LECTURE NOTES ON PROBABILITY, STATISTICS AND LINEAR ALGEBRA

C. H. Taubes
Department of Mathematics
Harvard University
Cambridge, MA 02138

Spring, 2010

CONTENTS

Preface

This is a very slight revision of the notes used for Math 19b in the Spring 2009 semester. These are written by Cliff Taubes (who developed the course), but re-formatted and slightly revised for Spring 2010. Any errors you might find were almost certainly introduced by these revisions and thus are not the fault of the original author.

I would be interested in hearing of any errors you do find, as well as suggestions for improvement of either the text or the presentation.

Peter M. Garfield
garfield@math.harvard.edu

CHAPTER

ONE

Data Exploration

The subjects of Statistics and Probability concern the mathematical tools that are designed to deal with uncertainty. To be more precise, these subjects are used in the following contexts:

- *To understand the limitations that arise from measurement inaccuracies.*
- *To find trends and patterns in noisy data.*
- *To test hypothesis and models with data.*
- *To estimate confidence levels for future predictions from data.*

What follows are some examples of scientific questions where the preceding issues are central and so statistics and probability play a starring role.

- An extremely large meteor crashed into the earth at the time of the disappearance of the dinosaurs. The most popular theory posits that the dinosaurs were killed by the ensuing environmental catastrophe. Does the fossil record confirm that the disappearance of the dinosaurs was suitably instantaneous?
- We read in the papers that fat in the diet is "bad" for you. Do dietary studies of large populations support this assertion?
- Do studies of gene frequencies support the assertion that all extent people are 100% African descent?
- The human genome project claims to have determined the DNA sequences along the human chromosomes. How accurate are the published sequences? How much variation should be expected between any two individuals?

Statistics and probability also play explicit roles in our understanding and modelling of diverse processes in the life sciences. These are typically processes where the outcome is influenced by many factors, each with small effect, but with significant total impact. Here are some examples:

Examples from Chemistry: What is thermal equilibrium? Does it mean stasis?

Why are chemical reaction rates influenced by temperature? How do proteins fold correctly? How stable are the folded configurations?

Examples from medicine: How many cases of flu should the health service expect to see this winter? How to determine cancer probabilities? Is hormone replacement therapy safe? Are anti-depressants safe?

An example from genomics: How are genes found in long stretches of DNA? How much DNA is dispensable?

An example from developmental biology: How does programmed cell death work; what cells die and what live?

Examples from genetics: What are the fundemental inheritance rules? How can genetics determine ancestral relationships?

An examples from ecology: How are species abundance estimates determined from small samples?

To summarize: There are at least two uses for statistics and probability in the life sciences. One is to tease information from noisy data, and the other is to develop predictive models in situations where chance plays a pivotal role. Note that these two uses of statistics are not unrelated since a theoretical understanding of the causes for the noise can facilitate its removal.

The rest of this first chapter focuses on the first of these two uses of statistics.

1.1 Snowfall data

To make matters concrete, the discussion that follows uses actual data on snowfall totals in Boston from 1890 through 2001. Table 1.1 gives snowfall totals (in inches) in Boston from the National Oceanic and Atmospheric Administration[1]. What we do with this data depends on what sort of questions we are going to ask. Noting the high snow falls

1890	42.6	1910	40.6	1930	40.8	1950	29.7	1970	57.3	1990	19.1
1891	46.8	1911	31.6	1931	24.2	1951	39.6	1971	47.5	1991	22.0
1892	66.0	1912	19.4	1932	40.6	1952	29.8	1972	10.3	1992	83.9
1893	64.0	1913	39.4	1933	62.7	1953	23.6	1973	36.9	1993	96.3
1894	46.9	1914	22.3	1934	45.4	1954	25.1	1974	27.6	1994	14.9
1895	38.7	1915	79.2	1935	30.0	1955	60.9	1975	46.6	1995	107.6
1896	43.2	1916	54.2	1936	9.0	1956	52.0	1976	58.5	1996	51.9
1897	51.9	1917	45.7	1937	50.6	1957	44.7	1977	85.1	1997	25.6
1898	70.9	1918	21.1	1938	40.3	1958	34.1	1978	27.5	1998	36.4
1899	25.0	1919	73.4	1939	37.7	1959	40.9	1979	12.7	1999	24.9
1900	17.5	1920	34.1	1940	47.8	1960	61.5	1980	22.3	2000	45.9
1901	44.1	1921	37.6	1941	24.0	1961	44.7	1981	61.8	2001	15.1
1902	42.0	1922	68.5	1942	45.7	1962	30.9	1982	32.7		
1903	72.9	1923	32.3	1943	27.7	1963	63.0	1983	43.0		
1904	44.9	1924	21.4	1944	59.2	1964	50.4	1984	26.6		
1905	37.6	1925	38.3	1945	50.8	1965	44.1	1985	18.1		
1906	67.9	1926	60.3	1946	19.4	1966	60.1	1986	42.5		
1907	26.2	1927	20.8	1947	89.2	1967	44.8	1987	52.6		
1908	20.1	1928	45.5	1948	37.1	1968	53.8	1988	15.5		
1909	37.0	1929	31.4	1949	32.0	1969	48.8	1989	39.2		

Table 1.1: NOAA Annual Snowfall Data (in inches)

in 1992, 1993 and 1995, I ask whether they indicate that winters in the more recent years are snowier than those in the first half of the record. Thus, I want to compare the snow falls in the years 1890-1945 with those in the years 1946–2001.

1.2 Data mining

One way to accomplish this is to just plot the snowfall amounts in the two cases and see if there is any evident difference in the two plots. These plots are shown in Figure 1.1. Mmmmm. These pictures don't help much. What I need are criteria for discerning when two data sets are distinctly different.

One approach is to introduce some numerical feature of a data set that can then be compared. There are two such invariants that you will often see used. The first is the *mean*, this the average of the data values. Thus, if a given data

[1] See the website `http://www.erh.noaa.gov/er/box/climate/BOS.SNW`.

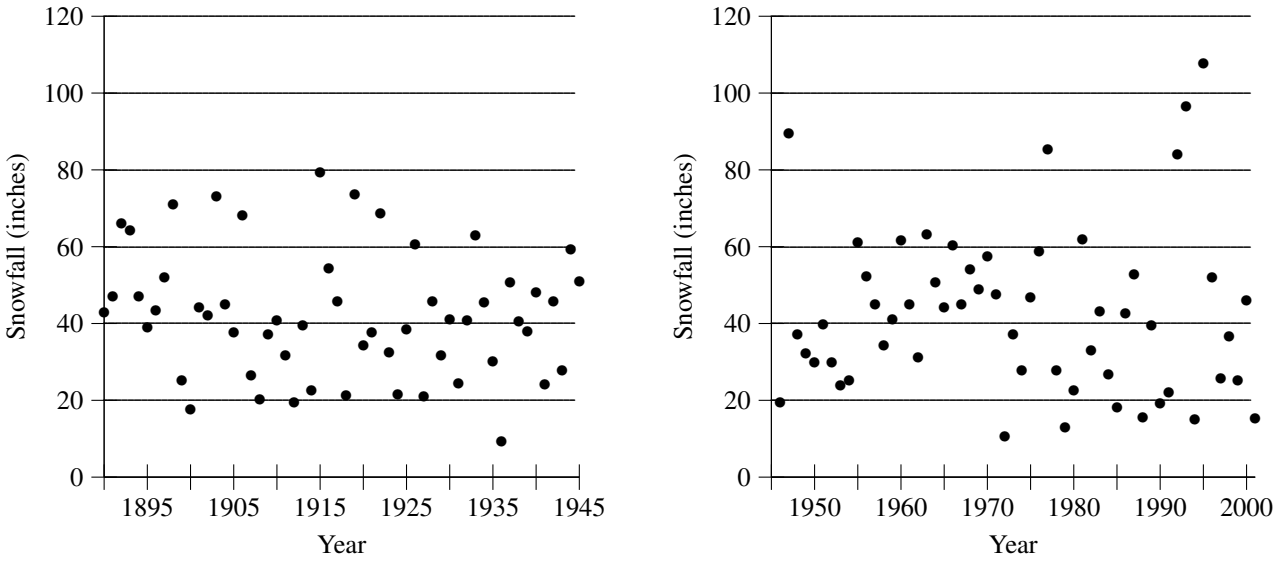

Figure 1.1: Snowfall Data for Years 1890–1945 (left) and 1946–2001 (right)

set consists of an ordered list of of some N numbers, $\{x_1, \ldots, x_N\}$, the mean is

$$\mu = \frac{1}{N}\left(x_1 + x_2 + \cdots + x_N\right) \tag{1.1}$$

In the cases at hand,

$$\mu_{\text{1890-1945}} = 42.4 \qquad \text{and} \qquad \mu_{\text{1946-2001}} = 42.3. \tag{1.2}$$

These are pretty close! But, of course, I don't know how close two means must be for me to say that there is no statistical difference between the data sets. Indeed, two data sets can well have the same mean and look very different. For example, consider that the three element sets $\{-1, 0, 1\}$ and $\{-10, 0, 10\}$. Both have mean zero, but the spread of the values in one is very much greater than the spread of the values in the other.

The preceding example illustrates the fact that means are not necessarily good criteria to distinguish data sets. Looking at this last toy example, I see that these two data sets are distinguished in some sense by the spread in the data; by how far the points differ from the mean value. The *standard deviation* is a convenient measure of this difference. It is defined to be

$$\sigma = \sqrt{\frac{1}{N-1}\big((x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_N-\mu)^2\big)}. \tag{1.3}$$

The standard deviations for the two snowfall data sets are

$$\sigma_{\text{1890-1945}} = 16.1 \qquad \text{and} \qquad \sigma_{\text{1946-2001}} = 21.4. \tag{1.4}$$

Well, these differ by roughly 5 inches, but is this difference large enough to be significant? How much difference should I tolerate so as to maintain that the snowfall amounts are "statistically" identical? How much difference in standard deviations signals a significant difference in yearly snowfall?

I can also "bin" the data. For example, I can count how many years have total snow fall less than 10 inches, then how many 10–20 inches, how many 20-30 inches, etc. I can do this with the two halves of the data set and then compare bin heights. Here is the result:

1890–1945:	1	2	11	11	16	5	6	4	0	0	0
1946–2001:	0	8	11	9	11	7	5	0	3	1	1

(1.5)

Having binned the data, I am yet at a loss to decide if the difference in bin heights really signifies a distinct difference in snow fall between the two halves of the data set.

What follows is one more try at a comparison of the two halves; it is called the *rank-sum* test and it works as follows: I give each year a number, between 1 and 112, by ranking the years in order of increasing total snow-fall. For example, the year with rank 1 is 1936 and the years with rank 109 and 110 are 1993 and 1995. I now sum all of the ranks for the years 1890-1945 to get the rank-sum for the first half of the data set. I then do the same for the years 1946-2001 to get the rank-sum for the latter half. I can now compare these two numbers. If one is significantly larger than the other, the data set half with the larger rank-sum has comparatively more high snowfall years than that with the smaller rank-sum. This understood, here are the two rank-sums:

$$\text{rank-sum}_{\text{1890-1945}} = 3137 \qquad \text{and} \qquad \text{rank-sum}_{\text{1946-2001}} = 3121. \tag{1.6}$$

Thus, the two rank-sums differ by 16. But, I am again faced with the following question: Is this difference significant? How big must the difference be to conclude that the first half of the 20'th century had, inspite of 1995, more snow on average, than the second half?

To elaborate now on this last question, consider that there is a hypothesis on the table:

> *The rank-sums for the two halves of the data set indicate that there is a significant difference between the snowfall totals from the first half of the data set as compared with those from the second.*

To use the numbers in (1.6) to analyze the validity of this hypothesis, I need an alternate hypothesis for comparison. The comparison hypothesis plays the role here of the control group in an experiment. This "control" is called the *null hypothesis* in statistics. In this case, the null-hypothesis asserts that the rankings are random. Thus, the null-hypothesis is:

> *The 112 ranks are distributed amongst the years as if they were handed out by a blindfolded monkey choosing numbers from a mixed bin.*

Said differently, the null-hypothesis asserts that the rank-sums in (1.6) are statistically indistinguishable from those that I would obtain I were to randomly select 56 numbers from the set $\{1, 2, 3, \ldots, 112\}$ to use for the rankings of the years in the first half of the data set, while using the remaining numbers for the second half.

An awful lot is hidden here in the phrase *statistically indistinguishable*. Here is what this phrase means in the case at hand: I should compute the probability that the sum of 56 randomly selected numbers from the set $\{1, 2, \ldots, 112\}$ differs from the sum of the 56 numbers that are left by *at least* 16. If this probability is very small, then I have some indication that the snow fall totals for the years in the two halves of the data set differ in a significant way. If the probability is high that the rank-sums for the randomly selected rankings differ by 16 or more, then the difference indicated in (1.6) should not be viewed as indicative of some statistical difference between the snowfall totals for the years in the two halves of the data set.

Thus, the basic questions are:

- *What is the probability in the case of the null-hypothesis that I should get a difference that is bigger than the one that I actually got?*
- *What probability should be considered "significant"?*

Of course, I can ask these same two questions for the bin data in (1.5). I can also ask analogs of these question for the two means in (1.2) and for the two standard deviations in (1.4). However, because the bin data, as well as the means and standard deviations deal with the snowfall amounts rather than with integer rankings, I would need a different sort of definition to use for the null-hypothesis in the latter cases.

In any event, take note that the first question is a mathematical one and the second is more of a value choice. The first question leads us to study the theory of probability which is the topic in the next chapter. As for the second question, I can tell you that it is the custom these days to take $\frac{1}{20} = 0.05$ as the cut-off between what is significant and what isn't. Thus,

If the probability of seeing a larger difference in values than the one seen is less than 0.05*, then the observed difference is deemed to be "significant".*

This choice of 5 percent is rather arbitrary, but such is the custom.

1.3 Exercises:

1. This exercise requires ten minutes of your time on two successive mornings. It also requires a clock that tells time to the nearest second.

   (a) On the first morning, before eating or drinking, record the following data: Try to estimate the passage of precisely 60 seconds of time with your eyes closed. Thus, obtain the time from the clock, immediately close your eyes and when you feel that 1 minute has expired, open them and immediately read the amount of time that has passed on the clock. Record this as your first estimate for 1 minute of time. Repeat this procedure ten time to obtain ten successive estimates for 1 minute.

   (b) On the second morning, repeat this part (a), but first drink a caffeinated beverage such as coffee, tea, or a cola drink.

   (c) With parts (a) and (b) completed, you have two lists of ten numbers. Compute the means and standard deviations for each of these data sets. Then, combine the data sets as two halves of a single list of 20 numbers and compute the rank-sums for the two lists. Thus, your rankings will run from 1 to 20. In the event of a tie between two estimates, give both the same ranking and don't use the subsequent ranking. For example, if there is a tie for fifth, use 5 for both but give the next highest estimate 7 instead of 6.

2. Flip a coin 200 times. Use n_1 to denote the number of heads that appeared in flips 1-10, use n_2 to denote the number that appeared in flips 11-20, and so on. In this way, you generate twenty numbers, $\{n_1, \ldots n_{20}\}$. Compute the mean and standard deviation for the sets $\{n_1, \ldots n_{10}\}$, $\{n_{11}, \ldots n_{20}\}$, and $\{n_1, \ldots n_{20}\}$.

3. The table that follows gives the results of US congressional elections during the 6th year of a President's term in office. (Note: he had to be reelected.) A negative number means that the President's party lost seats. Note that there aren't any positive numbers. Compute the mean and standard deviation for both the Senate and House of Representatives. Compare these numbers with the line for the 2006 election.

Year	President	Senate	House
2006	Bush	-6	-30
1998	Clinton	0	-5
1986	Reagan	-8	-5
1974	Ford	-5	-48
1966	Johnson	-4	-47
1958	Eisenhower	-13	-48
1950	Truman	-6	-29
1938	Roosevelt	-6	-71
1926	Coolidge	-6	-10
1918	Wilson	-6	-19

Table 1.2: Number of seats gained by the president's party in the election during his sixth year in office

CHAPTER

TWO

Basic notions from probability theory

Probability theory is the mathematics of chance and luck. To elaborate, its goal is to make sense of the following question:

> *What is the probability of a given outcome from some set of possible outcomes?*

For example, in the snow fall analysis of the previous chapter, I computed the rank-sums for the two halves of the data set and found that they differed by 16. I then wondered what the probability was for such rank sums to differ by more than 16 if the rankings were randomly selected instead of given by the data. We shall eventually learn what it means to be "randomly selected" and how to compute such probabilities. However, this comes somewhat farther into the course.

2.1 Talking the talk

Unfortunately for the rest of us, probability theory has its own somewhat arcane language. What follows is a list of the most significant terms. Treat this aspect of the course as you would any other language course. In any event, there are not so many terms, and you will soon find that you don't have to look back at your notes to remember what they mean.

Sample space: A *sample space* is the set of all possible outcomes of the particular "experiment" of interest. For example, in the rank-sum analysis of the snowfall data from the previous chapter, I should consider the sample space to be the set of all collections of 56 distinct integers from the collection $\{1, \ldots, 112\}$.

For a second example, imagine flipping a coin three times and recording the possible outcomes of the three flips. In this case, the sample space is

$$S = \{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH\}. \tag{2.1}$$

Here is a third example: If you are considering the possible birthdates of a person drawn at random, the sample space consists of the days of the year, thus the integers from 1 to 366. If you are considering the possible birthdates of two people selected at random, the sample space consists of all pairs of the form (j, k) where j and k are integers from 1 to 366. If you are considering the possible birthdates of three people selected at random, the sample space consists of all triples of the form (j, k, m) where j, k and m are integers from 1 to 366.

My fourth example comes from medicine: Suppose that you are a pediatrician and you take the pulse rate of a 1-year old child? What is the sample space? I imagine that the number of beats per minute can be any number between 0 and and some maximum, say 300.

To reiterate: The sample space is no more nor less than the collection of all possible outcomes for your experiment.

Events: An *event* is a subset of the sample space, thus a subset of possible outcomes for your experiment. In the rank-sum example, where the sample space is the set of all collections of 56 distinct integers from 1 through 112, here is one event: The subset of collections of 56 integers whose sum is 16 or more greater than the sum of those that remain. Here is another event: The subset that consists of the 56 consecutive integers that start at 1. Notice that the first event contains lots of collections of 56 integers, but the second event contains just $\{1, 2, \ldots, 56\}$. So, the first event has more elements than the second.

Consider the case where the sample space is the set of outcomes of two flips of a coin, thus $S = \{HH, HT, TH, TT\}$. For a small sample space such as this, one can readily list all of the possible events. In this case, there are 16 possible events. Here is the list: First comes the no element set, this denoted by tradition as $\emptyset$. Then comes the 4 sets with just one element, these consist of $\{HH\}$, $\{HT\}$, $\{TH\}$, $\{TT\}$. Next come the 6 two element sets, $\{HH, HT\}$, $\{HH, TH\}$, $\{HH, TT\}$, $\{HT, TH\}$, $\{HT, TT\}$, $\{TH, TT\}$. Note that the order of the elements is of no consequence; the set $\{HH, HT\}$ is the same as the set $\{HT, HH\}$. The point here is that we only care about the elements, not how they are listed. To continue, there are 4 distinct sets with three elements, $\{HH, HT, TH\}$, $\{HH, HT, TT\}$, $\{HH, TH, TT\}$ and $\{HT, TH, TT\}$. Finally, there is the set with all of the elements, $\{HH, HT, TH, TT\}$.

Note that a subset of the sample space can have no elements, or one element, or two, ..., up to and including all of the elements in the sample space. For example, if the sample space is that given in (1.1) for flipping a coin three times, then HTH is an event. Meanwhile, the event that a head appears on the first flip is $\{HTT, HHT, HTH, HHH\}$, a set with four elements. The event that four heads appears has zero elements, and the set where there are less than four heads is the whole sample space. No matter what the original sample space, the event with no elements is called the *empty set*, and is denoted by $\emptyset$.

In the case where the sample space consists of the possible pulse rate measurements of a 1-year old, some events are: The event that the pulse rate is greater than 100. The event that the pulse rate is between 80 and 85. The event that the pulse rate is either between 100 and 110 or between 115 and 120. The event that the pulse rate is either 85 or 95 or 105. The event that the pulse rate is divisible by 3. And so on.

By the way, this last example illustrates the fact that there are many ways to specify the elements in the same event. Consider, for example, the event that the pulse rate is divisible by 3. Let's call this event E. Another way to E is to provide a list of all of its elements, thus $E = \{0, 3, 6, \ldots, 300\}$. Or, I can use a more algebraic tone: E is the set of integers x such that $0 \leq x \leq 300$ and $x/3 \in \{0, 1, 2, \ldots, 100\}$. (See below for the definition of the symbol "$\in$".) For that matter, I can describe E accurately using French, Japanese, Urdu, or most other languages.

To repeat: Any given subset of a given sample space is called an *event*.

Set Notation: Having introduced the notion of a subset of some set of outcomes, you need to become familiar with some standard notation that is used in the literature when discussing subsets of sets.

(a) As mentioned above, $\emptyset$ is used to denote the "set" with no elements.

(b) If A and B are subsets, then $A \cup B$ denotes the subset whose elements are those that appear either in A or in B or in both. This subset is called the *union* of A and B.

(c) Meanwhile, $A \cap B$ denotes the subset whose elements appear in both A and B. It is called the *intersection* between A and B.

(d) If no elements are shared by A and B, then these two sets are said to be *disjoint*. Thus, A and B are disjoint if and only if $A \cap B = \emptyset$.

(e) If A is given as a subset of a set S, then A^c denotes the subset of S whose elements are not in A. Thus, A^c and A are necessarily disjoint and $A^c \cup A = S$. The set A^c is called the *complement* of A.

(f) If a subset A is entirely contained in another subset, B, one writes $A \subset B$. For example, if A is an event in a sample space S, then $A \subset S$.

(g) If an element, e, is contained in a set A, one writes $e \in A$. If e is not in A, one writes $e \notin A$.

What follows are some examples that are meant to illustrate what is going on. suppose that the sample space is the set of possible pulse rates of a 1-year old child. Lets us take this set to be $\{0, 1, \ldots, 300\}$. Consider the case where A is the set of elements that are at least 100, and B is the set of elements that are greater than 90 but less than 110. Thus, $A = \{100, 101, \ldots, 300\}$, and $B = \{91, \ldots, 109\}$. The union of A and B is the set $\{91, 92, \ldots, 300\}$. The intersection of A and B is the set $\{100, 101, \ldots, 109\}$. The complement of A is the set $\{0, 1, \ldots, 99\}$. The complement of B is the set $\{0, 1, \ldots, 90, 110, 111, \ldots, 300\}$. (Note that any given set is disjoint from its complement.) Meanwhile, $110 \in A$ but $110 \notin B$.

2.2 Axiomatic definition of probability

A probability function for a given sample space assigns the probabilities to various subsets. For example, if I am flipping a coin once, I would take my sample space to be the set $\{H, T\}$. If the coin were fair, I would use the probability function that assigns 0 to the empty set, $\frac{1}{2}$ to each of the subsets $\{H\}$ and $\{T\}$, and then 1 to the whole of S. If the coin were biased a bit towards landing heads up, I might give $\{H\}$ more than $\frac{1}{2}$ and $\{T\}$ less than $\frac{1}{2}$.

The choice of a probability function is meant to quantify what is meant by the term "at random". For example, consider the case for choosing just one number "at random" from the set $\{1, \ldots, 112\}$. If "at random" is to mean that that there is no bias towards any particular number, then my probability function should assign to each subset that consists of just a single integer. Thus, it gives to the subsets $\{1\}$, $\{2\}$, ..., etc. If I mean something else by my use of the term "at random", then I would want to use a different assignment of probabilities.

To explore another example, consider the case where the sample space represents the set of possible pulse rate measurements for a 1-year old child. Thus, S is the set whose elements are $\{0, 1, \ldots, 300\}$. As a pediatrician, you would be interested in the probability for measuring a given pulse rate. I expect that this probability is not the same for all of the elements. For example, the number 20 is certainly less probable than the number 90. Likewise, 190 is certainly less probable than 100. I expect that the probabilities are greatest for numbers between 80 and 120, and then decrease rather drastically away from this interval.

Here is the story in the generic, abstract setting: Imagine that we have a particular sample space, S, in mind. A probability function, P, is an assignment of a number no less than 0 and no greater than 1 to various subsets of S subject to two rules:

$$\begin{aligned} &\bullet\ \mathrm{P}(S) = 1. \\ &\bullet\ \mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) \qquad \text{when } A \cap B = \emptyset. \end{aligned} \tag{2.2}$$

Note that condition $\mathrm{P}(S) = 1$ says that there is probability 1 of at least something happening. Meanwhile, the condition $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$ when A and B have no points in common asserts the following: The probability of something happening that is in either A or B is the sum of the probabilities of something happening from A or something happening from B.

To give an example, consider rolling a standard, six-sided die. If the die is rolled once, the sample space consists of the numbers $\{1, 2, 3, 4, 5, 6\}$. If the die is *fair*, then I would want to use the probability function that assigns the value $\frac{1}{6}$ to each element. But, what if the die is not fair? What if it favors some numbers over others? Consider, for example, a probability function with $\mathrm{P}(\{1\}) = 0$, $\mathrm{P}(\{2\}) = \frac{1}{3}$, $\mathrm{P}(\{3\}) = \frac{1}{2}$, $\mathrm{P}(\{4\}) = \frac{1}{6}$, $\mathrm{P}(\{5\}) = 0$ and $\mathrm{P}(\{6\}) = 0$. If this probability function is correct for my die, what is the most probable number to appear with one roll? Should I expect to see the number 5 show up at all? What follows is a more drastic example: Consider the probability function where $\mathrm{P}(\{1\}) = 1$ and $\mathrm{P}(\{2\}) = \mathrm{P}(\{3\}) = \mathrm{P}(\{4\}) = \mathrm{P}(\{5\}) = \mathrm{P}(\{6\}) = 0$. If this probability function is correct, I should expect only the number 1 to show up.

Let us explore a bit the reasoning behind the conditions for P that appear in equation (2.2). To start, you should understand why the probability of an event is not allowed to be negative, nor is it allowed to be greater than 1. This is to conform with our intuitive notion of what probability means. To elaborate, think of the sample space as the suite of possible outcomes of an experiment. This can be any experiment, for example flipping a coin three times, or rolling a die once, or measuring the pulse rate of a 1-year old child. An event is a subset of possible outcomes. Let us suppose

that we are interested in a certain event, this a subset denoted by A. The probability function assigns to A a number, $\mathrm{P}(A)$. This number has the following interpretation:

> *If the experiment is carried out a large number of times, with the conditions and set up the same each time, then $\mathrm{P}(A)$ is a prediction for the fraction of those experiments where the outcome is in the set A.*

As this fraction can be at worse 0 (no outcomes in the set A), or at best 1 (all outcomes in the set A), so $\mathrm{P}(A)$ should be a number that is not less than 0 nor more than 1.

Why should $\mathrm{P}(S)$ be equal to 1 in all cases? Well, by virtue of its very definition, the set S is supposed to be the set of *all* possible outcomes of the experiment. The requirement for $\mathrm{P}(S)$ to equal 1 makes the probability function predict that each outcome must come from our list of all possible outcomes.

The second condition that appears in equation (2.2) is less of a tautology. It is meant to model a certain intuition that we all have about probabilities. Here is the intuition: The probability of a given event is the sum of the probabilities of its constituent elements. For example, consider the case where the sample set is the set of possible outcomes when I roll a fair die. Thus, the probability is $\frac{1}{6}$ for any given integer from $\{1, 2, 3, 4, 5, 6\}$ appearing. Let A denote the probability of $\{1\}$ appearing and B the probability of $\{2\}$ appearing. I expect that the probability of either 1 or 2 appearing, thus $A \cup B = \{1, 2\}$, is $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$. I would want my probability function to reflect this addititivity. The second condition in equation (2.2) asserts no more nor less than this requirement.

By the way, the condition for $A \cap B = \emptyset$ is meant to prevent over-counting. For an extreme example, suppose $A = \{1\}$ and B is also $\{1\}$. Thus both have probability $\frac{1}{6}$. Meanwhile, $A \cup B = \{1\}$ also, so $\mathrm{P}(A \cup B)$ should be $\frac{1}{6}$, not $\frac{1}{6} + \frac{1}{6}$. Here is a somewhat less extreme example: Suppose that $A = \{1, 2\}$ and $B = \{2, 3\}$. Both of these sets should have probability . Their union is $\{1, 2, 3\}$. I expect that this set has probability $\frac{1}{2}$, not $\frac{1}{3} + \frac{1}{3} = \frac{2}{3}$. The reason I shouldn't use the formula $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$ for the case where $A = \{1, 2\}$ and $B = \{2, 3\}$ is because the latter formula counts twice the probability of the shared integer 2; it counts it once from its appearance in A and again from its appearance in B.

For a second illustration, consider the case where S is the set of possible pulse rates for a 1-year old child. Take A to be the event $\{100, \ldots, 109\}$ and B to be the event $\{120, \ldots, 129\}$. Suppose that many years of pediatric medicine have given us a probability function, P, for this set. Suppose, in addition that $\mathrm{P}(A) = \frac{1}{4}$ and $\mathrm{P}(B) = \frac{1}{16}$. What should we expect for the probability that a measured pulse rate is either in A *or* in B? That is, what is the probability that the pulse rate is in $A \cup B$? Since A and B do not share elements ($A \cap B = \emptyset$), you might expect that the probability of being in either set is the sum of the probability of being in A with that of being in B, thus $\frac{5}{16}$.

Keeping this last example in mind, consider the set $\{109, 110, 111\}$. I'll call this set C. Suppose that our probability function says that $\mathrm{P}(C) = \frac{1}{64}$. I would not predict that $\mathrm{P}(A \cup C) = \mathrm{P}(A) + \mathrm{P}(C)$ since A and C both contain the element 109. Thus, I can imagine that $\mathrm{P}(A) + \mathrm{P}(C)$ over-counts the probability for $A \cup C$ since it counts the probability of 109 two times, once from its membership in A and again from its membership in C.

I started the discussion prior to equation (2.2) by asking that you imagine a particular sample space and then said that a probability function on this space is a rule that assigns to each event a number no less than zero and no greater than 1 to each subspace (event) of the sample space, but subject to the rules that are depicted in (2.2). I expect that many of you are silently asking the following question:

> *Who or what determines the probability function P?*

To make this less abstract, consider again the case of rolling a six-sided die. The corresponding sample space is $S = \{1, 2, 3, 4, 5, 6\}$. I noted above three different probability functions for S. The first assigned equal probability to each element. The second and third assigned different probabilities to different elements. The fact is that there are infinitely many probability functions to choose from. Which should be used?

To put the matter in even starker terms, consider the case where the sample space consists of the possible outcomes of a single coin flip. Thus, $S = \{H, T\}$. A probability function on S is no more nor less than an assignment of one number, $\mathrm{P}(H)$, that is not less than 0 nor greater than 1. Only one number is needed because the first line of (2.2) makes P assign 1 to S, and the second line of (2.2) makes P assign $1 - \mathrm{P}(H)$ to T. Thus, $\mathrm{P}(T) = 1 - \mathrm{P}(H)$. If you understand this last point, then it follows that there are as many probability functions for the set $S = \{H, T\}$ as there

are real numbers in the interval between 0 and 1 (including the end-points). By any count, there are infinitely many such numbers!

So, I have infinitely many choices for P. Which should I choose? So as to keep the abstraction to a minimum, let's address this question to the coin flipping case where $S = \{H, T\}$. It is important to keep in mind what the purpose of a probability function is: The probability function should accurately predict the relative frequencies of heads and tails that appear when a given coin is flipped a large number of times.

Granted this goal, I might proceed by first flipping the particular coin some large number of times to generate an "experimentally" determined probability. This I'll call P_E. I then use P_E to predict probabilities for all future flips of this coin. For example, if I flip the coin 100 times and find that 44 heads appear, then I might set $\mathrm{P}_E(H) = 0.44$ to predict the probabilities for all future flips. By the way, we instinctively use experimentally determined probability functions constantly in daily life. However, we use a different, but not unrelated name for this: We call it *experience*.

There is a more theoretical way to proceed. I might study how coins are made and based on my understanding of their construction, deduce a "theoretically" determined probability. I'll call this P_T. For example, I might deduce that $\mathrm{P}_T(H) = 0.5$. I might then use P_T to compute all future probabilities.

As I flip this coin in the days ahead, I may find that one or the other of these probability functions is more accurate. Or, I may suspect that neither is very accurate. How I judge accuracy will lead us to the subject of Statistics.

2.3 Computing probabilities for subsets

If your sample space is a finite set, and if you have assigned probabilities to all of the one element subsets from your sample space, then you can compute the probabilities for all events from the sample space by invoking the rules in (2.2). Thus,

> *If you know what P assigns to each element in S, then you know P on every subset: Just add up the probabilities that are assigned to its elements.*

We'll talk about the story when S isn't finite later. Anyway, the preceding illustrates the more intuitive notion of probability that we all have: It says simply that if you know the probability of every outcome, then you can compute the probability of any subset of outcomes by summing up the probabilities of the outcomes that are in the given subset.

For example, in the case where my sample space $S = \{1, \dots, 112\}$ and each integer in S has probability $\frac{1}{112}$, then I can compute that the probability of a blindfolded monkey picking either 1 or 2 is $\frac{1}{112} + \frac{1}{112} = \frac{1}{112}$. Here I invoke the second of the rules in (2.2) where A is the event that 1 is chosen and B is the event that 2 is chosen. A sequential use of this same line of reasoning finds that the probability of picking an integer that is less than or equal to 10 is $\frac{10}{112}$.

Here is a second example: Take S to be the set of outcomes for flipping a fair coin three times (as depicted in (2.1)). If the coin is fair and if the three flips are each fair, then it seems reasonable to me that the situation is modeled using the probability function, P, that assigns to each element in the set S. If we take this version of P, then we can use the rule in (2.2) to assign probabilities $\frac{1}{8}$ to any given subset of S. For example, the subset given by $\{HHT, HTH, THH\}$ has probability $\frac{3}{8}$ since

$$\mathrm{P}(\{HHT, HTH, THH\}) = \mathrm{P}(\{HHT, HTH\}) + \mathrm{P}(THH)$$

by invoking (2.2). Invoking it a second time finds

$$\mathrm{P}(\{HHT, HTH\}) = \mathrm{P}(HHT) + \mathrm{P}(HTH),$$

and so

$$\mathrm{P}(\{HHT, HTH, THH\}) = \mathrm{P}(HHT) + \mathrm{P}(HTH) + \mathrm{P}(THH) = \tfrac{3}{8}.$$

To summarize: If the sample space is a set with finite elements, or is a discrete set (such as the positive integers), then you can find the probability of any subset of the sample space if you know the probability for each element.

2.4 Some consequences of the definition

Here are some consequences of the definition of probability.

$$\begin{aligned}
&\text{(a) } \mathrm{P}(\emptyset) = 0.\\
&\text{(b) } \mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B).\\
&\text{(c) } \mathrm{P}(A) \leq \mathrm{P}(B) \text{ if } A \text{ is contained entirely in } B.\\
&\text{(d) } \mathrm{P}(B) = \mathrm{P}(B \cap A) + \mathrm{P}(B \cap A^c).\\
&\text{(e) } \mathrm{P}(A^c) = 1 - \mathrm{P}(A).
\end{aligned} \tag{2.3}$$

In the preceding, A^c is the set of elements that are *not* in A. The set A^c is called the *complement* of A.

I want to stress that all of these conditions are simply translations into symbols of intuition that we all have about probabilities. What follows are the respective English versions of (2.3).

Equation (2.3a):

> *The probability that no outcomes appear is zero.*

This is to say that if S is, as required, the list of *all* possible outcomes, then at least one outcome must occur.

Equation (2.3b):

> *The probability that an outcome is in either A or B is the probability that it is in A plus the probability that it is in B minus the probability that it is in both.*

The point here is that if A and B have elements in common, then one is overcounting to obtain $\mathrm{P}(A \cup B)$ by just summing the two probabilities. The sum of $\mathrm{P}(A)$ and $\mathrm{P}(B)$ counts twice the elements that are both in A and in B count twice. To see how this works, consider the rolling a standard, six-sided die where the probabilities of any given side appearing are all the same, thus . Now consider the case where A is the event that either 1 or 2 appears, while B is the event that either 2 or 3 appears. The probability assigned to A is $\frac{1}{3}$, that assigned to B is also $\frac{1}{3}$. Meanwhile, $A \cup B = \{1, 2, 3\}$ has probability $\frac{1}{2}$ and $A \cap B = \{2\}$ has probability $\frac{1}{6}$. Since $\frac{1}{2} = \frac{1}{3} + \frac{1}{3} - \frac{1}{6}$, the claim in (2.3b) holds in this case. You might also consider (2.3b) in a case where $A = B$.

Equation (2.3c):

> *The probability of an outcome from Ais no greater than that of an outcome from B in the case that all outcomes from A are contained in the set B.*

The point of (2.3c) is simply that if every outcome from A appears in the set B, then the probability that B occurs can not be less than that of A. Consider for example the case of rolling one die that was just considered. Take A again to be $\{1, 2\}$, but now take B to be the set $\{1, 2, 3\}$. Then $\mathrm{P}(A)$ is less than $\mathrm{P}(B)$ because B contains all of A's elements plus another. Thus, the probability of B occurring is the sum of the probability of A occurring and the probability of the extra element occurring.

Equation (2.3d):

> *The probability of an outcome from the set B is the sum of the probability that the outcome is in the portion of B that is contained in A and the probability that the outcome is in the portion of B that is not contained in A.*

This translation of (2.3d) says that if I break B into two parts, the part that is contained in A and the part that isn't, then the probability that some element from B appears is obtained by adding, first the probability that an element that is both in A and B appears, and then the probability that an element appears that is in B but not in A. Here is an example from rolling one die: Take $A = \{1, 2, 4, 5\}$ and $B = \{1, 2, 3, 6\}$. Since B has four elements and each has probability $\frac{1}{6}$, so B has probability $\frac{2}{3}$. Now, the elements that are both in B and in A comprise the set $\{1, 2\}$, and this set has probability $\frac{1}{3}$. Meanwhile, the elements in B that are not in A comprise the set $\{3, 6\}$. This set also has probability $\frac{1}{3}$. Thus (2.3d) holds in this case because $\frac{1}{3} + \frac{1}{3} = \frac{2}{3}$.

Equation (2.3e):

> *The probability of an outcome that is not in A is equal to 1 minus the probability that an outcome is in A.*

To see why this is true, break the sample space up into two parts, the elements in A and the elements that are not in A. The sum of the corresponding two probabilities must equal 1 since any given element is either in A or not. Consider our die example where $A = \{1, 2\}$. Then $A^c = \{3, 4, 5, 6\}$ and their probabilities do indeed add up to 1.

2.5 That's all there is to probability

You have just seen most of probability theory for sample spaces with a finite number of elements. There are a few new notions that are introduced later, but a good deal of what follows concerns either various consequences of the notions that were just introduced, or else various convenient ways to calculate probabilities that arise in common situations.

Before moving on, it is important to explicitly state something that has been behind the scenes in all of this: When you come to apply probability theory, the sample space and its probability function are chosen by you, the scientist, based on your understanding of the phenomena under consideration. Although there are often standard and obvious choices available, neither the sample space nor the probability function need be god given. The particular choice constitutes a *theoretical assumption* that you are making in your mental model of what ever phenomena is under investigation.

To return to an example I mentioned previously, if I flip a coin once and am concerned about how it lands, I might take for S the two element set $\{H, T\}$. If I think that the coin is fair, I would take my probability function P so that $\mathrm{P}(H) = \frac{1}{2}$ and $\mathrm{P}(T) = \frac{1}{2}$. However, if I have reason to believe that the coin is not fair, then I should choose P differently. Moreover, if I have reason to believe that the coin can sometimes land on its edge, then I would have to take a different sample space: $\{H, T, E\}$.

Here is an example that puts this choice question into a rather stark light: Given that the human heart can beat anywhere from 0 to 300 beats per minute, the sample space for the possible measurements of pulse rate is the set $S = \{0, 1, \ldots, 300\}$. Do you think that the probability function that assigns equal values to these integers will give reasonable predictions for the distribution of the measured pulse rates of you and your classmates?

2.6 Exercises:

1. Suppose an experiment has three possible outcomes, labeled 1, 2, and 3. Suppose in addition, that you do the experiment three successive times.

   (a) Give the sample space for the possible outcomes of the three experiments.

(b) Write down the subset of your sample space that correspond to the event that outcome 1 occurs in the second experiment.

(c) Write down the subset of your sample space that corresponds to the event that outcome 1 appears in at least one experiment.

(d) Write down the subset of your sample space that corresponds to the event that outcome 1 appears at least twice.

(e) Under the assumption that each element in your sample space has equal probability, give the probabilities for the events that are described in parts (b), (c) and (d) above.

2. Measure your pulse rate. Write down the symbol $+$ if the rate is greater than 70 beats per minute, but write down $-$ if the rate is less than or equal to 70 beats per minute. Repeat this four times and so generate an ordered set of 4 elements, each a plus or a minus symbol.

   (a) Write down the sample space for the set of possible 4 element sets that can arise in this manner.

   (b) Under the assumption that all elements of this set are equally likely, write down the probability for the event that precisely three of the symbols that appear in a given element are identical.

3. Let S denote the set $\{1, 2, \ldots, 10\}$.

   (a) Write down three different probability functions on S by giving the probabilities that they assign to the elements of S.

   (b) Write down a function on S whose values can not be those of a probability function, and explain why such is the case.

4. Four apples are set in a row. Each apple either has a worm or not.

   (a) Write down the sample space for the various possibilities for the apples to have or not have worms.

   (b) Let A denote the event that the apples are worm free and let B denote the event that there is at least two worms amongst the four. What is $A \cup B$ and $A^c \cap B$?

5. A number is chosen at random from 1 to 1000. Let A denote the event that the number is divisible by 3 and B the event that it is divisible by 5. What is $A \cap B$?

6. Some have conjectured that changing to a vegetarian diet can help lower cholesterol levels, and in turn lead to lower levels of heart disease. Twenty-four mostly hypertensive patients were put on vegetarian diets to see if such a diet has an effect on cholesterol levels. Blood serum cholesterol levels were measured just before they started their diets, and 3 months into the diet[1].

   (a) Before doing any calculations, do you think Table 2.1 shows any evidence of an effect of a vegetarian diet on cholesterol levels? Why or why not?

   The sign test is a simple test of whether or not there is a real difference between two sets of numbers. In this case, the first set consists of the 24 pre-diet measurements, and the second set consists of the 24 after diet measurements. Here is how this test works in the case at hand: Associate $+$ to a given measurement if the cholesterol level increased, and associate $-$ if the cholesterol decreases. The result is a set of 24 symbols, each either $+$ or $-$. For example, in this case, there are the number of $+$ is 3 and the number of $-$ is 21. One then ask whether such an outcome is likely given that the diet has no effect. If the outcome is unlikely, then there is reason to suspect that the diet makes a difference. Of course, this sort of thinking is predicated on our agreeing on the meaning of the term "likely", and on our belief that there are no as yet unknown reasons why the outcome appeared as it did. To elaborate on the second point, one can imagine that the cholesterol change is due not so much to the vegetarian nature of the diet, but to some factor in the diet that changed simultaneously with the change to a vegetarian diet. Indeed, vegetarian diets can be quite bland, and so it may be the case that people use more salt or pepper when eating vegetarian food. Could the cause be due to the change in condiment level? Or perhaps people are hungrier sooner after such a diet, so they treat themselves to an ice cream cone a few hours after dinner. Perhaps the change in cholesterol is due not to the diet, but to the daily ice cream intake.

[1] Rosner, Bernard. *Fundamentals of Biostatistics. 4th Ed.* Duxbury Press, 1995.

Subject	Before Diet	After Diet	Difference
1	195	146	−49
2	145	155	10
3	205	178	−27
4	159	146	−13
5	244	208	−36
6	166	147	−19
7	250	202	−48
8	236	215	−21
9	192	184	−8
10	224	208	−16
11	238	206	−32
12	197	169	−28
13	169	182	13
14	158	127	−31
15	151	149	−2
16	197	178	−19
17	180	161	−19
18	222	187	−35
19	168	176	8
20	168	145	−23
21	167	154	−13
22	161	153	−8
23	178	137	−41
24	137	125	−12

Table 2.1: Cholesterol levels before and three months after starting a vegetarian diet

(b) To make some sense of the notion of "likely", we need to consider a probability function on the set of possible lists where each list has 24 symbols with each symbol either $+$ or $-$. What is the sample space for this set?

(c) Assuming that each subject had a 0.50 probability of an increase in cholesterol, what probability does the resulting probability function assign to any given element in your sample space?

(d) Given the probability function you found in part (c), what is the probability of having no $+$ appear in the 24?

(e) With this same probability function, what is the probability of only one $+$ appear?

An upcoming chapter explains how to compute the probability of any number of $+$ appearing. Another chapter introduces a commonly agreed upon definition for "likely".

CHAPTER

THREE

Conditional probability

The notion of *conditional probability* provides a very practical tool for computing probabilities of events. Here is context where this notion first appears: You have a sample space, S, with a probability function, P. Suppose that A and B are subsets of S and that you have knowledge that the event represented by B has already occurred. Your interest is in the probability of the event A given this knowledge about the event B. This conditional probability is denoted by $\mathrm{P}(A \mid B)$; and it is often different from $\mathrm{P}(A)$.

Here is an example: Write down $+$ if you measure your pulse rate to be greater than 70 beats per minute; but write down $-$ if you measure it to be less than or equal to 70 beats per minute. Make three measurements of your pulse rate and so write down three symbols. The set of possible outcomes for the three measurements consists of the eight element set

$$S = \{+++, ++-, +-+, +--, -++, -+-, --+, ---\}. \tag{3.1}$$

Let A denote the event that all three symbols are $+$, and let B denote the event that the first symbol is $+$. Then $\mathrm{P}(A \mid B)$ is the probability that all symbols are $+$ *given* that the first one is also $+$. If each of the eight elements has the same probability, $\frac{1}{8}$, then it should be the case that $\mathrm{P}(A \mid B) = \frac{1}{4}$ since there are four elements in B but only one of these, $(+++)$, is also in A. This is, in fact, the case given the formal definition that follows. Note that in this example, $\mathrm{P}(A \mid B) \neq \mathrm{P}(A)$ since $\mathrm{P}(A) = \frac{1}{8}$.

Here is another hypothetical example: Suppose that you are a pediatrician and you get a phone call from a distraught parent about a child that is having trouble breathing. One question that you ask yourself is: What is the probability that the child is having an allergic reaction? Let's denote by A the event that this is, indeed, the correct diagnosis. Of course, it may be that the child has the flu, or a cold, or any number of diseases that make breathing difficult. Anyway, in the course of the conversation, the parent remarks that the child has also developed a rash on its torso. Let us use B to denote the probability that the child has a rash. I expect that the probability the child is suffering from an allergic reaction is much greater given that there is a rash. This is to say that $\mathrm{P}(A \mid B) > \mathrm{P}(A)$ in this case. Or, consider an alternative scenario, one where the parent does not remark on a rash, but remarks on a fever instead. In this case, I would expect that the probability of the child suffering an allergic reaction is rather small since the symptoms point more towards a cold or flu. This is to say that I now expect $\mathrm{P}(A \mid B)$ to be less than $\mathrm{P}(A)$.

3.1 The definition of conditional probability

As noted above, this is the probability that an event in A occurs given that you already know that an event in B occurs. The rule for computing this new probability is

$$\mathrm{P}(A \mid B) \equiv \mathrm{P}(A \cap B)/\mathrm{P}(B). \tag{3.2}$$

You can check that this obeys all of the rules for being a probability. In English, this says:

> *The probability of an outcome occurring from A given that the outcome is known to be in B is the probability of the outcome being in both A and B divided by the probability of the outcome being in B in the first place.*

Another way to view this notion is as follows: Since we are told that B has happened, one might expect that the probability that A occurs is the fraction of B's probability that is accounted for by the elements that are in both A and B. This is just what (3.2) asserts. Indeed, $\mathrm{P}(A \cap B)$ is the probability of the occurrence of an element that is in both A and B, so the ratio $\mathrm{P}(A \cap B)/\mathrm{P}(B)$ is the fraction of B's probability that comes from the elements that are both in A and B.

For a simple example, consider the case where we roll a die with each face having the same probability of appearing. Take B to be the event that an even number appears. Thus, $B = \{2, 4, 6\}$. I now ask: What is the probability that 2 appears given that an even number has appeared? Without the extra information, the probability that 2 appears is $\frac{1}{6}$. If I am told in advance that an even number has appeared, then I would say that the probability that 2 appears is $\frac{1}{3}$. Note that $\frac{1}{3} = \frac{1}{6}/\frac{1}{2}$; and this is just what is said in (3.2) in the case that $A = \{2\}$ and $B = \{2, 4, 6\}$.

To continue with this example, I can also ask for the probability that 1 or 3 appears given that an even number has appeared. Set $A = \{1, 3\}$ in this case. Without the extra information, we have $\mathrm{P}(A) = \frac{1}{3}$. However, as neither 1 nor 3 is an even number, $A \cap B = \emptyset$. This is to say that A and B do not share elements. Granted this obvious fact, I would say that $\mathrm{P}(A \,|\, B) = 0$. This result is consistent with (3.2) because the numerator that appears on the right hand side of (3.2) is zero in this case.

I might also consider the case where $A = \{1, 2, 4\}$. Here I have $\mathrm{P}(A) = \frac{1}{2}$. What should $\mathrm{P}(A \,|\, B)$ be? Well, A has two elements from B, and since B has three elements, each element in B has an equal probability of appearing, I would expect $\mathrm{P}(A \,|\, B) = \frac{2}{3}$. To see what (3.2) predicts, note that $A \cap B = \{2, 4\}$ and this has probability $\frac{1}{3}$. Thus, (3.2)'s prediction for $\mathrm{P}(A \,|\, B)$ is $\frac{1}{3}/\frac{1}{2} = \frac{2}{3}$ also.

What follows is another example with one die, but this die is rather pathological. In particular, imagine a six-sided die, so the sample space is again the set $\{1, 2, 3, 4, 5, 6\}$. Now consider the case where $\mathrm{P}(1) = \frac{1}{21}$, $\mathrm{P}(2) = \frac{2}{21}$, $\mathrm{P}(3) = \frac{3}{21}$, etc. In short, $\mathrm{P}(n) = \frac{n}{21}$ when $n \in \{1, 2, 3, 4, 5, 6\}$. Let B again denote the set $\{2, 4, 6\}$ and suppose that $A = \{1, 2, 4\}$. What is $\mathrm{P}(A \,|\, B)$ in this case? Well, A has two of the elements in B. Now B's probability is $\frac{2}{21} + \frac{4}{21} + \frac{6}{21} = \frac{12}{21}$ and the elements from A account for $\frac{6}{21}$, so I would expect that the probability of A given B is the fraction of B's probability that is accounted for by the elements of A, thus $\frac{6}{21}/\frac{12}{21} = \frac{1}{2}$. This is just what is asserted by (3.2).

What follows describe various common applications of conditional probabilities.

3.2 Independent events

An event A is said to be *independent* of an event B in the case that

$$\mathrm{P}(A \,|\, B) = \mathrm{P}(A). \tag{3.3}$$

In English: Events A and B are independent when the probability of A given B is the same as that of A with no knowledge about B. Thus, whether the outcome is in B or not has no bearing on whether it is in A.

Here is an equivalent definition: Events A and B are deemed to be independent whenever $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$. This is equivalent because $\mathrm{P}(A \,|\, B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$. Note that the equality between $\mathrm{P}(A \cap B)$ and $\mathrm{P}(A)\mathrm{P}(B)$ implies that $\mathrm{P}(B \,|\, A) = \mathrm{P}(B)$. Thus, independence is symmetric. Here is the English version of this equivalent definition: Events A and B are independent in the case that the probability of an event being both in A and in B is the product of the respective probabilities that it is in A and that it is in B.

For an example, take the sample space S as in (3.1), take A to be the event that $+$ appears in the third position, and take B to be the event that $+$ appears in the first position. Suppose that the chosen probability function assigns equal weight, $\frac{1}{8}$, to each element in S. Are A and B mutually independent? Well, $\mathrm{P}(A) = \frac{1}{2}$ is as is $\mathrm{P}(B)$. Meanwhile, $\mathrm{P}(A \cap B) = \frac{1}{4}$ which is $\mathrm{P}(A)\mathrm{P}(B)$. Thus, they are indeed independent. By the same token, if A is the event that $-$ appears in the third position, with B as before, then A and B are again mutually independent.

For a second example, consider A to be the event that a plus appears in the first position, and take B to be the event that a minus appears in the first position. In this case, no elements are in both A and B; thus $A \cap B = \emptyset$ and so $\mathrm{P}(A \cap B) = 0$. On the other hand, $\mathrm{P}(A)\mathrm{P}(B) = \frac{1}{4}$. As a consequence, these two events are not independent. (Are you surprised?)

Here is food for thought: Suppose that the sample space in (3.1) represents the set of outcomes that are obtained by measuring your pulse three times and recording $+$ or $-$ for the respective cases when your pulse rate is over 70 or no greater than 70. Do you expect that the event with the first measurement giving $+$ is independent from that where the third measurement gives $+$? I would expect that the third measurement is more likely to exceed 70 then not if the first measurement exceeds 70. If such is the case, then the assignment of equal probabilities to all elements of S does not provide an accurate model for real pulse measurements.

What follows is another example. Take the case of rolling the pathological die. Thus, $S = \{1, 2, 3, 4, 5, 6\}$ and if n is one of these numbers, then $\mathrm{P}(n) = \frac{n}{21}$. Consider the case where $B = \{2, 4, 6\}$ and A is $\{1, 2, 4\}$. Are these independent events? Now, $\mathrm{P}(A) = \frac{7}{21}$, $\mathrm{P}(B) = \frac{12}{21}$ and, as we saw $\mathrm{P}(A \,|\, B) = \frac{1}{2}$. Since $\frac{1}{2} \neq \frac{4}{21} = \mathrm{P}(A)\mathrm{P}(B)$, these events are not independent.

So far, you have seen pairs of events that are not independent. To see an example of a pair of independent events, consider flipping a fair coin twice. The sample space in this case consists of four elements, $S = \{HH, HT, TH, TT\}$. I give S the probability function that assigns $\frac{1}{4}$ to each event in S. Let A denote the event that the first flip gives heads and let B denote the event that the second flip gives heads. Thus, $A = \{HH, HT\}$ and $B = \{HH, TH\}$. Do you expect these events to be independent? In this case, $\mathrm{P}(A) = \frac{1}{2}$ since it has two elements, both with one-fourth probability. For the same reason, $\mathrm{P}(B) = \frac{1}{2}$. Since $A \cap B = \{HH\}$, so $\mathrm{P}(A \cap B) = \frac{1}{4}$. Therefore $\mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B)$ as required for A and B to be independent.

To reiterate, events A and B are independent when knowledge that B has occurred offers no hints towards whether A has also occurred. Here is another example: Roll two standard die. The sample space in this case, S, has 36 elements, these of the form (a, b) where $a = 1, 2, \ldots,$ or 6 and $b = 1, 2, \ldots,$ or 6. I give S the probability function that assigns probability to each element. Let B denote the set of pairs that sum to 7. Thus, $(a, b) \in B$ when $a + b = 7$. Let A denote the event that a is 1. Is A independent from B? To determine this, note that there are 6 pairs in B, these $(1, 6)$, $(2, 5)$, $(3, 4)$, $(4, 3)$, $(5, 2)$ and $(6, 1)$. Thus, $\mathrm{P}(B) = \frac{1}{6}$. Meanwhile, there are six pairs in A; these are $(1, 1)$, $(1, 2)$, $(1, 3)$, $(1, 4)$, $(1, 5)$ and $(1, 6)$. Thus $\mathrm{P}(A) = \frac{1}{6}$. Finally, $A \cap B = (1, 6)$ so $\mathrm{P}(A \cap B) = \frac{1}{36}$. Since this last is $\mathrm{P}(A) \cdot \mathrm{P}(B)$, it is indeed the case that A and B are independent.

Here is a question to ponder: If C denotes the set of pairs (a, b) with $a = 1$ or 2, are C and B independent? The answer is again yes since C has twelve elements so probability $\frac{1}{3}$. Meanwhile, $C \cap B = \{(1, 6), (2, 5)\}$ so $\mathrm{P}(C \cap B) = \frac{2}{36}$. Since this last ratio is equal to $\mathrm{P}(C) \cdot \mathrm{P}(B)$, it is indeed the case that C and B are independent.

3.3 Bayes theorem

Bayes theorem concerns the situation where you have knowledge of the outcome and want to use it to infer something about the cause. This is a typical situation in the sciences. For example, you observe a distribution of traits in the human population today and want to use this information to say something about the distribution of these traits in an ancestral population. In this case, the 'outcome' is the observed distribution of traits in today's population, and the 'cause' is the distribution of traits in the ancestral population.

To pose things in a mathematical framework, suppose that B is a given subset of S; and suppose that we know how to compute the conditional probabilities given B. Thus, $\mathrm{P}(A \,|\, B)$ for various events A. Suppose that we don't know that B actually occurred, but we do see a particular version of A. The question on the table is that of using this A's version of $\mathrm{P}(A \,|\, B)$ to compute $\mathrm{P}(B \,|\, A)$. Said prosaically: What does knowledge of the outcomes say about the probable 'cause'?

To infer causes from outcomes, use the equalities

$$\mathrm{P}(A \,|\, B) = \mathrm{P}(A \cap B)/\mathrm{P}(B) \qquad \text{and} \qquad \mathrm{P}(B \,|\, A) = \mathrm{P}(A \cap B)/\mathrm{P}(A)$$

to write

$$\mathrm{P}(B \,|\, A) = \frac{\mathrm{P}(A \,|\, B)\,\mathrm{P}(B)}{\mathrm{P}(A)}. \tag{3.4}$$

This is the simplest form of 'Bayes theorem'. It tells us the probability of cause B given that outcome A has been observed. What follows is a sample application.

Suppose that 1% of the population have a particular mutation in a certain protein, that 20% of people with this mutation have trouble digesting lactose, and that 5% of the population have trouble digesting lactose. If a classmate has trouble digesting lactose, what is the probability that the classmate has the particular mutation? Here, the outcome is a student with trouble digesting lactose, and we want to infer the probability that the cause is the mutant protein. To this end, let A denote the event that a person has trouble digesting lactose and let B denote the event that a person has the mutated protein. We are told that $\mathrm{P}(A) = 0.05$, that $\mathrm{P}(B) = 0.01$ and that $\mathrm{P}(A \mid B) = 0.2$. According to (3.4), the probability of B given that A occurs, $\mathrm{P}(B \mid A)$, is equal to 0.04. This is the probability that the classmate with lactose intolerance has the given mutation.

A similar application of Bayes' theorem is used when DNA evidence is invoked in a criminal investigation. Suppose, for example that 10% of the population has a certain sequence of paired DNA bases on a particular stretch of DNA, that 5% of the population have committed a felony, and that 20% of the felonies are committed by people with the given sequence of bases. An individual is charged with the crime. As it turns out, this individual does exhibit this special sequence of base pairs. What is the probability that the individual is guilty? To analyze this, let A denote the event that an individual has the given sequence of base pairs and let B denote the event that an individual has committed a felony. We are told that $\mathrm{P}(A) = 0.1$, that $\mathrm{P}(B) = 0.05$ and that $\mathrm{P}(A \mid B) = 0.2$. An application of (3.4) finds that the probability of interest, $\mathrm{P}(B \mid A)$, is equal to 0.1.

Applications of Bayes' theorem in medicine are very common. Suppose you, a pediatrician, see a child that is running a mild temperature but no other symptoms. You know that 95% of children with either a viral or bacterial infection run a temperature, that 5% of children run similar temperatures whether sick or not, and that 1% of children at any given time have some sort of infection. What is the probability that the child has an infection given that the child has a mild temperature? To answer this, we set A to denote the event that the child has a temperature, B to denote the event that the child has an infection. We are told that $\mathrm{P}(A) = 0.05$, $\mathrm{P}(B) = 0.001$ and $\mathrm{P}(A \mid B) = 0.95$. We are asking for $\mathrm{P}(B \mid A)$. Bayes' theorem finds this to equal 0.19.

What follows is an example with my pathological die. Thus $S = \{1, 2, 3, 4, 5, 6\}$ and $\mathrm{P}(n) = \frac{n}{21}$ when n is one of the integers from S. Let B denote the set $\{2, 4, 6\}$ and let A denote the set $\{2, 4\}$. We saw previously that $\mathrm{P}(A \mid B) = \frac{5}{6}$. What is $\mathrm{P}(B \mid A)$? This is the fraction of A's probability that is accounted for by the elements that are both in A and in B. Since every element in A is also in B (this is to say that $A \subset B$), all A's probability is accounted for by elements of B. Thus, I should conclude that $\mathrm{P}(B \mid A) = 1$. Meanwhile, Bayes' theorem finds that $\mathrm{P}(B \mid A) = \mathrm{P}(A \mid B) \cdot \mathrm{P}(B)/\mathrm{P}(A) = \frac{5}{6} \cdot \frac{12}{21} \Big/ \frac{10}{21} = 1$ as required.

3.4 Decomposing a subset to compute probabilities

It is often the case (as we will see in subsequent chapters) that it is easier to compute conditional probabilities first; then use them to compute unconditional probabilities of interest. In fact, this is a very common application of the notion of conditional probabilities.

What follows is a simple example. I have two coins in my pocket, one is fair so that the probability of heads is $\frac{1}{2}$. The other is not fair as the probability of heads is only $\frac{1}{4}$. I choose one of these coins while blind folded and then flip it. What is the probability of heads appearing? My logic here is based on the following reasoning: The probability of heads is equal to the sum of

$$(\text{the probability of heads given that the fair coin}) \times (\text{the probability that the coin is fair})$$

plus

$$(\text{the probability of heads given the unfair coin}) \times (\text{the probability that the coin is unfair}).$$

Thus, I would say that the probability of heads in this case is $(\frac{1}{2} \times \frac{1}{2}) + (\frac{1}{4} \times \frac{1}{2}) = \frac{3}{8}$. I trust that you notice here the appearance of conditional probabilities.

In general, the use of conditional probabilities to compute unconditional probabilities arises in the following situation: Suppose that a sample space, S, is given. In the coin example above, I took S to be set with four elements $\{(F, H), (F, T), (U, H), (U, T)\}$, where the symbols have the following meaning: First, (F, H) denotes the case where the coin is fair and heads appears and (F, T) that where the coin is fair and tails appears. Meanwhile, (U, H)

denotes the case where the coin is unfair and heads appears, and (U, T) that where the coin is unfair and tails appears. Keep this example in mind as we consider the abstract situation where S is just some given sample space of interest.

Now look for a convenient decomposition of S into subsets that do not share elements. Let me use N to denote the number of such sets. Thus, I write

$$S = B_1 \cup B_2 \cup \cdots \cup B_N,$$

where B_1, B_2, ..., B_N are subsets of S with $B_k \cap B_{k'} = \emptyset$ when $k \neq k'$. What I mean by 'convenient' is clarified in what follows. However, one criteria is that the probabilities of the various B_k should be easy to find.

In the case of my two coins, I used two sets for such a decomposition; I took B_1 to be the event that the coin is fair and B_2 to be the event that the coin is unfair. In this case, I told you that both $\mathrm{P}(B_1)$ and $\mathrm{P}(B_2)$ are equal to $\frac{1}{2}$.

To return now to the abstract situation, suppose that A is an event in S and I want the probability of A. If, for each $1 \leq k \leq N$, I know the conditional probability of A given B_k, then I can write

$$\mathrm{P}(A) = \mathrm{P}(A \,|\, B_1) \cdot \mathrm{P}(B_1) + \mathrm{P}(A \,|\, B_2) \cdot \mathrm{P}(B_2) + \cdots + \mathrm{P}(A \,|\, B_N) \cdot \mathrm{P}(B_N). \tag{3.5}$$

In words, this says the following:

> *The probability of an outcome from A is the probability that an outcome from A occurs given that B_1 occurs times the probability of B_1, plus the probability that an outcome from A occurs given that B_2 occurs times the probability of B_2, plus ... etc.*

The formula in (3.5) is useful only to the extent that the conditional probabilities $\mathrm{P}(A \,|\, B_1), \mathrm{P}(A \,|\, B_2), \ldots, \mathrm{P}(A \,|\, B_N)$ and the probabilities of each B_k are easy to compute. This is what I mean by the use of the descriptive 'convenient' when I say that one should look for a 'convenient' decomposition of S as $B_1 \cup B_2 \cup \cdots \cup B_N$.

By the way, do you recognize (3.5) as a linear equation? You might if you denote $\mathrm{P}(A)$ by y, each $\mathrm{P}(B_j)$ by x_j and $\mathrm{P}(A \,|\, B_j)$ by a_j so that this reads

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_N x_N \ .$$

Thus, linear systems arise!

Here is why (3.5) holds: Remember that $\mathrm{P}(A \,|\, B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$. Thus, $\mathrm{P}(A \,|\, B) \cdot \mathrm{P}(B) = \mathrm{P}(A \cap B)$. Therefore, the right hand side of (3.5) states that

$$\mathrm{P}(A) = \mathrm{P}(A \cap B_1) + \mathrm{P}(A \cap B_2) + \cdots + \mathrm{P}(A \cap B_N).$$

This now says that the probability of A is obtained by summing the probabilities of the parts of A that appear in each B_n. That such is the case follows when the B_n's don't share elements but account for all of the elements of S. Indeed, if, say B_1 shared an element with B_2, then that element would be overcounted on the right-hand side of the preceding equation. On the other hand, if the B_n's don't account for all elements in S, then there might be some element in A that is not accounted for by the sum on the right-hand side.

What follows is a sample application of (3.5): Suppose that I have six coins where the probability of heads on the first is $\frac{1}{2}$, that on the second is $\frac{1}{4}$, that on the third is $\frac{1}{8}$, that on the fourth is $\frac{1}{16}$, that on the fifth is $\frac{1}{32}$, and that on the last is $\frac{1}{64}$. Suppose that I label these coins by $1, 2, \ldots, 6$ so that the probability of heads on the m'th coin is 2^{-m}. Now I also have my pathological die, the one where the probability of the face with number $n \in \{1, 2, \ldots, 6\}$ is $\frac{n}{21}$. I roll my pathological die and the number that shows face up tells me what coin to flip. All of this understood, what is the probability of the flipped coin showing heads?

To answer this question, I first note that the relevant sample space has 12 elements, these of the form (n, H) or (n, T), where n can be 1, 2, 3, 4, 5, or 6. This is to say that (n, H) is the event that the n'th coin is chosen and heads appears, while (n, T) is the event that the n'th coin is chosen and tails appears. The set A in this case is the event that H appears. To use (3.5), I first decompose my sample space into 6 subsets, $\{B_n\}_{1 \leq n \leq 6}$, where B_n is the event that the n'th coin is chosen. This is a 'convenient' decomposition because I know $\mathrm{P}(B_n)$, this $\frac{n}{21}$. I also know $\mathrm{P}(A \,|\, B_n)$, this 2^{-n}. Granted this, then (3.5) finds that the probability of heads is equal to

$$\frac{1}{2} \cdot \frac{1}{21} + \frac{1}{4} \cdot \frac{2}{21} + \frac{1}{8} \cdot \frac{3}{21} + \frac{1}{16} \cdot \frac{4}{21} + \frac{1}{32} \cdot \frac{5}{21} + \frac{1}{64} \cdot \frac{6}{21} = \frac{5}{56}.$$

The next example comes from genetics. To start, suppose that I am breeding peas à la Gregor Mendel, and I have a variety that gives wrinkled peas and another that gives smooth peas. I know further that the gene for wrinkled peas is recessive. This is to say that a parent plant can have one of three 'genotypes', these denoted by ww, ws or ss. A plant with wrinkled peas must be of type ww. A plant with ws or ss has smooth peas. Now, I breed a two pea plants, not knowing whether they give wrinkled peas or smooth peas, and ask for the probability that the offspring is wrinkled. The assumption here is that the offspring inherits one gene from each parent. Thus, if a parent has genotype ws, then the offspring can inherit either w or s from that parent. If the parent has genotype ss, then the offspring inherits s from the parent.

My sample space for this problem consists of the possible triples that label the genotypes of Parent #1, Parent #2 and the offspring. For example, (ws, ws, ss) is an example of a triple that appears in the sample space. This labels the case where both parents have genotype ws and the offspring has genotype ss. Another triple from the sample space is (ws, ws, ws). Here is a third: (ss, ws, ss). A fourth is (ss, ws, sw). And so on. Note that only triples of the form $(--, --, ww)$ give rise to a plant with wrinkled peas.

Let A denote the event that the offspring has wrinkled peas. Let B_1 denote the event that both parents are of type ww, let B_2 denote the event that the first parent is ww and the second is ws, let B_3 denote the event that the first parent is ws and the second ww, let B_4 denote the event that the both parents are ws, and let B_5 denote the event that at least one parent is ss. Note that the collection $\{B_n\}_{1\leq n\leq 5}$ are such that their union is all of S, and no two share the same element. Also, I know what $\mathrm{P}(A \,|\, B_n)$ is for each n if I assume that a parent gives one of its two genes to the offspring, and that there is equal probability of giving one or the other. Thus, $\mathrm{P}(A \,|\, B_1) = 1$, $\mathrm{P}(A \,|\, B_2) = \mathrm{P}(A \,|\, B_3) = \frac{1}{2}$, $\mathrm{P}(A \,|\, B_4) = \frac{1}{4}$ and $\mathrm{P}(A \,|\, B_5) = 0$. As a consequence, (3.5) reads

$$\mathrm{P}(\text{wrinkled offspring}) = \mathrm{P}(B_1) + \tfrac{1}{2}\Big(\mathrm{P}(B_2) + \mathrm{P}(B_3)\Big) + \tfrac{1}{4}\mathrm{P}(B_4).$$

Thus, it is enough for me to know the probabilities for the possible genotypes of the parents.

The point here is that the conditional probabilities $\{\mathrm{P}(A \,|\, B_n)\}_{1\leq n\leq 5}$ are easy to compute. Note also that their computation is based on theoretical considerations. This is to say that we made the *hypothesis* that 'a parent gives one of its two genes to the offspring, and that there is equal probability of giving one or the other.' The formula given above for P(wrinkled offspring) should be viewed as a prediction to be confirmed or not by experiments.

What follows is another example from biology. Suppose that I am concerned with a stretch of DNA of length N, and want to know what the probability of *not seeing* the base guanine in this stretch. Let A denote the event of that there is no guanine in a particular length N stretch of DNA. Let B_1 denote the event that there is no guanine in the stretch of length $N-1$, and let B_2 denote the event that guanine does appear in this length $N-1$ stretch. In this case, $\mathrm{P}(A \,|\, B_2) = 0$ since A and B_2 are disjoint. What about $\mathrm{P}(A \,|\, B_1)$? This is the probability that guanine is not in the Nth site if none has appeared in the previous $N-1$ sites. Of course, the probability of guanine appearing in the Nth site may or may not be affected by what is happening in the other sites. Let us make the *hypothesis* that each of the four bases has equal probability to appear in any given site. Under this hypothesis, the probability of seeing no guanine in the Nth site is $\frac{3}{4}$ since there are four bases in all and only one of them, guanine, is excluded. Thus, under our hypothesis that each base has equal probability of appearing in any given site, we find that $\mathrm{P}(A \,|\, B_1) = \frac{3}{4}$. This understood, it then follows from (3.5) that $\mathrm{P}(A) = \frac{3}{4}\mathrm{P}(B_1)$.

Now we can compute $\mathrm{P}(B_1)$ in an analogous fashion by considering the relative probability of no guanine in the $(N-1)$st site given that none appears in the previous $N-2$ sites. Under our hypothesis of equal probabilities for the bases, this gives $\mathrm{P}(B_1) = \frac{3}{4}$ times the probability of no guanine in the first $N-2$ sites. We can use this trick again to compute the latter probability; it equals $\frac{3}{4}$ times the probability of no guanine in the first $N-3$ sites. Continuing in this vein finds $\mathrm{P}(A)$ to be equal to the product of N copies of $\frac{3}{4}$, thus $(\frac{3}{4})^N$.

This computation of $\mathrm{P}(A) = (\frac{3}{4})^N$ is now a theoretical prediction based on the hypothesis that the occurrence of any given base in any given DNA site has the same probability as that of any other base. You are challenged to think of an experiment that will test this prediction.

3.5 More linear algebra

What follows is meant as a second illustration of how linear algebra appears when considering probabilities. Let A_1, A_2, A_3 and A_4 denote the events that a given site in DNA has base A, G, C, and T. Let B_1, B_2, B_3 and B_4 denote the analogous event for the adjacent site to the $5'$ end of the DNA. (The ends of a DNA molecule are denoted $3'$ and $5'$ for reasons that have to do with a tradition of labeling carbon atoms on sugar molecules.) According to the rule in (3.5), we must have

$$\begin{aligned}
\mathrm{P}(A_1) &= \mathrm{P}\,(A_1 \,|\, B_1)\cdot\mathrm{P}(B_1) + \mathrm{P}\,(A_1 \,|\, B_2)\cdot\mathrm{P}(B_2) + \mathrm{P}\,(A_1 \,|\, B_3)\cdot\mathrm{P}(B_3) + \mathrm{P}\,(A_1 \,|\, B_4)\cdot\mathrm{P}(B_4),\\
\mathrm{P}(A_2) &= \mathrm{P}\,(A_2 \,|\, B_1)\cdot\mathrm{P}(B_1) + \mathrm{P}\,(A_2 \,|\, B_2)\cdot\mathrm{P}(B_2) + \mathrm{P}\,(A_2 \,|\, B_3)\cdot\mathrm{P}(B_3) + \mathrm{P}\,(A_2 \,|\, B_4)\cdot\mathrm{P}(B_4),\\
\mathrm{P}(A_3) &= \mathrm{P}\,(A_3 \,|\, B_1)\cdot\mathrm{P}(B_1) + \mathrm{P}\,(A_3 \,|\, B_2)\cdot\mathrm{P}(B_2) + \mathrm{P}\,(A_3 \,|\, B_3)\cdot\mathrm{P}(B_3) + \mathrm{P}\,(A_3 \,|\, B_4)\cdot\mathrm{P}(B_4),\\
\mathrm{P}(A_4) &= \mathrm{P}\,(A_4 \,|\, B_1)\cdot\mathrm{P}(B_1) + \mathrm{P}\,(A_4 \,|\, B_2)\cdot\mathrm{P}(B_2) + \mathrm{P}\,(A_4 \,|\, B_3)\cdot\mathrm{P}(B_3) + \mathrm{P}\,(A_4 \,|\, B_4)\cdot\mathrm{P}(B_4).
\end{aligned}$$

So, we have a 4×4 matrix M whose entry in row i and column j is $\mathrm{P}\,(A_i \,|\, B_j)$. Now write each $\mathrm{P}(A_i)$ as y_i and each $\mathrm{P}(B_j)$ as x_j, and these last four equations can be summarized by the assertion that

$$y_i = \sum_{1\le j\le 4} M_{ij}x_j \qquad \text{for each } i = 1,\, 2,\, 3,\, \text{and } 4.$$

This can be viewed as a system of four linear equations!

3.6 An iterated form of Bayes' theorem

Suppose that $S = B_1 \cup B_2 \cup \cdots \cup B_N$ is the union of N pairwise disjoint subsets. Suppose also that A is a given subset of S and we know that an outcome from A appears. Given this knowledge, what is the probability that the outcome was from some given B_k? Here is an example: Take S to be the set of all diseases and A diseases where the lungs fill with fluid. Take B_1 to be pneumonia, B_2 to be ebola viral infection, B_3 to be West Nile viral infection, etc. An old man's lungs fill with fluid. What is the probability that he has West Nile viral infection? We are therefore searching for $\mathrm{P}\,(B_3 \,|\, A)$.

Suppose that we know the probability of the lung filling with fluid with the disease that corresponds to B_k. This is to say that we know each $\mathrm{P}\,(A \,|\, B_k)$. Suppose we also know $\mathrm{P}(B_k)$; the probability of catching the disease that corresponds to B_k. How can we use this information to compute $\mathrm{P}\,(B_3 \,|\, A)$?

This is done using the following chain of equalities: I first write

$$\mathrm{P}\,(B_3 \,|\, A) = \mathrm{P}(B_3 \cap A)/\mathrm{P}(A) = \mathrm{P}\,(A \,|\, B_3)\cdot\mathrm{P}(B_3)/\mathrm{P}(A)$$

using the original form of Bayes' theorem. To compute $\mathrm{P}(A)$, I use (3.5). Together, these two equalities imply the desired one:

$$\mathrm{P}\,(B_3 \,|\, A) = \frac{\mathrm{P}\,(A \,|\, B_3)\,\mathrm{P}(B_3)}{\mathrm{P}\,(A \,|\, B_1)\cdot\mathrm{P}(B_1) + \cdots + \mathrm{P}\,(A \,|\, B_N)\cdot\mathrm{P}(B_N)}.$$

Of course, I make such an equation for any given $\mathrm{P}\,(B_k \,|\, A)$, and this gives the most general form of Bayes theorem:

$$\mathrm{P}\,(B_k \,|\, A) = \frac{\mathrm{P}\,(A \,|\, B_k)\,\mathrm{P}(B_k)}{\mathrm{P}\,(A \,|\, B_1)\cdot\mathrm{P}(B_1) + \cdots + \mathrm{P}\,(A \,|\, B_N)\cdot\mathrm{P}(B_N)}. \tag{3.6}$$

This equation provides the conditional probability of B_k given that A occurs from the probabilities of the various B_k and the conditional probabilities that A occurs given that any one of these B_k occur.

To see how this works in practice, return to the example I gave above where I have six coins, these labeled $\{1, 2, 3, 4, 5, 6\}$; and where the mth coin has probability 2^{-m} of landing with the head up when flipped. I also have my pathological six-sided die, where the probability of the nth face appearing when rolled is $\frac{n}{21}$. As before, I first roll the die and if the nth face appears, I flip the coin with the label n. I don't tell you which coin was flipped, but I do tell you that heads appeared. What is the probability that the coin #3 was flipped?

I can use (3.6) to compute this probability. For this purpose, let A denote the event that heads appears. For each $n \in \{1, 2, \ldots, 6\}$, let B_n denote the event that I flipped the coin labeled by n. Thus, my question asks for $\mathrm{P}(B_3 \mid A)$.

You have all of the ingredients to compute $\mathrm{P}(B_3 \mid A)$ via (3.6) since you are told that $\mathrm{P}(A|B_n) = 2^{-n}$ and $\mathrm{P}(B_n) = \frac{n}{21}$. In fact, we have already computed the sum that makes up the denominator in (3.6), this being $\frac{5}{56}$. As a consequence, the equality in (3.6) finds $\mathrm{P}(B_3 \mid A)$ equal to $\frac{1}{8} \cdot \frac{3}{21} / \frac{5}{56} = \frac{1}{5}$.

3.7 Exercises:

1. This exercise concerns the sample space S that is depicted in (3.1). If S represents the outcomes for three pulse rate measurements of a given individual, it is perhaps more realistic to take the following probability function: The function P assigns probability $\frac{1}{3}$ to $+++$ and to $---$ while assigning $\frac{1}{18}$ to each of the remaining elements.

   (a) Is the event that $+$ appears first independent of the event that $+$ appears last?

   (b) Is the event that $+$ appears second independent for the event that $+$ appears last?

   (c) What is the conditional probability that $+$ appears first given that $-$ appears second?

2. Suppose that that the probability of a student lying in the infirmary is 1%, and that the probability that a student has an exam on any given day is 5%. Suppose as well that 6% of students with exams go to the infirmary. What is the probability that a student in the infirmary has an exam on a given day?

3. Label the four bases that are used by DNA as $\{1, 2, 3, 4\}$.

   (a) Granted this labeling, write down the sample space for the possible bases at two given sites on the molecule.

   (b) Invent a probability function for this sample space.

   (c) Let A_j for $j = 1, 2, 3, 4$ denote the event in this two-site sample space that the first site has the base j, and let B_j for $j = 1, \ldots, 4$ denote the analogous event for the second site. Use the definition of conditional probability to explain why, for *any* probability function and for any k, $\mathrm{P}(A_1 \mid B_k) + \mathrm{P}(A_2 \mid B_k) + \mathrm{P}(A_3 \mid B_k) + \mathrm{P}(A_4 \mid B_k)$ must equal 1.

   (d) Is there a choice for a probability function on the sample space that makes $\mathrm{P}(A_1 \mid B_1) = \mathrm{P}(B_1 \mid A_1)$ in the case that A_1 and B_1 are not independent? If so, give an example. If not, explain why.

4. This problem refers to the scenario that I described above where I have six coins, these labeled $\{1, 2, 3, 4, 5, 6\}$; and where the mth coin has probability 2^{-m} of landing with the head up when flipped. I also have my pathological six-sided die, where the probability of the nth face appearing when rolled is $\frac{n}{21}$. As before, I first roll the die and if the nth face appears, I flip the coin with the label n. I don't tell you which coin was flipped, but I do tell you that heads appeared.

   (a) For $n = 1, 2, 4, 5$, and 6, give the probability that the coin with the label n was flipped.

   (b) For what n, if any, is the event that the nth face appears independent from the event that heads appears.

5. Suppose that A and B are subsets of a sample space with a probability function, P. Suppose in addition that $\mathrm{P}(A) = \frac{4}{5}$ and $\mathrm{P}(B) = \frac{3}{5}$. Explain why $\mathrm{P}(B \mid A)$ is at least $\frac{1}{2}$.

6. For many types of cancer, early detection is the key to successful treatment. Prostate cancer is one of these. For early detection, the National Cancer Institute suggests screening of patients using the Serum Prostate-Specific Antigen (PSA) Test. There is controversy due to the lack of evidence showing that early detection of prostate cancer and aggressive treatment of early cancers actually reduces mortality. Also, this treatment can often lead to complications of impotence and incontinence.

   Here is some terminology that you will meet if you go on to a career in medicine: The *sensitivity* of a test is the probability of a positive test when the patient has the disease, and the *specificity* of a test is the probability of a negative test when the patient does not have the disease. In the language of conditional probability,

- *Sensitivity* is the conditional probability of a positive test given that the disease is present.
- *Specificity* is the conditional probability of a negative test given that the disease is not present.

The standard PSA test to detect early stage prostate cancer has Sensitivity $= 0.71$ and Specificity $= 0.91$. Thus, 0.71 is the conditional probability of a positive test given that a person does have prostate cancer. And, 0.91 is the conditional probability of a negative test given that a person does not have prostate cancer. Note for what follows that roughly 0.7% of the male population is diagnosed with prostate cancer each year.

Granted this data, here is the question that this problem will answer:

If a patient receives a positive test for prostate cancer, what is the probability he truly has cancer?

To answer this question, let A denote the event that a person has a positive test, and let B denote the event that a person has prostate cancer. This question is asking for the conditional probability of B given A; thus $\mathrm{P}(B \mid A)$. The data above gives $\mathrm{P}(A \mid B) = 0.71$ and $\mathrm{P}(B) = 0.007$ and also $\mathrm{P}(A^c \mid B^c) = 0.91$ where A^c is the event that a person has a negative test, and B^c is the event that a person does not have cancer. As the set of questions that follow demonstrate, the information given is sufficient to answer the question posed above.

(a) Why is $\mathrm{P}(B \mid A) = (0.71) \times (0.007)/\mathrm{P}(A) \approx 0.005/\mathrm{P}(A)$?

If you answered this, then the task is to find $\mathrm{P}(A)$.

(b) Why is $\mathrm{P}(A) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^c)$?

(c) Why is $\mathrm{P}(A \cap B^c) = \mathrm{P}(B^c) - \mathrm{P}(A^c \cap B^c)$?

(d) Why is $\mathrm{P}(B^c) = 1 - \mathrm{P}(B)$?

If you answered these last three questions, you have discovered that $\mathrm{P}(A) = \mathrm{P}(A \cap B) + 1 - \mathrm{P}(B) - \mathrm{P}(A^c \cap B^c)$, and thus, using what you just derived, $\mathrm{P}(A) = 0.993 + \mathrm{P}(A \cap B) - \mathrm{P}(A^c \cap B^c)$.

(e) Why is $\mathrm{P}(A \cap B) = (0.71) \times (0.007) \approx 0.005$?

(f) Why is $\mathrm{P}(B^c \cap A^c) = (0.91) \times (0.993) \approx 0.904$?

If you answered (a)–(f), then you have found that $\mathrm{P}(A) \approx 0.094$ and so the question in italics asked above is ≈ 0.054.

For more info, see: `http://imsdd.meb.uni-bonn.de/cancernet/304727.html`

7. Many genetic traits are controlled by a single gene with two alleles, one dominant (A) and one recessive (a), that are passed on generation by generation. Albinism is one phenotype that can be described this way. A person is albino if he/she has gotten two copies of the recessive alleles (aa); one from each of his/her parents.

   Erin knows for certain that both her parents are carriers for the albino phenotype. That is, they both have the Aa genotype, one dominant allele and one recessive allele.

   (a) What is the sample space for the possible pairs of alleles that Erin could have inherited from her parents?

   Now assume each allele is equally likely to be passed from Erin's parents to Erin.

   (b) Explain how this information gives a probability function on the sample space that you found in (a).

   (c) Use the probability function from (b) to give the probability that Erin is albino.

   (d) Given that Erin is not albino, what is the probability that she has the albino allele?

   Mendel's paper: `http://www.mendelweb.org/Mendel.html`

CHAPTER

FOUR

Linear transformations

My purpose here is to give some examples of linear transformations that arise when thinking about probability as applied to problems in biology. As you should recall, a linear transformation on $\mathbf{R}^n$ can be viewed as the effect of multiplying vectors by a given matrix. If A is the matrix and $\vec{v}$ is any given vector, the transformation has the form $\vec{v} \to A\vec{v}$, where $A\vec{v}$ is the vector with components

$$(A\vec{v})_j = \sum_k A_{jk} v_k. \tag{4.1}$$

Thus,

$$\begin{aligned} (A\vec{v})_1 &= A_{11}v_1 + A_{12}v_2 + \cdots + A_{1n}v_n \\ (A\vec{v})_2 &= A_{21}v_1 + A_{22}v_2 + \cdots + A_{2n}v_n \\ &\vdots \\ (A\vec{v})_n &= A_{n1}v_1 + A_{n2}v_2 + \cdots + A_{nn}v_n \end{aligned} \tag{4.2}$$

4.1 Protein molecules

A protein molecule is made by attaching certain small molecules end to end. The small molecules are amino acids, and there are 20 that commonly appear. The amino acids that comprise a given protein can number in the thousands, but each is just one of these 20 varieties. The amino acids are numbered from one end. (The ends of the amino acids and thus the two ends of a protein are distinguished by their chemical composition.)

It is often the case that the protein found in one individual differs from that in another because the 10th or 127th or 535th amino acid from the start of the chain differ in the two versions. Some such substitutions drastically affect the protein function, others are benign.

Suppose that the 127th amino acid along a protein molecule can vary amongst a certain subset of 5 amino acids without effecting the protein's function. Now, imagine a population of individuals (mice, worms, fish, bacteria, whatever) that is watched over successive generations. Let $p_i(t)$ denote the probability that amino acid $i \in \{1, ..., 5\}$ occurs as the 127th amino acid in the tth generation of the given population. Even if a parent in generation t has amino acid numbered i in the 127th position, a mutation in the coding gene going from generation t to $t+1$ can change the coding so that the offspring has amino acid $j \neq i$ in the 127th position. There is some probability A_{ji} for this to happen. This number may depend on i and j. The reason is that the genetic code for an amino acid is 3 letters long, and so changing from amino acid i to j has the highest probability if only one letter separates their corresponding codes.

In any event, I can write

$$p_i(t+1) = \sum_j A_{ij} p_j(t). \tag{4.3}$$

This equation says that the probability of seeing amino acid i in the 127th position is obtained by using the formula in (3.5). In words:

The probability of seeing amino acid i in generation $t+1$ is the conditional probability that the amino acid at position 127 is i in the offspring given that it is 1 in the parent times the probability that amino acid 1 appears in generation t, plus the conditional probability that the amino acid at position 127 is i in the offspring given that it is 2 in the parent times the probability that amino acid 2 appears in generation t, plus ... etc.

The point here is that the vector $\vec{p}(t)$ with entries $p_i(t)$ is changed via $\vec{p}(t) \to \vec{p}(t+1) = A\vec{p}(t)$ after each generation. Here, A is the matrix whose ij entry is A_{ij}.

4.2 Protein folding

The long string of amino acids that comprise a protein does not appear in a cell as a straight string. Rather, the string is folded back on itself in a very complicated manner. The geometry of the folding helps determine the activity of the protein. A simple model has the folding occurring due to the fact that the angles of attachment of one amino acid to the next can vary. If the protein has n amino acids, and each can bond to its neighbor at some D possible angles, then there are $N = n^D$ possible configurations for the protein to fold into. Label the possible configurations by the integers starting at 1 and ending at N.

As it turns out, a protein configuration is rarely static. Rather, the protein fluctuates from one configuration to another over time. This fluctuation is due to a combination of effects, some classical (such as collisions between other molecules) and some quantum mechanical.

In any event, if i and j both label configurations, there is some probability, P_{ij}, of the protein being in configuration j at time t and configuration i at time $t+1$. Those of you with some chemical background may recall that thermodynamical considerations put

$$P_{ij} = e^{-(E(i)-E(j))/kT}$$

where $E(\ \cdot\ )$ is the free energy of the configuration, k is Boltzman's constant and T is the temperature in degrees Kelvin.

In any event, let $v_j(t)$ denote the probability that protein is in configuration i at time t. Then, $v_i(t+1) = \sum_j P_{ij} v_j(t)$. Thus, the vector $\vec{v}(t+1)$ in $\mathbf{R}^N$ is obtained from $\vec{v}(t)$ by the action of the linear transformation whose matrix has components P_{ij}.

CHAPTER

FIVE

How matrix products arise

What follows are some areas in biology and statistics where matrix products appear.

5.1 Genomics

Suppose that a given stretch of DNA coding for cellular product is very mutable, so that there are some number, N, of possible sequences that can appear in any given individual (this is called 'polymorphism') in the population. To elaborate, a strand of DNA is a molecule that appears as a string of small, standard molecules that are bound end to end. Each of these standard building blocks can be one of four, labeled C, G, A and T. The order in which they appear along the strand determines any resulting cellular product that the given part of the DNA molecule might produce. For example, AAGCTA may code for a different product than GCTTAA.

As it turns out, there are stretches of DNA where the code can be changed without damage to the individual. What with inheriting genes from both parents, random mutations over the generations can then result in a population where the codes on the given stretch of DNA vary from individual to individual. In this situation, the gene is called 'polymorphic'. Suppose that there are N different possible codes for a given stretch. For example, if one is looking at just one particular site along a particular DNA strand, there could be at most $N = 4$ possibilities at that site, namely C, G, A or T. Looking at two sites gives $N = 4 \times 4 = 16$ possibilities.

I am going to assume in what follows that sites of interest along the DNA is inherited from parent to child only from the mother or only from the father. Alternately, I will assume that I am dealing with a creature such as a bacteria that reproduces asexually. This assumption simplifies the story that follows.

Let us now label the possible sequences for the DNA site under discussion by integers starting from 1 and going to N. At any given generation t, let $p_j(t)$ denote the frequency of the appearance of the jth sequence in the population at generation t. These frequencies then change from one generation to the next in the following manner: The probability of any given sequence, say i, appearing in generation $t + 1$ can be written as a sum:

$$p_i(t+1) = \mathrm{P}\,(i \,|\, 1)\, p_1(t) + \mathrm{P}\,(i \,|\, 2)\, p_2(t) + \cdots + \mathrm{P}\,(i \,|\, N)\, p_N(t), \tag{5.1}$$

where each $\mathrm{P}\,(i \,|\, j)$ can be viewed as the conditional probability that a parent with sequence j produces an offspring with sequence i. This is to say that the probability of sequence i appearing in an individual in generation $t+1$ is equal to the probability that sequence 1 appears in the parent times the probability of a mutation that changes sequence 1 to sequence i, plus the probability that sequence 2 appears in the parent times the probability of a mutation that changes sequence 2 to sequence i, and so on.

We can write the suite of N versions of (5.1) using our matrix notation by thinking of the numbers $\{p_j(t)\}_{1 \le j \le N}$ as defining a column vector, $\vec{p}(t)$, in $\mathbf{R}^N$, and likewise the numbers $\{p_i(t+1)\}_{1 \le i \le N}$ as defining a second column vector, $\vec{p}(t+1)$ in $\mathbf{R}^N$. If I introduce the $N \times N$ matrix A whose entry in the jth column and ith row is $\mathrm{P}\,(i \,|\, j)$, then (5.1) says in very cryptic shorthand:

$$\vec{p}(t+1) = A\vec{p}(t). \tag{5.2}$$

I can sample the population at time $t = T =$ now, and thus determine $\vec{p}(T)$, or at least the proxy that takes $p_i(T)$ to

be the percent of people in the population today that have sequence i. One very interesting question is to determine $\vec{p}(T')$ at some point far in the past, thus $T' \ll T$. For example, if we find $\vec{p}(T')$ such that all $p_i(T')$ are zero but a very few, this then indicates that the population at time T' was extremely homogeneous, and thus presumably very small.

To determine $\vec{p}(T')$, we use (5.2) in an iterated form:

$$\vec{p}(T) = A\vec{p}(T-1) = AA\vec{p}(T-2)) = AAA\vec{p}(T-3) = \cdots = A\cdots A\vec{p}(T'), \tag{5.3}$$

where the final term has $T - T'$ copies of A multiplying one after the other.

On a similar vein, we can use (5.2) to predict the distribution of the sequences in the population at any time $T' > T$ by iterating it to read

$$\vec{p}(T') = A\cdots A\vec{p}(T). \tag{5.4}$$

Here, the multiplication is by $T' - T$ successive copies of the matrix A.

By the way, here is a bit of notation: This sort of sequence $\vec{p}\{(t)\}_{t=0,1,...}$ of vectors of probabilities is an example of a *Markov chain*. In general, a Markov chain is a sequence of probabilities, $\{\mathrm{P}(0), \mathrm{P}(1), \mathrm{P}(2), \ldots, \}$ where the Nth probability $\mathrm{P}(N)$ depends only on the probabilities with numbers that are less than N.

5.2 How bacteria find food

If you put certain sorts of bacteria on one side of a petri dish and put some sugar some distance away on the other, the bacteria will migrate towards the sugar. Apparently, the sugar diffuses in the petri dish, so that there is slight concentration everywhere, with most of the concentration in the original spot. The bacteria are sensitive to the different levels at their front and rear ends and tend to move in the direction of the greater level. At the expense of borrowing from a multivariable calculus course, the bacteria sense the direction of the *gradient* of the sugar concentration. Even so, their movement from low to high concentration has a certain randomness to it; at any given step there is some probability that they will move in the wrong direction.

What follows is a simplistic model for this: Imagine bacteria moving in steps along the x-axis along the segment that stretches from $x = 1$ to $x = N$. Here, N is some large integer. Suppose the sugar is placed initially where x = N, and the bacteria is placed initially at $x = 1$. Our model also supposes that the bacteria moves one unit per second, with probability $q \in (0, 1)$ of moving to the right at any given step, and probability $1 - q$ of moving to the left unless it is at the end of the interval. If it is at the $x = 1$ end, it moves to the right with probability q and stays put with probability $1 - q$. If it is at the $x = N$ end, it stays put with probability q and moves to the left with probability $1 - q$. In our model, q is independent of position, and x$\in \{1, ..., N\}$. Such would be roughly the case in the real petri dish where the sugar concentration gradient is relatively constant. We should take $q > \frac{1}{2}$ in the case that the bacteria is attracted to the sugar.

For each time step t, and $j \in \{1, ..., N\}$, let $p_j(t)$ denote the probability that the bacteria is at position j at time t. For example, $p_1(0) = 1$ and $p_j(0) = 0$ for $j > 1$. Our model then says that the probability of finding the bacteria at position $j \neq 1$ of N at time step $t > 0$ is equal to the probability that it is at position $j - 1$ at time $t - 1$ times the probability that it moves one step to the right, plus the probability that it is at position $j + 1$ at time $t - 1$ and moves one step to the left. Thus,

$$p_j(t+1) = qp_{j-1}(t) + (1-q)p_{j+1}(t) \qquad \text{when } 2 \leq j \leq N-1. \tag{5.5}$$

By the same token,

$$p_1(t+1) = (1-q)p_1(t) + (1-q)p_2(t) \qquad \text{and} \qquad p_N(t+1) = qp_{N-1}(t) + qp_N(t). \tag{5.6}$$

Let us introduce the vector $\vec{p}(t)$ in $\mathbf{R}^N$ whose jth component is $p_j(t)$. Then (5.5) and (5.6) assert that $\vec{p}(t+1)$ is obtained from $\vec{p}(t)$ by the action of a linear transformation: $\vec{p}(t+1) = A\vec{p}(t)$, where A is the $N \times N$ matrix whose only non-zero entries are:

$$A_{11} = 1 - q, A_{j\,j-1} = q \quad \text{for } 2 \leq j \leq N, \quad A_{j\,j+1} = 1 - q \quad \text{for } 1 \leq j \leq N-1, \quad \text{and} \quad A_{NN} = q. \tag{5.7}$$

For example, in the case that $N = 4$, the matrix A is

$$A = \begin{bmatrix} 1-q & 1-q & 0 & 0 \\ q & 0 & 1-q & 0 \\ 0 & q & 0 & 1-q \\ 0 & 0 & q & q \end{bmatrix}. \tag{5.8}$$

In any case, we can iterate the equation $\vec{p}(t+1) = A\vec{p}(t)$ to find that

$$\vec{p}(t) = A \cdots A\vec{p}(0) \tag{5.9}$$

where $A \cdots A$ signifies t copies of A multiplied one after the other. By the way, a common shorthand for some n copies of any given matrix, A, successively multiplying one after the other is A^n.

5.3 Growth of nerves in a developing embryo

Here is a very topical question in the study of development: Nerve cells connect muscles and organs to the brain. When an organism develops, how do its nerves know where to connect? A given nerve cell stretches out extremely long and thin 'appendages' called axons. Any given axon may end abutting a muscle cell, or another nerve cell in a chain of nerves that ends in the brain. How do the axons 'know' where they are supposed to attach? A simple model proposes that the tip of the growing axon in an embryo is guided by chemical gradients in much the same fashion as the bacteria in the previous discussion is guided to the sugar.

5.4 Enzyme dynamics

An enzyme is a protein molecule that facilitates a chemical reaction that would take a long time to occur otherwise. Most biological reactions are facilitated by enzymes. One typical mode of operation is for the enzyme to facilitate a reaction between molecules of type α and molecules of type β to produce a new molecule of type γ. It can do this if it simultaneously attracts the α and β molecules. In doing so, the enzyme holds the two sorts near each other for enough time that they can bind together to form the γ molecule. If this γ molecule does not have a strong attraction to the enzyme, it breaks away and so frees the enzyme to attract another α and another β to make more γ.

Now, it is often the case that the enzyme only works efficiently if it is folded in the right conformation. Folded in the wrong way, the parts of the molecule that attract either α or β molecules might find themselves covered by parts that are indifferent to α or β. Due to random collisions and other effects, any given molecule, and thus our enzyme, will change its fold configuration. Suppose that there is some probability, q, that the enzyme is folded correctly to attract α and β in any given unit of time, and thus probability $(1-q)$ that it is not. Suppose in addition, that when folded correctly, the enzyme makes 2 units of γ per unit of time. Meanwhile, suppose that the freely diffusing γ falls apart or is degraded by other enzymes at a rate of 1 unit per unit time.

For any integer j, let $p_j(t)$ denote the probability of finding level j of γ after some t units of time. Then the scenario just outlined finds

$$p_j(t+1) = qp_{j-1}(t) + (1-q)p_{j+1}(t) \qquad \text{when } j \geq 1 \text{ and} \qquad p_0(t+1) = (1-q)(p_1(t) + p_0(t)) \tag{5.10}$$

as long as α and β are well supplied. This last equation is another version of the one that is depicted in (5.2).

5.5 Exercises:

Exercises 1–3 concern the example above where the position of the bacteria at any given time t is labeled by an integer, $x(t)$, in the set $\{1, \ldots, N\}$. Don't assume that we know where the bacteria is at $t = 0$.

1. Suppose T is a positive integer.

   (a) What is the sample space for the collection of positions, $\{x(0), x(1), \ldots, x(T)\}$, of the bacteria at the times $t = 0, 1, \ldots, T$.

   (b) Fix some $j \in \{1, \ldots, N\}$ and $t \in \{0, 1, \ldots, T-1\}$. Let C denote the event that $x(t+1) = j$. For each $i \in \{1, \ldots, N\}$, let B_k denote the event that $x(t) = k$. Explain why the equation in (5.5) has the form $\mathrm{P}(C) = \mathrm{P}(C \,|\, B_1)\,\mathrm{P}(B_1) + \cdots + \mathrm{P}(C \,|\, B_N)\,\mathrm{P}(B_N)$.

2. Recall that events B and C are said to be independent when $\mathrm{P}(B \cap C) = \mathrm{P}(B)\mathrm{P}(C)$.

   (a) Explain why this definition is equivalent to the assertion that the conditional probability of B happening given C is equal to the probability of B happening.

   (b) Suppose that $k \in \{2, ..., N-1\}$ and that $t \geq 2$. Derive a relation between the respective probabilities that $x(t) = k-1$ and $x(t) = k+1$ so as to insure that the event that $x(t+1) = k$ is independent of the event that $x(t) = k-1$.

3. Suppose now that q_+ and q_- are non-negative numbers whose sum is less than 1. Change the bacteria model so that if the bacteria is at position $k \in \{2, ..., N-1\}$ at time t, then it moves to position $k+1$ with probability q_+, to position $k-1$ with probability q_- and stays put with probability $1 - q_+ - q_-$. If the bacteria is at position 1 at time t, then it stays put with probability $1 - q_+$ and moves to position 2 with probability q_+. If the bacteria is at position N at time t, then in stays put with probability $1 - q_-$ and moves to position $N-1$ with probability q_-.

   (a) Write down the analog of (5.5) for this model.

   (b) Write down the analog of the matrix in (5.8) for the $N = 4$ case.

   (c) Write down the probability that the bacteria is at position N at $t = N-1$ given that the bacteria starts at position 1 at time 0.

4. Write down a version of (5.10) that would hold in the case that the enzyme when folded correctly produces some $L = 1$, or $L > 2$ units of γ per unit time. To be more explicit, assume first that when the enzyme is folded correctly, it only makes 1 unit of γ per unit time to see how (5.10) will change. Then see how (5.10) must change if the enzyme makes 3 units per unit time. Finally, consider the case where it makes L units per unit time and write (5.10) in terms of this number L.

CHAPTER

SIX

Random variables

In favorable circumstances, the different outcomes of any given experiment have measurable properties that distinguish them. Of course, if a given outcome has a certain probability, then this is also the case for any associated measurement. The notion of a 'random variable' provides a mathematical framework for studying these induced probabilities on the measurements.

Here is a simple example: Suppose that I have some large number, say 100, coins, all identical and all with probability $\frac{1}{2}$ of landing heads when flipped. I am going to flip them all, and put a dollar in your bank account for each head that appears. You can't see me flipping the coin, but you can go to your bank tomorrow and measure the size of your bank account. You might be interested in knowing the probability of your account increasing by any given amount. The amount in your account is a random variable.

To explore this example a bit, note that the configuration space of possible outcomes from flipping 100 coins consists of sequences that are 100 letters long, each letter being either H or T. There are $2^{100} \approx 1.2 \times 10^{30}$ elements in the configuration space! If s is such a 100 letter sequence, let $f(s)$ denote the number of heads that appear in the sequence s. Thus, $f(s)$ can be any integer from 0 through 100. The assignment $s \to f(s)$ is an example of a random variable. It is a function of sorts on the configuration space. Of interest to you are the probabilities for the appearances of the various possible values of this function f. This is to say that your concern is the probability function on the 101 element set $\{0, \ldots, 100\}$ that gives the probability of your bank account increasing by any given amount.

Here is another example that is slightly less contrived: You are in charge of stocking a lake with trout. You put some large number, say N, of trout in a lake. Due to predation, the odds are 50–50 that any given trout will be eaten after one year. Meanwhile, the trout do not breed for their first year, so you are interested in the number of trout that survive to the second year. This number can be anywhere from 0 to N. What is the probability of finding a given number in this range? Note that in the case that $N = 100$, this question is essentially identical to the question just posed about your bank account.

The model for this is as follows: There is a sample space, S, whose elements consist of sequences of N letters, where each letter is either a D (for dead) or L (for live). Thus, S has 2^N elements. I assign to each element in S a number, this the number of L's in the sequence. This assignment of a number to each element in S is a function on S, and of interest to me are the probabilities for the possible values of this function. Note that these probabilities are not for the elements of S; rather they are for the elements in a different sample space, the set of integers from 0 through N.

6.1 The definition of a random variable

A random variable is no more nor less than a function on the sample space. In this regard, such a function assigns a number to each element in the sample space. One can view the function as giving the results of measurements of some property of the elements of the sample space.

Sometimes, the notion is extended to consider a function from the sample space to another set. For example, suppose that S is the set of possible 3 letter long sequences that are made from the 4 bases, guanine, cytosine, adenine and tyrosine, that appear in DNA molecules. Each such 3 letter sequence either codes for one of the 20 amino acids or

is the 'stop' signal. This is the genetic code. The code is thus a function from a set with 64 elements to one with 21 elements

Most often, random variables take real number values. For example, let S denote the 20 possible amino acids that can occupy the 127th position from the end of a certain enzyme (a type of protein molecule) that helps the cell metabolize the sugar glucose. Now, let f denote the function on S that measures the rate of glucose metabolism in growing bacteria with the given enzyme at the given site. In this case, f associates to each element in a 20 element set a real number.

6.2 Probability for a random variable

Suppose that S is our sample space and P is a probability function on S. If f is a random variable and r a possible value for f, then the probability that f takes value r is by definition, the probability of the subset of S where f is equal to r; thus P(Event that $f = r$). This is number is given by

$$\mathrm{P}(f = r) = \sum_{s \in S : f(s) = r} \mathrm{P}(s). \tag{6.1}$$

In words:

The probability that $f = r$ is the sum of the probabilities of those elements in S where f is equal to r.

For an example of what happens in (6.1), consider the situation that I described at the outset where I flip 100 coins and pay you one dollar for each head that appears. As noted, the sample space S is the 2^{100} element set whose typical element is a sequence, s, of 100 letters, each either H or T. The random variable, f, assigns to any given $s \in S$ the number of heads that appear in the sequence s.

To explore (6.1) in this case, let us agree that the probability of any given element in s is 2^{-100}. This is based on my telling you that each of the 100 coins is fair. It also assumes that the appearance of H or T on any one coin has no bearing on whether H or T appear on any other. (I can say this formally as follows: The event that H appears on any given coin is *independent* from the event that H appears on any other coin.) Thus, no matter what s is, the value P(s) that appears in (6.1) is equal to 2^{-100}. This understood, (6.1) asserts that the probability that f is equal to any given integer $r \in \{0, \dots, 100\}$ is obtained by multiplying 2^{-100} times the number of elements in s that are sequences with precisely r heads.

For example, $\mathrm{P}(f = 0) = 2^{-100}$ because there is just one element in s with no heads at all, this the element $TT \cdots T$. Thus, it is a good bet that you will get at least one dollar. On the other hand, $\mathrm{P}(f = 100)$ is also 2^{-100} since only $HH \cdots H$ has 100 heads. So, it is a good bet that I will lose less than 100 dollars. Consider next the probability for f to equal 1. There are 100 sequences from S with 1 head. These being $HTT \cdots T$, $THT \cdots T$, ..., $T \cdots THT$, $T \cdots TTH$. Thus, $\mathrm{P}(f = 1)$ is $100 2^{-100}$. This is still pretty small, on the order of 10^{-28}. How about the probability for 2 dollars? In this case, there are $\frac{1}{2}100 \cdot 99$ elements in S with two heads. If you buy this count, then $\mathrm{P}(f = 2)$ is $50 \cdot 99 \cdot 2^{-100}$. We shall learn in a subsequent chapter that $\mathrm{P}(f = r)$ is $(100 \times 99 \times \cdots \times (100 - r))/(1 \times 2 \times \cdots \times r) \times 2^{-100}$.

For a second example, take S to be the set of 20 possible amino acids at the 127th position from the end of the glucose metabolizing enzyme. Let f now denote the function from S to the 10 element set that is obtained by measuring to the nearest 10% the fraction of glucose used in one hour by the growing bacteria. Number the elements of S from 1 to 20, and suppose that P assigns the kth amino acid probability $\frac{1}{10}$ if $k \leq 5$, probability $\frac{1}{20}$ if $6 \leq k \leq 10$ and probability $\frac{1}{40}$ if $k > 10$. Meanwhile, suppose that $f(k) = 1 - \frac{k}{10}$ if $k \leq 10$ and $f(k) = 0$ if $k \geq 10$. This understood, it then follows using (6.1) that $\mathrm{P}(f = \frac{n}{10})$ is equal to

$$0 \text{ for } n = 10, \qquad \tfrac{1}{10} \text{ for } 5 \leq n \leq 9, \qquad \tfrac{1}{20} \text{ for } 1 \leq n \leq 4, \qquad \text{and} \qquad \tfrac{3}{10} \text{ for } n = 0. \tag{6.2}$$

By the way, equation (6.1) can be viewed (at least in a formal sense) as a matrix equation in the following way: Introduce a matrix by writing $A_{rs} = 1$ if $f(s) = r$ and $A_{rs} = 0$ otherwise. Then, $\mathrm{P}(f = r) = \sum_{s \in S} A_{rs} \mathrm{P}(s)$ is a

matrix equation. Of course, this is rather silly unless the set S and the possible values for f are both finite. Indeed, if S is finite, say with n elements, number them from 1 to n. Even if f has real number values, one typically makes its range finite by rounding off at some decimal place anyway. This understood, there exists some number, N, of possible values for f. Label the latter by the integers between 1 and N. Using these numberings of S and the values of f, the matrix A can be thought of as a matrix with n columns and N rows.

6.3 A probability function on the possible values of f

Return now to the abstract situation where S is a sample space and f is a function on S. As it turns out, the assignment $r \to \mathrm{P}(f = r)$ of a non-negative number to each of the possible values for f defines a probability function on the set of all possible values for f. Let us call this new sample space S_f, and the new probability function $\mathrm{P}_f(r)$. Thus, if $r \in S_f$, then $\mathrm{P}_f(r)$ is given by the sum on the right hand side of (6.1).

To verify that it is a probability function, observe that it is never negative by the nature of its definition. Also, summing the values of P_f over all elements in S_f gives 1. Indeed, using (6.1), the latter sum can be seen as the sum of the values of P over all elements of S.

The example in (6.2) illustrates this idea of associating a new sample space and probability function to a random variable. In the example from (6.2), the new sample space is the 11 element set of fractions of the form $\frac{n}{10}$ where $n \in \{0, \dots, 10\}$. The function P_f is that given in (6.2). You can verify on your own that $\sum_{0 \le n \le 10} \mathrm{P}(f = n) = 1$.

What follows is a less abstract example. There is a gambling game known as 'craps', that is played with two standard, six-sided dice. The dice are rolled, and the two numbers that appear are summed. When played in a casino, the one who roles the dice must pay a certain amount to play. The casino pays the player an amount that depends on the sum of the two numbers that appear. Thus, of interest are the probabilities for the various possible sums of two numbers, each chosen at random from the set $\{1, 2, \dots, 6\}$.

To analyze the probabilities for the various sums in the language of random variables, note first that the possible outcomes for rolling two fair dice is a set with 36 elements, these pairs of the form (a, b) with a and b from the set $\{1, 2, \dots, 6\}$. Let S denote this set of pairs of integers. If the diced are fair and if the roll of one is independent of that of the other, then I model this situation with the probability function on S that assigns probability $\frac{1}{36}$ to each pair from S. The sum of the two elements in a pair defines a function on S, this denoted in what follows by f. Thus, $f(a, b) = a + b$. By definition, this sum function is a random variable.

Of interest to casinos and to those who play craps are the probabilities for the various values of f. In this regard, note that f can be any integer from 2 through 12. To use (6.1) for computing the probability that $f = r$, note that this case is such that each $\mathrm{P}(s)$ that appears on the right hand side of (6.1) is equal to $\frac{1}{36}$. This understood, it then follows that $\mathrm{P}(f = r)$ is obtained by multiplying $\frac{1}{36}$ times the number of pairs in S whose components sum to r. Here is what results:

$$\begin{aligned}
\mathrm{P}(f = 2) = \mathrm{P}(f = 12) &= \tfrac{1}{36} \\
\mathrm{P}(f = 3) = \mathrm{P}(f = 11) &= \tfrac{2}{36} \\
\mathrm{P}(f = 4) = \mathrm{P}(f = 10) &= \tfrac{3}{36} \\
\mathrm{P}(f = 5) = \mathrm{P}(f = 9) &= \tfrac{4}{36} \\
\mathrm{P}(f = 6) = \mathrm{P}(f = 8) &= \tfrac{5}{36} \\
\mathrm{P}(f = 7) &= \tfrac{6}{36}.
\end{aligned}$$

These values define the probability function, P_f, on the set $S_f = \{2, \dots, 12\}$.

6.4 Mean and standard distribution for a random variable

Statisticians are partial to using a one or two numbers to summarize what might be a complicated story. The mean and standard deviation of a random variable are very commonly employed for this purpose. To some extent, the mean

of a random variable is the best guess for its value. However, the mean speaks nothing of the expected variation. The mean and standard deviation together give both an idea as to the expected value of the variable, and also some idea of the spread of the values of the variable about the expected value.

What follows are the formal definitions. To this end, suppose that f is a random variable on a sample space S, in this case just a function that assigns a number to each element in S. The *mean* of f is the 'average' of these assigned numbers, but with the notion of average defined here using the probability function. The mean is typically denoted as μ; here is its formula:

$$\mu = \sum_{s \in S} f(s)\mathrm{P}(s) \tag{6.3}$$

A related notion is that of the *standard deviation* of the random variable f. This is a measure of the extent to which f differs from its mean. The standard deviation is often denoted by the Greek letter σ and it is defined so that its square is the mean of $(f - \mu)$. To be explicit,

$$\sigma^2 = \sum_{s \in S} (f(s) - \mu)^2 \mathrm{P}(s). \tag{6.4}$$

Thus, the standard deviation is larger when f differs from its mean to a greater extent. The standard deviation is zero only in the case that f is the constant function. By the way, the right hand side of (6.4) can be written using (6.3) as

$$\sigma^2 = \sum_{s \in S} f(s)^2 \mathrm{P}(s) - \mu^2. \tag{6.5}$$

See if you can derive one version from the other.

It is a common mistake to write the mean as $\frac{1}{N}\sum_{s \in S} f(s)$ where N here denotes the number of elements in S. This last expression is correct if each element element in S has the same probability. Thus, when $\mathrm{P}(s) = \frac{1}{N}$ for all s. In general, the formula in (6.1) must be used. To see why, consider the following situation: I have a six-sided die, but one where the probability of the number 1 appearing is 1 and the probability of any other number appearing is zero. Thus, my sample space is the set $\{1, 2, \ldots, 6\}$, $\mathrm{P}(1) = 1$, and $\mathrm{P}(s) = 0$ for $s > 1$. Let's take the random variable, f, to be the function $f(s) = s$. In this case, (6.2) gives $\mu = 1$ as one might expect. Meanwhile, $\frac{1}{N}\sum_{s \in S} f(s)$ is equal to 3.5.

To see the utility of the standard deviation, consider the case where I have slightly less pathological six sided die, this where $\mathrm{P}(1) = \frac{1}{2}$, $\mathrm{P}(6) = \frac{1}{2}$ and $\mathrm{P}(s) = 0$ if s is neither 1 nor 6. As before, I take f so that $f(s) = s$. In this case the expression on the right side of (6.3) is equal to 3.5, this identical to what I would get for a fair die, one where the probability is for any given number appearing. On the other hand, the standard deviation for the case of a fair die is $\sqrt{\frac{35}{12}} \approx 1.7$, while the standard deviation for the case of this pathological die is 2.5.

To see how these definitions play out in another example, consider the case of the game of craps. Recall that the sample space in this example is the set, S, of pairs of the form (a, b) where a and b can be any integer from 1 through 6. In this case, each element in S has the same probability, this being $\frac{1}{36}$. The random variable here is the function, f, that assigns $a + b$ to the pair (a, b). The mean of f in this case is 7. The standard deviation is $\sqrt{\frac{35}{6}}$.

For a third example, consider again the case that is relevant to (6.2). The sum for the mean in this case is

$$1 \cdot 0 + \frac{9}{10} \cdot \frac{1}{10} + \frac{8}{10} \cdot \frac{1}{10} + \frac{7}{10} \cdot \frac{1}{10} + \frac{6}{10} \cdot \frac{1}{10} + \frac{5}{10} \cdot \frac{1}{10} + \frac{4}{10} \cdot \frac{1}{20} + \frac{3}{10} \cdot \frac{1}{20} + \frac{2}{10} \cdot \frac{1}{20} + \frac{1}{10} \cdot \frac{1}{20} + 0 \cdot \frac{3}{10}$$

which equals $\frac{2}{5}$. Thus, $\mu = \frac{2}{5}$. The standard deviation in this example is the number whose square is the sum

$$\frac{9}{25} \cdot 0 + \frac{25}{100} \cdot \frac{1}{10} + \frac{16}{100} \cdot \frac{1}{10} + \frac{9}{100} \cdot \frac{1}{10} + \frac{4}{100} \cdot \frac{1}{10} + \frac{1}{100} \cdot \frac{1}{10} + 0 \cdot \frac{1}{20} + \frac{1}{100} \cdot \frac{1}{20} + \frac{4}{100} \cdot \frac{1}{20} + \frac{9}{100} \cdot \frac{1}{20} + \frac{4}{25} \cdot \frac{3}{10}$$

which equals $\frac{11}{100}$. Thus, $\sigma = \frac{\sqrt{11}}{10} \approx 0.33$.

6.5 Random variables as proxies

Of ultimate interest are the probabilities for the points in S, but it is often only possible to directly measure the probabilities for some random variable, a given function on S. In this case, a good theoretical understanding of the

measurements and the frequencies of occurrences of the various measured values can be combined so as to make an educated guess for the probabilities of the elements of S.

To give an example, consider again the casino game called 'craps'. Recall that the game is played as follows: The player pays the casino a certain amount of money and then rolls two six sided dice. The two numbers showing are summed and the player's winnings are determined by this sum. This sum is a random variable on the sample space which in this case is the 36 element set of pairs of the form (a, b) where a and b can be any two numbers from 1 through 6. Suppose now that you are watching people play this game from a far off seat in the casino. You are too far away to see the pair (a, b) that arise on any given play, but you hear the casino croupier call out the sum. You make note in a ledger book of these values as you watch. So, in effect, each play of the game is an experiment and your noting the declared sum after each play is a measurement. Your interest here is in the probabilities for the various elements in the sample space, but you aren't privy to the frequency with which they occur. You are only privy to the frequency of occurrence of the values of the random variable. This is to say that after a day's worth of watching the game, the data in your ledger book provides the relative frequencies of the various possible values of the sum $a+b$. To elaborate, your data consists of a sequence of 11 numbers, $\{Y_2, Y_3, \dots, Y_{11}, Y_{12}\}$, where Y_k is the fraction of the time that the sum of the two numbers on the dice is k. Thus, each Y_k is some number from 0 through 1, and $Y_1 + Y_2 + \cdots + Y_{12} = 1$. For example, if the dice are fair and the face showing on one has no bearing on the face that shows on the other, then I would expect that $(Y_1, \dots, Y_{12})$ should be nearly $(\frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{6}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}, \frac{1}{36})$. If the casino is using loaded dice, then the values of the various Y_k may differ from those just listed.

The question is now: Can these relative frequencies be used to make an educated guess for the probabilities of the various pairs (a, b) that appear? Put starkly, can you use the data $\{Y_1, \dots, Y_{12}\}$ to decide if the dice being used are fair?

Here is how this is typically accomplished in the generic setting where you are doing some experiment to measure a random variable, f, on some sample space S. The experiment is done many times and the frequencies of occurrence of the possible values for f are then taken as a reasonable approximation for the probability function P_f. This is to say that we *declare* $\mathrm{P}(f = r)$ to equal the measured frequency that the value r was obtained for f. These experimentally determined values are substituted for $\mathrm{P}(f = r)$ in (6.1) to turn the latter equation where the right-hand side is known and the various $s \in S$ versions of $\mathrm{P}(s)$ on the left-hand side are desired. In short, the goal is to view (6.1) as a linear equation where the left-hand side is given by our experimental approximation for P_f and the various $s \in S$ versions of $\mathrm{P}(s)$ are the unknowns to be found. Granted that we can solve this equation, then the resulting solutions will give us a reasonable guess as to the true probability function on S.

Return to our hypothetical data from watching the game of craps. The version of (6.1) where $\{Y_2, \dots, Y_{12}\}$ are used on the left-hand side for the values of P_f yields 11 equations for 36 unknowns:

$$
\begin{aligned}
Y_2 &= X(1,1) \\
Y_3 &= X(1,2) + X(2,1) \\
Y_4 &= X(1,3) + X(2,2) + X(3,1) \\
Y_5 &= X(1,4) + X(2,3) + X(3,2) + X(4,1) \\
Y_6 &= X(1,5) + X(2,4) + X(3,3) + X(4,2) + X(5,1) \\
Y_7 &= X(1,6) + X(2,5) + X(3,4) + X(4,3) + X(5,2) + X(6,1) \\
Y_8 &= X(2,6) + X(3,5) + X(4,4) + X(5,3) + X(6,2) \\
Y_9 &= X(3,6) + X(4,5) + X(5,4) + X(6,3) \\
Y_{10} &= X(4,6) + X(5,5) + X(6,4) \\
Y_{11} &= X(5,6) + X(6,5) \\
Y_{12} &= X(6,6)
\end{aligned}
$$

Here $X(a, b)$ is our unknown proxy for the probability function, P, on the sample space of pairs of the form (a, b) where a and b are integers from 1 through 6.

To see how this works in the general case, suppose that S is a sample space and f a random variable on S. Suppose that there is some finite set of possible values for f, these labeled as $\{r_1, \dots, r_N\}$. When $k \in \{1, \dots, N\}$, let y_k denote the frequency that r_k appears as the value for f in our experiments. Label the elements in S as $\{s_1, \dots, s_n\}$.

Now introduce the symbol x_j to denote the unknown but desired $\mathrm{P}(s_j)$. Thus, the subscript j on x can be any integer in the set $\{1, \dots, n\}$. The goal is then to solve for the collection $\{x_j\}_{1\le j\le n}$ by writing (6.1) as the linear equation

$$\begin{aligned} y_1 &= a_{11}x_1 + \cdots + a_{1n}x_n \\ &\vdots \\ y_N &= a_{N1}x_1 + \cdots + a_{Nn}x_n, \end{aligned} \tag{6.6}$$

where $a_{kj} = 1$ if $f(s_j) = r_k$ and $a_{kj} = 0$ otherwise. Note that this whole strategy is predicated on two things: First, that the sample space is known. Second, that there is enough of a theoretical understanding to predict apriori the values for the measurement f on each element in S.

To see something of this in action, consider first the example from the game of craps. In this case there are 11 equations for 36 unknowns, so there are infinitely many possible choices for the collection $\{X(a,b)\}$ for any given set $\{Y_2, \dots, Y_{12}\}$. Even so, the equations determine $X(1,1)$ and $X(6,6)$. If we expect that $X(a,b) = X(b,a)$, then there are 21 unknowns and the equations now determine $X(1,2)$ and $X(5,6)$ also.

Consider next the example from (6.2). For the sake of argument, suppose that the measured frequency of $\mathrm{P}(f = \frac{n}{10})$ are exactly those given in (6.2). Label the possible values of f using $r_1 = 0$, $r_2 = \frac{1}{10}, \cdots, r_{11} = 1$. This done, the relevant version of (6.6)) is the following linear equation:

$$\begin{aligned} \tfrac{3}{10} &= x_{10} + \cdots + x_{20} \\ \tfrac{1}{20} &= x_9 \\ \tfrac{1}{20} &= x_8 \\ \tfrac{1}{20} &= x_7 \\ \tfrac{1}{20} &= x_6 \\ \tfrac{1}{10} &= x_5 \\ \tfrac{1}{10} &= x_4 \\ \tfrac{1}{10} &= x_3 \\ \tfrac{1}{10} &= x_2 \\ \tfrac{1}{10} &= x_1 \\ 0 &= 0 \end{aligned}$$

As you can see, this determines $x_j = \mathrm{P}(s_j)$ for $j \le 9$, but there are infinitely many ways to assign the remaining probabilities.

We shall see in subsequent lessons how to choose a 'best possible' solution of a linear equation that has more unknowns than knowns.

6.6 A biology example

Here is some background: It is typical that a given gene along a DNA molecule is read by a cell for its information only if certain nearby sites along the DNA are bound to certain specific protein molecules. These nearby sites are called 'promoter' regions (there are also 'repressor' regions) and the proteins that are involved are called 'promoters'. The promoter regions are not genes per se, rather they are regions of the DNA molecule that attract proteins. The effect of these promoter regions is to allow for switching behavior: The gene is 'turned on' when the corresponding promoter is present and the gene is 'turned off' when the promoter is absent. For example, when you go for a walk, your leg muscle cells do work and need to metabolize glucose to supply the energy. Thus, some genes need to be turned on to make the required proteins that facilitate this metabolism. When you are resting, these proteins are not needed—furthermore, they clutter up the cells. Thus, these genes are turned off when you rest. This on/off dichotomy is controlled by the relative concentrations of promoter (and repressor) proteins. A simplified version of the how this comes about is as follows: The nerve impulses to the muscle cell cause a change in the folding of a few particular

proteins on the cell surface. This change starts a chain reaction that ultimately frees up promoter proteins which then bind to the promoter regions of the DNA, thus activating the genes for the glucose metabolizing machinery. The latter then make lots of metabolic products for use while walking.

With this as the background, here is my example: Let S denote the set of positive integers up to some large number N, and let $\mathrm{P}(s)$ denote the probability that a given protein is attached to a given promoting stretch of DNA for the fraction of time $\frac{s}{N}$. We measure the values of a function, f, which is the amount of protein that would be produced by the cell were the promoter operative. Thus, we measure $\mathrm{P}(f = r)$, the frequencies of finding level r of the protein. A model from biochemistry might tell us how the value of f depends on the fraction of time that the promoter protein is attached to the promoter region of the DNA. With the model in hand, we could then write

$$P_f(r) = \sum_s a_{rs}\mathrm{P}(s), \tag{6.7}$$

where a_{rs} is obtained from the theoretical model. Note that our task then is to solve for the collection $\{\mathrm{P}(s)\}$, effectively solving a version of the linear equation in (6.7).

6.7 Independent random variables and correlation matrices

Before discussing these notions in the general context, I briefly describe an example from epidemiology. Let S denote the sample space that is the set of adults in the United States. Let f denote the function on S that assigns to each person that person's blood pressure. Let g denote the function on S that assigns to each person the level of salt taken on an average day. Question: Does knowledge of g say anything about the frequency of occurrence of the various values of f? Said differently, can the values of f and g vary independently, or are there correlations between the values of these two functions?

To explore these questions in a simple example, suppose that S is the sample space that consists of triples (a, b, c) where each entry is an integer from 1 through 3. Thus, S has 27 elements. Suppose, in addition, that S has the probability function, P, where all elements have the same probability. Thus, P assigns $\frac{1}{27}$ to each element in S.

Introduce two random variables on S. The first, f, is given by $f(a, b, c) = a + b$. The second, g, is given by $g(a, b, c) = b + c$. The values of both f and g can be any number from the 5 element set $\{2, 3, 4, 5, 6\}$. Here is a question: Does knowledge of the values of g give any information about those of f? This question can be reposed as follows: If r and ρ are integers from $\{2, 3, 4, 5, 6\}$, is the event that $f = r$ *independent* from the event that $g = \rho$? The random variables f and g are called *independent random variables* when such is the case for all r and all ρ.

Remember that sets A and B from a sample space S are called independent when $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$. Thus, to answer the question posed above, we should look at

$$\mathrm{P}(\text{Event that } f = r \text{ and } g = \rho) - \mathrm{P}(\text{Event that } f = r) \cdot \mathrm{P}(\text{Event that } g = \rho).$$

If this is zero for all values of r and ρ, then f and g are independent. Consider first the case where $r = 2$ and $\rho = 2$. Then the set where $f = 2$ and $g = 2$ consists of just the triple $(1, 1, 1)$ and so it has probability $\frac{1}{27}$. Meanwhile, the set where $f = 2$ consists of 3 triples, $(1, 1, 1)$, $(1, 1, 2)$ and $(1, 1, 3)$. Thus, it has probability $\frac{3}{27}$. Likewise, the set where $g = 2$ has probability $\frac{3}{27}$ since it consists of the three elements $(1, 1, 1)$, $(2, 1, 1)$ and $(3, 1, 1)$. The quantity $\mathrm{P}(f = 2 \text{ and } g = 2) - \mathrm{P}(f = 2)\mathrm{P}(g = 2)$ is equal to $\frac{2}{81}$. Since this is not zero, the random variables f and g are not independent.

For a second example, take S, P and f as just described, but now consider the case where g is the random variable that assigns $c - b$ to the triple (a, b, c). In this case, g can take any integer in the range from -2 to 2. Consider $f = 2$ and $g = 0$. In this case, the set where both $f = 2$ and $g = 0$ consists of $(1, 1, 1)$ and so has probability $\frac{1}{27}$. The set where $g = 0$ consists of 9 elements, these of the form (a, b, b) where a and b can be any integers from 1 through 3. Thus, $\mathrm{P}(g = 0) = \frac{9}{27} = \frac{1}{3}$. Since $\frac{1}{27} = \frac{1}{9} \cdot \frac{1}{3}$, the event that $a + b = 2$ is independent from the event that $c - b = 0$. Of course, to see if f and g are independent random variables, we need to consider other values for f and for g.

To proceed with this task, consider the case where $f = 2$ and $g = -2$. The event that $g = -2$ consists of three elements, these of the form $(a, 3, 1)$ where a can be any integer from 1 to 3. As a consequence, the event that $g = -2$

has probability $\frac{3}{27} = \frac{1}{9}$. Thus, the product of the probability of the event that $f = 2$ with that of the event that $g = -2$ is $\frac{1}{81}$. On the other hand, the event that both $f = 2$ and $g = -2$ is the empty set. As the empty set has probability 0, so f and g are not independent.

An example where f and a function g are independent is that where $g(a, b, c) = c$.

Turn now to the general case where S is any given sample space, P is a probability measure on S, and both f and g are random variables on S. We are again interested in whether knowledge of g tells us something about f. I can formalize the discussion just held by introducing the *correlation matrix* of the two random variables. The correlation matrix measures the extent to which the event that f has a given value is independent of the event that g has a given value.

To see how this matrix is defined, label the possible values for f as $\{r_1, \dots, r_N\}$ and label those of g as $\{\rho_1, \dots, \rho_M\}$. Here, N need not equal M, and there is no reason for the r's to be the same as the ρ's. Indeed, f can concern apples and g oranges: The r's might be the weights of apples, rounded to the nearest gram; and the ρ's might be the acidity of oranges, measured in pH to two decimal places. Or, the values of f might be blood pressure readings rounded to the first decimal place, and those of g the intake (in milligrams) of salt per day.

In any event, for a given f and g, the correlation matrix is the $N \times M$ matrix C with coefficients $(C_{k,j})_{1 \le k \le N, 1 \le j \le M}$ where

$$C_{kj} = \mathrm{P}(f = r_k \text{ and } g = \rho_j) - \mathrm{P}(f = r_k)\mathrm{P}(g = \rho_j). \tag{6.8}$$

In this last equation, $\mathrm{P}(f = r_k \text{ and } g = \rho_j)$ is the probability of the event that f has value r_k and g has value ρ_j; it is the sum of the values of P on the elements $s \in S$ where $f(s) = r_k$ and $g(s) = \rho_j$. Thus, $C_{kj} = 0$ if and only if the event that $f = r_k$ is independent from the event that $g = \rho_j$. If all entries are zero, the random variables f and g are said to be *independent random variables*. This means that the probabilities for the values of f have no bearing on those for g and vice-versa.

Consider this in our toy model from (6.2): Suppose that g measures the number of cell division cycles in six hours from our well-fed bacteria. Suppose, in particular, that the values of g range from 0 to 2, and that $g(k) = 2$ if $k \in \{1, 2\}$, that $g(k) = 1$ if $3 \le k \le 7$, and that $g(k) = 0$ if $k \ge 7$. In this case, the probability that g has value $\rho \in \{0, 1, 2\}$ is

$$\tfrac{2}{5} \text{ for } \rho = 0, \qquad \tfrac{2}{5} \text{ for } \rho = 1, \qquad \text{and} \qquad \tfrac{1}{5} \text{ for } \rho = 2. \tag{6.9}$$

Label the values of f so that $r_1 = 0$, $r_2 = \frac{1}{10}$, …, $r_{10} = \frac{9}{10}$, $r_{10} = 1$. Meanwhile, label those of g as in the order they appear above, $\rho_1 = 0$, $\rho_2 = 1$ and $\rho_3 = 2$. The correlation matrix in this case is an 11×3 matrix. For example, here are the coefficients in the first row:

$$C_{11} = \tfrac{11}{50}, \qquad C_{12} = -\tfrac{3}{25}, \qquad C_{13} = -\tfrac{3}{25}.$$

To explain, note that the event that $f = 0$ consists of the subset $\{10, \dots, 20\}$ in the set of integers from 1 to 20. This set is a subset of the event that g is zero since the latter set is $\{7, \dots, 20\}$. Thus, $\mathrm{P}(f = 0 \text{ and } g = 0) = \mathrm{P}(f = 0) = \frac{3}{10}$, while there are no events where f is 0 and g is either 1 or 2.

By the way, this example illustrates something of the contents of the correlation matrix: If $C_{kj} > 0$, then the outcome $f = r_k$ is relatively likely to occur when $g = \rho_j$. On the other hand, if $C_{kj} < 0$, then the outcome $f = r_k$ is unlikely to occur when $g = \rho_j$. Indeed, in the most extreme case, the function f is never r_k when g is ρ_j and so

$$C_{kj} = -\mathrm{P}(f = r_k)\mathrm{P}(g = \rho_j).$$

As an addendum to this discussion about correlation matrices, I say again that statisticians are want to use a single number to summarize behavior. In the case of correlations, they favor what is known as the *correlation coefficient*. The latter, $c(f, g)$, is obtained from the correlation matrix and is defined as follows:

$$c(f, g) = \frac{1}{\sigma(f)\sigma(g)} \sum_{k,j} (r_k - \mu(f))(\rho_j - \mu(g))C_{kj}. \tag{6.10}$$

Here, $\mu(f)$ and $\sigma(f)$ are the respective mean and standard deviation of f, while $\mu(g)$ and $\sigma(g)$ are their counterparts for g.

6.8 Correlations and proteomics

Let's return to the story about promoters for genes. In principle, a given protein might serve as a promoting protein for one or more genes, or it might serve as a promoter for some genes and a repressor for others. Indeed, one way a protein can switch off a gene is to bind to the DNA in such a way as to cause all or some key part of the gene coding stretch to be covered.

Anyway, suppose that f measures the level of protein #1 and g measures that of protein #2. The correlation matrix for the pair f and g is a measure of the extent to which the levels of f and g tend to track each other. If the coefficients of the matrix are positive, then f and g are typically both high and both low simultaneously. If the coefficients are negative, then the level of one tends to be high when the level of the other is low. This said, note that a reasonable approximation to the correlation matrix can be inferred from experimental data: One need only measure simultaneous levels of f and g for a cell, along with the frequencies that the various pairs of levels are observed.

By the way, the search for correlations in protein levels is a major preoccupation of cellular biologists these days. They use rectangular 'chips' that are covered with literally thousands of beads in a regular array, each coated with a different sort of molecule, each sort of molecule designed to bind to a particular protein, and each fluorescing under ultraviolet light when the protein is bound. Crudely said, the contents of a given kind of cell at some known stage in its life cycle are then washed over the chip and the ultraviolet light is turned on. The pattern and intensity of the spots that light up signal the presence and levels in the cell of the various proteins. Pairs of spots that tend to light up under the same conditions signal pairs of proteins whose levels in the cell are positively correlated.

6.9 Exercises:

1. A number from the three element set $\{-1, 0, 1\}$ is selected at random; thus each of -1, 0 or 1 has probability $\frac{1}{3}$ of appearing. This operation is repeated twice and so generates an ordered set (i_1, i_2) where i_1 can be any one of -1, 0 or 1, and likewise i_2. Assume that these two selections are done independently so that the event that i_2 has any given value is independent from the value of i_1.

   (a) Write down the sample space that corresponds to the possible pairs $\{i_1, i_2\}$.

   (b) Let f denote the random variable that assigns $i_1 + i_2$ to any given (i_1, i_2) in the sample space. Write down the probabilities $\mathrm{P}(f = r)$ for the various possible values of r.

   (c) Compute the mean and standard deviation of f.

   (d) Let g denote the random variable that assigns $|i_1| + |i_2|$ to any given (i_1, i_2). Write down the probabilities $\mathrm{P}(g = \rho)$ for the various possible values of ρ.

   (e) Compute the mean and standard deviation of g.

   (f) Compute the correlation matrix for the pair (f, g).

   (g) Which pairs of (r, ρ) with r a possible value for f and ρ one for g are such that the event $f = r$ is independent from the event $g = \rho$?

2. Let S denote the same sample space that you used in Problem 1, and let P denote some hypothetical probability function on S. Label the elements in S by consecutive integers starting from 1, and also label the possible values for f by consecutive integers starting from 1. Let x_j denote $\mathrm{P}(s_j)$ where s_j is the jth element of the sample space. Meanwhile, let y_k denote the $\mathrm{P}(f = r_k)$ where r_k is the kth possible value for f. Write down the linear equation that relates $\{y_k\}$ to $\{x_j\}$.

3. Repeat Problem 1b through 1e in the case that the probability of selecting either -1 or 1 in any given selection is $\frac{1}{4}$ and that of selecting 0 is $\frac{1}{2}$.

4. Suppose that N is a positive integer, and N selections are made from the set $\{-1, 0, 1\}$. Assume that these are done independently so that the probability of any one number arising on the kth selection is independent of any given number arising on any other selection. Suppose, in addition, that the probability of any given number arising on any given selection is $\frac{1}{3}$.

(a) How many elements are in the sample space for this problem?

(b) What is the probability of any given element?

(c) Let f denote the random variable that assigns to any given $(i_1, \ldots, i_N)$ their sum, thus: $f = i_1 + \cdots + i_N$. What are P($f = N$) and P($f = N - 1$)?

5. There are 32 teams in the National Football league, and each team plays 16 games.

   These teams are divided into 8 divisions, and any given team plays each of the other three teams in its division twice during the season. Now, give each team 1 point for a win and 0 for a loss and $\frac{1}{2}$ for a tie. Adding up these numbers for each team yields an ordered sequence of numbers, $\vec{n} = (n_1, n_2, \ldots, n_{32})$ where each n_k is non-negative, but no greater than 16. Also, $n_1 + n_2 + \cdots + n_{32} = 256$. Our sample space is the set, $\mathcal{S}$, of such ordered sequences, thus the set of all possible vectors $\vec{n} \in \mathbf{R}^{32}$ of the sort just described. For each integer, $k \in \{0, 1, \ldots, 16\}$, define a function, $f_k : S \to \{0, 1, \ldots, 32\}$, so that $f_k(\vec{n}) =$ number of teams with point total k or $k + \frac{1}{2}$ (remember that a team gets $\frac{1}{2}$ point for a tie). To see what these random variables look like, define a vector, $\vec{f}$ in $\mathbf{R}^{17}$ whose jth component is f_{j-1}. The columns in the table that follows gives the actual vector $\vec{f}$ for the past ten seasons in the NFL[1].

Wins	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	Count	Wins
0		1									1	0
1	1		1						1	1	4	1
2	1	2		1	1	1		1	1		8	2
3	1		1	1	1			1	1	2	8	3
4	2	2	4	2	5	2	4	2		2	25	4
5	3	3	2	2	3	4	6	2	3	3	31	5
6	1	1		3	4	5	3	2	4	2	25	6
7	3	2	7	3		3	3	5	6	3	35	7
8	5	5	4	8	1	4	2	3	2	2	36	8
9	5	5	2	4	4	4	1	6	2	4	37	9
10	3	2	5	3	3	3	6	4	3	4	36	10
11	3	4	2		6	1	1	3	3	4	27	11
12	1	4		2	1	2	4	3	2	3	22	12
13	2	1	3	2	2	1	1		2	1	15	13
14	1			1	1	1	1		1		6	14
15						1					1	15
16			1								1	16

Table 6.1: NFL teams counted by number of wins, by season

(a) Should you expect f_{15} and f_{16} to be independent random variables? Explain your answer.

(b) Should you expect f_8 and any f_k to be independent random variables? Explain your answer.

(c) From the table above, what are the mean and standard deviation of the data representing f_8 for the last ten years? See equations (1.1) and (1.3).

[1]Note: The Houston Texans did not start play until the 2002 season, so for the first two seasons in our table there are only 31 teams. Also, there were two tie games in this timespan: Eagles-Bengals in 2008 and Falcons-Steelers in 2002.

CHAPTER

SEVEN

The statistical inverse problem

Here is a basic issue in statistic: Experiments are often done to distinguish various hypothesis about the workings of a given system. This is to say that you have various models that are meant to predict the behavior of the system, and you do experiments to see which model best predicts the experimental outcomes. The ultimate goal is to use the observed data to find the correct model.

A more realistic approach is to use the observed data to generate a probability function on the set of models that are under consideration. The assigned probability should give the odds for the correctness of a given model. There are many ways to do this. Usually, any two of the resulting probability functions assign different probabilities to the models. For this reason, if for no other, some serious thought must be given as to which (if any) to use. In any event, I describe some of these methods in what follows.

To indicate the flavor of what is to follow, consider first an example. Here is some background: Suppose that a given stretch of DNA on some chromosome has N sites. If I look at this stretch in different people, I will, in general, not get the same sequence of DNA. There are typically some number of sites where the bases differ. Granted this basic biology, I then sample this site in a large number of people. If I find that one particular DNA sequence occurs more often then any other, I deem it to be the *consensus* sequence. Suppose for simplicity that there is, in fact, such a consensus sequence. My data gives me more than just a consensus sequence; I also have numbers $\{p_0, p_1, p_2, \dots, p_N\}$ where any given p_k is the fraction of people whose DNA differs at k sites from the consensus sequence. For example, p_0 is the fraction of people whose DNA for this N site stretch is identical to the consensus sequence. For a second example, p_{52} is the fraction of people whose DNA for this stretch differs at 52 sites from the consensus sequence.

I wish to understand how these number $\{p_0, p_1, \dots, p_N\}$ arise. If I believe that this particular stretch of DNA is 'junk DNA', and so has no evolutionary function, I might propose the following: There is some probability for a substituted DNA base to appear at any given site of this N site long stretch of DNA. The simplest postulate I can make along these lines is that the probability of a substitution is independent of the particular site. I am interested in finding out something about this probability of a substitution. To simplify matters, I look for probabilities of the form $\frac{m}{100}$ where $m \in \{0, 1, 2, \dots, 100\}$. So, my basic question is this:

> *Given the data $\{p_0, \dots, p_N\}$, what is the probability that a given $\theta = \frac{m}{100}$ is the true probability for a single site substitution in the given stretch of DNA?*

The interesting thing here is that if I know this probability, thus the number θ from the set $\{0, \frac{1}{100}, \frac{2}{100}, \dots, \frac{99}{100}, 1\}$, then I can predict the sequence $\{p_0, p_1, \dots, p_N\}$. We will give a general formula for this in a later chapter. However, for small values of N, you needn't know the general formula as we can work things out directly.

Consider first $N = 1$. Then there is just p_0 and p_1. The probability, given θ for one site change is θ (by definition), thus the probability for no site changes is $1 - \theta$. Hence, given θ, I would predict

$$\text{Probability of } 0 \text{ site changes (given } \theta) \text{ is } 1 - \theta.$$
$$\text{Probability of } 1 \text{ site change (given } \theta) \text{ is } \theta.$$

I can then compare these numbers with what I observed, p_0 and p_1. Doing so, I see that there is an obvious choice for θ, that with $p_0 = 1 - \theta$ and $p_1 = \theta$. Of course, if I restrict θ to fractions of 100, then the I should choose the closest

such fraction to p_1. Note, by the way, that $p_0 + p_1 = 1$ by virtue of their definition as fractions, so the two conditions on θ amount to one and the same thing.

Consider next the case $N = 2$. Given θ, I would then predict

$$\begin{aligned} &\text{Probability of 0 site changes (given } \theta) \text{ is } (1-\theta)^2. \\ &\text{Probability of 1 site change (given } \theta) \text{ is } 2\theta(1-\theta). \\ &\text{Probability of 2 site changes (given } \theta) \text{ is } \theta^2. \end{aligned} \tag{7.1}$$

To see why this is, think of flipping two coins, with heads = site change and tails = no site change; and with probability equal to θ for any given coin to have heads. The probability of the first coin/site having heads is θ and of it having tails is $(1-\theta)$. Likewise for the second coin/site. So, they both end up tails with probability $(1-\theta)^2$. Likewise, they both end up heads with probability θ^2. Meanwhile, there are two ways for one head to appear, either on the first coin/site, or on the second. Each such event has probability $\theta(1-\theta)$, so the probability of one coin/site to have heads is twice this.

This same sort of comparison with coin flipping leads to the following for the $N = 3$ case:

$$\begin{aligned} &\text{Probability of 0 site changes (given } \theta) \text{ is } (1-\theta)^3. \\ &\text{Probability of 1 site change (given } \theta) \text{ is } 3\theta(1-\theta)^2. \\ &\text{Probability of 2 site changes (given } \theta) \text{ is } 3\theta^2(1-\theta). \\ &\text{Probability of 3 site changes (given } \theta) \text{ is } \theta^3. \end{aligned} \tag{7.2}$$

However, note that in the case $N = 2$ the data $\{p_0, p_1, p_2\}$ need not lead to any obvious candidate for θ. This is because the equations

$$p_0 = (1-\theta)^2 \qquad \text{and} \qquad p_1 = 2\theta(1-\theta) \qquad \text{and} \qquad p_2 = \theta^2$$

need not have a solution. Even using the fact that $p_0 + p_1 + p_2 = 1$, there are still two conditions for the one unknown, θ. Consider, for example, $p_0 = \frac{9}{16}$, $p_1 = \frac{3}{16}$, $p_2 = \frac{1}{4}$. The equations $p_2 = \theta^2$ and $p_0 = (1-\theta)^2$ lead to very different choices for θ. Indeed, the equation $p_2 = \theta^2$ gives $\theta = \frac{1}{2}$ and the equation $p_0 = (1-\theta)^2$ gives $\theta = \frac{1}{4}$! This same problem gets worse as N gets ever larger.

What's to be done? Here is one approach: Use the data $\{p_0, \ldots, p_N\}$ to find a *probability function* on the possible values of θ. This would be a probability function on the sample space $\Theta = \{0, \frac{1}{100}, \frac{2}{100}, \ldots, \frac{99}{100}, 1\}$ that would give the probability for a given to be the true probability of a site substitution given the observed data $\{p_0, \ldots, p_N\}$ *and* given that my model of equal chance of a site substitution at each of the N sites along the DNA strand is correct.

I will use $\mathcal{P}$ to denote a probability function on Θ. This is to say that $\mathcal{P}(\frac{m}{100})$ is the probability that $\mathcal{P}$ assigns to any given θ. For example, $\mathcal{P}(\frac{3}{100})$ is the probability that $\frac{3}{100}$ is the true probability of a mutation occurring in any given cell. Likewise, $\mathcal{P}(\frac{49}{100})$ is the probability that $\frac{49}{100}$ is the true probability of a mutation occurring in any given cell.

Imagine for the moment that I have a probability function, $\mathcal{P}$, on the sample space Θ. I can then use $\mathcal{P}$ to make a theoretical prediction of what the observed data should be. This uses the notion of conditional probability. Here is how it works in the case $N = 2$:

$$\text{Prob}(k \text{ site changes}) = \mathrm{P}\left(k \,|\, 0\right) \mathcal{P}(0) + \mathrm{P}\left(k \,|\, \tfrac{1}{100}\right) \mathcal{P}(\tfrac{1}{100}) + \cdots + \mathrm{P}\left(k \,|\, 1\right) \mathcal{P}(1), \tag{7.3}$$

where the notation uses $\mathrm{P}\left(k \,|\, \theta\right)$ to denote the probability of there being k site changes given that θ is the correct probability. Note that I use the conditional probability notation from Chapter 3, although the context here is slightly different. Anyway, grant me this small abuse of my notation and agree to view $\mathrm{P}\left(k \,|\, \theta\right)$ as an honest conditional probability. Note in this regard that it obeys all of the required rules: Each $\mathrm{P}\left(k \,|\, \theta\right)$ is non-negative, and the sum $\mathrm{P}\left(0 \,|\, \theta\right) + \mathrm{P}\left(1 \,|\, \theta\right) + \cdots + \mathrm{P}\left(N \,|\, \theta\right) = 1$ for each θ.

The key point here is that I know what these conditional probabilities $\mathrm{P}\left(k \,|\, \theta\right)$ are. For example, they in the case $N = 2$, they are given in (7.1) and in the case $N = 3$ they are given in (7.3). Thus, the various versions of (7.3) for

$N = 2$ are

$$\begin{aligned}
\text{Prob}(0 \text{ site changes}) &= \sum_{m=0,1,\ldots 100} (1 - \tfrac{m}{100})^2 \mathcal{P}(\tfrac{m}{100}) \\
\text{Prob}(1 \text{ site change}) &= \sum_{m=0,1,\ldots 100} 2\tfrac{m}{100}(1 - \tfrac{m}{100}) \mathcal{P}(\tfrac{m}{100}) \\
\text{Prob}(2 \text{ site changes}) &= \sum_{m=0,1,\ldots 100} (\tfrac{m}{100})^2 \mathcal{P}(\tfrac{m}{100}).
\end{aligned} \tag{7.4}$$

Of course, I don't yet have a probability function $\mathcal{P}$. Indeed, I am looking for one; and I hope to use the experimental data $\{p_0, \ldots, p_N\}$ to find it. This understood, it makes some sense to consider only those versions of $\mathcal{P}$ that give the correct experimental results. This is to say that I consider as reasonable only those probability functions on the sample space that give Prob(k site changes) $= p_k$ for all choices of $k \in \{0, \ldots, N\}$. For example, in the case $N = 2$, these give the conditions

$$\begin{aligned}
p_0 &= \sum_{m=0,1,\ldots 100} (1 - \tfrac{m}{100})^2 \mathcal{P}(\tfrac{m}{100}) \\
p_1 &= \sum_{m=0,1,\ldots 100} 2\tfrac{m}{100}(1 - \tfrac{m}{100}) \mathcal{P}(\tfrac{m}{100}) \\
p_2 &= \sum_{m=0,1,\ldots 100} (\tfrac{m}{100})^2 \mathcal{P}(\tfrac{m}{100}).
\end{aligned} \tag{7.5}$$

To make it even more explicit, suppose again that I measured $p_0 = \frac{3}{4}$, $p_1 = \frac{3}{16}$ and $p_2 = \frac{1}{16}$. Then what is written in (7.5) would read:

$$\begin{aligned}
\tfrac{3}{4} &= \mathcal{P}(0) + (\tfrac{99}{100})^2 \mathcal{P}(\tfrac{1}{100}) + \cdots + (\tfrac{1}{100})^2 \mathcal{P}(\tfrac{99}{100}) \\
\tfrac{3}{16} &= 2(\tfrac{1}{100})\tfrac{99}{100}) \mathcal{P}(\tfrac{1}{100}) + 2(\tfrac{2}{100})(\tfrac{98}{100}) \mathcal{P}(\tfrac{2}{100}) \cdots + 2(\tfrac{99}{100})(\tfrac{1}{100}) \mathcal{P}(\tfrac{1}{100}) \\
\tfrac{1}{16} &= (\tfrac{1}{100})^2 \mathcal{P}(\tfrac{1}{100}) + \cdots + (\tfrac{99}{100})^2 \mathcal{P}(\tfrac{99}{100}) + \mathcal{P}(1).
\end{aligned} \tag{7.6}$$

Notice, by the way, that what is written here constitutes a system of 3 equations for the 101 unknown numbers, $\{\mathcal{P}(0), \mathcal{P}(\frac{1}{100}), \ldots, \mathrm{P}(1)\}$. As we know from our linear algebra, there will be infinitely many solutions. In the general case of a site with length N, the analog of (7.5) has $N + 1$ equations, with these reading

$$\begin{aligned}
p_0 &= \sum_{m=0,1,\ldots 100} \mathrm{P}\left(0 \,|\, \tfrac{m}{100}\right) \mathcal{P}(\tfrac{m}{100}) \\
p_1 &= \sum_{m=0,1,\ldots 100} \mathrm{P}\left(1 \,|\, \tfrac{m}{100}\right) \mathcal{P}(\tfrac{m}{100}) \\
&\vdots \\
p_N &= \sum_{m=0,1,\ldots 100} \mathrm{P}\left(N \,|\, \tfrac{m}{100}\right) \mathcal{P}(\tfrac{m}{100})
\end{aligned} \tag{7.7}$$

Given that I know what P $(k \,|\, \theta)$ is for any given k, this constitutes a system of $N + 1$ equations for the 101 unknowns $\{\mathcal{P}(0), \mathcal{P}(\frac{1}{100}), \ldots, \mathcal{P}(1)\}$. In particular, if $N > 100$, then there will be more equations than unknowns and so there may be *no* solutions! To make this look exactly like a system of linear equation, change the notation so as to denote the conditional probability P $\left(0 \,|\, \frac{m}{100}\right)$ as a_{0m}, P $\left(1 \,|\, \frac{m}{100}\right)$ as a_{01m}, etc. Use x_m for $\mathcal{P}(\frac{m}{100})$. And, use y_0 for p_0, y_1 for p_1, etc. This done, then (7.7) reads

$$\begin{aligned}
y_0 &= a_{00}x_0 + a_{01}x_1 + \cdots \\
y_1 &= a_{10}x_1 + a_{11}x_1 + \cdots \\
&\vdots
\end{aligned} \tag{7.8}$$

In any event, here is a lesson from this:

In general, the equations in (7.5) *for the case* $N = 2$*, or their analogs in* (7.7) *for* $N > 2$ *do not determine a God-given probability function on the sample space.*

The task of finding a probability function on the sample space from the experimental data, one that comes reasonably close to solving (7.7) *and* makes good biological sense, is an example of what I call the *statistical inverse* problem.

7.1 A general setting

This sort of problem arises in the following general setting: Suppose that I am interested in some system or phenomena, and have a set, Θ, of models of how it works. I shall assume in what follows that Θ is a finite set. In the example outlined above, the set Θ consists of the fractions of the form $\frac{m}{100}$ with $m \in \{0, 1, \dots, 100\}$. I plan to take data to explore the phenomena. Suppose that there is a set, K, of possible experimental outcomes. In the example of the introduction, K, consists of the numbers $0, 1, \dots, N+1$, where the integer k here denotes the number of sites on the given N-site DNA strand that differ from the consensus strand. Even in the general case, if there are but a finite set of possible outcomes, I can label them by consecutive integers starting from zero and this I now do.

As I know a competent model builder, I make sure that each model from Θ determines a probability function on K. This is to say any given model $\theta \in \Theta$ gives a predicted probability for any given outcome in K. I write θ's version of the probability of the kth outcome as $\mathrm{P}(k \,|\, \theta)$. Thus, $\mathrm{P}(k \,|\, \theta)$ is the probability of having the kth outcome were θ the correct model. In the $N = 2$ example from the introduction, $\mathrm{P}(0 \,|\, \theta) = (1-\theta)^2$, $\mathrm{P}(1 \,|\, \theta) = 2\theta(1-\theta)$ and $\mathrm{P}(2 \,|\, \theta) = \theta^2$.

I take lots of data and so generate an experimentally determined probability function, P_{exp}, on K. Here, $\mathrm{P}_{exp}(k)$ is the fraction of the experiments where kth outcome appeared. In the example from the introduction, the number $\mathrm{P}_{exp}(k)$ is what I called p_k.

The statistical inverse problem in this context is as follows: Use the experimentally determined probability function P_{exp} to determine a probability function, $\mathcal{P}$, on the set Θ. You are asked to use what you know, the data derived frequencies P_{exp} and the set of conditional probabilities, the various $\mathrm{P}(k \,|\, \theta)$, to find those model parameters from Θ that are more likely to be correct, and which are less likely to be correct.

Keep in mind that there is, in general, no God-given way to generate a probability function on the set Θ. Even so, there are various popular methods and one or more may well be useful in any given context.

7.2 The Bayesian guess

The Bayesian turns things upside down and reasons is as follows: Suppose that I have a conditional probability that tells me the probability that the model labeled by θ is correct given that the kth outcome in K appears. Suppose I have such a conditional probability for each outcome k and each parameter θ. I use $\mathfrak{P}(\theta \,|\, k)$ to denote this conditional probability. If I had such a collection of conditional probabilities, then I could use the general form of Bayes' theorem in (3.5) to write a probability function on the set Θ. This would be the function whose value on any given θ is

$$\mathcal{P}(\theta) = \mathfrak{P}(\theta \,|\, 0)\, \mathrm{P}_{exp}(0) + \mathfrak{P}(\theta \,|\, 1)\, \mathrm{P}_{exp}(1) + \cdots + \mathfrak{P}(\theta \,|\, N)\, \mathrm{P}_{exp}(N) \tag{7.9}$$

In our $N = 2$ case from the introduction, there are 101 versions of (7.9), one for each fraction $\{\frac{m}{100}\}_{m=0,1,\dots 100}$, and the mth version reads

$$\mathcal{P}(\tfrac{m}{100}) = \mathfrak{P}\left(\tfrac{m}{100} \,|\, 0\right) \mathrm{P}_{exp}(0) + \mathfrak{P}\left(\tfrac{m}{100} \,|\, 1\right) \mathrm{P}_{exp}(1) + \mathfrak{P}\left(\tfrac{m}{100} \,|\, 2\right) \mathrm{P}_{exp}(2) \tag{7.10}$$

Of course, this is wishful thinking without some scheme to obtain the various conditional probabilities $\mathfrak{P}(\theta \,|\, k)$.

The Bayesian makes the following rather ad-hoc proposal: Let's use for $\mathfrak{P}$ the formula

$$\mathfrak{P}_{\text{Bayes}}(\theta \,|\, k) = \tfrac{\mathcal{P}(k \,|\, \theta)}{Z(k)} \tag{7.11}$$

where $Z(k) = \sum_{\theta\in\Theta} \mathcal{P}(k \,|\, \theta)$ is needed so that the $\mathfrak{P}_{\text{Bayes}}(\theta \,|\, k)$ can be interpreted as a conditional probability. The point here is that the probability of *some* θ happening given k is 1, so with any k fixed, the sum of the various $\mathfrak{P}_{\text{Bayes}}(\theta \,|\, k)$ for all of the θ in Θ has to be 1. This is guaranteed with the factor $\frac{1}{Z(k)}$ present, but may not be the case without this factor. Using $\mathfrak{P}_{\text{Bayes}}$ in (7.9) gives what is called the *Bayesian* version of a probability function on Θ:

$$\mathcal{P}_{\text{Bayes}}(\theta) = \mathcal{P}(0 \,|\, \theta)\, \tfrac{1}{Z(0)} \mathrm{P}_{\text{exp}}(0) + \mathcal{P}(1 \,|\, \theta)\, \tfrac{1}{Z(1)} \mathrm{P}_{\text{exp}}(1) + \cdots . \tag{7.12}$$

Note that the presence of Z guarantees that the probabilities defined by the left-hand side of (7.12) sum to 1.

In effect, the Bayesian probability function on Θ is obtained by approximating the unknown conditional probability for θ given outcome k by the known conditional probability of outcome k given θ (times the factor $\frac{1}{Z(k)}$). This may or may not be a good approximation. In certain circumstances it is, and in others, it isn't.

7.3 An example

To see how this works in an example, consider the $N = 2$ case of the introduction. In this case, θ is one of the fractions from the set $\Theta = \{\frac{m}{100}\}_{m=0,1,\ldots 100}$ and (7.1) finds that $\mathcal{P}(0 \,|\, \theta) = (1-\theta)^2$, $\mathcal{P}(1 \,|\, \theta) = 2\theta(1-\theta)$ and $\mathcal{P}(2 \,|\, \theta) = \theta^2$. This understood, then

$$\begin{aligned} Z(0) &= 1 + (\tfrac{99}{100})^2 + (\tfrac{98}{100})^2 + \cdots + (\tfrac{1}{100})^2 \approx 33.835 \\ Z(1) &= 2(\tfrac{1}{100})(\tfrac{99}{100}) + 2(\tfrac{98}{100})(\tfrac{2}{100}) + \cdots + 2(\tfrac{99}{100})(\tfrac{1}{100}) \approx 33.33 \\ Z(2) &= (\tfrac{1}{100})^2 + (\tfrac{2}{100})^2 + \cdots + 1 \approx 33.835. \end{aligned} \tag{7.13}$$

Using this in (7.12) gives the Bayesian probability function whose value on $\frac{m}{100}$ is

$$\mathcal{P}_{\text{Bayes}}(\tfrac{m}{100}) = (1-\tfrac{m}{100})^2 \tfrac{1}{33.835} p_0 + 2(\tfrac{m}{100})(1-\tfrac{m}{100}) \tfrac{1}{33.33} p_1 + (\tfrac{m}{100})^2 \tfrac{1}{33.835} p_2. \tag{7.14}$$

For example, if it turned out that $p_0 = \frac{3}{4}$, $p_1 = \frac{3}{16}$ and $p_2 = \frac{1}{16}$, then the Bayesian guess for a probability function would be

$$\mathcal{P}_{\text{Bayes}}(\tfrac{m}{100}) = (1-\tfrac{m}{100})^2 \tfrac{1}{33.835} \tfrac{3}{4} + 2(\tfrac{m}{100})(1-\tfrac{m}{100}) \tfrac{1}{33.33} \tfrac{3}{16} + (\tfrac{m}{100})^2 \tfrac{1}{33.835} \tfrac{1}{16}.$$

Assume that you calculate $\mathcal{P}_{\text{Bayes}}$ in a given situation. You must then ask: What in the world does it tell me about my set of possible models? It gives a probability function on this set, but so what. There are lots of probability functions on any given set. The fact is that before you use $\mathcal{P}_{\text{Bayes}}$ (or any of other version of $\mathcal{P}$), you need to think long and hard about whether it is really telling you anything useful.

By the way, the formula in (7.14) for $\mathcal{P}_{\text{Bayes}}$ is another disguised version of what we are doing in linear algebra (as is the formula in (7.12)). To unmask this underlying linear algebra, agree to change to the linear algebra book's notation and so use x_0 to denote p_0, use x_1 to denote p_1 and x_2 to denote p_2. Then use y_m to denote $\mathcal{P}(\frac{m}{100})$ and use a_{m1} to denote the number $(1-\frac{m}{100})^2$, use a_{m2} to denote $2(\frac{m}{100})(1-\frac{m}{100})$ and a_{m3} to denote $(\frac{m}{100})^2$. This done, then the 101 versions of (7.14) read

$$\begin{aligned} y_0 &= a_{00}x_0 + a_{01}x_1 + a_{02}x_2 \\ y_1 &= a_{10}\,x_0 + a_{11}\,x_1 + a_{12}x_2 \\ &\cdots \quad \text{etc.} \end{aligned}$$

7.4 Gregor Mendel's peas

What follows gives another example of how the statistical inverse problem can arise. This involves a classic experiment done by Gregor Mendel. For those unfamiliar with the name, here is part of what Wikipedia has to say about Mendel (`http://en.wikipedia.org/wiki/Mendel`):

Gregor Johann Mendel (July 20[1], 1822 – January 6, 1884) was an Augustinian abbot who is often called the "father of modern genetics" for his study of the inheritance of traits in pea plants. Mendel showed that the inheritance of traits follows particular laws, which were later named after him. The significance of Mendel's work was not recognized until the turn of the 20th century. Its rediscovery prompted the foundation of genetics.

What follows describes the experiment: A given pea plant, when self-pollinated, can have either round or angular seeds. It can also have either yellow or green seeds. Mendel took 10 plants, where each plant can have either round or angular seeds, and either yellow or green seeds. Each plant was self-pollinated, and Mendel kept track of the number of seeds that were round and the number that were angular. He also kept track of the number that were yellow and the number that were green. Here are the published results (from `http://www.mendelweb.org/Mendel.html`):

	Experiment 1 (Shape)		**Experiment 2** (Color)	
Plant #	Round	Angular	Yellow	Green
1	45	12	25	11
2	27	8	32	7
3	24	7	14	5
4	19	10	70	27
5	32	11	24	13
6	26	6	20	6
7	88	24	32	13
8	22	10	44	9
9	28	6	50	14
10	25	7	44	18
Totals	336	101	355	123

Using the totals at the bottom finds that round seeds appear 77% of the time. Meanwhile, yellow seeds appear 74% of the time. Let us agree to use $p_r = 0.77$ for the experimentally determined probability for a seed to be round, and $p_y = 0.74$ for the experimentally determined probability for a seed to be yellow.

The conventional wisdom explains the numbers of round and angular seeds by postulating that each plant has two different genes that convey instructions for seed shape. These are denoted here by 'R' (for round) and 'a' (for angular). The parent in each case is hypothesized to be type Ra. The offspring inherits either type RR, Ra or aa, where one of the genes, either R or a, comes from the pollen and the other from the ovule. The hypothesis is that R is dominant so types RR and Ra come from round seeds. Meanwhile, offspring of type aa come from angular seeds.

There is a similar story for the color of the seed. Each plant also has two genes to control seed color. These are denoted here by 'Y' and 'g'. The parent is type Yg, and the offspring can be of type YY, Yg or gg. The Y is postulated to be the dominant of the two genes, so offspring of type YY or Yg have yellow seeds, while those of type gg have round seeds.

The conventional wisdom assigns probability $\frac{1}{2}$ for any given pollen grain or ovule to receive the dominant gene for shape, and likewise probability $\frac{1}{2}$ to receive the dominant gene for color. One should make it a habit to question conventional wisdom! In particular, suppose that we wish to compare the conventional probability model with another model, this giving some probability, θ, for any given pollen grain or ovule to have the dominant allele. To keep things simple, I will again assume that θ can be any number $\{\frac{m}{100}\}_{m\in\{0,1,\ldots,100\}}$. In this case, our set of models is again the 101 element set $\Theta = \{\frac{m}{100}\}_{m\in\{0,1,\ldots 100\}}$.

7.5 Another candidate for $\mathcal{P}(\theta)$: A maximum likelihood candidate.

What follows describes a simplistic thing that can be done with the data so as to determine a probability function on Θ. Look again at Mendel's data. In a sense, 915 different experiments are represented in this table, one for each

seedling with regards to round versus angular seed, and one from each seed with regards to yellow versus green color. To see how to make sense of this, consider a sample space S with 2^{915} elements, where each element is a vector in $\mathbf{R}^{915}$ whose entries are either 0 or 1. Each entry represents one of the 915 possible outcomes for the round/angular and yellow/green traits. The value 0 occurs in any given entry if the recessive trait is present, and the value 1 occurs if the dominant trait is present. Each $\theta \in \{\frac{m}{100}\}_{m\in\{0,\dots 100\}}$ gives us a probability function on this sample space S. For example, the probability that there all 915 cases have the recessive trait is $((1-\theta)^2)^{915}$, and the probability that all have the dominant trait is $(2\theta-\theta^2)^{915}$. Thus, when $\theta = \frac{1}{2}$, the vector in S with all 915 entries equal to 0 has probability $(\frac{1}{4})^{915}$, and the vector in S with all 915 entries equal to 1 has probability $(\frac{3}{4})^{915}$. Here is how to think of this probability function: Imagine flipping a coin 915 times keeping track of the number of heads and the number of tails; here the coin is biased so that a head appears with probability $(2\theta-\theta^2)$ and a tail appears with probability $(1-\theta)^2$. You are asked in a homework problem how to compute the probability of any given element in S.

Consider next the following random variable on S: Suppose that $\hat{a} = (a_1, \dots, a_{915}) \in S$. Set $f(\hat{a})$ to be the sum of the entries. Thus, $f(\hat{a}) = a_1 + a_2 + \cdots + a_{915}$. This is the total number of dominant traits that appear if the vector $\hat{a}$ represents the outcome of the 915 experiments. For example, a vector in S that gives Mendel's data, the value of f is 691.

Now, as remarked above, each θ from Θ determines a probability function on S. As described in the previous chapter, such a probability function on S can be used to obtain a probability function on the set of possible values for the random variable f. This set is the set of integers starting at 0 and ending at 915. Let me use $\mathrm{P}_f\,(k\,|\,\theta)$ to denote the probability function on the set $\{0, 1, \dots, 915\}$ that is determined by θ. Said differently, if $k \in \{0, 1, \dots, 915\}$, then $\mathrm{P}_f\,(k\,|\,\theta)$ is the probability *as determined by* θ of the set of elements in S whose entries sum to k. This can be viewed as a sort of conditional probability for f to equal k given that θ is the parameter used to determine the probabilities on S. This is why I use the notation $\mathrm{P}_f\,(k\,|\,\theta)$. Upcoming chapters will give us the tools to compute $\mathrm{P}_f\,(k\,|\,\theta)$. For now, just imagine this to be some function on the set $\{0, 1, \dots, 915\}$ whose values are non-negative and sum up to 1.

If, for a given θ, the probability $\mathrm{P}_f\,(691\,|\,\theta)$ is small relative to other values of θ, then this particular value of θ should have small probability of being correct. If, for a given θ, the probability $\mathrm{P}_f\,(691\,|\,\theta)$ is large relative to others, then I might expect that this particular θ should have greater probability of being true. These observations motivate the introduction of the following probability function on the set Θ:

$$\mathrm{P}_{\mathrm{ML}}(\theta) = \mathrm{P}_f\,(691\,|\,\theta)\,/Z \qquad \text{where} \qquad Z = \sum_{m=0,1,2,\dots 100} \mathrm{P}_f\left(691\,|\,\tfrac{m}{100}\right). \tag{7.15}$$

Note that the factor of $Z = \mathrm{P}_f\,(691\,|\,0) + \mathrm{P}_f\left(691\,|\,\frac{1}{100}\right) + \cdots + \mathrm{P}_f\,(691\,|\,1)$ is necessary so as to guarantee that the sum of the values of P_{ML} over all 101 elements in Θ is equal to 1. What is written in (7.15) is called by some a *maximum likelihood* probability function on the set Θ.

Of course, this is all pretty abstract if $\mathrm{P}_f\,(691\,|\,\theta)$ can't be computed. As we shall see in an upcoming chapter, it is, in fact, computable:

$$\mathrm{P}_f\,(691\,|\,\theta) = \frac{915\times 914\times\cdots\times 1}{(224\times 223\times\cdots\times 1)\,(691\times 690\times\cdots\times 1)}(2\theta-\theta^2)^{691}((1-\theta)^2)^{224}. \tag{7.16}$$

In any event, this probability function assigns the greatest probability to the model whose version of $\mathrm{P}_f\,(691\,|\,\theta)$ is largest amongst all models under consideration. This is to say that the model that P_{ML} makes most probable is the one that gives the largest probability to the number 691.

To see this approach in a simpler case, suppose that instead of 915 seedlings, I only had 4, and suppose that three of the four exhibited the dominant trait, and one of the four exhibited the recessive trait.

My sample space S for this simple example consists of $2^4 = 16$ elements, these

- $(0,0,0,0)$
- $(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1)$
- $(1,1,0,0), (1,0,1,0), (1,0,0,1), (0,1,1,0), (0,1,0,1), (0,0,1,1)$ (7.17)
- $(1,1,1,0), (1,1,0,1), (1,0,1,1), (0,1,1,1)$
- $(1,1,1,1)$.

To simplify the notation, set $\alpha = (1-\theta)^2$. This is the probability of seeing the recessive trait for the model where $\theta \in \{\frac{m}{100}\}_{m\in\{0,1,\dots,100\}}$ gives the probability that any given pollen grain or ovule has the dominant allele. Thus, $(1-\alpha)$ is the probability of seeing the dominant trait given the value of θ. For example, $\alpha = \frac{1}{4}$ when $\theta = \frac{1}{2}$. In any event with θ given, then the probability function on S that is determined by θ gives the following probabilities:

$$\begin{aligned}
&\mathrm{P} = \alpha^4 \text{ to } (0,0,0,0)\\
&\mathrm{P} = \alpha^3(1-\alpha) \text{ to everything in the second row of (7.17)}\\
&\mathrm{P} = \alpha^2(1-\alpha)^2 \text{ to everything in the third row of (7.17)}\\
&\mathrm{P} = \alpha(1-\alpha)^3 \text{ to everything in the fourth row of (7.17)}\\
&\mathrm{P} = (1-\alpha)^4 \text{ to } (1,1,1,1).
\end{aligned} \tag{7.18}$$

(You are asked to justify (7.18) in a homework problem.)

In this baby example, the possible values of f are $\{0,1,2,3,4\}$, and the set of elements in S that give the value $f = k$ consists of the kth row in (7.17). This being the case, then

$$\mathrm{P}_f(0\,|\,\theta) = \alpha^4, \qquad \mathrm{P}_f(1\,|\,\theta) = 4\alpha^3(1-\alpha), \qquad \mathrm{P}_f(2\,|\,\theta) = 6\alpha^2(1-\alpha)^2,$$
$$\mathrm{P}_f(3\,|\,\theta) = 4\alpha(1-\alpha)^3, \qquad \mathrm{P}_f(2\,|\,\theta) = (1-\alpha)^4.$$

In particular, $\mathrm{P}_f(3\,|\,\theta) = 4\alpha(1-\alpha)^3$. In terms of θ, this says that $\mathrm{P}_f(3\,|\,\theta) = 4(1-\theta)^2(2\theta-\theta^2)^3$.

The analog of (7.10) for this example sets

$$\mathcal{P}_{\mathrm{ML}}(\theta) = \tfrac{1}{20.3175}4(1-\theta)^2(2\theta-\theta^2)^3. \tag{7.19}$$

In this case $Z = 20.3175$. I'll leave it as an exercise for those who remember their one variable calculus to verify that the function $\theta \to \mathcal{P}(\theta)$ has its maximum at $\theta = \frac{1}{2}$.

As with the Bayesian probability function, you must still ask yourself whether the assigned probabilities to the models are useful or not. For example, the probability function in (7.19) finds $\mathcal{P}_{\mathrm{ML}}(\frac{1}{2}) \approx 0.02$. Meanwhile, $\mathcal{P}_{\mathrm{ML}}(\frac{2}{3}) \approx 0.015$. Is it really the case that the probabilities for $\theta = \frac{1}{2}$ and $\theta = \frac{2}{3}$ should be so nearly equal? The fact is that these probabilities are close is due to the fact that the data from 4 seedlings is not nearly enough to distinguish these two models. This is what I meant when I said at the outset that you must think hard about whether any probability function on your set of possible models is worth looking at. The fact that these two probabilities are so close really says nothing about peas and genetics, and everything about the fact that you don't have enough data to discriminate.

This gets to the heart of the matter with regards to using statistics: A good deal of common sense must be used to *interpret* the mathematics.

7.6 What to remember from this chapter

This has been somewhat of a rambling chapter, so I thought to tell you what are the important things to keep in mind:

- *A fundamental purpose of statistics is to provide tools that allow experimental data to assign probabilities to various proposed models for predicting the outcomes of the experiment.*
- *There are many such tools.* (7.20)
- *Depending on the circumstances, some of these tools will be more useful than others. In particular, some techniques can give misleading or nonsensical conclusions.*

Later chapters introduce other methods of assessing whether a given experimental model is likely to be valid. However, in using any such technique, it is important to keep your wits about you – use common sense and your understanding of the experimental protocol to decide whether a given statistical technique will be useful or not. Always remember:

Mathematics can't turn processed pig feed into gold.

The point is that the your data may not be sufficiently powerful to distinguish various models; and if this is the case, then no amount of fancy mathematics or statistics is going to provide anything useful.

7.7 Exercises:

1. Suppose that the probability that any given pollen grain or ovule has the recessive allele is some number $\alpha \in [0, 1]$, and so the probability that either has the dominant gene is $(1 - \alpha)$.

   (a) Consider an experiment where 4 seedlings are examined for the dominant or recessive trait. Thus, the outcome of such an experiment is a vector in $\mathbf{R}^4$ whose kth entry is 0 if the kth seedling has the recessive trait and it is 1 if the kth seedling has the dominant trait. Let S denote the $2^4 = 16$ element set of possible experimental outcomes. Let $s \in S$ denote an element with some $m \in \{0, \ldots, 4\}$ entries equal to 1 and the remaining entries equal to zero. Explain why s has probability $(1 - \alpha)^m \alpha^{4-m}$.

   (b) Consider now the analogous experiment where 915 seedlings are examined for the dominant or recessive trait. Thus, the outcome of such an experiment is a vector in $\mathbf{R}^{915}$ whose kth entry is 0 if the kth seedling has the recessive trait and it is 1 if the kth seedling has the dominant trait. Let S denote the 2^{915} element set of possible experimental outcomes. Let $s \in S$ denote an element with some $m \in \{0, \ldots, 915\}$ entries equal to 1 and the remaining entries equal to zero. Explain why the m has probability $(1 - \alpha)^m \alpha^{915-m}$.

   (c) In the case of part (a) above, define the random variable $f : S \to \{0, 1, 2, 3, 4\}$ where $f(s)$ is the sum of the entries of s. Compute the induced probability function P_f on the set $\{0, 1, 2, 3, 4\}$.

CHAPTER

EIGHT

Kernel and image in biology

Imagine a complicated cellular process that involves some n genes that are presumably 'turned on' by some N other genes that act at an earlier time. We want to test whether the effect of the early genes on the late genes involves a complicated synergy, or whether their affects simply add. To do this, I could try to derive the consequences of one or the other of these possibilities and then devise an experiment to see if the predicted consequences arise. Such an experiment could vary the expression level of the early genes from their normal levels (either $+$ or $-$) and see how the variation in the level of expression of the late genes from their normal levels changes accordingly.

To elaborate on this strategy, note that when a gene is expressed, its genetic code (a stretch of the DNA molecule) is used to code for a molecule much like DNA called mRNA. Here, the 'm' stands for 'messenger' and the RNA part is a sequence of small molecules strung end to end. Any of these can be one of four and the resulting sequence along the RNA string is determined by the original sequence on the coding stretch of DNA. This messenger RNA subsequently attaches to the protein making part of a cell (a 'ribosome') where its sequence is used to construct a particular protein molecule. In any event, the level of any given mRNA can be measured at any given time, and this level serves as a proxy for the level of expression of the gene that coded it in the first place.

One more thing to note: The level of expression of a gene can often be varied with some accuracy in an experiment by inserting into the cell nucleus certain tailored molecules to either promote or repress the gene expression. Such is the magic of modern biotechnology.

To make a prediction that is testable by measuring early and late gene expression, let us suppose that the affects of the early genes are simply additive and see where this assumption leads. For this purpose, label these early genes by integers from 1 to N, and let u_k to denote the deviation, either positive or negative, of the level of expression of the kth early gene from its normal level. For example, we can take u_k to denote the deviation from normal of the concentration of the mRNA that comes from the kth early gene.

Meanwhile, label the late genes by integers from 1 to n, and use p_j to denote the deviation of the latter's mRNA from its normal level. If the affects of the early genes on the late genes are simply additive, we might expect that any given p_j has the form

$$p_j = A_{j1}u_1 + A_{j2}u_2 + \cdots + A_{jN}u_N, \tag{8.1}$$

where each A_{jk} is a constant. This is to say that the level p_j is a sum of factors, the first proportional to the amount of the first early gene, the second proportional to the amount of the second early gene, and so on. Note that when A_{jk} is positive, then the kth early gene tends to promote the expression of the jth late gene. Conversely, when A_{jk} is negative, the kth early gene acts to repress the expression of the jth late gene.

If we use $\vec{u}$ to denote the N-component column vector whose kth entry is u_k, and if we use $\vec{p}$ to denote the n-component column vector whose jth component is p_j, then the equation in (8.1) is the matrix equation $\vec{p} = A\vec{u}$. Thus, we see a linear transformation from an N-dimensional space to an n-dimensional one.

By the way, note that the relation predicted by (8.1) can, in principle, be tested by experiments that vary the levels of the early genes and see if the levels of the late genes change in a manner that is consistent with (8.1). Such experiments will also determine the values for the matrix entries $\{A_{jk}\}$. For example, to find A_{11}, vary u_1 while keeping all $k > 1$ versions of u_k equal to zero. Measure p_1 as these variations are made and see if the ratio p_1/u_1 is constant as u_1 changes with each $k > 1$ version of u_k equal zero. If so, the constant is the value to take for A_{11}. If this ratio is not

constant as these variations are made, then the linear model is wrong. One can do similar things with the other p_j and u_k to determine all A_{jk}. One can then see about changing more than one u_k from zero and see if the result conforms to (8.1).

The question now arises as to the meaning of the kernel and the image of the linear transformation A from $\mathbf{R}^N$ to $\mathbf{R}^n$. To make things explicit here, suppose that n and N are both equal to 3 and that A is the matrix

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \tag{8.2}$$

As you can check, this matrix has kernel equal to the scalar multiples of the vector

$$\begin{bmatrix} 1 \\ -3 \\ 1 \end{bmatrix} \tag{8.3}$$

Meanwhile, its image is the span of the vectors

$$\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}. \tag{8.4}$$

Thus, it consists of all vectors in $\mathbf{R}^3$ that can be written as a constant times the first vector in (8.4) plus another constant times the second.

Here is the meaning of the kernel: Vary the early genes 1, 2 and 3 from their normal levels in the ratio $u_1/u_3 = 1$ and $u_2/u_3 = -3$ and there is no affect on the late genes. This is to say that if the expression levels of early genes 1 and 3 are increased by any given amount r while that of early gene 2 is decreased by $3r$, then there is no change to the levels of expression of the three late genes. In a sense, the decrease in the level by a factor of 3 of the second early gene exactly offsets the affect of increasing the equal increases in the levels of the first and third early genes.

As to the meaning of the image, what we find is that only certain deviations of the levels of expression of the three late genes from their background values can be obtained by modifying the expression levels of the three early genes. For example, both vectors in (8.4) are orthogonal to the vector

$$\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}. \tag{8.5}$$

Thus, values of p_1, p_2 and p_3 with the property that $p_1 + p_3 \neq p_2$ can not be obtained by any variation in the expression levels of the three early genes. Indeed, the dot product of the vector $\vec{p}$ with the vector in (8.5) is $p_1 - p_2 + p_3$ and this must be zero in the case that is a linear combination of the vectors in (8.4).

Granted that the matrix A is that in (8.2), then the preceding observation has the following consequence for the biologist: If values of p_1, p_2 and p_3 are observed in a cell with $p_1 + p_3 \neq p_2$, then the three early genes can not be the sole causative agent for the expression levels of the three late genes.

CHAPTER

NINE

Dimensions and coordinates in a scientific context

My purpose here is to give some toy examples where the notion of dimension and coordinates appear in a biological context.

9.1 Coordinates

Here is a hypothetical situation: Suppose that a cell has genes labeled $\{1, 2, 3\}$. The level of the corresponding product then defines vector in $\mathbf{R}^3$ where the obvious coordinates, x_1, x_2 and x_3, measure the respective levels of the products of gene 1, gene 2 and gene 3. However, this might not be the most useful coordinate system. In particular, if some subsets of genes are often turned on at the same time and in the same amounts, it might be better to change to a basis where that subset gives one of the basis vectors. Suppose for the sake of argument, that it is usually the case that the level of the product from gene 2 is three times that of gene 1, while the level of the product of gene 3 is half that of gene 1. This is to say that one usually finds $x_2 = 3x_1$ and $x_3 = x_1$. Then it might make sense to switch from the standard coordinate bases,

$$\vec{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad \vec{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \vec{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \tag{9.1}$$

to the coordinate system that uses a basis $\vec{v}_1$, $\vec{v}_2$ and $\vec{v}_3$ where

$$\vec{v}_1 = \begin{bmatrix} 1 \\ 3 \\ \frac{1}{2} \end{bmatrix}, \qquad \vec{v}_2 = \vec{e}_2, \qquad \vec{v}_3 = \vec{e}_3. \tag{9.2}$$

To explain, suppose I measure some values for x_1, x_2 and x_3. This then gives a vector,

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1\vec{e}_1 + x_2\vec{e}_2 + x_3\vec{e}_3. \tag{9.3}$$

Now, I can also write this vector in terms of the basis in (9.2) as

$$\vec{x} = c_1\vec{v}_1 + c_2\vec{v}_2 + c_3\vec{v}_3. \tag{9.4}$$

With $\vec{v}_1$, $\vec{v}_2$ and $\vec{v}_3$ as in (9.2), the coordinates c_1, c_2 and c_3 that appear in (9.5) are

$$c_1 = x_1, \qquad c_2 = x_2 - 3x_1, \qquad \text{and} \qquad c_3 = x_3 - x_1. \tag{9.5}$$

As a consequence, the coordinates c_2 describes the deviation of x_2 from its usual value of $3x_1$. Meanwhile, the coordinate c_3 describes the deviation of x_3 from its usual value of x_1.

Here is another example: Suppose now that there are again three genes with the levels of their corresponding products denoted as x_1, x_2, and x_3. Now suppose that it is usually the case that these levels are correlated in that x_3 is generally very close to $2x_2 + x_1$. Any given set of measured values for these products determines now a column vector as in (9.3). A useful basis in this case would by one where the coordinates c_1, c_2 and c_3 has

$$c_1 = x_1, \qquad c_2 = x_2, \qquad \text{and} \qquad c_3 = x_3 - 2x_2 - x_1. \tag{9.6}$$

Thus, c_3 again measures the deviation from the expected values. The basis with this property is that where

$$\vec{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \qquad \vec{v}_2 = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \qquad \text{and} \qquad \vec{v}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{9.7}$$

This is to say that if $\vec{v}_1$, $\vec{v}_2$, and $\vec{v}_3$ are as depicted in (9.7), and if $\vec{x}$ is then expanded in this basis as $c_1\vec{v}_1 + c_2\vec{v}_2 + c_3\vec{v}_3$, then c_1, c_2 and c_3 are given by (9.6).

9.2 A systematic approach

If you are asking how I know to take the basis in (9.7) to get the coordinate relations in (9.6), here is the answer: Suppose that you have coordinates x_1, x_2 and x_3 and you desire new coordinates, c_1, c_2 and c_3 that are related to the x's by a linear transformation:

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = A \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \tag{9.8}$$

where A is an invertible, 3×3 matrix. In this regard, I am supposing that you have determined already the matrix A and are simply looking now to find the vectors $\vec{v}_1$, $\vec{v}_2$ and $\vec{v}_3$ that allow you to write $\vec{x} = c_1\vec{v}_1 + c_2\vec{v}_2 + c_3\vec{v}_3$ with c_1, c_2 and c_3 given by (10.8). As explained in the linear algebra text, the vectors to take are:

$$\vec{v}_1 = A^{-1}\vec{e}_1, \qquad \vec{v}_2 = A^{-1}\vec{e}_2, \qquad \vec{v}_3 = A^{-1}\vec{e}_3. \tag{9.9}$$

To explain why (9.9) holds, take the equation $\vec{x} = c_1\vec{v}_1 + c_2\vec{v}_2 + c_3\vec{v}_3$ and act on both sides by the linear transformation A. According to (9.8), the left-hand side, $A\vec{x}$, is the vector $\vec{c}$ whose top component is c_1, middle component is c_2 and bottom component is c_3. This is to say that $A\vec{x} = c_1\vec{e}_1 + c_2\vec{e}_2 + c_3\vec{e}_3$. Meanwhile, the left-hand side of the resulting equation is $c_1A\vec{v}_1 + c_2A\vec{v}_2 + c_3A\vec{v}_3$. Thus,

$$c_1\vec{e}_1 + c_2\vec{e}_2 + c_3\vec{e}_3 = c_1A\vec{v}_1 + c_2A\vec{v}_2 + c_3A\vec{v}_3. \tag{9.10}$$

Now, the two sides of (9.10) are supposed to be equal for all possible values of c_1, c_2 and c_3. In particular, they are equal when $c_1 = 1$ and $c_2 = c_3 = 0$. For these choices, the equality in (9.10) asserts that $\vec{e}_1 = A\vec{v}_1$; this the left-most equality in (9.9). Likewise, setting $c_1 = c_3 = 0$ and $c_2 = 1$ in (9.10) gives the equivalent of the middle equality in (9.9); and setting $c_1 = c_2 = 0$ and $c_3 = 1$ in (9.10) gives the equivalent of the right-most equality in (9.9).

9.3 Dimensions

What follows is an example of how the notion of dimension arises in a scientific context. Consider the situation in Section 9.1, above, where the system is such that the levels of x_2 and x_3 are very nearly $x_2 \approx 3x_1$ and $x_3 \approx x_1$. This is to say that when we use the coordinate c_1, c_2 and c_3 in (9.5), then $|c_2|$ and $|c_3|$ are typically very small. In this case, a reasonably accurate model for the behavior of the three gene system can be had by simply assuming that c_2 and c_3 are always measured to be identically zero. As such, the value of the coordinate c_1 describes the system to great accuracy. Since only one coordinate is needed to describe the system, it is said to be '1-dimensional'.

A second example is the system that is described by c_1, c_2 and c_3 as depicted in (9.6). If it is always the case that x_3 is very close to $2x_2 + x_1$, then the system can be described with good accuracy with c_3 set equal to zero. This done,

then one need only specify the values of c_1 and c_2 to describe the system. As there are two coordinates needed, this system would be deemed '2-dimensional'.

In general, some sort of time dependent phenomena is deemed 'n-dimensional' when n coordinates are required to describe the behavior to some acceptable level of accuracy. Of course, it is typically the case that the value of n depends on the desired level of accuracy.

9.4 Exercises:

1. Suppose that four genes have corresponding products with levels x_1, x_2, x_3 and x_4 where x_4 is always very close to $x_1 + 4x_2$ while x_3 is always very close to $2x_1 + x_2$. Find a new set of basis vectors for $\mathbf{R}^4$ and corresponding coordinates c_1, c_2, c_3 and c_4 with the following property: The values of x_1, x_2, x_3 and x_4 for this four gene system are the points in the (c_1, c_2, c_3, c_4) coordinate system where c_3 and c_4 are nearly zero.

2. Suppose that two genes are either 'on' of 'off', so that there are affectively, just four states for the two gene system, $\{++, +-, -+, --\}$, where $++$ means that both genes are on; $+-$ means that the first is on and the second is off; etc. Assume that these four states have respective probabilities $\frac{1}{2}$, $\frac{1}{6}$, $\frac{1}{6}$, $\frac{1}{6}$.

   (a) Is the event that the first gene is on independent from the event that the second gene is on?

   Now suppose that these two genes jointly influence the levels of two different products. The levels of the first product are given by $\{3, 2, 1, 0\}$ in the respective states $++$, $+-$, $-+$, $--$. The levels of the second are $\{4, 2, 3, 1\}$ in these same states.

   (b) View the levels of the two products as random variables on the sample space S that consists of $\{++, +-, -+, --\}$ with the probabilities as stated. Write down the mean and standard deviations for these two random variables.

   (c) Compute the correlation matrix in equation (6.8) of Chapter 6 for these two random variables to prove that they are not independent.

CHAPTER

TEN

More about Bayesian statistics

The term 'Bayesian statistics' has different meanings for different people. Roughly, Bayesian statistics reverses 'causes' and 'effects' so as to make an *educated guess* about the causes given the known effects. The goal is deduce a probability function on the set of possible causes granted that we have the probabilities of the various effects.

Take note that I have underlined the words 'educated guess'. There are situations when the Bayesian strategy seems reasonable, and others where it doesn't.

10.1 A problem for Bayesians

There is a sample space of interest, S, with a known function (i.e. random variables) f to another finite set, W. A probability function for the set W is in hand, but what is needed is one for the set S.

Here is a simple situation that exemplifies this: Flip two distinct coins, coin #1 and coin #2. Move to the right one step ($x \to x+1$) for each heads that appears and to the left one step ($x \to x-1$) for each tails. Let W denote the set of possible positions after two flips, thus $W = \{-2, 0, 2\}$. Meanwhile, the sample space is $S = \{HH, HT, TH, TT\}$. We can do this experiment many times and so generate numbers $\mathrm{P}_{exp}(-2)$, $\mathrm{P}_{exp}(0)$ and $\mathrm{P}_{exp}(2)$ that give the respective frequencies that -2, 0 and 2 are the resulting positions. How can we use these frequencies to determine the probability of getting heads on coin #1, and also the probability of getting heads on coin #2? In this regard, we don't want to assume that these coins are fair, nor do we want to assume that the probability of heads for coin #1 is the same as that for coin #2.

10.2 A second problem

A six-sided die, hidden from view, is rolled twice and the resulting pair of numbers (each either $1, 2, \ldots, 6$) are added to obtain a single number, thus an integer that can be as small as 2 or as large as 12. We are told what this sum is, but not the two integers that appeared. If this is done many times, how can the relative frequencies for the various values of the sum be used to determine a probability function for the sample space?

10.3 Meet the typical Bayesian

To set the stage, remember that if P is any given probability function on S, then P induces one on W by the rule we saw in Chapter 5. Indeed, if the latter is denoted by P_f, the rule is that $\mathrm{P}_f(r)$ is the probability as measured by P of the subset of points in S where f has value r. Thus,

$$\mathrm{P}_f(r) = \sum_{s \in S \text{ with } f(s)=r} \mathrm{P}(s). \tag{10.1}$$

This last equation can be written in terms of conditional probabilities as follows:

$$\mathrm{P}_f(r) = \sum_{s \in S} \mathrm{P}\left(r \mid s\right) \mathrm{P}(s) \tag{10.2}$$

where $\mathrm{P}\left(r \mid s\right)$ is the conditional probability that $f = r$ given that you are at the point $s \in S$. Of course, this just says that $\mathrm{P}\left(r \mid s\right)$ is 1 if $f(s) = r$ and 0 otherwise.

The problem faced by statisticians is to deduce P, or a reasonable approximation, given only knowledge of some previously determined probability function, P_W, on the set W. In effect, we want to find a probability function P on S whose corresponding P_f is the known function P_W.

Your typical Bayesian will derive a guess for P using the following strategy:

Step 1: Assume that there is some conditional probability, $K(s; r)$, that gives the probability of obtaining any given s from S granted that the value of f is r. If such a suite of conditional probabilities were available, then one could take

$$\mathrm{P}_{\text{guess}}(s) = \sum_{r \in W} K(s; r) \mathrm{P}_W(r). \tag{10.3}$$

The problem is that the points in W are the values of a function of the points in S, not vice-versa. Thus, there is often no readily available $K(s; r)$.

Step 2: A Bayesian is not deterred by this state of affairs. Rather, the Bayesian plows ahead by using what we have, which is $\mathrm{P}\left(r \mid s\right)$. We know its values in all cases; it is 1 when $f(s) = r$ and zero otherwise. Why not, asks the Bayesian, take

$$K(s; r) = \frac{1}{Z(r)} \mathrm{P}\left(r \mid s\right), \tag{10.4}$$

where $Z(r)$ is the number of points in S on which f has value r. This, is to say that

$$Z(r) = \sum_{s \in S} \mathrm{P}\left(r \mid s\right). \tag{10.5}$$

To explain the appearance of $Z(r)$, remember that a conditional probability of the form $\mathrm{P}\left(A \mid B\right)$ is a probability function in its own right on the sample space S. Thus, $\mathrm{P}\left(S \mid B\right)$ must be 1 if S is the whole sample space. This need not be the case for $K(S; r)$ were the factor of $1/Z(r)$ absent.

Step 3: To summarize: Our typical Bayesian takes the following as a good guess for the probability function on S:

$$\mathrm{P}_{\text{Bayes}}(s) = \sum_{r \in W} \frac{1}{Z(r)} \mathrm{P}\left(r \mid s\right) \mathrm{P}_W(r). \tag{10.6}$$

Note that disentangling the definitions, there is really no summation involved in (10.6) because there is just one value of r that makes $\mathrm{P}\left(r \mid s\right)$ non-zero for any given s, this the value $r = f(s)$. Thus, (10.6) is a very roundabout way of saying that

$$\mathrm{P}_{\text{Bayes}}(s) = \frac{1}{Z(f(s))} \mathrm{P}_W(f(s)). \tag{10.7}$$

This is our Bayesian's guess for the probability function on S.

10.4 A first example

Consider the problem in Part a with flipping coin #1 and coin #2. As noted there, W denotes the set of possible positions after the two coin flips, thus $W = \{-2, 0, 2\}$. The set $S = \{HH, HT, TH, TT\}$. Suppose first that our

two coins have the same probability for heads, some number $q \in (0,1)$. Thus T has probability $1-q$, then the true probabilities for the elements in S are q^2, $q(1-q)$, $q(1-q)$ and $(1-q)^2$ in the order they appear above. These probability assignments give P_{true} on S. With these true probabilities, the frequencies of appearances of the three elements in W are $(1-q)^2$, $2q(1-q)$ and q^2. I'll take these for my definition of the probability function P_W on W.

Let's now see what the Bayesian would find for $\mathrm{P}_{\text{Bayes}}$. For this purpose, note that the only non-zero values of $\mathrm{P}(r \mid s)$ that appear in the relevant version of (9.6) are

$$\begin{aligned} &\bullet\ \mathrm{P}(-2, TT) = 1 \\ &\bullet\ \mathrm{P}(0, HT) = \mathrm{P}(0, TH) = 1 \\ &\bullet\ \mathrm{P}(2, HH) = 1 \end{aligned} \tag{10.8}$$

Thus, $Z(\pm 2) = 1$ and $Z(0) = 2$. Plugging this into (10.7) finds

$$\mathrm{P}_{\text{Bayes}}(HH) = q^2, \qquad \mathrm{P}_{\text{Bayes}}(TT) = (1-q)^2 \qquad \text{and} \qquad \mathrm{P}_{\text{Bayes}}(HT) = \mathrm{P}_{\text{Bayes}}(TH) = q(1-q). \tag{10.9}$$

Thus, the Bayesian guess for probabilities is the true probability.

10.5 A second example

Let us now change the rules in the coin flip game and consider the case where the first flip uses a fair coin (probability $\frac{1}{2}$) for either H or T, and the second uses a biased coin, with probability q for H and thus $(1-q)$ for T. In this case, the true probability on S is given by

$$\mathrm{P}_{\text{true}}(HH) = \tfrac{1}{2}q, \qquad \mathrm{P}_{\text{true}}(HT) = \tfrac{1}{2}(1-q), \qquad \mathrm{P}_{\text{true}}(TH) = \tfrac{1}{2}q, \qquad \text{and} \qquad \mathrm{P}_{\text{true}}(TT) = \tfrac{1}{2}(1-q). \tag{10.10}$$

The frequencies of appearance of the three positions in W are now $\frac{1}{2}(1-q)$, $\frac{1}{2}$, $\frac{1}{2}q$. I use these three numbers for the probabilities given by P_W. As the conditional probabilities in (10.8) do not change, we can employ then in (10.6) to find the Bayesian guess:

$$\mathrm{P}_{\text{Bayes}}(HH) = \tfrac{1}{2}q, \qquad \mathrm{P}_{\text{Bayes}}(HT) = \tfrac{1}{4}, \qquad \mathrm{P}_{\text{Bayes}}(TH) = \tfrac{1}{4}, \qquad \text{and} \qquad \mathrm{P}_{\text{Bayes}}(TT) = \tfrac{1}{2}(1-q). \tag{10.11}$$

Thus, the Bayesian guess goes bad when q deviates from $\frac{1}{2}$.

Roughly speaking, the Bayesian guess cannot distinguish between those points in the sample space that give the same value for random variable f.

10.6 Something traumatic

Let me show you something that is strange about the Bayesian's guess in (10.11). Suppose we ask for the probability as computed by $\mathrm{P}_{\text{Bayes}}$ that H appears on the first coin. According to our rules of probability,

$$\mathrm{P}_{\text{Bayes}}(\text{coin } \#1 = H) = \mathrm{P}_{\text{Bayes}}(HH) + \mathrm{P}_{\text{Bayes}}(HT) = \tfrac{1}{2}q + \tfrac{1}{4}. \tag{10.12}$$

This is also the probability $\mathrm{P}_{\text{Bayes}}(\text{coin } \#2 = H)$ since $\mathrm{P}_{\text{Bayes}}(HT) = \mathrm{P}_{\text{Bayes}}(TH)$. Now, note that

$$\mathrm{P}_{\text{Bayes}}(HH) \neq \mathrm{P}_{\text{Bayes}}(\text{coin } \#1 = H)\mathrm{P}_{\text{Bayes}}(\text{coin } \#2 = H) \tag{10.13}$$

unless $q = \frac{1}{2}$ since the left-hand side is $\frac{1}{2}q$ and the right is $(\frac{1}{2}q + \frac{1}{4})^2$. Thus, the Bayesian finds that the event of coin $\#1 = H$ is *not* independent of the event that coin $\#2 = H$!! (Remember that events A and B are deemed independent when $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$.)

10.7 Rolling dice

Consider here the case where the die is rolled twice and the resulting two integers are added. The sample space, S, consists of the 36 pairs of the form (a, b) where a and b are integers from the set $\{1, \ldots, 6\}$. The random variable (a.k.a. function of S) is the function that assigns $a + b$ to any given $(a, b) \in S$. Thus, the set of possible outcomes in $W = \{2, ..., 12\}$.

Suppose, for the sake of argument, that the true probabilities for rolling $1, 2, \ldots, 6$ on any given throw of the die are $\frac{1}{21}, \frac{2}{21}, \frac{3}{21}, \frac{4}{21}, \frac{5}{21}, \frac{6}{21}$. Were this the case, then the true probability, P_{true}, for any given pair (a, b) in S is

$$\mathrm{P}_{\text{true}}(a, b) = \tfrac{a}{21} \cdot \tfrac{b}{21} = \tfrac{ab}{441}. \tag{10.14}$$

If the die has these probabilities, then the probabilities that result for the outcomes are

$$\begin{aligned}
&\mathrm{P}_W(2) = \tfrac{1}{441}, && \mathrm{P}_W(3) = \tfrac{4}{441}, && \mathrm{P}_W(4) = \tfrac{10}{441}, && \mathrm{P}_W(5) = \tfrac{20}{441}, \\
&\mathrm{P}_W(6) = \tfrac{35}{441}, && \mathrm{P}_W(7) = \tfrac{56}{441}, && \mathrm{P}_W(8) = \tfrac{70}{441}, && \mathrm{P}_W(9) = \tfrac{76}{441}, \\
&\mathrm{P}_W(10) = \tfrac{73}{441}, && \mathrm{P}_W(11) = \tfrac{60}{441}, && \mathrm{P}_W(12) = \tfrac{36}{441}.
\end{aligned} \tag{10.15}$$

As indicated by my notation, I am using the preceding probabilities for P_W.

Now, given that we have P_W as just described, here is what the Bayesian finds for the probabilities of some of the elements in the sample space S:

$$\begin{aligned}
&\mathrm{P}_{\text{Bayes}}(1, 1) = \tfrac{1}{441} \\
&\mathrm{P}_{\text{Bayes}}(2, 1) = \mathrm{P}_{\text{Bayes}}(1, 2) = \tfrac{2}{441} \\
&\mathrm{P}_{\text{Bayes}}(3, 1) = \mathrm{P}_{\text{Bayes}}(2, 2) = \mathrm{P}_{\text{Bayes}}(1, 3) = \tfrac{1}{3}\ \tfrac{10}{441} \\
&\mathrm{P}_{\text{Bayes}}(4, 1) = \mathrm{P}_{\text{Bayes}}(3, 2) = \mathrm{P}_{\text{Bayes}}(2, 3) = \mathrm{P}_{\text{Bayes}}(1, 4) = \tfrac{5}{441} \\
&\mathrm{P}_{\text{Bayes}}(5, 1) = \mathrm{P}_{\text{Bayes}}(4, 2) = \mathrm{P}_{\text{Bayes}}(3, 3) = \mathrm{P}_{\text{Bayes}}(2, 4) = \mathrm{P}_{\text{Bayes}}(1, 5) = \tfrac{7}{441} \\
&\mathrm{P}_{\text{Bayes}}(6, 1) = \mathrm{P}_{\text{Bayes}}(5, 2) = \mathrm{P}_{\text{Bayes}}(4, 3) = \mathrm{P}_{\text{Bayes}}(3, 4) = \mathrm{P}_{\text{Bayes}}(2, 5) = \mathrm{P}_{\text{Bayes}}(1, 6) = \tfrac{1}{3}\ \tfrac{28}{441} \\
&\mathrm{P}_{\text{Bayes}}(6, 2) = \mathrm{P}_{\text{Bayes}}(5, 3) = \mathrm{P}_{\text{Bayes}}(4, 4) = \mathrm{P}_{\text{Bayes}}(3, 5) = \mathrm{P}_{\text{Bayes}}(2, 6) = \tfrac{14}{441} \\
&\text{and so on.}
\end{aligned} \tag{10.16}$$

10.8 Exercises:

1. (a) Complete the table in (10.16) by computing the values of $\mathrm{P}_{\text{Bayes}}$ on the remaining pairs in S.
   (b) According to $\mathrm{P}_{\text{Bayes}}$, is the event of the first roll comes up 1 independent from the event that second roll comes up 6? Justify your answer.
2. Compute the mean and standard deviation for the random variable $a + b$ first using P_{true} from (10.15) and then using $\mathrm{P}_{\text{Bayes}}$.
3. Consider now the same sample space for rolling a die twice, but now suppose that the die is fair, and so each number has probability of turning up on any given roll.
   (a) Compute the mean and standard deviation of the random variable $a + b$.
   (b) Compute the mean and standard deviation for the random variable ab.
   (c) Are the random variables $a + b$ and ab independent? In this regard, remember that two random variables, f and g, are said to be independent when $\mathrm{P}(f = r \text{ and } g = s) = \mathrm{P}(f = r)\mathrm{P}(g = s)$ for all pairs (r, s) where r is a possible value of f and s is a possible value of g. Justify your answer.

CHAPTER

ELEVEN

Common probability functions

There are certain probability functions that serve as models that are commonly used when trying to decide if a given phenomena is 'unexpected' or not. This chapter describes those that arise most often.

11.1 What does 'random' really mean?

Consider a bacterial cell that moves one cell length, either to the right or left, in each unit of time. If we start some large number of these cells at $x = 0$ and wait t units of time, we can determine a function, $p_t(x)$, which is the fraction of bacteria that end up at position $x \in \{0, \pm 1, \pm 2, \ldots\}$ at time t. We can now ask: What do we expect the function $x \to p_t(x)$ to look like?

Suppose that we think that the bacteria is moving 'randomly'. Two questions then arise:

- How do we translate our intuitive notion of the English term 'random' into a prediction for $p_t(x)$?
- Granted we have a prediction, for each t and x, then how far must $p_t(x)$ be from its predicted value before we must accept the fact that the bacteria is not moving according to our preconceived notion of random?

(11.1)

These questions go straight to the heart of what is called the 'scientific method'. We made a hypothesis: The bacterium moves left or right at random'. We want to first generate some testable predictions of the hypothesis (the first point in (11.1)), and then compare these predictions with experiment. The second point in (11.1) asks for criterion to use to evaluate whether the experiment confirms or rejects our hypothesis.

The first question in (11.1) is the provenance of 'probability theory' and the second the provenance of 'statistics'. This chapter addresses the first question in (11.1), while aspects of the second are addressed in some of the subsequent chapters.

11.2 A mathematical translation of the term 'random'

To say that an element is chosen from a given set at 'random' is traditionally given the following mathematical definition:

> Probabilities are defined using the probability function that assigns all elements the same probability. If the set has L elements, then the probability of any given element appearing is $\frac{1}{L}$. (11.2)

The probability function that assigns this constant value to all elements is called the *uniform probability* function.

Here is an archetypal example: A coin is flipped N times. Our sample space is the set S that consists of the 2^N possible sequences $(\pm 1, \pm 1, \ldots, \pm 1)$, where $+1$ is in the kth spot when the kth flip landed heads up, while -1 sits in this slot when the kth flip landed tails up. If I assume that the appearance of heads on any flip has probability , and that the appearance of heads on any subset of flips has no bearing on what happens in the other flips, then I would predict that the frequencies of appearance of any given sequence in S is 2^{-N}. This is to say that I would use the uniform probability function on S to predict the frequency of appearance of any subset of its elements.

Note that after setting $N = t$, this same sample space describes all of the possibilities for the moves of our bacterium from Section 11.1, above.

Here is another example: You go to a casino to watch people playing the game of 'craps'. Remember that this game is played by rolling two six-sided dice, and looking at the numbers that show on the top faces when the dice stop rolling. The sample space for one play of the game is the set of 36 elements where each is of the form (a, b) for a, b integers from 1 through 6. If I believe that the dice are 'fair' and that the appearance of any given number on one die has no bearing on what appears on the other, then I would use the uniform probability function on S to predict the frequency of appearance of any given outcome.

I might watch N games of craps played. In this case, the sample space is the set of 36^N possible sequence of the form $((a_1, b_1), \ldots, (a_N, b_N))$ where each a_k and each b_k is a number from 1 through 6. If I believe that the appearance of any face on the dice is as likely as any other, and if I believe that the appearance of any sequence in any given subset of the N games has no affect on what happens in the other games, then I would make my predictions for the frequencies using the uniform probability function on this 36^N element sample space.

What follows is yet one more example: Look at a single stranded DNA molecule that is composed of N bases strung end to end. As each cite on the DNA molecule can be occupied by one of 4 bases, the sample space that describes the various possibilities for this DNA molecule is the 4^N element set whose typical member is a sequence $(\theta_1, \ldots, \theta_N)$ where each θ_k is either A, C, G or T. If I believe that each such letter is equally likely to appear, and that the appearance of a given letter in any one slot has no bearing on what happens in the other slots, then I would use the uniform probability function to predict the frequencies of occurrences of the various letters in length N strand of DNA.

You might think that the uniform probability distribution is frightfully dull – after all, how much can you say about a constant?

11.3 Some standard counting solutions

The uniform probability distribution becomes interesting when you consider the probabilities for certain subsets, or the probabilities for the values of certain random variables. To motivate our interest in the uniform probability distribution, consider first its appearance in the case of bacteria. Here, our model of random behavior takes S as just described, the set of 2^N sequences of the form $(\alpha_1, \alpha_2, \ldots, \alpha_N)$, where each α_k is 1 or -1. Now define f so that $f(\alpha_1, \cdots, \alpha_N) = \alpha_1 + \alpha_2 + \cdots + \alpha_N$ and ask for the probabilities of the possible values of f. With regards to bacteria, f tells us the position of the bacteria after N steps in the model where the bacteria moves right or left with equal probability at each step. The probabilities for the possible values of f provide the theoretical predictions for the measured function $p_t(x)$ in the case that $t = N$. Thus, uniform probability function on S predicts that $p_{t=N}(x)$ is equal to 2^{-N} times the number of sequences in S whose entries add up to x.

Note that $p_N(x)$ is *not* the uniform probability function on the set of possible values for f. To see that such is the case, note that the possible values for x are the integers from the set $W = \{-N, -N+2, \ldots, N-2, N\}$. This set has $N + 1$ elements. Since only the sequence $(-1, -1, \ldots, -1)$ has entries that sum to $-N$, so $p_N(-N) = 2^{-N}$. Meanwhile, the only sequence from S whose entries sum to N is $(1, 1, \ldots, 1)$, and so $p_N(N) = 2^{-N}$ also. However there are N elements in S whose entries sum to $-N+2$, these the elements that have a single entry equal to $+1$. Thus, $p_N(-N+2) = N2^{-N}$. A similar count finds $p_N(N-2) = N2^{-N}$.

In general, if a set S has N elements and a subset $K \subset S$ has k elements, then the uniform probability distribution on S gives probability $\frac{k}{N}$ to the set K. Even so, it may be some task to count the elements in any given set. Moreover, the degree of difficulty may depend on the manner in which the set is described. There are, however, some standard counting formulae available to facilitate things. For example, let S denote the sample space with 2^N elements as

described above. For $n \in \{0, 1, \dots, N\}$, let K_n denote the subset of elements in S with n occurrences of $+1$, thus with $N - n$ occurrences of -1. Note that K_n has an alternate description, this as the set of elements on which $f(\alpha_1, \dots, \alpha_N) = \alpha_1 + \alpha_2 + \cdots + \alpha_N$ has value $-N + 2n$. Thus, the number of elements in K_n times 2^{-N} gives $p_{t=N}(x)$ in the case where $x = -N + 2n$.

In any event, here is a basic fact:

$$\text{The set } K_n \text{ has } \frac{N!}{n!(N-n)!} \text{ members.} \tag{11.3}$$

In this regard, remember that $k!$ is defined for any positive integer k as $k(k-1)(k-2)\cdots 1$. Also, 0! is defined to by equal to 1. For those who don't like to take facts without proof, I explain in the last section below how to derive (11.3) and also the formulae that follow.

By the way, arises often enough in counting problems to warrant its own symbol, this

$$\binom{N}{n}. \tag{11.4}$$

Here is another standard counting formula: Let $b \geq 1$ be given, and let S denote the set of b^N elements of the form $(\beta_1, \dots, \beta_N)$, where each β_k is now $\{1, \dots, b\}$. For example, if $b = 6$, then S is the sample space for the list of faces that appear when a six-sided die is rolled N times. If $b = 4$, then S is the sample space for the possible single strands of DNA with N bases.

If $N > b$, then each element in S has at least one β_k that is the same as another. If $N \leq b$, then there can be elements in S where no two β_k are the same. Fix b and let E_b denote the subset of those N-tuples $(\beta_1, \dots, \beta_N)$ where no two β_k are identical.

$$\text{The set } E_b \text{ has } \frac{b!}{(b-N)!} \text{ members.} \tag{11.5}$$

The case $n = N$ in (11.5) provides the following:

$$\text{There are } N! \text{ ways to order a set of } N \text{ distinct elements.} \tag{11.6}$$

Here, a set of elements is 'ordered' simply by listing them one after the other. For example, the set that consists of 1 apple and 1 orange has two orderings, (apple, orange) and (orange, apple). The set that consist of the three elements {apple, orange, grape} has six orderings, (apple, orange, grape), (apple, grape, orange), (orange, grape, apple), (orange, apple, grape), (grape, apple, orange), (grape, orange, apple).

Here is a direct argument for (11.6): To count the number of orderings, note that there are N possible choices for the first in line. With the first in line chosen, then $N - 1$ elements can be second in line. With the first two in line chosen, there are $N - 2$ left that can be third in line. Continuing in this reasoning leads to (11.6).

11.4 Some standard probability functions

The counting formulae just presented can be used to derive some probability functions that you will almost surely see again and again in your scientific career.

The Equal Probability Binomial: Let S denote the sample space with the 2^N elements of the form $(\alpha_1, \dots, \alpha_N)$ where each α_k can be 1 or -1. For any given integer $n \in \{0, 1, \dots, n\}$, let K_n denote the event that there are precisely n occurrences of $+1$ in the N-tuple $(\alpha_1, \dots, \alpha_N)$. Then the uniform probability function on S assigns K_n the probability

$$\mathcal{P}(n) = 2^{-N} \frac{N!}{n!\ (N-n)!} \tag{11.7}$$

The fact that K_n has probability $\mathcal{P}(n)$ follows from (11.3).

The assignment of $\mathcal{P}(n)$ to an integer n defines a probability function on the $N+1$ element set $\{0, 1, \dots, N\}$. This is a probability function that is not uniform. We will investigate some of its properties below.

Equation (11.7) can be used to answer the first question in (11.1) with regards to our random bacteria in Section 11.1. To see how this comes about, let S denotes the sample space with the 2^N elements of the form $(\pm 1, \pm 1, \dots, \pm 1)$. Let f again denote the random variable that assigns $\alpha_1 + \cdots + \alpha_N$ to any given $(\alpha_1, \dots, \alpha_N)$. Then, the event $f = -N + 2n$ is exactly our set K_n. This understood, if $x \in \{-N, -N+2, \dots, N-2, N\}$ then

$$\text{P(the event that } f = x) = 2^{-N} \frac{N!}{\left(\frac{N+x}{2}\right)! \left(\frac{N-x}{2}\right)!}. \tag{11.8}$$

If we believe that the bacteria chooses left and right with equal probability, and that moves made in any subset of the N steps have no bearing on those made in the remaining steps, then we should be comparing our experimentally determined $p_t(x)$ with the $N = t$ version of (11.8).

The Binomial Probability Function: The probability function P in (11.7) is an example of what is called the *binomial probability* function on the set $\{0, \dots, N\}$. The 'generic' version of the binomial probability distribution requires the choice of a number, $q \in [0, 1]$. With q chosen, the probability q-version assigns the following probability to an integer n:

$$\mathcal{P}_q(n) = q^n (1-q)^{N-n}. \tag{11.9}$$

The probability function in (11.9) arises from the sample space S whose elements are the N-tuples with elements $(\pm 1, \dots, \pm 1)$. In this regard, (11.9) arises in the case that the probability of seeing $+1$ in any given entry is q and that of -1 in a given entry is $1 - q$. Here, this probability function assumes that the occurrences of ± 1 in any subset of entries must have no bearing on the appearances of ± 1 in the remaining entries. The probability function in (11.9) describes the probability in this case for the set $K_n \subset S$ of elements with n occurrences of $+1$ and $N - n$ occurrences of -1.

One can also view the probability function in (11.9) as follows: Define a random variable, g, on S, that assigns to any given element the number of appearances of $+1$. Give S the probability function just described. Then (11.9) gives the probability that $g = n$. To see why this probability is (11.9), note that the event that $g = n$ is just our set K_n. With this new probability, each element in K_n has probability $q^n(1-q)^{N-n}$. As (11.3) gives the number of elements in K_n, its probability is therefore given by (11.9).

The probability function in (11.9) is relevant to our bacterial walking scenario when we make the hypothesis that the bacteria moves to the right at any given step with probability q, thus to the left with probability $1 - q$. I'll elaborate on this in a subsequent chapter.

Here is another example: Suppose you role a six sided die N times. I now ask fix an integer n from the set $\{0, 1, \dots, N\}$ and ask for the probability that precisely $n \leq N$ of the rolls are such that the number 6 appears. If I assume that the die is fair, and that the numbers that appear on any subset of rolls have no bearing on the numbers that appear in the remaining rolls, then the probability of n occurrences of 6 in N rolls of the die is given by the $q = \frac{1}{6}$ version of (11.9). Here is why: I make an N element sequence $(\alpha_1, \dots, \alpha_N)$ with each $\alpha_k = 1$ or -1 by setting $\alpha_k = +1$ when 6 appears on the kth roll of the die, and setting $\alpha_k = -1$ when 1, 2, 3, 4, or 5 appears on the kth roll of the die. The possible sequence of this form make a 2^N element set that I call S. As the probability is $\frac{1}{6}$ for any given α_k to equal to $+1$, so a given sequence from S with precisely n occurrences of $+1$ has probability $(\frac{1}{6})^n(\frac{5}{6})^{N-n}$. Meanwhile, the number of elements in S with n occurrences of $+1$ is given by (11.3). This understood, the probability of an element in S having n occurrences of $+1$ is equal to the product of (11.3) with $(\frac{1}{6})^n(\frac{5}{6})^{N-n}$, and this is just the $q = \frac{1}{6}$ version of (11.9).

To give an example, consider rolling a standard, six-sided die. If the die is rolled once, the sample space consists of the numbers $\{1, 2, 3, 4, 5, 6\}$. If the die is 'fair', then I would want to use the probability function that assigns the version of (11.9).

Here is a final example: Consider a single stranded DNA molecule made of N bases. Suppose that the probability that the base C appears in any given position is given by some $q \in [0, 1]$. If each base is equally likely, then $q = \frac{1}{4}$ to

each element. If the bases that appear in any given subset of the N stranded molecule have no bearing on those that appear in the remaining positions, then the probability that n of the N bases are C should be given by the expression in (11.9). Note that this is a definite prediction about the DNA molecules in a given cell, and can be tested in principle by determining the percent of DNA that has the base C (cytosine). If this percentage differs significantly from $\frac{1}{4}$, then at least one of our assumptions is incorrect: Either it is not true that the appearance of C in any given position has probability $\frac{1}{4}$, or it is not true that the bases that appear in a given subset of the molecule have no bearing on those that appear in the remainder of the molecule.

I talk later about the meaning of the term 'differs significantly from'.

The Poisson Probability Function: This is a probability function on the sample space, $\mathbf{N} = \{0, 1, \ldots\}$, the non-negative integers. As you can see, this sample space has an infinite number of elements. Even so, I define a probability function on $\mathbf{N}$ to be a function, P, with $\mathrm{P}(n) \in (0, 1)$ for each n, and such that

$$\sum_{n=0,1,\ldots} \mathrm{P}(n) = \mathrm{P}(0) + \mathrm{P}(1) + \mathrm{P}(2) + \cdots \tag{11.10}$$

is a convergent series with limit equal to 1.

As with the binomial probability function, there is a whole family of Poisson functions, one for each choice of a positive real number. Let $\tau > 0$ denote the given choice. The τ version of the Poisson function assigns to any given non-negative integer the probability

$$\mathrm{P}_\tau(n) = \frac{1}{n!}\tau^n e^{-\tau}. \tag{11.11}$$

You can see that $\sum_{n=0,1,\ldots} \mathrm{P}_\tau(n) = 1$ if you know about power series expansions of the exponential function. In particular, the function $\tau \to e^\tau$ has the power series expansion

$$e^\tau = 1 + \tau + \frac{1}{2}\tau^2 + \frac{1}{6}\tau^3 + \cdots + \frac{1}{n!}\tau^n + \cdots . \tag{11.12}$$

Granted (11.12), then $\sum_{n=0,1,\ldots} \frac{1}{n!}\tau^n e^{-\tau} = \left(\sum_{n=0,1,\ldots} \frac{1}{n!}\tau^n\right) e^{-\tau} = e^\tau e^{-\tau} = 1$.

The Poisson probability enters when trying to decide if an observed cluster of occurrences of a particular phenomenon is or is not due to chance. For example, suppose that on average, some number, Δ, of newborns in the United States carry a certain birth defect. Suppose that some number, n, of such births are observed in 2006. Does this constitute an unexpected clustering that should be investigated? If the defects are unrelated and if the causative agent is similar in all cases over the years, then the probability of n occurrences in a given year should be very close to the value of the $\tau = \Delta$ version of the Poisson function $\mathrm{P}_\tau(n)$.

Here is another example from epidemiology: Suppose that the university health service sees an average of 10 cases of pneumonia each winter. Granted that the conditions that prevail this coming winter are no different than those in the past, and granted that cases of pneumonia are unrelated, what is the probability of there being 20 cases of pneumonia this winter? If these assumptions are valid, then the $\tau = 10$ Poisson function should be used to compute this probability. In particular, $\mathrm{P}_{10}(20) \approx 0.0058$.

What follows is a final example, this related to discerning whether patterns are 'random' or not. Consider, for example, the sightings of 'sea monsters'. In particular, you want to know if more sea monster sightings have occurred in the Bermuda Triangle then can be accounted for by chance. One way to proceed is to catalogue all such sightings to compute the average number of sightings per day per ship. Let us denote this average by δ. Now, estimate the number of 'ship days' that are accounted for by ships while in the Bermuda Triangle. Let N denote this last number. If sightings of sea monsters are unrelated and if any two ships on any two days in any two parts of the ocean are equally likely to sight a sea monster, then the probability of n sightings in the Bermuda triangle is given by the $\tau = N\delta$ version of (11.11).

I give some further examples of how the Poisson function is used in a separate chapter.

By the way, the Poisson function is an $N \to \infty$ limit of the binomial function. To be more precise, the τ version of the Poisson probability $\mathrm{P}_\tau(n)$ is the $N \to \infty$ limit of the versions of (11.9) with q set equal to $1 - e^{-\tau/N}$. This is to say that

$$\frac{1}{n!}\tau^n e^{-\tau} = lim_{N\to\infty} \frac{N!}{n!\ (N-n)!} \left(1 - e^{-\tau/N}\right)^n e^{(-\tau/N)(N-n)}. \tag{11.13}$$

The proof that (11.13) holds takes us in directions that we don't have time for here. Let me just say that it uses the approximation to the factorial known as *Stirling's formula:*

$$n! = \sqrt{2\pi n} e^{-n} n^n \,(1 + \text{error}) \tag{11.14}$$

where the term designated as 'error' limits to zero as $n \to \infty$. For example, the ratio with $n!$ the numerator and $\sqrt{2\pi n} e^{-n} n^n$ as the denominator is as follows for various choices of n:

n	$\frac{n!}{\sqrt{2\pi n} e^{-n} n^n}$
2	7.7
5	1.016
10	1.008
20	1.004
100	1.0008

Thus, the approximation in (11.14) serves for most uses.

11.5 Means and standard deviations

Let me remind you that if P is a probability function on some subset S of the set of integers, then its mean, μ, is defined to be

$$\mu = \sum_{n\in S} n\mathrm{P}(n) \tag{11.15}$$

and the square of its standard deviation, σ, is defined to be

$$\sigma^2 = \sum_{n\in S} (n-\mu)^2 \mathrm{P}(n). \tag{11.16}$$

Here, one should keep in mind that when S is an infinite number of elements, then μ and σ are defined only when the corresponding sums on the right sides of (11.15) and (11.16) are those of convergent series. The mean and standard deviation characterize any given probability function to some extent. More to the point, both the mean and standard deviation are often used in applications of probability and statistics.

The mean and standard deviation for the binomial probability on $\{0, 1, \ldots, N\}$ are

$$\mu = Nq \qquad \text{and} \qquad \sigma^2 = Nq(1-q). \tag{11.17}$$

For the Poisson probability function on $\{0, 1, \ldots\}$, they are

$$\mu = \tau \qquad \text{and} \qquad \sigma^2 = \tau. \tag{11.18}$$

I describe a slick method for computing the relevant sums below.

To get more of a feel for the binomial probability function, note first that the mean for the $q = \frac{1}{2}$. version is $\frac{N}{2}$. This conforms to the expectation that half of the entries in the 'average' N-tuple $(\alpha_1, \ldots, \alpha_N)$ should be $+1$ and half should be -1. Meanwhile in the general version, the assertion that the mean is Nq suggests that the fraction q of the entries in the 'average' N-tuple should be $+1$ and the fraction $(1-q)$ should be -1.

By the way, one can ask for the value of n that makes $\mathcal{P}_q(n)$ largest. To see what this is, note that

$$\frac{\mathcal{P}_q(n+1)}{\mathcal{P}_q(n)} = \frac{N-n}{n+1}\frac{q}{1-q}. \tag{11.19}$$

This ratio is less than 1 if and only if

$$n < Nq - (1 - q). \tag{11.20}$$

Since $(1 - q) < 1$, this then means that the ratio in (11.19) peaks at a value of n that is within ± 1 of the mean.

In the case of the Poisson probability, the ratio $\frac{\mathcal{P}_\tau(n+1)}{\mathcal{P}_\tau(n)}$ is $\frac{1}{n+1}\tau$. If $\tau < 1$, then 0 has the greatest probability of occurring. If $\tau \geq 1$, then $\mathcal{P}_\tau(n)$ is greatest when n is the closest integer to $\tau - 1$. Thus, the mean in this case is also very nearly the integer with the greatest probability of occurring.

11.6 The Chebychev theorem

As I remarked in Chapter 5, the standard deviation indicates the extent to which the probabilities concentrate about the mean. To see this, consider the following remarkable fact:

The Chebychev Theorem. *Suppose that P is a given probability function on a subset of the integers, $\{\ldots, -1, 0, 1, \ldots\}$, one with a well defined mean μ and standard deviation σ. For any $R \geq 1$, the probability assigned to the set where $|n - \mu| > R\sigma$ is less than R^{-2}.*

For example, this says that the probability of being 2σ away from the mean is less than $\frac{1}{4}$, and the probability of being 3σ away is less than $\frac{1}{9}$.

This theorem justifies the focus in the literature on the mean and standard deviation, since knowing these two numbers gives you rigorous bounds for probabilities without knowing anything else about the probability function!!

If you remember only one thing from these lecture notes, remember the Chebychev Theorem.

Here is the proof of the Chebychev theorem: Let S denote the sample space under consideration, and let $E \subset S$ denote the set where $|n - \mu| > R\sigma$. The probability of E is then $\sum_{n \in E} \mathrm{P}(n)$). However, since $|n - \mu| > R\sigma$ for $n \in E$, one has

$$1 \leq \frac{|n - \mu|^2}{R^2\sigma^2} \tag{11.21}$$

on E. Thus,

$$\sum_{n \in E} \mathrm{P}(n) \leq \sum_{n \in E} \frac{|n - \mu|^2}{R^2\sigma^2} \mathrm{P}(n). \tag{11.22}$$

To finish the story, note that the right side of (11.22) is even larger when we allow the sum to include all points in S instead of restricting only to points in E. Thus, we learn that

$$\sum_{n \in E} \mathrm{P}(n) \leq \sum_{n} \frac{|n - \mu|^2}{R^2\sigma^2} \mathrm{P}(n). \tag{11.23}$$

The definition of σ^2 from (11.16) can now be invoked to identify the sum on the right hand side of (11.23) with R^{-2}.

Here is an example of how one might apply the Chebychev Theorem: I watch a six-sided die rolled 100 times and see the number 6 appear thirty times. I wonder how likely it is to see thirty or more appearances of the number 6 given that the die is fair. Under the assumption that the die is fair and that the faces that appear in any given subset of the 100 rolls have no bearing on those that appear in the remaining rolls, I can compute this probability using the $q = \frac{1}{6}$ version of the $N = 100$ binomial probability function by computing the sum $\mathcal{P}_{1/6}(30) + \mathcal{P}_{1/6}(31) + \cdots + \mathcal{P}_{1/6}(100)$. Alternately, I can get an upper bound for this probability by using the Chebychev Theorem. In this regard, I note that the mean of the $q = \frac{1}{6}$ and $N = 100$ binomial probability function is $\frac{100}{6} = \frac{50}{3} = 16\ \frac{2}{3}$ and the standard deviation is $\frac{5}{3}\sqrt{5} \approx 3.73$. Now, 30 is $13\ \frac{1}{3}$ from the mean, and this is ≈ 3.58 standard deviations. Thus, the probability of seeing thirty or more appearances of the number 6 in 100 rolls is no greater than $(3.58)^{-2} \approx 0.08$. As it turns out, the probability of seeing 6 appear thirty or more times is a good deal smaller than this.

It is usually the case that the upper bound given by the Chebychev Theorem is much greater than the true probability. Even so, the Chebychev Theorem is one of the most important things to remember from this course since it allows you to compute *some* upper bound with very little detailed knowledge of the probability function. Just two numbers, the mean and standard deviation, give you an upper bound for probabilities.

11.7 Characteristic functions

The slick computation of the mean and standard deviation that I mentioned involves the introduction of the notion of the *characteristic function*. The latter is a bona fide function on the real numbers that is determined by a given probability function on any sample space that is a subset of the non-negative integers, $\{0, 1, \dots\}$. This characteristic function encodes all of the probability function. Note, however, that this function is useful for practical purposes only to the extent that it isn't some complicated mess.

Before turning to the definition in abstract, consider the special case where the subset is $\{0, 1, 2, \dots, N\}$ and the probability function is the q binomial function from (11.9). In this case, the characteristic function is the function of x given by

$$\mathscr{P}(x) = (qx + (1 - q))^N . \tag{11.24}$$

Here is the connection between $\mathscr{P}(x)$ and the binomial probability function: Expand the product so as to write $\mathscr{P}$ as a polynomial of degree N in x. As I argue below, you will find that

$$\mathscr{P}(x) = \mathcal{P}_q(0) + x\mathcal{P}_q(1) + x^2\mathcal{P}_q(2) + \cdots + x^N\mathcal{P}_q(N). \tag{11.25}$$

Thus, the coefficient of x^n of this polynomial representation of $\mathscr{P}$ gives us the probabilities that $\mathcal{P}_q$ assigns to the integer n.

To see why (11.25) is the same as (11.24), consider multiplying out an N-fold product of the form:

$$(a_1x + b_1)(a_2x + b_2) \cdots (a_Nx + b_N) \tag{11.26}$$

A given term in the resulting sum can be labeled as $(\alpha_1, \dots, \alpha_N)$ where $\alpha_k = 1$ if the kth factor in (11.26) contributed a_kx, while $\alpha_k = -1$ if the kth factor contributes b_k. The power of x for such a term is equal to the number of α_k that are $+1$. This is $\frac{N!}{n!\,(N-n)!}$, the number of elements in the set K_n that appears in (11.3). Thus, $\frac{N!}{n!\,(N-n)!}$ terms contribute to the coefficient of x^n in (11.26). In the case of (11.25), all versions of a_k are equal to q, and all versions of b_k are equal to $(1 - q)$, so each term that contributes to x^n is $q^n(1 - q)^{N-n}x^n$. As there are $\frac{N!}{n!\,(N-n)!}$ of them, the coefficient of x^n in (11.24) is $\mathcal{P}_q(n)$ as (11.24) claims.

In general the characteristic function for a probability function, P, on a subset of $\{0, 1, \dots\}$ is the polynomial or infinite power series in x for which the coefficient of x^n is $\mathrm{P}(n)$. This is to say that

$$\mathscr{P}(x) = \mathrm{P}(0) + \mathrm{P}(1)x + \mathrm{P}(2)x^2 + \cdots = \sum_{n=0,1,2,\dots} \mathrm{P}(n)x^n. \tag{11.27}$$

This way of coding the probability function P is practical only to the extent that the series in (11.27) sums to a relatively simple function. I gave the q-binomial example above. In the case of the uniform probability function on $\{0, \dots, N\}$, the sum in (11.27) is the function $\frac{1}{N+1}\frac{1-x^{N+1}}{1-x}$. I argue momentarily that the τ version of the Poisson probability function on $\{0, 1, 2, \dots\}$ gives

$$\mathscr{P}(x) = e^{(x-1)\tau}. \tag{11.28}$$

In general, there are two salient features of a characteristic polynomial: First, the values of $\mathscr{P}$, its derivative, and its

second derivative at $x = 1$ are:

$$\begin{aligned} &\bullet\ \mathscr{P}(1) = 1 \\ &\bullet\ \left(\frac{d}{dx}\mathscr{P}\right)\Big|_{x=1} = \mu \\ &\bullet\ \left(\frac{d^2}{dx^2}\mathscr{P}\right)\Big|_{x=1} = \sigma^2 + \mu(\mu-1) \end{aligned} \tag{11.29}$$

Second, the values of the value of $\mathscr{P}$, its derivative and its higher order derivatives at $x = 0$ determine P since

$$\frac{1}{n!}\left(\frac{d^n}{dx^n}\mathscr{P}\right)\Big|_{x=0} = \mathrm{P}(n). \tag{11.30}$$

This is how the function $x \to \mathscr{P}(x)$ encodes all of the information that can be obtained from the given probability function P.

To explain how (11.29) follows from the definition in (11.27), note first that $\mathscr{P}(1) = \mathrm{P}(0)+\mathrm{P}(1)+\cdots$, and this is equal to 1 since the sum of the probabilities must be 1. Meanwhile, the derivative of $\mathscr{P}$ at $x = 1$ is $1\cdot\mathrm{P}(1) + 2\cdot\mathrm{P}(2) + \cdots$, and this is sum is the definition of the mean μ. With the help of (11.16), a very similar argument establishes the third point in (11.29).

I can use (11.29) to get my slick calculations of the mean and standard deviations for the binomial probability function. In this case, $\mathscr{P}$ is given in (11.24) and so

$$\frac{d}{dx}\mathscr{P} = Nq\left(qx + (1-q)\right)^{N-1}. \tag{11.31}$$

Set $x = 1$ here to find the mean equal to Nq as claimed. Meanwhile

$$\frac{d^2}{dx^2}\mathscr{P} = N(N-1)q^2\left(qx + (1-q)\right)^{N-1}, \tag{11.32}$$

Set x = 1 here finds the right-hand side of (11.32) equal to $N(N-1)q^2$. Granted this and the fact that $\mu = qN$, then the third point in (11.29) find σ^2 equal to

$$N(N-1)q^2 - N^2q^2 + Nq = Nq(1-q), \tag{11.33}$$

which is the asserted value.

For the Poisson probability function, the characteristic polynomial is the infinite power series

$$\mathrm{P}_\tau(0) + x\mathrm{P}_\tau(1) + x^2\mathrm{P}_\tau(2) + \cdots = \left(1 + x\tau + \frac{1}{2}x^2\tau^2 + \frac{1}{6}x^3\tau^3 + \cdots + \frac{1}{n!}x^n\tau^n + \cdots\right)e^{-\tau}. \tag{11.34}$$

As can be seen by replacing τ in (11.12) with $x\tau$, the sum on the right here is $e^{x\tau}$. Thus, $\mathscr{P}(x) = e^{(x-1)\tau}$ as claimed by (11.28). Note that the first and second derivatives of this function at $x = 1$ are both equal to τ. With (11.29), this last fact serves to justify the claim that the both the mean and standard deviation for the Poisson probability are equal to τ.

The characteristic polynomial for a probability function is used to simplify seemingly hard computations in the manner just illustrated in the cases where the defining sum in (11.27) can be identified with the power series expansion of a well known function. The characteristic function has no practical use when the power series expansion is not that of a simple function.

11.8 Loose ends about counting elements in various sets

My purpose in this last section is to explain where the formula in (11.3) and (11.5) come from. To start, consider (11.5). There are b choices for β_1. With β_1 chosen, there are $b-1$ choices for β_2, one less than that for β_1 since we are not

allowed to have these two equal. Given choices for β_1 and β_2, there are $b-2$ choices for β_3. Continuing in this vein finds $b-k$ choices available for β_{k+1} if $(\beta_1,\dots,\beta_k)$ have been chosen. Thus, the total number of choices is $b(b-1)\cdots(b-N+1)$, and this is the claim in (11.5).

To see how (11.3) arises, let me introduce the following notation: Let $m_n(N)$ denote the number of elements in the $(\alpha_1,\dots,\alpha_N)$ version of K_n. If we are counting elements in this set, then we can divide this version of K_n into two subsets, one where $\alpha_1=1$ and the other where $\alpha_1=-1$. The number of elements in the first is $m_{n-1}(N-1)$ since the $(N-1)$-tuple $(\alpha_2,\dots,\alpha_N)$ must have $n-1$ occurrences of $+1$. The number in the second is $m_n(N-1)$ since in this case, the $(N-1)$-tuple $(\alpha_2,\dots,\alpha_N)$ must have all of the n occurrences of $+1$. Thus, we see that

$$m_n(N) = m_{n-1}(N-1) + m_n(N-1). \tag{11.35}$$

I will now make this last formula look like a matrix equation. For this purpose, fix some integer $T\geq 1$ and make a T-component vector, $\vec{m}(N)$, whose coefficients are the values of m_n for the cases that $1\leq n\leq T$. This equation asserts that $\vec{m}(N)=A\vec{m}(N-1)$ where A is the matrix with $A_{k,k}$ and $A_{k,k-1}$ both equal to 1 and all other entries equal to zero. Iterating the equation $\vec{m}(N)=A\vec{m}(N-1)$ finds

$$\vec{m}(N) = A^{N-1}\vec{m}(1), \tag{11.36}$$

where $\vec{m}(1)$ is the vector with top component 1 and all others equal to zero.

Now, we don't have the machinery to realistically compute A^{N-1}, so instead, lets just verify that the expression in (11.3) gives the solution to (11.35). In this regard, note that $\vec{m}(N)$ is completely determined from $\vec{m}(1)$ using (11.36), and so if we believe that we have a set $\{\vec{m}(1),\vec{m}(2),\dots\}$ of solutions, then we need only plug in our candidate and see if (11.36) holds. This is to say that in order to verify that (11.3) is the correct, one need only check that the formula in (11.36) holds. This amounts to verifying that

$$\frac{N!}{n!\,(N-n)!} = \frac{(N-1)!}{(n-1)!\,(N-n)!} + \frac{(N-1)!}{n!\,(N-n-1)!}. \tag{11.37}$$

I leave this task to you.

11.9 A Nobel Prize for the clever use of statistics

The 1969 Nobel Prize for Physiology and Medicine was given to Max Delbruck and Salvador Luria for work that had a crucial statistical component. What follows is a rough description of their Nobel Prize winning work.

Prior to Delbruck and Luria's Nobel Prize winning work[1] in 1943, there were two popular explanations for the evolution of new traits in bacteria. Here is the first: Random mutations occur as reproduction proceeds in a population. Therefore, no population of size greater than 1 is completely homogeneous. As time goes on, environmental stresses favor certain variations over others, and so certain *pre-existing* variants will come to dominate any given population. This view of evolution is essentially that proposed by Darwin. There is, in contrast, the so called Lamarckian proposal: Environmental stresses cause *new* traits to arise amongst some individuals in any given population, and these are subsequently favored. What follows summarizes these two proposals:

Darwin: *Evolution occurs through the selection due to environmental stresses of the most favorable traits from a suite of pre-existing traits.*

Lamarck: *Evolution occurs when environmental stresses force new, more favorable traits to arise.*

Max Delbruck and Salvador Luria shared the 1969 Nobel Prize in Physiology and Medicine for an experiment, done in 1943, that distinguished between these two proposals as explanations for the evolution of immunity in bacterium. The results conclusively favored Darwin over Lamarck.

[1] Luria, SE, Delbruck, M., "Mutations of Bacteria from Virus Sensitivity to Virus Resistance" Genetics **28** (1943) 491-511.

What follows is an idealized version of what Luria and Delbruck did. They were working with a species of bacteria that was susceptible to a mostly lethal virus. (A virus that attacks bacteria is called a 'phage'.) They started a large number of colonies, each with one of these bacteria, and fed these colonies well for some length of time T. Here, it is convenient to measure time in units so that a change of one unit of time is the mean time between successive generations. With these units understood, each colony contained roughly $\mathcal{N} = 2^T$ bacteria. Each of their colonies was then exposed to the virus. Shortly after exposure to the virus, the number of members in each colony were counted. I use $\mathcal{K}$ to denote the number of colonies. The count of the number of living bacteria in each colony after phage infection gives a list of numbers that I write as $(z_1, z_2, \dots, z_{\mathcal{K}})$. This set of numbers constitutes the experimental data. With regards to this list, it is sufficient for this simplified version of the story to focus only on

$$\mu_{\text{exp}} = \frac{1}{\mathcal{K}} \sum_{1 \le j \le \mathcal{K}} z_j \qquad \text{and} \qquad \sigma^2_{\text{exp}} = \frac{1}{\mathcal{K}} \sum_{1 \le j \le \mathcal{K}} (z_j - \mu_{\text{exp}})^2. \tag{11.38}$$

Luria and Delbruck realized that μ_{exp} and σ^2_{exp} can distinguish between the Darwin and Lamarck proposals. To explain their thinking, consider first what one would expect were the Lamarck proposal true. Let p denote the probability that exposure to the virus causes a mutation in any given bacterium that allows the bacterium to survive the infection. The probability of n surviving bacteria in a colony with $\mathcal{N}$ members should be given by the binomial probability function from (11.9) as defined using p and $\mathcal{N}$ in lieu of q and N. Thus, this probability is $\frac{\mathcal{N}!}{n!\,(\mathcal{N}-n)!} p^n (1-p)^{\mathcal{N}-n}$. In particular, the mean number of survivors should be $p\mathcal{N}$ and the square of the standard deviation should be $p(1-p)\mathcal{N}$. This is to say that the Lamarck proposal predicts that the experimental data

$$\mu_{\text{exp}} \approx pN \qquad \text{and} \qquad \sigma^2_{\text{exp}} \approx p(1-p)N. \tag{11.39}$$

Not knowing the value for p, one can none the less say that the Lamarck proposal predicts the following:

$$\frac{\sigma^2_{\text{exp}}}{\mu_{\text{exp}}} \approx 1 - p. \tag{11.40}$$

Note in particular that this number is independent of $\mathcal{N}$ and $\mathcal{K}$; and in any event it is very close to 1 when p is small. Small p is expected.

Now consider the Darwinian proposal: In this case, there is some small, but non-zero probability, I'll call it $\mathfrak{p}$, for the founding bacterium of any given colony to have a mutation that renders it immune to the virus. If the founder of a colony has this mutation, then it is unaffected by the virus as are all of its descendents. So, a colony with an immune founder should have population $2^T = \mathcal{N}$ after viral attack. If a colony is founded by a bacterium without this mutation, then its population after the viral attack should be very small. Thus, in this ideal situation, the Darwinian proposal predicts that each z_j should be either nearly zero or nearly $\mathcal{N}$.

If $\mathcal{K}$ colonies are started, then the probability that n of them are founded by an immune bacterium is given by the binomial probability function from (11.9) as defined now using $\mathfrak{p}$ and $\mathcal{K}$ in lieu of q and N. Thus, this probability is $\frac{\mathcal{K}}{n!\,(\mathcal{K}-n)!} \mathfrak{p}^n (1-\mathfrak{p})^{\mathcal{K}-n}$. Note that the mean of this is $\mathfrak{p}\mathcal{K}$ and the square of the standard deviation is $\mathfrak{p}(1-\mathfrak{p})\mathcal{K}$. As a consequence, the Darwin proposal predicts that μ_{exp} and σ^2_{exp} should be roughly

$$\begin{aligned} &\bullet\ \mu_{\text{exp}} \approx \sum_{0 \le n \le \mathcal{K}} (\text{Probability of } n \text{ immune founders})(\mathcal{K}^{-1} n \mathcal{N}) \\ &\bullet\ \sigma^2_{\text{exp}} \approx \sum_{0 \le n \le \mathcal{K}} (\text{Probability of } n \text{ immune founders})(\mathcal{K}^{-1} n \mathcal{N}^2) - \mu^2_{\text{exp}}. \end{aligned} \tag{11.41}$$

These sums give the predictions $\mu_{\text{exp}} \approx \mathfrak{p}\mathcal{N}$ and $\sigma^2_{\text{exp}} \approx \mathfrak{p}(1-\mathfrak{p})\mathcal{N}^2$. As a consequence, the Darwin proposal predicts the ratio

$$\frac{\sigma^2_{\text{exp}}}{\mu_{\text{exp}}} \approx (1-\mathfrak{p})\mathcal{N}. \tag{11.42}$$

Note in particular that if $\mathfrak{p}$ is small, then this number is roughly $\mathcal{N} = 2^T$ while that in (11.40) is roughly 1.

Delbruck and Luria saw statistics that conclusively favored the Darwin proposal over the Lamarck one.

11.10 Exercises:

1. Let A denote the 4×4 version of the matrix in (11.36). Thus,

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

   (a) Present the steps of the reduced row echelon reduction of A to verify that it is invertible.

   (b) Find A^{-1} using Fact 2.3.5 of Bretscher's book Linear Algebra with Applications.

2. Let τ denote a fixed number in $(0, 1)$. Now define a probability function, P, on the set $\{0, 1, 2, \ldots\}$ by setting $\mathrm{P}(n) = (1 - \tau)\tau^n$.

   (a) Verify that $\mathrm{P}(0) + \mathrm{P}(1) + \cdots = 1$, and thus verify that P is a probability function.
   (If you haven't seen how to do this, set $V(n) = 1+\tau+\tau^2+\cdots+\tau^n$. Prove that $\tau V(n) = V(n)-1+\tau^{n+1}$. Then rearrange things to write $V(n)(1-\tau) = (1-\tau^{n+1})$. Now solve for $V(n)$ and consider what happens when $n \to \infty$.)

   (b) Sum the series $\mathrm{P}(0) + x\mathrm{P}(1) + x^2\mathrm{P}(2) + \cdots$ to verify that the characteristic function is the $\mathscr{P}(x) = \frac{1-\tau}{1-x\tau}$.

   (c) Use the formula in (11.28) to compute the mean and standard deviation of P.

   (d) In the case $\tau = \frac{1}{2}$, the mean is 1 and the standard deviation is $\sqrt{2}$. As $6 \geq \mu+3\sigma$, the Chebychev Theorem asserts that the probability for the set $\{6, 7, \ldots\}$ should be less than $\frac{1}{9}$. Verify this prediction by summing $\mathrm{P}(6) + \mathrm{P}(7) + \cdots$.

   (e) In the case $\tau = \frac{2}{3}$, the mean is 2 and $\sigma = \sqrt{6}$. Verify the prediction of the Chebychev Theorem that $\{7, 8, \ldots\}$ has probability less than $\frac{1}{4}$ by computing the sum $\mathrm{P}(7) + \mathrm{P}(8)) + \cdots$.

3. This exercise fills in some of the details in the verification of (11.3).

   (a) Multiply both sides of (11.37) by $(n-1)!(N-n-1)!$ and divide both sides of the result by $(N-1)!$ Give the resulting equation.

   (b) Use this last result to verify (11.37).

4. Suppose that I have a bag with 6 red grapes and 5 green ones. I reach in with my eyes closed and pick a grape at random. After looking at its color, I replace the grape, shake up the bag and redo this experiment, 10 times in all. Let n be an integer between 0 and 10. Assume that any given grape is chosen at random each time, and that my choices in any given subset of the experiments have no bearing on those of the remaining experiments. What is the probability of choosing exactly n green grapes in the 10 experiments?

5. The purpose of this exercise is to explore the $q = \frac{1}{2}$ version of the binomial probability function, thus the function $\mathcal{P}(n)$ that appears in (11.7).

   (a) As explained in the text, $\mathcal{P}(n)$ is largest when $n = \frac{N}{2}$ in the case that N is even. Use Stirling's formula to justify the claim that $\mathcal{P}(\frac{N}{2}) \approx \sqrt{\frac{2}{\pi N}}$ when N is even and very large. Compare this last number with the true value given by (11.7) for $N = 10$ and $N = 100$.

   (b) Suppose that N is a positive, even integer. The mean for the $q = \frac{1}{2}$ binomial probability function for the set $\{0, 1, \ldots, N\}$ is $\mu = \frac{1}{2}N$ and the standard deviation is $\sigma = \frac{1}{2}\sqrt{N}$. Use the Chebychev theorem in the cases $N = 36$ and 64 to find an upper bound for the probability of the set of $n \leq 10$. Then, use a calculator to compute this probability using the formula in (11.7).

6. The purpose of this problem is to explore the Poisson probability function in the context of discerning patterns. Suppose that I am a police captain in Boston in charge of a 10 block by 10 block neighborhood. Thus a neighborhood of 100 squares, each 1 block on a side. I note that in over the years, my force was asked to respond to an average of 500 calls from the neighborhood, thus an average of 5 per square block per year. The

police station is in the square at one corner of the neighborhood. There were no calls from this square last year. The square furthest from the police station recorded 15 calls. Under the assumption that a call is as likely to come from any one square as any other, and that whether or not a call has come in from a square has no bearing on when the next call will come in from that square or any other square, give

(a) the probability that there are no calls from the police station's square, and

(b) the probability that there are 15 calls from the square furthest from the police station's square.

7. This problem is a continuation of the preceding one. The purpose here is to let you see how an event with small probability to occur in any one place can have a reasonably large probability of appearing somewhere. In particular, the goal for this exercise is to compute the probability that there is at least one square with no calls. For this purpose, let q denote your answer to Problem 6a.

   (a) Explain why the probability that every square has at least one call is $(1-q)^{100}$.

   (b) Explain why the probability that at least one square has no calls is $1-(1-q)^{100}$.

8. This is a problem whose purpose is to get you to think about counting and probability. The point is that being able to make an accurate count of the various ways that a given outcome can arise is crucial to making a correct calculation of probabilities.

   Here is the scenario: In baseball's World Series, the two league champions in a given year meet to see which team is the best in baseball for that year. The teams play games sequentially until one or the other wins four games. The team that wins four games is declared the World Champion. Thus, the World Series can involve anywhere from 4 to 7 games. Since 1923, there have been 16 years where the series ended after 4 games, 16 years when it ended after 5 games, 18 years when it ended after 6 games and 33 years when it ended after 7 games. See `http://mlb.mlb.com/NASApp/mlb/mlb/history/postseason/mlb_ws.jsp?feature=recaps_index`.

   Thus, the relative frequencies for 4, 5, 6 and 7 game series are:

$$\begin{aligned} \text{4 games:} &\quad \tfrac{16}{83} \approx 0.19 \\ \text{5 games:} &\quad \tfrac{16}{83} \approx 0.19 \\ \text{6 games:} &\quad \tfrac{18}{83} \approx 0.22 \\ \text{7 games:} &\quad \tfrac{33}{83} \approx 0.40 \end{aligned}$$

   Here is a question: What is the expected frequency of 4, 5, 6 and 7 game series if we assume that the two teams have a 50-50 chance of winning any given game, and if we assume that the event of winning any one game is independent of the event of winning any other. To answer this question, let A and N denote the two teams.

   (a) What is the probability for A to win the first 4 games? Likewise, what is the probability for N?

   (b) Explain why the probability for the World series to last 4 games is $\frac{1}{8} = 0.125$.

   (c) Explain why the probability for the World Series to last 5 games is equal to

$$2\binom{4}{3}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right) = \frac{1}{4} = 0.25.$$

   (d) Explain why the probability for the World Series to last 6 games is

$$2\binom{5}{3}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right) = \frac{5}{16} = 0.3125.$$

   (e) Explain why the probability for the World Series to last 7 games is

$$2\binom{6}{3}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right) = \frac{5}{16} = 0.3125.$$

CHAPTER

TWELVE

P-values

My purpose in this chapter is to describe a criterion for deciding when data is "surprising". For example, suppose that you flip a coin 100 times and see 60 heads. Should you be surprised? Should you question your belief that the coin is fair? More generally, how many heads should appear before you question the coin's fairness? These are the sorts of questions that I will address below.

12.1 Point statistics

Suppose that you do an experiment N times and find that a certain event occurs m times out of the N experiments. Can one determine from this data a probability function for the event to occur? For example, I flip a coin 100 times and see 60 heads. Can I determine a probability function for heads to appear?

If we assume that the N experiments are identical in set up, and that the appearance of the event in any one has no bearing on its appearance in any other, then we can propose the following hypothesis: The event occurs in any given experiment with probability q (to be determined) and so the probability that some $n \leq N$ events occurs in N experiments is given by the q-version of the binomial function, thus

$$\mathrm{P}_q(n) = \frac{N!}{n!(N-n)!} q^n (1-q)^{N-n}. \tag{12.1}$$

The question now is: What value should be used for q?

The use of experimental data to estimate a single parameter – q in this case – is an example of what are called *point statistics*. Now, it is important for you to realize that there are various ways to obtain a "reasonable" value to use for q. Here are some:

- Since we actually found m events in N trials, take the value of q that gives m for the mean of the probability function in (12.1). With reference to (11.17) in the previous chapter, this choice for q is $\frac{m}{N}$.
- Take q to so that $n = m$ is the integer with the maximal probability. If you recall (11.19) from the previous chapter, this entails taking q so that both

  $$\frac{\mathrm{P}_q(m)}{\mathrm{P}_q(m+1)} > 1 \qquad \text{and} \qquad \frac{\mathrm{P}_q(m-1)}{\mathrm{P}_q(m)} < 1. \tag{12.2}$$

  This then implies that $\frac{m}{N+1} < q < \frac{m+1}{N+1}$. Note that $q = \frac{m}{N}$ satisfies these conditions.

With regards to my example of flipping a coin 100 times and seeing 60 heads appear, these arguments would lead me to postulate that the probability of seeing some number, n, of heads, is given by the $q = 0.6$ version of (12.1).

12.2 P-value and bad choices

A different approach asks for the *bad* choices of q rather than the "best" choice. The business of ruling out various choices for q is more in the spirit of the scientific method. Moreover, giving the unlikely choices for q is usually much more useful to others than simply giving your favorite candidate. What follows explains how statisticians often determine the likelihood that a given choice for q is realistic.

For this purpose, suppose that we have some value for q in mind. There is some general agreement that q is not a reasonable choice when the following occurs:

> *There is small probability as computed by the q-version of* (12.1) *of there being the observed m occurrences of the event of interest.*

To make the notion of *small probability* precise, statisticians have introduced the notion of the "P-value" of a measurement. This is defined with respect to some hypothetical probability function, such as our q-version of (12.1). In our case, the P-value of m is the probability for the subset of numbers $n \in \{0, 1, \ldots, N\}$ that are at least as far from the mean, qN, as is m. For example, m has P-value $\frac{1}{2}$ in the case that (12.1) assigns probability $\frac{1}{2}$ to the set of integers n that obey $|n - Nq| \geq |m - Nq|$. A P-value that is less than 0.05 is deemed *significant* by statisticians. This is to say that if m has such a P-value, then q is deemed unlikely to be incorrect.

In general, the definition of the P-value for any measurement is along the same lines:

Definition: *Suppose that a probability function on the set of possible measurements for some experiment is proposed. The P-value of any given measurement is the probability for the subset of values that lie as far or farther from the mean of this probability function than the given measurement. The P-value is deemed significant if it is smaller than* 0.05.

Note here that this definition requires computing the probability of being both greater than the mean and less than then the mean. There is also the notion of a *one-sided* P-value. This computes the probability of being on the same side of the mean as the observed value, and at least as far from the mean as the observed value. Thus, if the observed value is greater than the mean, the 1-sided P-value computes the probability that a measurement will be as large or larger than the observed value. If the observed value is less than the mean, then the 1-sided P-value computes the probability of being as small or smaller than the observed value. There are other versions of P-value used besides that defined above and the 1-sided P-values. However, in these lecture notes, P-value means what is written in the preceding definition. But, keep in mind when reading the literature or other texts that there are alternate definitions.

Consider now the example where I flip a coin 100 times and find 60 heads. If I want to throw doubt on the hypothesis that the coin is fair, I should compute the P-value of 60 using the $q = \frac{1}{2}$ version of the binomial probability function in (12.1). This means computing $P_{1/2}(60) + P_{1/2}(61) + \cdots + P_{1/2}(100) + P_{1/2}(40) + P_{1/2}(39) + \cdots + P_{1/2}(1)$. The latter sum gives the probability of seeing a number of heads appear that is at least as far from the mean, 50, as is 60. My calculator finds that this sum is 0.057. Thus, the P-value of 60 is greater than 0.05 and so I hesitate to reject the hypothesis that the coin is fair.

An *upper bound* for the P-value that uses only the mean and standard deviation of a probability function can be had using the Chebychev theorem in Chapter 11. As you should recall, this theorem asserts that the probability of finding a measurement with distance $R\sigma$ from the mean is less than R^{-2}. Here, σ denotes the standard deviation of the given probability function. Granted this, a measurement that differs from the mean by 5σ or more has probability less than 0.04 and so has a significant P-value. Such being the case, the 5σ bound is often used in lieu of the 0.05 bound.

To return to our binomial case, to say that m differs from the mean, Nq, by at least 5σ, is to say that

$$|m - Nq| \geq 5(Nq(1-q))^{1/2}. \tag{12.3}$$

We would consider q to be a "bad" choice in the case that (12.3) holds.

In the case of my coin flip example, with $N = 100$ and $q = \frac{1}{2}$, the standard deviation is 5. Thus, 60 heads is only 2 standard deviations from the mean, which is less than the 5σ criterion used above.

When using the Chebychev theorem to get an upper bound for a P-value, remember the following:

If an outcome differs from the mean by 5σ or more, then it's P-value is necessarily smaller than 0.05.

The converse of this last statement is not true. An event can have P-value that is smaller than 0.05 yet differ from the mean by less than 5σ. Take for example the case where I flip a coin 100 times and now see 61 heads instead of 60. Under the assumption that the coin is fair, the P-value of 61 heads is 0.0452. This is less than 0.05. However, 61 is 2.2σ from the mean.

12.3 A binomial example using DNA

As you may recall, a strand of a DNA molecule consists of a chain of smaller molecules tied end to end. Each small molecule in the chain is one of four types, these labeled A, T, G and C. Suppose we see that G appears some n times on some length N strand of DNA. Should this be considered unusual?

To make this question precise, we have to decide how to quantify the term *unusual*, and this means choosing a probability function for the sample space whose elements consist of all length N strings of letters, where each letter is either A, C, G or T. For example, the assumption that the appearances of any given molecule on the DNA strand are occurring at random suggests that we take the probability of any given letter A, C, G or T appearing at any given position to be $\frac{1}{4}$. Thus, the probability that G does not appear at any given location is $\frac{3}{4}$, and so the probability that there are n appearances of G in a length N string (if our random model is correct) would be given by the $q = \frac{3}{4}$ version of the binomial function in equation (12.1).

This information by itself is not too useful. A more useful way to measure whether n appearances of G is unusual is to ask for the probability in our $q = \frac{1}{4}$ binomial model for more (or less) appearances of G to occur. This is to say that if we think that there are too many G's for the appearance to be random then we should consider the probability *as determined by our $q = \frac{1}{4}$ binomial function* of there being at least this many G's appearing. Thus, we should be computing the P-value of the measured number, n. In the binomial case with $q = \frac{1}{4}$, this means computing

$$\sum_{k\in B} \frac{N!}{k!(N-k)!} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{N-k}, \tag{12.4}$$

where the sum is over all integers k from the set, B, of integers in $\{0, \dots, N\}$ that obey $|b - \frac{N}{4}| \geq |n - \frac{N}{4}|$.

Suppose, for instance that we have a strand of $N = 20$ bases and see 10 appearances of G. Does this suggest that our hypothesis is incorrect about G's appearance being random in the sense just defined? To answer this question, I would compute the P-value of 10. This means computing the sum in (12.4) with $N = 20$ and B the subset in $\{0, \dots, 20\}$ of integers b with either $b = 0$ or $b \geq 10$. As it turns out, this sum is close to 0.017. The P-value is less than 0.05 and so we are told to doubt the hypothesis that the G appears at random.

As the sum in (12.4) can be difficult to compute in any given case, one can often make due with the upper bound for the P-value that is obtained from the Chebychev theorem. In this regard, you should remember the Chebychev theorem's assertion that the probability of being R standard deviations from the mean is less than R^{-2}. Thus, Chebychev says that a measurement throws significant doubt on a hypothesis when it is at 5 or more standard deviations from the mean.

In the case under consideration in (12.4), the standard deviation, σ, is $\frac{\sqrt{3N}}{4}$. In this case, the Chebychev theorem says that the set of integers b that obey $|b - \frac{N}{4}| > R\frac{\sqrt{3N}}{4}$ has probability less than R^{-2}. Taking $R = 5$, we see that a value for n has P-value less that 0.05 if it obeys $|n - \frac{N}{4}| \geq \frac{5}{4}\sqrt{3N}$. This result in the DNA example can be framed as follows: The measured fraction, $\frac{n}{N}$, of occurrences of G has significant P-value in our $q = \frac{1}{4}$ random model if

$$\left|\frac{n}{N} - \frac{1}{4}\right| > \frac{5}{4}\sqrt{\frac{3}{N}}. \tag{12.5}$$

You should note here that as N gets bigger, the right-hand side of this last inequality gets smaller. Thus, as N gets bigger, the experiment must find the ratio $\frac{n}{N}$ ever closer to $\frac{1}{4}$ so as to forestall the death of our hypothesis about the random occurrences of the constituent molecules on the DNA strand.

As I wrote earlier, the Chebychev theorem gives only an upper bound for the P-value. Indeed, in the example where $N = 20$ and $n = 10$, the actual P-value was found to be 0.017. Even so, the inequality in (12.5) is not obeyed since the left-hand side is $\frac{1}{4}$ and the right-hand side is roughly 0.48. Thus, the Chebychev upper bound for the P-value is larger than 0.05 even though the true P-value is less than 0.05.

12.4 An example using the Poisson function

All versions of the Poisson probability function are defined on the set of non-negative integers, $\mathbf{N} = \{0, 1, 2, \ldots\}$. As noted in the previous chapter, a particular version is determined by a choice of a positive number, τ. The Poisson probability for the given value of τ is:

$$P_\tau(n) = \frac{1}{n!}\tau^n e^{-\tau} \tag{12.6}$$

Here is a suggested way to think about P_τ:

$$\text{$P_\tau(n)$ gives the probability of seeing n occurrences of a particular event in any given unit time interval when the occurrences are unrelated and they average τ per unit time.} \tag{12.7}$$

What follows is an example that doesn't come from biology but is none-the-less dear to my heart. I like to go star gazing, and over the years, I have noted an average of 1 meteor per night. Tonight I go out and see 5 meteors. Is this unexpected given the hypothesis that the appearances of any two meteors are unrelated? To test this hypothesis, I should compute the P-value of $n = 5$ using the $\tau = 1$ version of (12.6). Since the mean of P_τ is τ, this involves computing

$$\left(\sum_{m\geq 5}\frac{1}{m!}\right)e^{-1} = 1 - \left(1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24}\right)e^{-1} \tag{12.8}$$

My trusty computer can compute this, and I find that $P(5) \leq 0.004$. Thus, my hypothesis of the unrelated and random occurrence of meteors is unlikely to be true.

What follows is an example from biology, this very relevant to the theory behind the "genetic clocks" that predict the divergence of modern humans from an African ancestor some $100,000$ years ago. I start the story with a brief summary of the notion of a *point mutation* of a DNA molecule. This occurs as the molecule is copied for reproduction when a cell divides; it involves the change of one letter in one place on the DNA string. Such changes, cellular typographical errors, occur with very low frequency under non-stressful conditions. Environmental stresses tend to increase the frequency of such mutations. In any event, under normal circumstances, the average point mutation rate per site on a DNA strand, per generation has been determined via experiments. Let μ denote the latter. The average number of point mutations per generation on a segment of DNA with N sites on it is thus μN. In $T \geq 1$ generations, the average number of mutations in this N-site strand is thus μNT.

Now, make the following assumptions:

- The occurrence of any one mutation on the given N-site strand has no bearing on the occurrence of another.
- Environmental stresses are no different now than in the past, (12.9)
- The strand in question can be mutated at will with no effect on the organism's reproductive success.

Granted the latter, the probability of seeing n mutations in T generations on this N-site strand of DNA is given by the $\tau = \mu NT$ version of the Poisson probability:

$$\frac{1}{n!}(\mu NT)^n e^{\mu NT}. \tag{12.10}$$

The genetic clock idea exploits this formula in the following manner: Suppose that two closely related species diverged from a common ancestor some unknown number of generations in the past. This is the number we want to estimate. Call it R. Today, a comparison of the N site strand of DNA in the two organisms finds that they differ by mutations at n sites. The observed mutations have arisen over the course of $T = 2R$ generations. That is, there are R generations worth of mutations in the one species and R in the other, so $2R$ in all. We next say that R is a reasonable guess if the $\tau = \mu N(2R)$ version of the Poisson function gives n any P-value that is *greater* than 0.05. For this purpose, remember that the mean of the τ version of the Poisson probability function is τ.

We might also just look for the values of R that make n within 5 standard deviations of the mean for the $\tau = \mu N(2R)$ version of the Poisson probability. Since the square of the standard deviation of the τ version of the Poisson probability function is also τ, this is equivalent to the demand that $|2\mu NR - n| \leq 5\sqrt{2\mu NR}$. This last gives the bounds

$$\frac{2n + 25 - 5\sqrt{2n + 25}}{4\mu N} \leq R \leq \frac{2n + 25 + 5\sqrt{2n + 25}}{4\mu N}. \tag{12.11}$$

12.5 Another Poisson example

There is a well known 1998 movie called *A Civil Action* that tells a fictionalized account of a true story. The movie stars John Travolta and Robert Duvall. Here is a quote about the movie and the case from `www.civil-action.com`:

> *In the early 1980s, a leukemia cluster was identified in the Massachusetts town of Woburn. Three companies, including W. R. Grace & Co., were accused of contaminating drinking water and causing illnesses. There is no question that this tragedy had a profound impact on everyone it touched, particularly the families of Woburn.*

John Travolta and Robert Duvall play the roles of the lawyers for the folks in Woburn who brought forth the civil suit against the companies.

Here are the numbers involved: Over twenty years, there were 20 cases of childhood leukemia in Woburn. On average, there are $3,500$ children of the relevant ages in Woburn each year, so we are talking about 20 cases per $3,500 \times 20 = 70,000$ person-years. Given these numbers, there are two questions to ask:

(a) Is the number of cases, 20, so high as to render it very likely that the cases are *not* random occurrences, but do to some underlying cause?

(b) If the answer to (a) is yes, then are the leukemias caused by pollution from the companies that are named in the suit?

We can't say much about question (b), but statistics can help us with question (a). This is done by testing the significance of the hypothesis that this large number of cases is a random event. For this purpose, note that the sort of leukemia that is involved here occurs with an expected count of 13 per $100,000$ person-years which is 9.1 per $70,000$ years. Thus, we need to ask whether 20 per $70,000$ person-years is significant. If the probabilities here are well modeled by a Poisson probability then the probability of seeing $n \in \{0, 1, \ldots\}$ cases per $100,000$ person-years is

$$\frac{1}{n!}(9.1)^n e^{-9.1}. \tag{12.12}$$

In particular, of interest here is the P-value of 20. This is the probability of being 10.9 or more from the mean. The mean of this Poisson function is 9.1 and the standard deviation is $\sqrt{9.1} \approx 3$. Thus, 20 is roughly 4 standard deviations, which isn't enough to use the Chebychev theorem. This being the case, I can still try to compute the P-value directly using its definition as the probability of getting at least as far from the mean as is 20. This is to say that

$$P\text{-value}(20) = \sum_{n \geq 20} \frac{1}{n!}(9.1)^n e^{-9.1} \approx 0.0012. \tag{12.13}$$

This number is less than 0.05, so indicates a significant P-value.

Now, if I were an advocate for the companies that are involved in the suit, I would assert that this P-value is irrelevant, and for the following reason: Looking at the web site `www.citypopulation.de/USA-Massachusetts.html`, I see that there are 44 towns and cities in Massachusetts with population $30,000$ or greater. (Woburn has a population of roughly $37,000$). I would say that the relevant statistic is not the the leukemia P-value 0.0012 for Woburn, but the P-value for at least one of the 44 towns or cities to have its corresponding leukemia P-value either 0.0012 or less.

The point here is that given any sufficiently large list of towns, random chance will find one (or more) with enough cases of childhood leukemia to give it a leukemia P-value less than or equal to 0.0012. An unethical lawyer could scour the leukemia records of all towns in Massachusetts (or the US!) so as to find one with a very small leukemia P-value. That lawyer might then try to convince the folks in that town to pay the lawyer to sue those local companies that could conceivably pollute the water supply.

Anyway, to see how to compute the P-value for at least one of 44 towns or cities to have leukemia P-value 0.0012, note that this number is identical to the probability of getting at least 1 tail when a coin is flipped 44 times, and where the probability of tails is 0.0012. To compute this probability, remark to start that it is 1 minus the probability of getting all heads. The probability of all heads is $(0.9988)^{44} \approx 0.9485$. Thus, the probability of at least 1 tail is 0.0515. This is barely bigger than our 0.05 cut-off for a significant P-value. This being the case, you can imagine that the lawyers might have brought to court rival "expert statistician witnesses" to argue for or against significance. I don't know if they did or not. You can read more about the actual events at: `http://lib.law.washington.edu/ref/civilaction.htm`.

12.6 A silly example

What follows is a challenge of sorts: Morning after morning, you have woken and seen that daylight invariably arrives. Thus, morning after morning, you have experimental evidence that the sun exists. Based on this, your own experience, give a lower bound for the probability that the sun will exist tomorrow.

How to answer this question? Here is one way: Suppose that there is some probability, $\tau \in [0,1]$, for the sun to exist on any given day. If you are 20 years old, then you have seen the sun has existed on roughly $20 \times 365 \approx 7730$ days in a row. Now, suppose that I have a coin with probability τ for heads and so $(1-\tau)$ for tails. What is the smallest value for τ such that the event of flipping the coin 7730 times and getting all heads has P-value at least? I will take this smallest τ for a lower bound on the probability of sun existing tomorrow.

To calculate this P-value, I imagine flipping this coin until the first tails appears. The sample space for this problem is $S = \{0, 1, 2, \dots\}$, this the set of non-negative integers. The relevant probability function on S is given by

$$\mathrm{P}(n) = (1-\tau)\tau^n. \tag{12.14}$$

Do you see why (12.14) gives the probability? Here is why: The probability for the first flip to land heads is τ, that for the first and second is $\tau \times \tau = \tau^2$, that for the first three to land heads is $\tau \times \tau \times \tau = \tau^3$, and so on; thus you find that the probability for the first n to land heads is τ^n. Now if there are precisely n heads in a row and then tails, the probability is $\tau^n \times (\text{probability of tails}) = \tau^n(1-\tau)$. If you did Problem 2a in the previous chapter, you will have verified that these probabilities sum to 1 as they should.

If you did Problem 2d in the previous chapter, you will have found that the mean of this probability function is

$$\mu = \sum_{n=0,1,2,\dots} n(1-\tau)\tau^n = \frac{\tau}{1-\tau}. \tag{12.15}$$

Note that the mean increases when τ increases and it decreases when τ decreases.

To see what sort of value for τ gives a small P-value for 7730, I start by considering value of τ where the mean, μ, is less than half of 7730. Note that I can solve (12.15) for τ in terms of μ to see that this means looking for τ between 0 and $\frac{3865}{3866}$. There are two reasons why I look at these values first:

- If I find τ in this set, I needn't look further since (12.15) shows that the mean, μ, when viewed as a function of τ, increases with increasing τ.

- There are no non-negative integers n that are both less than μ and further from μ than 7730. This means that the P-value for 7730 as defined by $\tau < \frac{3865}{3866}$ is computed by summing the probability, $(1-\tau)\tau^n$, for those integers $n \geq 7730$.

This is to say that the P-value of 7730 as defined using any given $\tau < \frac{3865}{3866}$ is

$$\sum_{n \geq 7730} (1-\tau)\tau^n = \tau^{7730} \sum_{n \geq 0} (1-\tau)\tau^n = \tau^{7730}. \tag{12.16}$$

Since I want to find the smallest τ where the P-value of 7730 is 0.05, I solve

$$\tau^{7730} = 0.05 \tag{12.17}$$

which is to say $\tau = (0.05)^{1/7730} \approx 0.9996$. Note that this is less than $\frac{3865}{3866} \approx 0.9997$.

Thus, if you are 20 years old, you can say, based on your personal experience, that the probability of the sun existing tomorrow is likely to be greater than 0.9996.

12.7 Exercises:

1. Define a probability function, P, on $\{0, 1, 2, \dots\}$ by setting $\mathrm{P}(n) = \frac{1}{10}(\frac{9}{10})^n$. What is the P-value of 5 using this probability function?

2. Suppose we lock a monkey in a room with a word processor, come back some hours later and see that the monkey has typed N lower case characters. Suppose this string of N characters contains precisely 10 occurrences of the letter "e". The monkey's word processor key board allows 48 lower case characters including the space bar.

   (a) Assume that the monkey is typing at random, and give a formula for the probability that 10 occurrences of the letter "e" appear in the N characters.

   (b) Use the Chebychev theorem to estimate how big to take N so that 10 appearances of the letter "e" has significant P-value.

   Here is a rather more difficult thought problem along the same line: Suppose that you pick up a copy of Tolstoy's novel *War and Peace*. The paperback version translated by Ann Dunnigan has roughly 1,450 pages. Suppose that you find as you read that if you take away the spacing between letters, there is an occurrence of the string of letters "professortaubesisajerk". Note that spaces can be added here to make this a bona fide English sentence. You might ask whether such a string is likely to occur at random in a book of 1,450 pages, or whether Tolstoy discovered something a century or so ago that you are only now coming to realize is true. (Probabilistic analysis of the sort introduced in this problem debunks those who claim to find secret, coded prophesies in the Bible and other ancient texts.)

3. The Poisson probability function is often used to distinguish "random" from less than random patterns. This problem concerns an example that involves spatial patterns. The issue concerns the distribution of appearances of the letter "e" in a piece of writing. To this end, obtain two full page length columns of text from a Boston Globe newspaper.

   (a) Draw a histogram on a sheet of graph paper whose bins are labeled by the nonnegative integers, $n = 0, 1, \dots$. Make the height of the bin numbered by any given integer n equal to the number of lines in your two newspaper columns that have n appearances of the letter "e".

   (b) Compute the total number appearances of the letter "e", and divide the latter number by the total number of lines in your two newspaper columns. Call this number τ. Plot on your graph paper (in a different color), a second histogram where the height of the nth bin is the function $\mathrm{P}_\tau(n)$ as depicted in (12.6).

   (c) Compute the standard deviation of your observed data as defined in (1.3) of Chapter 1 and give the ratio of this observed standard deviation to the standard deviation, $\sqrt{\tau}$, for $\mathrm{P}_\tau(n)$.

(d) Count the number of lines of your two newspaper columns where the number of appearance of the letter "e" differs from τ by more than $5\sqrt{\tau}$. Is the number of such columns more or less than 5% of the total number of lines?

4. This problem concerns a very standard epidemiology application of P-value. If you go to work for the Center for Disease Control, then you may well see a lot of applications of the sort that are considered here. Creutzfeld-Jacob disease is a degenerative brain disease with symptoms very much like that of mad cow disease. It is not known how the disease is propagated, and one conjecture is that arises spontaneously due to chance protein misfolding. Suppose that this is true, and that the average number of cases seen in England per year is 10. Suppose that there were 15 cases seen last year in England. Does appearance of so many cases in one year shed significant doubt on the hypothesis that the disease arises spontaneously?

   (a) Use a calculator to compute the actual P-value of 15 under the assumption that the probability of seeing n cases in one year in England is determined by the $\tau = 10$ version of the Poisson probability function. In this regard, it is almost surely easier to first compute the probability of $\{6, 7, \ldots, 14\}$ and then subtract the latter number from 1. However, if you do the computation in this way, you have to explain why the resulting number is the P-value of 15.

   (b) Compute the Chebychev theorem's upper bound for the P-value of 15.

5. Suppose that cells from the skin are grown in low light conditions, and it is found that under these circumstances, there is some probability, p, for any given cell to exhibit some chromosome abnormality. Granted this, suppose that 100 skin cells are taken at random from an individual and 2 are found to have some chromosome abnormality. How small must p be to deem this number significant?

6. This problem returns to the World Series data that is given in Problem 8 of Chapter 11. The question here is this: Does the data make it unlikely that, on average, each team in the World Series has a 50% probability of winning any given game. To answer this question,

   (a) Use the actual data to prove that the average number of games played in the World Series is 5.8.

   (b) Use the probabilities given in Problem 8(b)–(e) of Chapter 11 to justify the following:

   > If each team has a 50-50 chance of winning any given game, then the mean of the number of games played is 5.8125.

   Thus, the average numbers of games are almost identical.

   (c) Here is another approach: If the probability of a 7-game series is $\frac{5}{16}$, one can ask for the P-value of the fact that there were 33 of them in 83 years. Explain why this question can be answered by studying the binomial probability function on the sample space $\{0, \ldots, 83\}$ with $q = \frac{5}{16}$. In particular, explain why the P-value of 33 is

$$\sum_{n=33,34,\ldots,83} \binom{83}{n} \left(\frac{5}{16}\right)^n \left(\frac{11}{16}\right)^{83-n} + \sum_{n=0,1,\ldots,19} \binom{83}{n} \left(\frac{5}{16}\right)^n \left(\frac{11}{16}\right)^{83-n}.$$

   As it turns out, this number is roughly 0.12 which is not significant.

   (d) There are other statistics to try. For example, lets agree on the following terminology: A *short series* is one that lasts 4 or 5 games, and a *long series* is one that lasts 6 or 7 games.

      (i) Under the assumption that each team has a 50-50 chance of winning any given game, explain why the probability of a short series is $\frac{3}{8}$, and so that of a long series is $\frac{5}{8}$.

      (ii) Explain why the expected number of short series over 83 years is 31.125.

      (iii) The real data finds 32 short series. Do you think this has a significant P-value? Explain why the probability of having $n \in \{0, 1, \ldots, 83\}$ short series in 83 years is $\binom{83}{n} \left(\frac{3}{8}\right)^n \left(\frac{5}{8}\right)^{83-n}$.

CHAPTER

THIRTEEN

Continuous probability functions

So far, we have only considered probability functions on finite sets and on the set of non-negative integers. The task for in this chapter is to introduce probability functions on the whole real line, $\mathbf{R}$, or on a subinterval in $\mathbf{R}$ such as the interval where $0 \leq x \leq 1$. Let me start with an example to motivate why such a definition is needed.

13.1 An example

Suppose that we have a parameter that we can vary in an experiment, say the concentration of sugar in an airtight, enclosed Petri dish with photosynthesizing bacteria. Varying the initial sugar concentration, we measure the amount of oxygen produced after one day. Let x denote the sugar concentration and y the amount of oxygen produced. We run some large number, N, of versions of this experiment with respective sugar concentrations $x_1, \ldots, x_N$ and measure corresponding oxygen concentrations, $y_1, \ldots, y_N$.

Suppose that we expect, on theoretical grounds, a relation of the form $y = cx + d$ to hold. In order to determine the constants c and d, we find the least squares fit to the data $\{(x_j, y_j)\}_{1 \leq j \leq N}$.

Now, the differences,

$$\Delta_1 = y_1 - cx_1 - d, \qquad \Delta_2 = y_2 - cx_2 - d, \qquad \text{etc} \tag{13.1}$$

between the actual measurements and the least squares measurements should not be ignored. Indeed, these differences might well carry information. Of course, you might expect them to be spread "randomly" on either side of 0, but then *what does it mean for a suite of real numbers to be random?* More generally, how can we decide if their distribution on the real line carries information?

13.2 Continuous probability functions

As the example just given illustrates, a notion of "random" is needed for drawing numbers from some interval of real numbers. This means introducing probability functions for sample spaces that are not finite or discrete sets. What follows is the definition when the sample space is a bounded or unbounded in interval, $[a, b] \subset \mathbf{R}$ where $-\infty \leq a < b \leq \infty$. A continuous probability function on such a sample space is any function, $x \to p(x)$, that is defined for those points x in the interval, and that obeys the following two conditions:

- $p(x) \geq 0$ for all $x \in [a, b]$.
- $\displaystyle\int_a^b p(x)\, dx = 1.$ (13.2)

I'll say more about these points momentarily. Here is how you are supposed to use $p(x)$ to find probabilities: If $U \subset [a,b]$ is any given subset, then

$$\int_{x \in U} p(x)\, dx \tag{13.3}$$

is meant to give the probability of finding the point x in the subset U. Granted this use of $p(x)$, then the first constraint in (13.2) forbids negative probabilities; meanwhile, the second guarantees that there is probability 1 of finding x *somewhere* in the given interval $[a,b]$.

A continuous probability function is often called a "probability distribution" since it signifies how probabilities are distributed over the relevant portion of the real line. Note in this regard, that people often refer to the "cumulative distribution function". This function is the anti-derivative of $p(x)$. It is often denoted as $P(x)$ and is defined by

$$P(x) = \int_a^x p(s)\, ds. \tag{13.4}$$

Thus, $P(a)$ is zero, $P(b)$ is one, and $P'(x) = p(x)$. In this regard, $P(x)$ is the probability that p assigns to the interval $[a,x]$. It is the probability of finding a point that is less than the given point x.

The functions $p(x) = 1$, $p(x) = 2x$, $p(x) = 2(x - x^2)$ and $p(x) = \sin(\pi x)$ are all probability functions for the interval $[0,1]$.

13.3 The mean and standard deviation

A continuous probability function for an interval on the real line can have a *mean* and a *standard deviation*. The mean, μ, is

$$\mu = \int_a^b xp(x)\, dx. \tag{13.5}$$

This is the "average" value of x where $p(x)$ determines the meaning of average. The standard deviation, σ, has its square given by

$$\sigma^2 = \int_a^b (x-\mu)^2 p(x)\, dx. \tag{13.6}$$

Note that in the case that $|a|$ or b is infinite, one must worry a bit about whether the integrals actually converge. We won't be studying examples in this course where this is an issue.

By the way, novices in probability theory often forget to put the factor of $p(x)$ into the integrands when they compute the mean or standard deviation. Many of you will make this mistake at some point.

13.4 The Chebychev theorem

As with the probability functions studied previously, there is a fixation on the mean and standard deviation that is justified by a version of Chapter 11's Chebychev theorem:

Theorem 1. *(Chebychev Theorem) Let $x \to p(x)$ denote a probability function on the interval $[a,b]$ where $|a|$ or $|b|$ can be finite or infinite. Suppose now that $R \geq 1$. Then, the probability as defined by $p(x)$ for the points x with $|x - \mu| \geq R\sigma$ is no greater than $\frac{1}{R^2}$.*

Note that this theorem holds for any $p(x)$ as long as both μ and σ are defined. Thus, the two numbers μ and σ give you enough information to obtain *upper bounds* for probabilities without knowing anything more about $p(x)$.

The Chebychev theorem justifies the ubiquitous focus on means and standard deviations.

13.5 Examples of probability functions

Three examples of continuous probability functions appear regularly in the scientific literature.

The uniform probabilities: The simplest of the three is the uniform probability function on some finite interval. Thus, a and b must be finite. In this case,

$$p(x) = \frac{1}{b-a}. \tag{13.7}$$

This probability function asserts that the probability of finding x in an interval of length $L < b - a$ inside the interval $[a, b]$ is equal to $\frac{L}{b-a}$; thus it is proportional to L.

Here is an example where this case can arise: Suppose we postulate that bacteria in a petri dish can not sense the direction of the source of a particular substance. We might then imagine that the orientation of the axis of the bacteria with respect to the xy-coordinate system in the plane of the petri dish should be "random". This is to say that the head end of a bacteria is pointed at some angle, $\theta \in [0, 2\pi]$, and we expect that the particular angle for any given bacteria is "random". Should we have a lot of bacteria in our dish, this hypothesis implies that we must find that the percent of them with head pointed between angles $0 \le \alpha < \beta \le 2\pi$ is equal to $\frac{\beta-\alpha}{2\pi}$.

The mean and standard deviation for the uniform probability function are

$$\mu = \frac{1}{2}(b + a) \qquad \text{and} \qquad \sigma = \frac{b-a}{\sqrt{12}}. \tag{13.8}$$

In this regard, note that the mean is the midpoint of the interval $[a, b]$ (are you surprised?). For example, the uniform probability distribution on the interval $[0, 1]$ has mean $\frac{1}{2}$ and standard deviation $\frac{1}{\sqrt{12}}$.

The Gaussian probabilities: These are probability functions on the whole of $\mathbf{R}$. Any particular version is determined with the specification of two parameters, μ and σ. Here, μ can be any real number, but σ must be a positive real number. The (μ, σ) version the Gaussian probability function is

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-|x-\mu|^2/(2\sigma^2)}. \tag{13.9}$$

If you have a graphing calculator and graph this function for some numerical choices of μ and σ, you will see that the graph is the famous "bell-shaped" curve, but centered at the point μ and with the width of the bell given by σ. In fact, μ is the mean of this probability function and σ is its standard deviation. Thus, small σ signifies that most of the probability is concentrated at points very close to μ. Large σ signifies that the probability is spread out.

There is a theorem called the "Central Limit Theorem" that explains why the Gaussian probability function appears as often as it does. This is a fantastically important theorem that is discussed momentarily.

The exponential probabilities: These are defined on the half line $[0, \infty)$. There are various versions and the specification of any one version is determined by the choice of a positive real number, μ. With μ chosen,

$$p(t) = \frac{1}{\mu} e^{-t/\mu}. \tag{13.10}$$

This one arises in the following context: Suppose that you are waiting for some particular "thing" to happen and you know the following:

- On average, you will have to wait for μ minutes.
- The conditional probability that the "thing" occurs at times greater than t given that it has not happened after some previous t' depends only on the elapsed time, $t - t'$.

(13.11)

If $0 \leq a < b \leq \infty$, you can ask for the probability that the "thing" occurs when $a \leq t < b$. This probability is given by integrating $p(t)$ in (13.10) over the interval where $a \leq t < b$. Thus, it is $e^{-a/\mu} - e^{-b/\mu}$.

The mean of the exponential is μ and the standard deviation is also equal to μ.

To illustrate when you might use this probability function, suppose that you are waiting for a bus, and you know that the mean waiting time is 10 minutes. You have been waiting for 6 minutes, so you know that no bus has appeared during times $t' \leq 6$. You want to know the probability that you will have to wait at least 6 more minutes for the bus. You would use the $\mu = 10$ version of the exponential probability function to compute this probability if you make the following assumption about buses: Let A denote the event that the bus appears after time t. Supposing that $t' < t$, let B denote the event that the bus appears after time t'. Then the conditional probability of A given B, thus $\mathrm{P}(A \,|\, B)$, is depends only on the $t - t'$. Granted this assumption, then the probability for waiting at least 6 more minutes is $\mathrm{P}(A \,|\, B)$ as computed for the case that $A = [12, \infty)$ and $B = [6, \infty)$. Compute this number using the formula $\mathrm{P}(A \,|\, B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$. For this purpose, note that $A \cap B = [12, \infty)$, so $\mathrm{P}(A \cap B)$ is obtained by integrating the function $e^{-t/10}$ from 12 to ∞. The result is $e^{-6/5}$. Meanwhile, $\mathrm{P}(B)$ is obtained by integrating this same function from 6 to ∞. The result is $e^{-3/5}$. Thus, the probability of waiting at least 6 more minutes for a bus is $e^{-3/5} \approx 0.55$.

13.6 The Central Limit Theorem: Version 1

The Central Limit Theorem explains why the "bell shaped" curve arises in so many different contexts. Here is a typical situation: You do the same experiment some large number of times, each time measuring some given quantity. The result is a suite of N numbers, $\{x_1, \dots, x_N\}$. Suppose now that we look at the average of the N measurements,

$$\overline{x} = \frac{1}{N}\,(x_1 + \cdots + x_N)\,. \tag{13.12}$$

Even if the x_ks have only a finite set of possible values, the set of possible values of $\overline{x}$ becomes ever larger as $N \to \infty$. The question one might ask is what is the probability of $\overline{x}$ having any given value? More to the point, what is the probability function for the possible values of $\overline{x}$?

The Central Limit Theorem answers this question when the following assumption is valid: There is some probability function, p, on the set of possible values for any given x_k, and it is the same for each k. The Central Limit Theorem also assumes that the values that are measured in any subset of the N experiments have no bearing on the values that are measured for the remaining experiments. The Central Limit Theorem then asserts that the probability function for the possible values of the average, $\overline{x}$, depends only on the mean and standard deviation of the probability function p. The detailed ups and downs of p are of no consequence, only its mean, μ, and standard deviation, σ. Here is the theorem:

Central Limit Theorem. *Under the assumptions just stated, the probability that the value of $\overline{x}$ is in some given interval $[a, b]$ is well-approximated by*

$$\int_a^b \frac{1}{\sqrt{2\pi}\;\sigma_N} e^{-|x-\mu|^2/(2\sigma_N^2)}\; dx \tag{13.13}$$

where $\sigma_N = \frac{1}{\sqrt{N}}\sigma$. This is to say that for very large N the probability function for the possible values of $\overline{x}$ is very close to the Gaussian probability function with mean μ and with standard deviation $\sigma_N = \frac{1}{\sqrt{N}}\sigma$.

Here is an example: Suppose a coin is flipped some N times. For any given $k \in \{1, 2, \dots, N\}$, let $x_k = 1$ if the kth flip is heads, and let $x_k = 0$ if it is tails. Let $\overline{x}$ denote the average of these values, this as defined via (13.12). Thus, $\overline{x}$ can take any value in the set $\{0, \frac{1}{N}, \frac{2}{N}, \cdots, 1\}$. According to the Central Limit Theorem, the probabilities for the values of $\overline{x}$ are, for very large N, essentially determined by the Gaussian probability function

$$p_N(x) = \frac{1}{\sqrt{2\pi}} 2\sqrt{N} e^{-2N|x-\frac{1}{2}|^2}. \tag{13.14}$$

Here is a second example: Suppose that N numbers are randomly chosen between 0 and 100 with uniform probability in each case. This is to say that the probability function is that given in (13.7) using $b = 100$ and $a = 0$. Let $\overline{x}$ denote the average of these N numbers. Then for very large N, the probabilities for the values of $\overline{x}$ are are very close to those determined by the Gaussian probability function

$$p_N(x) = \frac{1}{\sqrt{2\pi}} \frac{\sqrt{12N}}{1000} e^{-3N|x-50|^2/5000}. \tag{13.15}$$

By the way, here is something to keep in mind about the Central Limit Theorem: As N gets larger, the mean for the Gaussian in (13.13) is unchanged, it is the same as that for the original probability function p that gives the probabilities for the possible values of any given measurement. However, the standard deviation shrinks to zero in the limit that $N \to \infty$ since it is obtained from the standard deviation, σ, of p as $\frac{1}{\sqrt{N}}\sigma$. Thus, the odds of finding the average, $\overline{x}$, some fixed distance from the mean μ decreases to zero in the limit that $N \to \infty$. The Chebychev inequality also predicts this. Indeed, the Chebychev inequality in this context asserts the following:

> Fix a real number, r, and let $p_N(r)$ denote the probability that the average, $\overline{x}$, of N measurements obeys $|\overline{x} - \mu| > r$. Then $p_N(r) \leq \frac{1}{N}\left(\frac{\sigma}{r}\right)^2$ when N is very large. (13.16)

The use of (13.13) to approximate $p_N(r)$ when N is large suggests that $p_N(r)$ is much smaller than the Chebychev upper bound from (13.16). Indeed, the sum of the integrals that appears in (13.13) in the case $a = \mu + r$, $b = \infty$ and in the case $a = -\infty$ and $b = \mu - r$ can be proved no greater than

$$\sqrt{\frac{2}{\pi}} \, \frac{1}{\sqrt{N}} \, \frac{\sigma}{r} e^{-Nr^2/(2\sigma^2)}. \tag{13.17}$$

To derive (13.17), I am using the following fact: Suppose that both κ and r are positive real numbers. Then

- $\displaystyle\int_{\mu+r}^{\infty} \frac{1}{\sqrt{2\pi}\,\kappa} e^{-(x-\mu)^2/(2\kappa^2)}\, dx \leq \frac{1}{\sqrt{2\pi}} \frac{\kappa}{r} e^{-r^2/(2\kappa^2)}.$
- $\displaystyle\int_{-\infty}^{\mu-r} \frac{1}{\sqrt{2\pi}\,\kappa} e^{-(x-\mu)^2/(2\kappa^2)}\, dx \leq \frac{1}{\sqrt{2\pi}} \frac{\kappa}{r} e^{-r^2/(2\kappa^2)}.$

(13.18)

Here, I use κ as a generic stand-in for "σ", since we will be applying this formula in instances where $\kappa = \sigma$ (such as in (13.17)), and others where $\kappa = \sigma/\sqrt{N}$ (such as when using the Central Limit Theorem). You are walked through the proof of the top inequality of (13.18) in one of the exercises at the end of this chapter. The bottom inequality is obtained from the top by changing variables of integration from x to $2\mu - x$.

13.7 The Central Limit Theorem: Version 2

There is another, even more useful version of the Central Limit Theorem that has the version just given as a special case. This more general version asserts that when a measurement is affected by lots of small, random perturbations, the probabilities for the values of the measurement are well approximated by those that are determined via a Gaussian probability function.

For example, consider measuring a certain quantity some large number of times, thus generating a sequence of numbers, $\{x_1, \ldots, x_N\}$, where x_k is the result of the kth measurement. Consider the function

$$f_N(x) = \text{fraction of measurements } j \text{ with } x_j < x. \tag{13.19}$$

The possible values of $f_N(x)$ are $\{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\}$. We are interested in the large N version of this function. What almost always happens is that as N gets large, the graph of $f_N(x)$ gets closer and closer to the graph of the cumulative

distribution function for a Gaussian probability. To be explicit, the following phenomena is often observed:

$$\begin{aligned}&\text{As } N \to \infty, \text{ the graph of } x \to f_N(x) \text{ approaches that of the function}\\ &x \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(s-\mu)^2/(2\sigma)}\, ds \text{ for an appropriate choice of } \mu \text{ and } \sigma.\end{aligned} \tag{13.20}$$

Note that the constant μ to use is the $N \to \infty$ limit of the average, $\overline{x} = \frac{1}{N}(x_1 + \cdots x_N)$.

Another version of the Central Limit Theorem provides a mathematical explanation for why this phenomena arises. As the assumptions of the version that follows are rather technical and the proof is not something we will cover, I just give you the following somewhat vague statement:

Central Limit Theorem II. *Suppose that the variations of some measurement are due to a very large number of small perturbations. If the probability of the value of any given perturbation is a function with mean zero, and if the collection of these probability functions are not unreasonable, then the probability for any given measurement is well approximated by a Gaussian probability function.*

Note that this version does not assume that the probability function for any given source of perturbation is the same as that of any of the others. The term "not unreasonable" means that the sum of the squares of the standard deviations for the probability functions of the perturbations is finite, as is the sum of the average values of x^4 for these probability functions.

13.8 The three most important things to remember

The three most important things to remember here are the following:

- A Gaussian probability function offers a good approximation to reality when a measurement is affected by a very large number of small perturbations.
- The probabilities for the values of the average of some N measurements are determined to a very good approximation when N is very large by a Gaussian probability function of the form $\frac{1}{\sqrt{2\pi}\,\sigma}\sqrt{N}e^{-N(x-\mu)^2/(2\sigma^2)}$, where μ and σ are independent of N.
- When using a Gaussian probability measure, remember the following very useful inequality: $\int_{\mu+r}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/(2\sigma^2)}\, dx \leq \frac{\sigma}{\sqrt{2\pi}\,r} e^{-r^2/(2\sigma^2)}$.

(13.21)

The top two points tell you that Gaussian probability functions arise in most contexts, and the bottom point tells you how to find upper bounds for probabilities as determined by a Gaussian probability function.

13.9 A digression with some comments on Equation (13.1)

What follows constitutes a digression to point out that the suite of numbers $\{\Delta_j\}$ that appear in (13.1) cannot be completely arbitrary by virtue of the fact that their sum is zero: $\Delta_1 + \Delta_2 + \cdots + \Delta_N = 0$. This conclusion is a consequence of the least squares definition of the constants c and d. To see how this comes about, remember that the least squares is defined by first introducing the matrix, A, with 2 columns and N rows whose jth row is $(x_j, 1)$. The constants c and d are then

$$\begin{pmatrix} c \\ d \end{pmatrix} = \left(A^T A\right)^{-1} A^T \vec{y}, \tag{13.22}$$

where $\vec{y} \in \mathbf{R}^N$ is the vector whose jth entry is y_j. Granted (13.22), the vector $\vec{\Delta} \in R^n$ whose jth entry is Δ_j is

$$\vec{\Delta} = \vec{y} - A\left(A^T A\right)^{-1} A^T \vec{y}. \tag{13.23}$$

Now, save (13.23) for the moment, and note that if $\vec{v} \in \mathbf{R}^N$ is any vector, then

$$A^T \vec{v} = \begin{pmatrix} x_1 v_1 + \cdots + x_N v_N \\ v_1 + \cdots + v_N \end{pmatrix}. \tag{13.24}$$

In the case that $\vec{v} = \vec{\Delta}$, the top and bottom components in (13.24) are

$$\Delta_1 x_1 + \cdots + \Delta_N x_N \qquad \text{and} \qquad \Delta_1 + \cdots + \Delta_N. \tag{13.25}$$

Now, let us return to (13.23) to see what $A^T \vec{\Delta}$ turns out to be. For this purpose, multiply both sides by A^T to find

$$A^T \Delta = A^T \vec{y} - A^T A \left(A^T A\right)^{-1} A^T \vec{y}. \tag{13.26}$$

Next, use the fact that $A^T A \left(A^T A\right)^{-1}$ is the identity matrix to conclude that $A^T \vec{\Delta} = 0$. Thus, both sums in (13.25) are zero, the right-hand one in particular.

13.10 Exercises:

1. (a) Sketch a graph of the Gaussian probability function for the values $\mu = 0$ and $\sigma = 1$, 2, 4, and 8.
   (b) Prove the inequality that is depicted in the top line of (13.18) by the following sequence of arguments:
      (i) Change variables of integration using the substitution $y = x - \mu$ so write the integral as
      $$\int_r^\infty \frac{1}{\sqrt{2\pi}\,\kappa} e^{-y^2/(2\kappa^2)}\, dy$$
      (ii) Note that the integration range has $y \geq r$, so the integral is no greater than
      $$\int_r^\infty \frac{1}{\sqrt{2\pi}\,\kappa} \frac{y}{r} e^{-y^2/(2\kappa^2)}\, dy$$
      (iii) Change variables in this integral by writing $u = \frac{y^2}{2\kappa^2}$. Thus, $du = \frac{y}{\kappa^2}\, dy$ and so as to write this last integral as
      $$\frac{\kappa}{\sqrt{2\pi}\, r} \int_{r^2/(2\kappa^2)}^\infty e^{-u}\, du.$$
      (iv) Complete the job by evaluating the integral in (iii).

2. This exercise works with the exponential probability function in (13.10).
   (a) Verify by integration that the mean and standard deviations for $p(x)$ in (13.10) are both μ.
   (b) Use (13.10) to compute the probability that the desired event does not occur prior to time $t' > 0$.
   (c) Suppose t', a and b are given with $0 < t' < a < b$. Prove that the conditional probability that the time of the desired event is in the interval $[a, b]$ given that the event time is not before t' is given by $e^{-(a-t')/\mu} - e^{-(b-t')/\mu}$.

3. Consider the uniform probability function on the interval $[0, \pi]$. Compute the mean and standard deviation for the random variable $x \to \sin(x)$. In this regard, remember that when $x \to f(x)$ is a random variable, then its mean is $\mu_f \equiv \int_a^b f(x)p(x)\, dx$, and the square of its standard deviation is $\sigma^2 \equiv \int_a^b (f(x) - \mu_f)^2 p(x)\, dx$.

4. This problem concerns the example in Section 13.1 above where we run N versions of an experiment with sugar concentration x_k in the kth experiment and measure the corresponding oxygen concentration y_k. We then do a least squares fit of the data to a line of the form $y = ax + b$. Then $\Delta_j = y_j - ax_j - b$ tells us how far y_j is from the value predicted by the equation $y = ax + b$. We saw that the average of these Δ_j's is zero. That is, the sum $\sum_j \Delta_j$ is 0. If we assume that the probability for any Δ_j landing in any given interval $U \subset (-\infty, \infty)$ is determined by a Gaussian probability function, thus equal to
   $$\int_{x \in U} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-|x-\mu|^2/(2\sigma^2)}\, dx.$$
   Find formulae for μ and σ in terms of the quantities $\{\Delta_j\}_{1 \leq j \leq N}$. Hint: See Chapter 1.

5. Assume that a given measurement can take any value between 0 and π.

   (a) Suppose that the probability function for its possible values is $p(x) = \alpha \sin(x)$. Find α, the mean, the standard deviation, and the probability for the measurement to lie between 0 and $\frac{\pi}{4}$. Finally, write down the cumulative distribution function.

   (b) Suppose instead that another probability function is proposed, this one whose cumulative distribution function has the form $\beta + \gamma e^x$. Find the constants β, γ and the probability function. What is the probability in this case for the measurement to lie between 0 and $\frac{\pi}{4}$.

6. Of interest to mathematicians is whether the digits that appear in the decimal expansion for the number π occur at random. What follows are the first 101 digits in this expansion:

$$\pi = 3.1415926535\ 8979323846\ 2643383279\ 5028841971\ 6939937510$$
$$5820974944\ 5923078164\ 0628620899\ 8628034825\ 3421170679$$

   If the digits occur at random, then each digit should have probability $\frac{1}{10}$ of occurring at any given decimal place and one might expect that the probability of a given digit appearing n times in the expression above is given by the version of the binomial function in (11.9) of Chapter 11 with $N = 101$ and $q = \frac{1}{10}$. Count the number of appearances of each digit from the set $\{0, \ldots, 9\}$, then use Chapter 11's Chebychev Theorem to get an upper bound the P-value of its appearance.

7. Throw away the 3 to the left of the decimal place in the decimal expansion of π given in the previous problem. Group the 100 remaining digits into 10 groups of 10, by moving from left to right. Thus, the first group is $\{1415926535\}$ and the tenth group is $\{3421170679\}$. Let μ_k for $k \in \{1, \ldots, 10\}$ denote the mean of the kth group.

   (a) Compute each μ_k.

   (b) Let S denote the set $\{0, \ldots, 9\}$ and let p denote the probability function on S that assigns $\frac{1}{10}$ to each element. Compute the mean and standard deviation of the numbers in S as determined by p.

   (c) Compute the mean and standard deviation of the set $\{\mu_1, \ldots, \mu_{10}\}$

   (d) What does the Central Limit Theorem predict for the mean and the standard deviation of the set $\{\mu_1, \ldots, \mu_{10}\}$ if it is assumed that the probability of a digit appearing in any given decimal place is $\frac{1}{10}$ and that the digits that appear in any subset of the 100 decimal places have no bearing on those that appear in the remaining places?

8. Let $p(x)$ denote the version of the Gaussian probability function in (13.9) with $\mu = 0$ and $\sigma = 1$.

   (a) Use the relevant version of (13.18) to obtain an upper bound for the probability that p assigns to the set of points on the line where $|x| \geq 5$. These are the points that differ from the mean by 5 standard deviations.

   (b) Use the Chebychev Theorem to obtain an upper bound for the probability that p assigns to the set of points where $|x| \geq 5$.

   (c) Use a calculator to compute the ratio $\frac{\alpha}{\beta}$ where α is your answer to 8(a) and β is your answer to 8(b). (You will see that the Chebychev upper bound for the Gaussian probability function is much greater than the bound obtained using (13.18).)

CHAPTER

FOURTEEN

Hypothesis testing

My purpose in this chapter is to say something about the following situation: You repeat some experiment a large number, N, times and each time you record the value of a certain key measurement. Label these values as $\{z_1, \ldots, z_N\}$. A good theoretical understanding of both the experimental protocol and the biology should provide you with a hypothetical probability function, $x \to p(x)$, that gives the probability that a measurement has value in any given interval $[a, b] \subset (-\infty, \infty)$. Here, $a < b$ and $a = -\infty$ and $b = \infty$ are allowed. For example, if you think that the variations in the values of z_j are due to various small, unrelated, random factors, then you might propose that $p(x)$ is a Gaussian, thus a function that has the form

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/(2\sigma^2)}. \tag{14.1}$$

for some suitable choice of μ and σ.

In any event, let's suppose that you have some reason to believe that a particular $p(x)$ should determine the probabilities for the value of any given measurement. Here is the issue on the table:

> Is it likely or not that N experiments will obtain a sequence $\{z_1, \ldots, z_N\}$ if the probability of any one measurement is really determined by $p(x)$? (14.2)

If the experimental sequence, $\{z_1, \ldots, z_N\}$ is "unlikely" for your chosen version of $p(x)$, this suggests that your understanding of the experiment is less than adequate.

I describe momentarily two ways to answer the question that is posed in (14.2).

14.1 An example

By way of an example, I ask you to consider the values of $\sin(x)$ for x an integer. As the function $x \to \sin(x)$ has its values between -1 and 1, it is tempting to make the following hypothesis:

> The values of $\sin(x)$ on the set of integers are distributed in a random fashion in the interval $[-1, 1]$.

As a good scientist, I now proceed to test this hypothesis. To this end, I do 100 experiments, where the kth experiment amounts to computing $\sin(k)$. Thus, the outcome of the kth experiment is $z_k = \sin(k)$. For example, here are $z_1, \ldots, z_{10}$ to three significant figures:

$$\{0.841, 0.909, 0.141, -0.757, -0.959, -0.279, 0.657, 0.989, 0.412, -0.544\}\,.$$

The question posed in (14.2) here asks the following: How likely is it that the 100 numbers $\{z_k = \sin(k)\}_{1 \le k \le 100}$ are distributed at random in the interval $[-1, 1]$ with "random" defined using the uniform probability function that has $p(x) = \frac{1}{2}$ for all x?

14.2 Testing the mean

To investigate the question that is posed in (14.2), let μ denote the mean for $p(x)$. I remind you that

$$\mu = \int_a^b xp(x)\ dx, \tag{14.3}$$

where $-\infty \leq a < b \leq \infty$ delineate the portion of $\mathbf{R}$ where $p(x)$ is defined. If the probabilities for the measurements $(z_1, \ldots, z_N)$ are determined by our chosen $p(x)$, then you might expect their average,

$$\overline{z} = \frac{1}{N}(z_1 + z_2 + \cdots + z_N) \tag{14.4}$$

to be reasonably close to μ. The size of the difference between $\overline{z}$ and μ shed light on the question posed in (14.2).

To elaborate, I need to introduce the standard deviation, σ, for the probability function $p(x)$. I remind you that

$$\sigma^2 = \int_a^b (x-\mu)^2\, p(x)\ dx. \tag{14.5}$$

I will momentarily invoke the Central Limit Theorem. To this end, suppose that N is a positive integer. Let $\{x_1, \ldots, x_N\}$ denote the result of N measurements where no one measurement is influenced by any other, and where the probability of each is actually determined by the proposed function $p(x)$. Let

$$\overline{x} \equiv \frac{1}{N} \sum_{1 \leq j \leq N} x_j \tag{14.6}$$

denote the average of the N measurements. The Central Limit Theorem asserts the following: If $R \geq 0$, then the probability that $\overline{x}$ obeys $|\overline{x} - \mu| \geq R\frac{1}{\sqrt{N}}\sigma$ is approximately

$$2\frac{1}{\sqrt{2\pi}} \int_R^\infty e^{-s^2/2}\ ds \tag{14.7}$$

when N is very large. Note that this last expression is less than

$$2\frac{1}{\sqrt{2\pi}}\ \frac{1}{R}\ e^{-R^2/2}, \tag{14.8}$$

as can be seen by applying the version of (13.18) with $\mu = 0$, $\kappa = 1$ and $r = R$.

With the proceeding understood, I will use the Central Limit Theorem to estimate the probability that $\overline{x}$ is further from the mean of $p(x)$ than the number $\overline{z}$ that is depicted in (14.4). This probability is the P-value of $\overline{z}$ under the hypothesis that $p(x)$ describes the variations in the measurements of the N experiments. The Central Limit Theorem's estimate for this probability is given by the $R = \frac{1}{\sigma}\sqrt{N}|\mu - \overline{z}|$ version of (14.7).

If the number obtained by the $R = \frac{1}{\sigma}\sqrt{N}|\mu - \overline{z}|$ version of (14.7) is less than $\frac{1}{20}$, and N is very large, then the P-value of our experimental mean $\overline{z}$ is probably "significant" and suggests that our understanding of our experiment is inadequate.

To see how this works in an example, consider the values of $\sin(x)$ for x an integer. The $N = 100$ version of $\overline{z}$ for the case where $z_k = \sin(k)$ is -0.0013. Since I propose to use the uniform probability function to determine probabilities on $[-1, 1]$, I should take $\mu = 0$ and $\sigma = \frac{2}{\sqrt{12}} = \frac{1}{\sqrt{3}}$ for determining R for use in (14.7). Granted that $N = 100$, I find that $R = \frac{1}{\sigma}\sqrt{N}|\mu - \overline{z}| \approx 0.0225$. As this is very much smaller than $\sqrt{2}$, I can't use (14.8) as an upper bound for the integral in (14.7). However, I have other tricks for estimating (14.7) and find it very close to 0.982. This is a good approximation to the probability that the average of 100 numbers drawn at random from $[-1, 1]$ is further from 0 than 0.0013. In particular, this result doesn't besmirch my hypothesis about the random nature of the values of $\sin(x)$ on the integers.

14.3 Random variables

The subsequent discussion should be easier to follow after this primer about random variables. To set the stage, suppose that you have a probability function, $p(x)$, defined on a subset $[a, b]$ of the real numbers. A random variable in this context is no more nor less than a function, $x \to f(x)$, that is defined when x is between a and b. For example, $f(x) = x$, $f(x) = x^2$, and $f(x) = e^{\sin(x)} - 5x^3$ are examples of random variables.

A random variable can have a mean and also a standard deviation. If $f(x)$ is the random variable, then its mean is often denoted by μ_f, and is defined by

$$\mu_f = \int_a^b f(x)p(x)\, dx. \tag{14.9}$$

This is the "average" of f as weighted by the probability $p(x)$. The standard deviation of f is denoted by σ_f, it is non-negative and defined by setting

$$\sigma_f^2 = \int_a^b \left(f(x) - \mu_f\right)^2 p(x)\, dx. \tag{14.10}$$

The standard deviation measures the variation of f from its mean, μ_f. I assume in what follows that the integrals in (14.9) and (14.10) are finite in the case that $a = -\infty$ or $b = \infty$. In general, this need not be the case.

By way of example, when $f(x) = x$, then μ_f is the same as the mean, μ, of p that is defined in (14.3). Meanwhile, σ_f in the case that $f(x) = x$ is the same as the standard deviation that is defined in (14.5).

Here is another example: Take $p(x)$ to be the Gaussian probability function for the whole real line. Suppose that k is a positive integer. Then $x \to x^k$ is a random variable whose mean is zero when k is odd and equal to $\frac{k!}{\frac{k}{2}!2^{k/2}}$ when k is even. Meanwhile, the square of the standard deviation for this random variable is $\frac{(2k)!}{k!2^k}$ when k is odd and equal to $\frac{1}{2^k}\left(\frac{(2k)!}{k!} - \left(\frac{k!}{\frac{k}{2}!}\right)^2\right)$ when k is even.

There is no need for you to remember the numbers from these examples, but it is important that you understand how the mean and standard deviation for a random variable are defined.

By the way, here is some jargon that you might run into: Suppose that $p(x)$ is a probability function on a part of the real line, and suppose that μ is its mean. Then the average, $\int_a^b (x-\mu)^k p(x)\, dx$, for positive integer k is called the *kth order moment of* $p(x)$.

14.4 The Chebychev and Central Limit Theorems for random variables

The mean and standard deviation for any given random variable are important for their use in the Chebychev Theorem and the Central Limit Theorem. For example, the Chebychev Theorem in the context of a random variable says the following:

Chebychev Theorem for Random Variables. *Let $x \to p(x)$ denote a probability function on a part of the real line, and let $x \to f(x)$ denote a random variable on the domain for p with mean μ_f and standard deviation σ_f as determined by p. Let R be a positive number. Then, $\frac{1}{R^2}$ is greater than the probability that p assigns to the set of points where $|f(x) - \mu_f| \geq R\sigma_f$.*

The proof of this theorem is identical, but for a change of notation, to the proof given previously for the original version of the Chebychev Theorem.

Here is the context for the random variable version of the Central Limit Theorem: You have a probability function, $x \to p(x)$, that is defined on some or all of the real line. You also have a function, $x \to f(x)$, that is defined where p is. You now interpret p and f in the following way: The possible "states" of a system that you are interested in are labeled by the points in the domain of p, and the function p characterizes the probability for the system to be in the state labeled by x. Meanwhile, $f(x)$ is the value of a particular measurement when the system is in the state labeled by x.

Granted these interpretations of p and f, imagine doing some large number, N, of experiments on the system and make the corresponding measurement each time. Generate in this way N numbers, $\{f_1, \dots, f_N\}$. Let $\overline{f} = \frac{1}{N}(f_1 + \cdots + f_N)$ denote the average of the N measurements. These measurements should be independent in the sense that the values that are obtained in any subset of the N measurements have no bearing on the values of the remaining measurements. Under these circumstances, the Central Limit Theorem asserts the following:

Central Limit Theorem. *When N is large, the probability that $\overline{f}$ has values in any given subset of the real line is very close to the probability that is assigned to the subset by the Gaussian function $\frac{1}{\sqrt{2\pi}\,\sigma_f} e^{-(x-\mu_f)^2/(2\sigma_f)}$. For example, if $-\infty \le s < r \le \infty$ and the subset in question is the interval where $s < x < r$, then the probability that $s < \overline{f} < r$ is very close when N is large to*

$$\int_s^r \frac{1}{\sqrt{2\pi}\,(\sigma_f/\sqrt{N})} e^{-(x-\mu_f)/(2N\sigma_f)}\, dx.$$

For example, if $R \ge 0$, then the probability that $\overline{x}$ obeys $|\overline{x} - \mu_f| \ge R\frac{\sigma_f}{\sqrt{N}}$ is approximately

$$2 \cdot \frac{1}{\sqrt{2\pi}} \int_R^\infty e^{-s^2/2}\, ds \tag{14.11}$$

when N is very large.

14.5 Testing the variance

Return now to the question posed in (14.2). If our experimental mean is reasonably close to hypothesized mean, μ, for $p(x)$, then our hypothesis that $p(x)$ describes the variations in the measurements $\{z_1, \dots, z_N\}$ can be tested further by seeing whether the *variation* of the z_j about their mean is a likely or unlikely occurrence.

Here is one way to do this: Let $f(x)$ denote the function $x \to (x-\mu)^2$. I view f as a random variable on the domain for $p(x)$ that takes its values in $[0, \infty)$. Its mean, μ_f, as determined by $p(x)$, is

$$\mu_f = \int_a^b (x-\mu)^2 p(x)\, dx. \tag{14.12}$$

Here, I write the interval where p is defined as $[a, b]$ with $-\infty \le a < b \le \infty$. You might recognize μ_f as the square of the standard deviation of $p(x)$. I am calling it by a different name because I am thinking of it as the mean of the random variable $f(x) = (x-\mu)^2$.

The random variable $f(x) = (x-\mu)^2$ can have a standard deviation as well as a mean. According to our definition in (14.10), this standard deviation has square

$$\sigma_f^2 = \int_a^b \left((x-\mu)^2 - \mu_f\right)^2 p(x)\, dx = \int_a^b (x-\mu)^4 p(x)\, dx - \mu_f^2. \tag{14.13}$$

Put μ_f and σ_f away for a moment to bring in our experimental data. Our data gives us the N numbers $\{(z_1 - \mu)^2, (z_2-\mu)^2, \dots, (z_N-\mu)^2\}$. The plan is to use the random variable version of the Central Limit Theorem to estimate a P-value. This is the P-value for the average of these N numbers,

$$\mathrm{Var} \equiv \frac{1}{N} \sum_{1 \le j \le N} (z_j - \mu)^2, \tag{14.14}$$

under the assumption that $p(x)$ determines the spread in the numbers $\{z_j\}_{1 \le j \le N}$. I call Var the "experimentally determined variance".

According to the Central Limit Theorem, if the spread in the numbers $\{z_j\}_{1 \le j \le N}$ are really determined by $p(x)$, then the P-value of our experimentally determined variance is well approximated by the integral in (14.11) for the case that

$$R = \frac{\sqrt{N}}{\sigma_f} \left|\mathrm{Var} - \mu_f\right|. \tag{14.15}$$

To summarize: Our hypothetical probability function $p(x)$ gives the expected variance, this the number that we compute in (14.12). The Central Limit Theorem then says that our experimentally determined variance, this the number from (14.14), should approach the predicted one in (14.12) as $N \to \infty$. Moreover, the Central Limit Theorem gives an approximate probability, this the expression given using (14.15) in (14.11), for any measured variance to differ from the theoretical one by more than the experimentally determined variance. In particular, if the integral in (14.11) is less than $\frac{1}{20}$ using R from (14.15), then there is a significant chance that our theoretically determined $p(x)$ is not correct.

I am hoping that the example using the values of $\sin(x)$ on the integers makes this less abstract. To return to this example, recall that the probability function $p(x)$ I use is the uniform probability function on the interval $[-1, 1]$. This version of $p(x)$ finds $\mu = 0$ and $\mu_f = \frac{1}{3}$ for the case $f(x) = (x - \mu)^2 = x^2$. The relevant version version of (14.13) uses $a = -1$, $b = 1$ and $p(x) = \frac{1}{2}$; and the resulting integral in (14.13) finds $\sigma_f^2 = \frac{4}{45} \approx 0.089$. Meanwhile, the $N = 100$ and $x_k = \sin(k)$ version of (14.14) finds $\mathrm{Var} \approx 0.503$. Granted this, then I find the right hand side of (14.15) equal to 5.69. Since this value for R is larger than $\sqrt{2}$, I can use (14.8) as an upper bound for the value of the integral in (14.11). This upper bound is less than 4.2×10^{-7}! This is an upper bound for the P-value of the number $\mathrm{Var} \approx 0.503$ under the hypothesis that the values of $\{\sin(k)\}_{1 \le k \le 100}$ are distributed uniformly in the interval $[-1, 1]$. This tiny P-value puts a very dark cloud over my hypothesis that the values of $\sin(x)$ for x an integer are uniformly distributed between -1 and 1.

14.6 Did Gregor Mendel massage his data?

To reach conclusions about inheritance, Gregor Mendel ran a huge number of experiments on pea plants[1] In fact, he raised over $20,000$ plants from seed over an 8 year period. From this data, he was able to glean the "simple" inheritance of dominant-recessive genes. In modern language, here is what this means: Some traits are controlled by two different versions of a particular gene. The different versions are called "alleles". One allele gives the "dominant" trait or phenotype, and one gives the "recessive" trait. An individual plant has 2 copies of each gene, and shows the recessive trait only when both the copies are the recessive allele. Each pollen grain receives only one gene, and each ovule receives only one. A seed is formed when a pollen grain merges with an ovule; thus each offspring has again two copies of the gene. Granted this fact, if it is assumed that each allele has probability to be passed to any given pollen grain or ovule, one should expect about a 3:1 ratio of dominant phenotypes to recessive phenotypes.

This explanation for the inheritance of traits from parent plant to seedling predicts also predicts the following: A individual plant with the dominant phenotype should have 2 dominant alleles of the time; the other of the time should have one copy of each of the two alleles. A plant with the dominant trait is called a "homozygous" dominant when it has two dominant alleles, and it is called a "heterozygous" dominant when it has one dominant allele and one recessive allele. For hundreds of dominant phenotype plants, Mendel was able to classify whether they were homozygous or heterozygous by noting whether the given plant's clonal offspring (the plant was self-pollinated) showed any recessive traits. From these experiments, the theory predicts a 2:1 ratio of heterozygous dominant to homozygous dominant plants. Below is a chart of data from Mendel's work with regards to this homozygous/heterozygous experiment.

Mendel was accused, long after he died, of having massaged his data. One of the complains was that Mendel's data does not have enough spread about the mean. R. A. Fisher[2] analyzed this complaint along the following lines: Imagine flipping a coin some number, say N times. Here, the coin is such that the probability of heads is $\frac{1}{3}$. Use the $q = \frac{1}{3}$ binomial probability functions to see that the mean number of heads for N flips is $\frac{N}{3}$ and the standard deviation is $\frac{1}{3}\sqrt{2N}$. Now, consider Mendel's data as presented above. There are 8 separate experiments, where the values of N differ. In particular, the values of the mean and the standard deviation for the 8 different values of N are: Here, I have rounded μ and σ off to the nearest integer. The column that is headed as "h", is a copy of the column that is headed by "Homozygous Dominant" in the previous table. Thus, this column contains the actual data, the number of homozygous dominant offspring. The far right column takes the number of homozygous dominant offspring, subtracts the mean, divides the result by σ and, after taking absolute values, rounds to the nearest 10th. For example, the number in the far right column for round/wrinkled seeds is $\frac{193-188}{11} = \frac{5}{11} \approx 0.5$. The number in the far right column for green/yellow seeds is $\frac{173-166}{10.7} \approx 0.7$. Thus, the far right column gives the distance of the observed value from the expected value

[1] Mendel, G., 1866 Versuche über Pflanzen-Hybriden. Verh. Naturforsch. Ver. Brünn **4**: 3–47 (first English translation in 1901, J. R. Hortic. Soc. **26**: 1–32; reprinted in *Experiments in Plant Hybridization*, Harvard University Press, Cambridge, MA, 1967).

[2] Fisher, R. A., 1936, Has Mendel's work been rediscovered? Ann. Sci. **1**: 115–137.

	Homozygous Dominant	Heterozygous Dominant	Ratio
round/wrinkled seed	193	372	1.93
green/yellow seed	166	353	2.13
colored/white flower	36	64	1.78
tall/dwarf plant	28	72	2.57
constricted pod or not	29	71	2.45
green/yellow pod #1	40	60	1.50
terminal flower or not	33	67	2.03
yellow/green pod #2	35	65	1.86
Total	560	1124	2.01

Table 14.1: Tests of Heterozygosity of Dominant Trait Phenotypes

	N	μ	σ	h	$\lvert h-\mu\rvert/\sigma$
round/wrinkled seed	565	188	11	193	0.5
green/yellow seed	519	173	11	166	0.6
colored/white flower	100	33	5	36	0.6
tall/dwarf plant	100	33	5	28	1
constricted pod or not	100	33	5	29	0.8
green/yellow pod #1	100	33	5	40	1.4
terminal flower or not	100	33	5	33	0
yellow/green pod #2	100	33	5	35	0.4

as measured in units where 1 means 1 standard deviation.

Those who complained about the spread of Mendel's data presumably made a table just like the one just given, and then noticed the following interesting fact: In 7 of the 8 cases, the observed value for h has distance less than or equal to 1 standard deviation from the mean. They must have thought that this clustering around the mean was not likely to arise by chance.

What follows describes a method to estimate a P-value for having 7 of 8 measurements land within one standard deviation of the mean. To do this invoke the Central Limit theorem in each of the eight cases to say that the probability that h has a given value, say x, should be determined to great accuracy by the Gaussian probability function:

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/(2\sigma^2)}. \tag{14.16}$$

Granted that the Central Limit theorem is applicable, then the probability of h having value within 1 standard deviation from the mean is

$$q = \int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/(2\sigma^2)}\, dx \approx 0.68. \tag{14.17}$$

This has the following consequence: The probability of getting $n \in \{0, 1, \ldots, 8\}$ out of 8 measurements to lie within 1 standard deviation is given by the $N = 8$ binomial probability function using $q = 0.68$. This is $p(n) = \frac{8!}{n!\,(8-n)!}(0.68)^n(0.32)^{8-n}$. Note that the mean this binomial function is $8(0.68) = 5.44$. Thus, the P-value for getting 7 of 8 measurements within 1 standard deviation of the mean is

$$p(0) + p(1) + p(2) + p(3) + p(7) + p(8).$$

As it turns out this ≈ 0.58, thus greater than 0.05.

Looking back on this discussion, note that I invoked binomial probability functions in two different places. The first was to derive the right most column in the table. This used a $q = \frac{1}{3}$ binomial with values of N given by the left-most column of numbers in the table. The value of came from the hypothesis about the manner in which genes are passed from parent plant to seedling. The second application of the binomial probability function used the $q = 0.68$ version with $N = 8$. I used this version to compute a P-value for having 7 of 8 measurements land within 1 standard deviation of the mean. The value 0.68 for q was justified by an appeal to the Central Limit theorem.

14.7 Boston weather 2008

This last part of the chapter briefly discusses a case where hypothesis testing might say something interesting. The following chart from the Sunday, January 4 New York Times gives the daily temperature extremes in Boston for the year 2008: The chart claims that the average temperature for the year was 0.5 degrees above normal. The grayish

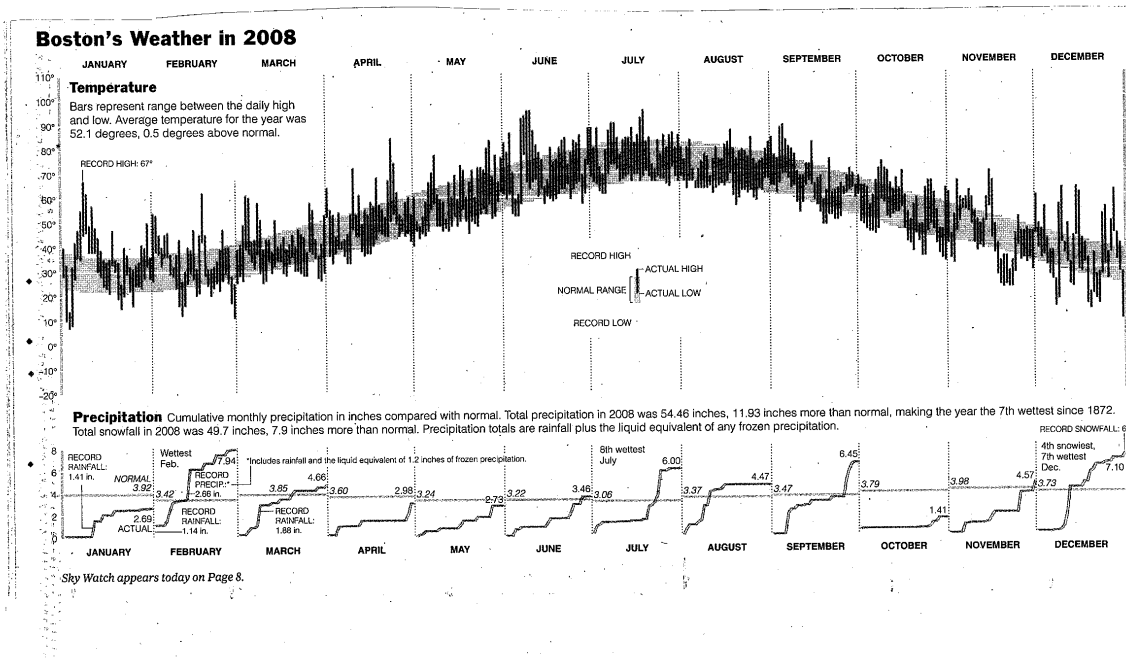

Figure 14.1: Temperature Extremes in Boston (NY Times, Sunday, January 4, 2009)

swath in the background is meant to indicate the "normal range". (Nothing is said about what this means.)

We hear all the time about "global warming". Does this chart give evidence for warming in Boston during the year 2008? What follows describe what may be an overly simplistic way to tease some answer to this question from the chart. Let us assume that the center of the grey swath on each day represents the mean temperature on the given day as computed by averaging over some large number of years. This gives us an ordered set of 366 numbers, $\{T_1, \ldots, T_{366}\}$. Meanwhile, the chart gives us, for each day in 2008, the actual mean temperature on that day, this half of the sum of the indicated high and low temperature on that day. This supplies us with a second sequence of 366 numbers, $\{t_1, \ldots, t_{366}\}$. Subtracting the first sequence from the second gives us numbers $(x_1 = t_1 - T_1, \ldots, x_{366} = t_{366} - T_{366})$. Take this set of 366 numbers to be the experimentally measured data.

The chart tells us that the average, $\mu_{\text{real}} = \frac{1}{366}(x_1 + \cdots + x_{366})$, of the experimental data is equal to 0.5. We can then ask if this average is consistent with the hypothesis that the sequence $(x_1, \ldots, x_{366})$ is distributed according to a Gaussian distribution with mean zero and some as yet unknown standard deviation. Unfortunately, the chart doesn't say anything about the standard deviation. This understood, let us proceed with the assumption that the standard deviation is half the vertical height of the grey swath, which is to say roughly 6° degrees Fahrenheit. If we make this assumption, then we are comparing the number 0.5 with what would be expected using the Central Limit theorem with input a Gaussian with mean zero and standard deviation 6°. The Central Limit theorem asserts that the P-value of the measured mean 0.5 as computed using $N = 366$ and this hypothetical Gaussian is ≈ 0.14, which is not significant.

We can also ask whether the variance of the data set $(x_1, \ldots, x_{366})$ is consistent with the assumption that these numbers are distributed according a Gaussian with mean $\mu = 0$ and standard deviation $\sigma = 6^\circ$. This is to say that we are comparing the experimentally variance $\sigma^2_{\text{real}} = \frac{1}{366}(x_1^2 + \cdots + x_{366}^2)$ with the average of 366 distances from the mean as computed using the Central Limit theorem for the Gaussian with mean $\mu = 0$ and $\sigma = 6^\circ$.

Of course, we can go further. For example, when I look at the graph, I sense that there are more peaks that are very much above the grey swath than there are peaks that are very much below the grey swath. I can test whether this is consistent with the Gaussian assumption by using the central limit theorem for the average of $N = 366$ measurements of the random variable $f(x) = \frac{1}{2}(x + |x|)$. (Note that $f(x) = 0$ when $x \leq 0$ and $f(x) = x$ when $x > 0$.)

14.8 Exercises:

1. Suppose that we expect that the x-coordinate of bacteria in our rectangular petri dish should be any value between -1 and 1 with equal probability in spite of our having coated the $x = 1$ wall of the dish with a specific chemical. We observe the positions of 900 bacteria in our dish and so obtain 900 values, $\{z_1, \ldots, z_{900}\}$, for the x-coordinates.

   (a) Suppose the average, $\overline{z} = \frac{1}{900} \sum_{1 \leq k \leq 900} z_k$, is 0.1. Use the Central Limit Theorem to obtain a theoretical upper bound based on our model of a uniform probability function for the probability that an average of 900 x-coordinates differs from 0 by more than 0.1.

   (b) Suppose that the average of the squares, $\text{Var} = \sum_{1 \leq k \leq 900} z_k^2$, equals 0.36. Use the Central Limit Theorem and (13.21) to obtain a theoretical upper bound based on our model of a uniform probability function for the probability that an average of the squares of 900 x-coordinates is greater than or equal to 0.36. (Note that I am not asking that it differ by a certain amount from the square of the standard deviation for the uniform probability function. If you compute the latter, you will be wrong by a factor of 2.)

2. Use Stirling's formula in Equation (11.14) to give approximate formulae for both $\frac{(2k)!}{k!2^k}$ and $\frac{1}{2^k}\left(\frac{(2k)!}{k!} - \left(\frac{k!}{\frac{k}{2}!}\right)^2\right)$ when k is large.

3. R. A. Fisher (see also reference[2] above) discussed a second criticism of Mendel's experimental data. This involved the manner in which a given dominant phenotype plant was classified as being "homozygous dominant" or "heterozygous dominant". According to Fisher, Mendel used the following method on any given plant: He germinated 10 seeds from the plant via self-pollination, and if all 10 of the resulting seedlings had the dominant phenotype, he then labeled the plant as "homozygous dominant". If one or more of the 10 seedlings had the recessive phenotype, he labeled the plant as "heterozygous dominant". Fisher pointed out that if Mendel really did the classification in this manner, then he should have mislabeled some heterozygous dominant plants as homozygous dominant. The following questions walk you through some of Fisher's arguments.

   (a) What is the probability for a heterozygous dominant plant to produce a seedling with the dominant phenotype?

   (b) What binomial probability function should you use to compute the probability that a heterozygous plant produces 10 consecutive dominant seedlings.

   (c) Use the binomial probability function for (b) to compute the probability of any given heterozygous dominant plant to produce 10 consecutive dominant seedlings.

   (d) If any given plant has probability to be homozygous dominant and thus to be heterozygous dominant, what is the probability that Mendel would label any given plant as "homozygous dominant"? (To answer this, you can use conditional probabilities. To this end, suppose that you have a sample space of N plants. Use A to denote the subset of plants that are homozygous dominant, B to denote the subset that are heterozygous dominant, and C to denote the subset that Mendel designates are homozygous dominant.)

   (e) Redo the second table in this chapter based on your answer to (c) of this chapter.

4. The 2006 election for the United States Senator in Virginia had the following outcome (according to `cnn.com`):

   Webb: $1,172,671$ votes
   Allen: $1,165,440$ votes

The total votes cast for the two candidates was $2,238,111$, and the difference in the vote totals was $7,231$. This was considered by the press to be an extremely tight election. Was it unreasonably close? Suppose that an election to choose one of two candidates is held with $2,238,111$ voters. Suppose, in addition, that each voter casts his or her vote at random. If you answer correctly the questions (b) and (c) below, you will obtain an estimate for the probability as determined by this "vote at random" model that the difference between the two candidates is less than or equal to $7,231$.

(a) Set $N = 2,238,111$. Let S denote the sample space whose elements are sequences of the form $\{z_1, \dots, z_N\}$ where each z_k is either 1 or -1. Use f to denote the random variable on S that is given by the formula $f(z_1, \dots, z_N) = z_1 + z_2 + \cdots + z_N$. What is the mean and what is the standard deviation of f?

(b) Use the Central Limit Theorem to find a Gaussian probability function that can be used to estimate the probability that $\frac{1}{N} f$ has value in any given interval on the real line.

(c) Use the Gaussian probability from (b) to compute the probability that $|\frac{1}{N} f| \leq \frac{7,231}{N}$. Note that this is the probability (as computed by this Gaussian) for $|f|$ to be less than or equal to $7,231$. (You can use a calculator if you like to compute the relevant integral.)

CHAPTER

FIFTEEN

Determinants

This chapter is meant to provide some cultural background to the story told in the linear algebra text book about determinants. As the text explains, the determinant of a square matrix is non-zero if and only if the matrix is invertible. The fact that the determinant signals invertibility is one of its principle uses. The other is the geometric fact observed by the text that the absolute value of the determinant of the matrix is the factor by which the linear transformation expands or contracts n-dimensional volumes. This chapter considers an invertibility question of a biological sort, an application to a protein folding problem.

Recall that a protein molecule is a long chain of smaller molecules that are tied end to end. Each of these smaller molecules can be any of twenty so called amino acids. This long chain appears in a cell folded up on itself in a complicated fashion. In particular, its interactions with the other molecules in the cell are determined very much by the particular pattern of folding because any given fold will hide some amino acids on its inside while exhibiting others on the outside. This said, one would like to be able to predict the fold pattern from knowledge of the amino acid that occupies each site along the chain.

In all of this, keep in mind that the protein is constructed in the cell by a component known as a "ribosome", and this construction puts the chain together starting from one end by sequentially adding amino acids. As the chain is constructed, most of the growing chain sticks out of the ribosome. If not stabilized by interactions with surrounding molecules in the cell, a given link in the chain will bend this way or that as soon as it exits the ribosome, and so the protein would curl and fold even as it is constructed.

To get some feeling for what is involved in predicting the behavior here, make the grossly simplifying assumption that each of the amino acids in a protein molecule of length N can bend in one of 2 ways, but that the probability of bending, say in the $+$ direction for the nth amino acid is influenced by the direction of bend of its nearest neighbors, the amino acids in sites $n-1$ and $n+1$. One might expect something like this for short protein molecules in as much as the amino acids have electrical charges on them and so feel an electric force from their neighbors. As this force is grows weaker with distance, their nearest neighbors will affect them the most. Of course, once the chain folds back on itself, a given amino acid might find itself very close to another that is actually some distance away as measured by walking along the chain.

In any event, let us keep things very simple and suppose that the probability, $p_n(t)$, at time t of the nth amino acid being in the $+$ fold position evolves as

$$p_n(t+1) = a_n + A_{n,n}p_n(t) + A_{n,n-1}p_{n-1}(t) + A_{n,n+1}p_{n+1}(t). \tag{15.1}$$

Here, a_n is some fixed number between 0 and 1, and the numbers $\{a_n, A_{n,n}, A_{n,n\pm1}\}$ are constrained so that

$$0 \le a_n + A_{n,n}x + A_{n,n-1}\,x_- + A_{n,n+1}x_+ \le 1 \tag{15.2}$$

for any choice between 0 and 1 of values for x, x_- and x_+ with $x + x_- + x_+ \le 1$. This constraint is necessary to guarantee that $p_n(t+1)$ is between 0 and 1 if each of $p_n(t)$, $p_{n+1}(t)$ and $p_{n-1}(t)$ is between 0 and 1. In this regard, let me remind you that $p_n(t+1)$ must be between 0 and 1 if it is the probability of something. My convention here takes both $A_{1,0}$ and $A_{N,N+1}$ to be zero.

As for the values of the other coefficients, I will suppose that knowledge of the amino acid type that occupies site n

is enough to determine a_n and A_{nn}, and that knowledge of the respective types that occupy sites $n-1$ and $n+1$ is enough to determine $A_{n,n-1}$ and $A_{n,n+1}$. In this regard, I assume access to a talented biochemist.

Granted these formulæ, the N-component vector $\vec{p}(t)$ whose nth component is $p_n(t)$ evolves according to the rule:

$$\vec{p}(t+1) = \vec{a} + A\vec{p}(t) \tag{15.3}$$

where $\vec{a}$ is that N-component vector whose nth entry is a_n, and where A is that $N \times N$ matrix whose only non-zero entries are $\{A_{n,n}, A_{n,n\pm1}\}_{1\leq n\leq N}$.

We might now ask if there exists an *equilibrium* probability distribution, an N-component vector, $\vec{p}$, with non-negative entries that obeys

$$\vec{p} = \vec{a} + A\vec{p} \tag{15.4}$$

If there is such a vector, then we might expect its entries to give us the probabilities for the bending directions of the various links in the chain for the protein. From this, one might hope to compute the most likely fold pattern for the protein.

To analyze (15.4), let us rewrite it as the equation

$$(I - A)\vec{p} = \vec{a}, \tag{15.5}$$

where I here denotes the identity matrix; this the matrix with its only entries on the diagonal and with all of the latter equal to 1. We know now that there is some solution, $\vec{p}$, when $\det(I-A) \neq 0$. It is also unique in this case. Indeed, were there two solutions, $\vec{p}$ and $\vec{p}'$, then

$$(I - A)(\vec{p} - \vec{p}') = 0. \tag{15.6}$$

This implies $(I-A)$ has a kernel, which is forbidden when $I-A$ is invertible. Thus, to understand this version of the protein folding problem, we need to consider whether the matrix $I-A$ is invertible. As remarked at the outset, this is the case if and only if it has non-zero determinant.

By the way, we must also confirm that the solution, $\vec{p}$, to (15.5) has its entries between 0 and 1 so as to use the entries as probabilities.

In any event, to give an explicit example, consider the 3×3 case. In this case, the matrix $I - A$ is

$$\begin{pmatrix} 1-A_{11} & -A_{12} & 0 \\ -A_{21} & 1-A_{12} & -A_{23} \\ -A_{31} & -A_{32} & 1-A_{12} \end{pmatrix}. \tag{15.7}$$

Using the formulae from Chapter 6 of the linear algebra text book, its determinant is found to be

$$\det(I-A) = (1-A_{11})(1-A_{22})(1-A_{33}) - (1-A_{11})A_{23}A_{32} - (1-A_{33})A_{12}A_{21}. \tag{15.8}$$

CHAPTER

SIXTEEN

Eigenvalues in biology

My goal in this chapter is to illustrate how eigenvalues can appear in problems from biology.

16.1 An example from genetics

Suppose we have a fixed size N population of cells that reproduce by fission. The model here is that there are N cells that divide at the start, thus producing $2N$ cells. After one unit of time, half of these die and half survive, so there are N cells to fission at the end of 1 unit of time. These N survivors divide to produce $2N$ cells and so start the next run of the cycle. In particular, there are always N surviving cells at the end of one unit of time and then $2N$ just at the start of the next as each of these N cells splits in half.

Now, suppose that at the beginning, $t = 0$, there is, for each $n \in \{0, 1, \dots, N\}$, a given probability which I'll call $p_n(0)$ for n of the N initial cells to carry a certain trait. Suppose that this trait (red color as opposed to blue) is neutral with respect to the cell's survival. In other words, the probability is for any given red cell to survive to reproduce, and the probability is for any given blue cell to survive to reproduce.

Here is the key question: Given the initial probabilities, $\{p_0(0), p_1(0), \dots, p_N(0)\}$, what are the corresponding probabilities after some t generations? Thus, what are the values of $\{p_0(t), p_1(t), \dots, p_N(t)\}$ where now $p_n(t)$ denotes the probability that n cells of generation t carry the red color?

To solve this, we can use our tried and true recourse to conditional probabilities by noting that

$$\begin{aligned} p_n(t) = {} & \mathrm{P}\,(n \text{ red survivors} \,|\, 0 \text{ red parents}) \cdot p_0(t-1) \\ & + \mathrm{P}\,(n \text{ red survivors} \,|\, 1 \text{ red parent}) \cdot p_1(t-1) \\ & + \mathrm{P}\,(n \text{ red survivors} \,|\, 2 \text{ red parents}) \cdot p_2(t-1) + \cdots \end{aligned} \tag{16.1}$$

where $\mathrm{P}\,(n \text{ red survivors} \,|\, m \text{ red parents})$ is the conditional probability that there are n red cells at the end of a cycle given that there were m such cells the end of the previous cycle. If the ambient environmental conditions don't change, one would expect that these conditional probabilities are independent of time. As I explain below, they are, in fact, computable from what we are given about this problem. In any event, let me use the shorthand A to denote the square matrix of size $N + 1$ whose entry in row n and column m is $\mathrm{P}\,(n \text{ red survivors} \,|\, m \text{ red parents})$. In this regard, note that n and m run from 0 to N, not the from 1 to $N + 1$. Let $\vec{p}(t)$ denote the vector in $\mathbf{R}^{N+1}$ whose nth entry is $p_n(t)$. Then (16.1) reads

$$\vec{p}(t) = A\vec{p}(t-1). \tag{16.2}$$

This last equation would be easy to solve if we knew that $\vec{p}(0)$ was an *eigenvector* of the matrix A. That is, if it were the case that

$$A\vec{p}(0) = \lambda\vec{p}(0) \qquad \text{with } \lambda \text{ some real number.} \tag{16.3}$$

Indeed, were this the case, then the $t = 1$ version of (16.2) would read $\vec{p}(1) = \lambda\vec{p}(0)$. We could then use the $t = 2$ version of (16.2) to compute $\vec{p}(2)$ and we would find that $\vec{p}(2) = \lambda A\vec{p}(0) = \lambda^2\vec{p}(0)$. Continuing in the vein, we would find that $\vec{p}(t) = \lambda^t\vec{p}(0)$ and our problem would be solved.

Now, it is unlikely that $\vec{p}(0)$ is going to be an eigenvector. However, even if $\vec{p}(0)$ is a linear combination of eigenvectors,

$$\vec{p}(0) = c_1\vec{e}_1 + c_2\vec{e}_2 + \cdots , \tag{16.4}$$

we could still solve for $\vec{p}(t)$. Indeed, let λ_k denote the eigenvalue for any given $\vec{e}_k$. Thus, $A\vec{e}_k = \lambda_k\vec{e}_k$. Granted we know these eigenvalues and eigenvectors, then we can plug (16.4) into the $t = 1$ version of (16.2) to find

$$\vec{p}(1) = c_1\lambda_1\vec{e}_1 + c_2\lambda_2\vec{e}_2 + \cdots . \tag{16.5}$$

We could then plug this into the $t = 2$ version of (16.2) to find that

$$\vec{p}(2) = c_1\lambda_1^2\vec{e}_1 + c_2\lambda_2^2\vec{e}_2 + \cdots , \tag{16.6}$$

and so on. In general, this then gives

$$\vec{p}(t) = c_1\lambda_1^t\vec{e}_1 + c_2\lambda_2^t\vec{e}_2 + \cdots . \tag{16.7}$$

16.2 Transition/Markov matrices

The matrix A that appears in (16.2) has entries that are conditional probabilities, and this has the following implications:

- All entries are non-negative.
- $A_{0,m} + A_{1,m} + \cdots + A_{N,m} = 1$ for all $m \in \{0, 1, \ldots, N\}$.

(16.8)

The last line above asserts that the probability is 1 of there being *some* number, either 0 or 1 or 2 or ... or N of red cells in the subsequent generation given m red cells in the initial generation.

A square matrix with this property is called a *transition matrix*, or sometimes a *Markov matrix*. When A is such a matrix, the equation in (16.2) is called a *Markov process*.

Although we are interested in the eigenvalues of A, it is amusing to note that the transpose matrix, A^T, has an eigenvalue equal to 1 with the corresponding eigenvector being proportional to the vector, $\vec{a}$, with each entry equal to 1. Indeed, the entry in the mth row and nth column of A^T is A_{nm}, this the entry of A in the nth row and mth column. Thus, the mth entry of $A^T\vec{a}$ is

$$(A^T\vec{a})_m = A_{0,m}a_0 + A_{1,m}a_1 + \cdots + A_{N,m}a_N. \tag{16.9}$$

If the lower line in (16.10) is used and if each a_k is 1, then each entry of $A^T\vec{a}$ is also 1. Thus, $A^T\vec{a} = \vec{a}$.

As a "cultural" aside, what follows is the story on A_{nm} in the example from Section 16.1. First, $A_{nm} = 0$ if n is larger than $2m$ since m parent cells can spawn at most $2m$ survivors. For $n \le m$, consider that you have $2N$ cells of which $2m$ are red and you ask for the probability that a choice of N from the $2N$ cells results in n red ones. This is a counting problem that is much like those discussed in Section 11.3 although more complicated. The answer here is:

$$A_{nm} = \frac{(N!)^2}{(2N)!}\,\frac{(2m)!}{n!\,(2m-n)!}\,\frac{(2N-2m)!}{(N-m)!\,(N+n-2m)!} \qquad \text{when } 0 \le n \le 2m. \tag{16.10}$$

16.3 Another protein folding example

Here is another model for protein folding. As you may recall from previous chapters, a protein is made of segments tied end to end as a chain. Each segment is one of 20 amino acids. The protein is made by the cell in a large and complex molecule called a *ribosome*. The segments are attached one after the other in the ribosome and so the chain grows, link by link. Only a few segments are in the ribosome, and the rest stick out as the protein grows. As soon as a joint between two segments is free of the ribosome, it can bend if it is not somehow stabilized by surrounding

molecules. Suppose that the bend at a joint can be in one of n directions as measured relative to the direction of the previously made segment. A simplistic hypothesis has the direction of the bend of any given segment influenced mostly by the bend direction of the previously made segment.

To model this, let me introduce, for $k, j \in \{1, \dots, n\}$, the conditional probability, $A_{k,j}$ that the a given segment has bend direction k when the previously made segment has bend direction j. Next, agree to label the segments along the protein by the integers in the set $\{1, \dots, N\}$. If z denotes such an integer, let $p_k(z)$ denote the probability that the zth segment is bent in the kth direction relative to the angle of the $z-1$st segment. Then we have

$$p_k(z) = \sum_{j=1}^{n} A_{k,j} p_j(z-1). \tag{16.11}$$

I can now introduce the vector $\vec{p}(z)$ in $\mathbf{R}^n$ whose kth component is $p_k(z)$, and also the square matrix A with the components $A_{k,j}$. Then (16.11) is the equation

$$\vec{p}(z) = A\vec{p}(z-1). \tag{16.12}$$

Note that as in the previous case, the matrix A is a Markov matrix. This is to say that each A_{kj} is non-negative because they are conditional probabilities; and

$$A_{1,j} + A_{2,j} + \cdots + A_{n,j} = 1 \tag{16.13}$$

because the segment must be at some angle or other.

Here is an explicit example: Take $n = 3$ so that A is a 3×3 matrix. In particular, take

$$A = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}. \tag{16.14}$$

As it turns out, A has the following eigenvectors:

$$\vec{e}_1 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \vec{e}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \qquad \text{and} \qquad \vec{e}_3 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} \tag{16.15}$$

with corresponding eigenvalues $\lambda_1 = 1$, $\lambda_2 = \frac{1}{4}$ and $\lambda_3 = \frac{1}{4}$. Note that the vectors in (16.15) are mutually orthogonal and have norm 1, so they define an orthonormal basis for $\mathbf{R}^3$. Thus, any given z version of $\vec{p}(z)$ can be written as a linear combination of the vectors in (16.15). Doing so, we write

$$\vec{p}(z) = c_1(z)\vec{e}_1 + c_2(z)\vec{e}_2 + c_3(z)\vec{e}_3 \tag{16.16}$$

where each $c_k(z)$ is some real number. In this regard, there is one point to make straight away, which is that $c_1(z)$ must equal $\frac{1}{\sqrt{3}}$ when the entries of $\vec{p}(z)$ represent probabilities that sum to 1. To explain, keep in mind that the basis in (16.15) is an *orthonormal* basis, and this implies that $\vec{e}_1 \cdot \vec{p} = c_1(z)$. However, since each entry of $\vec{e}_1$ is equal to $\frac{1}{\sqrt{3}}$, this dot product is $\frac{1}{\sqrt{3}}(p_1(z) + p_2(z) + p_3(z))$ and so equals $\frac{1}{\sqrt{3}}$ when $p_1 + p_2 + p_3 = 1$ as is the case when these coefficients are probabilities.

In any event, if you plug the expression in (16.16) into the left side of (16.12) and use the analogous $z-1$ version on the right side, you will find that the resulting equation holds if and only if the coefficients obey

$$c_1(z) = c_1(z-1), \qquad c_2(z) = \frac{1}{4}c_2(z-1) \qquad \text{and} \qquad c_3(z) = \frac{1}{4}c_3(z-1). \tag{16.17}$$

Note that the equality here between $c_1(z)$ and $c_1(z-1)$ is heartening in as much as both of them are supposed to equal $\frac{1}{\sqrt{3}}$. Anyway, continue by iterating (16.17) by writing the $z-1$ versions of c_k in terms of the $z-2$ versions, then the latter in terms of the $z-3$ versions, and so on until you obtain

$$c_1(z) = c_1(0), \qquad c_2(z) = \left(\frac{1}{4}\right)^z c_2(0) \qquad \text{and} \qquad c_3(z) = \left(\frac{1}{4}\right)^z c_3(0). \tag{16.18}$$

Thus, we see from (16.18) we have

$$\vec{p}(z) = \frac{1}{3}\begin{pmatrix}1\\1\\1\end{pmatrix} + \left(\frac{1}{4}\right)^z \begin{pmatrix} -\frac{1}{\sqrt{2}}c_2(0) + \frac{1}{\sqrt{6}}c_3(0) \\ \frac{1}{\sqrt{2}}c_2(0) + \frac{1}{\sqrt{6}}c_3(0) \\ \sqrt{\frac{2}{3}}c_3(0) \end{pmatrix}. \tag{16.19}$$

By the way, take note of how the probabilities for the three possible fold directions come closer and closer to being equal as z increases even if the initial $z = 0$ probabilities were drastically skewed to favor one or the other of the three directions. For example, suppose that

$$\vec{p}(0) = \begin{pmatrix}1\\0\\0\end{pmatrix} \tag{16.20}$$

As a result, $c_1(0) = \frac{1}{\sqrt{3}}$, $c_2(0) = -\frac{1}{\sqrt{2}}$ and $c_3(0) = \frac{1}{\sqrt{6}}$. Thus, we have

$$\vec{p}(z) = \frac{1}{3}\begin{pmatrix}1\\1\\1\end{pmatrix} + \left(\frac{1}{4}\right)^z \begin{pmatrix} \frac{2}{3} \\ -\frac{1}{3} \\ -\frac{1}{3} \end{pmatrix}. \tag{16.21}$$

Thus, by $z = 4$, the three directions differ in probability by less than 1%.

16.4 Exercises:

1. Multiply the matrix A in (16.14) against the vector $\vec{p}(z)$ in (16.19) and verify that the result is equal to $\vec{p}(z+1)$ as defined by replacing z by $z+1$ in (16.19).
2. Let A denote the 2×2 matrix $\begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix}$. The vectors $\vec{e}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}$ and $\vec{e}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}-1\\1\end{pmatrix}$ are eigenvectors of A.
   (a) Compute the eigenvalues of $\vec{e}_1$ and $\vec{e}_2$.
   (b) Suppose $z \in \{0, 1, \dots\}$ and $\vec{p}(z+1) = A\vec{p}(z)$ for all z. Find $\vec{p}(z)$ if $\vec{p}(0) = \begin{pmatrix} \frac{1}{4} \\ \frac{3}{4} \end{pmatrix}$.
   (c) Find $\vec{p}(z)$ if $\vec{p}(0) = \begin{pmatrix}0\\1\end{pmatrix}$.
   (d) Find $\vec{p}(z)$ in the case that $\vec{p}(z+1) = \begin{pmatrix}1\\-1\end{pmatrix} + A\vec{p}(z)$ for all z and $\vec{p}(0) = \begin{pmatrix}0\\1\end{pmatrix}$.
3. Suppose that you have a model to explain your data that predicts the probability of a certain measurement having any prescribed value. Suppose that this probability function has mean 1 and standard deviation 2.
   (a) Give an upper bound for the probability of a measurement being greater than 5.
   (b) Suppose that you average some very large number, N, of measurements that are taken in unrelated, but identical versions of the same experimental set up. Write down a Gaussian probability function that you can use to estimate the probability that the value of this average is greater than 5. In particular, give a numerical estimate using this Gaussian function for $N = 100$.
   (c) Let us agree that you will throw out your proposed model if it predicts that the probability for finding an average value that is greater than your *measured* average for 100 measurements is less than $\frac{1}{20}$. If your measured average is 1.6 for 100 experiments, should you junk your model?

4. Suppose that you repeat the same experiment 100 times and each time record the value of a certain key measurement. Let $\{x_1, \ldots, x_{100}\}$ denote the values of the measurement in N experiments. Suppose that $\sum_{k=1}^{100} x_k = 0$ and $\frac{1}{99} \sum_{k=1}^{100} x_k^2 = 1$. Suppose, in addition, that five of the x_k obey $x_k > 3$. The purpose of this problem is to walk you through a method for obtaining an upper bound for the likelihood of having five measurements that far from the mean. In this regard, suppose that you make the hypothesis that the spread of the measured numbers $\{x_k\}_{1 \leq k \leq 100}$ is determined by the Gaussian probability function with mean 0 and standard deviation equal to 1.

   (a) Let p denote the probability using this Gaussian probability function for a measurement x with $x \geq 3$. Explain why $p \leq e^{-9/2} \leq 0.0015$.

   (b) Use the binomial probability function to explain why the probability of finding five or more values for k out of 100 that have $x_k > 3$ is equal to

$$\sum_{k=5}^{100} \frac{100!}{k!\,(100-k)!} p^k (1-p)^{100-k}.$$

   (c) Let $k \geq 5$ and let $z \to f_k(z) = z^k (1-z)^{100-k}$. Show that f is an increasing function of z when $z < \frac{1}{20}$. (Hint: Take its derivative with respect to z.)

   (d) Since $p \leq 0.0015 \leq 0.05 = \frac{1}{20}$, use the result from part (c) to conclude that the probability of finding 5 or more of the x_k's out of 100 with $x_k > 3$ has probability less than

$$\sum_{k=5}^{100} (0.0015)^k (0.9985)^{100-k}.$$

   (e) We saw in Equation (11.19) of Chapter 11 that the terms in the preceding sum get ever smaller as k increases. Use a calculator to show that the $k = 5$ term and thus all higher k terms are smaller than 5×10^{-7}. Since there are less than $95 \leq 100$ of these terms prove that the sum of these terms is no greater than 0.00005.

CHAPTER

SEVENTEEN

More about Markov matrices

The notion of a Markov matrix was introduced in Chapter 15. By way of a reminder, this is an $N \times N$ matrix A that obeys the following conditions:

$$\begin{aligned} &\bullet\ A_{jk} \geq 0 \ \text{ for all } j \text{ and } k. \\ &\bullet\ A_{1k} + A_{2k} + \cdots + A_{Nk} = 1 \text{ for all } k. \end{aligned} \tag{17.1}$$

These two conditions are sufficient (and also necessary) for the interpretation of the components of A as conditional probabilities. To this end, imagine a system with N possible states, labeled by the integers starting at 1. Then A_{jk} can represent the conditional probability that the system is in state j at time t if it is in state k at time $t-1$. In particular, one sees Markov matrices arising in a dynamical system where the probabilities for the various states of the system at any given time are represented by an N-component vector, $\vec{p}(t)$, that evolves in time according to the formula

$$\vec{p}(t) = A\vec{p}(t-1). \tag{17.2}$$

Here is a 3×3 Markov matrix

$$\begin{bmatrix} \frac{1}{4} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{4} & \frac{1}{3} & \frac{1}{6} \end{bmatrix}. \tag{17.3}$$

All entries are non-negative, and the entries in each column sum to 1. Note that there is no constraint for the sum of the entries in any given row.

A question that is often raised in the general context of (17.1) and (17.2) is whether the system has an equilibrium probability function, thus some non-zero vector $\vec{p}_*$, with non-negative entries that sum to 1, and that obeys $A\vec{p}_* = \vec{p}_*$. If so, there is the associated question of whether the $t \to \infty$ limit of $\vec{p}(t)$ must necessarily converge to the equilibrium $\vec{p}_*$.

Here are some facts that allow us to answer these questions.

- Any Markov matrix always has an eigenvector with eigenvalue 1.
- If each entry of a Markov matrix A is strictly positive, then every non-zero vector in the kernel of $A - I$ has either all positive entries or all negative entries. Here, I denotes the identity matrix.
- If each entry of a Markov matrix A is strictly positive, there is a vector in the kernel of the matrix $A - I$ whose entries are positive and sum to 1.
- If each entry of a Markov matrix A is strictly positive, then the kernel of $A - I$ is one dimensional. Furthermore, there is unique vector in the kernel of $A - I$ whose entries are positive and sum to 1.
- If each entry of A is strictly positive, all other eigenvalues have absolute value strictly less than 1. Moreover, the entries of any eigenvector for an eigenvalue that is less than 1 must sum to zero.

(17.4)

I elaborate on these points in Section 17.2, below. Accept them for the time being.

17.1 Solving the equation

Let me suppose that A is a Markov matrix and that none of its entries are zero. This allows me to use all of the facts that are stated in (17.4). Let me also suppose that A has a basis of eigenvectors. In this case, (17.2) can be solved using the basis of eigenvectors. To make this explicit, I denote this basis of eigenvectors as $\{\vec{e}_1, \dots, \vec{e}_n\}$ and I use λ_k to denote the eigenvector of $\vec{e}_k$. Here, my convention has $\lambda_1 = 1$ and I take $\vec{e}_1$ to be the eigenvector with eigenvalue 1 whose entries sum to 1. Note that I am using the first four facts in (17.4) to conclude that A must have a unique eigenvector with eigenvalue 1 and positive entries that sum to 1.

Now, the solution to (17.2) depends on the starting vector, $\vec{p}(0)$. In the context where this is a vector of probabilities, it can not have a negative entry. Moreover, its entries must sum to 1. As explained below, this requires that

$$\vec{p}(0) = \vec{e}_1 + \sum_{k=2}^{n} c_k \vec{e}_k. \tag{17.5}$$

The point is that the coefficient in front of $\vec{e}_1$ is necessarily equal to 1. The coefficient, c_k, in front of any $k \geq 2$ version of $\vec{e}_k$ is not so constrained.

Here is why (17.5) must hold: In general, I can write $\vec{p}(0) = c_1\vec{e}_1 + \sum_{k=2}^{n} c_k\vec{e}_k$ where $c_1, c_2, \dots, c_n$ are real numbers because I am assuming that A is diagonalizable, and the eigenvectors of any diagonalizable matrix comprise a basis. It follows from this representation of $\vec{p}(0)$ that the sum of its entries is obtained by adding the following numbers: First, c_1 times the sum of the entries of $\vec{e}_1$, then c_2 times the sum of the entries $\vec{e}_2$, then c_3 times the sum of the entries of $\vec{e}_3$, and so on. However, the last point of (17.4) asserts that the sum of the entries of each $k \geq 2$ version of $\vec{e}_k$ is zero. This means that the sum of the entries of $\vec{p}(0)$ is c_1 times the sum of the entries of $\vec{e}_1$. Since the sum of the entries of $\vec{e}_1$ is 1 and since this is also the sum of the entries of $\vec{p}(0)$, so c_1 must equal 1. This is what is asserted by (17.5).

By way of an example, consider the 2×2 case where

$$A = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} \\ \frac{3}{4} & \frac{1}{2} \end{bmatrix}. \tag{17.6}$$

The eigenvalues in this case are 1 and $-\frac{1}{4}$ and the associated eigenvectors $\vec{e}_1$ and $\vec{e}_2$ in this case can be taken to be

$$\vec{e}_1 = \begin{bmatrix} \frac{2}{5} \\ \frac{3}{5} \end{bmatrix} \qquad \text{and} \qquad \vec{e}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{17.7}$$

For this 2×2 example, the most general form for $\vec{p}(0)$ that allows it to be a vector of probabilities is

$$\vec{p}(0) = \begin{bmatrix} \frac{2}{5} \\ \frac{3}{5} \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{2}{5} + c_2 \\ \frac{3}{5} - c_2 \end{bmatrix}. \tag{17.8}$$

where c_2 can be any number that obeys $-\frac{2}{5} \leq c_2 \leq \frac{3}{5}$.

Returning to the general case, it then follows from (17.2) and (17.4) that

$$\vec{p}(t) = \vec{e}_1 + \sum_{k=2}^{n} c_k \lambda_k^t \vec{e}_k. \tag{17.9}$$

Thus, as $t \to \infty$, we see that in the case where each entry of A is positive, the limit is

$$\lim_{t \to \infty} \vec{p}(0) = \vec{e}_1. \tag{17.10}$$

Note in particular that $\vec{p}(t)$ at large t is very nearly $\vec{e}_1$. Thus, if you are interested only in the large t behavior of $\vec{p}(t)$, you need only find one eigenvector!

In our 2×2 example,

$$\vec{p}(t) = \begin{bmatrix} \frac{2}{5} + \left(-\frac{1}{4}\right)^t c_2 \\ \frac{3}{5} - \left(-\frac{1}{4}\right)^t c_2 \end{bmatrix}. \tag{17.11}$$

In the example provided by (17.3), the matrix has eigenvalues 1, 0 and $-\frac{1}{4}$. If I use $\vec{e}_2$ for the eigenvector with eigenvalue 0 and $\vec{e}_3$ for the eigenvector with eigenvalue $-\frac{1}{4}$, then the solution with $\vec{p}(0) = \vec{e}_1 + c_2\vec{e}_2 + c_3\vec{e}_3$ is

$$\vec{p}(t) = \vec{e}_1 + c_3 \left(-\tfrac{1}{4}\right)^t \vec{e}_3 \qquad \text{for } t > 0. \tag{17.12}$$

As I remarked above, I need only find $\vec{e}_1$ to discern the large t behavior of $\vec{p}(t)$; and in the example using the matrix in (17.3),

$$\vec{e}_1 = \begin{bmatrix} \frac{4}{15} \\ \frac{7}{15} \\ \frac{4}{15} \end{bmatrix}. \tag{17.13}$$

17.2 Proving things about Markov matrices

My goal here is to convince you that a Markov matrix obeys the facts that are stated by the various points of (17.4).

Point 1: As noted, an equilibrium vector for A is a vector, $\vec{p}$, that obeys $A\vec{p} = \vec{p}$. Thus, it is an eigenvector of A with eigenvalue 1. Of course, we have also imposed other conditions, such as its entries must be non-negative and they should sum to 1. Even so, the first item to note is that A does indeed have an eigenvector with eigenvalue 1. To see why, observe that the vector $\vec{w}$ with entries all equal to 1 obeys

$$A^T \vec{w} = \vec{w} \tag{17.14}$$

by virtue of the second condition in (17.1). Indeed, if k is any given integer, then the kth entry of A^T is $A_{1k} + A_{2k} + \cdots + A_{Nk}$ and this sum is assumed to be equal to 1, which is the kth entry of $\vec{w}$.

This point about A^T is relevant since $\det(A^T - \lambda I) = \det(A - \lambda I)$ for any real number λ. Because $A^T - I$ is not invertible, $\det(A^T - I)$ is zero. Thus, $\det(A - I)$ is also zero and so $A - I$ is not invertible. Thus it has a positive dimensional kernel. Any non-zero vector in this kernel is an eigenvector for A with eigenvalue 1.

Point 2: I assume for this point that all of A's entries are positive. Thus, all A_{jk} obey the condition $A_{jk} > 0$. I have to demonstrate to you that there is no vector in the kernel of $A - I$ with whose entries are not all positive or all negative. To do this, I will assume that I have a vector $\vec{v}$ that violates this conclusion and demonstrate why this last assumption is untenable. To see how this is going to work, suppose first that only the first entry of $\vec{v}$ is zero or negative and the rest are non-negative with at least one positive. Since $\vec{v}$ is an eigenvector of A with eigenvalue 1,

$$v_1 = A_{11}v_1 + A_{12}v_2 + \cdots + A_{1n}v_n. \tag{17.15}$$

As a consequence, (17.15) implies that

$$(1 - A_{11})v_1 = A_{12}v_2 + \cdots + A_{1n}v_n \tag{17.16}$$

Now, A_{11} is positive, but it is less than 1 since it plus $A_{21} + A_{31} + \cdots + A_{n1}$ give 1 and each of $A_{21}, \ldots, A_{n1}$ is positive. This implies that the left-hand side of (17.16) is negative if it were the case that $v_1 < 0$, and it is zero if v_1 were equal to zero. Meanwhile, the right-hand side to (17.16) is strictly positive. Indeed, at least one $k > 1$ version of v_k is positive and its attending A_{1k} is positive, while no $k > 1$ versions of v_k are negative nor are any A_{jk}. Thus, were it the case that v_1 is not strictly positive, then (17.16) equates a negative or zero left-hand side with a positive right-hand side. Since this is nonsense, I see that I could never have an eigenvector of A with eigenvalue 1 that had one negative or zero entry with the rest non-negative with one or more positive.

Here is the argument when two or more of the v_ks are negative or zero and the rest are greater than or equal to zero with at least one positive: Let's suppose for simplicity of notation that v_1 and v_2 are either zero or negative, and that rest of the v_ks are either zero or positive. Suppose also that at least one $k \geq 3$ version of v_k is positive. Along with (17.15), we have

$$v_2 = A_{21}v_1 + A_{22}v_2 + A_{23}v_3 + \cdots + A_{2n}v_n. \tag{17.17}$$

Now add this equation to (17.15) to obtain

$$v_1 + v_2 = (A_{11} + A_{21})v_1 + (A_{12} + A_{22})v_2 + (A_{13} + A_{23})v_3 + \cdots + (A_{1n} + A_{2n})v_n. \tag{17.18}$$

According to (17.1), the sum $A_{11} + A_{21}$ is positive and strictly less than 1 since it plus the strictly positive $A_{31} + \cdots + A_{n1}$ is equal to 1. Likewise, $A_{21} + A_{22}$ is strictly less than 1. Thus, when I rearrange (17.18) as

$$(1 - A_{11} - A_{21})v_1 + (1 - A_{21} - A_{22})v_2 = (A_{13} + A_{23})v_3 + \cdots + (A_{1n} + A_{2n})v_n, \tag{17.19}$$

I again have an expression where the left-hand side is negative or zero and where the right hand side is greater than zero. Of course, such a thing can't arise, so I can conclude that the case with two negative versions of v_k and one or more positive versions can not arise.

The argument for the general case is very much like this last argument so I walk you through it in one of the exercises.

Point 3: I know from Point 1 that there is at least one non-zero eigenvector of A with eigenvalue 1. I also know, this from Point 2, that either all of its entries are negative or else all are positive. If all are negative, I can multiply it by -1 so as to obtain a new eigenvector of A with eigenvalue 1 that has all positive entries. Let r denote the sum of the entries of the latter vector. If I now multiply this vector by $\frac{1}{r}$, I get an eigenvector whose entries are all positive and sum to 1.

Point 4: To prove this point, let me assume, contrary to the assertion, that there are two non-zero vectors in the kernel of $A - I$ and one is not a multiple of the other. Let me call them $\vec{v}$ and $\vec{u}$. As just explained, I can arrange that both have only positive entries and that their entries sum to 1, this by multiplying each by an appropriate real number. Now, if $\vec{v}$ is not equal to $\vec{u}$, then some entry of one must differ from some entry of the other. For the sake of argument, suppose that $v_1 < u_1$. Since the entries sum to 1, this then means that some other entry of $\vec{v}$ must be *greater* than the corresponding entry of $\vec{u}$. For the sake of argument, suppose that $v_2 > u_2$. As a consequence the vector $\vec{v} - \vec{u}$ has negative first entry and positive second entry. It is also in the kernel of $A - I$. But, these conclusions are untenable since I already know that every vector in the kernel of $A - I$ has either all positive entries or all negative ones. The only escape from this logical nightmare is to conclude that $\vec{v}$ and $\vec{u}$ are equal.

This then demonstrates two things: First, there is a unique vector in the kernel of $A - I$ whose entries are positive and sum to 1. Second, any one vector in kernel $A - I$ is a scalar multiple of any other and so this kernel has dimension 1.

Point 5 : Suppose here that λ is an eigenvalue of A and that $\lambda > 1$ or that $\lambda \leq -1$. I need to demonstrate that this assumption is untenable and I will do this by deriving some patent nonsense by taking it to be true. Let me start by supposing only that λ is *some* eigenvector of A with out making the assumption about its size. Let $\vec{v}$ now represent some non-zero vector in the kernel of $A - \lambda I$. Thus, $\lambda\vec{v} = A\vec{v}$. If I sum the entries on both sides of this equation, I find that

$$\lambda\,(v_1 + \cdots + v_n) = (A_{11} + \cdots + A_{n1})v_1 + (A_{12} + \cdots + A_{n2})v_2 + \cdots + (A_{1n} + \cdots + A_{nn})v_n. \tag{17.20}$$

As a consequence of the second point in (17.1), this then says that

$$\lambda(v_1 + \cdots + v_n) = v_1 + \cdots + v_n. \tag{17.21}$$

Thus, either $\lambda = 1$ or else the entries of $\vec{v}$ sum to zero. This is what is asserted by the second sentence of the final point in (17.4).

Now suppose that $\lambda > 1$. In this case, the argument that I used in the discussion above for Point 2 can be reapplied with only minor modifications to produce the ridiculous conclusion that something negative is equal to something positive. To see how this works, remark that the conclusion that $\vec{v}$'s entries sum to zero implies that has at least one negative entry and at least one positive one. For example, suppose that the first entry of $\vec{v}$ is negative and the rest are either zero or positive with at least one positive. Since $\vec{v}$ is an eigenvector with eigenvalue λ, we have

$$\lambda v_1 = A_{11}v_1 + A_{12}v_2 + \cdots + A_{1n}v_n, \tag{17.22}$$

and thus

$$(\lambda - A_{11})v_1 = A_{12}v_2 + \cdots + A_{1n}v_n. \tag{17.23}$$

Note that this last equation is the analog in the $\lambda > 1$ case of (17.15). Well since $\lambda > 1$ and $A_{11} < 1$, the left-hand side of (17.23) is negative. Meanwhile, the right-hand side is positive since each A_{1k} that appears here is positive and since at least one $k \geq 2$ version of v_k is positive.

Equation (17.19) has its $\lambda > 1$ analog too, this where the $(1 - A_{11} - A_{21})$ is replaced by $(\lambda - A_{11} - A_{21})$ and where $(1 - A_{12} - A_{22})$ is replaced by $(\lambda - A_{12} - A_{22})$. The general case where has some $m < n$ negative entries and the rest zero or positive is ruled out by these same sorts of arguments.

Consider now the case where $\lambda \leq -1$. I can rule this out using the trick of introducing the matrix $A^2 = A \cdot A$. This is done in three steps.

Step 1: If A obeys (17.1) then so does A^2. If all entries of A are positive, then this is also the case for A^2. To see that all of this is true, note that the first point in (17.1) holds since each entry of A^2 is a sum of products of the entries of A and each of the latter is positive. As for the second point in (17.1), note that

$$\sum_{m=1}^{n}(A^2)_{mk} = \sum_{m=1}^{n}\left(\sum_{1\leq j\leq n} A_{mj}A_{jk}\right). \tag{17.24}$$

Now switch the orders of summing so as to make the right-hand side read

$$\sum_{m=1}^{n}(A^2)_{mk} = \sum_{j=1}^{n}\left(\sum_{m=1}^{n} A_{mj}A_{jk}\right). \tag{17.25}$$

The sum inside the parentheses is 1 for each j because A obeys the second point in (17.1). Thus, the right-hand side of (17.24) is equal to

$$\sum_{j=1}^{n} A_{jk}, \tag{17.26}$$

and such a sum is equal to 1, again due to the second point in (17.1).

Step 2: Now, if $\vec{v}$ is an eigenvector of A with eigenvalue λ, then $\vec{v}$ is an eigenvector of A^2 with eigenvalue λ^2. In the case that $\lambda < -1$, then $\lambda^2 > 1$. Since A^2 obeys (17.1) and all of its entries are positive, we know from what has been said so far that it does *not* have eigenvalues that are greater than 1. Thus, A has no eigenvalues that are less than -1.

Step 3: To see that -1 is not an eigenvalue for A, remember that if $\vec{v}$ were an eigenvector with this eigenvalue, then its entries would sum to zero. But $\vec{v}$ would also be an eigenvector of A^2 with eigenvalue 1 and we know that the entries of any such eigenvector must either be all positive or all negative. Thus, A can't have -1 as an eigenvalue either.

17.3 Exercises:

1. The purpose of this exercise is to walk you through the argument for the second point in (17.4). To start, assume that obeys $A\vec{v} = \vec{v}$ and that the first $k < n$ entries of are are negative or zero and the rest either zero or positive with at least one positive.

   (a) Add the first k entries of the vector $A\vec{v}$ and write the resulting equation asserting that the latter sum is equal to that of the first k entries of $\vec{v}$. In the case $k = 2$, this is (17.18).

   (b) Rewrite the equation that you got from (a) so that all terms that involve v_1, v_2, ..., and v_k are on the left-hand side and all terms that involve v_{k+1}, ..., v_n are on the right-hand side. In the case $k = 2$, this is (17.19).

   (c) Explain why the left-hand side of the equation that you get in (b) is negative or zero while the right-hand side is positive.

   (d) Explain why the results from (c) forces you to conclude that every eigenvector of A with eigenvalue 1 has entries that are either all positive or all negative.

2. (a) Consider the matrix $A = \begin{bmatrix} \frac{2}{3} & a & b \\ a & \frac{2}{3} & c \\ b & c & \frac{2}{3} \end{bmatrix}$. Find all possible values for a, b and c that make this a Markov matrix.

   (b) Find the eigenvector for A with eigenvalue 1 with positive entries that sum to 1.

   (c) As a check on you work in (a), prove that your values of a, b, c are such that A also has eigenvalue $\frac{1}{2}$. Find two linearly independent eigenvectors for this eigenvalue.

3. This problem plays around with some of our probability functions.

   (a) The exponential probability function is defined on the half line $[0, \infty)$. The version with mean μ is the function $x \to p(x) = e^{-x/\mu}$. The standard deviation is also μ. If $R > 0$, what is the probability that $x \geq (R+1)\mu$?

   (b) Let $Q(R)$ denote the function of R you just derived in (a). We know *a priori* that $Q(R)$ is no greater than $\frac{1}{R^2}$ and so $R^2Q(R) \leq 1$. What value of R maximizes the function $R \to R^2Q(R)$ and give the value of $R^2Q(R)$ to two decimal places at this maximum. You can use a calculator for this last part.

   (c) Let $p(x)$ denote the Gaussian function with mean zero and standard deviation σ. Thus, $p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-x^2/(2\sigma^2)}$. We saw in (13.21) that the probability, P(R), that x is greater than $R\sigma$ is less than $\frac{1}{\sqrt{2\pi}\,R} e^{-R^2/2}$. We also know from the Chebychev Theorem that know that P(R) $\leq \frac{1}{R^2}$. The ratio of these two upper bounds is R. What value of R is this ratio at its largest value? Use a calculator to write this largest value.

   (d) Let $L > 0$ and let $x \to p(x)$ denote the uniform probability function on the interval where $-L \leq x \leq L$. This probability has mean 0 and standard deviation L. Suppose that R is larger than 1 but smaller than $\sqrt{3}$. What is the probability that x has distance $\frac{RL}{2\sqrt{3}}$ or more from the origin?

(e) Let $R \to Q(R)$ denote the function of R that gives the probability from (d) that x has distance at least $\frac{RL}{2\sqrt{3}}$ from the origin. We know that $R^2 Q(R) \leq 1$. What value of R in the interval $[1, \sqrt{3}]$ maximizes this function and what is its value at its maximum?

CHAPTER

EIGHTEEN

Markov matrices and complex eigenvalues

The previous chapter analyzed Markov matrices in some detail, but left open the question as to whether such a matrix can have complex eigenvalues. My purpose here is to explain that such can be the case. I will then describe some of their properties.

18.1 Complex eigenvalues

As it turns out, there are no 2×2 Markov matrices with complex eigenvalues. You can argue using the following points:

- A matrix with real entries has an even number of distinct, complex eigenvalues since any given complex eigenvalue must be accompanied by its complex conjugate.
- There are at most 2 eigenvalues for a 2×2 matrix: Either it has two real eigenvalues or one real eigenvalue with algebraic multiplicity 2, or two complex eigenvalues, one the complex conjugate of the other.
- The number 1 is always an eigenvalue of a Markov matrix.

On the other hand, here is a 3×3 Markov matrix with complex eigenvalues:

$$A = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}. \tag{18.1}$$

If you compute the characteristic polynomial, $\mathscr{P}(\lambda) = \det(A - \lambda I)$, you will find that it is equal to

$$\mathscr{P}(\lambda) = -\left(\lambda^3 - \tfrac{3}{2}\lambda^2 + \tfrac{3}{4}\lambda + \tfrac{1}{4}\right). \tag{18.2}$$

Ordinarily, I would be at a loss to factor a generic cubic polynomial, but in this case, I know that 1 is a root, so I know that $\lambda - 1$ divides $\mathscr{P}(\lambda)$ to give a quadratic polynomial. I can do this division and I find that

$$\mathscr{P}(\lambda) = -(\lambda - 1)\left(\lambda^2 - \tfrac{1}{2}\lambda + \tfrac{1}{4}\right). \tag{18.3}$$

The roots of the quadratic polynomial $\lambda \to \lambda^2 - \frac{1}{2}\lambda + \frac{1}{4}$ are roots of $\mathscr{P}$. The roots of the quadratic polynomial can be found (using the usual formula) to be

$$\frac{\frac{1}{2} \pm \sqrt{\frac{1}{4} - 1}}{2} = \frac{1}{4} \pm \frac{\sqrt{3}}{4} i. \tag{18.4}$$

You might complain that the matrix A here has some entries equal zero, and it would be more impressive to see an example where all entries of A are positive. If this is your attitude, then consider the Markov matrix

$$A = \begin{bmatrix} \frac{1}{2} & \frac{1}{16} & \frac{7}{16} \\ \frac{7}{16} & \frac{1}{2} & \frac{1}{16} \\ \frac{1}{16} & \frac{7}{16} & \frac{1}{2} \end{bmatrix} \tag{18.5}$$

whose characteristic polynomial is $-\left(\lambda^3 - \frac{3}{2}\lambda^2 + \frac{171}{256}\lambda - \frac{43}{256}\right)$. The roots of the latter are 1 and $\frac{1}{4} \pm \frac{3\sqrt{3}}{16}i$.

18.2 The size of the complex eigenvalues

I demonstrated in the previous chapter that a Markov matrix with no zero entries has a single real eigenvalue equal to 1 and that all of its remaining real eigenvalues have absolute value less than 1. An argument very much along the same lines will demonstrate that the absolute value of any complex eigenvalue of such a matrix is less than 1 also. For example, the absolute value of the complex eigenvalues for the matrix in (18.5) is $\sqrt{\frac{31}{64}}$.

To see how this works in the general case, let's again use A to denote our Markov matrix with all $A_{jk} > 0$. If λ is a complex eigenvalue for A, then it must be a complex eigenvalue for A^T. Let $\vec{v}$ denote a corresponding complex eigenvector; thus $A^T\vec{v} = \lambda\vec{v}$. In terms of components, this says that

$$A_{1k}v_1 + A_{2k}v_2 + \cdots + A_{nk}v_n = \lambda v_k \tag{18.6}$$

for any $k \in \{1, 2, \ldots, n\}$. Taking absolute values of both sides in (18.6) finds the inequality

$$A_{1k}|v_1| + A_{2k}|v_2| + \cdots + A_{nk}|v_n| \geq |\lambda||v_k| \tag{18.7}$$

Here, I have used two facts about absolute values: First, the absolute value of λv_k is the product of $|\lambda|$ and $|v_k|$. Indeed, if a and b are any two complex numbers, then $|ab| = |a|\ |b|$ which you can see by writing both a and b in polar form. Thus, write $a = re^{i\theta}$ and $b = se^{i\varphi}$ with s and r non-negative. Then $ab = rse^{i(\theta+\varphi)}$ and so the absolute value of ab is rs which is also $|a|\ |b|$. Meanwhile, I used the fact that $|a + b| \leq |a| + |b|$ in an iterated fashion to obtain

$$\begin{aligned} |A_{1k}v_1 + A_{2k}v_2 + \cdots + A_{nk}v_n| &\leq A_{1k}|v_1| + |A_{2k}v_2 + \cdots + A_{nk}v_n| \\ &\leq A_{1k}|v_1| + A_{2k}|v_2| + |A_{3k}v_3\ + \cdots + A_{nk}v_n| \\ &\vdots \\ &\leq A_{1k}|v_1| + A_{2k}|v_2| + \cdots + A_{nk}|v_n|. \end{aligned} \tag{18.8}$$

to deduce that the expression on the left side of (18.7) is no less than that on the right side. By the way, the fact that $|a + b| \leq |a| + |b|$ holds for complex numbers is another way to say that the sum of the lengths of any two sides to a triangle is no less than the length of the third side.

Consider the inequality depicted in (18.7) in the case that k is chosen so that

$$|v_k| \geq |v_j| \qquad \text{for all } j \in \{1, \ldots, n\}. \tag{18.9}$$

Thus, v_k has the largest absolute value of any entry of $\vec{v}$. In this case, each $|v_j|$ that appears on the left side of (18.7) is no larger than $|v_k|$, so the left-hand side is even larger if each $|v_j|$ is replaced by $|v_k|$. This done, then (18.7) finds that

$$(A_{1k} + A_{2k} + \cdots + A_{nk})|v_k| \geq |\lambda||v_k|, \tag{18.10}$$

Since $A_{1k} + A_{2k} + \cdots + A_{nk} = 1$, this last expression finds that $|v_k| \geq |\lambda|\ |v_k|$ and so $1 \geq |\lambda|$.

Now, to see that $|\lambda|$ is actually less than 1, let us see what is required if every one of the inequalities that were used to go from (18.6) to (18.7) and from (18.7) to (18.10) are equalities. Indeed, if any one them is a strict inequality, then $1 > |\lambda|$ is the result. Let's work this task backwards: To go from (18.7) to (18.10) with equality requires that each

v_j have the same norm as v_k. To go from (18.6) to (18.7), we used the triangle inequality roughly n times, this the assertion that $|a+b| \le |a| + |b|$ for any two complex numbers a and b. Now this is an equality if and only if $a = rb$ with $r > 0$; thus if and only if the triangle is degenerated to one where the $a+b$ edge contains both the a and b edges as segments.

In the cases at hand, this means that $A_{jk}v_j = rA_{kk}v_k$ for each j. Thus, not only does each v_j have the same norm as v_k, each is a multiple of v_k with that multiple being a positive real number. This means that the multiple is 1. Thus, the vector is a multiple of the vector whose entries are all equal to 1. As we saw in Handout 14, this last vector is an eigenvector of A with eigenvalue 1 so if $|\lambda| = 1$, then $\lambda = 1$ and so isn't complex.

18.3 Another Markov chain example

The term *Markov chain* refers to an unending sequence, $\{\vec{p}(0), \vec{p}(1), \vec{p}(2), \dots\}$ of vectors that are obtained from $\vec{p}(0)$ by successive applications of a Markov matrix A. Thus,

$$\vec{p}(t) = A\vec{p}(t-1) \qquad \text{and so} \qquad \vec{p}(t) = A^t\vec{p}(0). \tag{18.11}$$

I gave an example from genetics of such a Markov chain in Chapter 15. What follows is a hypothetical example from biochemistry.

There is a molecule that is much like DNA that plays a fundamental role in cell biology, this denoted by RNA. Whereas DNA is composed of two strands intertwined as a double helix, a typical RNA molecule has just one long strand, usually folded in a complicated fashion, that is composed of standard segments linked end to end. As with DNA, each segment can be one of four kinds, the ones that occur in RNA are denoted as G, C, A and U. There are myriad cellular roles for RNA and the study of these is arguably one of the hottest items these days in cell biology. In any event, imagine that as you are analyzing the constituent molecules in a cell, you come across a long strand of RNA and wonder if the sequence of segments, say AGACUA$\cdots$, is "random" or not.

To study this question, you should know that a typical RNA strand is constructed by sequentially adding segments from one end. Your talented biochemist friend has done some experiments and determined that in a test tube (*in vitro*, as they say), the probability of using one of A, G, C, or U for the tth segment depends on which of A, C, G or U has been used for the $(t-1)$st segment. This is to say that if we label A as 1, G as 2, C as 3 and U as 4, then the probability, $p_j(t)$ of seeing the segment of the kind labeled $j \in \{1, 2, 3, 4\}$ in the tth segment is given by

$$p_j(t) = A_{j1}p_1(t-1) + A_{j2}p_2(t-1) + A_{j3}p_3(t-1) + A_{j4}p_4(t-1) \tag{18.12}$$

where A_{jk} denotes the conditional probability of a given segment being of the kind labeled by j if the previous segment is of the kind labeled by k. For example, if your biochemist friend finds no bias toward one or the other base, then one would expect that each A_{jk} has value $\frac{1}{4}$. In any event, A is a Markov matrix, and if we introduce $\vec{p}(t) \in \mathbf{R}^4$ to denote the vector whose kth entry is $p_k(t)$, then the equation in (18.12) has the form of (18.11).

Now, those of you with some biochemistry experience might argue that to analyze the molecules that comprise a cell, it is rather difficult to extract them without breakage. Thus, if you find a strand of RNA, you may not be seeing the whole strand from start to finish and so the segment that you are labeling as $t = 0$ may not have been the starting segment when the strand was made in the cell. Having said this, you would then question the utility of the 'solution', $\vec{p}(t) = A^t\vec{p}(0)$ since there is no way to know $\vec{p}(0)$ if the strand has been broken. Moreover, there is no way to see if the strand was broken.

As it turns out, this objection is a red herring of sorts because one of the virtues of a Markov chain is that the form of $\vec{p}(t)$ is determined solely by $\vec{p}(t-1)$. This has the following pleasant consequence: Whether our starting segment is the original $t = 0$ segment, or some $t = N > 0$ segment makes no difference if we are looking at the subsequent segments. To see why, let us suppose that the strand was broken at segment N and that what we are calling strand t was originally strand $t + N$. Not knowing the strand was broken, our equation reads $\vec{p}(t) = A^t\vec{p}(0)$. Knowing the strand was broken, we must relabel and equate our original $\vec{p}(t)$ with the vector $\vec{p}(t+N)$ that is obtained from the starting vector, $\vec{p}(0)$, of the unbroken strand by the equation $\vec{p}(t+N) = A^{t+N}\vec{p}(0)$.

Even though our equation $\vec{p}(t) = A^t\vec{p}(0)$ has the tth power of A while the equation $\vec{p}(t+N) = A^{t+N}\vec{p}(0)$ has the $(t+N)$th power, these two equations make the identical predictions. To see that such is the case, note that the equation for $\vec{p}(t+N)$ can just as well be written as $\vec{p}(t+N) = A^t\vec{p}(N)$ since $\vec{p}(N) = A^N\vec{p}(0)$.

18.4 The behavior of a Markov chain as $t \to \infty$

Suppose that we have a Markov chain whose matrix A has all entries positive and has a basis of eigenvectors. In this regard, we can allow complex eigenvectors. Let us use $\vec{e}_1$ to denote the one eigenvector whose eigenvalue is 1, and let $\{\vec{e}_2, \ldots, \vec{e}_n\}$ denote the others. We can then write our starting $\vec{p}(0)$ as

$$\vec{p}(0) = \vec{e}_1 + \sum_{k=2}^{n} c_k\vec{e}_k \tag{18.13}$$

where c_k is real if $\vec{e}_k$ has a real eigenvalue, but complex when $\vec{e}_k$ has a complex eigenvalue. With regards to the latter case, since our vector $\vec{p}(0)$ is real, the coefficients c_k and $c_{k'}$ must be complex conjugates of each other when the corresponding $\vec{e}_k$ and $\vec{e}_{k'}$ are complex conjugates also.

I need to explain why $\vec{e}_1$ has the factor 1 in front. This requires a bit of a digression: As you may recall from the previous chapter, the vector $\vec{e}_1$ can be assumed to have purely positive entries that sum to 1. I am assuming that such is the case. I also argued that the entries of any eigenvector with real eigenvalue less than 1 must sum to zero. This must also be the case for any eigenvector with complex eigenvalue. Indeed, to see why, suppose that $\vec{e}_k$ has eigenvalue $\lambda \neq 1$, either real or complex. Let $\vec{v}$ denote the vector whose entries all equal 1. Thus, $\vec{v}$ is the eigenvector of A^T with eigenvalue 1. Note that the dot product of $\vec{v}$ with any other vector is the sum of the other vector's entries. Keep this last point in mind. Now, consider that the dot product of $\vec{v}$ with $A\vec{e}_k$ is, on the one hand, $\lambda\vec{v}\cdot\vec{e}_k$, and on the other $(A^T\vec{v})\cdot\vec{e}_k$. As $A^T\vec{v} = \vec{v}$, we see that $\lambda\vec{v}\cdot\vec{e}_k = \vec{v}\cdot\vec{e}_k$ and so if $\lambda \neq 1$, then $\vec{v}\cdot\vec{e}_k = 0$ and so the sum of the entries of is zero.

Now, to return to the factor of 1 in front of $\vec{e}_1$, remember that $\vec{p}(0)$ is a vector whose components are probabilities, and so they must sum to 1. Since the components of the vectors $\vec{e}_2, \ldots, \vec{e}_n$ sum to zero, this constraint on the sum requires the factor 1 in front of $\vec{e}_1$ in (18.13).

With (18.13) in hand, it then follows that any given $t > 0$ version of $\vec{p}(t)$ is given by

$$\vec{p}(t) = \vec{e}_1 + \sum_{k=2}^{n} c_k\lambda_k^t\vec{e}_k. \tag{18.14}$$

Since $|\lambda^t| = |\lambda|^t$ and each λ that appears in (18.14) has absolute value less than 1, we see that the large t versions of $\vec{p}(t)$ are very close to $\vec{e}_1$. This is to say that

$$\lim_{t\to\infty} \vec{p}(t) = \vec{e}_1. \tag{18.15}$$

This last fact demonstrates that as t increases along a Markov chain, there is less and less memory of the starting vector $\vec{p}(0)$. It is sort of like the aging process in humans: As $t \to \infty$, a Markov chain approaches a state of complete senility, a state with no memory of the past.

18.5 Exercises:

1. Any 2×2 Markov matrix has the generic form $\begin{bmatrix} a & 1-b \\ 1-a & b \end{bmatrix}$, where $a, b \in [0, 1]$. Compute the characteristic polynomial of such a matrix and find expressions for its roots in terms of a and b. In doing so, you will verify that it has only real roots.

2. Let A denote the matrix in (18.1). Find the eigenvectors for A and compute $\vec{p}(1)$, $\vec{p}(2)$ and $\lim_{t\to\infty} \vec{p}(t)$ in the case that $\vec{p}(t) = A\vec{p}(t-1)$ and $\vec{p}(0) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$. Finally, write $\vec{p}(0)$ as a linear combination of the eigenvectors for A.

3. Repeat Problem 2 using for A the matrix in (18.5).

CHAPTER

NINETEEN

Symmetric matrices and data sets

Suppose that you take some large number of measurements of various facets of a system that you are studying. Lets say that there are n facets and you take $N \gg n$ measurements under different conditions; thus, you generate a collection of N vectors in $\mathbf{R}^n$ which you can label as $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$.

An issue now is whether some of the n facets that you measure are dependent on the rest. For example, suppose that one facet is the temperature of the sample, and if all of the other facets of the system are completely determined by the temperature, then all of these N vectors will lie on very near some curve in the n-dimensional space, a curve parameterized as $T \to \vec{x}(T)$ by the temperature. On the other hand, if no facets are determined by any collection of the others, then the N vectors could be spread in a more or less random fashion through a region of $\mathbf{R}^n$. The point here is that if you are interested in discovering relations between the various facets, then you would like to know if the distribution of the N vectors is spread out or concentrated near some lower dimensional object – a curve or a surface or some such. Any such concentration towards something less spread out indicates relations between the various facets.

For example, suppose n = 2. If you plot the endpoints of the vectors $\{\vec{x}_k\}$ in $\mathbf{R}^2$ and see the result as very much lying near a particular curve (not necessarily a line), this says that the two facets are not able to vary in an independent fashion. Indeed, if $y = f(x)$ is the equation for the curve, it says that the variation in x determines the variation in y. Now, for this $n = 2$ case, you can go and plot your N vectors and just look at the picture. However, for $n > 2$, this is going to be hard to do. How then can you discern relationships when $n > 2$?

19.1 An example from biology

Here is a topical example: The technology is such that it is possible to monitor the levels of huge numbers of proteins in a cell as it goes about its business. These proteins are typically interacting with each other, and so it is of crucial import to determine which are influencing which. Let n denote the number of proteins involved. Discerning the levels of these n proteins at some N times during the cell cycle with various environmental factors altered at various times gives some very large data set of the sort described above, N vectors in $\mathbf{R}^n$. Of interest is how these vectors distribute themselves – is it random through some region, or are the vectors concentrated near some intrinsically much thinner set such as a curve, surface or what ever in $\mathbf{R}^n$. If the latter is the case, then the structure of this thinner set, a curve, a surface or whatever, carries information about the complicated chemical interactions in a cell.

19.2 A fundamental concern

I don't think that there is a completely foolproof way to discern relationships between a bunch of vectors in $\mathbf{R}^n$. What follows is an $n = 2$ example to keep in mind. As is typically the case, the vectors $\{\vec{x}_k\}$ can not be too big; they are constrained to lie in some region of fixed size in $\mathbf{R}^n$. Let me suppose in this example that no vector from this collection has norm greater than 1 and so all lie in the disk of radius 1 about the origin.

Now, consider the curve in this disk given in the parametric form by

$$t \to (x = ct\cos(t), y = ct\sin(t)). \tag{19.1}$$

Here $c > 0$ is a constant and $0 \leq t \leq \frac{1}{c}$ is the parameter for the curve. For any given c, this curve is a spiral. The radius, $r(t)$, is equal to ct and the angle is t as measured from the positive x-axis in the anti-clockwise direction.

As c gets smaller, the spiral gets tighter and tighter; there are more and more turns before the curve hits the boundary of the disk. Indeed, when c is very large, the spiral hardly turns and stays very close to the x-axis. When $c = \frac{1}{2\pi}$, the spiral makes one complete turn before exiting the disk. When $c = \frac{1}{2\pi m}$ and m is an integer, the spiral makes m turns before it exits.

Now, here is the problem: Suppose that our vectors are near some very small c version of (19.1). Since this spiral makes a lot of turns in the disk, all points in the disk are pretty close to the spiral and so even a random collection of points in the disk will find each point close to the spiral. In particular, our collection of vectors $\{\vec{x}_k\}$ will be close to all small c versions of the spiral no matter what!!

Here is another example: Suppose that the points $\{\vec{x}_k\}$ are distributed randomly in a thin strip of width $r \ll 1$ along the diagonal $y = x$. Thus, each $\vec{x}_k$ has coordinates x and y that obey $|x - y| \leq r$. So, inside this strip, the points are spread at random. Outside the strip, there are no points at all. If your experimental error is on the order of r, then I would be happy to conclude that the points lie on the line $y = x$ and thus the experiment indicates that the y-measurement is completely determined by the x-measurement. On the other hand, if the experimental error is much less than r, then I would not agree that the concentration near the $x - y$ line signifies that y is determined by x. Maybe some part of y, but there is some part left over that is independent.

19.3 A method

Suppose we have data $\{\vec{x}_k\}_{1\leq k\leq N}$, each a vector in $\mathbf{R}^n$. Here is a method to analyze whether the data near any given $\vec{x}_j$ is concentrated near some lower dimensional subspace of $\mathbf{R}^n$.

Step 1: You must choose a number, r, that is a reasonable amount greater than your experimental error. In this regard, r should also be significantly less than the maximum distance between any two points in $\{\vec{x}_k\}$. Thus, $r \ll \max_{j,k} |\vec{x}_j - \vec{x}_k|$. This number r determines the scale on which you will be looking for the clustering of the vectors.

Step 2: Let $\vec{x}_j \in \{\vec{x}_k\}$. To see if the data is clustering around a lower dimensional subspace near $\vec{x}_j$, take all points in $\{\vec{x}_k\}$ that have distance r or less from $\vec{x}_j$. Let m denote the number of such points. Allow me to relabel these points as $\{\vec{y}_1, \ldots, \vec{y}_m\}$. These are the only points from the collection $\{\vec{x}_k\}$ that will concern us while we look for clustering around the given point $\vec{x}_j$ at the scale determined by r.

Step 3: Let $\vec{a}$ denote the vector $\frac{1}{m}(\vec{y}_1 + \cdots + \vec{y}_m)$. This vector should be viewed as the center of the collection $\{\vec{y}_1, \ldots, \vec{y}_m\}$. For each index $i = 1, \ldots, m$, set $\vec{z}_i = \vec{y}_i - \vec{a}$. This just shifts the origin in $\mathbf{R}^n$.

Step 4: View each $i \in \{1, \ldots, m\}$ version of $\vec{z}_i$ as a matrix with 1 column and then introduce the transpose, $\vec{z}_i^T$, which is a matrix with 1 row. Note that if ever is a vector in $\mathbf{R}^n$ viewed as a 1-column matrix, then $\vec{z}$ can multiply $\vec{z}^T$ to give the square matrix $\vec{z}\vec{z}^T$. For example, if $\vec{z}$ has top entry 1 and all others 0, then $\vec{z}\vec{z}^T$ has top left entry 1 and all others zero. In general, the entry $(\vec{z}\vec{z}^T)_{ik}$ is the product of the ith and kth entries of $\vec{z}$.

Granted the preceding, introduce the matrix

$$A = \frac{1}{m}\left(\vec{z}_1\vec{z}_1^T + \cdots + \vec{z}_m\vec{z}_m^T\right). \tag{19.2}$$

This is a symmetric, $n \times n$ matrix, so it has n real eigenvalues which I will henceforth denote by $\{\lambda_1, \ldots, \lambda_n\}$.

Step 5: As I explain below, none of the eigenvalues has absolute value greater than r^2. I also explain below why A has no negative eigenvalues. Granted this, then A's eigenvalues can be anywhere from 0 to r^2. Those eigenvalues that are much smaller than r^2 correspond to 'pinched' directions. If $n - d \leq n$ of the eigenvalues are much smaller than r^2, this indicates that the distribution of the vectors from $\{\vec{x}_k\}$ with distance r or less from $\vec{x}_j$ is concentrated near a subspace of $\mathbf{R}^n$ whose dimension is d. To quantify this, I am going to say that the vectors near $\vec{x}_j$ cluster around a subspace whose dimension is no greater than d when A has $n - d$ or more eigenvalues that are smaller than $\frac{1}{2(n+2)}r^2$. In this regard, note that I am not going to expect much accuracy with this prediction unless m is large.

I give some examples below to illustrate what is going on. At the end, I outline my reasons for choosing the factor $\frac{1}{2(n+2)}$ to distinguish between small and reasonably sized eigenvalues.

19.4 Some loose ends

To tie up some loose ends, let me show you why all of A's eigenvalues are in the interval between 0 and r^2. To this end, suppose that $\vec{v}$ is some vector. To see what $A\vec{v}$ looks like, the first thing to realize is that if $\vec{z}$ is any given vector, then $\vec{z}^T\vec{v}$ is a 1×1 matrix, thus a number and that this number is the dot product between $\vec{z}$ and $\vec{v}$. This implies that

$$A\vec{v} = \frac{1}{m}(\vec{z}_1 \cdot \vec{v})\vec{z}_1 + \cdots + \frac{1}{m}(\vec{z}_m \cdot \vec{v})\vec{z}_m. \tag{19.3}$$

Therefore, if I take the dot product of $A\vec{v}$ with $\vec{v}$, I find that

$$\vec{v} \cdot (A\vec{v}) = \frac{1}{m}(\vec{z}_1 \cdot \vec{v})^2 + \cdots + \frac{1}{m}(\vec{z}_m \cdot \vec{v})^2. \tag{19.4}$$

Note that this is a sum of non-negative terms, so $\vec{v} \cdot (A\vec{v}) \geq 0$ for any vector $\vec{v}$.

Now suppose that $\vec{v}$ is an eigenvector with eigenvalue λ. Then $A\vec{v} = \lambda\vec{v}$ and so $\vec{v} \cdot (A\vec{v}) = \lambda|\vec{v}|^2$. As this is non-negative, we see that $\lambda \geq 0$.

To see that $\lambda \leq r^2$, use the fact that

$$|\vec{z} \cdot \vec{v}| \leq |\vec{z}|\ |\vec{v}| \tag{19.5}$$

for any given vector on each term on the right hand side of (19.5) to conclude that

$$\vec{v} \cdot (A\vec{v}) \leq \frac{1}{m}|\vec{z}_1|^2\ |\vec{v}|^2 + \cdots + \frac{1}{m}|\vec{z}_m|^2\ |\vec{v}|^2. \tag{19.6}$$

When I pull out the common factor $\frac{1}{m}|\vec{v}|^2$ on the right of this last inequality, I find that

$$\vec{v} \cdot (A\vec{v}) \leq \frac{1}{m}\left(|\vec{z}_1|^2 + \cdots + |\vec{z}_m|^2\right)|\vec{v}|^2 \tag{19.7}$$

since each $\vec{z}_k$ has norm less than r, the right-hand side of (19.7) is no larger than $r^2|\vec{v}|^2$. Thus,

$$\vec{v} \cdot (A\vec{v}) \leq r^2|\vec{v}|^2 \tag{19.8}$$

for any vector $\vec{v}$.

Now, if $\vec{v}$ is an eigenvector with eigenvalue λ, then the left-hand side of this last equation is $\lambda|\vec{v}|^2$ and so $\lambda \leq r^2$ as claimed.

19.5 Some examples

What follows are a few examples to try to convince you that my attempt at estimating a dimension is not unreasonable..

Example 1: Suppose that all $\vec{z}_k$ sit very close to the origin, for example, suppose that each $\vec{z}_k$ has $|\vec{z}_k| < \varepsilon r$ with $\varepsilon^2 \leq \frac{1}{2(n+2)}$. To see what this implies, return now to (19.7) to conclude that

$$\vec{v} \cdot (A\vec{v}) \leq \varepsilon^2 r^2 |\vec{v}|^2. \tag{19.9}$$

This then means A's eigenvalues are no larger than $\varepsilon^2 r^2$ which is smaller than $\frac{1}{2(n+2)} r^2$. Thus, we would predict the vectors from $\{\vec{x}_k\}$ are clustering around a point near $\vec{x}_j$.

Example 2: Suppose that m is large and that the vectors $\{\vec{z}_k\}$ all sit on a single line, and that they are evenly distributed along the line. In particular, let $\vec{z}$ denote the unit tangent vector to the line, and suppose that $\vec{z}_k = (-1 + \frac{2k}{m+1}) r\vec{z}$. Thus, $\vec{z}_1 = -\frac{m-1}{m+1} r\vec{z}$ and $\vec{z}_m = \frac{m-1}{m+1} r\vec{z}$. This being the case

$$A = \frac{r^2}{m} \sum_{k=1}^{m} \left(-1 + \tfrac{2k}{m+1}\right)^2 \vec{z}\vec{z}^T. \tag{19.10}$$

As it turns out, this sum can be computed in closed form and the result is

$$A = \frac{r^2}{3} \, \frac{m-1}{m+1} \vec{z}\vec{z}^T. \tag{19.11}$$

The matrix A has the form $c\vec{z}\vec{z}^T$ where c is a constant. Now any matrix $\hat{A} = c\vec{z}\vec{z}^T$ has the following property: If $\vec{v}$ is any vector in $\mathbf{R}^n$, then

$$\hat{A}\vec{v} = c\vec{z}\,(\vec{z} \cdot \vec{v})\,. \tag{19.12}$$

Thus, $\hat{A}\vec{v} = 0$ if $\vec{v}$ is orthogonal to $\vec{z}$, and $\hat{A}\vec{z} = c\vec{z}$. Hence $\hat{A}$ has 0 and c as its eigenvalues, where 0 has multiplicity $n-1$ and c has multiplicity 1.

In our case, this means that there is one eigenvalue that is on the order of $\frac{r^2}{3}$ and the others are all zero. Thus, we would say that the clustering here is towards a subspace of dimension 1, and this is precisely the case.

Example 3: To generalize the preceding example, suppose that $d \geq 1$, that $V \subset \mathbf{R}^n$ is a d-dimensional subspace, and that the vectors $\{\vec{z}_k\}_{1 \leq k \leq m}$ all lie in V.

In this case, one can immediately deduce that A will have $n-d$ orthonormal eigenvectors that have zero as eigenvalue. Indeed, any vector in the orthogonal complement to V is in the kernel of A and so has zero eigenvalue. As this orthogonal subspace has dimension $n-d$, so A has $n-d$ linearly independent eigenvalues with eigenvalue 0.

Thus, we see predict here that the clustering is towards a subspace whose dimension is no greater than d.

19.6 Small versus reasonably sized eigenvalues

I am going to give some indication here for my choice of the factor $\frac{1}{2(n-d+2)}$ to distinguish the small eigenvalue of A. Note here that you or others might want some other, smaller factor. For example, I am probably being conservative with this factor.

To explain where this factor $\frac{1}{2(n+2)}$ is coming from, I have to take you on a digression that starts here with the introduction of the uniform probability function on the ball of radius r in $\mathbf{R}^d$. This is to say that the probability of choosing a point in any given subset of the ball is proportional to the d-dimensional volume of the subset. To each vector $\vec{u}$, in this ball (thus, to each $\vec{u}$ with $|\vec{u}| \leq r$), we can assign the square matrix $\vec{u}\vec{u}^T$. This can be viewed as a random variable that maps vectors in the ball to $d \times d$ matrices. For example, in the case $d = 2$, one has

$$\vec{u}\vec{u}^T = \begin{bmatrix} x^2 & xy \\ xy & y^2 \end{bmatrix}. \tag{19.13}$$

For any d, the mean of the random variable $\vec{u} \to \vec{u}\vec{u}^T$ is the matrix whose entry in the jth row and kth column is the average of $u_j u_k$ over the ball. These averages can be computed and one finds that the mean of $\vec{u}\vec{u}^T$ is $\frac{r^2}{d+2} I_d$, where I_d is the $d \times d$ identity matrix.

Keeping all of this in mind, go back to the formula for A in (19.2). Suppose that I consider a set of m vectors, $\{\vec{u}_k\}_{1 \le k \le m}$ that all lie in a d-dimensional subspace. Suppose further that I sprinkle these vectors at random in the radius r ball inside this subspace. I can then view

$$\hat{A} = \frac{1}{m}\left(\vec{u}_1\vec{u}_1^T + \cdots + \vec{u}_m\vec{u}_m^T\right) \tag{19.14}$$

as the average of m identical, but unrelated random variables, this being m maps of the form $\vec{u} \to \vec{u}^T$. According to the Central Limit Theorem, when m is large, the probability that $\hat{A}$ differs from $\frac{r^2}{d+2} I_d$ is very small.

If in my actual data, the vectors $\{\vec{z}_k\}$, are truly clustering around a d-dimensional subspace of $\mathbf{R}^n$ and are more or less randomly sprinkled in the radius r ball of this set, then I should expect the following:

> When m is very large, the Central Limit Theorem tells me that the matrix A in (19.2) is very close to the matrix $\frac{r^2}{d+2} P$, where P here denotes the orthogonal projection on to the subspace in question.

Thus, it should have d eigenvalues that are very close to $\frac{r^2}{d+2}$ and $n - d$ eigenvalues that are much smaller than this number. In particular, when m is large, then under the hypothesis that the vectors $\{\vec{z}_k\}_{1 \le k \le m}$ are sprinkled at random in a d-dimensional subspace, the matrix A in (19.2) will be very likely to have d eigenvalues that are on the order of $\frac{r^2}{d+2}$ and $n - d$ eigenvalues that are zero.

This application of the Central Limit Theorem explains my preference for the size distinction $\frac{1}{2(n+2)} r^2$ between small eigenvalues of A and eigenvalues that are of reasonable size. I think that this cut-off is rather conservative and one can take a somewhat larger one.

19.7 Exercises:

1. The purpose of this exercise is to compute the average of the matrix in (19.13) over the disk of radius r in the xy-plane. This average is the matrix, U, whose entries are

$$U_{11} = \frac{1}{\pi r^2}\iint x^2\, dx\, dy, \qquad U_{22} = \frac{1}{\pi r^2}\iint y^2\, dx\, dy, \qquad \text{and} \qquad U_{12} = U_{21} = \frac{1}{\pi r^2}\iint xy\, dx\, dy.$$

   To compute U, change to polar coordinates (ρ, θ) where $\rho \ge 0$ and $\theta \in [0, 2\pi]$ using the formula $x = \rho\cos(\theta)$ and $y = \rho\sin(\theta)$. Show that

   (a) $U_{11} = \frac{1}{\pi r^2}\int_0^{2\pi}\int_0^r \rho^2\cos^2(\theta)\rho\, d\rho\, d\theta$,

   (b) $U_{22} = \frac{1}{\pi r^2}\int_0^{2\pi}\int_0^r \rho^2\sin^2(\theta)\rho\, d\rho\, d\theta$, and

   (c) $U_{12} = U_{21} = \frac{1}{\pi r^2}\int_0^{2\pi}\int_0^r \rho^2\sin(\theta)\cos(\theta)\rho\, d\rho\, d\theta$.

   Next, use the formula $\cos(2\theta) = 2\cos^2(\theta) - 1 = 1 - 2\sin^2(\theta)$ and $\sin(2\theta) = 2\sin(\theta)\cos(\theta)$ to do the angle integrals first and so find that

   (d) $U_{11} = U_{22} = \frac{1}{r^2}\int_0^r \rho^3\, d\rho$, and

   (e) $U_{12} = U_{21} = 0$.

   Finally, do the ρ-integral to find that $U_{11} = U_{22} = \frac{1}{4} r^2$ and $U_{12} = U_{21} = 0$. Note that this means that U is the $d = 2$ version of $\frac{r^2}{d+2} I_d$.