Solution Manual
for

Mathematical Models in Biology: An Introduction

Elizabeth S. Allman
John A. Rhodes

Cambridge University Press, 2003

Preface

Despite our best efforts, there is little chance that these solutions are error-free. Please let us know of any mistakes you find, so that we can correct them.

Special thanks to the public libraries of Berlin, Maryland; Columbia, North Carolina; and Clarksville, Virginia for their hospitality, and to the governments and private benefactors that fund them. They allowed these solutions to be written under more pleasant conditions than we expect most students will experience.

Elizabeth Allman
eallman@maine.edu
John Rhodes
jrhodes@bates.edu
Assateague Island, Maryland
Cape Hatteras, North Carolina
Buggs Island Lake, Virginia

Contents

CHAPTER 1

Dynamic Modeling with Difference Equations

1.1. The Malthusian Model

1.1.1. a.

t	0	1	2	3	4	5
P_t	100	300	900	2700	8100	24300

b. $P_{t+1} = 3P_t$, $\Delta P = 2P_t$

c. $f - d = 2$

1.1.2. a. $P_{t+1} = 2P_t$, $\Delta t = .5$ hr

b. In the following table, t is measured in half-hours.

t	0	2	4	6	8	10
P_t	1	4	16	64	256	1024

t	12	14	16	18	20	22
P_t	4096	16384	65536	262144	1048576	4194304

c. According to the model, the number of cells after ten hours is over one million. Since the observed number is around 30,000, this suggests that the model only fits well at the early stages of cell division, and that during the first ten hours (or twenty time steps) the rate of cell division has slowed. Understanding how and why this slow down occurs could be biologically interesting.

1.1.3. a.

t	0	1	2	3	4	5	6
P_t	1	1.3	1.69	2.197	2.8561	3.7129	4.8268

t	0	1	2	3	4	5	6
N_t	10	8	6.4	5.12	4.096	3.2768	2.6214

t	0	1	2	3	4	5	6
Z_t	10	12	14.4	17.28	20.736	24.8832	29.8598

1.1.4. The first sequence of MATLAB commands has the user iteratively multiply P_t by 1.3. The values are stored as a row vector $\mathtt{x} = [\mathtt{P_0}\ \mathtt{P_1}\ \cdots\ \mathtt{P_t}]$. The second sequence of commands works similarly, but uses a '`for`'-loop to do the iteration automatically.

1.1.5. Experimentally, 9, 18, and 27 time steps are required.
Since $P_t = 1.3^t$, then $P_t \approx 10$ when $\ln 10 \approx t \ln 1.3$. Thus $t \approx \ln 10 / \ln 1.3 \approx 8.8$. Similarly, $P_t \approx 100$ when $t = 17.6$; $P_t \approx 1000$ when $t = 26.3$. Since t must be an integer, the first times when P_t exceeds 10, 100, and 1000 are 9, 18, and 27, respectively.
Notice these times are equally spaced. A characteristic of exponential growth is that the time required for an increase by a factor m is always the same. Here, the time required for an increase by a factor of 10 is always 9 time steps.

1.1.6. By calculating the ratios P_{t+1}/P_t, it is clear that a geometric model does not fit the data well. The finite growth is fast at first, then slows down. It is not

constant as a geometric model would require. If you graph the data, you can see these growth trends and that an exponential growth curve is not a good fit to the data.

However, for the first few time steps (say, $t = 0, 1, 2, 3$) $P_{t+1}/P_t \approx 1.5$, so a geometric model is not a bad one for those initial steps.

1.1.7. a. $k > 1$ and $r > 0$

b. $0 \leq k < 1$ and $-1 \leq r < 0$

c. $k = 1$ and $r = 0$

1.1.8. If $r < -1$, then in a single time step the population must decrease by more than Q_t. This is impossible, since it would result in a negative population size.

1.1.9.

t	0	1	2	3	4	5
N_t	.9613	1.442	2.163	3.2444	4.8667	7.3

1.1.10. a. $\Delta P = 0$

b. once the population size is P^*, it does not change again, but remains P^*.

c. Yes, but only if $r = 0$.

1.1.11. If $\Delta P = rP$, then $P_{t+1} = (1 + r)P_t$. Thus, over each time step the population is multiplied by a factor of $(1 + r)$. Over t time steps, P_0 has been multiplied by a factor of $(1 + r)^t$, giving the formula.

1.1.12. $\Delta P = (b - d + i - e)P$ so $r = (b - d + i - e)$

1.1.13. a. The equation is precisely the statement that the amount of light penetrating to a depth of $d + 1$ meters is proportional to the amount of light penetrating to d meters.

b. $k \in (0, 1)$. The constant of proportionality k can not be greater than 1 since less light penetrates to a depth of $d + 1$ meters than to a depth of d meters. Also, k can not be negative since it does not make sense that an amount of light be negative.

c. The plot shows a rapid exponential decay.

d. The model is probably less applicable to a forest canopy, but it would depend on the makeup of the forest. Many trees have a thick covering of leaves at the tops of their trunks, but few leaves and branches closer to the bottom. This means that it is more difficult for light to penetrate near the tops of trees than it is near the bottom.

1.1.14. a. A plot of the data reveals that it is not well-fit by an exponential model. While the population is constantly increasing, the growth rate is slowing down up until 1945, when the population begins to grow rapidly. The Great Depression and World War II are probably responsible for the slow growth rate. In particular, the tiny growth between 1940 and 1945 is surely due to World War II. The rapid growth after World War II is commonly known as the baby boom. There is a particularly large increase in the US population between 1945 and 1950, though after 1950, even with rapid growth, the growth slows down from the post-war high.

b. The growth rate between 1920 and 1925 is $\lambda = 1.0863$, leading to a model $P_{t+1} = 1.0863P_t$ with time steps of 5 years. This is a poor model to describe the US population and grossly overestimates the population, as a graph shows. A table of values from the model is given below, for comparison purposes. The US population is given in thousands.

year	1920	1925	1930	1935	1940
P_t	106630	115829	125822	136676	148467

year	1945	1950	1955	1960
P_t	161276	175189	190303	206720

c. Answers may vary. Here is one option. The mean of all the ratios P_{t+1}/P_t is 1.0685, leading to the model $P_{t+1} = 1.0685P_t$. This is not a particularly good model either, since it does not capture the growth variations. It fits particularly poorly around the war years. No simple exponential model can do a very good job of fitting this data.

1.1.15. The equation $P_{t+1} = 2P_t$ states the population doubles each time step. This is true regardless of whether the population in measured in individuals or thousands of individuals.

Alternately, if $N_{t+1} = 2N_t$ then $P_{t+1} = N_{t+1}/1000 = 2N_t/1000 = 2P_t$.

1.1.16. a.

t	0	1	2	3	4	5	6
P_t	A	$\sqrt{2}$A	2A	$2\sqrt{2}$A	4A	$4\sqrt{2}$A	8A

Thus $P_{t+1} = \sqrt{2}P_t$.

b. The start of a table for Q_t is below. You can see that $Q_{10} = N_1 =$A and that $Q_5 = P_1$.

t	0	1	2	3	4	5
Q_t	A	$2^{\frac{1}{10}}$A	$2^{\frac{2}{10}}$A	$2^{\frac{3}{10}}$A	$2^{\frac{4}{10}}$A	$\sqrt{2}$A

t	6	7	8	9	10	...
Q_t	$2^{\frac{6}{10}}$A	$2^{\frac{7}{10}}$A	$2^{\frac{8}{10}}$A	$2^{\frac{9}{10}}$A	2A	...

Thus $Q_{t+1} = \sqrt[10]{2}Q_t$.

c. $R_{t+1} = 2^h R_t$

d. If $N_{t+1} = kN_t$ where the time step is one year, and time steps for P_t are chosen as h years, then $1/h$ time steps must pass for P_t to change by a factor of k. Thus in each time step, P_t should change by a factor of $k^{\frac{1}{1/h}} = k^h$. Thus $P_{t+1} = k^h P_t$.

Alternately, if N_t changes by a factor of k each year, then N_t changes by a factor of k^h every h years. Since h years is one time step for P_t, then P_t changes by a factor of k^h each time step. Thus $P_{t+1} = k^h P_t$.

e. For example, suppose $k = 5$, then $\ln k \approx 1.6094$. A table of approximations is given below.

h	.1	-.1	.01	-.01	.001	-.001
$\frac{5^h-1}{h}$	1.7462	1.4866	1.6225	1.5966	1.6107	1.6081

f. By separation of variables,

$$\frac{dP}{dt} = (\ln k)P \implies \frac{dP}{P} = \ln k\, dt \implies$$
$$\int \frac{1}{P}\, dP = \int \ln k\, dt \implies \ln P = t \ln k + C \implies$$
$$P(t) = P_0 e^{t \ln k} \implies P(t) = P_0 k^t.$$

The discrete model gives $N(t) = N_0 k^t$. Thus the discrete (difference equation) model with finite growth rate k agrees with the continuous (differential equation) model with growth rate $\ln k$.

1.2. Nonlinear Models

1.2.1.

t	0	1	2	3	4	5
P_t	1	2.17	4.3788	7.5787	9.9642	10.0106

t	6	7	8	9	10
P_t	9.9968	10.0010	9.9997	10.0001	10.0000

The graph shows typical logistic growth at first, but there is a very slight overshoot past $K = 10$, followed by oscillations that decay in size.

1.2.2. ΔP will be positive for any value of $P < 10$ and ΔP will be negative for any value of $P > 10$. Assuming $P > 0$ so that the model has a meaningful biological interpretation, we see that a population increases in size if it is smaller than the carrying capacity $K = 10$ of the environment, and decreases when it is larger than the environment's carrying capacity.

1.2.3. The MATLAB commands use a '`for`'-loop to iterate the model, storing all population values in a row vector `x`.

1.2.4. For $r = .2$ and .8, the model produces typical logistic growth with the graph from $r = .8$ progressing to the equilibrium more quickly than the graph from $r = .2$. The value $r = 1.3$ also appears to produce typical logistic growth in the early time steps, but later the values of P overshoot (or undershoot) the carrying capacity during a single time step, so there is some oscillation as P_t approaches equilibrium. When $r = 2.2$, surprisingly, the values of P_t do not approach the equilibrium value of 10. Instead, the values ultimately oscillate in a regular fashion above and below K. The values jump between roughly 7.5 and 11.6. For $r = 2.5$, the values of P_t appear to fall into a *four cycle*, that is, they cycle between four values (about 5.4, 11.6, 7, 12.25) above and below $K = 10$. For the values $r = 2.9$ and $r = 3.1$, it is hard to find any patterns to the oscillation of the population values P_t. We will address the effect of changing r on the behavior of the model in the next section.

1.2.5. a. $\Delta P = 2P(1 - P/10)$; $\Delta P = 2P - .2P^2$; $\Delta P = .2P(10 - P)$; $P_{t+1} = 3P_t - .2P_t^2$

b. $\Delta P = 1.5P(1 - P/(7.5))$; $\Delta P = 1.5P - .2P^2$; $\Delta P = .2P(7.5 - P)$; $P_{t+1} = 2.5P_t - .2P_t^2$

1.2.6. b. The MATLAB commands `x=[0:.1:12]`, `y=x+.8*x.*(1-x/10)`, `plot(x,y)` work.

c. The cobweb diagram should fit well with the table below.

t	0	1	2	3	4	5
P_t	1	1.72	2.8593	4.4927	6.4721	8.2988

However, it is hard to cobweb very accurately by hand, so you shouldn't be too surprised if your diagram matches the table poorly. Errors tend to compound with each additional step.

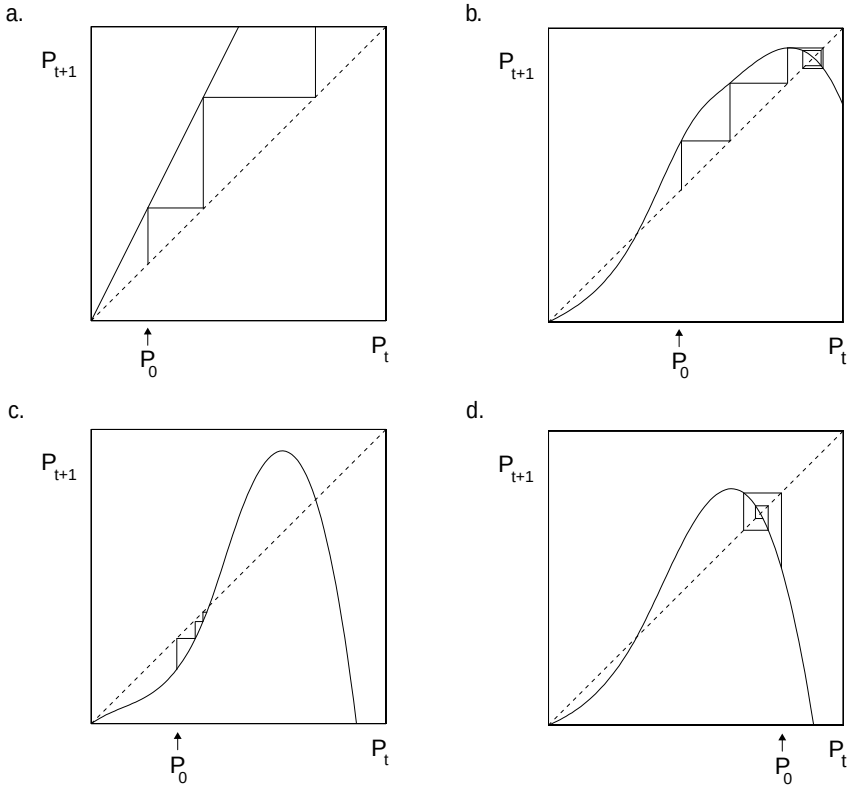
1.2.7. After graphing the data, a logistic equation seems like a reasonable choice for the model. Estimating from the table and graph, $K \approx 8.5$ seems like a good choice for the carrying capacity. Since $P_2/P_1 \approx 1.567$, a reasonable choice for r is .567. However, trial and error shows that increasing the r value a bit appears to give an even better logistic fit. Here is one possible answer: $\Delta P = .63P(1 - P/8.5)$.

1.2.8. a. $M_{t+1} = M_t + .2M_t(1 - \frac{M_t}{200})$, where M_t is measured in thousands of individuals. Notice that the carrying capacity is $K = 200$ thousands, rather than

200,000 individuals. In addition, observe that if the model had been exponential, $M_{t+1} = M_t + .2M_t$, that changing the units would have no effect on the formula expressing the model.

b. $L_{t+1} = L_t + .2L_t(1 - L_t)$, where L_t is measured in units of 200,000 individuals.

1.2.9.



1.2.10. Since the graph appears to be a straight line through the origin with slope 2, $P_{t+1} = 2P_t$. This is an exponential growth model.

1.2.11. a. The equation states the change in the amount N of chemical 2 is proportional to the amount of chemical 2 present. Values of r that are reasonable are $0 \leq r \leq 1$ and $N_0 = 0$. (However, if $r = 1$, then all of chemical 1 is converted to chemical 2 in a single time step.) A graph of N_t as a function of t looks like an exponential decay curve that has been reflected about a horizontal axis, and moved upward so that it has a horizontal asymptote at $N = K$. Thus, N_t is an increasing function, but its rate of increase is slowing for all time.

b. The equation states the amount of chemical 2 created at each time step is proportional to both the amount of chemical 1 and the amount of chemical 2 present. This equation describes a discrete logistic model, and the resulting time plot of N_t shows typical logistic growth. Note that with a small time interval, r should be small, and so the model should not display oscillatory behavior as it approaches equilibrium. If N_0 equals zero, then the chemical reaction will not take place, since at least a trace amount of chemical 2 is necessary for this particular reaction. The shape of a logistic curve makes a lot

of sense for an autocatalytic reaction since it shows that the chemical reaction is slow at first when little of chemical 2 is present, speeds up as the reaction progresses and both chemicals are present in significant amounts, and then slows down again as the amount of chemical 1 diminishes.

1.3. Analyzing Nonlinear Models

1.3.1. a. stable
b. stable
c. unstable
d. unstable

1.3.2. If m denotes the slope of the line, the equilibrium is stable if $|m| < 1$ and unstable if $|m| > 1$.

1.3.3. The equilibrium is stable if $|f'(P^*)| < 1$, and unstable if $|f'(P^*)| > 1$.

1.3.4. The model shows: simple approach to equilibrium without oscillations for $0 < r \leq 1$; approach to equilibrium with oscillations for $1 < r < 2$; 2-cycle behavior for $2 < r < 2.45$; 4-cycle behavior for $2.45 < r < 2.55$. (The values of r for the 2- and 4-cycle behavior are only approximate.)

1.3.5. a. If $\Delta N = rN(1-N)$ and $r \neq 0$, then $\Delta N = 0$, only if $N_t = 0$ or $N_t = 1$, regardless of the value of r.
b. $N_{t+2} = 10.24N_t - 29.568N_t^2 + 30.976N_t^3 - 10.648N_t^4$
c. Set $N_{t+2} = N_t = N$ in part (b) and try to solve $N = 10.24N - 29.568N^2 + 30.976N^3 - 10.648N^4$ or $g(N) = 9.24N - 29.568N^2 + 30.976N^3 - 10.648N^4 = 0$. There are several problems: theoretically there should be four zeros N^* since this is a fourth degree polynomial. We already know that $N^* = 0$ and $N^* = 1$, are roots so the remaining roots should be the points in the 2-cycle. This means we should factor out $N(N-1)$ and use the quadratic formula on the remaining quadratic polynomial. This is a bit difficult to do without the aid of a computer algebra system. In any event, $g(N) = N(N-1)(-10.648N^2 + 20.328N - 9.24) = -10.648N(N-1)(N - .7462465593)(N - 1.16284435)$, so the 2-cycle must be the values .7462465593 and 1.16284435.

1.3.6. a. $P^* = 0, 15$
b. $P^* = 0, 44$
c. $P^* = 0, 20$
d. $P^* = 0, a/b$
e. $P^* = 0, (c-1)/d$

1.3.7. a. At $P^* = 0$, the linearization is $p_{t+1} \approx 1.3p_t$. Since $|1.3| > 1$, $P^* = 0$ is unstable. At $P^* = 15$, the linearization process gives

$$\begin{aligned} 15 + p_{t+1} &= 1.3(15 + p_t) - .02(15 + p_t)^2 \implies \\ 15 + p_{t+1} &= [1.3(15) - .02(15)^2] + 1.3(p_t) - .02(30p_t + {p_t}^2) \implies \\ p_{t+1} &= 1.3(p_t) - .02(30p_t + {p_t}^2) \implies \\ p_{t+1} &\approx 1.3(p_t) - .02(30p_t) = .7p_t. \end{aligned}$$

Since $|.7| < 1$, $P^* = 15$ is stable.
b. $P^* = 0$ is unstable since $|3.2| > 1$; $P^* = 44$ is unstable since $|-1.2| > 1$.
c. $P^* = 0$ is unstable since $|1.2| > 1$; $P^* = 20$ is stable since $|.8| < 1$.

d. $P^* = 0$ is stable if $|1 + a| < 1$ (i.e., $-2 < a < 0$) and unstable if $|1 + a| > 1$ (i.e., $a < -2$ or $a > 0$); $P^* = a/b$ is stable if $|1 - a| < 1$ (i.e., $0 < a < 2$) and unstable if $|1 - a| > 1$ (i.e., $a < 0$ or $a > 2$).
e. $P^* = 0$ is stable if $|c| < 1$ and unstable if $|c| > 1$; $P^* = (c - 1)/d$ is stable if $|2 - c| < 1$ (i.e., $1 < c < 3$) and unstable if $|2 - c| > 1$ (i.e., $c < 1$ or $c > 3$).

1.3.8. The equilibria are $P^* = 0$ and $P^* = 1$. For $P^* = 0$, $p_{t+1} \approx (1 + r)p_t$, thus $P^* = 0$ is stable if $|1 + r| < 1$ and unstable if $|1 + r| > 1$. Equivalently, $P^* = 0$ is stable if $-2 < r < 0$ and unstable if $r < -2$ or $r > 0$.
For $P^* = 1$, $p_{t+1} \approx (1 - r)p_t$, thus $P^* = 0$ is stable if $|1 - r| < 1$ and unstable if $|1 - r| > 1$. Equivalently, $P^* = 0$ is stable if $0 < r < 2$ and unstable if $r < 0$ or $r > 2$. Of course, we have seen that the logistic model falls into a 2-cycle, for r just a little bigger than two.

1.3.9. If $f(P) = P + rP(1 - P) = (1 + r)P - rP^2$, then $f'(P) = (1 + r) - 2rP$. So, $f'(0) = 1 + r$ and $f'(1) = 1 - r$. Thus, $P^* = 0$ is stable if $|1 + r| < 0$ and unstable if $|1 + r| > 0$. $P^* = 1$ is stable if $|1 - r| < 0$ and unstable if $|1 - r| > 0$. This is exactly the same as in the last problem.

1.3.10. Using the point-slope formula for a line with point $(0, 0)$ and slope $f'(0) = 1+r$, the equation of the tangent line at $P^* = 0$ is $y = (1 + r)P$. Using this tangent line approximation, $P_{t+1} \approx (1 + r)P_t$. Thus if P_t is near 0, it changes by a factor of about $1 + r$ with each time step. Thus it will get closer to 0, making the equilibrium stable, provided $|1 + r| < 1$.
Similarly, using the point-slope formula with $(1, 1)$ and $f'(1) = 1 - r$ gives that the equation of the tangent line at $P^* = 1$ is $P_{t+1} - 1 = (1 - r)(P_t - 1)$. Thus, the offset from equilibrium, $P_t - 1$ changes by a factor of $1 - r$ each time step. The offset shrinks, and the equilibrium is stable, provided $|1 - r| < 1$.
Thus, we reach the same conclusion as in the last two problems.

1.3.11. a. Since the concentration of oxygen in the blood stream can not be more than that of the lung, B can not change by more than half the difference $(L - B)$; thus, $0 < r \leq .5$.
b. $\Delta B = r(K - 2B)$
c. If we choose an initial value $0 < B_0 < .5$ for the oxygen concentration in the bloodstream, then B steadily increases up to $B^* = .5K$. The rate of increase slows as B gets close to $.5K$. If r is increased to values just slightly smaller than .5, then B approaches equilibrium quite quickly, much more quickly than with $r = .1$.
d. $B^* = K/2$. (Note that the denominator is the total volume of the two compartments, and B^* has the correct units.) This answer makes sense in that the equilibrium concentration for B (and for L) would be (amount of oxygen)/(total volume).
e. $\Delta b = r(K - 2(K/2 + b_t)) = -2rb_t$. Equivalently, $b_{t+1} = (1 - 2r)b_t$.
f. $b_t = (1 - 2r)^t b_0$. $B_t = K/2 + (1 - 2r)^t b_0$. Note that $b_0 < 0$, since we assume that $L > B$ initially. So, since $0 \leq 1 - 2r < 1$, B increases up to its equilibrium value of $K/2$.
g. Suppose the volume of the lung is V_L and the volume of the bloodstream is V_B, then the total amount of oxygen $K = LV_L + BV_B$ and $L = (K - BV_B)/V_L$. The equation for ΔB then becomes $\Delta B = r((K - BV_B)/V_L - B)$ and the equilibrium is $B^* = K/(V_L + V_B)$, etc.

1.4. Variations on the Logistic Model

1.4.1. a. increase; decrease

b. No. If the relative growth rate is 0, then $P_{t+1} = 0$ and the population has died out. While this is possible, it misses the point of comparing the size of the population in consecutive time steps, since here the extinction of a species is probably of more interest than the relative growth rate. If the relative growth rate were negative, then P_{t+1} and P_t must have opposite signs. This means that either P_{t+1} or P_t represents a population of a negative number of organisms, which is clearly rubbish, if you are trying to model population dynamics.

c. The exponential model $P_{t+1} = \alpha P_t$ has relative growth rate α; the logistic model $P_{t+1} = P_t + rP_t(1 - \frac{P_t}{K})$ has relative growth $(1 + r) - \frac{r}{K}P_t$; the Ricker model $P_{t+1} = P_t e^{r(1-P_t/K)}$ has relative growth $e^{r(1-P_t/K)}$; the fourth model $P_{t+1} = \frac{\lambda P_t}{(1+aP_t)^\beta}$ has relative growth rate $\frac{\lambda}{(1+aP_t)^\beta}$.

d. The graph of the relative growth rate for the exponential model is a horizontal line; the graph of the relative growth rate for the logistic model, assuming r and K are both positive, is a decreasing line with slope $-r/K$ and y-intercept $(1+r)$; the graph of the relative growth rate for the Ricker model with $r, K > 0$ is an exponential decay curve; the graph of the relative growth rate for the fourth model depends on particular parameter choices.

1.4.2. According to the Allee effect, a population must reach a critical number if it is to survive and thrive. If the population is too small, then it dies out. One explanation for the Allee effect is that a species needs a certain number of members to gather enough food to survive or to protect itself from predators or environmental hazards. Another possibility is that a species needs to reach a certain population size in order to breed successfully and in numbers large enough to sustain a population. (Some species are so endangered, that intervention by humans has been necessary to sustain their numbers.) For modeling purposes, if a population dies out when $0 < P_t < L$, then the interval $[0, L]$ is sometimes called the *pit of extinction*.

1.4.3. a. The equations say the population declines if it is too small or if it is too large. The population will grow if is larger than some critical number L, but not so large that resource limitations affect the population adversely.

b. The graph of the polynomial $y = P(K - P)(P - L)$ has horizontal-axis intercepts at $P = 0$, $P = K$, and $P = L$. Since $0 < L < K$, the polynomial's values are negative for $0 < P < L$ (when exactly one factor is negative) and $K < P$ (when all three factors are negative), and positive for $L < P < K$ (when exactly two factors are negative).

Thus for $0 < P < L$, $\Delta P/P < 0$ so the per-capita growth rate is negative; the population suffers due to the Allee effect. For $L < P < K$, $\Delta P/P > 0$ so the per-capita growth rate is positive and the population grows. Finally, for $P > K$, $\Delta P/P < 0$ and the population declines due to scarcity of resources.

c. MATLAB experimentation.

d. For P much greater than K, the cubic gives value below -1, which is not possible for a per-capita growth rate. A better model might have the curve asymptotically decay down to $y = -1$, so that the per-capita growth rate is always at least -1. Similarly, for $0 < P < L$, it is best that the per-capita growth rate never drop below -1, though it may with the given cubic. Also, the maximum of the cubic can be unrealistically large, depending on the values

of L and K. All of these features could be improved using a more complicated formula.

1.5. Comments on Discrete and Continuous Models

1.5.1. a. To verify that $N(t) = K(1 + Ce^{-rt})^{-1}$ is a solution to the differential equation we compute

$$\begin{aligned} N'(t) &= K(-1)(1 + Ce^{-rt})^{-2}(Ce^{-rt})(-r) \\ &= rK(1 + Ce^{-rt})^{-1}\left(\frac{Ce^{-rt}}{1 + Ce^{-rt}}\right) \\ &= rN\left(\frac{1}{K}\right)\left(\frac{KCe^{-rt}}{1 + Ce^{-rt}}\right) \\ &= rN\left(\frac{N}{K}\right)Ce^{-rt} \\ &= rN\left(\frac{N}{K}\right)\left(\frac{K}{N} - 1\right) \\ &= rN\left(1 - \frac{N}{K}\right). \end{aligned}$$

Moreover,

$$N(0) = \frac{K}{1 + C} = \frac{K}{1 + \frac{K - N_0}{N_0}} = \frac{KN_0}{N_0 + K - N_0} = N_0.$$

(Note that the given solution can be found from the differential equation by separation of variables.)

b. Be careful to take N_0 close to zero to get the full logistic curve.

c. For small positive N_0, increasing r makes the population tend to the equilibrium value of 1 more rapidly. If $N_0 > 1$, then the decrease to the equilibrium is also more rapid with larger r values. Of course, this makes sense since r is considered the intrinsic growth rate of the logistic model.

Note that even for very large r, no cycle or chaotic behavior occurs with this model.

CHAPTER 2

Linear Models of Structured Populations

2.1. Linear Models and Matrix Algebra

2.1.1. a. $\begin{pmatrix} 0 \\ 17 \end{pmatrix} = (0, 17)$

b. $(-1, 11, -18)$

c. $\begin{pmatrix} 0 & -8 \\ 17 & 30 \end{pmatrix}$

d. $\begin{pmatrix} -1 & -2 & 7 \\ 11 & 7 & -8 \\ -18 & -1 & -1 \end{pmatrix}$

2.1.2. The matrix on the left has 1 column, but the matrix on the right has 2 rows. For multiplication to have been possible, these numbers would have had to have been equal.

2.1.3. a. $\begin{pmatrix} 4 & 1 \\ -3 & 3 \end{pmatrix}$

b. $\begin{pmatrix} -1 & 3 \\ -5 & 3 \end{pmatrix}$

c. $\begin{pmatrix} 4 & 5 \\ -4 & -2 \end{pmatrix}$

d. $\begin{pmatrix} -1 & 4 \\ -2 & -1 \end{pmatrix}$

e. $\begin{pmatrix} 2 & 4 \\ -2 & 2 \end{pmatrix}$

f. Both sides equal $\begin{pmatrix} -7 & 8 \\ -6 & 9 \end{pmatrix}$.

2.1.4. a. $\begin{pmatrix} 4 & 2 & -2 \\ 0 & 1 & 2 \\ -1 & 0 & 1 \end{pmatrix}$

b. $\begin{pmatrix} 3 & 3 & -2 \\ 4 & 4 & 0 \\ -5 & 0 & 1 \end{pmatrix}$

c. $\begin{pmatrix} 8 & 1 & -1 \\ -4 & 2 & -2 \\ -3 & 0 & -2 \end{pmatrix}$

d. $\begin{pmatrix} 2 & -1 & 1 \\ 4 & 1 & -2 \\ 3 & -1 & 5 \end{pmatrix}$

e. $\begin{pmatrix} 2 & 0 & -2 \\ 4 & 2 & 0 \\ -2 & 2 & -4 \end{pmatrix}$

f. Both sides equal $\begin{pmatrix} 2 & 2 & -4 \\ -9 & -3 & 5 \\ 11 & 5 & -9 \end{pmatrix}$

2.1.5. $A(c\mathbf{x}) = \begin{pmatrix} r(cx) + s(cy) \\ t(cx) + u(cy) \end{pmatrix}$, $c(A\mathbf{x}) = \begin{pmatrix} c(rx + sy) \\ c(tx + uy) \end{pmatrix}$

2.1.6. Rounding to 4 decimal digits, $P^2 = \begin{pmatrix} .9852 & .0247 \\ .0148 & .9753 \end{pmatrix}$, $P^3 = \begin{pmatrix} .9779 & .0368 \\ .0221 & .9632 \end{pmatrix}$, $P^{500} = \begin{pmatrix} .6250 & .6250 \\ .3750 & .3750 \end{pmatrix}$. The matrices are the transition matrices for the forest succession model if the time steps were taken to be two years, three years, or five hundred years respectively. Interestingly, the columns of P^{500} are identical and the column entries are in the same ratio as the equilibrium ratio of A trees to B trees that we saw in the text.

2.1.7. All initial vectors with nonnegative entries will tend towards an equilibrium state of $(625, 375)$.

2.1.8. a. The transition matrix is $P = \begin{pmatrix} 0 & 0 & 73 \\ .04 & 0 & 0 \\ 0 & .39 & 0 \end{pmatrix}$ with $\mathbf{x}_t = (E_t, L_t, A_t)$.

b. $P^2 = \begin{pmatrix} 0 & 28.47 & 0 \\ 0 & 0 & 2.92 \\ .0156 & 0 & 0 \end{pmatrix}$, $P^3 = \begin{pmatrix} 1.1388 & 0 & 0 \\ 0 & 1.1388 & 0 \\ 0 & 0 & 1.1388 \end{pmatrix}$. The matrices represent the transition matrices describing what happens to the population classes over two and three time steps.

c. All the diagonal entries of P^3 are 1.1388. In the text, we argued that the adult insect population would grow exponentially by a factor of 1.1388 every three time steps. This diagonal matrix shows that all three classes of insect grow at the same exponential rate over three time steps, and that over three time steps there is no interaction among the three class sizes.

2.1.9. a. The transition matrix is $P = \begin{pmatrix} 0 & 0 & 73 \\ .04 & 0 & 0 \\ 0 & .39 & .65 \end{pmatrix}$ with $\mathbf{x}_t = (E_t, L_t, A_t)$.

b. $P^2 = \begin{pmatrix} 0 & 28.47 & 47.45 \\ 0 & 0 & 2.92 \\ .0156 & .2535 & .4225 \end{pmatrix}$, $P^2 = \begin{pmatrix} 1.1388 & 18.5055 & 30.8425 \\ 0 & 1.1388 & 1.898 \\ .01014 & .164775 & 1.413425 \end{pmatrix}$. Notice that in P^3 there are now non-zero off-diagonal entries (signifying interaction among the sizes of the classes) and that the $(3, 3)$ entry is larger than in the last problem. These are the effects of 65% of the adults living on to the next cycle and reproducing again.

c. All three populations appear to grow roughly exponentially. There is some oscillation in the population values that is particularly noticeable for a small number of iterations. Of course, if 65% of the adults live on into the next time step to produce eggs, the populations should grow even faster than in the previous problem.

2.2. Projection Matrices for Structured Models

2.2.1. The matrix for the first insect model is a Leslie matrix, and the matrix for the more complicated insect model is an Usher matrix, where the addition of .65 in the $(3,3)$ position is for the 65% of the adult population that live on into the next reproductive cycle. See problems 2.1.8(a) and 2.1.9(a) for the matrices.

2.2.2. Ultimately, all ten classes settle into what appears to be exponential growth, possibly after some initial oscillation. The class of individuals ages 0–4 is the most populous, followed by the class of individuals ages 5–9, etc.

2.2.3. Letting A, B, and C be the matrices in the order given, $\det A = -1$, $A^{-1} = \begin{pmatrix} -3 & 2 \\ 2 & -1 \end{pmatrix}$; $\det B = 8$, $B^{-1} = \begin{pmatrix} 3/8 & 1/8 \\ -1/4 & 1/4 \end{pmatrix}$; $\det C = 0$, so C has no inverse.

2.2.4. Letting A, B, and C be the matrices in the order given, $\det A = -5$, $A^{-1} = \begin{pmatrix} 2/5 & 1/5 & -1/5 \\ -4/5 & 3/5 & 2/5 \\ -3/5 & 1/5 & -1/5 \end{pmatrix}$; $\det B = 8$, $B^{-1} = \begin{pmatrix} 1/4 & -1/8 & 1/2 \\ 1/4 & 3/8 & -1/2 \\ 1/4 & 3/8 & 1/2 \end{pmatrix}$; $\det C = 0$, so C has no inverse.

2.2.5. a. 3

b. 50%

c. 20% of the organisms in the immature class remain in the immature class with each time step.

d. 30% of the organisms in the immature class progress into the adult class with each time step.

2.2.6. a. $P^{-1} = \begin{pmatrix} -.625 & 3.75 \\ .375 & -.25 \end{pmatrix}$

b. $\mathbf{x}_0 = (1000, 300)$, $\mathbf{x}_2 = (1570, 555)$,

2.2.7. a. A^{100} is the transition matrix for the model in which the time steps are one hundred times as large as they were taken for A. For instance, if $\mathbf{x}_n$ is a population vector and $\mathbf{x}_{n+1} = A\mathbf{x}_n$ is the new population after one year, then $\mathbf{x}_{n+100} = A^{100}\mathbf{x}_n$ is the population vector after one hundred years. If, instead, $\mathbf{x}_n$ is multiplied by $(A^{100})^{-1}$, then the resulting vector is $\mathbf{x}_{n-100}$, the population vector for a time one hundred years earlier.

b. $(A^{-1})^{100}$ is the hundredth power of the transition matrix that take you back one time step; thus, this matrix multiplies a population vector to create a population vector for a time one hundred time steps earlier. In other words $(A^{-1})^{100}\mathbf{x}_n = \mathbf{x}_{n-100}$.

c. Both matrices represent the transition matrix for calculating population vectors one hundred time steps earlier. Since there is nothing special about 100, more generally $(A^n)^{-1} = (A^{-1})^n$ since both are used to project populations n time steps into the past.

2.2.8. .11 represents the percentage of pups that remain pups after one year. (Pups can not give birth.) One possible explanation for some pups living but not progressing into the yearling stage after one year is that coyotes are born over several months throughout the year. The .15 entries indicate that on average each yearling and adult gives birth to .15 pups each year. The percentage of pups that progress into the yearling stage is 30% each year, so $1-.11-.30 = 59\%$ of pups die. While 60% of the yearlings progress into the adult stage, the remaining 40% die. Finally, each year 40% of the adult coyotes die, but 60% live on into the next time step.

2.2.9. a. Both $A\mathbf{x}$ and $A\mathbf{y}$ equal $(17, 51)$, though $\mathbf{x} \neq \mathbf{y}$. Notice that A has no inverse.
b. If A^{-1} exists, then $A\mathbf{x} = A\mathbf{y}$ implies $A^{-1}A\mathbf{x} = A^{-1}A\mathbf{y}$ or $\mathbf{x} = \mathbf{y}$.

2.2.10. a. $(AB)^{-1} = B^{-1}A^{-1} = \begin{pmatrix} -7 & 9 \\ 4 & -5 \end{pmatrix}$, $A^{-1}B^{-1} = \begin{pmatrix} -8 & 3 \\ 11 & -4 \end{pmatrix}$.
b. Answers may vary.
c. Answers may vary.

2.2.11. a. By associativity, $(B^{-1}A^{-1})(AB) = B^{-1}(A^{-1}A)B = B^{-1}B = I$. This shows that (AB) has a left inverse, but if a left inverse exists for a square matrix, then it also serves as a right inverse.
b. $\mathbf{x}_1 = W^{-1}\mathbf{x}_2$; $\mathbf{x}_0 = D^{-1}x_1$. Thus, to find $\mathbf{x}_0$ from $\mathbf{x}_2$, it is necessary to multiply first by W^{-1}, and then by D^{-1}: $\mathbf{x}_0 = D^{-1}W^{-1}\mathbf{x}_2$. This shows that the inverse of (WD) is the product $D^{-1}W^{-1}$ by indicating how to obtain $\mathbf{x}_0$ back from $\mathbf{x}_2 = WD\mathbf{x}_0$. Another way to explain this is that if you want to undo the action of a dry year followed by a wet year, you first undo the action of the recent wet year, then undo the action of the initial dry year.

2.2.12. a. $A_{t+1} = 2/3A_t + 1/4B_t$, $B_{t+1} = 1/3A_t + 3/4B_t$
b. $P = \begin{pmatrix} 2/3 & 1/4 \\ 1/3 & 3/4 \end{pmatrix}$, with $\mathbf{x}_t = (A_t, B_t)$.
c. $P^2 = \begin{pmatrix} 19/36 & 17/48 \\ 17/36 & 31/48 \end{pmatrix}$ so using decimal approximations $A_{t+1} = .5278A_t + .3542B_t$, $B_{t+1} = .4722A_t + .6458B_t$.
d. $P^{-1} = \begin{pmatrix} 9/5 & -3/5 \\ -4/5 & 8/5 \end{pmatrix}$ so $A_{t-1} = 1.8A_t - .6B_t$, $B_{t+1} = -.8A_t + 1.6B_t$.
e. The values of the populations are given in the table below. The populations seem to be stabilizing with $A_t \approx 85.7$ and $B_t \approx 114.3$.

t	0	1	2	3	4	5
A_t	100.0000	91.6667	88.1944	86.7477	86.1449	85.8937
B_t	100.0000	108.3333	111.8056	113.2523	113.8551	114.1063

t	6	7	8	9	10
A_t	85.7890	85.7454	85.7273	85.7197	85.7165
B_t	114.2110	114.2546	114.2727	114.2803	114.2835

e. If the initial populations A_0 and B_0 are non-negative and sum to 200, then they tend toward an equilibrium of around $(85.7, 114.3)$.

2.3. Eigenvectors and Eigenvalues

2.3.1. The model does behave as expected, showing slow exponential growth in both classes, with decaying oscillations superposed.

2.3.2. MATLAB finds that the eigenvector corresponding to eigenvalue 1.0512 is $(-.8852, -.4653)$ and the eigenvector corresponding to eigenvalue $-.9512$ is $(-.9031, .4295)$. These are essentially the same eigenvectors that were given in the text, since any scalar multiple of these are also eigenvectors. The text has simply multiplied them by -1. Note that MATLAB calculates eigenvectors (x, y) with $x^2 + y^2 = 1$

2.3.3. The eigenvalues of the plant model are approximately 1.1694, $-.7463$, $-.0738$, and .1107. The dominant eigenvalue is larger than one and the figure shows that the populations grow exponentially, as expected from an eigenvalue analysis. Since two of the eigenvalues are negative but smaller than 1 in absolute

value, there is some decaying oscillation superposed on the overriding trend of exponential growth.

2.3.4. a. In zeroing out the first row, no new ungerminated seeds are added to the population. Since the $(2,1)$ entry has been replaced with zero, no ungerminated seeds progress into the class of sexually immature plants. This eliminates the class of ungerminated seeds from the population. (One reason for considering this model would be to understand the effect of ungerminated seeds on the population dynamics, by imagining what would happen in their absence.)

b. The dominant eigenvalues of the model in the text is 1.1694 and the dominant eigenvalue of the altered matrix is 1.1336. This means that both models predict exponential growth, though the growth rate for the model with no ungerminated seeds is slightly slower. If the ungerminated seed entry of the dominant eigenvectors is discarded, there is also little difference in the stable stage vector for the two models.

c. The ungerminated seeds might be gathered by animals and spread throughout a region, possibly germinating in a later year and spreading the plant species. Also, if the plants have a bad year (due to factors not included in the model, such as drought, extreme cold, fire, etc.) and many fail to survive, the ungerminated seeds still remain in the area despite the temporary adverse growing conditions. If they then germinate at a later date, this may help the population recover. Even though they have little effect on the 'normal year' population dynamics, the ungerminated seeds may well be important.

2.3.5. a. The model should produce slow exponential growth. One way to see this is to notice that after one time step 40% of the first class survives to reproduce and 30% remain in the first class. Of the 30%, the model indicates that 40%, or $(.3)(.4) = 12\%$ will survive to reach the reproduction stage after a second time step. This means that at least $.4 + .12 = 52\%$ of the first class will survive to reproduce. Since on average, each adult produces two offspring, we should expect at least $(.52)2 = 1.04 > 1$ offspring produced by individual members of the first class on average. Thus, the population will grow slowly. In fact, the growth rate should be a little larger than 1.04, since $(.3)^2(.4) = .036 = 3.6\%$ of the first class progress into the second stage after three time steps and then reproduce. Similarly, for four, five, ... time steps. Clearly, the situation is somewhat complicated and an eigenvalue analysis can help us understand the growth trend more easily.

b. The eigenvalue 1.0569 is dominant with eigenvector $(.9353, .3540)$. The other eigenvalue is $-.7569$ with corresponding eigenvector $(-.8841, .4672)$.

c. The intrinsic growth rate is 1.0569, a number a little bit bigger than 1.04 as anticipated by (a). The stable stage distribution is $(2.6423, 1)$.

d. Using eigenvectors calculated by MATLAB, $(5,5) = 9.0100(.9353, .3540) + 3.8757(-.8841, .4672)$.

e. $\mathbf{x}_t = 9.0100(1.0569)^t(.9353, .3540) + 3.8757(-.7569)^t(-.8841, .4672)$.

2.3.6. a. Since, as discussed in the text, a model with 0 replacing the .65 results in growth by a factor 1.1388 every 3 time steps, we expect a greater growth here (greater than $\sqrt[3]{1.1388} = 1.0443$ for each time step).

b. The dominant eigenvalue is 1.3118 with $\mathbf{v}_1 = (.9994, .0305, .0180)$ its eigenvector. The other eigenvalues are complex, $-.3309 + .8710i$ and $-.3309 - .8710i$,

with corresponding eigenvectors, $\mathbf{v}_2 = (-.6521+.7568i, .0403+.0146i, -.0061-.0112i)$ and $\mathbf{v}_3 = (-.6521-.7568i, .0403-.0146i, -.0061+.0112i)$.

c. The intrinsic growth rate is 1.3118, while the stable stage distribution is $(55.6492, 1.6969, 1)$. You can see the large number of members of the first class compared to the other two classes.

d. $(100, 10, 1) = 135.2502\mathbf{v}_1 + (61.9629 - 30.1548i)\mathbf{v}_2 + (61.9629 + 30.1548i)\mathbf{v}_3$

e. $\mathbf{x}_t = 135.2502(1.3118)^t\mathbf{v}_1 + (61.9629 - 30.1548i)(-.3309 + .8710i)^t\mathbf{v}_2 + (61.9629 + 30.1548i)(-.3309 - .8710i)^t\mathbf{v}_3$

2.3.7. The dominant eigenvalue is .6791 so the coyote population will decline rather rapidly. The stable stage distribution is $(2.2636, 1, 7.5877)$.

2.3.8. The intrinsic growth rate is 1.0818, describing growth. The stable age distribution is $(.4332, .3991, .3682, .3397, .3130, .2882, .2649, .2430, .2243, .2039)$. To find the intrinsic annual growth rate, it is necessary to take the fifth root: $\sqrt[5]{1.0818} = 1.0159$.

2.3.9. a. The transition matrix $P = \begin{pmatrix} 0 & 5 \\ 1/6 & 1/4 \end{pmatrix}$ is for an Usher model.

b. The dominant eigenvalue is 1.0464 with eigenvector $(.9788, .2048)$. The other eigenvalue is $-.7964$ with eigenvector $(-.9876, .1573)$.

c. The intrinsic growth rate is 1.0464 and the population will grow. The stable stage distribution is $(4.7783, 1)$.

2.3.10.

$$\begin{aligned} |(a+bi)(c+di)| &= |(ac-bd)+(ad+bc)i| = \sqrt{(ac-bd)^2+(ad+bc)^2} \\ &= \sqrt{a^2c^2 - 2abcd + b^2d^2 + a^2d^2 + 2abcd + b^2c^2} \\ &= \sqrt{a^2(c^2+d^2) + b^2(c^2+d^2)} = \sqrt{a^2+b^2}\sqrt{c^2+d^2} \\ &= |a+bi||c+di|. \end{aligned}$$

2.4. Computing Eigenvectors and Eigenvalues

2.4.1. a. A: $\lambda_1 = 1$ and $\lambda_2 = .6$; B: $\lambda_1 = -1$ and $\lambda_2 = 5$; C: $\lambda_1 = -3$ and $\lambda_2 = 2$

b. A: $\mathbf{v}_1 = (3, 1)$, $\mathbf{v}_2 = (1, -1)$; B: $\mathbf{v}_1 = (-2, 1)$, $\mathbf{v}_2 = (1, 1)$; C: $\mathbf{v}_1 = (-3, 2)$, $\mathbf{v}_2 = (1, 1)$

2.4.2. The answers should agree. However, since the power method creates a dominant eigenvector with the largest entry equal to 1 and your calculation in the last problem may not have, it may be necessary to rescale to get agreement.

2.4.3. Yes, the answers agree, but rescaling may be necessary since MATLAB and the power method choose different ways of scaling an eigenvector.

2.4.4. a. $\lambda_1 = \lambda_2 = 2$, $\mathbf{v}_1 = (1, 0)$, and $\mathbf{v}_2 = (0, 1)$

b. $\lambda_1 = \lambda_2 = 2$ and $\mathbf{v}_1 = (1, 0)$. However, it is impossible to find a second eigenvector, since $B - 2I = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and the only solutions to $(B-2I)\mathbf{x} = \mathbf{0}$ are $(c, 0) = c\mathbf{v}_1$. (The 1 in the $(1, 2)$ entry forces the second entry of any nonzero eigenvector to be zero.)

2.4.5. The power method is an algorithm for finding a dominant eigenvector that depends on the truth of the Strong Ergodic Theorem. This method works almost always since a vector $\mathbf{v}$ picked at random is likely to contain a component of the dominant eigenvector, i.e., a random vector can be expressed as a linear combination of the dominant eigenvector and other vectors, where the coefficient

in front of the dominant eigenvector is nonzero. Thus, $\mathbf{v}$ will be drawn to the dominant eigenvector, when multiplied repeatedly by the matrix and rescaled. If we think of the matrix as describing a population model, the power method is essentially to pick an initial population at random and then iterate the model until the stable stage distribution becomes clear. This gives the dominant eigenvector. The rescaling at each step just keeps the numbers in a reasonable range.

CHAPTER 3

Nonlinear Models of Interactions

3.1. A Simple Predator–Prey Model

3.1.1. The prey population peaks first, with the predator population peaking slightly later. Similarly, the prey population bottoms out slightly before the predator population. This is biologically reasonable, since when the prey population peaks, there will be a slight time lag as the predator population grows to its peak through consuming the prey. Any changes in prey population size should be reflected in the predator population size slightly later.

3.1.2. a. The peaks on the graph of $P_t = \cos t$ lead those on the graph of $Q_t = \sin t$ (by a time interval of $\pi/2$), similar to those of the prey and predator.
b. The plotted points lie on a circle centered at the origin, starting on the x-axis when $t = 0$ and proceeding counterclockwise around the circle.
c. Because the oscillations in the first figure get smaller as time increase (i.e., they are damped), the spiral in the second figure goes inward.

3.1.3. For increased s, the equilibrium value of the prey, P^*, appears unchanged, though that of the predator, Q^*, is reduced. Since larger s means the predator-prey interaction is harming the prey population more, this is a bit surprising, since the prey population size is all that is ultimately reduced. On further reflection, this is not unreasonable biologically, as a result of 'feedback' to the predator.
As v is increased, the equilibrium P^* is reduced, and Q^* increased. Since larger v means the prey benefits more from the predator-prey interaction, it is reasonable that Q^* would be larger, resulting in smaller P^*.

3.1.4. The effect of increasing s and v on the stable equilibrium is discussed in the last problem, so here we describe only the movement toward or away from the equilibrium.
If s or v is increased by a small amount, simulations often show qualitatively-similar damped oscillations of populations toward an equilibrium, with counter-clockwise motion in the phase plane. If either parameter is increased excessively, orbits become more likely to leave the phase plane, signifying extinction. This is reasonable since increasing either of the parameters increases the effect of the predator-prey interaction, which seems likely to destabilize things. Large increases in the parameters can also lead to growing population oscillations (still with counterclockwise orbits), which also result in ultimate extinction.

3.1.5. The oscillatory behavior appears for most values of r, indicating their origin in the predator-prey interaction. For instance, when $r = .3$, which would not produce oscillations in the one-population logistic model, the predator-prey oscillations seem to take even longer to damp out than for the original value of r.

3.1.6. With the other parameters as in the text, $r = 2.1$ results in what appears to be a stable equilibrium, despite the fact that in the logistic model it leads to a 2-cycle. This illustrates that a predator-prey interaction can have a stabilizing effect on otherwise complex dynamics. A real-world example of this involves deer populations and hunters. (Can you find parameter values for which the logistic model is chaotic, yet the predator-prey model has an apparently stable equilibrium?)

3.1.7. The parameter w represents the size of the prey population that can be protected in the refuge. If $P > w$, then $P-w$ is the part of the prey population not in the refuge, which is therefore subject to the predator-prey interaction. The given interaction terms thus describe the predation appropriately. If $P < w$, however, the terms are not correct. Replacing the $P - w$ with $\max(0, P - w)$ would be better.

3.1.8. The introduction of a refuge typically results in a larger equilibrium value of the prey. If the refuge is small, the equilibrium value for the predator may increase. For a larger refuge, typically the equilibrium value of the predator is reduced. Oscillations may also tend to damp out faster. Using the model parameters of the text with $w = 0, .1, .2, .3$ gives good examples. Be careful to only consider orbits where P stays larger than w.

3.1.9. $P(1 - e^{-vQ})$, $\sqrt{PQ}$, and $P + Q$ all increase if either P or Q is increased. Only the first two of these are reasonable as interaction terms, though, since the there is no interaction between P and Q in $P + Q$.

3.1.10. In the absence of predators ($Q = 0$), this is the Ricker model of the prey. In the absence of prey ($P = 0$), the predators immediately die out. The factor e^{-sQ} in the formula for P is 1 when $Q = 0$ and decreases if Q is increased. Thus the larger Q_t is, the more this factor shrinks the size of P_{t+1}, as predation should. The factor $(1 - e^{-vQ})$ in the formula for Q is 0 when $P = 0$ and increases toward 1 as Q is increased. Thus Q_{t+1} will be larger for larger Q_t and fixed P_t, but will never exceed uP. This means the predator population can be at most a constant multiple of the prey population. These modeling equations are probably more reasonable in most situations than those used in the text, but they are a bit more complicated to analyze.

3.1.11. Behavior is qualitatively similar to model of text, at least for some parameter values (e.g., $r = 1.3$, $K = 1$, $s = .5$, $u = 5$, $v = 1.6$). Varying parameters produces interesting, yet reasonable results (e.g., changing to $s = 1.5$).

3.2. Equilibria of Multipopulation Models

3.2.1. If $u/v = 1$, the vertical line of the Q-nullcline joins the sloping line of the P-nullcline on the P-axis at $P = 1$. Then the only equilibria are $(0, 0)$ and $(1, 0)$. If $u/v > 1$, the vertical line lies even further to the right, and intersects the sloping line below the P-axis. The resulting equilibrium has $Q^* < 0$, so $(0, 0)$ and $(1, 0)$ are the only two biologically meaningful equilibria..

3.2.2. By problem 3.2.1, we need only discuss situations with $u/v < 1$. If u is increased, or v is decreased, the vertical line in the Q-nullcline moves right, causing the equilibrium at the intersection of it and the sloping line to move right and down. Thus P^* increases and Q^* decreases. Since increasing u means the predator dies more quickly, and decreasing v means the predator benefits

less from the predator-prey interaction, it is biologically reasonable that Q^* should decrease and therefor P^* should increase.

3.2.3. The nullclines are described in problem 3.2.1. For both $u/v = 1$ and $u/v > 1$ the region under the sloping line should have arrows pointing down and to the right. The region to the right of the vertical line should have arrows pointing up and to the left. The remaining region should have arrows pointing down and to the left. MATLAB experiments confirm this.

3.2.4. a. Yes; b. Yes, interpreting $\Delta P = 0$ as meaning the orbit must move vertically up or down; c. Yes, interpreting $\Delta Q = 0$ as meaning the orbit must move horizontally left or right; In (b) and (c), note that when an orbit jumps over a nullcline, the lines drawn don't change direction until they get to the next population values.

3.2.5. a. Pick a point with P and Q both large, then $1 - P < 0$, so $rP(1 - P) < 0$ and $-sPQ < 0$, so $\Delta P = rP(1 - P) - sPQ < 0$. Thus arrows point left.
b. Pick a point with P very small, so $-u+vP < 0$ and $\Delta Q = Q(-u+vP) < 0$. Thus arrows point down.
c. Pick a point with P very large, so $-u+vP > 0$ and $\Delta Q = Q(-u+vP) > 0$. Thus arrows point up.

3.2.6. a. The P-nullcline is composed of the Q-axis ($P = 0$) and the line $Q = (-r/sK)P + r/s$. The Q-nullcline is given by the P-axis ($Q = 0$) and $P = Q/(u(1 - e^{-vQ}))$, which can be graphed by computer for specific values of u and v, or analyzed using calculus. This last curve approaches the P-axis at $1/uv$, moving upward and to the right (concave down) , and is asymptotic to $Q = uP$ for large P and Q. For $1/uv < K$, the nullclines and direction arrows produce a figure qualitatively like Figure 3.4, with the vertical line replaced by one curving to the right.
b. Two equilibria are $(0, 0)$, and $(K, 0)$. Assuming $1/uv < K$, there is a third biologically meaningful one that is the solution to the two equations $Q = (-r/sK)P + r/s$ and $P = Q/(u(1 - e^{-vQ}))$. While these can be solved numerically for specific values of the parameters, there is not a simple formula for the solution.

3.2.7. a. Both predator and prey are follow the logistic model in the absence of the other, but the extra terms mean the predator benefits and prey is harmed from the predator-prey interaction.
b. The P-nullcline is as in Figure 3.4. The Q-nullcline is the P-axis ($Q = 0$) together with the line $P = (u/v)(Q - 1)$ which goes through $(0, 1)$ and slopes upward. If $r/s > 1$, the two sloping lines of the nullclines intersect, and produce four regions. If $r/s \leq 1$, there are only three regions. Below $P = (u/v)(Q - 1)$ arrows point up; above it they point down. Above the line $Q = (r/s)(1 - P)$ arrows point to the left; below it they point to the right.
c. Equilibria are at $(0, 0)$, $(1, 0)$ and $((r - s)u/(ru + vs), (u + v)r/(ru + vs))$. The third equilibrium is only biologically meaningful if $r/s > 1$.
d. For $r/s > 1$ you might expect orbits to move counterclockwise around the third equilibrium, provided they begin close enough to it. Whether they spiral inward or outward is not yet clear.

3.3. Linearization and Stability

3.3.1. At $(0,0)$, linearization produces $\begin{pmatrix} p_{t+1} \\ q_{t+1} \end{pmatrix} \approx \begin{pmatrix} 2.3 & 0 \\ 0 & .3 \end{pmatrix} \begin{pmatrix} p_t \\ q_t \end{pmatrix}$. The eigenvalues are 2.3 and .3, so the equilibrium is a saddle and unstable. This is biologically reasonable, since small prey populations with no predators will move away from this equilibrium, while small predator populations with no prey will move toward it.

At $(1,0)$, $\begin{pmatrix} p_{t+1} \\ q_{t+1} \end{pmatrix} \approx \begin{pmatrix} -.3 & -.5 \\ 0 & 1.9 \end{pmatrix} \begin{pmatrix} p_t \\ q_t \end{pmatrix}$. The eigenvalues are $-.3$ and 1.9, so the equilibrium is a saddle and unstable. This is biologically reasonable since if there are no predators, we expect a nearby orbit to move toward this equilibrium, while if there are some predators, it might move away.

Numerical experiments confirm these results.

3.3.2. a. Equilibria are $(0,0)$, $(1,0)$, and $(.05,.19)$.

b. The first two appear to be saddles (so unstable) and the last as unstable.

c. Linearization at $(0,0)$ produces $\begin{pmatrix} p_{t+1} \\ q_{t+1} \end{pmatrix} \approx \begin{pmatrix} 1.8 & 0 \\ 0 & .9 \end{pmatrix} \begin{pmatrix} p_t \\ q_t \end{pmatrix}$. The eigenvalues are 1.8 and .9, so the equilibrium is a saddle and unstable.

Linearization at $(1,0)$ produces $\begin{pmatrix} p_{t+1} \\ q_{t+1} \end{pmatrix} \approx \begin{pmatrix} .2 & -4 \\ 0 & 2.9 \end{pmatrix} \begin{pmatrix} p_t \\ q_t \end{pmatrix}$. The eigenvalues are .2 and 2.9, so the equilibrium is a saddle and unstable.

Linearization at $(.05,.19)$ produces $\begin{pmatrix} p_{t+1} \\ q_{t+1} \end{pmatrix} \approx \begin{pmatrix} .96 & -.2 \\ .38 & 1 \end{pmatrix} \begin{pmatrix} p_t \\ q_t \end{pmatrix}$. The eigenvalues are $.98 \pm \sqrt{.0756}i$, with absolute value approximately 1.0178, so the equilibrium is unstable.

3.3.3. a. Equilibria are $(0,0)$, $(1,0)$, and $(1.167,-2.667)$, so only the first two are biologically meaningful.

b. The first appears to be a saddle (so unstable) and the second appears to be stable.

c. Linearization at $(0,0)$ produces $\begin{pmatrix} p_{t+1} \\ q_{t+1} \end{pmatrix} \approx \begin{pmatrix} 2.6 & 0 \\ 0 & .3 \end{pmatrix} \begin{pmatrix} p_t \\ q_t \end{pmatrix}$. The eigenvalues are 2.6 and .3, so the equilibrium is a saddle and unstable.

Linearization at $(1,0)$ produces $\begin{pmatrix} p_{t+1} \\ q_{t+1} \end{pmatrix} \approx \begin{pmatrix} -.6 & -.1 \\ 0 & .9 \end{pmatrix} \begin{pmatrix} p_t \\ q_t \end{pmatrix}$. The eigenvalues are $-.6$ and .9, so the equilibrium is stable.

3.3.4. The surface of a bump, or mountain top, with the high point being the unstable equilibrium; the surface of a bowl or depression, withe the low point being the stable equilibrium.

3.3.5. Substituting $P_t = P^{\#} + p_t$ and $Q_t = Q^{\#} + q_t$ into the model equations, and then discarding all terms of degree greater than 1, leaves both constant terms and terms of degree 1. The constants prevent the model from being expressed as a simple matrix equation. Rather than getting a linear approximation, we get an affine one.

3.3.6. a. Initial populations of the form $(P_0,0)$, with P_0 small, will move away from the origin, since in the absence of predators, the prey behaves logistically. Initial populations of the form $(0,Q_0)$ will move toward the origin, since in the absence of prey, the predators will die out. Thus the origin must be a saddle equilibrium.

b. Computing the characteristic polynomial of the matrix, and finding its roots, shows the eigenvalues are $1+r$ and $1-u$. Since $r > 0$ and $0 < u < 1$ are positive, $1 + r > 1$ and $0 < 1 - u < 1$, so the equilibrium is a saddle.

3.3.7. a. $\begin{pmatrix} 2.3 - 2.6P_t - .5Q_t & -.5P_t \\ 1.6Q_t & .3 + 1.6P_t \end{pmatrix}$

b. This yields the matrix obtained in the text.

3.3.8. a. $\begin{pmatrix} 1 + r - 2rP - sQ & -sP \\ vQ & 1 - u + vP \end{pmatrix}$

b. At $(0,0)$, $\begin{pmatrix} 1+r & 0 \\ 0 & 1-u \end{pmatrix}$; at $(1,0)$, $\begin{pmatrix} 1-r & -s \\ 0 & 1-u+v \end{pmatrix}$; at $(\frac{u}{v}, \frac{r}{s}(1-\frac{u}{v}))$, $\begin{pmatrix} 1 - \frac{ru}{v} & -\frac{su}{v} \\ \frac{vr}{s}(1-\frac{u}{v}) & 1 \end{pmatrix}$

3.4. Positive and Negative Interactions

3.4.1. Since $(1 - e^{-wQ})$ is 0 when $Q = 0$ and increases toward 1 as Q is increased, the term $sP(1 - e^{-wQ})$ means the per capita benefit each individual in P might receive from the interaction with Q is between 0 and s, and increases with the size of Q.

3.4.2. a. $\begin{pmatrix} L_{t+1} \\ P_{t+1} \\ A_{t+1} \end{pmatrix} = \begin{pmatrix} 0 & 0 & f \\ \tau_{1,2} & 0 & 0 \\ 0 & 1 & \tau_{3,3} \end{pmatrix} \begin{pmatrix} L_t \\ P_t \\ A_t \end{pmatrix}$

b. The factor $e^{-c_{EA}A_t}$ describes cannibalism of eggs by adults; the factor $e^{-c_{PA}A_t}$ describes cannibalism of pupae by adults.

3.4.3. Since $r = u$ and $s < v$ for the first choice of parameters, we expect the immune system to have the advantage. Simulations show it typically 'wins', eliminating the infectious agent, though initially the infectious agent may grow. For the second choice of parameters, we expect the infectious agent to have the advantage. It typically 'wins' in simulations by growing to infinity, unless initially the immune agents are much more plentiful. The first set of parameters gives a more desirable outcome if your own immune response is being modeled.

3.4.4. The P-nullcline is the P-axis and the vertical line $P = r/s$. The Q-nullcline is the P-axis and the vertical line $P = u/v$.

For most parameter choices, the equilibria are all points on the P-axis. If $r/s = u/v$, then all points on the vertical line $P = r/s$ are also equilibria. To the left of the line $P = u/v$ arrows point up, and to the right of it they point down. To the left of the line $P = r/s$ arrows point to the right, and to the right of it they point left. This model can be in equilibrium only if the infectious agent is irradicated, but any amount of immune agent may remain.

For $u/v < r/s$ we might expect orbits to typically move toward an equilibrium. Small amounts of infectious and immune agents might grow for a bit, and then the immune agents might continue to grow while reducing the infectious ones toward 0.

For $u/v > r/s$ we might expect orbits to typically move toward $P = r/s$ as Q goes to infinity.

However, these scenarios are not guaranteed, as orbits may 'jump' by amounts that the direction arrows are not sufficient to predict.

3.4.5. While the concepts of equilibria and stability have some value for this model, they do not play as important a role as in, for example, the predator-prey

model. The precise value of an equilibrium P^* is not important. The main issue for this model is does Q grow to infinity, or shrink to 0?
Interestingly, when an equilibrium value $(P^*, 0)$ is perturbed to (P^*, q) for some small q, it may move to a *different* equilibrium. (What are the eigenvalues of the linearized model when this happens?)

CHAPTER 4

Modeling Molecular Evolution

4.2. An Introduction to Probability

4.2.1. a-c. Answers may vary.
d. Generally, the estimates from experiments with more flips should be better, though that is not always the case.

4.2.2. a. $(\frac{1}{2})^{10} = .0009765625$
b. $(\frac{1}{2})^{10} = .0009765625$
c. People tend to believe (falsely) that a string of ten tails is less likely than any particular run of heads and tails such as that in part (a).

4.2.3. a. $p_A \approx .4$, $p_G \approx .6$, $p_C \approx 0$, $p_T \approx 0$
b. $p_A \approx .4$, $p_G \approx .3$, $p_C \approx .1$, $p_T \approx .2$
c. $p_A \approx .4$, $p_G \approx .2$, $p_C \approx .25$, $p_T \approx .15$
d. There are more G's at the beginning of the sequence and more C's towards the end. The A's and T's are more evenly distributed.

4.2.4. a. $\mathcal{P}(A) \approx .05$, $\mathcal{P}(G) \approx .4$, $\mathcal{P}(C) \approx .3$, $\mathcal{P}(T) \approx .25$
b. $\mathcal{P}(purine) \approx .45$, $\mathcal{P}(pyrimidine) \approx .55$
c. G, a purine, is the most likely base. This may, at first, appear to contradict part (b) which shows the base is most likely to be a pyrimidine. However, there is no real contradiction: While G is the most likely base, the probability of either a C or T is higher than that of an A or G.

4.2.5. a. (F,F,F), (M,F,F), (F,M,F), (F,F,M), (F,M,M), (M,F,M), (M,M,F), (M,M,M), all with probability $(1/2)^3 = .125$
b. $\{(F,F,F),(F,M,F),(F,F,M),(F,M,M)\}$, $4(.125) = .5$
c. $\{(F,M,M),(M,F,M),(M,M,F)\}$, $3(.125) = .375$
d. 'the family is either all male or has at least two daughters', $\{(F,F,F)$, (M,F,F), (F,M,F), (F,F,M), $(M,M,M)\}$, $5(.125) = .625$
e. $\{(F,F,F),(M,F,F),(F,M,F),(F,F,M),(M,M,F),(F,M,M)$, $(M,F,M)\}$, $7(.125) = .875$

4.2.6. a. $\{A,G,C,T\}$, $\{A,G,C\}$, $\{A,G,T\}$, $\{A,C,T\}$, $\{G,C,T\}$, $\{A,G\}$, $\{A,C\}$, $\{A,T\}$, $\{G,C\}$, $\{G,T\}$, $\{C,T\}$, $\{A\}$, $\{G\}$, $\{C\}$, $\{T\}$, $\{\}$
b. To form an event, each of the n possible outcomes is either included or not. Thus picking an event is equivalent to picking n times between the 2 possibilities "include" or "don't include", for a total of 2^n different events.

4.2.7. a. not mutually exclusive, independent
b. mutually exclusive, dependent
c. not mutually exclusive, dependent

4.2.8. Two mutually exclusive events E_1 and E_2 with positive probabilities can not be independent since $0 = \mathcal{P}(E_1 \cap E_2) \neq \mathcal{P}(E_1)\mathcal{P}(E_2) > 0$. More informally, if

the events cannot occur together, then knowing whether one has occurred does give us information as to whether the other has.

4.2.9. $E \cap F = \{\}$ is equivalent to saying both events cannot occur at the same time, since there is no outcome that is in both.

4.2.10. a. If E and F are disjoint, then $\mathcal{P}(E \cap F) = 0$
b. $\mathcal{P}(E_{mult\ 3} \cup E_{<4}) = \mathcal{P}(\{1,2,3,6\}) = \frac{2}{3}$. This equals $\mathcal{P}(E_{mult\ 3}) + \mathcal{P}(E_{<4}) - \mathcal{P}(E_{mult\ 3} \cap E_{<4}) = \frac{1}{3} + \frac{1}{2} - \frac{1}{6}$.

4.2.11. If knowledge of E gives no information about the probability of F occurring, then the same should be true of the complementary events. After all, knowledge of whether E occurred is equivalent to knowledge of whether E' occurred, and similarly for F and F'.

4.2.12. a. $\mathcal{P}(\textit{change},\ \textit{no change}) + \mathcal{P}(\textit{no change},\ \textit{change}) = .02955$
b. (*no change*, *no change*, *nochange*); (*change*, *change*, *no change*); (*change no change*, *change*); (*no change*, *change*, *change*)
c. $4(.985)(.015)^2 = .0008865$

4.3. Conditional Probabilities

4.3.1. a. $\{(F,F),(F,M),(M,F),(M,M)\}$, all with probability .25
b. 3/4
c. 1/2
d. 2/3
e. 1
f. No. Knowledge that one child is female effects the likelihood that the youngest child is female, since $1/2 \neq 2/3$. Alternately, $3/4 \neq 1$ shows that knowledge that the youngest child is female affects the likelihood that one child is female.

4.3.2. a. $\mathcal{P}(E_{odd} \cap E_{\leq 2}) = \mathcal{P}(E_1) = \frac{1}{6} = \mathcal{P}(E_{odd})\mathcal{P}(E_{\leq 2}) = \frac{1}{2} \cdot \frac{1}{3}$
b. $\mathcal{P}(E_{odd} \cap E_{\leq 3}) = \mathcal{P}(E_1 \cup E_3) = \frac{1}{3} \neq \mathcal{P}(E_{odd})\mathcal{P}(E_{\leq 3}) = \frac{1}{2} \cdot \frac{1}{2}$
c. $E_{\leq 2} = \{1,2\}$ and $E'_{\leq 2} = \{3,4,5,6\}$. Both $E_{\leq 2}$ and $E'_{\leq 2}$ contain equal numbers of evens and odds. However, $E_{\leq 3}$ contains two odds and one even, while $E'_{\leq 3}$ contains one odd and two evens. Knowledge of whether the roll is less than or equal to 3 effects the probability that the roll is even or odd.

4.3.3. a. Sensitivity is $\mathcal{P}(+ \text{ result} \mid \text{disease})$; Specificity is $\mathcal{P}(- \text{ result} \mid \text{no disease})$.
b. False positive: $\mathcal{P}(+ \text{ result} \mid \text{no disease})$; False negative: $\mathcal{P}(- \text{ result} \mid \text{disease})$.
c. Sensitivity = 22/30 = .7333; Specificity = 1739/1790 = .9715.

4.3.4. a.

	Healthy Persons	Diseased Persons
Negative Result	98901	1
Positive Result	999	99

b. $\mathcal{P}(\text{Diseased} \mid +) = 99/(999+99) = .0902$. Thus only about 9 of every 100 individuals testing positive actually have the disease, despite the high specificity of the test.

4.3.5. a. The conditional probabilities $\mathcal{P}(S_0 = i \mid S_1 = j)$ below differ from those in the table in the text.

$S_1 \diagdown S_0$	A	G	C	T
A	.778	0	.111	.111
G	.083	.75	.167	0
C	0	.182	.636	.182
T	.125	0	.125	.75

b. $\mathcal{P}(S_1 = i \mid S_0 = j)$ is the conditional probability that given a j in S_0 it mutates to become an i in S_1. However, $\mathcal{P}(S_0 = i \mid S_1 = j)$ is the conditional probability that given a j in the descendent, it came from an i in the ancestor. The first is found by dividing an entry in the table by its column sum, the second by dividing by its row sum.

4.3.6. a. The diagonal entries correspond to no mutation occurring. These are likely to be the largest, since point mutations are rare.
b. Transitions: entries $(1,2)$, $(2,1)$, $(3,4)$, $(4,3)$; Transversions: entries $(1,3)$, $(1,4)$, $(2,3)$, $(2,4)$, $(3,1)$, $(3,2)$, $(4,1)$, $(4,2)$. This table does not support the hypothesis that transitions are more common than transversions.

4.3.7. a. The distribution of bases in S_0 is estimated by $p_A = .225$, $p_G = .275$, $p_C = .275$, $p_T = .225$.
b. The distribution of bases in S_1 is estimated by $p_A = .225$, $p_G = .3$, $p_C = .275$, $p_T = .2$.

4.3.8. a. $\mathcal{P}(S_0 = A) = .225$, $\mathcal{P}(S_0 = G) = .275$, $\mathcal{P}(S_0 = C) = .275$, $\mathcal{P}(S_0 = T) = .225$, $\mathcal{P}(S_1 = A) = .225$, $\mathcal{P}(S_1 = G) = .3$, $\mathcal{P}(S_1 = C) = .275$, $\mathcal{P}(S_1 = T) = .2$.
b. No, since $\mathcal{P}(S_1 = i \text{ and } S_0 = j) \neq \mathcal{P}(S_0 = i)\mathcal{P}(S_1 = j)$. For instance, since $\mathcal{P}(S_1 = i \text{ and } S_0 = j) = (1/40)(\text{the } (j,i) \text{ entry of the table})$, we find $\mathcal{P}(S_1 = A \text{ and } S_0 = A) = 7/40 = .175 \neq (.225)(.225) = .050625$.
c. Since the sequences are related and mutations are rare, the appearance of a particular base at a site in S_0 means it is highly probable that the same base would appear at the same site in S_1, i.e. the events $\{S_0 = i\}$ and $\{S_1 = j\}$ are not independent.

4.3.9. a. Since there is no relationship between the two sequences, knowing information about one should convey nothing about the other.
b. All the columns would be the same.

4.3.10. a. The formula calculates the conditional probability of a purine occurring in S_2 given a purine occurred in S_0 by accounting for either a purine or a pyrimidine occurring in the intermediate sequence S_1.
$\mathcal{P}(S_2 = pyr \mid S_0 = pur) = \mathcal{P}(S_2 = pyr \mid S_1 = pur) \cdot \mathcal{P}(S_1 = pur \mid S_0 = pur) + \mathcal{P}(S_2 = pyr \mid S_1 = pyr) \cdot \mathcal{P}(S_1 = pyr \mid S_0 = pur)$, etc.
b. $\mathcal{P}(S_2 = pur \mid S_0 = pur) = .9606$; $\mathcal{P}(S_2 = pyr \mid S_0 = pur) = .0394$; $\mathcal{P}(S_2 = pur \mid S_0 = pyr) = .0197$; $\mathcal{P}(S_2 = pyr \mid S_0 = pyr) = .9803$
c. Note that with the given assumptions

$$\begin{aligned}
&\mathcal{P}(S_2 = pur \mid S_1 = pur) \cdot \mathcal{P}(S_1 = pur \mid S_0 = pur) \\
&\quad = \mathcal{P}(S_2 = pur \mid S_1 = pur \text{ and } S_0 = pur) \cdot \mathcal{P}(S_1 = pur \mid S_0 = pur) \\
&\quad = \frac{\mathcal{P}(S_2 = pur \text{ and } S_1 = pur \text{ and } S_0 = pur)}{\mathcal{P}(S_1 = pur \text{ and } S_0 = pur)} \cdot \frac{\mathcal{P}(S_1 = pur \text{ and } S_0 = pur)}{\mathcal{P}(S_0 = pur)} \\
&\quad = \frac{\mathcal{P}(S_2 = pur \text{ and } S_1 = pur \text{ and } S_0 = pur)}{\mathcal{P}(S_0 = pur)}.
\end{aligned}$$

Similarly,

$$\mathcal{P}(S_2 = pur \mid S_1 = pyr) \cdot \mathcal{P}(S_1 = pyr \mid S_0 = pur) = \frac{\mathcal{P}(S_2 = pur \text{ and } S_1 = pyr \text{ and } S_0 = pur)}{\mathcal{P}(S_0 = pur)}.$$

Therefore

$$\begin{aligned}&\mathcal{P}(S_2 = pur \mid S_1 = pur) \cdot \mathcal{P}(S_1 = pur \mid S_0 = pur) \\ &\qquad + \mathcal{P}(S_2 = pur \mid S_1 = pyr) \cdot \mathcal{P}(S_1 = pyr \mid S_0 = pur) \\ &= \frac{\mathcal{P}(S_2 = pur,\ S_1 = pur,\ S_0 = pur) + \mathcal{P}(S_2 = pur,\ S_1 = pyr,\ S_0 = pur)}{\mathcal{P}(S_0 = pur)} \\ &= \frac{\mathcal{P}(S_2 = pur,\ S_0 = pur)}{\mathcal{P}(S_0 = pur)} = \mathcal{P}(S_2 = pur \mid S_0 = pur).\end{aligned}$$

4.3.11. a. Given E_1, either E_2 takes place or it does not.
b. $\mathcal{P}(E_2 \mid E_1) + \mathcal{P}(E_2' \mid E_1) = \frac{\mathcal{P}(E_2 \cap E_1)}{\mathcal{P}(E_1)} + \frac{\mathcal{P}(E_2' \cap E_1)}{\mathcal{P}(E_1)} = \frac{\mathcal{P}(E_2 \cap E_1) + \mathcal{P}(E_2' \cap E_1)}{\mathcal{P}(E_1)} = \frac{\mathcal{P}((E_2 \cap E_1) \cup (E_2' \cup E_1))}{\mathcal{P}(E_1)} = \frac{\mathcal{P}(E_1)}{\mathcal{P}(E_1)} = 1$

4.3.12. b. `sum((S0=='C') & (S1=='G'))`
c. `sum((S0=='A') | (S0=='G'))`
d. `sum(((S0=='A') | (S0=='G')) & ((S1=='C') | (S1=='T')))`

4.3.13. a. `colsum` is the column sums, `N` the total sum of the entries in `F`. Since the column sums give the number of times each base occurs is S_0, `colsum/N` gives the fraction of sites with each base in S_0.
b. `rowsum=F*[1;1;1;1], N=[1,1,1,1]*rowsum, p1=rowsum/N`
c. D is a diagonal matrix whose entries represent the number of times the bases appear in S_0, in the order A, G, C, T. Then each entry of $M * D$, the product of the matrix of conditional probabilities and the count data stored in D, represents a frequency count. For example, the $(2, 3)$ entry of $M * D$ is $\mathcal{P}(S_1 = G \mid S_0 = C)(\#$ of C's in $S_0)$, the (average) number of C's in S_0 that mutate to become G's in S_1.

4.4. Matrix Models of Base Substitution

4.4.1. a. A plot in the forest can be in the state "occupied by an A tree" or the state "occupied by a B tree."
b. All the entries are non-negative and the column sums are one.
c. $(1, 1)$ entry: the conditional probability that a spot which is occupied by an A tree in one year remains occupied by an A tree the next year; $(1, 2)$ entry: the conditional probability that a spot which is occupied by a B tree in one year is occupied by an A tree the next year; $(2, 1)$ entry: the conditional probability that a spot which is occupied by an A tree in one year is occupied by an B tree the next year; $(2, 2)$ entry: the conditional probability that a spot which is occupied by a B tree in one year remains occupied by an B tree the next year
d. $(.01, .99)$

4.4.2. The fecundities are often greater than 1 and (regardless of their values) cannot be interpreted as probabilities. While survival coefficients can be interpreted as probabilities, the columns of the matrix generally do not sum to 1.

4.4.3. a. About 27 steps to be within .05; about 67 steps to be within .01.

c. $\mathbf{p}_0$ is unchanged by multiplying by M; it is an equilibrium. Notice that $\mathbf{p}_0$ is an eigenvector of M with eigenvalue 1.

d. The initial vector is drawn towards the stable equilibrium $(.25, .25, .25, .25)$. This $\mathbf{p}_0$ corresponds to an initial sequence comprised entirely of G's.

4.4.4. a. $\alpha = .06$ is faster.

b. Yes.

c. The larger the value of α, the more mutation occurs and the quicker any initial vector $\mathbf{p}_0$ will move towards equilibrium.

4.4.5. Because mutation is rare, the conditional probabilities describing no change should be largest.

4.4.6. a-c. Answers may vary, but the experimentally-determined equilibrium should be very close to an eigenvector with eigenvalue 1.

d. A Markov matrix with all positive entries will always have 1 as its dominant eigenvalue with corresponding eigenvector having all non-negative entries.

4.4.7. $M = \begin{pmatrix} \delta & \gamma & \beta & \gamma \\ \gamma & \delta & \gamma & \beta \\ \beta & \gamma & \delta & \gamma \\ \gamma & \beta & \gamma & \delta \end{pmatrix}$, where $\delta = 1 - 2\gamma - \beta$. This matrix is different since a purine and a pyrimidine have been interchanged, though it represents the same model.

4.4.8. a. The first theorem applies to M, but the second does not since M has some zero entries. (However, since M^2 has all non-zero entries, you can apply the second theorem to it.)

b. $(.1849, .3946, .2819, .1386)$

4.4.9. a. $\mathbf{p}_0 = (.3, .225, .25, .225)$, $M = \begin{pmatrix} .833 & 0 & 0 & .111 \\ .083 & .889 & 0 & 0 \\ 0 & .111 & 1 & .111 \\ .083 & 0 & 0 & .778 \end{pmatrix}$

b. $\mathbf{p}_0$ is reasonable close to $(.25, .25, .25, .25)$. M may seem less close to a Jukes-Cantor matrix than you might expect, because of the variation in the off-diagonal entries. One way to estimate α is to average the off-diagonal entries to estimate $\alpha/3$. This gives $\alpha/3 = .0416$, so $\alpha = .1248$.

4.4.10. a. The Jukes-Cantor model is more appropriate for the pair S_0', S_1', since a particular base seems to mutate to any of the other three bases with roughly the same frequency. Note also that the bases in S_0' are in roughly equal numbers.

b. The Kimura 2-parameter model is more appropriate for the pair S_0, S_1, since the data shows that transitions are more likely than transversions. Note also that the bases in S_0 are in roughly equal numbers.

4.4.11. `u0, u1` were made with a Jukes-Cantor model; `s0, s1` were made with a Kimura 2-parameter model; and `t0, t1` were made with a general Markov model. Both `s0` and `u0` have roughly equal numbers of all bases, while `t0` does not. All mutations are roughly equally likely for `u0, u1`; transitions occur in roughly equal numbers and are more likely than transversions for `s0, s1`; general patterns are hard to recognize in the frequency data for `t0, t1`.

4.4.12. a. The 20×20 Markov matrix would have diagonal entries $1-\alpha$ and off-diagonal entries $\alpha/19$. The initial base distribution would be $\mathbf{p}_0 = (.05, .05, \ldots, .05)$.

b. The general Markov model would have $(19)(20) = 380$ parameters in the matrix.

4.4.13. a. Answers will vary. However, in general the Markov matrix recovered from the two sequences does not look much like the original Jukes-Cantor matrix that was used to create the two sequences. The base distribution vector also varies somewhat from equidistribution.

b. Answers will vary. For a sequence of length 1000, the diagonal entries of the recovered Markov matrix are generally close to .9, though the off-diagonal entries in any column vary a bit about the average of .0333. The base distribution is more closely uniform than for sequences of length 10 or 100. The Markov matrix recovered from sequence data of length 100 usually less closely resembles the true Markov matrix used in creating the data than does the one from longer sequences.

c. Stochastic error in short sequences (i.e., a small number of trials of the same probabilistic process) can hide the true underlying process governing sequence mutation. In long sequences this is less problematic.

4.4.14. Yes. In general, the longer the sequences the better able one is to recover the true Markov matrix underlying the simulation.

4.4.15. a. $\mathbf{p}_0 = (.15, .25, .35, .25)$ is not an equilibrium base distribution for the Jukes-Cantor matrix $M = \begin{pmatrix} .7 & .1 & .1 & .1 \\ .1 & .7 & .1 & .1 \\ .1 & .1 & .7 & .1 \\ .1 & .1 & .1 & .7 \end{pmatrix}$

b. $\mathbf{p}_0 = (.19, .25, .31, .25)$ is not an equilibrium base for the transition matrix $M = \begin{pmatrix} .5526 & .06 & .0484 & .06 \\ .1316 & .7 & .0806 & .1 \\ .1842 & .14 & .7903 & .14 \\ .1316 & .1 & .0806 & .7 \end{pmatrix}$, which is not Jukes-Cantor.

4.4.16. a.The graph of $y = 1 - \frac{4}{3}\alpha$ is a straight line between $(0, 1)$ and $(1, -1/3)$.

b. Since $|1 - \frac{4}{3}\alpha| < 1$ for $0 < \alpha \leq 1$, then $(1 - \frac{4}{3}\alpha)^t \to 0$ as $t \to \infty$. Thus, all the entries of M^t tend to .25 as $t \to \infty$.

c. If $\alpha = 0$ in the Jukes-Cantor model, then no mutation takes place and M is the identity matrix. Thus $M^t = I \to I$ as $t \to \infty$.

4.4.17. You should find experimentally that for large t the columns of M^t are approximately multiples of the dominant eigenvector.

4.4.18. $\alpha_3 = (\alpha_1 + \alpha_2) - \frac{4}{3}\alpha_1\alpha_2$

4.4.19. If the first Kimura 3-parameter matrix has parameters β, γ, and δ and the second has parameters β', γ', and δ', then their product is Kimura 3-parameter with parameters $(1-\beta-\gamma-\delta)\beta' + \beta(1-\beta'-\gamma'-\delta') + \gamma\delta' + \delta\gamma'$, $(1-\beta-\gamma-\delta)\gamma' + \gamma(1-\beta'-\gamma'-\delta') + \beta\delta' + \delta\beta'$, $(1-\beta-\gamma-\delta)\delta' + \delta(1-\beta'-\gamma'-\delta') + \gamma\beta' + \beta\gamma'$.

4.4.20. $\lambda_1 = 1$, $\lambda_2 = 1 - 2\gamma - 2\delta$, $\lambda_3 = 1 - 2\beta - 2\delta$, $\lambda_4 = 1 - 2\beta - 2\gamma$

4.4.21. Since $\mathbf{e}_1 = .25\mathbf{v}_1 + .25\mathbf{v}_2 + .25\mathbf{v}_3 + .25\mathbf{v}_4$, the first column of M^t is

$$.25\begin{pmatrix}1\\1\\1\\1\end{pmatrix} + .25(1-2\gamma-2\delta)^t\begin{pmatrix}1\\1\\-1\\-1\end{pmatrix} + .25(1-2\beta-2\delta)^t\begin{pmatrix}1\\-1\\1\\-1\end{pmatrix} + .25(1-2\beta-2\gamma)^t\begin{pmatrix}1\\-1\\-1\\1\end{pmatrix}$$
$$= \begin{pmatrix}.25+.25(1-2\gamma-2\delta)^t+.25(1-2\beta-2\delta)^t+.25(1-2\beta-2\gamma)^t\\ .25+.25(1-2\gamma-2\delta)^t-.25(1-2\beta-2\delta)^t-.25(1-2\beta-2\gamma)^t\\ .25-.25(1-2\gamma-2\delta)^t+.25(1-2\beta-2\delta)^t-.25(1-2\beta-2\gamma)^t\\ .25-.25(1-2\gamma-2\delta)^t-.25(1-2\beta-2\delta)^t+.25(1-2\beta-2\gamma)^t\end{pmatrix}.$$

4.4.22. a. $1-\alpha$ is the conditional probability that if the base at a site agrees with the original base, then there is no change at the site over the next time step; α is the conditional probability that if the base at a site agrees with the original base then a base substitution away from the original base occurs over the next time step; $\frac{\alpha}{3}$ is the conditional probability that if the base currently disagrees with the original base then a base substitution occurs back to the original base; $1-\frac{\alpha}{3}$ is the conditional probability that if the base currently disagrees with the original base then it continues to disagree, by either not changing from the current base (probability $1-\alpha$), or by changing to another base that still does not agree with the original base (probability $\frac{2\alpha}{3}$).

b. $(1/4, 3/4)$

c. $\lambda = 1-4\alpha/3$ with eigenvector $(1,-1)$.

d. Since $\begin{pmatrix}q_t\\p_t\end{pmatrix} = M^t\begin{pmatrix}1\\0\end{pmatrix}$, and $\begin{pmatrix}1\\0\end{pmatrix} = 1\begin{pmatrix}1/4\\3/4\end{pmatrix} + \frac{3}{4}\begin{pmatrix}1\\-1\end{pmatrix}$, then $q_t = \frac{1}{4} + \frac{3}{4}(1-\frac{4\alpha}{3})^t$ and $p_t = \frac{3}{4} - \frac{3}{4}(1-\frac{4\alpha}{3})^t$.

4.4.23. a. The expression $(1-\alpha)q_t$ is the probability that at time t the base at a site agrees with the original base and does not mutate by time $t+1$, and the expression $\frac{\alpha}{3}(1-q_t)$ is the probability that it is a different base from the original and mutates back to the original base at time $t+1$. $q_0 = 1$.

b. $q^* = 1/4$, as is expected from other developments of the Jukes-Cantor model.

c. Substituting yields

$$q^* + \epsilon_{t+1} = \frac{\alpha}{3} + \left(1-\frac{4\alpha}{3}\right)(q^*+\epsilon_t) \implies$$
$$q^* + \epsilon_{t+1} = \left[\frac{\alpha}{3} + \left(1-\frac{4\alpha}{3}\right)q^*\right] + \left(1-\frac{4\alpha}{3}\right)\epsilon_t \implies$$
$$\epsilon_{t+1} = \left(1-\frac{4\alpha}{3}\right)\epsilon_t.$$

d. Since $q_0 = 1$, $\epsilon_0 = 3/4$.

e. $\epsilon_t = (1-\frac{4\alpha}{3})^t\epsilon_0$

f. $q_t = q^* + e_t = \frac{1}{4} + \frac{3}{4}(1-\frac{4\alpha}{3})^t$

4.5. Phylogenetic Distances

4.5.1. .1367

4.5.2. a. The Jukes-Cantor distance is .1102158097.

b. The Kimura 2-parameter distance is .1102165081.

c. Since the two distance calculations agree to several decimal spaces, we might hypothesize (if the problem had not already told us) that the data is fit reasonably well by the Jukes-Cantor model. Notice that the Kimura 2-parameter distance reports more mutations (including hidden mutations) than the Jukes-Cantor distance.

4.5.3. a. .2224580274

b. .2308224444

c. The Kimura 2-parameter distance is probably a better choice (assuming we did not already know that the sequences were created with the Kimura 2-parameter model). The frequency table shows a definite pattern of more transitions than transversions. Notice too that the distances differ in the second decimal position.

4.5.4. Jukes-Cantor simulation: $d_{K3} = .1104707856$ and $d_{LD} = .1105916542$. Notice these are about the same as the Jukes-Cantor distance, since that model is a special case of the more general ones.

Kimura 2-parameter simulation: $d_{K3} = .2308544863$ and $d_{LD} = .2337622488$. Notice these are about the same as the Kimura 2-parameter distance, since that model is a special case of the more general ones.

4.5.5. Graph the Jukes-Cantor distance on a graphing calculator or computer.

a. If the sequences are identical, then $p = 0$. This means the Jukes-Cantor distance is $-.75\log(1) = 0$.

b., c. Mathematically, if two sequences differ in more than 3/4 of the sites, then $p > 3/4$. Then the Jukes-Cantor distance formula requires taking the logarithm of a negative number, which is impossible. This is not a limitation with real data. If we took two sequences that were in no way related, we would expect that about 1/4 of the sites agree and about 3/4 of the sites disagree, since with a uniform distribution of bases about 25% of the time the two sequences should agree if everything is chosen at random. For related sequences the formulas for the Jukes-Cantor model derived in the last section show p is at most 3/4, and in practice p is usually much less than 3/4. Notice that the Jukes-Cantor distance gets huge as the values of p get close to .75. This is desirable, since distances should be large when comparing sequences that appear almost unrelated.

4.5.6.

$$p = \frac{3}{4} - \frac{3}{4}(1 - \frac{4}{3}\alpha)^{t} \implies (1 - \frac{4}{3}p) = (1 - \frac{4}{3}\alpha)^{t} \implies$$
$$\ln(1 - \frac{4}{3}p) = t\ln(1 - \frac{4}{3}\alpha) \implies t = \frac{\ln(1 - \frac{4}{3}p)}{\ln(1 - \frac{4}{3}\alpha)}.$$

4.5.7. Substituting $(1-q)$ for p yields $d_{JC} = -\frac{3}{4}\ln(1 - \frac{4}{3}p) = -\frac{3}{4}\ln(1 - \frac{4}{3}(1-q)) = -\frac{3}{4}\ln(\frac{4}{3}q - \frac{1}{3}) = -\frac{3}{4}\ln(\frac{4q-1}{3})$.

4.5.8. Some numerical comparisons are given in the table below. The graphs of $y = \ln(1+x)$ and $y = x$ are very close when x is around 0. In fact, they are tangent to one another at the point $(0, 0)$.

x	$-.1$	$-.01$	$-.001$	$-.0001$
$\ln(1+x)$	$-.1053605157$	$-.0100503359$	$-.0010005003$	$-.0001000050$
x	.0001	.001	.01	.1
$\ln(1+x)$	.0000999950	.0009995003	.0099503309	.0953101798

4.5.9. Since $f'(x) = \frac{1}{1+x}$, the slope of the tangent line is $f'(0) = 1$. The tangent line passes through the point $(0, 0)$. Using the point-slope formula, the equation of the tangent line to $f(x) = \ln(1+x)$ is $g(x) = x$. Thus, for values of x near 0, $f(x) \approx g(x) = x$.

4.5.10. $d(S_0, S_1) + d(S_0, S_2) = d(S_1, S_0) + d(S_0, S_2)$ by symmetry. By additivity, this equals $d(S_1, S_2)$.

4.5.11. Two transitions at a particular site will result in a return to the original base and thus a hidden mutation. For example, $A \to G \to A$ is a hidden mutation. Two transversions may produce a hidden mutation, but often don't (e.g., $A \to C \to G$). If transitions are more likely than transversions, then hidden mutations are more likely.

4.5.12. a. If there are a lot of point mutations at a site, then hidden mutations are more likely. Thus p, the proportion of observed point mutations, is an underestimate of the true proportion of point mutations.

b. When p is small, few point mutations are observed. Since little mutation is observed, it's reasonable to assume little occurred, and therefore that few mutations have been hidden. Thus p should be a good estimate of the proportion of point mutations.

4.5.13. The Kimura 3-parameter distance is given by $d_{K3} = -\frac{1}{4}(\ln(1-2\beta-2\gamma) - \ln(1-2\beta-2\delta) - \ln(1-2\beta-2\gamma))$. Substituting $\alpha/3$ for β, γ, and δ gives

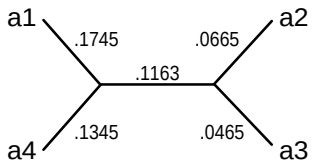
$$\begin{aligned} d &= -\frac{1}{4}(\ln(1-2\alpha/3-2\alpha/3) - \ln(1-2\alpha/3-2\alpha/3) - \ln(1-2\alpha/3-2\alpha/3)) \\ &= \frac{1}{4}(3\ln(1-4\alpha/3)) = d_{JC}. \end{aligned}$$

4.5.14. The distance from the Jukes-Cantor formula is not equal to .4 and may occasionally even be quite far away. Lots of factors are responsible for the discrepancy: the length of the sequences is relatively short; simulated data is always an imperfect reflection of the underlying model; the larger p is, the greater effect a small variation in it has on the reconstructed value of αt; etc.

4.5.15. a. The Jukes-Cantor distances are given in the table below.

	a1	a2	a3	a4
a1		.3721	.3648	.3091
a2			.1125	.2958
a3				.2763

b. Answers will vary. One possibility is



This tree was constructed by observing that a2 and a3 were closest and so perhaps should have an immediate common ancestor. (This also means a2 and

a4 must be joined.) Then, from looking at the data, a1 seemed to be furthest from the other three taxa, so instead of dividing the $d(a1, a4)$ in half, more length was assigned to the edge leading to a1 and a little less than half to a4. Similarly, for the other neighbors. Finally, the length on the internal edge is an average of the differences between the distances in the table and the length assigned to the tree thus far. Clearly, this method is *ad hoc*. A distance table constructed from this tree is given below, for comparison purposes.

	a1	a2	a3	a4
a1		.3573	.3373	.309
a2			.113	.3173
a3				.2973

4.5.16. a. Substituting γ for δ into the Kimura 3-parameter distance formula gives

$$\begin{aligned} d &= -\frac{1}{4}\left(\ln(1-2\beta-2\gamma) - \ln(1-2\beta-2\gamma) - \ln(1-2\gamma-2\gamma)\right) \\ &= \frac{1}{2}\ln(1-2\beta-2\gamma) - \frac{1}{4}\ln(1-4\gamma). \end{aligned}$$

b. The parameter β represents the probability that a transition takes place at a site. Estimating this with p_1, the observed fraction of sites with transitions, is quite reasonable. In the general Kimura 3-parameter model, the parameters γ and δ represent the probability that a transversion of a particular type takes place at a site and $\gamma+\delta$ is the probability of a transversion of any γ. Counting the fraction of sites with observed transversions, of all types, estimates the probability of a transversion, or 2γ in terms of model parameters. Thus, $d_{K2} = \frac{1}{2}\ln(1-2\beta-2\gamma) - \frac{1}{4}\ln(1-4\gamma) \approx \frac{1}{2}\ln(1-2p_1-p_2) - \frac{1}{4}\ln(1-2p_2)$.

4.5.17. a.

$$\begin{aligned} &1-2\beta'-2\delta' = \\ &1-2\left(\frac{1}{4}+\frac{1}{4}(1-2\gamma-2\delta)^t - \frac{1}{4}(1-2\beta-2\delta)^t - \frac{1}{4}(1-2\beta-2\gamma)^t\right) \\ &\qquad -2\left(\frac{1}{4}-\frac{1}{4}(1-2\gamma-2\delta)^t - \frac{1}{4}(1-2\beta-2\delta)^t + \frac{1}{4}(1-2\beta-2\gamma)^t\right) \\ &\qquad\qquad = (1-2\beta-2\delta)^t. \end{aligned}$$

The other two derivations are similar.

b. From (a),

$$\begin{aligned} &\ln(1-2\beta'-2\delta') + \ln(1-2\beta'-2\gamma') + \ln(1-2\gamma'-2\delta') \\ &\qquad = \ln(1-2\beta-2\delta)^t + \ln(1-2\beta-2\gamma)^t + \ln(1-2\gamma-2\delta)^t \\ &\qquad = t\left(\ln(1-2\beta-2\delta) + \ln(1-2\beta-2\gamma) + \ln(1-2\gamma-2\delta)\right). \end{aligned}$$

c. Using the approximation $\ln(1+x) \approx x$ gives

$$\begin{aligned} &\ln(1-2\beta'-2\delta') + \ln(1-2\beta'-2\gamma') + \ln(1-2\gamma'-2\delta') \\ &\qquad \approx t\left((-2\beta-2\delta) + (-2\beta-2\gamma) + (-2\gamma-2\delta)\right) \\ &\qquad = -4t(\beta+\gamma+\delta). \end{aligned}$$

d. The sum $\beta+\gamma+\delta$ is the sum of the probabilities of all three types of point mutations. If these parameters are considered rates, with units (number of base

substitutions per site per unit time), then this sum can be interpreted as the total rate of base substitution.

Thus $d_{K3} \approx t(\beta + \gamma + \delta)$, the product of the elapsed time with the total rate of base substitution, gives a measure of the total amount of mutation.

4.5.18. a. Since the base distribution is uniform and the Jukes-Cantor model exactly fits the data, the N sites of S_0 must contain an equal number of each of the four bases. Thus $\mathbf{f}_0 = (N/4, N/4, N/4, N/4)$.

Recall that the first row of $M(\alpha)$ is $\begin{pmatrix}\mathcal{P}_{A|A} & \mathcal{P}_{A|G} & \mathcal{P}_{A|C} & \mathcal{P}_{A|T}\end{pmatrix}$. Thus, in F, the (A, A) entry, namely the number of A's in S_0 that remain A's in S_1, is the product of $\mathcal{P}_{A|A}$ with the number of A's occurring in S_0, or $\frac{N}{4}\mathcal{P}_{A|A}$. Similarly, the (A, G) entry of F, the number of G's in S_0 that become A's in S_1, is $\frac{N}{4}\mathcal{P}_{A|G}$, where $N/4$ represents the number of G's in S_0. Similar reasoning gives the values $\frac{N}{4}\mathcal{P}_{A|C}$ and $\frac{N}{4}\mathcal{P}_{A|T}$ to complete the entries of the first row of F. The other three rows can be computed similarly. Thus, multiplying each entry of $M(\alpha)$ by $N/4$ gives F.

b. That $\mathbf{f}_0 = (N/4, N/4, N/4, N/4)$ was explained in the solution to part (a). Since a Jukes-Cantor matrix has $(1, 1, 1, 1)$ as an eigenvector, if the sequence S_0 has an equal number of each base, so will S_1. Thus the same explanation shows why $\mathbf{f}_1 = (N/4, N/4, N/4, N/4)$.

c. First calculate that $g_0 = g_1 = \left(\frac{N}{4}\right)^4$ since all four entries of $\mathbf{f}_0$ and $\mathbf{f}_1$ equal $\frac{N}{4}$, and recall that the eigenvalues of a Jukes-Cantor matrix are 1 and a triple eigenvalue $(1 - \frac{4}{3}\alpha)$. Then

$$\begin{aligned}
d_{LD}(S_0, S_1) &= -\frac{1}{4}\left(\ln\left(\det(F)\right) - \frac{1}{2}\ln\left(g_0 g_1\right)\right) \\
&= -\frac{1}{4}\left(\ln\left(\left(\frac{N}{4}\right)^4 \det(M(\alpha))\right) - \frac{1}{2}\ln\left(\left(\frac{N}{4}\right)^4\left(\frac{N}{4}\right)^4\right)\right) \quad (a) \\
&= -\frac{1}{4}\ln\left(\left(\frac{N}{4}\right)^4 (1)\left(1 - \frac{4}{3}\alpha\right)^3\right) + \frac{1}{8}\ln\left(\left(\frac{N}{4}\right)^8\right) \qquad (b) \\
&= -\ln\left(\frac{N}{4}\right) - \frac{1}{4}\ln\left(1 - \frac{4}{3}\alpha\right)^3 + \ln\left(\frac{N}{4}\right) \\
&= -\frac{3}{4}\ln\left(1 - \frac{4}{3}\alpha\right) = d_{JC}(S_0, S_1).
\end{aligned}$$

Equality (a) uses fact (i) and equality (b) uses fact (ii).

4.5.19. As in the last problem, we can show that for the Kimura 3-parameter model $\mathbf{f}_0 = \mathbf{f}_1 = (N/4, N/4, N/4, N/4)$, and $F = \frac{N}{4}M(\beta, \gamma, \delta)$. Since the eigenvalues of $M(\beta, \gamma, \delta)$ are 1, $1 - 2\beta - 2\gamma$, $1 - 2\beta - 2\delta$, and $1 - 2\gamma - 2\delta$. Then

$$\begin{aligned}
&d_{LD}(S_0, S_1) = -\frac{1}{4}\left(\ln\left(\det(F)\right) - \frac{1}{2}\ln\left(g_0 g_1\right)\right) \\
&= -\frac{1}{4}\left(\ln\left(\left(\frac{N}{4}\right)^4 \det(M(\beta, \gamma, \delta))\right) - \frac{1}{2}\ln\left(\left(\frac{N}{4}\right)^4\left(\frac{N}{4}\right)^4\right)\right) \\
&= -\frac{1}{4}\ln\left(\left(\frac{N}{4}\right)^4 (1)(1 - 2\beta - 2\gamma)(1 - 2\beta - 2\delta)(1 - 2\gamma - 2\delta)\right) + \frac{1}{8}\ln\left(\left(\frac{N}{4}\right)^8\right).
\end{aligned}$$

Routine algebra simplifies this to the formula for d_{K3} in problem 4.5.17.

4.5.20. a. Suppose F is the frequency table for evolution from a sequence S_0 to S_1 and $\mathbf{f}_0$ and $\mathbf{f}_1$ are the base distributions of S_0 and S_1 respectively. If we interchange the ancestor and descendent sequences, then the frequency table for evolution from S_1 to S_0 is the transpose F^T. Thus, $d_{LD}(S_1, S_0) = -\frac{1}{4}(\ln(\det(F^T)) - \frac{1}{2}\ln(\det(g_1 g_0)) = -\frac{1}{4}(\ln(\det(F)) - \frac{1}{2}\ln(\det(g_0 g_1))$ which is $d_{LD}(S_1, S_0)$.
b. Each entry of the product $M_{0\to 1}\mathbf{p}_0$ results from multiplying a row of $M_{0\to 1}$ by $\mathbf{p}_0 = (p_A, p_G, p_C, p_T)$. For example, the third entry is

$$\begin{pmatrix}\mathcal{P}_{C|A} & \mathcal{P}_{C|G} & \mathcal{P}_{C|C} & \mathcal{P}_{c|T}\end{pmatrix}\begin{pmatrix}p_A\\ p_G\\ p_C\\ p_T\end{pmatrix} = \mathcal{P}_{C|A}p_A + \mathcal{P}_{C|G}p_G + \mathcal{P}_{C|C}p_C + \mathcal{P}_{c|T}p_T,$$

which is the probability that a C occurs in the sequence S_1.
c.

$$\begin{aligned} N_{1\to 2}N_{0\to 1} &= D_2^{-1}M_{1\to 2}D_1D_1^{-1}M_{0\to 1}D_0 = D_2^{-1}M_{1\to 2}M_{0\to 1}D_0 \\ &= D_2^{-1}M_{0\to 2}D_0 = N_{0\to 2}. \end{aligned}$$

Taking determinants and natural logarithms yields

$$\begin{aligned} \ln(\det(N_{1\to 2}N_{0\to 1})) &= \ln(\det(N_{0\to 2})) \\ \ln(\det(N_{1\to 2})) + \ln(\det(N_{0\to 1})) &= \ln(\det(N_{0\to 2})). \end{aligned}$$

d. Note

$$D_j N_{i\to j} D_i = D_j D_j^{-1} M_{i\to j} D_i D_i = M_{i\to j} D_i^2.$$

The (m,n) entry of this product is the (m,n) entry of $M_{i\to j}$ times the nth diagonal entry of D_i^2. Thus it is $\mathcal{P}(S_j = m \mid S_i = n)\mathcal{P}(S_i = n) = \mathcal{P}(S_j = m \text{ and } S_i = n)$. This is precisely the (m,n) entry of the relative frequency array comparing S_i to S_j.
Taking determinants and natural logarithms gives $\ln(\det(G)) = \ln(\det(D_j)) + \ln(\det(N_{i\to j})) + \ln(\det(D_i)) = \ln(\det(N_{i\to j})) + \ln(\det(D_i)) + \ln(\det(D_j))$.
e. Observe that $g_s = \det(ND_s^2)$ by fact (iii). Thus $g_s = N^4\det(D_s^2)$ by fact (i) of problem 4.5.18. Thus $g_s = N^4\det(D_s)^2$ by fact (ii), and so

$$\begin{aligned} d_{LD}(S_i, S_j) &= -\frac{1}{4}\left(\ln(\det(NG_{i\to j})) - \frac{1}{2}\ln(g_i g_j)\right) \\ &= -\frac{1}{4}\left(\ln(N^4\det(G_{i\to j})) - \frac{1}{2}\ln(N^8\det(D_i)^2\det(D_j)^2\right) \\ &= -\frac{1}{4}(4\ln N + \ln(\det(N_{i\to j}) + \ln(\det(D_i)) + \ln(\det(D_j)) \\ &\qquad - 4\ln N - \ln(\det D_i) - \ln(\det D_j)) \\ &= -\frac{1}{4}\ln(\det(N_{i\to j})). \end{aligned}$$

Using this formula,

$$\begin{aligned} d_{LD}(S_0, S_2) &= -\frac{1}{4}\ln(\det(N_{0\to 2})) \\ &= -\frac{1}{4}\ln(\det(N_{0\to 1}N_{1\to 2})) \\ &= -\frac{1}{4}\ln(\det(N_{0\to 1})\det(N_{1\to 2})) \\ &= -\frac{1}{4}\left(\ln(\det(N_{0\to 1})) + \ln(\det(N_{1\to 2}))\right) \\ &= d_{LD}(S_0, S_1) + d_{LD}(S_1, S_2). \end{aligned}$$

CHAPTER 5

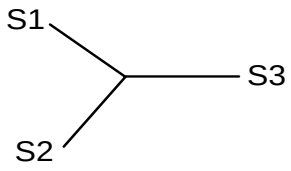
Constructing Phylogenetic Trees

5.1. Phylogenetic Trees

Warning: Trees are not drawn to scale.

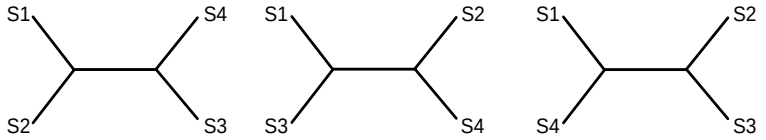
5.1.1. a. $\{T_2, T_3\}$
b. $\{T_2, T_3, T_5\}$
c. $\{T_1, T_6\}$, $\{T_2, T_3, T_5\}$
d. $\{T_1, T_2, T_3, T_4, T_5, T_6\}$
e. T_4, T_6

5.1.2. a.

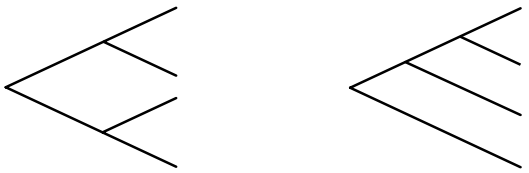


b. In the tree in part (a), the root can be placed along the edge joining the internal node to S1, S2, or S3.

5.1.3. a.



b. In each of the three trees in part (a), the root can be located on any of the five edges.
Equivalently, for the tree below on the left there are three distinct labelings (from top to bottom) of the leaves: {S1, S2, S3, S4}, {S1, S3, S2, S4}, {S1, S4, S2, S3}, and for the tree on the right there are twelve distinct labelings: {S1, S2, S3, S4}, {S1, S2, S4, S3}, {S1, S3, S2, S4}, {S1, S3, S4, S2}, {S1, S4, S2, S3}, {S1, S4, S3, S2}, {S2, S3, S1, S4}, {S2, S3, S4, S1}, {S2, S4, S1, S3}, {S2, S4, S3, S1}, {S3, S4, S1, S2}, {S3, S4, S2, S1}.



5.1.4.

n	3	4	5	6	7	8	9	10
$\frac{(2n-5)!}{2^{(n-3)}(n-3)!}$	1	3	15	105	945	10395	135135	2027025

5.1.5.

n	2	3	4	5	6	7	8	9	10
$\frac{(2n-3)!}{2^{(n-2)}(n-2)!}$	1	3	15	105	945	10395	135135	2027025	34459425

5.1.6. a. When we add a new edge to an existing tree, the edge count increases by two: one for the new edge, and one more since an existing edge is split into two edges where the new one is attached.

b. By part (a), each time we add an edge the edge count increases by two. Thus, we see the pattern:

n	2	3	4	5	$\cdots$	n
e	1	3	5	7	$\cdots$	$2n-3$

Alternatively, $e = 1+2(n-2)$ counts 1 edge for the first two terminal vertices, plus 2 edges for each of the other $(n-2)$ terminal vertices successively attached to the tree.

c. An unrooted tree with n terminal vertices has $2n-3$ edges. To create an unrooted tree with $n+1$ terminal vertices from such a tree, a new edge with the new terminal vertex can be attached to any of the $2n-3$ existing edges. Thus, if there are m unrooted trees with n terminal vertices, there are $m(2n-3)$ unrooted trees with $n+1$ terminal vertices.

d. Iterating the result of part (c) gives the formula:

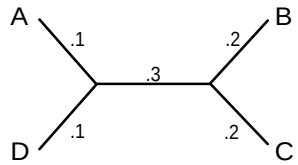
n	no. unrooted trees
2	1
3	$1(2\cdot 2-3)=1$
4	$1(2\cdot 3-3)=1\cdot 3$
5	$1\cdot 3\cdot(2\cdot 4-3)=1\cdot 3\cdot 5$
$\vdots$	$\vdots$
n	$(1)(3)(5)\cdots(2n-5)$

e. The denominator contains as factors all the even numbers between 2 and $(2n-6)$, canceling out the even numbers in $(2n-5)!$

f. Imagine a rooted tree with n terminal vertices. Attaching a new edge at the root location creates an unrooted tree with $n+1$ terminal vertices.

5.1.7. The most accurate estimate, produced by writing a brief computer program to find the product, is 4.89×10^{296}.

5.1.8. a. Approaches may vary, but the only tree fitting the data is:



b. There is no way to determine the root, without making some additional assumptions. If you assume a molecular clock, then the root belongs on the central branch, .2 away from the node joining A and D.

5.2. Tree Construction: Distance Methods – Basics

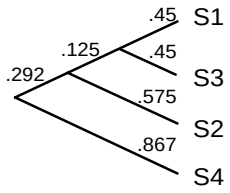
Warning: Trees are not drawn to scale.

5.2.1.

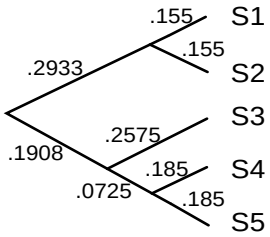
	S1	S2	S3	S4
S1		.425	.27	.55
S2			.425	.55
S3				.55

While the distance between the first two taxa to be joined, $d(\text{S1}, \text{S3})$, agrees exactly with the original distance table, the other distances are only close to the original distances. The duplication of some table entries reflects the molecular clock hypothesis, since certain subsets of taxa will be equidistant from a common ancestor.

5.2.2.



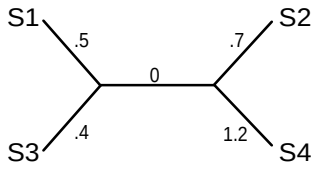
5.2.3.



Topologically, the rooted UPGMA tree is the same as the unrooted FM tree. However, the metric distances are not the same; you can see the molecular clock hypothesis at work in the UPGMA tree.

5.2.4. a. There are several algebraic approaches: Either *ad hoc* algebra or methodical elimination of variables can be used, or matrix algebra. The nicest solution (since it makes the formulas memorable) is a geometric one: $d_{AB}+d_{AC}$ includes the edge x twice, and the edges y and z once, so subtracting d_{BC} gives $2x$, etc.
b. $x = .555$, $y = .079$, $z = .772$

5.2.5.



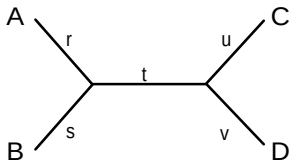
Topologically, the trees are the same as unrooted trees. They are not the same metrically. Note, for instance, FM assigns a branch length of 0 to the internal edge, while UPGMA assigns .125.

5.2.6. a. In order for a molecular clock hypothesis to hold, all the terminal vertices would have to be equidistant from the root. This is impossible. The root cannot be placed at the internal node since the edge lengths are different. Moreover, the root cannot be placed on any of the three edges since no two of the edges have the same length.

b. Since the two shortest edge lengths are equal to .1, it is possible to assume a molecular clock. The root would have to be placed on the edge of length .2 at a distance of .05 from the internal node. Then all terminal vertices are .15 from the root.

c. Here two of the edge lengths are equal, but their length .2 is larger than the length of the third edge. This means it is not possible to locate the root on either of the longer edges nor the shorter edge and achieve equal distances from the root. Of course, the internal node could not serve as a root either, if a molecular clock is to be assumed.

5.2.7. a.



b. $d_{AB} = r+s$, $d_{AC} = r+t+u$, $d_{AD} = r+t+v$, $d_{BC} = s+t+u$, $d_{BD} = s+t+v$, $d_{CD} = u+v$; As this is a system of six equations in only five unknowns, in general there will not be a solution.

c. Answers may vary; one possibility follows. For the distances $d_{AB} = .2$, $d_{AC} = .3$, $d_{AD} = 1.33$, $d_{BC} = .29$, $d_{BD} = 1.3$, $d_{CD} = 1.19$, the system does not have a solution, whereas for the distances $d_{AB} = .17$, $d_{AC} = .32$, $d_{AD} = 1.33$, $d_{BC} = .29$, $d_{BD} = 1.3$, $d_{CD} = 1.19$, the system has a solution.

5.2.8. a. For calculating these measures of errors, the length b was assigned to zero.

	s_{FM}	s_F	s_{TNT}
FM tree	.4699	.6370	.2592
UPGMA tree	.4933	.8968	.3515

The FM tree is a better fit to the data according to all three of these measures.

b. All of these formulas give 0 if a tree exactly fits the data.

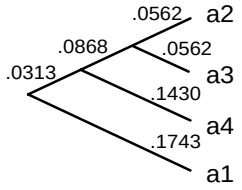
s_F simply sums the absolute value of the difference between the tree lengths and the original distance data. The absolute value prevents negative differences from canceling with positive ones. All deviations of tree lengths from the data are treated identically.

s_{TNT} sums the squares of the differences between tree lengths and distance data (again preventing cancelation), then takes the square root. This is reminiscent of the formula for standard deviation. This measure penalizes large differences more than s_F does, while weighing small differences less: If $|d_{ij} - e_{ij}| < 1$, then $(d_{ij} - e_{ij})^2$ is even smaller, while if $|d_{ij} - e_{ij}| > 1$, then $(d_{ij} - e_{ij})^2$ is larger.

s_{FM} measures the differences between tree lengths and data as proportions. Other than that, it is similar to s_{TNT} in that in penalizes large differences

in the proportions much more than small ones. When tree edge lengths vary greatly in size, s_{FM} will, unlike the other two measures, not allow greater proportional errors in the short edges than the long ones.

5.2.9. a.

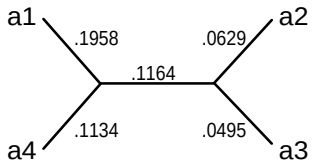


b.

	a1	a2	a3	a4
a1		.3486	.3486	.3486
a2			.1124	.2860
a3				.2860

From the table, the distance between the first pair of taxa joined, a2 and a3, agrees with the original distance data (up to rounding error). The other distances approximate the original distances, and in fact represent averages, with duplication occurring because of the molecular clock hypothesis. Since we know the sequences were created assuming a molecular clock hypothesis, however, we should assume that this tree more accurately reflects the relationships between the sequences than a FM tree does.

5.2.10. a.



b.

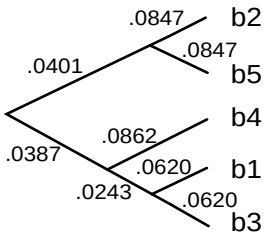
	a1	a2	a3	a4
a1		.3751	.3617	.3092
a2			.1124	.2927
a3				.2793

The distances in this table match up reasonably well with the original distances. In particular, since FM joins a pair of taxa at each step, you find several matches (within rounding error) between the distance tables.

Even though the FM tree produced here appears to match the distance table better than the UPGMA tree of the last problem, since these sequences were created with a molecular clock hypothesis, the tree FM produces should not be preferred to the UPGMA tree. This is an example of *overfitting* the data, by using a more general approach than actually best describes the simulated evolution.

This is analogous to an issue raised in Chapter 8 while studying the method of least squares: while it is possible to fit a degree five polynomial exactly to six data points, if the points are approximately linearly related, this may disguise the true trend of the data.

5.2.11. a.

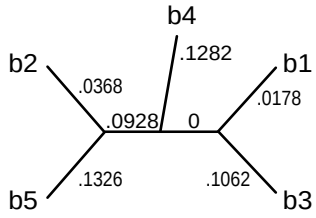


b.

	b1	b2	b3	b4	b5
b1		.2497	.1240	.1724	.2497
b2			.2497	.2497	.1694
b3				.1724	.2497
b4					.2497

The distances computed from the UPGMA tree are not in very close agreement with the data. Since we know that the data was produced without a molecular clock, we should not expect UPGMA to be able to fit the data well. UPGMA makes an assumption that we know is incorrect.

5.2.12. a.



b.

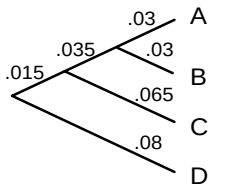
	b1	b2	b3	b4	b5
b1		.1474	.1240	.1460	.2432
b2			.2358	.2578	.1694
b3				.2344	.3316
b4					.3536

The tree distances agree reasonably well with the distance data. If we believe a molecular clock hypothesis is invalid, then the FM tree might be a better reconstruction of evolutionary relationships than the UPGMA tree. (Recall however, that they differ only metrically, not topologically.)

5.2.13. a. There is no way that all four taxa can be equidistant from a root: Since A and C are not equidistant from the internal node to which they are both joined, if a molecular clock is assumed, the root would have to be on the edge leading to taxon C. If this were the case, then it would be impossible for B and D to be equidistant from the root.

b.

	A	B	C	D
A		.06	.12	.14
B			.14	.12
C				.22



Note: If the labels C and D are exchanged in this tree, then the resulting tree would also be a UPGMA tree, since there were two minima in the collapsed data table.

c. Notice that the UPGMA tree constructed has the wrong topological structure. Since A and B are closest in distance, they are joined first by UPGMA, even though this results in the wrong topology. Neighbor joining, introduced in the next section, will not make this mistake.

d. FM creates the same topological tree as UPGMA, so it too will construct an incorrect tree topology.

5.3. Tree Construction: Distance Methods – Neighbor Joining

Warning: Trees are not drawn to scale.

5.3.1. a. Let G the group of all taxa other than Si and Sj. Then

$$\begin{aligned} d(\mathrm{S}i, V) &= \frac{1}{2}(d(\mathrm{S}i, G) + d(\mathrm{S}i, \mathrm{S}j) - d(\mathrm{S}j, G)) \\ &= \frac{(\sum_{k\neq j} d(\mathrm{S}i, \mathrm{S}k))}{2(N-2)} + \frac{d(\mathrm{S}i, \mathrm{S}j)}{2} - \frac{(\sum_{l\neq j} d(\mathrm{S}j, \mathrm{S}l))}{2(N-2)} \\ &= \frac{d(\mathrm{S}i, \mathrm{S}j)}{2} + \frac{R_i - R_j}{2(N-2)}. \end{aligned}$$

Notice the $d(\mathrm{S}i, \mathrm{S}j)$ terms cancel in the expression $R_i - R_j$ to make the last equality hold.

b. This follows from determining edge lengths for the three-taxa tree with Sk, Si, Sj as terminal nodes and V as internal vertex.

5.3.2. a. $R_1 = 1.52$, $R_2 = 2.52$, $R_3 = 1.48$, $R_4 = 1.86$, and M is given by the table:

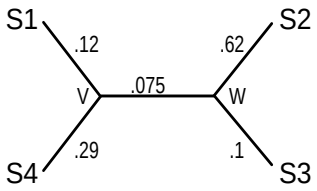
	S2	S3	S4
S1	−2.38	−2.44	−2.56
S2		−2.56	−2.44
S3			−2.38

b. $d(\mathrm{S1}, V) = .12$, $d(\mathrm{S4}, V) = .29$

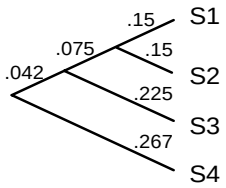
c. $d(\mathrm{S2}, V) = .695$, $d(\mathrm{S3}, V) = .175$

d. $d(\mathrm{S2}, W) = .62$, $d(\mathrm{S3}, W) = .1$, $d(V, W) = .075$

e.



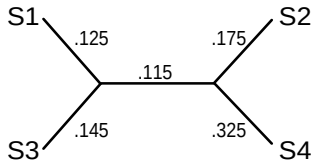
5.3.3. a.



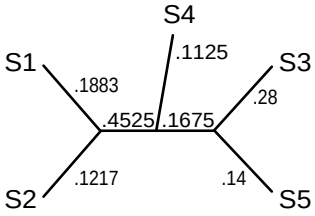
The UPGMA tree does not recover the correct topology. Note: Another UPGMA tree has taxa S3 and S4 interchanged above.

b. Neighbor Joining does recover the correct metric and topological tree.

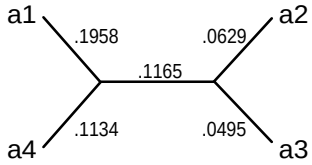
5.3.4. a. The Neighbor Joining tree (shown below) has the same unrooted topological structure as the UPGMA one, but a different metric structure.



b. The Neighbor Joining tree (shown below) differs both topologically and metricly from the FM one.

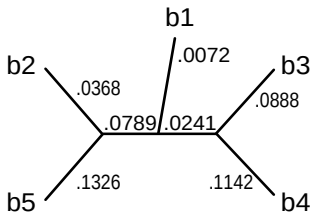


5.3.5. a.



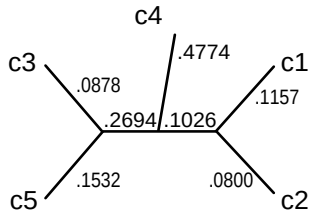
The unrooted topological structure of the UPGMA, FM, and NJ trees are all the same. All trees show that `a1` is furthest from the neighbors `a2` and `a3`. The metric features of the FM and NJ trees are essentially the same, and differ from the UPGMA tree. (However, since this data was simulated with a molecular clock, we might still prefer the UPGMA tree to the others.)

b.



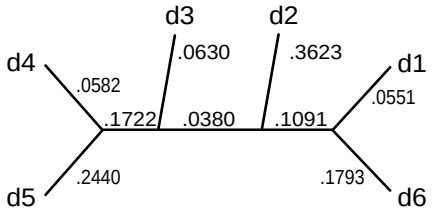
The NJ tree is topologically and metrically different from the UPGMA and FM trees. In particular, NJ chose `b3` and `b4` to be neighbors, even though `b3` is closer in distance to `b1` than `b4`. FM could not do this. Notice that NJ, like FM, created branches of quite different lengths. As we know this data was not simulated with a molecular clock, we should prefer the NJ tree to either of the others.

5.3.6. a. A frequency table for every pair of sequences shows that transitions are more common than transversions. In addition, it appears that all transitions occur at about the same rate, and all transversions occur at about the same rate.
b.

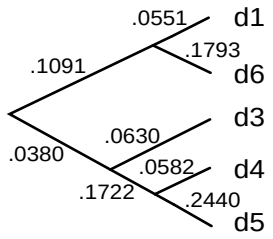


c. The tree does not appear to support a molecular clock hypothesis. Assuming a molecular clock, the best location for the root is either along the edge joining `c4` to the main tree or along the edge of length .2694. However, with either choice there is still much variation in distances between the root and taxa.

5.3.7. a. Frequency tables for pairs of sequences show transitions are more common than transversions. In addition, it appears that all transitions occur at about the same rate, and all transversions occur at about the same rate. Thus using the Kimura 2-parameter distance seems a reasonable choice.
b.



c. The outgroup appears to be d2.

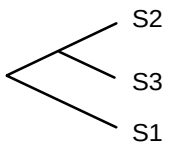


5.4. Tree Construction: Maximum Parsimony

5.4.1. a. Both trees have parsimony score 3.
b. The most parsimonious trees have score 2.
c. Since there are only four bases, we can always find a tree that requires three substitutions. For example, if we create one joined cluster of taxa with A's, another cluster with G's, another with C's, and a last with T's, then join up these clusters, the resulting tree will have parsimony score 3.

5.4.2. a. The tree on the left has parsimony score 7; the tree on the right has parsimony score 8.
b. The third unrooted tree has parsimony score 10. Therefore, the tree pictured on the left is the most parsimonious unrooted tree.

5.4.3. a. Sites 3, 6, 8, and 11
b. S1 and S4 are neighbors and S2 and S3 are neighbors. The parsimony score for the rooted tree relating them is 5.
c.



5.4.4. Suppose there are n sequences. If, at a particular site, $n-1$ sequences are in agreement and the remaining sequence disagrees with these, then the mutation count must be increased by one. If there are n_1 such sites, then the mutation count must be augmented by n_1. If, at a particular site, $n-2$ of the sequences are in agreement and the two remaining sequences disagree with each other and all the other sequences, then the count is increased by two. If there are a total of n_2 such sites, then the mutation count is augmented by $2n_2$. Similarly, for $3n_3$.

5.4.5. Both trees require three mutations.

5.4.6. In order for a notion of informative sites to make sense, there must be at least four sequences being compared, since an informative site is one for which at least two bases occur twice each. Since parsimony scores measure the fitness of unrooted trees and there is only one unrooted tree relating three taxa, there is no need for informative sites when we want to compare three taxa.

5.4.7. a. There are n sites and one of 4 bases must occur at each site. For the first sequence, there are 4 possibilities for the base occurring there; for the second sequence there are also 4 possibilities. This gives a total of $4^2 = 16$ possible

patterns for two sequences. If we consider a third sequence, since there are 4 possibilities for the base occurring at the site, there are four times as many patterns or $4^3 = 4(16)$ total patterns for three sequences. In general, for n sequences, there are 4^n possible patterns.

b. There are n possible ways to select the sequence that does not agree with the other $n-1$ sequences. There are 4 bases which can occur at this particular sequence and 3 remaining choices for the base that occurs in the $n-1$ remaining sequences. This gives a total of $(4)(3)n$ possibilities.

c. There are $\frac{n(n-1)}{2}$ ways to select the two sequences that disagree with the other $n-2$ sequences (n choices for the first sequence, $n-1$ for the second, divide by 2! since order does not matter). There are 4 ways to choose the base appearing at the first of these, 3 ways to choose the base occurring at the second of these, and 2 ways to choose the base occurring at the other $n-2$ sequences. Thus, there are $(4)(3)n(n-1)$ possible ways to obtain this pattern.

d. This is similar to part (c). There are $\frac{n(n-1)(n-2)}{3!}$ ways to select the three sequences that disagree with the other $n-3$ sequences. There are 4 ways to choose the base appearing at the first of these, 3 ways to choose the base occurring at the second of these, 2 ways to choose the base occurring at the third of these, and 1 way to choose the base occurring at the other $n-3$ sequences. Thus, there are $(4)n(n-1)(n-2)$ possible ways to obtain this pattern.

e. The number of informative patterns is 4^n - $((4)(3)n + (4)(3)n(n-1) + (4)n(n-1)(n-2))$. Since 4^n grows much more rapidly than n^3, most patterns are informative.

5.4.8. The parsimony score would be $\sum f_{pattern} p_{pattern}$, where the sum is taken over all patterns.

5.4.9. a. Informative patterns for four taxa contain two bases, each occurring twice. There are 3 ways to make such patterns without regard to base choices.

b. 25

5.4.10. a. 8.44% or 38/450

b. There are three unrooted trees relating four taxa.

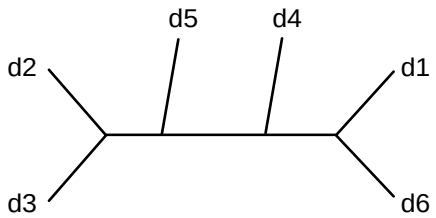
c. Using the first ten informative sites, `a1` and `a4` are neighbors, as are `a2` and `a3`. This branching structure is in agreement with both the UPGMA and the NJ trees.

5.4.11. a. 30.8% or 244/792

b. 105

c. For the tree topology reported in Problem 5.3.7b, the parsimony score is 18.

d. Answers may vary. Interchanging the taxa `d2` and `d3` on the tree of part (c) results in a parsimony score of 17. For the tree below, the parsimony score is 20.



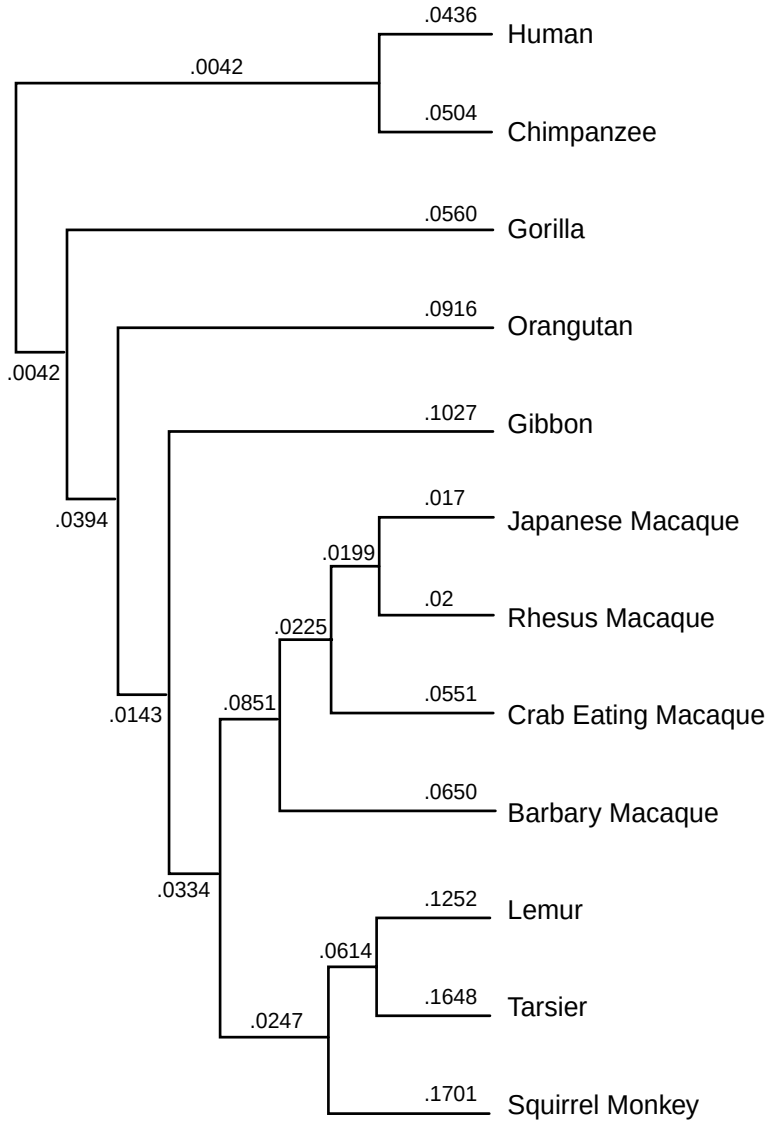
e. You have calculated parsimony scores for only five out of 105 possible trees (4.76%), using only ten out of 244 informative sites (4.10%). It is hard to have

much confidence in your answer using such a small subset of the data and only testing a small number of trees.

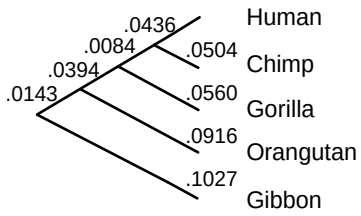
5.6. Applications and Further Reading

Warning: Trees are not drawn to scale.

5.6.1.

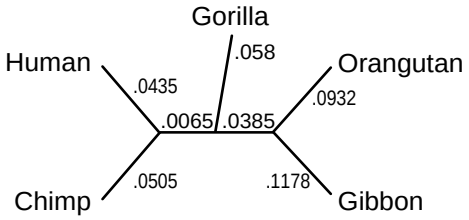


5.6.2.



5.6.3. There are $654,729,075$ unrooted trees that might relate the twelve primates and 105 rooted trees that might relate the five hominoids.

5.6.4.



This tree agrees with the full primate tree topologically. There are very small metric differences, but none of any importance. Notice that the NJ algorithm joined the taxa in a different order than it did in Problem 5.6.1. This is responsible for the small numerical differences, as the collapsing steps were done differently.

5.6.5. a. While all 4 distance arrays are reasonable close, those in `Dist_primates` are closest to Jukes-Cantor (maximum difference .050), next closest to Kimura 2-parameter (maximum difference .011), and furthest from log-det (maximum difference .026). All three of these distances tend to give larger values than those in `Dist_primates`.

b. The topology of all four trees is the same. However, NJ with `DJC`, `DK2`, and `DLD` connects the neighbors orangutan and gibbon to gorilla in the second step, whereas gorilla is the last taxa joined to the tree using `Dist_hominoids`. The metric features of the trees are quite close, especially if you compare relative relationships of edge lengths rather than absolute edge lengths. All this indicates that for this data the phylogeny inferred by NJ is not very sensitive to the choice of distance.

5.6.6. Using `compseq` to compare sequences shows that, in general, transitions are considerably more frequent than transversions, so a Jukes-Cantor model seems a poor choice. Perhaps a Kimura 2-parameter model may be a reasonable, but note that not all transitions occur equally often.

Calculating the frequencies of each base in each of the sequences shows a dearth of G's, so the Kimura 2-parameter assumption of uniform base distribution is far from valid.

Also, many of the Markov matrices have zero entries, have diagonal entries that differ by more than .1, and do not much resemble a Kimura 2-parameter or Jukes-Cantor matrix.

All this seems to point toward choosing a different model, though further analysis would be needed to decide if it is necessary to do so.

5.6.7. The observations of Problem 5.6.5 continue to hold, though the maximum differences between distances are slightly larger. As in Problem 5.6.6, using a distance other than Jukes-Cantor or Kimura 2-parameter seems justified — either the log-det distance, or another more general model.
Regardless of which of the four distances is used, NJ produces the same topological tree.

5.6.8. The tree with neighbor pairs (chimpanzee-gorilla) and (orangutan-gibbon) is the most parsimonious of the three, with parsimony score 15. The tree with neighbor pairs (human-chimpanzee) and (orangutan-gibbon) has parsimony score 16, while the tree with neighbor pairs (human-gibbon) and (chimpanzee-gorilla) has parsimony score 18. Notice that the most parsimonious tree here is not in agreement with the tree from neighbor joining.

5.6.9. The following answer was constructed using informative sites at locations 20, 109, 223, 265, 382, 442, 558, 716, 773, and 842 in the sequences. These sites are evenly distributed among the ninety informative sites (indices 4, 13, 22, 31, 40, 49, 58, 67, 76, 85 from among the ninety informative sites).
For these informative sites, the tree with neighbor pairs (human-chimpanzee) and (orangutan-gibbon) and the tree with neighbor pairs (chimpanzee-gorilla) and (orangutan-gibbon) tie for most parsimonious, with scores of 17. The third tree has a parsimony score of 20.
Use of these sites might be more sound than use of the first ten, since because they are more spread out in the sequence, it is more reasonable that substitutions at them are independent of one another.

5.6.10. Many biologists use PAUP* or PHYLIP.

5.6.11. a. The frequency tables vary quite a bit. In many of the pairs, there are substantially more transitions than transversions so the Jukes-Cantor model seems inappropriate. However, the sequences deviation from a uniform base distribution makes it difficult to justify the choice of the Kimura 2-parameter models.
b. There are many fewer noncoding sites than coding ones (only 205 out of 898), so caution is necessary when drawing conclusions from the frequency tables. These tables have a lot of zeros and small numbers in the off-diagonal positions, yet the proportion of base changes seems to be lower than for the coding sites. This may be surprising, since we might think that there would be more mutations in noncoding sequences as these have no effect on the viability of the species.
Since transitions appear more frequently than transversions, but the base distribution in each sequence is far from uniform, a model different from either the Jukes-Cantor or Kimura 2-parameter is probably needed.
Moreover, since substitution rates and the base distribution for noncoding sites is a bit different from that for coding sites, it may be desirable to use a different model for each set of sites.

5.6.12. a. Performing neighbor joining with the log-det distances on only the coding sites gives exactly the same topological tree as using all the sites.
b. Neighbor joining using the noncoding sites with the log-det distances gives a different topological tree than the one constructed using all the sites. For example, this tree creates the neighbor pairs (Japanese Macaque-Rhesus Macaque) and (Crab Eating Macaque-Barbary Macaque) which are then joined. Then (Orangutan-Gibbon) are joined before being linked to the previous cluster.

Since the sequences of noncoding sites are relatively short, we shouldn't draw too strong a conclusion here, but this has at least shown the need for more analysis. Perhaps further knowledge of DNA sequences, fossil records, or morphological data can help determine if the tree constructed using only the coding sites is the most believable.

CHAPTER 6

Genetics

6.1. Mendelian Genetics

6.1.1. F_n will have 2^n copies of each gene.

6.1.2.

	DW	dW	Dw	dw
DW	$DDWW$	$DdWW$	$DDWw$	$DdWw$
dW	$DdWW$	$ddWW$	$DdWw$	$ddWw$
Dw	$DDWw$	$DdWw$	$DDww$	$Ddww$
dw	$DdWw$	$ddWw$	$Ddww$	$ddww$

Genotype proportions are 1/16 for $DDWW$, $ddWW$, $DDww$, and $ddww$; 1/8 for $DdWW$, $Ddww$, $DDWw$, and $ddWw$; 1/4 for $DdWw$. Phenotype 7proportions are 1/16 for dwarf wrinkled-seed; 3/16 for dwarf round-seed; 3/16 for tall wrinkled-seed; 9/16 for tall round-seed.

6.1.3. a. $\mathcal{P}(\text{tall wrinkled-seed}) = \mathcal{P}(\text{tall})\mathcal{P}(\text{wrinkled-seed})$. But

$$\begin{aligned}\mathcal{P}(\text{tall}) &= 1 - \mathcal{P}(\text{dwarf}) = 1 - \mathcal{P}(dd)\\ &= 1 - \mathcal{P}(d \text{ from first parent})\mathcal{P}(d \text{ from second parent})\\ &= 1 - (1/2)(1) = 1/2,\end{aligned}$$

$$\begin{aligned}\mathcal{P}(\text{wrinkled-seed}) &= \mathcal{P}(ww)\\ &= \mathcal{P}(w \text{ from first parent})\mathcal{P}(w \text{ from second parent})\\ &= (1/2)(1/2) = 1/4.\end{aligned}$$

Therefore $\mathcal{P}(\text{tall wrinkled-seed}) = (1/2)(1/4) = 1/8$.

b. Using the calculations in part (a),

$$\begin{aligned}\mathcal{P}(\text{tall round-seed}) &= \mathcal{P}(\text{tall})\mathcal{P}(\text{round-seed})\\ &= \mathcal{P}(\text{tall})(1 - \mathcal{P}(\text{wrinkled-seed}))\\ &= (1/2)(1 - 1/4) = 3/8.\end{aligned}$$

6.1.4. a. Since the probability of having the allele is 1/31 for the male and also 1/31 for the female, assuming these are independent the probability is $(1/31)^2 \approx .00104$.

b. Since the child must inherit the recessive allele from each parent, the probability is $(1/2)(1/2) = 1/4$.

c. $(1/31)^2(1/4) \approx .0002601$.

6.1.5. a. Four – ABC, aBC, AbC, and abC.

b. 9 genotypes are possible: $AABBCC$, $AaBBCC$, $aaBBCC$, $AABbCC$, $AaBbCC$, $aaBbCC$, $AAbbCC$, $AabbCC$, $aabbCC$. 4 phenotypes are possible: the offspring could have the dominant or recessive phenotype from either of the first two genes, but must have the dominant phenotype from the third.

6.1.6. a. 2^n

b. 3^n genotypes and 2^n phenotypes

c. There are $3^k 2^{l-k}$ possible genotypes (3 possibilities for each of genes 1–k, 2 possibilities for each of genes $(k+1)$–l, and only 1 possibility for the remaining genes). These give 2^l different phenotypes, since genes 1–l might give dominant or recessive traits while the remaining ones must be recessive. (Note we are using that at genes $(k+1)$–l the first individual is homozygous recessive.)

6.1.7. a. $AA \times aa$ dominant:recessive=1:0; $Aa \times aa$ dominant:recessive=1:1; $aa \times aa$ dominant:recessive=0:1.

b. $DdwwYY$

c. Crossing with a homozygous recessive allows all parental alleles to manifest themselves in phenotypes of the progeny, whereas crossing with a homozygous dominant would result only in progeny of the dominant phenotypes. Quantitatively, the ratios in part (a) are all different, so the parental phenotype can be distinguished, while for a cross with a homozygous dominant, all ratios would be dominant:recessive=1:0. The parental phenotype has no effect.

6.1.8. a. From $BBRR \times bbrr$, all offspring have genotype $BbRr$, with black and normal length fur.

b. In F_2, 1/2 of the rabbits will be homozygous for the color gene (BB or bb) and 1/4 will be homozygous for both genes. Of the black rabbits, 1/6 will be homozygous for both genes. (For all the rabbits, BB:Bb:bb=1:2:1, but only BB and Bb are black, so $1/(1+2) = 1/3$ of the black rabbits are BB. Of these 1/2 are homozygous at the second gene with rr or RR.)

c. Black rabbits with normal length fur have genotypes $BBRR$, $BbRR$, $BBRr$, or $BbRr$. In the entire F_2 population these occur in proportions 1/16, 2/16, 2/16, and 4/16, for a total of 9/16. Thus the genotype ratios for black rabbits with normal length fur homozygous for both genes is 1:2:2:4 giving proportions $1/9, 2/9, 2/9, 4/9$.

6.1.9. a. Genotype $WwGg$, with round yellow seed phenotype.

b. If the genes assort independently, F_2 should be: 1/16 with wrinkled green seeds, 3/16 with wrinkled yellow seeds, 3/16 with round green seeds, and 9/16 with round yellow seeds.

c. If Mendel's data did not exactly match these proportions, he should not necessarily doubt the independent assortment hypothesis. After all, these proportions are really probabilities, so only for very large amounts of data should the fit be very close. The more data he collected, the closer he should expect his data to match these proportions if the hypothesis is valid. Deciding how close is close enough for a match, taking into account the amount of data collected, will be discussed in the next section.

6.1.10. The cross is $Y^l y \times Y^l y$. Embryo genotype ratios will be $Y^l Y^l$:$Y^l y$:yy=1:2:1, but only the last two genotypes will be born. Thus the viable progeny will have genotypes $Y^l y$ or yy, with respective phenotypes yellow and agouti, in proportions 2/3 and 1/3.

6.1.11. a. Since one child is homozygous recessive, and neither parent is, both parents must be heterozygous.

b. From a cross of two heterozygotes, the probability that a child is not homozygous recessive is 3/4.

c. Since the sons do not have sickle-cell anemia, the possible genotypes are homozygous dominant or heterozygous. Since all offspring have probabilities

1/4 and 1/2 of these genotypes, for a disease-free son the probabilities are $(1/4)/(1/4+1/2) = 1/3$ and $(1/2)/(1/4+1/2) = 2/3$.

6.1.12. a. The trait is dominant. If it were recessive, all children of the parents would exhibit brachydactyly. The parents must each be heterozygotes, since one child has normal length fingers. The child with normal length fingers is a homozygous recessive. The child with short fingers is either homozygous dominant or heterozygous.

b. The probability that one child has normal fingers is 1/4, and since the two children's phenotypes are independent, the probability that both have normal length fingers is $(1/4)^2 = 1/16$.

6.1.13. a. $(1/4)^3 = 1/64 = .015625$

b. $(1/2)^3 = 1/8 = .125$

c. The proportion homozygous only for the first gene is (1/2)(1/2)(1/2)=1/8. Similarly, 1/8 is homozygous only for the second, and 1/8 is homozygous only for the third. Thus 1/8+1/8+1/8=3/8 is homozygous for exactly one of the genes.

d. The proportion homozygous for at least one gene is

$$1 - (\text{proportion heterozygous for all three}) = 1 - (1/2)^3 = 7/8 = .875.$$

6.1.14. F_1 has genotype Ww only, with pink flower phenotype. F_2 has genotypes WW, Ww, and ww in proportions 1/4, 1/2, and 1/4, with red flower, pink flower, and white flower phenotypes, respectively.

6.1.15. $a_1a_3 \times a_2a_3$ produces genotypes a_1a_2, a_1a_3, a_2a_3, and a_3a_3 in frequencies 1/4, 1/4, 1/4, and 1/4. This gives phenotypes associated to a_1, a_2, and a_3 in frequencies 1/2, 1/4, and 1/4.

6.1.16. a. Type A: I^AI^A, I^AI^O; Type B: I^BI^B, I^BI^O; Type AB: I^AI^B; Type O: I^OI^O

b. $I^AI^A \times I^BI^O$ produces offspring with genotypes I^AI^B and I^AI^O with equal probability. Thus type AB and type A blood occur with relative frequencies 1/2 and 1/2.

c. From $I^AI^O \times I^BI^O$ we expect 1/4 of the progeny to have type O blood (genotype I^OI^O). So out of four children, we would expect one to have type O blood. However, any number might have this blood type. The probability of any one child having it is 1/4, and it is really only in a very large number of trials (much greater than 4) that we can be reasonably confident that close to 1/4 of the trials will produce this outcome. For instance, there is a probability of $(1/4)^4 = 1/256$ that all four children will have type O blood, and of $(3/4)^4 = 81/256$ that none of them will. (See the next section for a more careful definition of the word 'expect'.)

6.1.17. a. $RRpp \times rrpp$ produces only $Rrpp$, for a rose comb phenotype. $rrPP \times rrpp$ produces only $rrPp$, for a pea comb phenotype.

b. $RRpp \times rrPP$ produces an F_1 of $RrPp$ which have walnut comb phenotype. Interbreeding to produce F_2 gives 3/16 rose comb (1/16 $RRpp$ and 2/16 $Rrpp$), 3/16 pea comb (1/16 $rrPP$ and 2/16 $rrPp$), 1/16 single comb (1/16 $rrpp$), and 9/16 walnut comb (1/16 $RRPP$, 2/16 $RrPP$, 2/16 $RRPp$, and 4/16 $RrPp$).

6.2. Probability Distributions in Genetics

6.2.1. $HHHTT$, $HHTHT$, $HTHHT$, $THHHT$, $HHTTH$, $HTHTH$, $THHTH$, $HTTHH$, $THTHH$, $TTHHH$; $\binom{5}{3} = \frac{5\cdot4\cdot3\cdot2\cdot1}{(3\cdot2\cdot1)(2\cdot1)} = \frac{5\cdot4}{2} = 10$.

6.2.2. $\mathcal{P}(\text{exactly 0 tails in 3 flips}) = \binom{3}{0}(1/2)^0(1/2)^3 = 1/8$,
$\mathcal{P}(\text{exactly 2 tails in 3 flips}) = \binom{3}{2}(1/2)^2(1/2)^1 = 3/8$,
$\mathcal{P}(\text{exactly 3 tails in 3 flips}) = \binom{3}{3}(1/2)^3(1/2)^0 = 1/8$;
The sum of the four probabilities is 1, since the outcomes are mutually exclusive and exhaust all possibilities.

6.2.3. Use $\mathcal{P}(i) = \binom{4}{i}(3/4)^i(1/4)^{4-i}$.

6.2.4. The values of $\binom{10}{k}$ for $k = 1, 2, \ldots 10$ are 1, 10, 45, 120, 210, 252, 210, 120, 45, 10, 1.
a. The value is smallest when $k = 0$ or 10. If we choose no objects, or all 10 objects, there is only one way to do so. If we choose any number from 1 to 9, there is more than one way to do so.
b. The value is largest for $k = 5$. It does seem intuitively reasonable that there are more ways to choose exactly half of the objects than there are to choose fewer, or more. With fewer or more, there is less freedom in varying what is chosen.
c. $\binom{n}{k}$ increases with k until k is half of n and then decreases, $\binom{n}{k} = \binom{n}{n-k}$, $\binom{n}{1} = n$; these patterns hold for all n.

6.2.5. a. In choosing only one object, the various ways are: choose the first, choose the second, ... , choose the nth. Thus $\binom{n}{1} = n$. Choosing $n-1$ objects is equivalent to picking the one object *not* chosen, so $\binom{n}{n-1} = \binom{n}{1} = n$.
b. There is only one way to choose no objects, so $\binom{n}{0} = 1$. To choose n objects from n, we must choose them all, so $\binom{n}{n} = 1$.

6.2.6. a. $\binom{6}{4}(\frac{1}{2})^4(\frac{1}{2})^2 = 15\frac{1}{64} = \frac{15}{64} \approx .2344$
b. $\mathcal{P}(\text{exactly } i \text{ boys in 6 children}) = \binom{6}{i}(\frac{1}{2})^6$, so for $i = 0, 1, 2, \ldots 6$, the values are: .0156, .0938, .2344, .3125, .2344, .0938, .0156.
$\mathcal{P}(\text{exactly } i \text{ girls in 6 children})$ has exactly the same values.
c. The expected number of boys is $\sum_{i=0}^{6} i\mathcal{P}(\text{exactly } i \text{ boys in 6 children}) = 0(.0156) + 1(.0938) + 2(.2344) + 3(.3125) + 4(.2344) + 5(.0938) + 6(.0156) = 3$.
Alternately, for a binomial distribution, the expected value is $n \cdot p = 6 \cdot \frac{1}{2} = 3$.
d. $\mathcal{P}(4 \text{ or more girls of 6 children}) = \mathcal{P}(4 \text{ girls}) + \mathcal{P}(5 \text{ girls}) + \mathcal{P}(6 \text{ girls}) = .2344 + .0938 + .0156 = .3438$.

6.2.7. a. $\mathcal{P}(\text{exactly 30 agouti in 40 offspring}) = \binom{40}{30}(\frac{3}{4})^{30}(\frac{1}{4})^{10} \approx .1444$;
$\mathcal{P}(\text{exactly 300 agouti in 400 offspring}) = \binom{400}{300}(\frac{3}{4})^{300}(\frac{1}{4})^{100} \approx .0460$
b. Even though these results indicate the probability of having exactly 3/4 of the offspring with agouti fur decreases as the number of offspring increases, these results are consistent with expecting 3/4 of a large number of offspring to have agouti fur. We don't expect *exactly* 3/4 to have agouti fur, but rather that for a very large number of offspring, the proportion with agouti fur is likely to be close to 3/4, and is *on average* 3/4.

6.2.8. $1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$

6.2.9. a. The outcomes $k = 2, 3, \ldots 7$ can each occur in $k-1$ ways, namely $1+(k-1)$, $2+(k-2)$, ... $(k-1)+1$. The outcomes $k = 7, 8, \ldots, 12$ can each occur in $13-k$ ways, namely $6+(k-6)$, $5+(k-5)$, ... , $(k-6)+6$. Each of the individual outcomes listed occurs with probability $(1/6)^2 = 1/36$. Therefore the probabilities of the sum being 2, 3, ... , 12 are: 1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 5/36, 4/36, 3/36, 2/36, 1/36. The expected value of the sum is

$2(1/36)+3(2/36)+4(3/36)+5(4/36)+6(5/36)+7(6/36)+8(5/36)+9(4/36)+10(3/36)+11(2/36)+12(1/36)=7$.

b. By problem 6.2.8, the expected value for one die toss is 3.5. Letting X_1 and X_2 denote the random variables for the two dice, $E(X_1)=E(X_2)=3.5$ so $E(X_1+X_2)=E(X_1)+E(X_2)=7$.

6.2.10. a. $p=1/6$, $q=5/6$, $k=3$, $n=10$; $\binom{10}{3}(\frac{1}{6})^3(\frac{5}{6})^7\approx .1550$

b. $p=5/6$, $q=1/6$, $k=7$, $n=10$; $\binom{10}{7}(\frac{5}{6})^7(\frac{1}{6})^3\approx .1550$

6.2.11. a. Choosing k objects out of n is exactly equivalent to designating the $n-k$ objects that are *not* chosen. Since 'choose' and 'designate' mean essentially the same thing here, this means $\binom{n}{k}=\binom{n}{n-k}$.

b. $\binom{n}{n-k}=\frac{n!}{(n-(n-k))!(n-k)!}=\frac{n!}{k!(n-k)!}=\frac{n!}{(n-k)!k!}=\binom{n}{k}$.

6.2.12. a. $1/2$

b. $(1/2)(1/2)=1/4$

c. $\binom{2}{1}(1/2)(1/2)=1/2$

d. $\binom{2}{1}(1/2)(1/2)+\binom{2}{2}(1/2)^2(1/2)^0=3/4$, or, computing the probability that it is not the case that no children are albinos, $1-\binom{2}{0}(1/2)^0(1/2)^2=3/4$.

e. Using the formula for the expected value of a binomial random variable, $2(1/2)=1$, or, using the definition of expected value, $0\cdot\binom{2}{0}(1/2)^0(1/2)^2+1\cdot\binom{2}{1}(1/2)^1(1/2)^1+2\cdot\binom{2}{2}(1/2)^0(1/2)^2=1$.

6.2.13. a. The probability of any particular offspring being fat with agouti fur is $(3/4)(1/4)=3/16$, assuming these genes assort independently. The number of progeny in 25 with this phenotype is a binomial random variable. Thus the expected value of it is $(25)(3/16)=75/16=4.6875$

b. $\binom{25}{4}(3/16)^4(13/16)^{21}\approx .1997$

c. $\sum_{i=0}^{4}\binom{25}{i}(3/16)^i(13/16)^{25-i}\approx .4837$

d. $1-\sum_{i=0}^{3}\binom{25}{i}(3/16)^i(13/16)^{25-i}\approx .7160$

6.2.14. a. $\mathcal{P}(\text{age at death}=0)=1/2$;
$\mathcal{P}(\text{age at death}=1)=(1/2)(3/4)=3/8$;
$\mathcal{P}(\text{age at death}=2)=(1/2)(1/4)(3/4)=3/32$;
$\mathcal{P}(\text{age at death}=3)=(1/2)(1/4)(1/4)(1)=1/32$.
These probabilities add to 1 since the events are disjoint and exhaust all possibilities.

b. $0(1/2)+1(3/8)+2(3/32)+3(1/32)=.65625$

6.2.15. a. Recall from problem 6.1.10, that the probability an offspring is yellow is 2/3. Then the probability 5 of 12 have normal coloring is $\binom{12}{5}(1/3)^5(2/3)^7\approx .1908$.

b. $\sum_{i=10}^{12}\binom{12}{i}(2/3)^i(1/3)^{12-i}\approx .1811$

c. $\sum_{i=0}^{3}\binom{12}{i}(2/3)^i(1/3)^{12-i}\approx .0039$

6.2.16. a. Since the probability that any given child in the family will develop Huntington disease is 1/2, the probability that none of 4 do is $\binom{4}{0}(1/2)^0(1/2)^4=1/16$.

b. The probability that at least one of the 4 develops the disease is $1-\mathcal{P}(\text{none of 4})=1-1/16=15/16$.

c. The probability that 3 or more develop the disease is $\binom{4}{3}(1/2)^3(1/2)^1+\binom{4}{4}(1/2)^4(1/2)^0=5/16$.

6.2.17. For each trait, the probability an individual exhibits the dominant phenotype is 3/4, so the probability of an individual exhibiting the dominant phenotype for all three traits is $(3/4)^3=27/64$. The probability that 20 of 30 progeny will

exhibit all three dominant phenotypes is therefore $\binom{30}{20}(27/64)^{20}(37/64)^{10} \approx .0040$.

The probability an individual exhibits the dominant phenotype for at least one trait is $1 - \mathcal{P}(\text{recessive phenotype for all 3 traits}) = 1 - (1/4)^3 = 63/64$. The probability that at least 2 of 30 progeny exhibit the dominant phenotype for at least one trait is $1-\binom{30}{0}(63/64)^0(1/64)^{30}-\binom{30}{1}(63/64)^1(1/64)^{29} \approx 1-(1.2339\times 10^{-51}) \approx 1$.

6.2.18. a. There are n choices for the first ball. The remaining $n-1$ balls give $n-1$ choices for the second ball. Then there are $n-2$ choices for the third ball, etc., so there are $n-l+1$ choices for the lth ball.

b. To see how many ways k balls could be chosen (in order), we simply multiply the number of possible choices at each successive picking of a ball. This gives $n(n-1)(n-2)\cdots(n-k+1)$.

c. Picking k of k balls (in order), by the reasoning in (a) and (b), can be done in $k(k-1)(k-2)\cdots(2)1 = k!$ ways.

d. If the various choices of ordered balls counted in (b) are grouped according to the *unordered* set of balls chosen, then each group will have in it the count in (c). Thus the number of unordered sets of balls that could be chosen is $n(n-1)(n-2)\cdots(n-k+1)/(k!)$.

e. $\frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} = \frac{n(n-1)(n-2)\cdots(n-k+1)(n-k)(n-k-1)\cdots(2)1}{(k!)(n-k)(n-k-1)\cdots(2)1} = \frac{n!}{k!(n-k)!}$.

6.2.19. a. $(x+y)^2 = x^2+2xy+y^2 = \binom{2}{0}x^2+\binom{2}{1}xy+\binom{2}{2}y^2$

$(x+y)^3 = x^3+3x^2y+3xy^2+y^3 = \binom{3}{0}x^3+\binom{3}{1}x^2y+\binom{3}{2}xy^2+\binom{3}{3}y^3$

$(x+y)^4 = x^4+4x^3y+6x^2y^2+4xy^3+y^4 = \binom{4}{0}x^4+\binom{4}{1}x^3y+\binom{4}{2}x^2y^2+\binom{4}{3}xy^3+\binom{4}{4}y^4$

b. $(x+y)^n = (x+y)(x+y)\cdots(x+y)$. To multiply this out, we must multiply each term in the individual factors by the terms in other factors in all possible ways. Since there are n factors, a term x^ky^{n-k} will be produced for every way we can choose k of the n factors to contribute an x, with the remaining $n-k$ factors contributing a y. But $\binom{n}{k}$ by definition gives the number of ways these choices can be made, so the product will contain exactly $\binom{n}{k}$ copies of x^ky^{n-k}. Collecting these produces the given formula.

c. $\sum_{i=0}^n \binom{n}{i} = 2^n$ for all n, since $\sum_{i=0}^n \binom{n}{i} = \sum_{i=0}^n \binom{n}{i}1^i1^{n-i} = (1+1)^n = 2^n$

6.2.20. a. $E = \sum_{i=0}^n i\frac{n!}{i!(n-i)!}p^iq^{n-i}$

b. Use straightforward algebra.

c. Note the $i=0$ term in E is 0, so

$$\begin{aligned}
E &= \sum_{i=1}^n i\frac{n!}{(n-i)!i!}p^iq^{n-i} = pn\sum_{i=1}^n \frac{(n-1)!}{(n-i)!(i-1)!}p^{i-1}q^{(n-1)-(i-1)} \\
&= pn\sum_{j=0}^{n-1} \frac{(n-1)!}{(n-1-j)!j!}p^jq^{(n-1)-j} \qquad \text{(replacing } i \text{ with } j+1) \\
&= pn\sum_{j=0}^{n-1} \binom{n-1}{j}p^jq^{(n-1)-j} \\
&= pn(p+q)^{n-1} = pn(1)^{n-1} = pn
\end{aligned}$$

6.2.21. a.

$$\begin{aligned} E(X_1 + X_2) &= \sum_k k\mathcal{P}(X_1 + X_2 = k) \\ &= \sum_k \sum_{i,j \text{ with } i+j=k} (i+j)\mathcal{P}(X_1 = i)\mathcal{P}(X_2 = j) \\ &= \sum_{i,j} (i+j)\mathcal{P}(X_1 = i)\mathcal{P}(X_2 = j) \end{aligned}$$

Note the last step simply reorders the sum.
b.

$$\begin{aligned} &\sum_i \sum_j (i+j)\mathcal{P}(X_1 = i)\mathcal{P}(X_2 = j) \\ &\quad = \sum_i \sum_j i\mathcal{P}(X_1 = i)\mathcal{P}(X_2 = j) + \sum_j \sum_i j\mathcal{P}(X_1 = i)\mathcal{P}(X_2 = j) \\ &\quad = \sum_i i\mathcal{P}(X_1 = i) \sum_j \mathcal{P}(X_2 = j) + \sum_j j\mathcal{P}(X_2 = j) \sum_i \mathcal{P}(X_1 = i) \end{aligned}$$

c. The two summations mentioned each give 1, so the result in part (b) is $E(X_1) + E(X_2)$.

6.2.22. a. Progeny of the specified cross should display the dominant phenotype with probability 3/4. The number of progeny in 1000 that display the dominant phenotype is a random variable with binomial distribution, so its expected value is (1000)(3/4)=750. We thus expect 750 dominant and 250 recessive phenotypes.
If 700 dominant and 300 recessive phenotypes are observed, then

$$\chi^2 = \frac{(700 - 750)^2}{750} + \frac{(300 - 250)^2}{250} = 13.333.$$

With $\alpha = .05$ and 1 degree of freedom, the critical value is $\chi^2_{crit} = 3.841$. Since this is less than our computed value, we find the data is not in accord with the Mendelian model at the .05 significance level.
b. If instead 725 dominant phenotypes are observed in 1000 progeny, $\chi^2 = 3.333$. This is less than the critical value in part (a), so the data is in accord with the Mendelian model at the .05 significance level.
c. If N dominant phenotypes are observed in 1000 progeny, then

$$\chi^2 = \frac{(N - 750)^2}{750} + \frac{(1000 - N - 250)^2}{250} = \frac{4(N - 750)^2}{750}.$$

Thus the data is in accord with the Mendelian model at the .05 significance level when

$$\begin{aligned} \frac{4(N - 750)^2}{750} &< 3.841 \\ (N - 750)^2 &< 720.1875 \\ |N - 750| &< 26.8363 \\ 723.1637 < N &< 776.8363 \end{aligned}$$

Since N must be an integer, this means $724 \leq N \leq 776$.

6.2.23. In the table, moving across a row means α decreases (with the degrees of freedom not changing). Since a smaller α means a χ^2 value computed from data is *less likely* to exceed the critical value, the critical value must get larger. Moving down a column means the degrees of freedom increases (with α held fixed). Since a larger degrees of freedom means more terms are added to compute χ^2, we expect χ^2 values to typically be larger. Thus for a fixed significance level, the critical value must be larger in order that computed χ^2 values exceed it the same percentage of times.

6.2.24. With 556 individuals included in the table, and independent assortment of the genes implying expected proportions of 9/16, 3/16, 3/16, and 1/16 of the phenotypes in the order listed, the expected frequencies are 312.75, 104.25, 104.25, and 34.75. This gives $\chi^2 = .47$. This is well below the critical value if α is taken to be .01, .05, or even .10. Thus the data is judged to be in accord with independent assortment at any of these significance levels.

6.2.25. a. For a significance level of .01, the critical value would appear further to the right than the value shown in Figure 6.2. It should be located so that the area under the curve to the right of it is one fifth of the shaded area.
b. The shape of the curve shows that typically the values of χ^2 from data described by a model will be small, but not too small. Most values will lie in the region on the horizontal axis that is below the 'hump'. Very large values of χ^2 are rare when the model describes the production of the data well, and the use of a cut-off critical value in the goodness-of-fit test simply formalizes how rare. Note also that extremely small values of χ^2 are also rare.

6.3. Linkage

6.3.1. a. Since Albert was not a hemophiliac, he must have had genotype X^+Y. Since so many of Victoria's children carried the hemophilia allele, it is unlikely they all were the result of new mutations, so Victoria had genotype X^+X^h.
b. 1/2; 0 ; 1/2
c. $\binom{9}{3}(1/2)^3(1/2)^6 \approx .1641$

6.3.2. F_1 is composed of X^+X^w and X^+Y genotypes, so the probability of each of the genotypes X^+X^+, X^+Y, X^+X^w and X^wY for an individual in F_2 is 1/4. Thus we expect the phenotypes listed in the order in Table 6.9 to occur in proportions 1/2, 0, 1/4, and 1/4. With 4252 progeny, we'd expect 2126, 0, 1063, and 1063 of the phenotypes. While the model is in rough agreement with the data, it's not that close. Perhaps the white-eyed males have reduced viability, lowering the count for that group below the theoretical prediction. On the other hand, even if the model is correct, this experiment's results may simply be due to random fluctuations.

6.3.3. a. X^wX^w and X^+Y
b. F_1 is composed of X^+X^w, which are red-eyed females, and X^wY, which are white-eyed males, in proportions 1/2 and 1/2.
c. F_2 is composed of X^+X^w (red-eyed females), X^wX^w (white-eyed females) X^+Y (red-eyed males) and X^wY (white-eyed males) in equal proportions.

6.3.4. The expected proportion of each of the four phenotypes is 1/4, so with 435 progeny, we'd expect 108.75 of each. Thus $\chi^2 = 17.4598$. With $\alpha = .05$, and three degrees of freedom, $\chi^2_{critical} = 7.81473$. Thus we do not find the data in accord with the model at the .05 significance level.

Notice, however, that this discrepancy could be explained if the white-eyed allele also results in reduced viability, since the white-eyed progeny of both sexes appeared in smaller numbers than expected. It is important to look deeper than the bald result of the χ^2 test in order to form new hypotheses.

6.3.5. a. Since the parents are X^+X^d and X^+Y, the probability a son has the disease is $1/2$.

b. The probability a daughter is heterozygous is $1/2$.

c. Note that daughters can not be homozygous for the disease-causing allele, so the probability two daughters are carriers is $(1/2)^2 = 1/4$.

6.3.6. a. The first cross is $X^+X^+ \times X^cY$. A daughter of this cross must be X^+X^c. Her offspring are from $X^+X^c \times X^+Y$, and so her sons are X^+Y and X^cY with equal probability. Thus the probability that a son is color blind is $1/2$.

b. Her daughters are X^+X^+ and X^+X^c with equal probability, and so (assuming the color-blindness allele is recessive) the probability of a daughter being color blind is 0.

c. $\binom{3}{2}(1/2)^2(1/2)^1 = 3/8$.

6.3.7. a. For expression in all male offspring, the mother must be X^aX^a. For expression in no female offspring, the father must be X^+Y and the a allele must be recessive.

b. For expression in 50% of male offspring, the mother must be X^+X^a. For expression in 50% of female offspring, the father must be X^aY if the a allele is recessive, or X^+Y if it is dominant.

c. For expression in no male offspring, the mother must be X^+X^+. For expression in all female offspring, the father must be X^aY and the a allele must be dominant.

d. For expression in 50% of male offspring, the mother must be X^+X^a. For expression in no female offspring, the father must be X^+Y and the a allele must be recessive.

e. Expression in 25% of the progeny can only occur through i) expression in no males and 50% of females, ii) expression in 25% of males and 25%of females, or iii) expression in 50% of males and no females. Case (i) cannot occur, since expression in no males requires the mother is X^+X^+, and then expression in a single female requires the father be X^aY and the allele dominant, which would then result in expression in all females. Case (ii) cannot occur either, since the model can only produce expression in all, half, or none of a sex. Case (iii), which is analyzed in part (d) must occur.

6.3.8. The standard model of one sex-linked gene cannot explain such data. One possibility is a cross $X^+X^a \times X^aY$ with the mutant phenotype displayed only by those individuals with 2 mutant alleles, so X^aY, X^+Y and X^+X^a all display wildtype phenotype while X^aX^a displays mutant phenotype. Other possibilities include not having a single type of cross due to a parental population of several genotypes, or involvement of multiple genes. Other valid answers no doubt exist.

6.3.9. a. Letting b denote the gene for black body color, and X^v that for vermillion eye color, the cross is $BbX^+X^v \times bbX^vY$. Analyzing the genes separately, $1/2$ the progeny will have black bodies and $1/2$ gray bodies, while both males and females will be $1/2$ vermillion-eyed and $1/2$ red-eyed. Thus each of the 8 phenotypes will occur in proportion $1/8$.

b. From $BbX^vX^v \times BBX^+Y$, all progeny will have gray bodies, 1/2 will be red-eyed females and 1/2 vermillion-eyed males.

6.3.10. The expected number of each phenotype is 312.25, so $\chi^2 = 425.3603$, which is considerably larger than the critical value. Thus the data is not consistent, at the .05 significance level, with the assumption of independent assortment of genes.

6.3.11. a. The probability is 6/7, since regardless of what chromosome the first gene lies on, the probability the second is not on that chromosome is 6/7.

b. The probability that two genes chosen at random assort independently would be greater than 6/7, since all genes on different chromosomes assort independently, and those far apart on the same chromosome do as well.

6.3.12. a.

Phenotype	Number
tall, normal sheaf	303
tall, white sheaf	163
dwarf, normal sheaf	169
dwarf, white sheaf	290
Total	925

b. $(163 + 169)/925 = .3589$, so the genetic distance is estimated as 35.89 cM.

c. This does not agree with the genetic distance of 37 cM in the text, since by collapsing the table, we lost information on the double crossovers. As a result, we undercounted recombinants, and got a distance that is smaller than the true one.

6.3.13. Assuming independent assortment, this cross would produce the 4 phenotypes in roughly equal proportions. Since the observed proportions are far from equal, there is evidence for linkage. The recombination frequency is $(39 + 35)/(198 + 228 + 39 + 35) = 74/500 = .1480$.

6.3.14. a. sn^+m^+ and $sn\,m$ each in proportion .425, sn^+m and $sn\,m^+$ each in proportion .075.

b. Wildtype bristles and wings in proportion

$$3(.425^2) + 4(.425)(.075) + 2(.075^2) = .680625;$$

Wildtype bristles and miniature wings in proportion

$$2(.425)(.075) + .075^2 = .069375;$$

Singed bristles and wildtype wings in proportion

$$2(.425)(.075) + .075^2 = .069375;$$

Singed bristles and miniature wings in proportion $.425^2 = .180625$

6.3.15. The genes on different autosomes assort independently, and thus can be analyzed separately. An $a^+b^+/a\,b$ produces gametes a^+b^+, a^+b, $a\,b^+$, and $a\,b$ in proportions .45, .05, .05, and .45. Crossing with a homozygous recessive thus yields the 4 phenotypes associated to these alleles in the same proportions. Similarly, the cross will yield phenotypes associated to c^+d^+, c^+d, $c\,d^+$, and $c\,d$ in proportions .43, .07, .43, and .07. Thus the 16 phenotypes and their proportions will be: $a^+b^+c^+d^+$, (.45)(.43)=.1935; $a^+b^+c^+d$, (.45)(.07)=.0315; $a^+b^+c\,d^+$, .0315; $a^+b^+c\,d$, .1935; $a^+b\,c^+d^+$, (.05)(.43)=.0215; $a^+b\,c^+d$, (.05)(.07)=.0035; $a^+b\,c\,d^+$, .0035; $a^+b\,c\,d$, .0215; $a\,b^+c^+d^+$, .0215; $a\,b^+c^+d$, .0035; $a\,b^+c\,d^+$, .0035; $a\,b^+c\,d$, .0215; $a\,b\,c^+d^+$, .1935; $a\,b\,c^+d$, .0315; $a\,b\,c\,d^+$, .0315; $a\,b\,c\,d$, .1935.

6.3.16. A trans configuration can be used in genetic mapping, though then the recombinant phenotypes are those that are wildtype for both traits or mutant for both traits. Even if it was not known that the heterozygous parent had a cis or trans configuration, the sizes of the phenotypic classes resulting from the cross with a homozygous recessive would indicate which one.

6.3.17. The physical distance separating genes a and b is likely to be larger than that separating c and d, since for the probability of a crossover occuring between them to be the same, near the centromere the physical distance would usually need to be larger.

6.3.18. With a tetrad drawn as in Figure 6.3 of the text, label the strands 1,2,3, and 4 from top to bottom. Then have strands 1 and 3 crossover between the 2 genes, and strands 2 and 4 crossover as well.

No two-gene configuration can produce 3 recombinant and one parental type since two strands must crossover to produce two of the recombinants, and a third strand must crossover with another strand to produce the third. This last strand cannot be one of the first two (why not?) so it must be the fourth. However, this would produce 4 recombinants.

Note that there are three-gene configurations producing three recombinants and one parental type.

6.3.19. The rare phenotypes are those associated to $(a^+b^+c^+)$ and $(a\,b\,c)$ so these must come from double crossovers. Since the parental types were (a^+b^+c) and $(a\,b\,c^+)$, we deduce the gene order $a\,c\,b$.

6.3.20. a. Any of $cl^+dp^+rd^+/cl\,dp\,rd$, $cl^+dp^+rd/cl\,dp\,rd^+$, $cl^+dp\,rd^+/cl\,dp^+rd$, or $cl^+dp\,rd\,/cl\,dp^+rd^+$ could be crossed with a homozygous recessive.

b. Since $cl^+dp\,rd$ is the result of a double crossover, the gene order must be $dp\,cl\,rd$.

6.3.21. The phenotypes normal-eyes, hairy-legs, prickly-antennae and enlarged-eyes, hairless-legs, smooth-antennae both occur with frequency (1/2)(.88)(.85)=.374; the phenotypes enlarged-eyes, hairy-legs, prickly-antennae and normal-eyes, hairless-legs, smooth-antennae both occur with frequency (1/2)(.12)(.85)=.051; the phenotypes normal-eyes, hairy-legs, smooth-antennae and enlarged-eyes, hairless-legs, prickly-antennae both occur with frequency (1/2)(.88)(.15)=.066; the phenotypes normal-eyes, hairless-legs, prickly-antennae and enlarged-eyes, hairy-legs, smooth-antennae both occur with frequency (1/2)(.12)(.15)=.009.

6.3.22. a. Since the male parent in this cross is recessive for all genes, the offspring, whether male or female, will display phenotypes associated with the maternal gamete. Thus the proportion of each phenotype should be the same in the male and female offspring.

b. The data does give evidence of linkage since the sizes of the phenotype classes are far from equal.

c. The rare phenotypes arise from maternal gametes ct^+s^+v and $ct\,s\,v^+$, so these result from double crossovers. Given the genotype of the mother, the gene order must be $ct\,v\,s$.

The genetic distance from ct to v is estimated as $(8+125+105+5)/1919 = 243/1919 \approx .1266 = 12.66cM$. The genetic distance between v and s is estimated as $(8+71+106+5)/1919 = 190/1919 \approx .0990 = 9.9cM$

6.3.23. a. $(.082)(.125)(2000) = 20.5$

b. $c = 3/20.5 = .1463$

c. No interference would mean $c = 1$, so $I = 0$. At the other extreme, if interference is so great that no double crossovers occur, then $c = 0$ so $I = 1$. In general I increases if interference produces fewer crossovers.
c. $I = .8537$

6.4. Gene Frequency in Populations

6.4.1. a. Let p be the frequency of ct in the population. Then $p^2 = 9/450$, so $p \approx .1414$.
b. The percentage of the population heterozygous for the gene is $2p(1-p) \approx .2428$.

6.4.2. a. The color-blindness allele occurs with frequency $p = .08$, so $q = .92$.
b. About $.08^2 = .0064$ of the female population is color blind, while about $2(.08)(.92) = .1472$ of the females have normal vision but carry the color-blindness allele.

6.4.3. a. $2p(1-p) = .4$ implies $p^2 - p + .2 = 0$, so $p = (1 \pm \sqrt{1-.8})/2 = (1 \pm \sqrt{.2})/2 \approx .2764$ or $.7236$, with q being the other value.
For $2p(1-p) = H$, the values of p and q are $(1 \pm \sqrt{1-2H})/2$.
b. $H = 2p(1-p)$ is maximized when $p = 1/2$, $q = 1/2$. This can be seen either by graphing the parabola, or by using calculus.

6.4.4. a. $(p+q)^2 = p^2 + 2pq + q^2$. Assuming p is the frequency of alleles a and q the frequency of allele A, the terms in this expansion give the frequencies of aa, Aa, and AA genotypes produced in a population with random mating.
b. $(p+q+r)^2 = p^2 + q^2 + r^2 + 2pq + 2pr + 2qr$. Assuming p, q, r are frequencies of the alleles a_1, a_2, a_3 of a triallelic gene in a randomly mating population, these terms are the frequencies of the genotypes of the next generation a_1a_1, a_2a_2, a_3a_3, a_1a_2, a_1a_3 and a_2a_3.
c. Yes, the Hardy-Weinberg equilibrium concept still makes sense. For instance, in the next generation of gametes, the frequency of allele a_1 is $p^2 + (1/2)2pq + (1/2)2pr = p(p+q+r) = p \cdot 1 = p$. The frequencies of the other alleles are similarly shown to be constant.

6.4.5. a. Let p, q, and r denote the frequencies of the alleles I^A, I^B, and I^O. Assuming random mating in the population,

$$p^2 + 2pr = .32,\ q^2 + 2qr = .15,\ 2pq = .04,\ r^2 = .49.$$

Solving these gives $r = .7$, $p = .2$, and $q = .1$. Note that even though there are 4 equations in only 3 unknowns here, these values makes all equations hold.
b. The equations to be solved are

$$p^2 + 2pr = .40,\ q^2 + 2qr = .11,\ 2pq = .05,\ r^2 = .44.$$

From the last we find $r = .6633$. Then the first gives $p = .2532$, and the second gives $q = .0783$. With these values $2pq = .0396$, so the third equation is *not* satisfied. (Also, $p + q + r \neq 1$.) Thus the system has no exact solution.
It could be that the population is not in a Hardy-Weinberg equilibrium, or that the data is flawed. Given the relative ease of collecting bloodtype data, and the doubtfulness of the random mating assumption applying to the U.S. population, the first is more likely.

6.4.6. Assume allele a_1 is dominant over a_2 and a_3, and a_2 dominant over a_3. Let p_1, p_2, and p_3 denote their frequencies in the population. Then knowing the frequency of the phenotype associated to a_3 would let us solve for p_3. Knowing

the frequency of the phenotype associated to a_2 along with p_3 would let us solve for p_2. Since $p_1 + p_2 + p_3 = 1$, we could then determine p_1. Thus knowing two phenotype frequencies is sufficient. Since there are two independent variables (p_2 and p_3, say), knowing a single phenotype frequency is not enough.
In general, for a gene with n alleles, at least $n-1$ phenotype frequencies must be known.

6.4.7. a. $p = (2Np_1 + 2Np_2)/(4N) = (p_1 + p_2)/2$
b.After the flood, a^+a^+ has frequency $(p_1^2 + p_2^2)/2$, a^+a has frequency $p_1(1 - p_1) + p_2(1 - p_2)$, and aa has frequency $((1 - p_1)^2 + (1 - p_2)^2)/2$.
A Hardy-Weinberg equilibrium would predict the three frequencies were $(p_1 + p_2)^2/4$, $(p_1 + p_2)(2 - p_1 - p_2)/2$, and $(2 - p_1 - p_2)^2/4$.
These disagree (for most values of p_1, p_2) since the population has not yet undergone random mating. There is no reason to expect a Hardy-Weinberg equilibrium.

6.4.8. The model becomes

$$p_{t+1} = \frac{p_t^2 + p_t q_t}{p_t^2 + 2p_t q_t + q_t^2} = \frac{p_t(p_t + q_t)}{(p_t + q_t)^2} = p_t.$$

This means the frequencies are unchanging, and in Hardy-Weinberg equilibrium. Note that here there is no selection operating, but mating is still random.

6.4.9. a. The model shows a gradual increase in frequency of A, toward fixation at $p = 1$. Thus a is eliminated ultimately. It appears that $p = 1$ is a stable equilibrium, and $p = 0$ an unstable one.
b. The model shows a gradual decrease in frequency of A, toward elimination at $p = 0$. Thus A is eliminated ultimately. It appears that $p = 0$ is a stable equilibrium, and $p = 1$ an unstable one.
c. If $p_0 > .5$, the frequency of A increases, toward $p = 1$; if $p_0 < .5$, the frequency of A decreases, toward $p = 0$. Thus the model shows a gradual increase in the frequency of whichever allele is initially more common. Eventually that allele is fixed in the population, while the other dies out. There are stable equilibria at $p = 0$ and 1, and an unstable one at $p = .5$.
d. If $p_0 > .5$, the frequency of A decreases, toward $p = .5$; if $p_0 < .5$, the frequency of A increases, toward $p = .5$. Thus the model shows movement toward an equal proportion of both alleles. While $p = .5$ is a stable equilibria, there are unstable ones at $p = 0$ and 1.

6.4.10. With $w_{AA} = 0$, $w_{Aa} = w_{aa} = 1$, simulations show that the frequency of the allele declines to 0, and that eventually the allele is eliminated from the population. (See problem 6.4.11.)

6.4.11. a. The parameters indicate homozygous dominants do not reproduce, while heterozygotes have no selective disadvantage relative the homozygous recessives. (See problem 6.4.10.)
b. $p_{t+1} = \dfrac{p_t q_t}{2p_t q_t + q_t^2} = \dfrac{p_t}{2p_t + q_t} = \dfrac{p_t}{p_t + 1}$, if $q_t \neq 0$ (or $p_t \neq 1$).
c. $p_1 = \dfrac{p_0}{p_0 + 1}$, so $p_2 = \dfrac{\frac{p_0}{p_0+1}}{\frac{p_0}{p_0+1} + 1} = \dfrac{p_0}{p_0 + (p_0 + 1)} = \dfrac{p_0}{2p_0 + 1}$. In general, if $p_t = \dfrac{p_0}{tp_0 + 1}$, then $p_{t+1} = \dfrac{\frac{p_0}{tp_0+1}}{\frac{p_0}{tp_0+1} + 1} = \dfrac{p_0}{p_0 + (tp_0 + 1)} = \dfrac{p_0}{(t+1)p_0 + 1}$.
Note that this shows that as $t \to \infty$, $p_t \to 0$ so such an allele will die out under random mating.

6.4.12. a. The homozygous recessives have no progeny, while heterozygotes are at no relative advantage to homozygous dominants.
b. $p_{t+1} = 1/(2 - p_t)$
c. $p_t = (t-(t-1)p_0)/((t+1)-tp_0)$, thus as $t \to \infty$, $p_t \to 1$ and the dominant allele becomes fixed.

6.4.13. a. If p is an equilibrium, then

$$p = ((w_{AA}p^2 + w_{Aa}p(1-p))/(w_{AA}p^2 + 2w_{Aa}p(1-p) + w_{aa}(1-p)^2),$$

so

$$w_{AA}p^3 + 2w_{Aa}p^2(1-p) + w_{aa}p(1-p)^2 - w_{AA}p^2 - w_{Aa}p(1-p) = 0.$$

b. $p = 0$ and 1 are two of the three equilibria.
c. The cubic polynomial in (a) factors as

$$p(p-1)((w_{AA} - 2w_{Aa} + w_{aa})p + (w_{Aa} - w_{aa})) = 0.$$

d. The third equilibrium satisfies $(w_{AA} - 2w_{Aa} + w_{aa})p + (w_{Aa} - w_{aa}) = 0$, so straightforward algebra shows it's given by the formula stated.

6.4.14. a. If either $w_{aa} - w_{Aa}$ or $w_{AA} - w_{Aa}$ is 0, then we get one of the two first two equilibria, 0 or 1. Otherwise, since the third equilibrium can be written $p = 1/\left(1 + \dfrac{w_{AA} - w_{Aa}}{w_{aa} - w_{Aa}}\right)$, for it to lie between 0 and 1 we must have $\dfrac{w_{AA} - w_{Aa}}{w_{aa} - w_{Aa}} > 0$. This means the numerator and denominator have the same sign, which is equivalent to the given condition.
b. Again thinking in terms of the signs of the factors in (a), these are seen to be equivalent.

6.4.15. a. Homozygote advantage results in a trend toward fixation of whichever allele is most common initially in the population, and elimination of the other.
b. Heterozygote advantage results in a trend toward non-zero proportions of both alleles in the population. If $w_{AA} = w_{aa}$, then equal proportions of each will occur in the long run, but more generally the allele with the greater homozygous relative fitness is the more common.

6.4.16. The formula for mean fitness shows it is a weighted average of the relative fitness parameters, weighted by the frequencies of the corresponding genotypes in the population. The result that $\overline{w}_{t+1} \geq \overline{w}_t$ says that the mean fitness of a population can only increase. This is a quantitative statement of "survival of the fittest."

6.4.17. For low population sizes, the graphs are quite jagged, and often result in the fixation of one allele and elimination of the other. There is some tendency for the initially more common allele to be fixed, but many exceptions occur. For midsize populations, the size of the fluctuations (as a percentage of population) is much smaller, and fixation/elimination occurs more rarely. For large populations, the fluctuations are quite small (though still present) and one rarely observes fixation/elimination unless initially one allele was quite rare. Despite the fluctuations, the overall trend for large populations is that the frequencies remain roughly constant. This supports the idea that genetic drift is only a significant factor for small populations (or rare alleles).

6.4.18. Introducing selection into the genetic drift model results in biasing the drift along the trends that would occur in a selection model without drift. For example, with dominant advantage, seldom does drift result in elimination of the

dominant allele, although it does occasionally. An interesting case is homozygous advantage, where the tendency of drift to lead to fixation/elimination is tempered, so that both alleles persist for much longer.

6.4.19. By the definition of expected value, $E = 0(.0625) + 1(.25) + 2(.375) + 3(.25) + 4(.0625) = 2$, or, since the random variable has a binomial distribution, $E = 4(1/2) = 2$. This expected value is exactly the Hardy-Weinberg equilibrium. Even in the presence of genetic drift, the idea of a Hardy-Weinberg equilibrium is still valid for expected values of frequencies of alleles.

6.4.20. a. H should tend toward 0, regardless of which allele is fixed, since $H = 2pq$ and either p or q approaches 0.

b. H declines to 0 exponentially. The larger the population size N, the slower the decline. Varying the initial value of H_0 does mean it may take more or less time for H to reach any specified value, but does not affect the rate of exponential decay.

c. $H_t = (1 - \frac{1}{2N})^t H_0$

CHAPTER 7

Infectious Disease Modeling

7.1. Elementary Epidemic Models

7.1.1. a. For larger values of S_0, the orbits first show an increase in I and a decrease in S, but later show I declining to 0 while S decreases to some limiting value (which depends on the initial values). For smaller values of S_0, only the later behavior is observed. For the given parameter values, $S_0 = 50$ appears to be the dividing line between these behaviors.

b. The behavior is roughly the same as α is varied. However, a larger value of α causes orbits to give early increases in I and decreases in S for more values of S_0. For instance $\alpha = .002$ gives this behavior if $S_0 > 25$. Larger α also causes I to have a larger maximum value, the maximum I to be reached in fewer time steps, and the number of susceptibles remaining after the epidemic subsides to be smaller. These are all signs of a more severe epidemic. Smaller α has the opposite effect.

c. Varying γ also produces roughly the same behavior. However, a smaller value of γ causes orbits to give early increases in I and decreases in S for more values of S_0. For instance $\gamma = .03$ gives this behavior if $S_0 > 30$. Smaller γ also causes I to have a larger maximum value, the maximum I to be reached in fewer time steps, and the number of susceptibles remaining after the epidemic subsides to be smaller. These are all signs of a more severe epidemic. Larger γ has the opposite effect.

At least for the given parameter values, varying γ seems to produce a smaller change than varying α.

7.1.2. Since $S + I + R = N$ is constant, $S + I \leq N$, so initial values must be below the line $S + I = N$.

7.1.3. The SIR model is not appropriate for malaria, since contact between infected and susceptible humans is not the source of new infections. A good malaria model would require tracking both humans and mosquitoes.

7.1.4. a. 1/8; b. $\gamma = 1/m$

7.1.5. a. $S_0 = 99$, $I_0 = 1$, $R_0 = 0$, so $S_1 = 0$, $I_1 \geq 99$ (depending on value of γ). On later days, $S = 0$, and I will have declined, while R grew. Since $\alpha = 1$, a single infective can in one time step infect the full population.

b. If $\alpha = 1$ and $I_0 = 5$, then $S_0 = 95$ so $S_1 = 95 - 1(95)(5) = -380$. However, S should be between 0 and 100 to make sense biologically.

c. Since we need $0 \leq S_0 - \alpha(S_0)(I_0) = S_0(1 - \alpha I_0)$, we must have $I_0 \leq 1/\alpha = 1/.1 = 10$.

d. $\alpha \leq 1/I_0$.

7.1.6. a. Instead of αSI, use $\alpha S((1 - q)I) = \alpha(1 - q)SI$. When $q = 0$ the model is the usual SIR.

b. $\alpha' = (1 - q)\alpha$

c. As q is increased, the model behavior changes as described in problem 7.1.1 when α is increased. This is as expected, since quarantining should reduce contact between susceptibles and infectives, thus slowing or stopping disease transmission. For the given parameter values, when $q \geq .5$, there is no increase in I, regardless of the value of I_0.

7.1.7. a. The vaccinated are put into the removed class initially, so vaccinations must occur before $t = 0$.

b. $R_0 = qN = 100q$, $S_0 = N - I_0 - R_0 = (1-q)N - 1$. When $q = 0$ the model is the usual SIR.

c. As q increases, we are effectively using a smaller value of S_0. For larger q, we see fewer time steps in which I increases in the orbit, as we might expect if a significant part of the population is vaccinated. For the given parameter values, if $q \geq .49$ then I never increases, regardless of the value of I_0.

7.2. Threshold Values and Critical Parameters

7.2.1. a. Since $S + I + R = N$ is constant, if S and I don't change, neither does R.

b. $\Delta S = 0$ implies $S = 0$ or $I = 0$. $\Delta I = 0$ implies $\alpha S - \gamma = 0$ or $I = 0$. Thus, assuming $\alpha, \gamma \neq 0$, the equilibria are $I = 0$ and with S having any value. (What if $\alpha = 0$ or $\gamma = 0$?) The points where $I = 0$ are obviously equilibria, since if there are no infectives, no susceptible can fall ill, and no one can enter the removed class.

c. These equilibria might not be stable: if $S > \gamma/\alpha$ and I is perturbed to a small positive value then an epidemic will begin. However, if $S < \gamma/\alpha$, then a perturbation will not begin an epidemic, but the population values may move toward a different but nearby equilibrium $(S^*, 0)$. While this is technically not stability, it is close in spirit.

7.2.2. a. 1/37; b. 7/37

7.2.3. From phase plane plots, or time plots, the value of S when I peaks appears to be around 170 to 180. From numerical population values, for $N = 300$, we find $S = 177.28$. Since $\rho = 178.5$ there is not exact agreement. This is because for a discrete model, S is unlikely to exactly hit the threshold value, so I will increase until S exceeds the threshold value. The threshold will only be within the range of S values one time step from when the numerical maximum of I is observed.

7.2.4. a. $\Delta I/I = \alpha S - \gamma$

b. The graph is an upward sloping line, crossing the $\Delta I/I$-axis at $-\gamma$ and the S-axis at γ/α. This shows the per capita growth rate for infectives is only positive when $S > \gamma/\alpha = \rho$, so only then can an epidemic occur.

7.2.5. a. $\gamma = 1/4$, using time steps of one day.

b. Since $I_0 = 1$, $S_0 = 99$, so an epidemic occurs if $\rho < 99$. Thus for $\alpha = \gamma/\rho > .25/99 \approx .002525$ an epidemic will occur.

c., d. Let T denote the number of time steps until I reaches its maximum. Then the table shows T decreases as α increases.

α	.003	.005	.01	.0125
T	33	19	9	8
$\mathcal{R}_0$	1.188	1.98	3.96	4.95
ρ	83.33	50	25	20

7.2.6. a. $\gamma = .3$

b., c. The plotted epidemics are numbered in order of decreasing maximum values of I. Estimates may vary from those given here.

Epidemic	1	2	3	4
ρ	2500	4100	6000	7500
α	.00012	.000073	.00005	.00004
$\mathcal{R}_0$	4	2.44	1.67	1.33
T	10	14	22	40(?)
S_∞	100	1050	3050	5400

Increasing the transmission coefficient α results in a greater peak number of infectives, a shorter time until that peak is reached, and a smaller number of individuals remaining disease free through all times. All of these things are as one might expect if disease transmission is made more likely.

7.2.7. a., b. The plotted epidemics are numbered in order of decreasing maximum values of I. m denotes mean infectious period. Estimates may vary from those given here.

Epidemic	1	2	3	4
ρ	800	2600	3600	6500
γ	.064	.208	.288	.520
m	15.625	4.808	3.472	1.923
$\mathcal{R}_0$	12.487	3.842	2.775	1.537
T	12	16 (?)	18(?)	large
S_∞	near 0	350	1000	4200

Increasing the removal rate γ results in a smaller peak number of infectives, a longer time until that peak is reached, and a larger number of individuals remaining disease free through all times. All of these things are as one might expect if sick individuals recover (or die) sooner.

7.2.8. a. While the infective class size does not change on the first step, it decreases on the second and later steps.

b. If $\Delta I = 0$ initially, there is a balance between infectives recovering and susceptibles becoming infectives. At the next time step, there are fewer susceptibles and the same number of infectives, so this balance is no longer maintained. Fewer susceptibles become ill than the number of infectives that are removed, and the disease begins to die out.

c. If at time 0, $\Delta I = 0$, then $S_0 = \gamma/\alpha$, so $S_1 = S_0 - \alpha S_0 I_0 < S_0 = \gamma/\alpha$. Thus at time 1, $\Delta I < 0$.

7.2.9. a., b. The S-nullcline is the two coordinate axes $S = 0$ and $I = 0$. The I-nullcline is the S-axis ($I = 0$) and the vertical line $S = \gamma/\alpha$, which is 984.83 for part (a). To the left of this line, arrows point to the left and down. To the right of the line, they point to the left and up.

7.2.10. The effective relative removal rate is $\rho' = \gamma/((1-q)\alpha)$. Increasing q toward 1 causes ρ' to grow toward infinity. For any fixed number of susceptibles S_0, we could increase q so that $S_0 < \rho'$, and then no epidemic could occur.

7.2.11. With enough data from an epidemic, plots such as those in Figure 7.3 could be created. With γ estimated from the mean infectious period (which could be estimated by careful surveillance of a relatively small number of individuals), we could proceed as in problem 7.2.6: ρ could be estimated from the plot, and α from ρ and γ. This is all very indirect, so we should not be too confident of the precise value obtained.

Collecting data to get a good plot as in Figure 7.3 is hard, as it requires good figures at many times on the size of least two of the three classes. Counting infectives might be easiest (if the illness has clear symptoms throughout the infective period and no social stigma is associated with it). Counting susceptibles or removed is much harder. Should self-reporting be accepted as accurate? Can medical tests confirm who has had the disease in the past? Are there naturally immune individuals who we have no way of detecting?
Note that collecting data on a special population (e.g., students at one college) may be easier, but transmission and removal parameters for such a population may not be correct for a different population.

7.3. Variations on a Theme

7.3.1. If $\alpha > 0$ and $I_0 > 0$, the model shows what appears to be logistic growth in I, toward an equilibrium of $I = N$. Larger values of α produce quicker movement toward the equilibrium. The phase plane plot shows the orbit lying on the straight line $S + I = N$.

7.3.2. a. The equilibria are the points of the form $(0, N)$ and $(N, 0)$. For the first, the entire population has become infected, and for the second, everyone is disease free.
b. The nullclines are the two axes. Arrows point left and up. This suggests the equilibria $(S, 0)$ are unstable, and that all orbits starting elsewhere will move toward equilibria $(0, I)$ in which the entire population is infected.

7.3.3. a. $I_{t+1} = I_t + \alpha(N - I_t)I_t = I_t + \alpha N I_t(1 - I_t/N)$. Notice this second form is clearly the logistic model.
b. Time plots from `onepop` are the same as produced by `twopop` for the class I. There is no phase plane plot produced, but it was not particularly helpful in providing understanding anyway.

7.3.4. Typically the model moves toward a constant non-zero value of both S and I, though these values depend on the parameters. Time plots appear similar to those produced by the logistic model as an equilibrium is approached, though not initially. Increasing α or decreasing γ tends to lower the value which S approaches, as well as make it approach more quickly.

7.3.5. a. The equilibria are $(N, 0)$ and $(\gamma/\alpha, N-\gamma/\alpha)$. The first of these simply means the entire population is free of disease. The second represents an endemic level of disease, in which new infections must be balanced by recoveries.
b. The nullclines for both S and I are the S-axis and the vertical line $S = \gamma/\alpha$. Both nullclines are identical since $S + I = N$ is constant. To the left of the vertical line, arrows point right and down; to the right of the line they point left and up. This suggests most orbits tend to the equilibria on the vertical line, representing endemic disease. Note that since $S + I = N$ is constant, all orbits will lie on this straight line, and we could analyze the model by focusing on S or I alone, never using a phase plane.

7.3.6. a. $S_{t+1} = S_t + (\gamma - \alpha S_t)(N - S_t)$
c. With $N = 1$, $\alpha = .1$, $\gamma = .05$, simulations typically approach the equilibrium of $\gamma/\alpha = .5$ without overshooting. If both α and γ are increased by the same factor f, the equilibrium doesn't change, but the behavior does. For f around 20 to 40, oscillatory approaches to equilibrium occur. For $f = 50$ a 4-cycle

appears. The logistic model appears to be lurking somewhere. See problem 7.3.10.

7.3.7. a. There is no analogous threshold for the SI model, since all populations with $I_0 > 0$ show I increasing to the full population size.
b. For SIS, $\Delta I > 0$ exactly when $S > \gamma/\alpha$, just as for SIR. Note however that while knowing when I will increase is important for the SIS model, usually it is secondary to understanding the endemic level of infection $I = N - \gamma/\alpha$.

7.3.8. The same definition works as for SIR.

7.3.9. If a cure is developed for the disease, then infectives might be able to either move back into the susceptible class, for an SIS model, or into a removed class, for an SIR model.

7.3.10. a. Use $S = N - I$ in the formula for I_{t+1}.
b. The formula in (a) is a logistic one, with $K = N - \gamma/\alpha$. Thus 0 and $N - \gamma/\alpha$ are equilibria. Since the second is between 0 and N, it represents an endemic level of infection, with γ/α being the number of susceptibles at that equilibrium.
c. Since $\alpha N - \gamma$ plays the role of r in the logistic model, the endemic equilibrium will be stable if $|1 - \alpha N + \gamma| < 1$, or $0 < \alpha N - \gamma < 2$.
d. There will be an oscillatory approach to equilibrium if $1 < \alpha N - \gamma < 2$. It is conceivable that such behavior might occur naturally, though the SIS model itself is such an oversimplification of a real diseases dynamics that its probably best not to make too much of this.

7.3.11. a. $\sigma = N\alpha/\gamma$
b. $\mathcal{R}_0/\sigma = S_0/N$
c. $\mathcal{R}_0/\sigma \approx 1$, though it is actually slightly less than 1.
d. $\mathcal{R}_0$ is slightly smaller than σ.

7.3.12. $\Delta i > 0$ exactly when $s > \gamma/\beta$, so $\gamma/\beta = 1/\sigma$ is the threshold for s_0.

7.3.13. a. $\sigma = 2$
b. $1 - 1/\sigma = .5$, so half the population must be vaccinated.
c. Solving $(1 - .9q) < .5$ gives $q > .5556$, so about 56% of the population must be vaccinated.

7.3.14. To prevent I from growing, we need $S < \rho$, so at least $N - \rho = N - \gamma/\alpha$ individuals must be immunized.

7.3.15. With a contact number less than one, no epidemic will occur even if no vaccinations are given. $1 - 1/\sigma$ will be negative exactly when $0 < \sigma < 1$.

7.3.16. a. $\Delta s = -\beta(1-q)is$; $\Delta I = \beta(1-q)is - \gamma i$; $\Delta r = \gamma i$
b. An epidemic occurs exactly when $s > \gamma/((1-q)\beta)$.
c. To prevent epidemics for all values of $s \leq 1$, we need to make $\gamma/((1-q)\beta) \geq 1$, or $q \geq 1 - \gamma/\beta$.

7.3.17. For an SIS disease, previous infection does not confer immunity, so the usual idea behind vaccination simply doesn't apply. However, one could imagine an immunization that provided temporary protection from infection, though this would require a more complicated $SIRS$ model.

7.3.18. a. These diseases are not transmitted person-to-person but rather from a natural reservoir of disease to a person. If a person had no contact with other humans, the risk of getting these diseases would not necessarily be reduced.
b. If vaccines are cheap and safe of side-effects and the risk of infection is high, vaccinate everyone. If vaccines are expensive or dangerous, vaccinate only those most at risk. These strategies are directed solely at individual protection.

7.3.19. A few factors to consider are seriousness of the disease, health risks from the vaccination, sizes of α and γ, whether subpopulations are at greater risk of either being infected or infecting others.

7.3.20. a. $\pm\beta si$ are the usual mass action terms describing new infectives arising from susceptibles who came in contact with infectives. $\pm\gamma i$ are the usual terms describing infectives recovering and returning to the susceptible class. $-(\mu+\nu)i$ describes the death of infectives, with μi occuring due to natural causes and νi due to the disease. μi also describes births to infectives, but $p\mu i$ are disease-free births, and the remaining $(1-p)\mu i$ are born with the disease.

b. Since $\Delta s+\Delta i = -\nu i \neq 0$, the total population size does not remain constant.

c. This is a modified SIS model, since infectives can be 'removed' back into the susceptible class. If we kept track of those who died in another class r, it would have features of an SIR model as well.

d. Deaths due to natural causes of susceptibles would be described by $-\mu s$, while μs would describe births to susceptibles. Since all these births would be disease free, both μs and $-\mu s$ would be added to the formula for Δs, producing no net change in the formula.

7.3.21. a. Answers may vary. This problem is unreasonably open-ended, and could make a substantial project.

b. The MMR vaccination need not be given earlier since infants are in the M class.

7.4. Multiple Populations and Differentiated Infectivity

7.4.1. Other sexually transmitted diseases in heterosexual populations might also be described by this model, provided treatment or recovery is possible and former infection provides no future immunity. Examples include syphilis and genital lice. Many other STDs, such as herpes or AIDS fail to meet the basic assumptions.

7.4.2. a. Since $\alpha^m > \alpha^f$, from any one contact with an infective males are more likely to catch the disease than females. Since $\gamma^m > \gamma^f$, males recover faster than females.

b. $N^f/\rho^f = 1.2857$; $N^m/\rho^m = 1.8$. These values are both greater than 1, so an endemic level of infection should exist. Computer simulation also indicates it exists.

c. $N^f N^m - \rho^f\rho^m = 8.52 \times 10^7$. Endemic equilibrium values are $I^f = 3739.8$ and $I^m = 4646.5$, so $S^f = 6260.2$ and $S^m = 10353.5$. A computer simulation indicates this as well.

d.The equilibrium appears stable.

7.4.3. a. Females are more likely to become infected than males when either has contact with an infective of the other sex. Females recover more quickly, though, with a mean infectious period of 2 days, rather than the 5 days of males.

b. $N^f/\rho^f = 2.2$; $N^m/\rho^m = .46$. While these are not both greater than 1, their product is, so the disease may remain endemic at an equilibrium. (The equilibrium values, however, are quite close to 0, so a hastily read computer simulation can be misleading.)

If $N^f = 50$ and $N^m = 450$, then $N^f/\rho^f = 1.1$; $N^m/\rho^m = .5175$, so the product is less than 1 and no endemic equilibrium exists. If $N^f = 250$ and $N^m = 250$,

then $N^f/\rho^f = 5.5$; $N^m/\rho^m = .2875$, so the product is greater than 1 and an endemic equilibrium exists.

7.4.4. a. Let $N = N^e + N^y$. Then $\Delta S^e = -\alpha^e S^e(N - S^e - S^y) + \gamma^e(N^e - S^e)$, $\Delta S^y = -\alpha^y S^y(N - S^e - S^y) + \gamma^y(N^y - S^y)$.

b. Since elderly are more likely to be infected from contact with an infective, $\alpha^e = .0003$ and $\alpha^y = .0001$. Since the young recover more quickly, $\gamma^y = .21$ and $\gamma^e = .05$.

c. The orbits quickly move toward an endemic equilibrium, with $S^e \approx 112$ and $S^y \approx 684$.

7.4.5. a. Solving for I^f in $\alpha^f(N^f - I^f)I^m - \gamma^f I^f = 0$ yields the formula. By symmetry (or similar work) $I^m = N^m I^f/(I^f + \rho^m)$ at equilibrium.

b. Substituting yields

$$I^f = \frac{N^f \frac{N^m I^f}{I^f + \rho^m}}{\frac{N^m I^f}{I^f + \rho^m} + \rho^f},$$

and then solving for I^f produces the formula in the text.

c. Simply replace all superscripts m with f and f with m.

7.4.6. A reasonable expectation is that $(0, 0)$ be an unstable equilibrium if an endemic equilibrium exists, and a stable one if there is no endemic equilibrium.

Linearizing at $(0, 0)$ yields $\begin{pmatrix} i^f_{t+1} \\ i^m_{t+1} \end{pmatrix} = \begin{pmatrix} 1-\gamma^f & \alpha^f N^f \\ \alpha^m N^m & 1-\gamma^m \end{pmatrix} \begin{pmatrix} i^f_t \\ i^m_t \end{pmatrix}$. The eigenvalues of this matrix are the roots of

$$\lambda^2 - (2 - \gamma^f - \gamma^m)\lambda + ((1-\gamma^f)(1-\gamma^m) - \alpha^f \alpha^m N^f N^m) = 0.$$

Using the quadratic formula, and a little algebra, the roots are

$$1 - \frac{\gamma^f + \gamma^m}{2} \pm \sqrt{\left(\frac{\gamma^f + \gamma^m}{2}\right)^2 + \alpha^f \alpha^m (N^f N^m - \rho^f \rho^m)}.$$

Thus one of the roots will be larger than 1 if $N^f N^m - \rho^f \rho^m > 0$. This is exactly the criteria for an endemic equilibrium, so if such an equilibrium exists, $(0, 0)$ is unstable.

We leave additional analysis as a challenge to the student.

CHAPTER 8

Curve Fitting and Biological Modeling

8.1. Fitting Curves to Data

8.1.1. $f(t) = 200e^{\ln \frac{129}{200} t} = 200e^{-.4385t}$; error $\approx (200, 129, 58, 33) - (200, 129, 53.7, 34.6)$ $\approx (0, 0, 4.33, -1.62)$; $TD \approx 0 + 0 + 4.33 + 1.62 \approx 5.95$; $SSE \approx 0^2 + 0^2 + 4.33^2 + (-1.62)^2 \approx 21.4$. Thus using TD we judge f to be a better fit than f_1, and using SSE we judge f to be a better fit than f_2.

8.1.2. a. One approach is to use k and each data point to get an estimate for $\ln a$, and then average these estimates to get one that might be better. Using $k = -.467$ from the text, and $\ln y = kt + \ln a$, the four data points yield 5.30, 5.33, 5.46, and 5.36 as estimates of $\ln a$. The average is approximately 5.36, giving $a \approx 213.3$.
b. The curve $y = 213.3e^{-.467t}$ has error $(-13.3, -4.71, 5.45, 0.0596)$ so $SSE \approx$ 228.8. This makes it a worse fit, by SSE measure, than f_2.

8.1.3. a. The third line, $y = 3x + 1.1$, appears to fit the data slightly better.
b. The SSEs for the three lines are 2.66, 2.69, and 1.89, which quantifies the judgment that the third line is the best fit.
c. From the graph, we might think increasing the slope of the third line slightly would improve the fit. Indeed, the guess $y = 3.05x + 1.1$ produces an SSE of 1.51, for a better fit.

8.1.4. a. $y_1 = (1 - r)y_0$, so $y_2 = (1 - r)y_1 = (1 - r)^2 y_0$ and, continuing this pattern shows $y_t = (1 - r)^t y_0$.
b. Letting $k = \ln(1 - r)$ and $a = y_0$, we have $1 - r = e^k$, so $y_t = y_0(1 - r)^t = a(e^k)^t = ae^{kt}$.
c. If r is the percentage of drug absorbed in one time step, it must be between 0 and 100%=1. Since $0 < r < 1$, then $0 < 1 - r < 1$, so $k = \ln(1 - r) < 0$.

8.1.5. a. $y = -35.5t + 164.5$ goes through the middle two points. The SSE for it is 1370.5, so it is a much worse fit than f_2.
b. One scheme would be to find the slopes between each pair of consecutive data points, and then average these to give a possibly better choice of m in $y = mt + b$. Then use each data point to calculate an estimate of b, and average these for a possibly better estimate. This gives slopes of -71, -35.5, and -25, for an average of -43.83. Using $y = -43.83t + b$, the data points give estimates of b 200, 172.83 189.49, and 208.32, for an average of 192.66. The line $y = -43.83t + 192.66$ has $SSE = 702.39$. While this line is a better fit than that of part (a), it's still considerably worse than f_2.

8.1.6. a. The exponential or power model appears to be better than the linear one.
b. The semilog plot appears to be roughly linear, with slope approximately $k = .82$.
c. The log-log plot also appears roughly linear, with slope approximately $n = 2.3$. (However, comparison to the plot in (b) suggests the exponential model is more appropriate.)

8.1.7. a. If a horizontal line lies below all three points, moving it upward until it passes through the bottom two will decrease the TD, since it decreases the deviation at each point.
b. Similar to (a).
c. Lowering any horizontal line between the points increases the deviation from $(1, C)$, but decreases the deviation from both $(0, 0)$ and $(2, 0)$. Since all three changes have the same magnitude, the net effect is a decrease in TD from lowering the line. In formulas, for the line $y = b$ with $0 \le b \le C$, $TD = b + (C - b) + b = C + b$, and this is made smallest by taking $b = 0$.
d. Regardless of the value of C, $y = 0$ minimizes TD for this data. Thus the height of the middle data point has no effect on the choice of best-fit line using TD.
e. Hint: Repeat the argument above, but at each step consider the effect on TD of tilting the line.

8.1.8. a. Similar to problem 8.1.7(a) and (b).
b. The line $y = b$ with $0 \le b \le 1$ has $TD = b + (1 - b) + (1 - b) + b = 2$.
c. Any line $y = b$ with $0 \le b \le 1$ minimizes TD, so all infinitely-many of these are best-fit according to TD.

8.2. The Method of Least Squares

8.2.1. The equations $1 = -m + b$, $3 = 0 + b$, and $4 = m + b$ yield

$$\begin{pmatrix} 1 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix}.$$

The normal equations are $\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 3 \\ 8 \end{pmatrix}$, with solution $(m, b) = (\frac{3}{2}, \frac{8}{3})$. The least-squares best-fit line is $y = \frac{3}{2}x + \frac{8}{3}$.

8.2.2. With $A = \begin{pmatrix} 3 & 1 \\ 4 & 1 \\ 5 & 1 \\ 6 & 1 \\ 7 & 1 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 120 \\ 116 \\ 114 \\ 109 \\ 106 \end{pmatrix}$,
a. The MATLAB command `inv(A'*A)*A'*b`, gives output -3.5; 130.5, for a best-fit line of $y = -3.5x + 130.5$.
b. The MATLAB command `A\b` produces the same output as in (a). This is an easier way to find least-squares solutions with MATLAB.
c. The command `polyfit([3,4,5,6,7],[120,116,114,109,106],1)` produces the output -3.5, 130.5, giving the same line.

8.2.3. a. $\ln y = -.44t + 5.31$
b. Exponentiating (a) gives $y = e^{5.31}e^{-.44t} = 202.35e^{-.44t}$.

8.2.4. a. The semilog plot is roughly linear, indicating exponential growth. The model $P_{t+1} = \lambda P_t$ produces exponential growth.
b. $\ln y = .5554t + 5.1301$
c. $y = 169e^{.5554t} = 169(1.7426)^t$.
d. 1.7426

8.2.5. a. Answers will vary, but three experiments produced best-fit lines with $(m, b) = (.6915, 2.0208)$, $(.6844, 2.2170)$, and $(.7260, 2.0246)$. None of these give exactly

the line $y = .7x + 2.1$, but the slopes and intercepts are quite close. Notice they can be larger or smaller than the 'true' values.

b. The '.3' controls the size of the random error introduced. Replacing it with '3', the points lie less close to a line, and the recovered least-squares line is considerably less similar to $y = .7x + 2.1$.

c. If the '10' is replaced with a '3', the least-squares line tends to be less similar to the 'true' line, while a '30' tends to make it more similar. This is reasonable, since the more data points are used, the more accurate our recovery of the underlying trend should be.

8.2.6. a. Given three data points, plugging them into a quadratic equation would give 3 equations (one for each data point) in 3 unknowns (the 3 coefficients of the quadratic). That leads to a 3×3 matrix equation $A\mathbf{c} = \mathbf{b}$. Since most square matrices have inverses, there is probably one and only one solution, $\mathbf{c} = A^{-1}\mathbf{b}$. The key point here is that the matrix is square.

b. For n data points, an $(n-1)$th degree polynomial can usually be found whose graph passes through the points. This is because an $(n-1)$th degree polynomial has n coefficients, so we are led to a square matrix equation.

8.2.7. a. The least squares line is $y = -.1611x + 7.4264$.

b. $(\bar{x}, \bar{y}) \approx (5.75, 6.5)$. This lies on the least-squares line, to within round-off error.

c. All experiments lead to $(\bar{x}, \bar{y})$ lying on the least-squares line. See problem 8.2.11.

8.2.8. Similar to development in text.

8.2.9. a. $\frac{d}{dm}SSE(m, \hat{b}) = -2\sum(y_i - mx_i - \hat{b})x_i$. Since this must be 0 when $m = \hat{m}$, we obtain $\sum(y_i - \hat{m}x_i - \hat{b})x_i = 0$.

b. Similar to (a).

8.2.10.

$$\begin{pmatrix} \hat{m} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{pmatrix}^{-1} \begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix}$$
$$= \frac{1}{n(\sum x_i^2) - (\sum x_i)^2} \begin{pmatrix} n & -\sum x_i \\ -\sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix}$$

8.2.11. Note $\bar{x} = \frac{\sum x_i}{n}$, $\bar{y} = \frac{\sum y_i}{n}$, and using the values of $\hat{m}$ and $\hat{b}$ from problem 8.2.10, algebra shows $\hat{m}\bar{x} + \hat{b} = \bar{y}$.

8.2.12. Passing from the first to the second line uses the fact that for invertible matrices $(MN)^{-1} = N^{-1}M^{-1}$. However, if A is not square (as it usually isn't in applications) then A and A^T cannot have inverses, so the step is invalid. Note that A^TA is square even if A isn't, so $(A^TA)^{-1}$ can exist even when A^{-1} and $(A^T)^{-1}$ don't.

8.3. Polynomial Curve Fitting

8.3.1. Of the three plots, the log-log appears most linear, indicating a power function model is likely to be a good choice. Since the slope appears very close to 3, a cubic function $y = at^3$ is appropriate.

8.3.2. MATLAB work should produce output agreeing with the text.

8.3.3. a. $\begin{pmatrix} 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \\ 16 & 4 & 1 \\ 25 & 5 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1.1 \\ 8.7 \\ 19.8 \\ 39.5 \\ 64.7 \end{pmatrix}$

b. $\begin{pmatrix} 979 & 225 & 55 \\ 225 & 55 & 15 \\ 55 & 15 & 5 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 2463.60 \\ 559.40 \\ 133.80 \end{pmatrix}$

c. $y = 3.13x^2 - 2.97x + 1.26$

8.3.4. a. $\begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 2 \\ 9 \\ 1 \\ 4 \end{pmatrix}$; $\begin{pmatrix} 30 & 10 \\ 10 & 4 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 39 \\ 16 \end{pmatrix}$; $y = -.20x + 4.5$; $SSE = 37.8$

b. $\begin{pmatrix} 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \\ 16 & 4 & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 2 \\ 9 \\ 1 \\ 4 \end{pmatrix}$; $\begin{pmatrix} 354 & 100 & 30 \\ 100 & 30 & 10 \\ 30 & 10 & 4 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 111 \\ 39 \\ 16 \end{pmatrix}$; $y = -1x^2 + 4.8x - .5$; $SSE = 33.8$

c.

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 8 & 4 & 2 & 1 \\ 27 & 9 & 3 & 1 \\ 64 & 16 & 4 & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 2 \\ 9 \\ 1 \\ 4 \end{pmatrix};$$

$$\begin{pmatrix} 4890 & 1300 & 354 & 100 \\ 1300 & 354 & 100 & 30 \\ 354 & 100 & 30 & 10 \\ 100 & 30 & 10 & 4 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 357 \\ 111 \\ 39 \\ 16 \end{pmatrix};$$

$y = 4.33x^3 - 33.5x^2 + 77.17x - 46$; $SSE = 0$

8.3.5. a. A linear polynomial seems likely to fit well.

c. The polynomials of degree 2,3,4 don't appear to fit too much better than the linear one. While the degree 5 one passes through all data points, it curves around a lot to do so. The simplicity of the linear polynomial, together with the fact that it captures the data trend well, makes it the most reasonable choice (barring some other reason to think the data might not be linear).

8.3.6. a. The cubic curve is the first curve to appear to come very close to all the data points. The linear and quadratic curves are clearly worse fits, and the higher degree curves do not appear to be much better than the cubic. The cubic both has simplicity (low degree) and appears to fit well.

b. There is a more dramatic decrease (two orders of magnitude) in the SSE in passing from the quadratic to the cubic than in any other increase in degree, except 5th to 6th. However, with 7 data points we know a 6th degree polynomial can fit them exactly, so we shouldn't find that so compelling.

8.3.7. The pattern to the graphs is much like the preceeding problem. The cubic appears to fit better than lower degree curves, but little noticeable improvement in fit results from higher degree curves. For the SSE we see a decrease by an order of magnitude from linear to quadratic and again from quadratic to cubic. Higher degree curves produce much smaller changes in the SSE. Again, the cubic has a desirable combination of simplicity and good fit.

8.3.8. a. The equations to be solved are $3 = c$, $5 = c$, and $10 = c$, or $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} c = \begin{pmatrix} 3 \\ 5 \\ 10 \end{pmatrix}$. The normal equation is $3c = 18$, so $y = 6$ is the best-fit horizontal line.
b. $(3 + 5 + 10)/3 = 6$.
c. For y-coordinate data values $y_1, y_2, \dots y_n$, we'd like to solve $\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} c = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$. The normal equation is $nc = \sum y_i$, so the solutions has $c = \frac{\sum y_i}{n} = \bar{y}$, the average y-coordinate.

8.3.9. a. The best-fit cubic is $y = 0.13t^3 - 0.89t^2 + 61.84t - 9.74$. It's graph appears to fit the data remarkably well.
b. The data for subsequent times appears to be a reasonably well fit for about another 50 months or so, but then diverges from the curve quite seriously. The trend in the data seems very different from the curve in the long term.
c. The dynamics of the disease transmission might have changed, due to medical advances, changes in behavior, or changing social conditions; or perhaps enough of the high-risk population had already been infected to change the trend. A change in the surveillance definition of a case of AIDS in 1987 and 1993 makes comparisons across those dates difficult.
d. A 3rd or 4th degree polynomial can fit the data reasonably well (though it is a different 3rd degree than found in part (a)).

8.3.10. a. Since $S_t = N - I_t$, in the early stages of an epidemic I_t is small relative to N so $S_t \approx N$. Thus $I_{t+1} \approx I_t + \alpha N I_t = (1 + N\alpha) I_t$.
b. The approximate model in (a) is a linear one, leading to exponential growth: $I_t \approx (1 + N\alpha)^t I_0$ for small t.
c. For SIR, $I_{t+1} = I_t - \gamma I_t + \alpha I_t S_t$, with $S_t = N - I_t - R_t$. At the beginning of an epidemic, I_t and R_t are small, so $S_t \approx N$, and $I_{t+1} \approx (1 - \gamma + \alpha N) I_t$, a linear model. For SIS, the formula is the same (though $R_t = 0$ throughout).