3);nops(%);
> numbcomb(6,3);
> 6!/(3!*(6-3)!);

  MATLAB
  % combinations of six things taken three at a time
> comb6t3=perm6t3/prod(n3)
```

2. Five patients need heart transplants and three hearts for transplant surgery are available. How many ways are there to make a list of recipients? How many ways are there to make a list of the two of the five who must wait for further donors? (The answer to the previous two questions should be the same.) How many lists can be made for the possible recipients in the order in which the surgery will be performed?

```
  MAPLE
> with(combinat):
> numbcomb(5,3); numbcomb(5,2);
> numbperm(5,3);

  MATLAB
  % combinations of five things taken two at a time
> comb5t2=prod(1:5)/(prod(1:2)*prod(1:3))
> comb5t3=prod(1:5)/(prod(1:3)*prod(1:2))
> perm5t3=prod(1:5)/prod(1:2)
```

3. Choose an integer in the interval [1, 6]. If a single die is thrown 300 times, one would expect to get the number chosen about 50 times. Do this experiment and record how often each face of the die appears.

MAPLE

```
> with(stats): with(describe):
> die:=rand(1..6);
> for i from 1 to 6 do
  count[i]:=0
  od:
> for i from 1 to 300 do
  n:=die():
  count[n]:=count[n]+1:
  od:
> for i from 1 to 6 do
  print(count[i]);
  od;
> i:='i':
```

MATLAB

```
  % rand(1,300) is a random vector with components between 0 and up to but not including 1; then 6
  %   times this gives numbers from 0 up to 6; add 1 and get numbers 1 up to 7; finally, fix() truncates
  %   the fractional part
> die=fix(6*rand(1,300)+1);
  % now count the number of 3s
> count3s=1./(die-3); % gives infinity at every 3
> count3s=isinf(count3s); % 1 for infinity, 0 otherwise
> number3s=sum(count3s)
```

4. Simulate throwing a pair of dice for 360 times using a random number generator and complete Table 2.8.2 using the sums of the top faces.

Table 2.8.2.

Sums	Predicted	Simulated
2	10	
3	20	
4	30	
5	40	
6	50	
7	60	
8	50	
9	40	
10	30	
11	20	
12	10	

Calculate the mean and standard deviation for your sample using the appropriate equations of Section 2.7 and compare this with the outcome probabilities. Draw a histogram for the simulated throws on the same graph as the normal distribution defined by (2.8.11); use the mean and the standard deviation you just calculated. The following syntax may help:

MAPLE

```
> with(stats): with(describe):
> red:=rand(1..6):
  blue:=rand(1..6):
> for i from 2 to 12 do
  count[i]:=0:
  od:
```

```
> for i from 1 to 360 do
  n:=red()+blue():
  count[n]:=count[n]+1;
  od:
> for i from 2 to 12 do
  print(count[i]);
  od;
> inter:=seq(n-1/2..n+1/2,n=2..12);
> throws:=[seq(Weight(inter[i-1],count[i]),i=2..12)];
> mean(throws)=evalf(mean(throws));
> standarddeviation(throws)=evalf(standarddeviation(throws));
> theory:=[Weight(inter[1],10), Weight(inter[2],20), Weight(inter[3],30), Weight(inter[4],40),
           Weight(inter[5],50), Weight(inter[6],60), Weight(inter[7],50), Weight(inter[8],40),
           Weight(inter[9],30), Weight(inter[10],20), Weight(inter[11],10)];
> mu:=mean(theory);
> sigma:=standarddeviation(theory);evalf(sigma);
> y:=x->360*exp(-(x-mu)^2/(2*sigma^2))/(sigma*sqrt(2*Pi));
> J:=statplots[histogram](throws):
> K:=plot([x,y(x),x=0..14]):
> plots[display]({J,K});

  MATLAB
> red=fix(6*rand(1,360)+1);
> blue=fix(6*rand(1,360)+1);
> pairDice=red+blue;
> x=2:12;
> hist(pairDice,x)
> hold on
> h=hist(pairDice,x)
> mu=dot(x,h)/sum(h)
  % weight each int by its fraction of outcomes, add
> v=(x-mu).^2; % vector of diffs squared
> var = dot(v,h)/sum(h); % variance
> sigma=sqrt(var)
> t=linspace(2,12);
> y=360*exp(-(t-mu).^2/(2*sigma^2))/(sigma*sqrt(2*pi));
> plot(t,y)
  % the theoretical probability for seeing 2 is 1/36, same for 12, for seeing 3 is 2/36, same for 11, etc.,
  % for seeing 7 is 6/36.
  % compare with h above
> theory=[10 20 30 40 50 60 50 40 30 20 10];
> mu=dot(x,theory)/sum(theory)
> v=(x-mu).^2;var=dot(v,theory)/sum(theory);
> sigma=sqrt(var)
> y=360*exp(-(t-mu).^2/(2*sigma^2))/(sigma*sqrt(2*pi));
> hold off
> plot(t,y); hold on
> hist(pairDice,x)
```

5. This exercise is a study of independent events. Suppose a couple's genetic makeup makes the probability that a child they conceive will have brown eyes equal to $\frac{3}{4}$. Assume that the eye color for two children is a pair of independent events.

(a) What is the probability that the couple will have two blue-eyed children? One blue-eyed and one brown-eyed? Two brown-eyed children? What is the sum of these probabilities?

```
  MAPLE
> binomial(2,0)*1/4*1/4;
> binomial(2,1)*3/4*1/4;
> binomial(2,2)*3/4*3/4;
> sum(binomial(2,j)*(3/4)^j*(1/4)^(2-j),j=0..2);

  MATLAB
  % #ways for two blue eyed is C(2,2)
```

```
  % (2 choose 2)=2!/(2!*0!) so the probability is that times (1/4)^2, etc.
> blublu=prod(1:2)/(prod(1:2)*1)*(1/4)^2 % blu/blu children
> Bwnblu=prod(1:2)/(prod(1:1)*prod(1:1))*(3/4)*(1/4) % Bwn/Blu
> BwnBwn=prod(1:2)/(1*prod(1:2))*(3/4)^2 % Bwn/Bwn children
> blublu+Bwnblu+BwnBwn
```

(b) Suppose that the couple have five children. What is the probability that among the five, exactly two will have brown eyes?

```
  MAPLE
> binomial(5,2)*(3/4)^2*(1/4)^3;

  MATLAB
  % exactly two are brown eyed is (5 choose 2)*(3/4)^2*(1/4)^3
> exact2=prod(1:5)/(prod(1:2)*prod(1:3))*(3/4)^2*(1/4)^3
```

(c) What is the probability that among the five children, there are at least two with brown eyes?

```
  MAPLE
> sum(binomial(5,j)*(3/4)^j*(1/4)^(5-j),j=2..5);

  MATLAB
> exact3=prod(1:5)/(prod(1:3)*prod(1:2))*(3/4)^3*(1/4)^2
> exact4=prod(1:5)/(prod(1:4)*prod(1:1))*(3/4)^4*(1/4)^1
> exact5=prod(1:5)/(prod(1:5)*1)*(3/4)^5
> atleast2=exact2+exact3+exact4+exact5
```

References and Suggested Further Reading

[1] AIDS CASES IN THE U.S.:
HIV/AIDS Surveillance Report, Division of HIV/AIDS, Centers for Disease Control, U.S. Department of Health and Human Services, Atlanta, GA, July, 1993.

[2] CUBIC GROWTH OF AIDS:
S. A. Colgate, E. A. Stanley, J. M. Hyman, S. P. Layne, and C. Qualls, Risk-behavior model of the cubic growth of acquired immunodeficiency syndrome in the United States, *Proc. Nat. Acad. Sci. USA*, **86** (1989), 4793–4797.

[3] IDEAL HEIGHT AND WEIGHT:
S. R. Williams, *Nutrition and Diet Therapy*, 2nd ed., Mosby, St. Louis, 1973, 655.

[4] GEORGIA TECH EXERCISE LABORATORY:
P. B. Sparling, M. Millard-Stafford, L. B. Rosskopf, L. Dicarlo, and B. T. Hinson, Body composition by bioelectric impedance and densitometry in black women, *Amer. J. Human Biol.*, **5** (1993), 111–117.

[5] CLASSICAL DIFFERENTIAL EQUATIONS:
E. Kamke, *Differentialgleichungen Lösungsmethoden und Lösungen*, Chelsea, New York, 1948.

[6] THE CENTRAL LIMIT THEOREM:
R. Hogg and A. Craig, *Introduction to Mathematical Statistics*, Macmillan, New York, 1965.

[7] MORTALITY TABLES FOR ALABAMA:
Epidemiology Report IX (*Number* 2), Alabama Department of Public Health, Montgomery, AL, February, 1994.

[8] BASIC COMBINATORICS:
R. P. Grimaldi, *Discrete and Combinatorial Mathematics*, Addison–Wesley, New York, 1998.