Statistics and Probability Primer for Computational Biologists

by

Peter Woolf

Christopher Burge

Amy Keating

Michael Yaffe

Massachusetts Institute of Technology

BE 490/ Bio7.91

Spring 2004

Index

Index

Introduction

Chapter 1:Common Statistical Terms

1.1 Mean, Median, and Mode

1.2 Expectation Values

1.3 Standard Deviation and Standard Error

Chapter 2: Probability

2.1 Definition of Probability

2.2 Joint Probabilities and Bayes' Theorem

Chapter 3: Discrete Distributions

3.0 Geometric Distribution

3.1 Binomial Distribution & Bernoulli Trials

3.3 Poisson Distribution

3.4 Multinomial Distribution

3.5 Hypergeometric Distribution

Chapter 4: Continuous Distributions

4.0 Continuous Distributions

4.1 Gaussian Distribution

4.2 Standard normal distribution and z-scores

4.3 Exponential distribution

Chapter 5: Statistical Inference

5.0 Statistical Inference

5.1 Confidence Intervals

5.2 P-values

5.3 Comparing two distributions

5.4 Correlation coefficients

Appendix A: Dirichlet Distribution

A.1 Dirichlet Distribution

Appendix B: Bayesian Networks

B.1 Bayesian Networks

Appendix C: Solutions

Introduction

Why do you need to know the theory behind statistics and probability to be proficient in computational biology? While it is true that software to analyze biological data often comes prepackaged and in a working form (sometimes), it is important to know how these tools work if we are able to interpret the meaning of their results properly. Even more important, often the software does not currently exist to do what we want, so we are forced to write algorithms ourselves. For this latter case especially, we need to be intimately familiar not only with the language of statistics and probability, but also with examples of how these approaches can be used.

In this primer, our goal is to convey the essential information about statistics and probability that is required to understand algorithms in modern computational biology. We start with an introduction to basic statistical terms such as mean and standard deviation. Next we turn our attention to probability and describe ways that probabilities can be manipulated and interconverted. Following this, we introduce the concept of a discrete distribution, and show how to derive useful information from this type of model. As an extension of discrete distributions we then introduce continuous distributions and show how they can be used. Finally, we then discuss a few of the more advanced tools of statistics, such as p-values and confidence intervals. Appendix A and B cover optional material that related to more complex probability distributions and a discussion of Bayesian networks. If you are already comfortable with basic statistical terms, such as mean and variance, then feel free to skip the first chapter.

In each section, our goal is to provide both intuitive and formal descriptions of each approach. To aid in providing a more intuitive understanding of the material, we have included a number of worked out examples that relate to biologically relevant questions. In addition, we also include example problems at the end of the chapter as an exercise for the reader. Answers to these problems are listed in Appendix C. Working the problems at the end of each chapter is essential to ensure comprehension of the material covered. More difficult optional problems that may require significant mathematical expertise are marked with a star next to their number.

Introduction

Although this primer is designed to stand alone, we recommend that students also use other resources to find additional information. To aid in this goal, each chapter ends with a reference to suggested further reading. One book that we will reference often is “The Cartoon Guide to Statistics” by Larry Gonick and Woollcott Smith. This cartoon guide provides a lighthearted but thorough grounding in many of the tools discussed in this work.

Finally, this primer is a work in progress, so your feedback would be appreciated. If you find any errors or have suggestions, please email these directly to the authors so that we can incorporate your changes into future editions.

Chapter 1: Common Statistical Terms

1.1 *Mean, Median, and Mode*

Communicating statistical concepts requires a solid understanding of the language of basic statistics. In this section, we will provide a basic overview of some of the statistical terms that will be used throughout the primer and the class.

How do we describe a dataset? For example, imagine that we are measuring the transfection efficiency of a plasmid into a cell line. If we transfect 12 identical plates of cells and measure the fraction successfully transfected, we might get values like those shown in Table 1.1.1.

0.39	0.12	0.29	0.41
0.62	0.33	0.39	0.37
0.51	0.12	0.12	0.28

Table 1.1.1: Transfection efficiency expressed as the fraction of cells transfected for 12 independent measurements.

One way to describe this data would be to take the average, or **mean**, of the data. The mean can be thought of as the most likely value for the true transfection efficiency given the available data. Formally the mean is defined as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1.1.1)$$

For the transfection example in Table 1.1.1, the mean would be calculated as:

$$\bar{x} = \frac{1}{12}(0.39 + 0.12 + 0.29 + ... + 0.12 + 0.28) = 0.33 \qquad (1.1.2)$$

Another measure of the middle value is the **median**. The median value of a distribution is the data point that is separated from the minimum and maximum data points by the same number of entries. To find the median value, we sort our data from smallest to largest. After successive eliminations of the maximum and minimum values (or just taking the middle value), we end up with either one value if there are an odd number of points, or two values if there are an even number of points. This resulting value, or the average of the two resulting values, is the median. For the dataset in Table 1.1.1, we find the median first by sorting, then eliminating as is shown below:

0.62	0.51	0.41	0.39	0.39	0.37	0.33	0.29	0.28	0.12	0.12	0.12
	0.51	0.41	0.39	0.39	0.37	0.33	0.29	0.28	0.12	0.12	
		0.41	0.39	0.39	0.37	0.33	0.29	0.28	0.12		
			0.39	0.39	0.37	0.33	0.29	0.28			
				0.39	0.37	0.33	0.29				
					0.37	0.33					
			median		0.35						

The primary use of the median is to find the middle of a distribution even when there are outliers, or data points that are very much larger or smaller than the rest of the values and would tend to strongly skew the mean. For example, if the data in Table 1.1.1 represented the signal deviation from a control experiment, and we took one more measurement and observed a value of –1.23, then this likely erroneous entry would bias the mean from 0.33 to 0.21 , while the median would only change from 0.35 to 0.33.

A final measure of the middle of the data is the **mode**, which represents the number that we find most often, or most frequently. From our dataset in Table 1.1.1, we find that the mode is 0.12, because it is present three times in the dataset.

Based on these descriptions of mean, median, and mode, we find that they all describe different properties of a probability density. Therefore in most cases mean, median, and mode will not be equal. Only for rare cases such as a symmetric distribution (e.g. a bell shaped Gaussian distribution) do all of these measurements align.

1.2 Expectation values

A more general form of the mean that will be used throughout this primer is the **expectation value**. The expectation value of a distribution is the sum (or integral for continuous data) of the outcome times the probability of that outcome. The expectation value of a variable or function is written by enclosing the variable or function in angle brackets, < >, or using the notation E[x]. For discrete cases, the expectation is defined as

$$E[x] = \langle x \rangle = \sum xp(x) \tag{1.2.1}$$

where the sum is taken over all possible values of the variable x. For a continuous distribution, the expectation value is

$$E[x] = \langle x \rangle = \int xp(x)dx \tag{1.2.2}$$

over the possible range of x. A specific example of the expectation value is the mean, where the probability of each value is assumed to be equal. In this case p(x) is just 1/N. For this case:.

$$E[x] = \langle x \rangle = \sum xp(x) = \sum x\frac{1}{N} = \frac{1}{N}\sum x \tag{1.2.3}$$

However, in general all outcomes will not have the same probability, so this general definition of expectation should be used.

Example 1.2.1: Expected payout

A casino is considering offering a new, somewhat boring, card game and wants to know if it will make money. Under the current rules, each player pays $3 to play one game. In this game the player draws a single card from a full 52 card deck. If the card is a king, then the player is paid $20, while if the card is a spade, the player is paid $10. If the player chooses the king of spades, he is paid $50. From these rules, what is the expected profit per game for the casino?

Solution

Our goal is to calculate the total profit to the casino based on the probability of each event. We calculate the probability by finding the expectation value using Equation 1.2.1. To solve this expression we need to find the probabilities for each outcome.

- There are 3 non-spade kings in the pack, so the probability of drawing a non-spade king is 3/52.
- There are 12 non-king spades in the pack, so the probability of drawing a non-king spade is 12/52.
- There is only one king of spades, to the probability of drawing a king of spades is 1/52.

Multiplying these probabilities by their associated payouts, we generate the following expression

$$\langle \text{payout} \rangle = \sum (\text{event cost})\text{p}(\text{event}) = (\$20)\frac{3}{52} + (\$10)\frac{3}{52} + (\$50)\frac{1}{52} = \$2.69$$

Thus each time a player plays the game, the casino will pay out an average of $2.69, while receiving a payment from the player of $3.00, making a net profit for the casino of $0.31 per play.

1.3 *Standard Deviation and Standard Error*

We also might want to know about the spread in a dataset. For example, the distribution in the dataset in Table1.1.1 could be due to an experimental error. If we modify the experimental protocol, we would like to know whether this modification increases or decreases errors, so it would be helpful to have a quantitative measure. A common method to quantify the spread of a dataset is to use a **standard deviation**. Intuitively, the standard deviation describes the average separation of the data from the mean value. Thus, a large standard deviation means that the data is spread out, while a small standard deviation means that the data is tightly clustered. Formally, the sample standard deviation is defined as:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (1.3.1)$$

For our example in Table 1.1, the standard deviation can be calculated in the following way:

$$s = \sqrt{\frac{1}{12-1}\left[(0.39-0.33)^2 + (0.11-0.33)^2 + ... + (0.28-0.33)^2\right]} = 0.16 \qquad (1.3.2)$$

Note that the standard deviation of a sample is defined with a denominator of n-1. If values for the entire population have been measured, then the expression changes to

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \qquad (1.3.3)$$

where σ is the population standard deviation, and μ is the population mean. Also in this population description the denominator changes from n-1 to n.

A more general form of the standard deviation is written using the expectation value introduced in section 1.2. In general, the standard deviation is written as

$$\sigma = \sqrt{\left\langle (x - \langle x \rangle)^2 \right\rangle} = \sqrt{\left\langle (x - \mu)^2 \right\rangle} \qquad (1.3.4)$$

For the discrete case where all values are weighted equally we can recover our previous expression of the standard deviation using this definition.

$$\sigma = \sqrt{\left\langle (x - \mu)^2 \right\rangle} = \sqrt{\sum (x - \mu)^2 p((x - \mu)^2)} = \sqrt{\sum (x - \mu)^2 \frac{1}{N}} = \sqrt{\frac{1}{N} \sum (x - \mu)^2} \qquad (1.3.5)$$

Another metric that we are often interested in is the standard deviation of the mean, or **standard error**. For any stochastic (i.e. random) system, we will always measure a value with errors, no matter how many data points we take. However, as we take more and more measurements, our certainty of the mean value increases. It can be shown that the error around the mean (e.g. difference between the measured and true values) decreases as one over the square root of the total number of measurements according to the following relationship

$$\sigma_\mu = \frac{\sigma}{\sqrt{n}} \qquad (1.3.6)$$

Where σ_μ is the standard error of the mean. The above relationship has profound implications for computational and systems biology, for it states that if we want to reduce the error of our estimate of a mean value by a factor of ten, we have to gather one hundred times more data.

For more information, see

- Chapters 1 and 2 of “The Cartoon Guide to Statistics” by L. Gonick and W. Smith
- Chapters 1 and 2 in “Elementary Statistics” by M. F. Triola

Problems

1) Three measurements of gene expression yield the values 1.34, 3.23, and 2.11. Find the mean and standard deviation of these values. Next, find the standard deviation of the mean. Show all of your work.
2) Draw a histogram where the mean, median, and mode are all different. Mark the approximate location of these three quantities and explain why you put them there.
3) Imagine that we are trying to find the "average" value rolled on a fair 6-sided die. (a) We start by rolling a die 10 times to get the following data: 1,3,4,6,2,2,1,5,4,1. What are the mean, standard deviation, and standard error for this dataset? (b) Next we run 10 more trails to get the following data: 3,2,2,4,6,5,1,1,5,3. What are the mean, standard deviation, and standard error for the two datasets together? (c) How do you expect that the standard error would change if we gathered 2000 more data points? Why?

Chapter 2: Probability

2.1 *Definition of Probability*

Probability theory describes the likelihood of a particular outcome for an experiment. Common examples of probability theory include finding the probability that a coin toss will come up heads, or the probability of rain next week. There are a number of important applications of probability theory to biology. For example, what is the probability that a gene is up-regulated given a set of microarray data? What is the probability of carrying a gene mutation that causes albinism? What is the probability that a given genomic region binds to a particular transcription factor?

An example of probability theory applicable to molecular biology is a very simple model of a DNA or protein sequence. In this case, the outcome is one of four bases of DNA or one of 20 amino acids of a protein at a particular position. For each of these possible outcomes, we can assign a probability p_i. Thus, for example, the probability of finding a T at a particular position in a DNA sequence might be p_T=0.23, or 23%. In general, probabilities have the following two properties:

$$p_i \geq 0 \qquad (2.1.1)$$

and

$$\sum_{i=1}^{N} p_i = 1 \qquad (2.1.2)$$

where there are N possible outcomes. The first property says that probabilities cannot be negative, as that would make no sense. The second property says that if we consider the probabilities of all possible outcomes, then the sum of their probabilities must equal 1.0.

<ins>*Example 2.1.1*</ins>*: Motif alignment*

A common process in bioinformatics is searching for motifs or patterns in DNA sequences. These motifs may indicate the presence of regulatory sites or coding regions, for example. A common method for characterizing a motif is by creating a weight matrix based on a set of aligned occurrences of the motif. The weight matrix assigns a probability of finding each base pair at a particular position in a sequence. Imagine that we had aligned the following 10 eight base pair sequences:

TATGCACT

AATGCACT

TTTGCACT

TATGGACT

TATGCACT

CATGCACT

TATGCACT

TATGTACT

CATGCACT

TCTGCACT

Overall these sequences are fairly similar, but there are a few positions where they vary. If we focus only on the first base, we can count up the number of occurrences of each base in this site and approximate its frequency or probability:

$$p(S_1 = T) = \frac{7}{10} = 0.7 \qquad p(S_1 = A) = \frac{1}{10} = 0.1$$

$$p(S_1 = C) = \frac{2}{10} = 0.2 \qquad p(S_1 = G) = \frac{0}{10} = 0.0$$

Note that this probability estimate assigns a probability of zero to finding G in the first site. Although this estimate reflects the given data, we probably do not have enough information in ten sequences to make such a strong statement. One way to handle small datasets like this one is to use *pseudocounts* which derive from a Bayesian prior as is addressed in Appendix A.

2.2 *Joint Probabilities and Bayes' Theorem*

How do we calculate the probability that multiple events take place? For example, what is the probability that two coin tosses will yield two heads? Similarly, what is the probability that two polymorphic positions in a chromosome will both contain the most common allele? To answer these questions, we have to introduce some new terms and notation.

The probability of two or more events taking place is called the **joint probability.** Mathematically, the joint probability of two events, A and B, is written as:

$$P(A \ \& \ \ B) \tag{2.2.1}$$

or

$$P(A,B) \tag{2.2.2}$$

Both ways of writing this probability are equivalent, although the second is generally preferred as it is shorter.

If the two events are **independent**, then the outcome of the one event does not influence the outcome of the other event. To calculate the joint probability of N independent events, we take the product of their individual probabilities as is shown below:

$$P(e_1,e_2,...e_N) = \prod_{i=1}^{N} p(e_i) \tag{2.2.3}$$

Assumptions of independence are common in probability theory, and accurately describe events such as tossing coins and rolling dice. In biological analysis, assumptions of independence are often used as a first pass model or as a null hypothesis. For example, we can model DNA as a random sequence of A, T, G, and C, assuming that the bases at different positions are independent of each other. Although such a model would not be appropriate for coding regions of DNA (as these contain codon triplets which are not independent), it is appropriate for some intergenic regions.

Two events are **dependent** if the outcome of one event gives information about the outcome of the other event. In this way, there is no general way to decouple the probabilities of the individual events, and they must be analyzed jointly or conditionally (discussed below). For example, the expression of two genes might be tightly coupled because both are governed by the same transcription factor. In this case, the expression level of one gene provides information about the expression level of the other gene, making these two measurements dependent. An extreme biological example would be the likelihood of finding an A paired to a T in a DNA sequence. In this case, knowing the sequence identity of the base on one strand provides essentially complete information about the identity of the base in the same position on the complementary strand.

An important concept often used in probability theory is **conditional probability**. Conditional probability describes the likelihood of particular event given that we know the outcome of another event. The conditional probability of event A given that the outcome of event B is known is written as

$$P(A \mid B) \tag{2.2.4}$$

This expression is read "the probability of A given B." If A and B are independent, then this expression simplifies to

$$P(A \mid B) = P(A) \tag{2.2.5}$$

because by definition the outcome of event A does not depend on the outcome of B if A and B are independent.

To calculate the joint probability that two events take place, A and B, we can use the definition of conditional probability to expand this definition to

$$P(A,B) = P(B \mid A)P(A) \tag{2.2.6}$$

Equation 2.2.6 can be interpreted as the probability that both A and B are true equals the probability that A is true times the probability that B is true conditioned on the requirement that A is true. If A and B are independent then this definition simplifies to

$$P(A,B) = P(A)P(B) \tag{2.2.7}$$

To find the probability of an event independent of other events we can **marginalize** over the other events to calculate the **marginal probability**. The term "marginal" probability refers to the way these probabilities were calculated as sums written in the margins of a probability table, as is illustrated in Example 2.2.1. Marginalization involves summing over all possible configurations or states of the other variables to obtain a weighted average probability. Formally this can be written for one variable as

$$P(A) = \sum_{B} P(A,B) = \sum_{B} P(A \mid B)P(B) \tag{2.2.8}$$

where the sums are take over all possible outcomes for the event B. To marginalize over more variables, we need only include more sums. For example, marginalizing over two variables requires two sums:

$$P(X) = \sum_{Y}\sum_{Z} P(X,Y,Z) = \sum_{Y}\sum_{Z} P(X \mid Y,Z)P(Y,Z) \tag{2.2.9}$$

A common use of marginalization is for removing *nuisance parameters* or parameters that are an intermediate for the calculation that are not particularly meaningful or useful in themselves. As an example, often we are interested in finding the probability of a model given data. This model intrinsically has parameters in it, but we may not care about the specific values of these parameters, so we marginalize them out as shown below

$$P(Model \mid Data) = \sum_{Parameters} P(Model, Parameters \mid Data) \qquad (2.2.10)$$

Example 2.2.1: Predicting membrane proteins

A researcher hypothesizes that it is possible to detect membrane proteins using the fraction of hydrophobic residues alone. To test this model, the researcher creates a library of 7500 proteins and scores each of these proteins based on their fraction of hydrophobic residues and whether they are membrane proteins. The results of this analysis are shown below

	Majority hydrophobic	*Majority hydrophilic*
Membrane Bound	2911	961
Cytosolic	713	2915

Given this information, we wish to calculate the likelihood that a novel protein that is primarily hydrophobic is also a membrane protein.

Solution

To solve this problem, we will first summarize our data as a table of all of the possible combinations of possibilities.

	H	*NOT H*
M	*P(H, M)*	*P(NOT H,M)*
NOT M	*P(H,NOT M)*	*P(NOT H, NOT M)*

In this table, H identifies hydrophobic proteins and M membrane proteins. Next we also include the sums of the probabilities in the margins to calculate the marginal probabilities

	H	*NOT H*	*Sum*
M	*P(H, M)*	*P(NOT H,M)*	*P(M)*
NOT M	*P(H,NOT M)*	*P(NOT H, NOT M)*	*P(NOT M)*
Sum	*P(H)*	*P(NOT H)*	1

The values in this table can be filled directly from the given data. For example,

$$P(H,M) = \frac{2911}{7500} = 0.388$$

Note that the sum in the lower right corner must equal one, for it is the sum of all possible outcomes of each variable. Filling in the remaining values, we calculate the following probabilities

	H	*NOT H*	*Sum*
M	*0.388*	*0.128*	*0.516*
NOT M	*0.095*	*0.389*	*0.484*
Sum	*0.483*	*0.517*	1

By rearranging our definition of conditional probability, we can now answer the question of the likelihood of a novel protein being membrane bound given that it is hydrophobic:

$$P(M \mid H) = \frac{P(H,M)}{P(H)} = \frac{0.388}{0.483} = 0.803$$

Finally, if we want to calculate the joint probability of two dependent events, then we would expect that the joint probability would remain the same independent of the order of the events. Stated formally this says that

$$P(A,B) = P(B,A) \tag{2.2.11}$$

If we expand these two terms out using the definitions above we obtain

$$P(B \mid A)P(A) = P(A \mid B)P(B) \tag{2.2.12}$$

This equivalence can be rearranged to produce one form of **Bayes' rule**

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{2.2.13}$$

Bayes' rule or more generally Bayesian statistics are widely used in probability theory and computational biology as is shown in the examples in this chapter.

Example 2.2.2: Rare Diseases

A test for a rare disease claims that it will report a positive result for 99.5% of people with the disease, and will report a negative result for 99.9% of those without the disease. We know that the disease is present in the population at 1 in 100,000. Knowing this information, what is the likelihood that an individual who tests positive will actually have the disease?

Solution

This problem provides a simple example of how Bayes' rule can be useful. As before, we begin by establishing our notation.

$P(+test\,|+disease)$ ← probability of a positive test result given that the patient has the disease. From the data this probability is 0.995.

$P(-test\,|-disease)$ ← probability of a negative test result given that the patient does not have the disease. From the data this probability is 0.999.

$P(+disease)$ ← probability that the patient has the disease given no other information. From the population average, this value is 0.00001.

$P(-disease)$ ← probability that the patient does not have the disease given no other information. Calculated from the population average, this value is 1-0.00001 or 0.99999.

What we want to find is

$P(+disease\,|+test)$ ← probability that a patient has the disease given a positive test result.

This unknown can be found directly using Bayes' rule

$$P(+disease\,|+test) = \frac{P(+test\,|+disease)P(+disease)}{P(+test)}$$

To evaluate this expression, we need to know the probability of a positive test, which can be found by marginalizing over all possible disease states (+disease and –disease)

$$\begin{aligned} P(+test) &= \sum_{disease\ \ states} P(+test, disease) = \sum_{disease\ \ states} P(+test\,|\,disease)P(disease) \\ &= P(+test\,|+disease)P(+disease) + P(+test\,|-disease)P(-disease) \\ &= P(+test\,|+disease)P(+disease) + (1 - P(-test\,|-disease))P(-disease) \end{aligned}$$

For our given data this works out to be

$$P(+test) = (0.995)(0.00001) + (1.0 - 0.999)(0.99999) = 0.00100994$$

Now, all of the elements are known and can be directly calculated.

$$P(+disease \mid +test) = \frac{P(+test \mid +disease)P(+disease)}{P(+test)}$$

$$= \frac{(0.995)(0.00001)}{0.00100994} = 0.0099$$

Thus, if 100 people get a positive result, only one of these people will really have the disease.

<u>Example 2.2.3</u>: *The occasionally dishonest casino*

A classic example of how Bayes's rule can be used in probability theory is the case of the occasionally dishonest casino. In this example, a casino uses two kinds of dice. One kind of die is fair and is used 99% of the time. The unfair die rolls a six 50% of the time and is used for the rest of the time. (a) If we pick up a single die at random, how likely is it that we will roll a six? (b) If we roll a single die chosen at random for a few trials and observe the outcomes, how can we use this information to predict if the die is fair or unfair?

Solution

The first step in such a problem is to define the key probabilities that we do and do not know. Below we list these probabilities both in formal notation and in words.

$P(six \mid D_{fair})$ ← probability of rolling a six given that the die is fair

for a six sided fair die, this is 1 out of 6, or one sixth.

$P(six \mid D_{unfair})$ ← probability of rolling a six given that the die is unfair

from the given information, this is 50% or one half.

$P(D_{fair})$ ← probability that a randomly chosen die will be fair

from the given information, this is 99% if we have no other information

$P(D_{unfair})$ ← probability that a randomly chosen die will be unfair

from the given information, this is 100%-99% = 1% if we have no other information

(a) We can calculate the likelihood of rolling a six given a random die choice by calculating the marginal probability of rolling a six

$$P(six) = \sum_{fair\ \ and\ \ unfair} P(six \mid Die)P(Die) = P(six \mid D_{fair})P(D_{fair}) + P(six \mid D_{unfiar})P(D_{unfair})$$

$$= \frac{1}{6}\left(\frac{99}{100}\right) + \frac{1}{2}\left(\frac{1}{100}\right) = 0.17$$

(b) Next, we run a series of experiments on a single die by making rolls and observing the outcome. The results of these trials are shown below

Roll #	Result
1	3
2	1
3	6
4	6
5	6
6	6
7	4
8	6

From these data, we can now use Bayes' rule to update our prior belief about the identity of the die. With no experimental information, we only knew that the die was 99% likely to be fair. After the first roll of a 3, we can obtain a posterior probability that the die is fair

$$P(D_{fair} \mid Data) = \frac{P(Data \mid D_{fair})P(D_{fair})}{P(Data)}$$

similarly, we can do the same for the probability that the die is unfair

$$P(D_{unfair} \mid Data) = \frac{P(Data \mid D_{unfair})P(D_{unfair})}{P(Data)}$$

Note that in both cases, the probability of the data given the die is the probability of not rolling a six and therefore is independent of the specific roll value (e.g. 1,2,3,4 and 5 all have the same value). Therefore, the outcome probabilities for a *single* roll are

$$P(roll\ \ non-six\,|\,D_{fair}) = \frac{5}{6} \qquad P(roll\ \ six\,|\,D_{fair}) = \frac{1}{6}$$

$$P(roll\ \ non-six\,|\,D_{unfair}) = \frac{1}{2} \qquad P(roll\ \ six\,|\,D_{unfair}) = \frac{1}{2}$$

In both *P(D$_{fair}$ | Data)* and *P(D$_{unfair}$ | Data)* we are left with the probability of the data term, which can be calculated by noting that the two likelihoods add up to one

$$P(D_{fair}\,|\,Data) + P(D_{unfair}\,|\,Data) = 1$$

therefore by rearrangement

$$P(Data) = \sum_{fair\ \ and\ \ unfair} P(Data,Die)$$
$$= P(Data\,|\,D_{fair})P(D_{fair}) + P(Data\,|\,D_{unfair})P(D_{unfiar})$$

This last probability term can be understood as just the probability of the data marginalized over all possible kinds of dice.

Finally, we can now calculate the probabilities for both cases

$$P(D_{fair}\,|\,Data) = \frac{\frac{5}{6}\left(\frac{99}{100}\right)}{\frac{5}{6}\left(\frac{99}{100}\right) + \frac{1}{2}\left(\frac{1}{100}\right)} = 0.994$$

$$P(D_{unfair}\,|\,Data) = (1-0.994) = 0.006$$

Thus, the first roll increases the probability that the die is fair.

Next, we will examine the probability that the die is fair based on all eight rolls. In this case, 5 out of 8 of the rolls are sixes. In this situation, our prior beliefs about the fairness of the die are unchanged. What does change is the likelihood of the data given a particular die. Thus, for the fair die,

$$P(Data\,|\,D_{fair}) = \left(\frac{5}{6}\right)^3\left(\frac{1}{6}\right)^5 = 7.44\times 10^{-5}$$

and for the unfair die

$$P(Data\,|\,D_{unfair}) = \left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^5 = 3.91\times 10^{-3}$$

From this information we can calculate the updated posterior probability of the die identity given data as

$$P(D_{fair} \mid Data) = \frac{7.44 \times 10^{-5}\left(\frac{99}{100}\right)}{7.44 \times 10^{-5}\left(\frac{99}{100}\right) + 3.91 \times 10^{-3}\left(\frac{1}{100}\right)} = 0.6535$$

$$P(D_{unfair} \mid Data) = \frac{3.91 \times 10^{-3}\left(\frac{1}{100}\right)}{7.44 \times 10^{-5}\left(\frac{99}{100}\right) + 3.91 \times 10^{-3}\left(\frac{1}{100}\right)} = 0.3465$$

Therefore, after only eight experiments we are now ~34 times more confident that this die is unfair.

Note that this same model could be used to search genomic DNA for locally enriched patterns, such as CpG islands. In this case, instead of six and non-six rolls, we could search a sequence for Gs and Cs versus As and Ts. At every position in a sequence, we could calculate the probability that we were in a CpG island, just as we can calculate the probability that the casino is using a loaded die.

For more information, see

- Chapter 3 of “The Cartoon Guide to Statistics” by L. Gonick and W. Smith

Problems

1. In Example 2.1.1, what is the conditional probability that the first base is T given that the second base is A? i.e. what is $P(s_1=T \mid s_2=A)$?
2. Using the data in Example 2.2.1, calculate the probability that a protein is not hydrophobic given that it is a membrane bound protein.
3. The expression of two genes, A and B, are thought to be co-regulated based on a large body of expression array data. In all of these datasets, both genes are up-regulated 33% of the time, both are down 57% of the time, A is up while B is down 4% of the time, and B is down and A is up 6% of the time. (a) From this data, what is the probability that A is up-regulated independent of B? (b) What is the probability that B is down-regulated independent of A?
4. In example 2.2.2, we found that a single positive test result of a rare disease will increase the odds that a patient has the disease from 1 in 100,000 to 1 in 100. (a) What is the probability that the patient actually has the disease if he or she tests positive for the disease twice? (b) What if the patient tests positive for the disease three times in a row? Assume that the error in the test is random.
5. In Example 2.2.3, (a) what is the probability of rolling 4 sixes in a row, independent of which die is used? (b) What is the conditional probability that the unfair die was used given that four sixes were rolled in a row?
6. In a really dishonest casino, they use loaded dice 50% of the time. We roll a single die from this casino 10 times and observe the following sequence: 2,6,4,6,3,4,6,6,4,1 Assuming that all of the other parameters are the same from Example 2.2.3 except for $P(D_{unfair})=P(D_{fair})=0.5$, what is the probability that the die we just rolled was fair?

Chapter 3: Discrete Distributions

In the previous section, we explored how probability theory could be used to analyze systems with only two outcomes (e.g. heads or tails, six or non-six, and diseased or healthy). In this section, we will show how these results can be generalized for repeated trials, and where each event has two or more outcomes. Starting with the simplest example of the geometric distribution, we will then move to the binomial and Poisson distributions. Finally, we will show how the discrete distributions introduced in this chapter can be further extended to systems with more than two outcomes using multinomial distributions.

3.0 Geometric Distribution

The geometric distribution is used to describe the probability of success at a given trial, and is best illustrated with an example. Imagine that we are trying to stably transfect a cell line with a reporter construct. From previous experience with this system, we know that we have a 30% chance of success on each experiment (or trial). If each experiment is independent, then we can calculate the probability that the experiment will succeed on any given experiment. For example, in the first experiment we know that we have a 30% chance of success, so the probability of success is 0.3. If we succeed, we stop. However if we fail, then we continue on to the second experiment. Thus the probability that the second experiment will succeed is equal to (0.7)(0.3), which is the probability that the first experiment failed and the second experiment succeeded. This procedure can then be iterated for larger and larger numbers of experiments in the following way

Experiment #	p(success on this experiment)
1	0.30
2	(0.30)(0.70)
3	$(0.30)(0.70)^2$
4	$(0.30)(0.70)^3$
…	….
N	$(0.30)(0.70)^{N-1}$

Generalizing this procedure to an arbitrary number of steps, we then obtain the geometric distribution in the form shown below

$$P(N) = f(1-f)^{N-1} \tag{3.0.1}$$

where N is the trial or experiment number and *f* is the probability of success on any one experiment. This distribution is plotted for two values of f in Figure 3.0.1.

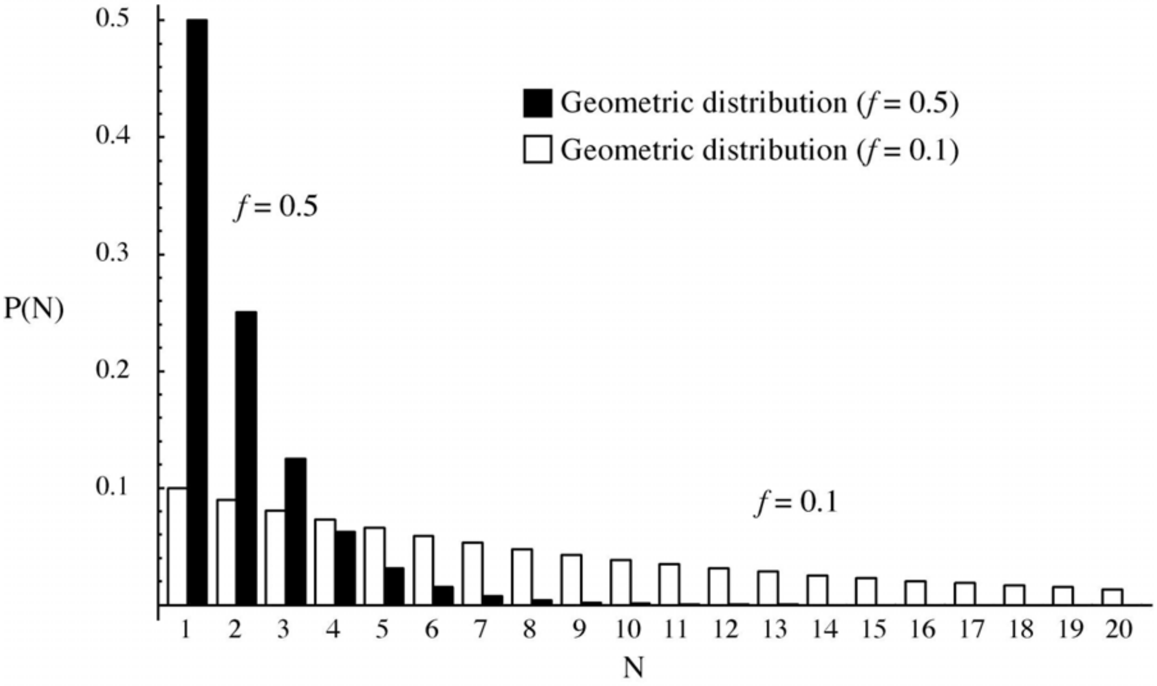


Figure 3.0.1: The geometric distribution plotted for *f*=0.5 and *f*=0.1 as a function of the number of trials or experiments, N.

To find the probability of success at a given trial, N, *or earlier*, we need to sum the geometric distribution. Fortunately, this expression has a simple closed form solution, as shown below

$$P(N \;\; or \;\; earlier) = \sum_{j=0}^{N} f(1-f)^{j-1} = 1-(1-f)^{N} \tag{3.0.2}$$

The sum up to a particular value in a probability distribution is known as the **cumulative distribution function** (or **CDF**) for that distribution. Unfortunately, the CDFs for most probability distributions are complicated and often can not be expressed analytically.

Example 3.0.1: *Drug screening*

Many of the pharmaceuticals on the market today were found using high-throughput screening assays. In these assays, ~1 million random molecules are tested and of these only a few show appreciable activity. If we assume that the success rate in these screens is one in ten thousand (p=0.0001), then how large of a library do we need to be 99% sure that we will find at least one active molecule?

Because this problem is asking when we will find our first success from a series of random, independent trials we can model this problem using a geometric distribution. We can find the size of the library using the distribution function from Eqn. 3.0.2 above

$$P(N\ \ or\ \ earlier) = 0.99 = 1 - (1 - 0.0001)^N$$

Solving by rearrangement, we find that our library size must contain 46,049 compounds if we are to find at least one active compound with 99% probability.

3.1 Binomial Distribution & Bernoulli Trials

Imagine that we know the fraction of Gs + Cs in the genome, and want to use this information to build a model of a piece of DNA of arbitrary length. If we call the fraction of Gs + Cs in the genome *f,* then choosing a G or C versus an A or T is akin to tossing a weighted coin that comes up heads with a probability of *f* and tails with a probability of (1-*f)*. Thus, the probability that a randomly chosen base is G or C is *f*, and the probability that it is an A or T is (1-*f*).

What if we choose a sequence that is two bases long? Here the probability of each event can be written in tree form as is shown below

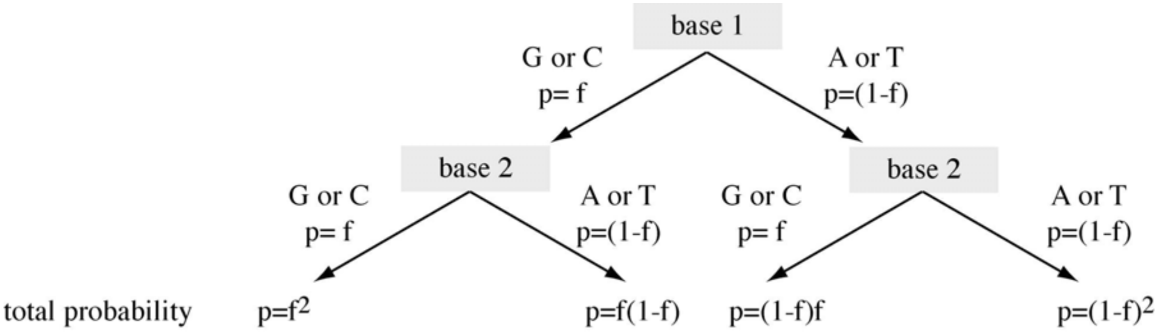


Figure 3.1.1: Tree representation of how probabilities distribute across a two-base sequence.

If we do not care what order these bases come in, then the middle two outcomes in the figure above can be merged. For example, if we are asking what is the probability of finding a two base pair sequence that has one A or T and one G or C, then the total probability can be found by adding the two middle probabilities to yield 2(1-*f)f.*

To generalize this problem, we can ask what is the probability of finding a sequence with N G or C bases and M A or T bases? Assuming that the bases occur independently of one another (as if generated by independent coin flips), then the probability of finding such a sequence is proportional to the product of the probabilities:

$$P(N_{G/C}, M_{A/T}) = kf^{N_{G/C}}(1-f)^{M_{A/T}} \qquad (3.1.1)$$

In this expression, k is a combinatorial term that represents the number of distinct sequences composed of N Gs and Cs and M As and Ts. We can generate values for k by just iterating the tree structure above, yielding Pascal's triangle. Alternatively, we can express k as the number of possible combinations where order does not matter:

$$k = \binom{N_{G/C} + M_{A/T}}{N_{G/C}} = \frac{(N_{G/C} + M_{A/T})!}{N_{G/C}!\,M_{A/T}!} \qquad (3.1.2)$$

In Eqn 3.1.2, ! is the factorial operator, which is defined as x!=x(x-1)(x-2)…(2)(1). The numerator represents the number of ways that we can rearrange any sequence of length (N+M) divided by the number of ways that we can rearrange M and N identical items. For an illustration of this operation, see example 3.1.1.

<u>Example 3.1.1</u>: *96 well plate combinations*

In a 96 well plate, we add cells to 10 wells, and leave the rest of the wells empty. How many different configurations of plates could we make?

In this problem, we will assume that all of the wells with cells are equivalent, so we are asking how many different ways can we arrange 10 identical things in 96 slots. We could evaluate this quantity by numbering each well and writing out every combination e.g.

#1	1,2,3,4,5,6,7,8,9,10
#2	1,2,3,4,5,6,7,8,9,11
#3	1,2,3,4,5,6,7,8,9,12

However, this approach would take a very long time. Alternatively we could use the Equation 3.1.2 to evaluate the number of possible combinations. Here, the total number of wells is 96, while the number of empty wells after adding cells is 86. Thus the total number of configurations can be calculated as

$$\frac{96!}{86!10!} = \frac{96\cdot 95\cdot 94\cdot 93\cdot 92\cdot 91\cdot 90\cdot 89\cdot 88\cdot 87}{10\cdot 9\cdot 8\cdot 7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1} \approx 1.13\times 10^{13}$$

Thus our final expression for the probability of N Gs and Cs and M As and Ts is

$$P(N_{G/C}, M_{A/T}) = \frac{(N_{G/C} + M_{A/T})!}{N_{G/C}!M_{A/T}!} f^{N_{G/C}} (1-f)^{M_{A/T}} \qquad (3.1.3)$$

or

$$P(N_{G/C}, L) = \frac{L!}{N_{G/C}!(L - N_{G/C})!} f^{N_{G/C}} (1-f)^{L-N_{G/C}} \qquad (3.1.4)$$

for $N_{G/C}$ and $M_{A/T}$ non-negative integers, and $L=N_{G/C}+M_{A/T}$. Equation 3.1.3 and 3.1.4 are both forms of the **Binomial Distribution**.

For given values of *f* and L we can then plot the Binomial probability distribution as is in Figure 3.1.2. An important property of all well formed probability distributions is that the area under its curve sums to one, indicating that all possible configurations are included. Said more formally for the Binomial distribution the following must be true:

$$\sum_{N=0}^{L} \frac{L!}{N!(L-N)!} f^{N} (1-f)^{L-N} = 1 \qquad (3.1.5)$$

for $0 \le f \le 1$ and $L \ge 0$.

The mean and standard deviation of a binomial distribution also have simple closed form solutions and are listed below

$$\begin{aligned} \mu &= E[N] = \langle N \rangle = Lf \\ \sigma &= \sqrt{E[(N-E[N])^2]} = \sqrt{\langle N^2 \rangle - \langle N \rangle^2} = \sqrt{Lf(1-f)} \end{aligned} \qquad (3.1.6)$$

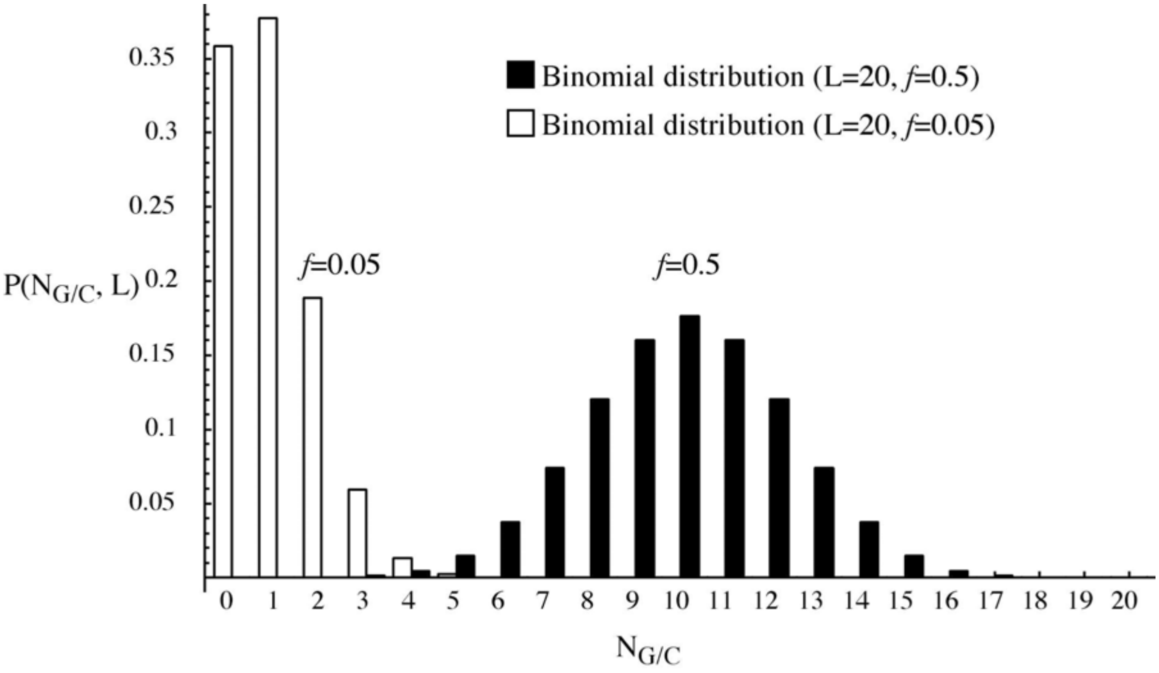


Figure 3.1.2: Plot of the Binomial distribution at L=20 and two different values of *f* as a function of $N_{G/C}$.

The experiment used to generate the data for a binomial distribution is called a **Bernoulli trial**. In a Bernoulli trial, we repeat an experiment for some number X times. At the end of each experiment, the result is one of two outcomes, such as heads vs. tails, absent vs. present, or success vs. failure. Two requirements of this test are that the result of each experiment be independent of the results of previous experiments, and that the underlying probability of the event remains the same throughout the trial.

Example 3.1.2: Probabilistic design of siRNA

According to standard practice, silencing RNA or siRNA should be between 21 and 23 base pairs long and should ideally have a GC fraction of approximately 45%-55%. Human genomic DNA has an average GC fraction of 41% for the whole genome. If we are making a siRNA that is 22 bases long, what is the probability that a sequence randomly chosen from the genome will have exactly 11 (50%) G/C bases? What is the probability that a 22 base pair sequence will have 10-12 GC bases (45%-55%)?

This problem can be answered using the binomial distribution. The probability mass function for this system can be written as

$$P(N_{G/C}) = \frac{N_{total}!}{N_{G/C}!(N_{total} - N_{G/C})!} f^{N_{G/C}} (1-f)^{(N_{total}-N_{G/C})}$$

where N_{total} is the total number of nucleotides in the sequence and is equal to 22. Substituting in for the values given in the problem statement we obtain the expression

$$P(N_{G/C}) = \frac{22!}{N_{G/C}!(22 - N_{G/C})!} 0.41^{N_{G/C}} (1-0.41)^{(22-N_{G/C})}$$

A plot of this distribution (Figure 3.1.3) exhibits the bell shaped curve that characterizes a binomial distribution when *f* is near 0.5.

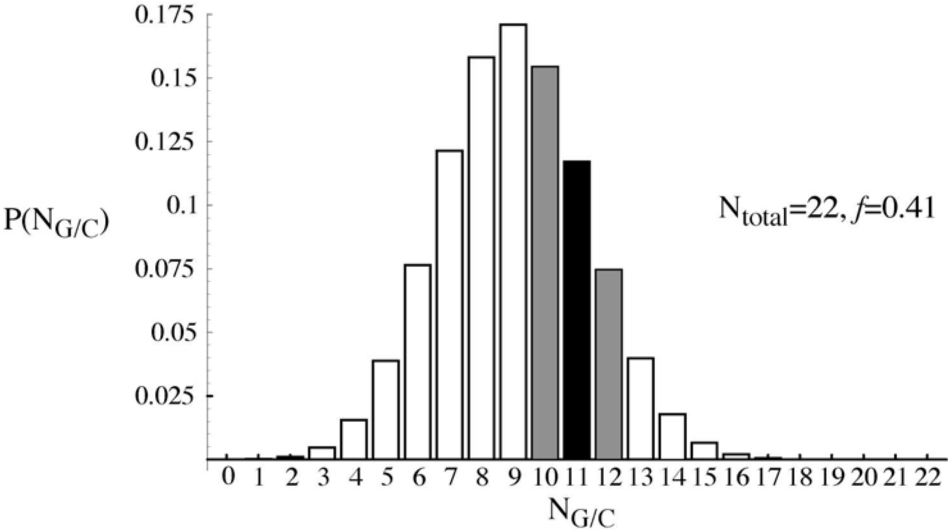


Figure 3.1.3: A histogram of the binomial distribution with N_{total}=22 and *f*=0.41. The probability of a sequence with 50% G/C content is shown in black, and those that are approximately 45% and 55% are shown in gray.

To calculate the probability of finding a sequence with 50% G/C content (11 Gs or Cs in a 22 base sequence), we plug in $N_{G/C}$=11 and calculate a probability of 0.117, or 11.7%. To calculate the probability of a range of values we take the sum of the individual probabilities

$$P(55\% > GC > 45\%) = \sum_{possible\ \ values} P(N_{G/C}) = P(10) + P(11) + P(12)$$
$$= 0.154 + 0.117 + 0.076 \approx 0.35$$

Self Test:

What is the probability of finding a sequence with 2 or fewer Gs or Cs?

Answer: 0.0116 or 1.16%

Example 3.1.3: *Mutant search.*

As part of a large genetic study, it is found that 550 out of 1000 people carry a mutation that may make them more susceptible to a rare disease. (a) Assuming that this sample was taken at random, and assuming that people either do or do not have the mutation, write out the appropriate probability distribution to describe this experiment. (b) From this data, what is the probability that fewer than 50% of the people in the greater population have the mutation?

Answer:

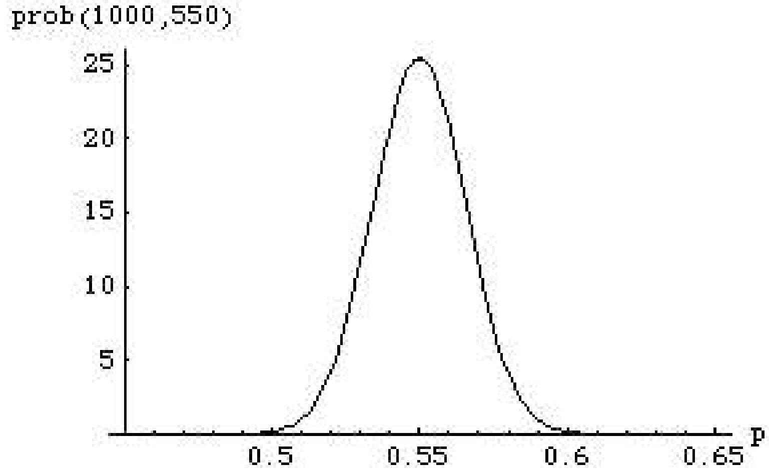
(a) As there are only two outcomes possible (mutant or non-mutant) and the sample is random, the results of this experiment are best described by a binomial distribution. The general form for this distribution is:

$$prob(n,k) = \binom{n}{k} p^k (1-p)^{n-k}$$

For the given parameters this expression can be rewritten as:

$$prob(1000,550) = \frac{1000!}{550!(450)!} p^{550} (1-p)^{450}$$

A plot of this distribution shows its general shape



(b) To find the probability that fewer than 50% of the people have the mutation, we need to find the area under the curve above from p=0 to p=0.5, and divide that by the total area under the curve. This answer assumes that *a priori*, *i.e*. before collecting the data on the 1000 people, all allele frequencies between 0 and 1 were equally likely and is an example of Bayesian statistics. In mathematical terms, this can be stated as

$$prob(p \le 0.5) = \frac{\int_0^{0.5} prob(1000{,}550)dp}{\int_0^{1.0} prob(1000{,}550)dp} = \frac{\left(\frac{1000!}{550!450!}\right)\int_0^{0.5} p^{550}(1-p)^{450}\,dp}{\left(\frac{1000!}{550!450!}\right)\int_0^{1.0} p^{550}(1-p)^{450}\,dp}$$

This integral can be performed numerically or analytically to yield a probability of 0.000781. Thus it is very unlikely that less than 50% of the population has the mutation.

The method used in this example to determine if a statement is statistically relevant is a form of statistical inference, and will be discussed in greater detail in Chapter 5.

3.3 Poisson Distribution

In some cases, limiting versions of the Binomial distribution can be more useful for describing a system. For example, what happens if we do not know the total number of events that take place but only know the expected number of successes? For example, if we know the mean (expected) number of stem cells in a tissue sample, then how can we predict the probability that more than twice that number will be present due to random chance in a replicate sample? To make such a measure, we begin with the binomial distribution

$$P_p(k \mid N) = \frac{(N)!}{k!(N-k)!} p^k (1-p)^{N-k} \tag{3.3.1}$$

with this distribution we can rescale to the expected number of successes, μ, which is defined as

$$\mu \equiv Np \tag{3.3.2}$$

Substituting for p we find

$$P_{\mu/N}(k \mid N) = \frac{(N)!}{k!(N-k)!} \left(\frac{\mu}{N}\right)^k \left(1 - \frac{\mu}{N}\right)^{N-k} \tag{3.3.3}$$

If we allow the sample size, N, to increase without bound while keeping μ finite, the distribution approaches

$$P_\mu(k) = \frac{\mu^k e^{-\mu}}{k!} \tag{3.3.4}$$

which is known as the **Poisson distribution**. Note that the sample size is no longer a part of the expression. A plot of the Poisson distribution is shown in Figure 3.3.1

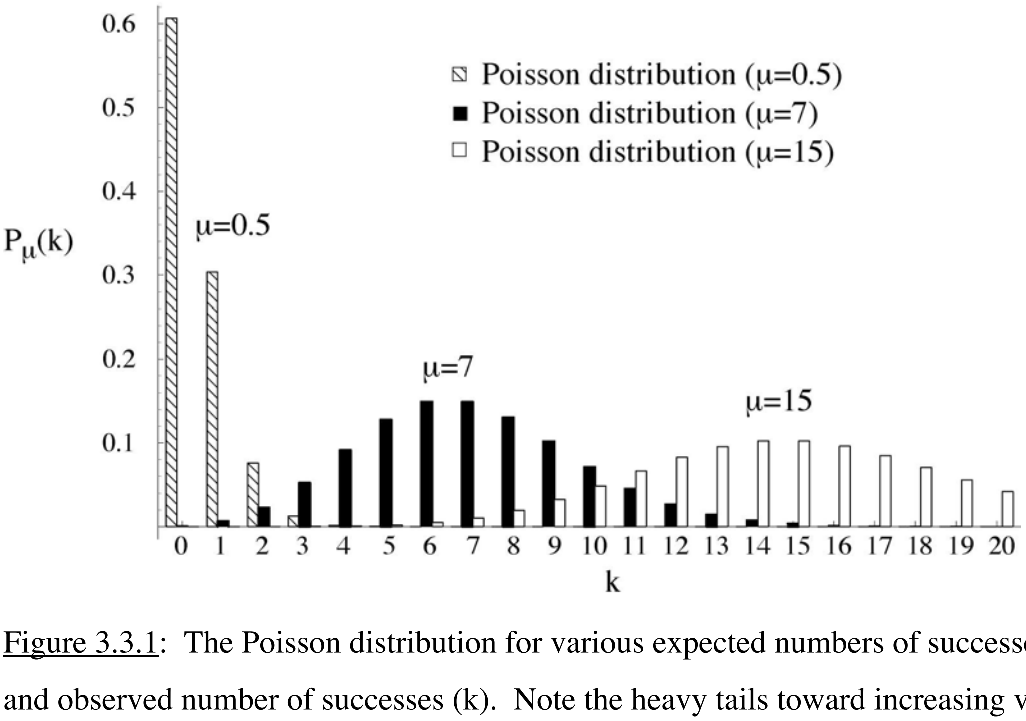


Figure 3.3.1: The Poisson distribution for various expected numbers of successes (μ), and observed number of successes (k). Note the heavy tails toward increasing values of k.

By definition, a **Poisson process** generates a Poisson distribution. A Poisson process is characterized by the following three traits

1) Outcomes are discrete.
2) The number of successes, k, in any interval is independent of the number of successes in any other interval.
3) The probability of two or more successes over a sufficiently small interval is essentially zero.

These requirements do not mean that Poisson processes only model rare events. However, it does require that if we reduce the interval (e.g. window of our observation), either in time or space, then we will see less than two events. Thus, whether or not a process is Poisson distributed depends on how the question is asked in many cases. For example, repeated binary trials such as a coin toss are in general binomially distributed. However, if we ask what is the probability of seeing 10 heads tossed in a row in a series of 100 tosses, this is Poisson distributed. Biological examples of uses of the Poisson distribution include calculating the number of proteins that decay over a given time interval or the number of stem cells in a given sample volume.

Example 3.3.1: Rare cell types

Some cells types in a tissue represent only a very small portion of total cell population, such as hematopoetic stem cells in blood. Imagine that a particular cell type is present at a rate of 1 cell in 100,000. If we are given a sample with 50,000 cells, what is the probability of finding exactly 1 of these rare cells? If we need at least 20 of these cells with 95% confidence, how many total cells must we collect?

Rare cell types satisfy the three requirements for a Poisson process

1) Cell numbers are discrete
2) Rare cells are randomly distributed among samples
3) If the interval size is made smaller (e.g. our sample size is reduced), the probability of finding two or more rare cells goes to zero.

Therefore we will assume that the number of rare cells is distributed according to a Poisson distribution. To answer the first question, we begin by finding the expected number of rare cells

$$\mu = Np = (50{,}000)\left(\frac{1}{100{,}000}\right) = 0.5$$

Thus we expect to find half a cell per 50,000 cells sampled. Next we use the expression for the Poisson distribution to calculate the probability of finding exactly one rare cell

$$P_\mu(k) = \frac{\mu^k e^{-\mu}}{k!} = \frac{(0.5)^1 e^{-0.5}}{1!} = 0.303$$

This result means that approximately 30% of the time we will find one rare cell in a sample of 50,000 cells.

The second problem is more difficult, as we are now calculating N. We begin as before by calculating the expected number of rare cells

$$\mu = Np = \frac{N}{100{,}000}$$

Next we write down an expression for the probability of finding 20 or more rare cells for a given μ value.

$$P_\mu(k \ge 20) = 0.95 = \sum_{k=20}^{\infty} \frac{\mu^k e^{-\mu}}{k!} = 1.0 - \sum_{k=0}^{20} \frac{\mu^k e^{-\mu}}{k!} = 1.0 - \sum_{k=0}^{20} \left[\left(\frac{N}{100{,}000}\right)^k \frac{e^{-\left(\frac{N}{100{,}000}\right)}}{k!} \right]$$

The simplest method to solve this expression is to use a numerical solver or to just try various values of N. For example, the Sum[ ] function in Mathematica is well suited to directly solve such expressions. Using such an approach, we find that we need approximately 2.8 million cells to be 95% sure that we will have at least 20 rare cells in our sample.

Example 3.3.2: Bacteriophages

1 ml of a bacteriophage suspension is mixed with 20 ml of a bacterial culture and 50% of the phages adsorb. We know that the bacteriophage suspension had a concentration of 1 x 10^{10} viruses per ml, and the bacterial culture had a concentration of 3

x 10^8 bacteria per ml. How many viruses are adsorbed per cell (multiplicity of infection)? What fraction of the cells is uninfected? Singly infected? Multiply infected?

To begin, we first calculate the total number of viruses and bacteria in the system.

Total viruses

1 ml (1 x 10^{10} viruses per ml) = 1 x 10^{10} viruses

Total bacteria

20 ml (3 x 10^8 bacteria per ml) = 6 x 10^9 bacteria

Next we calculate the average number of viruses absorbed per bacteria

$$\mu = \frac{\left(1\times 10^{10}\ \text{viruses}\right)\left(0.5\ \text{adsorb}\right)}{6\times 10^{9}\ \text{bacteria}} = 0.833\ \frac{\text{viruses}}{\text{bacteria}}$$

which is the multiplicity of infection for this system.

Next, assume a Poisson process, and calculate the fraction of bacteria that are uninfected (k=0), singly infected (k=1), and multiply infected (k>1) from the following Poisson distribution

$$P_{\mu}(k) = \frac{\mu^k e^{-\mu}}{k!} = \frac{(0.833)^k e^{-0.833}}{k!}$$

where the expected number of successes, μ, is the multiplicity of infection calculated above. Below are the results of each calculation

$$P_{\mu=0.833}(k=0) = \frac{(0.833)^0 e^{-0.833}}{0!} = 0.435$$

$$P_{\mu=0.833}(k=1) = \frac{(0.833)^1 e^{-0.833}}{1!} = 0.362$$

$$P_{\mu=0.833}(k>1) = 1.0 - \sum_{k=0}^{1}\frac{(0.833)^k e^{-0.833}}{k!} = 1.0 - 0.435 - 0.362 = 0.203$$

3.4 Multinomial Distribution

By definition, the binomial distribution can only model events with two outcomes. However, this definition can be generalized to describe events with an arbitrary number of outcomes to form a **multinomial distribution**. Consider an event that has k different outcomes, each of which is observed n_i times in a series of N trials. The probability density function for the multinomial distribution is then

$$P(n_1,n_2,..n_k) = \frac{N!}{n_1!n_2!\ldots n_k!}\prod_{i=1}^{k} p_i^{n_i} \tag{3.4.1}$$

where p_i is the probability of seeing outcome i. For the limited case of k=2 we recover the following expression

$$P(n_1,n_2) = \frac{(n_1+n_2)!}{n_1!n_2!}p_1^{n_1}p_2^{n_2} = \frac{N!}{n_1!(N-n_1)!}p_1^{n_1}(1-p_1)^{N-n_1} \tag{3.4.2}$$

which is identical to the binomial distribution.

Example 3.4.1: Multinomial analysis of STAT3 phosphorylation

The human transcription factor STAT3 is thought to play a central role in embryonic stem cell differentiation. One way the activity of STAT3 is regulated is by phosphorylation on two different amino acids: tyr705 and ser727. An experiment is run on a large sample to measure which sites are phosphorylated on STAT3 in an embryonic stem (ES) cell line, generating the following data:

Phosphorylation		
Tyr705	Ser727	frequency
N	N	0.013
N	Y	0.267
Y	N	0.031
Y	Y	0.689

Next we use a new mass-spectrometry tool to make 100 measurements of the individual phosphorylation state of STAT3 molecules in two differentiated cell lines to yield the following data

Cell line #1

Phosphorylation		
Tyr705	Ser727	number
N	N	38
N	Y	41
Y	N	1
Y	Y	20

Total		100

Cell line #2

Phosphorylation		
Tyr705	Ser727	number
N	N	30
N	Y	21
Y	N	6
Y	Y	43
Total		100

Which cell line is more like the undifferentiated ES cell based on its STAT3 phosphorylation profile? By how much?

Solution

In this case, we have a total of four distinct phosphorylation states, and as such a multinomial distribution is appropriate. The probability of any measurement can be written as

$$P(n_1,n_2,n_3,n_4) = \frac{N!}{n_1!n_2!n_3!n_4!} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4}$$

where each n_i represents the total number of proteins in a particular phosphorylation state i, and p_i is the probability of that state and N is the total number of observations. Calculation of the probability of any particular state can be achieved by simply plugging in the given states. Calculating for each cell line we find

Cell line #1

$$P(38,41,1,20) = \frac{100!}{38!41!1!20!} 0.013^{38} 0.267^{41} 0.031^{1} 0.689^{20} = 2.6 \times 10^{-55}$$

Cell line #2

$$P(30,21,6,43) = \frac{100!}{30!21!6!43!} 0.013^{30} 0.267^{21} 0.031^{6} 0.689^{43} = 3.7 \times 10^{-35}$$

Thus, cell line #2 is more like the ES cell line because it has a higher probability score. Because we have a quantitative measure of the probability that each cell line came from an ES cell population, we can describe the relative likelihood that each cell population is ES cell-like by taking the ratio of their probabilities. For our data, cell line #2 is ~10^{20} times more likely to be an ES cell.

Note: The intermediate values of calculations like these can pose numerical difficulties in many calculators due to their extremely large or small sizes. This problem can be circumvented by either using mathematics software packages that maintain arbitrary levels of precision (such as Mathematica) or by using approximations like Sterling's approximation (see http://mathworld.wolfram.com/StirlingsApproximation.html).

Part II

Assuming a total of 100 STAT3 proteins in a cell, what is the probability of the following conditions?

Phosphorylation		
Tyr705	Ser727	number
N	N	$n_1 = 0$
N	Y	$n_2 \leq 20$
Y	N	$n_3 \leq 2$
Y	Y	$n_4 = 100 - n_2 - n_3$
Total		100

Solution

Here we are asked to calculate the probability a family of configurations given some constraints. Some of the information given in the problem can be included directly into the multinomial probability expression as is shown below

$$P(0,n_2,n_3,100-n_2-n_3) = \frac{100!}{n_2!n_3!(100-n_2-n_3)!} p_1^0 p_2^{n_2} p_3^{n_3} p_4^{100-n_2-n_3}$$

A plot of this distribution is shown below in Figure 3.4.1.

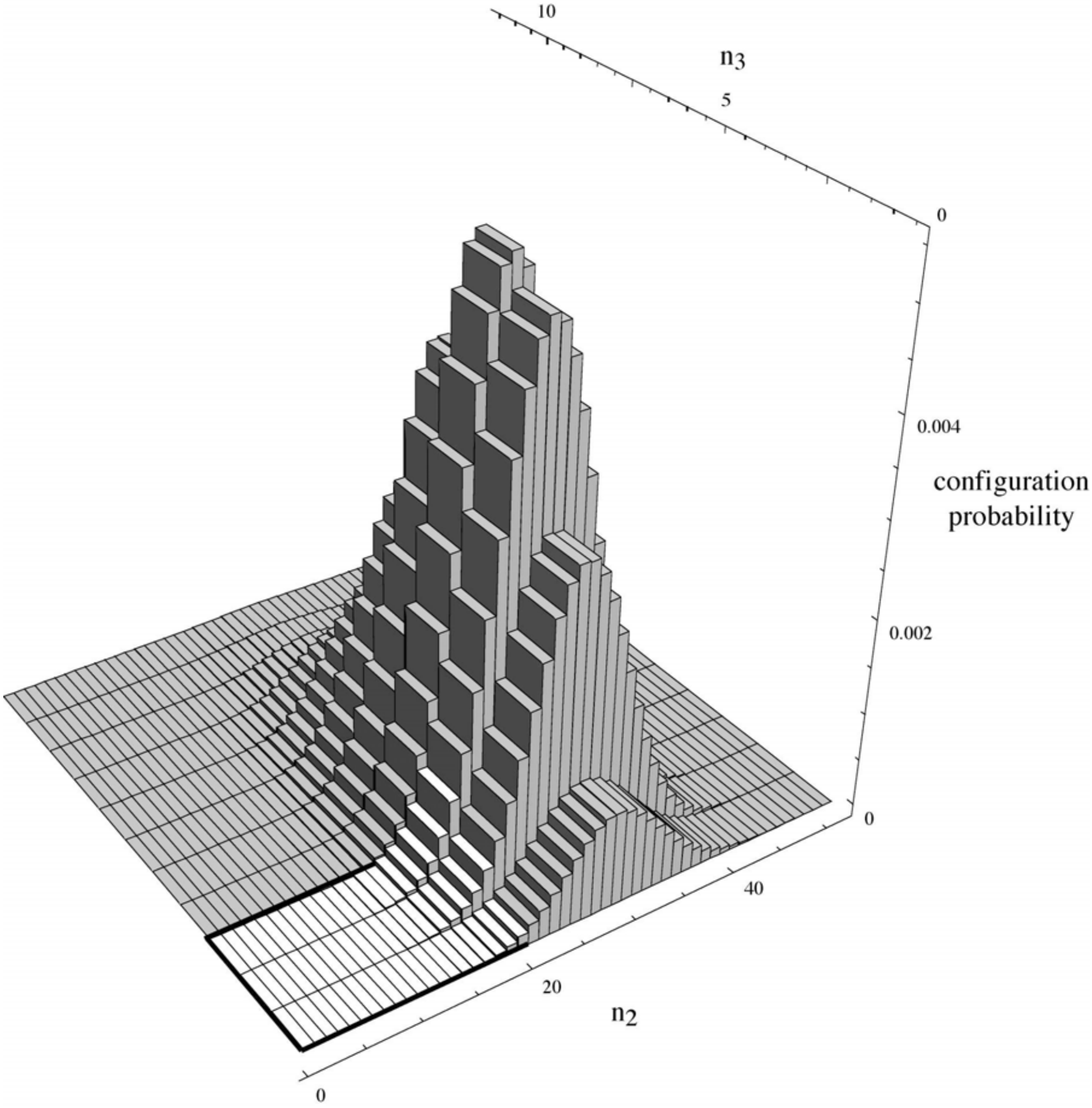


Figure 3.4.1: Multinomial distribution for the probability of finding n_2 and n_3 STAT3 proteins in a particular phosphorylation state. The probability of finding $n_2 \leq 20$ and $n_3 \leq 2$ is the volume under the lower left hand side of this surface (shaded).

As before, to find the total probability of a range of values, we sum the possible combinations. Because we have two variables that are changing, we sum over two variables

$$P(0,n_2 \le 20,n_3 \le 2,100-n_2-n_3) = \sum_{n_2=0}^{20}\sum_{n_3=0}^{2} \frac{100!}{n_2!n_3!(100-n_2-n_3)!} p_1^0 p_2^{n_2} p_3^{n_3} p_4^{100-n_2-n_3}$$

This double sum yields a net probability of 0.0057.

3.5 Hypergeometric Distribution

Consider the case where we sample without replacement from a set of N items. In this set there are K items that are different (say phosphorylated or dead). If we take a sample of size n without replacement, the probability of finding k different items is described by a hypergeometric distribution:

$$P(X=k) = \frac{\begin{pmatrix} K \\ k \end{pmatrix}\begin{pmatrix} N-K \\ n-k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}, \quad k = 0,1,2,...,n \qquad (3.5.1)$$

where the quantities in brackets are the numbers of possible combinations, just as in Eqn. 3.1.2. Note that this format assumes that our sample size is smaller than K. If our sample size is larger than K, then the number of possible different items observed, k, can be no greater than K. We can interpret this distribution as the number of possible ways to arrange k of K items times the number of ways to arrange the non-k items, divided by the number of ways to arrange all of the items together.

<u>Example 3.5.1</u>: *Pocket change*

In our pocket we know that we have 100 pennies and 10 dimes. If we randomly draw 5 coins from our pocket (without replacement), what is the probability of finding exactly 1 dime? 0 dimes?

Because we are drawing without replacement from a pool that has binary outcomes (dime or penny), we can model this process using a hypergeometric distribution with N=110, K=10, n=5, and k=0 or 1. For k=0

$$P(X=0)=\frac{\binom{10}{0}\binom{100}{5}}{\binom{110}{5}}=\frac{\left(\frac{10!}{10!0!}\right)\left(\frac{100!}{5!95!}\right)}{\left(\frac{110!}{5!105!}\right)}=0.615$$

and for k=1

$$P(X=1)=\frac{\binom{10}{1}\binom{100}{4}}{\binom{110}{5}}=\frac{\left(\frac{10!}{9!1!}\right)\left(\frac{100!}{4!96!}\right)}{\left(\frac{110!}{5!105!}\right)}=0.320$$

For more information, see

- Chapter 5 of "The Cartoon Guide to Statistics" by L. Gonick and W. Smith
- Chapter 3 of "Elements of Engineering Probability and Statistics" by Rodger E. Ziemer (1997).
- Eric Weissteins's World of Mathematics at mathworld.wolfram.com
- Chapter 11 of "Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids" by R. Durbin *et al.*

Problems

1. If we toss a particular thumbtack, it will land point down 35% of the time. (a) If we toss this same thumbtack 50 times in a row, what is the expected (mean) number of times that the tack will land point down? (b) What is the standard deviation?
2. The human genome contains approximately 30,000 genes, of which approximately 800 are G-protein coupled receptors. (a) If we were to choose genes at random from the genome *with replacement*, what is the probability of finding at least one G-protein coupled receptor in the first 10 samples? (b) If we were to choose genes at random from the genome *without replacement*, what is the probability of finding at least one G-protein coupled receptor in the first 10 samples?
3. When searching the human genome for a particular motif, we find that motif an average of 5 times per 300,000 base pairs. Assuming the motif is randomly distributed, what is the probability of finding (a) the motif 0 times in 300,000 base pairs? (b) the motif 10 times in 300,000 base pairs
4. Mendelian genetics suggests that crossing two heterozygous mutants will produce offspring that are 25% wild type, 50% heterozygous, and 25% homozygous for the mutation. Experimentally, we find that one such crossing yields 6 wild type, 25 heterozygous, and 11 homozygous offspring. What is the probability of this particular outcome assuming a Mendelian model?

5.* Show that for the following form of the Binomial distribution

$$\frac{(N)!}{k!(N-k)!}p^k(1-p)^{N-k}$$

the mean is

$$\mu =< k >= Np$$

Hint: The average or *expectation value* of a function is defined as

$$\langle f(x)\rangle = \sum_x f(x)P(x)$$

thus the average value of k can be expressed as

$$\langle k\rangle = \sum_{k=0}^{N}\left((k)\left(\frac{(N)!}{k!(N-k)!}p^k(1-p)^{N-k}\right)\right)$$

Chapter 4: Continuous Distributions

In the previous sections, we described cases where we could count the number of discrete events that took place and described them with the distributions such as the binomial, multinomial, or Poisson. However, in many cases we either cannot directly count the number of events taking place or that number is too large to be practical. In other cases, variables such as the time, length, and weight are naturally continuous rather than discrete. Therefore we need a different kind of probability model to describe these situations.

Fortunately, there exist a large family of continuous functions that lend themselves to describe probability distributions. In some cases, these distributions are just continuous analogs of discrete distributions, while in other cases the distributions are empirical, and as such based only on data. Operationally, a key difference between continuous and discrete distributions is that summing in discrete distributions is done with sums, while summing a continuous distribution means integrating. By allowing us to use the tools of calculus, continuous distributions provide many advantages not available to discrete distributions.

Common to all continuous distributions are the concepts of a **probability distribution function** (**PDF**) and a **cumulative distribution function** (**CDF**). A probability distribution function describes the instantaneous probability for a particular point. From the PDF, or a plot of this function, one can identify which regions are more and less likely to occur or be observed by looking for large PDF values. A general property of PDFs is that the area under the curve must be equal to exactly one, meaning that the total probability of finding any possible value must be 100%. A cumulative distribution function is the integral of a PDF from negative infinity up to the value on the ordinate. The CDF is useful for it describes the cumulative probability of observing a value equal to or less than any other value. In general, a CDF will increase as the variable of interest increases, eventually reaching a maximum value of 1 at positive infinity. Plots of two different PDFs and CDFs are shown in Figure 4.0.1.

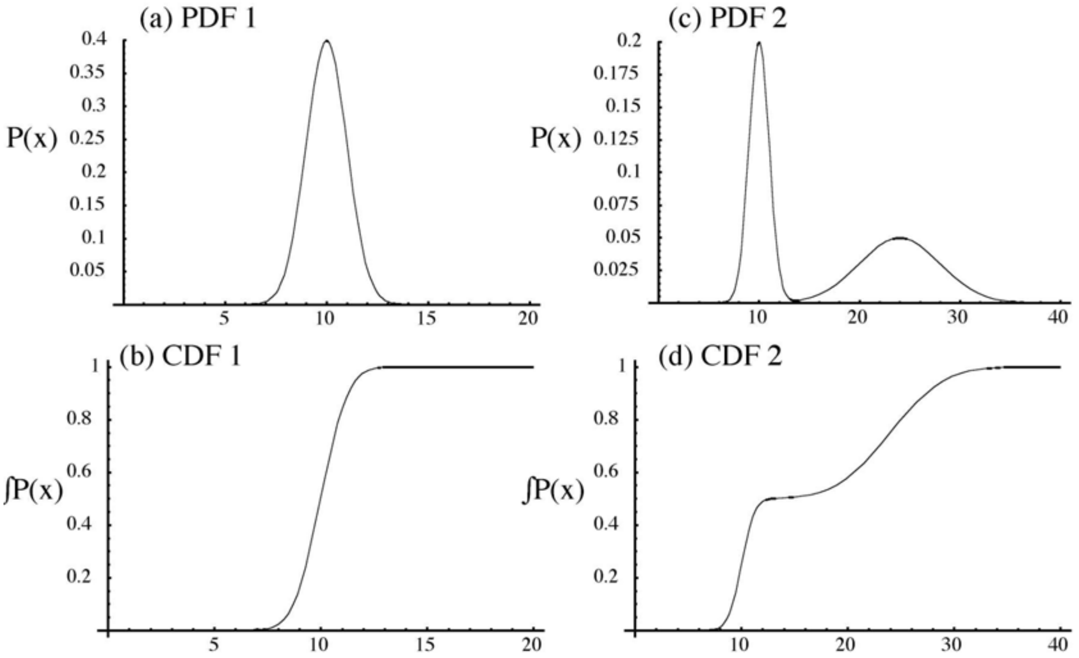


Figure 4.0.1: Plots of the probability distribution function, (a) and (c), and cumulative distribution function, (b) and (d) for two different distributions.

In this chapter we will show some of the more common continuous distributions, where they come from, and how they are used in computational biology.

4.1 Gaussian Distribution

The **Gaussian distribution**, or Normal distribution, is probably the most commonly encountered continuous distribution. Each time you take a set of data, average it and calculate the standard deviation of that data, one implicitly assumes that the underlying distribution is Gaussian.

The analytical expression of the Gaussian distribution can derived from the binomial distribution using the *De Moivre-Laplace theorem.* The proof of this relationship is lengthy, and as such we direct the interested reader to the original work by Uspensky (1937) listed at the end of this chapter. The result of this calculation is an analytical expression of the Gaussian distribution

$$P(k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(k-\mu)^2/2\sigma^2} \qquad (4.1.1)$$

This relationship between the binomial distribution and the Gaussian distribution can also be shown graphically as in Figure 4.1.1.

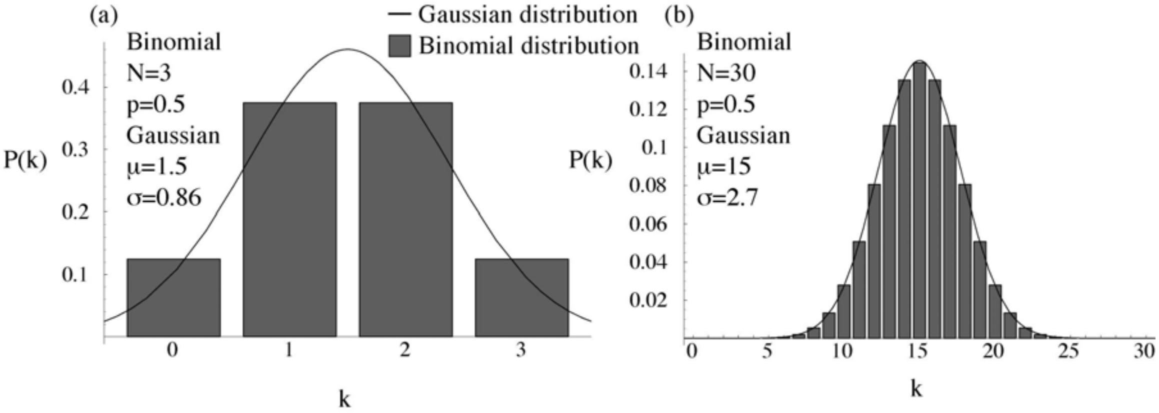


Figure 4.1.1: Relationships between the binomial distribution and Gaussian distribution. (a) For small values of N, the binomial distribution is only poorly approximated by the Gaussian distribution. (b) For intermediate to large values of N the Gaussian and binomial distributions are nearly identical.

To obtain a probability estimate from a Gaussian distribution, we need to integrate the Gaussian PDF over some finite range

$$P(k_1 < x < k_2) = \int_{k_1}^{k_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx = \frac{1}{2}\left(Erf\left[\frac{k_2-\mu}{\sigma\sqrt{2}}\right] - Erf\left[\frac{k_1-\mu}{\sigma\sqrt{2}}\right]\right) \qquad (4.1.2)$$

where *Erf* is the error function which is a function much like sin or log and is available on many scientific calculators and as the function Erf[ ] in Mathematica. Using a similar approach we can derive the CDF for the Gaussian distribution as

$$CDF_{Gaussian} = P(-\infty < x < k) = \int_{-\infty}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx = \frac{1}{2}\left(Erf\left[\frac{k-\mu}{\sigma\sqrt{2}}\right] + 1\right) \qquad (4.1.3)$$

Example 4.1.1: Weighing mice

We are interested in the weight of a particular mouse, yet we have only crude scales with which to weigh the mouse. After trying out many scales, we obtain an average weight of 16 grams with a standard deviation of 3 grams. From this data, what is the probability that the mouse actually weighs more than 20 grams?

Solution

One approach to solve this problem is to assume that the underlying distribution of the measurements is Gaussian and integrate. Using the expression in Equation 4.4 above, we find

$$P(20 \le x < \infty) = \int_{20}^{\infty} \frac{1}{\sqrt{2\pi(3)^2}} e^{-(x-16)^2/2(3)^2} dx = \frac{1}{2}\left(Erf\left[\infty\right] - Erf\left[\frac{20-16}{3\sqrt{2}}\right]\right) = 0.0912$$

Thus, there is a 9.12% chance that the mouse actually weighs more than 20 grams.

Note that we assumed that the underlying distribution was Gaussian when we calculate a mean and standard deviation, but this approximation is not always accurate. For example, here a Gaussian distribution also predicts that there is a 0.000004% chance that the mouse could actually have a *negative* weight. Therefore, be aware that the Gaussian distribution is not always the best choice for a distribution.

Example 4.1.2: Boron in plants

The healthy range of boron concentrations in most plant leaves ranges from 25 to 200 ppm. Beyond this range, most plants become stunted or sick. A random sample of 100 plants yields a symmetric distribution of readings, centered at 40 ppm with a standard deviation of 10 ppm. From this sample, what percentage of plants likely suffers from a boron deficiency?

Solution

The boron measurement in this problem is a continuous value and the distribution is symmetric, therefore we will model boron concentrations as a Gaussian distribution. To find the percent of plants that are boron deficient, we can directly use the Gaussian CDF

$$CDF_{Gaussian} = P(-\infty < x < 25) = \frac{1}{2}\left(Erf\left[\frac{25-40}{10\sqrt{2}}\right] + 1 \right) = 0.06681$$

Thus, ~7% of the plants likely suffer a boron deficiency. Note that the CDF integrates from minus infinity to 25, while ppm measurements only go as low as zero. If we had used the more general form of the CDF in 4.1.2, and integrated from 0 to 25 we would have obtained a probability score of 0.06678, which is nearly identical to the CDF result.

4.2 Standard normal distribution and z-scores

A special case of the Gaussian distribution is one that is centered at zero (μ=0) with a standard deviation of exactly one (σ=1). This distribution is called the **standard normal distribution**, and provides a standardized way to discuss Gaussian distributions. Given a standard normal distribution, the distance from the origin is measured by the variable z, which is also called a **z-score**. The z-score is convenient because it is a single number that can be related to any non-standard normal distribution through the following transformation

$$z = \frac{x - \mu}{\sigma} \tag{4.2.1}$$

Tables in statistics textbooks often have pre-calculated tables that show how the z-score varies with the probability density.

Historically, the z-score was used in place of exact calculations using complicated functions such as the error function shown in 4.1.2, however the concept is still useful for making quick estimates. For example, if given the mean and standard deviation of a normally distributed sample, one can easily determine the range of values that encompass 90%, 95%, and 99% of the probability density by using z-values of 1.645, 1.960, and 2.575 respectively. Other z-score values can be looked up in most statistics textbooks or calculated using the CDF in equation 4.1.3.

Example 4.2.1: E. coli doubling times

After running repeated experiments, we find that the doubling time for a particular strain of *E. coli* is 58 minutes with a standard deviation of 10 minutes. Using z-scores, determine the range of expected doubling times at the 90% and 99% confidence levels.

Solution

We start by rearranging the definition of the z-score transform in equation 4.2.1 to solve for x

$$x_{upper} = \mu + z\sigma$$

$$x_{lower} = \mu - z\sigma$$

Here we provide both the upper and lower limit by symmetrically adding or subtracting from the mean. To find the upper limit of the 90% confidence interval we plug in a z-value of 1.645 to yield

$$x_{upper} = 58 + 1.645(10) = 74.45 \text{ min}$$

$$x_{lower} = 58 - 1.645(10) = 41.55 \text{ min}$$

At the 99% confidence interval we use a z-score of 2.575

$$x_{upper} = 58 + 2.575(10) = 83.75 \text{ min}$$

$$x_{lower} = 58 - 2.575(10) = 32.25 \text{ min}$$

From this calculation we are 95% sure that the doubling times will take between 41.55 and 74.45 minutes, and 99% sure that they will take place between 32.35 and 83.75 minutes.

4.3 Exponential distribution

When modeling the waiting time between successive changes, such as cell differentiation or DNA mutations, or the distance between successive events, such as the distance between simple repeats along the genome, one often uses the **exponential distribution**. The exponential distribution has the following probability distribution function

$$P(x) = \lambda e^{-\lambda x} \qquad (4.3.1)$$

The cumulative distribution function for the exponential distribution function has the form

$$CDF_{\exp}(x) = 1 - e^{-\lambda x} \qquad (4.3.2)$$

where λ is the unit rate of change.

Example 4.3.1: Mammalian mutation rates

The average mammalian genome mutation rate is 2.2x10^{-9} per base pair per year (Kumar, 2002). Given this rate, what is the probability that the interval between successive mutations at a particular base pair is 80 years or less (or one human lifespan)? Assuming a genome of 3 billion base pairs and independent mutation events, how many bases are expected to be mutated over this time span?

Solution

This problem is best described by an exponential distribution because we are given a continuous rate of change and asked the probability of an two events being separated by an interval in time. The probability that a mutation will take place in 80 years or less can be calculated directly using the CDF in equation 4.3.2.

$$CDF_{\exp}(x=80) = 1 - e^{-\lambda x} = 1 - e^{-(2.2\times10^{-9})(80)} = 1.76\times10^{-7}$$

Assuming 3x10^{9} bases, we would expect that after 80 years of life 538 bases would have mutated. Note that we ignore the possibility that a site might have mutated twice or more as the probability of a double-mutation is extremely low for this case.

The reader may have noticed that the exponential distribution and the Poisson distribution are related. The Poisson distribution describes the probability that a certain number of events have taken place *over a fixed interval*, such as in a set window of time or space. In contrast, the exponential distribution describes the *distribution of intervals between successive events*. If we take a genomic example, the Poisson distribution would describe the probability of seeing 0,1, 2, 3 … features in a segment of DNA that is L bases long. An exponential distribution would model the distribution of the number of DNA base-pairs between successive observations of the feature.

For more information, see

- Chapter 5 of “The Cartoon Guide to Statistics” by L. Gonick and W. Smith
- Eric Weissteins’s World of Mathematics at mathworld.wolfram.com

- Chapter 11 of “Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids” by R. Durbin *et al.*
- Kumar, S., Subramanian, S. “Mutation rates in mammalian genomes” *Proc. Natl. Acad Sci USA*, 99(2):803-8 (2003).
- Uspensky, J. V. "Approximate Evaluation of Probabilities in Bernoullian Case." Ch. 7 in *Introduction to Mathematical Probability*. New York: McGraw-Hill, pp. 119-138, 1937.

Problems

1) According to manufacturer literature, an experimental assay produces data with a standard deviation approximately equal to 30% of the observed value. We make one measurement using this assay and record a value of 10.0. (a) What is the expected standard deviation for this measurement? (b) From this single point of data is the probability of measuring a value of 11 or greater?
2) (a) Calculate the probability of tossing a coin 10 times and finding three or fewer heads using the binomial distribution. (b) Next, calculate the mean and standard deviation of this binomial distribution. (c) Using the mean and standard deviation from (b), use a Gaussian distribution to calculate the probability of finding 3 or fewer heads. (d) Why do they differ? (e) Would data from more experiments make the binomial and Gaussian predictions more or less similar?
3) When measuring the mRNA expression level of a gene, we find a mean expression level of 1000 units with a standard deviation of 50 units. Using z-scores, find the 90% confidence interval for this measurement.
4) After repeated measurements of the percent change of the weight of a rat liver, we find our data is described by a Gaussian distribution with a mean change of 0% with a standard deviation of 1%. Find the lower limit of the 90% confidence interval using the CDF in equation 4.1.3.
5) In a particular stage of development, neural precursor stem cells are observed to differentiate at a rate of 10 per day per 10,000 cells. Given a population of 20,000 undifferentiated cells, what is the probability of seeing one cell differentiate in 1 hour?

Chapter 5: Statistical Inference

In the previous sections we have introduced tools for statistical descriptions of data, and shown how probability theory can be used to make some inference or prediction based on data. In this chapter will introduce some commonly used statistical tools for inference and show how they can be used in biological problems.

5.1 Confidence Intervals

When validating an experimental or computational result, we are often faced with the question of how sure we are of a particular measurement or value. A statistical method to evaluate this sureness is a **confidence interval**. A confidence interval simply says that we are sure that the true value is somewhere within an interval with a given probability (typically 95% or 99%). The concept of a confidence interval is illustrated in the plot in Figure 5.1.0.

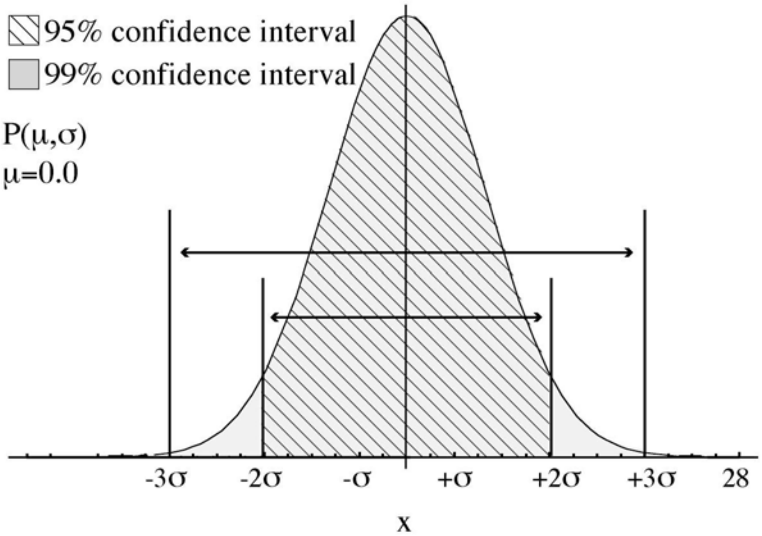


Figure 5.1.0: Illustration of a confidence interval around the mean value of a distribution. Note that the shaded area encompasses the confidence interval region.

To find the range of values within a confidence interval, we integrate the probability distribution function to find a symmetric region that covers the confidence interval. Introductory statistics texts focus primarily on calculating confidence intervals only for Gaussian distributions, however the same approach applies equally well for any distribution of any shape, although the mathematics can be more challenging.

Example 5.1.1: Methylated DNA

We are interested in determining the fraction of genomic DNA that is methylated in a newly discovered organism. As a rapid assay, we sample 100 random segments of DNA from this organism and determine the average fraction of methylated bases to be 0.05 with a standard deviation of 0.01. To within 95% confidence, what is the expected range of future measurements in this system, assuming the distribution is Gaussian? To within 95% confidence, what is the expected error on the mean fraction of methylated DNA in the organism?

Solution

In this example we assume that the underlying distribution is Gaussian, which is generally acceptable for large, aggregate samples. For the distribution of measurements, we are given the mean and standard deviation of 0.05 and 0.01. From this information we can immediately write down the Gaussian distribution

$$P(f)df = \frac{1}{\sqrt{2\pi(0.01)^2}} e^{-(f-0.05)^2/2(0.01)^2} df$$

where f is the fraction of methylated base pairs. We want to find the range of f values that cover 95% of the area under this curve. This range can be found by integrating the probability distribution between the mean minus some unknown value x, to the mean plus some value x, as is shown below.

$$\int_{0.05-x}^{0.05+x} \frac{1}{\sqrt{2\pi(0.01)^2}} e^{-(f-0.05)^2/2(0.01)^2} df = 0.95$$

By solving the resulting expression for x numerically (for example, using Mathematica's FindRoot[ ] function), we find that a value of x=0.02 satisfies this expression. Thus, in future experiments we would expect 95% of our measurements to fall within the interval 0.05±0.02.

To find the expected error of the mean, we need to find the standard error of the mean, as was introduced in chapter 1. To find this error we divide our given standard deviation by the square root of the number of samples

$$\sigma_\mu = \frac{\sigma}{\sqrt{n}} = \frac{0.01}{\sqrt{100}} = 0.001$$

Using this standard deviation of the mean as our standard deviation, we can now perform the same calculation as before

$$\int_{0.05-x}^{0.05+x} \frac{1}{\sqrt{2\pi(0.001)^2}} e^{-(f-0.05)^2/2(0.001)^2} df = 0.95$$

Here we find that x=0.002 satisfies our equation. Thus we would expect that the true mean for the whole genome lies somewhere within the range of 0.05±0.002.

Example 5.1.2: Cell death

In response to radiation treatment of a particular cell line, 1 in 500 cells die each minute. If we start with 2,500 cells, how many will likely die after one minute of irradiation? To within a 90% confidence interval, what is the range of expected number of cells to die after 1 minute of irradiation?

Solution

The data in this problem is given as a frequency of cell death per unit time, which suggests that a Poisson distribution may be a good model for this system. We recall the form of the Poisson distribution from Chapter 3

$$P_\mu(k) = \frac{\mu^k e^{-\mu}}{k!}$$

where ν is the expected number of successes and k is the observed number of successes. From the given data we find μ as

$$\mu = Np = 2500\frac{1}{500} = 5$$

Thus, we expect that on average 5 cells would die in response to this irradiation procedure.

To calculate the error on this prediction, we need to find the boundaries that encompass 90% of the probability density. This requirement can be expressed in the following expression

$$\sum_{i=5-x}^{5+x} \frac{\mu^k e^{-\mu}}{k!} = \sum_{i=5-x}^{5+x} \frac{5^k e^{-5}}{k!} = 0.90$$

This equation can be solved by trial and error with an x value of 3. Thus we are 90% sure that after one minute of irradiation, between 2 and 8 cells will die from this population. This range corresponds to the shaded area shown in Figure 5.2.2.

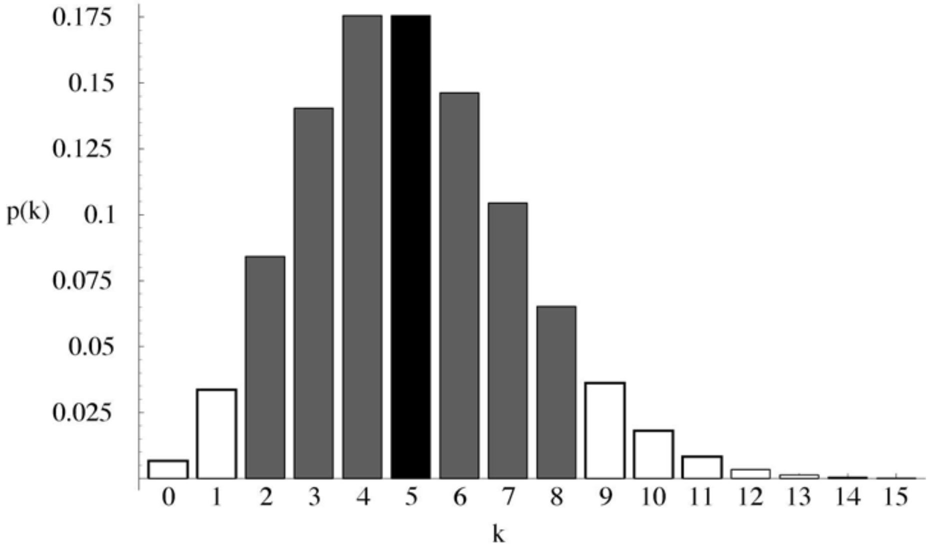


Figure 5.2.2: A Poisson distribution describing cell death for a population of 2,500 cells. Shaded areas represent the 90% confidence region surrounding the mean at k=5.

5.2 P-values

Another common question in statistics is to ask what is the probability that a specific assertion is right or wrong. To make such a clear result, we have to be careful in specifying what we mean by right and wrong. We begin by deciding on a suitable **null hypothesis,** commonly designated as $\mathbf{H_0}$. A null hypothesis is a default case that contradicts our hypothesis. For example, if we compare two means, our null hypothesis may be that the means were actually drawn from the same distribution and the separation is coincidental. Similarly, if we are searching for patterns in a DNA sequence, we might use a random sequence as our null hypothesis. Our hypothesis is called the **alternative hypothesis, or $\mathbf{H_1}$**. For the comparison of two means, the alternative hypothesis may be that the two means are distinct.

Given this background, a **p-value** is defined as the probability that a set of observations would assume a value greater than or equal to the observed value strictly by chance. Thus, large p-values indicate that the result is likely to have happened by chance, while low p-values indicate that random chance played less of a roll. In many experimental contexts, p-values of less than 0.05 or 0.01 are considered statistically significant, while larger values are too likely to have happened by random chance. This cutoff is an arbitrary rule of thumb and may be too strict for some cases and not strict enough for others, depending on the situation.

To calculate a p-value, we need to integrate the distribution generated by the measurements, much like we did earlier to calculate confidence intervals. Examples of how p-values would be calculated are shown in the following examples.

Example 5.2.1: *Geriatric Worms*

The nematode C. elegans has a mean lifetime of 18 days with a standard deviation of 10 days. After performing a silencing RNA (siRNA) screen to knockdown the expression of combinations of genes, you find that worms exposed to a particular set of siRNAs live on average 21 days. In total 37 worms were followed. Do these conditions cause the nematodes to live longer? How significant is this finding?

Solution

To begin, we must choose which distribution to use to model our data. Because the measurement is continuous (lifetime as measured in days), a continuous model is most appropriate. As a first approximation we may also assume that the distribution is symmetric and normally distributed, in which case we can write the distribution down directly

$$P(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

This distribution describes how the normal worm lifespan varies based on known data. However, we are interested in seeing if the *mean* of this distribution is significantly lower than 21 days, so we must use the standard error, or error of the mean to construct this distribution

$$\sigma_{\mu} = \frac{\sigma}{\sqrt{N}} = \frac{10}{\sqrt{37}} = 1.64$$

Thus the distribution describing our alternative hypothesis is

$$P(\mu,\sigma_{\mu}) = \frac{1}{\sqrt{2\pi(1.64)^2}} e^{-(x-18)^2/2(1.64)^2}$$

To find the probability, or p-value, that the siRNA treatment results in longer nematode life, we need to calculate the area under the probability distribution curve that is 21 days or more, as is shown in the figure below. This area can be calculated directly by integrating the distribution, which can be easily done for the Gaussian distribution as was shown in the previous chapter (Eqn. 4.1.4) and is shown below

$$P(x \geq 21) = \int_{21}^{\infty} P(\mu,\sigma_{\mu})dx = \frac{1}{2}\left(Erf\left[\frac{\infty - \mu}{\sigma_{\mu}\sqrt{2}}\right] - Erf\left[\frac{21-\mu}{\sigma_{\mu}\sqrt{2}}\right]\right)$$

$$= \frac{1}{2}\left(1 - Erf\left[\frac{21-18}{(1.64)\sqrt{2}}\right]\right) = 0.034$$

Therefore the lifetime of the nematodes is longer than average, with a p-value of 0.034. This p-value is less than the 0.05 criteria used by many researchers, and therefore may be considered significant.

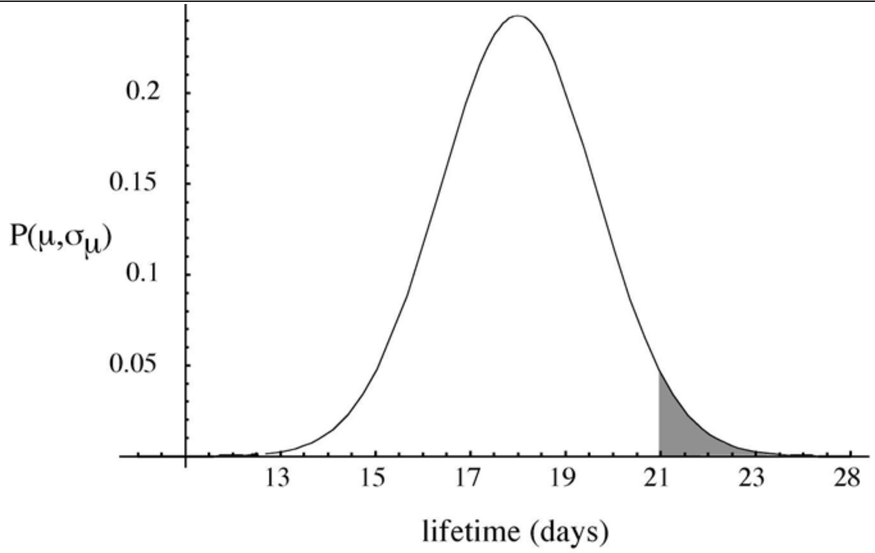


Figure 5.2.1: A graphical example of a p-value calculation. In this case, the p-value represents the probability that the lifetime of the worms is 21 days or more by chance. This value is proportional to the shaded area on the right of the figure.

Note that this example is what is called a **one-tailed test**, meaning that only one side of the distribution was integrated. We only integrate one side because we are only interested in the probability that our observed data means that worms live longer.

Example 5.2.2: Blood glucose levels

Blood glucose levels in humans have a mean of 170 mg/dL and a standard deviation of 40 mg/dL. In a sample 100 of obese men and women, we find a mean glucose level of 180 mg/dL. What is the probability that the obese population has a normal glucose level?

Solution

In this case our null hypothesis is that the obese population has a mean blood glucose level of 170 mg/dL, and the deviation that we saw from this value is a result of random chance. The alternative hypothesis is that the obese population does not have a mean glucose concentration of 170 mg/dL. These two assertions are summarized below

$$H_0 \qquad \mu=170 \text{ mg/dL}$$

$$H_1 \qquad \mu\neq170 \text{ mg/dL}$$

To test this assertion, we will perform a **two-tailed test**, because we are allowing for the possibility that the average blood glucose level could be too high or too low.

As before, we will assume a Gaussian distribution for blood glucose levels. Our main interest in this problem is not the full distribution, but more how the distribution of the mean changes, thus we cast our distribution in terms of the standard error

$$\sigma_\mu = \frac{\sigma}{\sqrt{N}} = \frac{40}{\sqrt{100}} = 4$$

The probability distribution is then

$$P(\mu,\sigma_\mu) = \frac{1}{\sqrt{2\pi\sigma_\mu^2}} e^{-(x-\mu)^2/2\sigma_\mu^2} = \frac{1}{\sqrt{32\pi}} e^{-(x-170)^2/32}$$

To calculate the right tail of the distribution we integrate this distribution from 180 to infinity as is shown below

$$P(x \geq 190) = \int_{190}^{\infty} P(\mu,\sigma_\mu)dx = \frac{1}{2}\left(Erf\left[\frac{\infty - 170}{4\sqrt{2}}\right] - Erf\left[\frac{180-170}{4\sqrt{2}}\right]\right)$$

$$= \frac{1}{2}\left(1 - Erf\left[\frac{10}{4\sqrt{2}}\right]\right) = 0.0062$$

For the left tail, we integrate the symmetric left side from negative infinity to 160. Because the Gaussian distribution is symmetric however, we know that we will get the same result of 0.0062. Therefore the two-tailed probability that the null hypothesis is true is 2(0.0062)= 0.0124, indicating that the obese population is likely significantly different.

Example 5.2.3: Finding transcription factors

You have just developed a novel way to identify transcription factors based on protein sequence alone. To test your software, you construct a database of 100 proteins known to be transcription factors and 900 proteins that are known not to be transcription factors. In this database, your software correctly identifies 64 of the known transcription factors, however it also misidentifies 51 of the non-transcription factor proteins as transcription factors. How well does your software compare to a random sampling of proteins? E.g. What is the probability that a random selection of 64+51=115 proteins from the database would have 64 or more transcription factors?

Solution

This is a common problem in bioinformatics where we make a prediction and then need to quantitatively assess how well that prediction compares to a random prediction, or null hypothesis. In this case, for the null hypothesis we assume that we are drawing at random from a set of 1000 proteins without replacement. In for each draw there are two possible outcomes (transcription factor or non-transcription factor), suggesting that we should describe the probability of each outcome with a hypergeometric distribution (introduced in Chapter 3). The general form of the hypergeometric distribution is

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \quad k = 0,1,2,...,n$$

For our problem, N=1000, K=100, and n=115. We are asked then to calculate the probability that k is 64 or greater. Because there are only 100 transcription factors present in our database, the maximum number that could be selected would be 100. Therefore to find the probability of drawing between 64 and 100 transcription factors we need to calculate the following sum

$$P(64 \le k \le 100) = \sum_{k=64}^{100} \frac{\binom{100}{k}\binom{900}{115-k}}{\binom{1000}{115}} = 3.9 \times 10^{-43}$$

Therefore, although the software produces almost as many false positives as true positives, it is far better than a random selection.

What would be the expected number of transcription factors correctly identified if this were a random selection process?
Answer: 11.5

5.3 Comparing two distributions

Another common situation in computational biology is the comparison of two distributions. In this case, we are given two bodies of data and are asked what the

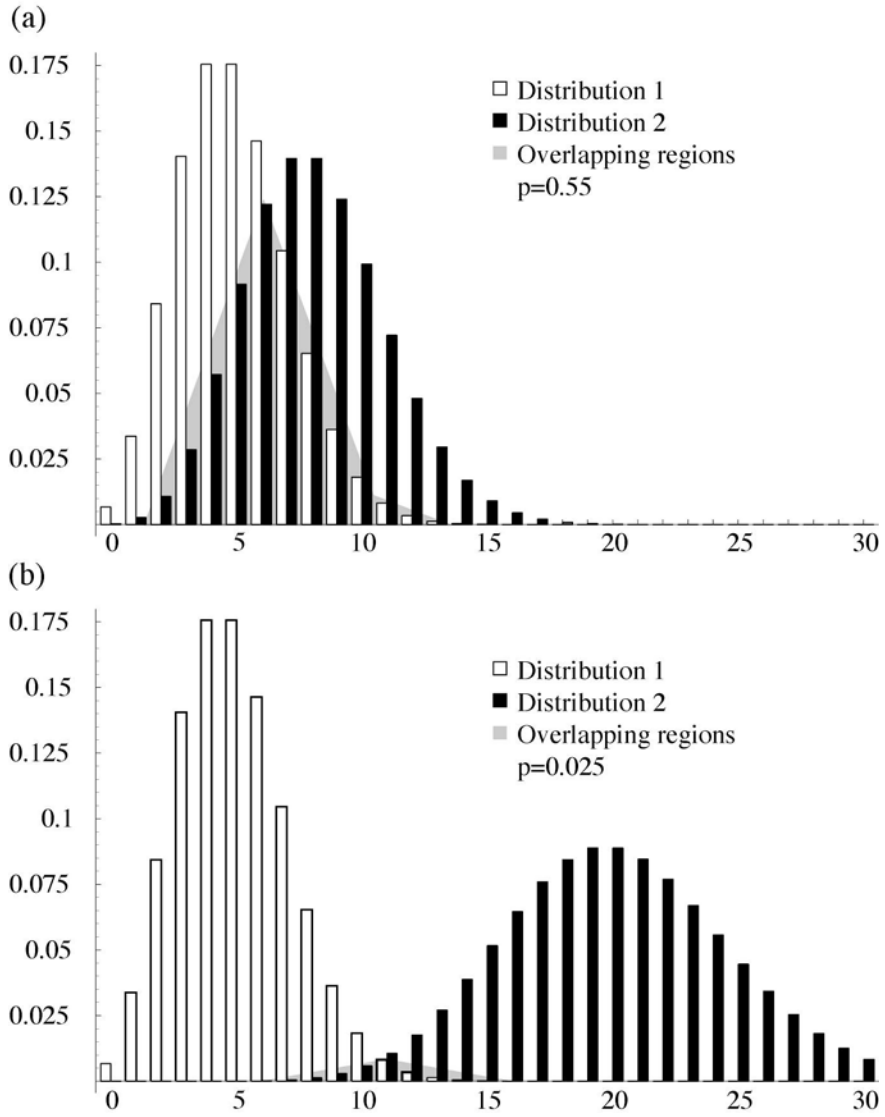


Figure 5.3.1: Graphical representation of the probability of overlap for two cases. The overlapping regions are shaded and are proportional to the common probability measure for the distributions. (a) Here both distributions strongly overlap, thus they share a large common area and have a large probability value. (b) In contrast, in this case, both distributions do not significantly overlap and as such this system has a small probability value.

probability is that both of these sets of data were drawn from the same distribution. This probability is similar to the p-value described in section 5.2, but is more general. Graphically, the probability that two distributions are the same is the overlapping region of two probability distributions, as is shown in Figure 5.3.1.

In general to calculate a probability value from two distributions, we need to integrate or sum the area encompassed by both distributions. This can be stated explicitly using a piece wise function as

$$P(overlap) = \int_{-\infty}^{\infty} \min\begin{cases} p_1(x \mid \theta_1) \\ p_2(x \mid \theta_2) \end{cases} dx \qquad (5.3.1)$$

where p_1 and p_2 are two continuous probability distributions, and θ_1 and θ_2 are the parameters for those distributions. Similarly for two discrete distributions

$$P(overlap) = \sum_{k=0}^{\infty} \min\begin{cases} p_1(k \mid \theta_1) \\ p_2(k \mid \theta_2) \end{cases} \qquad (5.3.2)$$

Essentially, we are making an explicit measure of what area these two distributions have in common, as is illustrated in the **Venn diagrams** in Figure 5.1.2 below.

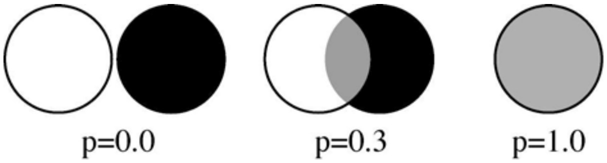


Figure 5.3.2: Venn diagram schematic of how probability values can be calculated for arbitrary distributions. Note that the greater the overlapping area, the larger the probability value.

Example 5.3.1: Protein expression

Imagine that we are running an experiment to see if the presence of a drug causes the up regulation of a particular enzyme within the p450 family. We make 30 measurements of the protein expression level in the absence and presence of the drug, for a total of 60 experiments. From these experiments, we find that the protein expression level with the drug has a mean value of 615 units with a standard deviation of 45 units. In the absence of the drug, the expression level is 575 units with a standard deviation of 60 units. What is the probability that the drug causes the enzyme to up regulate?

Solution

In this example, our null hypothesis is that the measurements are due to random fluctuations in the measurements and our alternative hypothesis is that the distributions reflect distinct states.

To assess the probability of the null hypothesis, we calculate the probability that both means are drawn from the same distribution. We do this first by transforming the standard deviations to standard errors, as is introduced in Chapter 1. Thus for the case with drug

$$\sigma_{\mu,D+} = \frac{\sigma_{D+}}{\sqrt{n}} = \frac{45}{\sqrt{30}} = 8.215$$

and without the drug

$$\sigma_{\mu,D-} = \frac{\sigma_{D-}}{\sqrt{n}} = \frac{60}{\sqrt{30}} = 10.95$$

Using distributions characterized by these standard errors, we then calculate the probability value using the expression in Equation 5.3.1

$$\int_{-\infty}^{\infty} \min\left\{\begin{array}{l} \dfrac{1}{\sqrt{2\pi(\sigma_{u,D+})^2}} e^{-(x-\mu_{D+})^2/2(\sigma_{u,D+})^2} \\ \dfrac{1}{\sqrt{2\pi(\sigma_{u,D-})^2}} e^{-(x-\mu_{D-})^2/2(\sigma_{u,D-})^2} \end{array}\right. dx = \int_{-\infty}^{\infty} \min\left\{\begin{array}{l} \dfrac{1}{\sqrt{2\pi(8.215)^2}} e^{-(x-615)^2/2(8.215)^2} \\ \dfrac{1}{\sqrt{2\pi(10.95)^2}} e^{-(x-575)^2/2(10.95)^2} \end{array}\right. dx$$

This integral can be calculated numerically to find a probability of 0.041. Because this value is less than 0.05, many would consider this result significant.

This finding illustrates a number of key points. First, the two distributions of measurements overlap significantly, making it difficult to distinguish them. However, the large number of replicate measurements reduced the error associated with the mean such that they could be rigorously differentiated. In general, the closer two points are the more data is required to differentiate them.

Second, explicit calculations of the probability that an assertion is true can be computationally expensive, but is tractable using standard software packages.

5.4 *Correlation coefficients*

Sometimes we encounter situations where we are given two sets of data that seem to scale with each other, but not perfectly. In this case, we can assess how well these two variables correlate with each other by using the **correlation coefficient**, or as it is sometimes called **Pearson's correlation**. Most of us have encountered the correlation coefficient as the r^2 fit of data to a line. Values of the correlation coefficient that are near one indicate a strong correlation, while values near zero indicate a poor correlation.

The correlation coefficient is defined as

$$r^2 = \frac{ss_{xy}^2}{ss_{xx} ss_{yy}} \tag{5.4.1}$$

where ss_{xy}, ss_{xx}, and ss_{yy} are defined as

$$\begin{aligned} ss_{xy} &\equiv \sum (x - x)(y - y) \\ ss_{xx} &\equiv \sum (x - x)^2 \\ ss_{yy} &\equiv \sum (y - y)^2 \end{aligned} \tag{5.4.2}$$

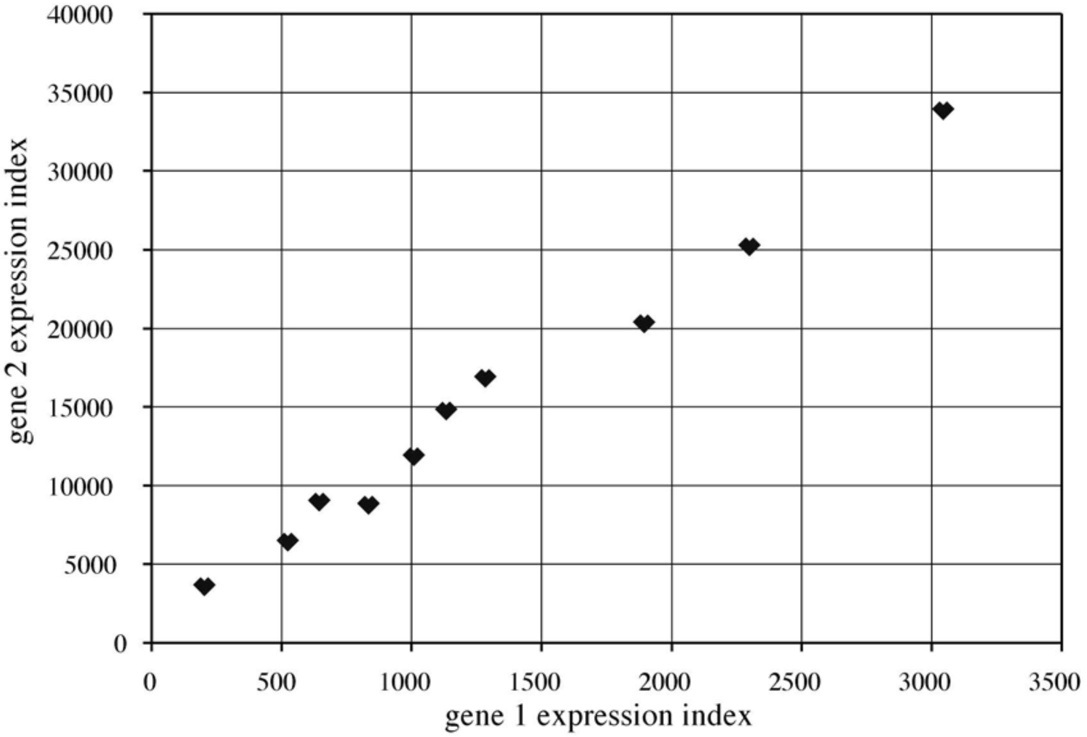
Here, bars over variables indicate an average value. Thus ss_{xx} is the sum of the squared deviation of x from its mean.

Example 5.4.1: Gene clustering

A widely used approach in analyzing expression array data is to cluster genes that behave similarly under a wide variety of physiological conditions. One method to quantify how well two genes track each other is to calculate their correlation coefficient. Thus, imagine that we have gathered the following mRNA expression data for two genes over ten different conditions. The raw data for from these experiments are listed below:

Experiment #	expression index of gene 1	expression index of gene2
1	645	9045
2	1284	16943
3	523	6534
4	3045	33928
5	203	3698
6	1009	11960
7	1132	14850
8	1894	20394
9	834	20394
10	2300	25290

$$ss_{xy} \equiv (645-1286.9)(9045-15153.6)+\ldots+(2300-1286.9)(25290-15153.6)=7.3\times10^{7}$$
$$ss_{xx} \equiv (645-1286.9)^{2}+\ldots+(2300-1286.9)^{2}=7.0\times10^{6}$$
$$ss_{yy} \equiv (9045-15153.6)^{2}+\ldots+(25290-15153.6)^{2}=7.8\times10^{8}$$



For more information, see

- Chapter 7 of “The Cartoon Guide to Statistics” by L. Gonick and W. Smith
- Eric Weissteins’s World of Mathematics at mathworld.wolfram.com
- Chapter 11 of “Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids” by R. Durbin *et al*.
- Chapter 6 in “Elementary Statistics” by M. F. Triola

Problems

1) If we toss a fair coin four times, we expect to see two heads and two tails. If we allow that we might see 1,2, or 3 heads, then what confidence interval are we covering? Said another way, what percent of the time will we see between 1 and 3 heads?
2) The average migration speed of an isolated neutrophil is 20 μm/min. After treatment with an ERK inhibitor, we observe 25 neutrophils and find an average migration speed on 18 μm/min with a standard deviation of 6 μm/min. Is this speed significantly slower than average for neutrophils? What is the p-value?
3) Imagine a simple saw-tooth type distribution defined by

$$p_1(x) = \begin{cases} 4x & \text{if } 0 \le x < 0.5 \\ 4(1-x) & \text{if } 0.5 \le x \le 1.0 \end{cases}$$

   (a) Plot this distribution.
   (b) Show that the total area under this distribution exactly equals one.
   (c) Find the expectation value or mean of this distribution (show your work).
   (d) Calculate the range around the mean that encompasses a 99% confidence region.
4) (a) Plot by hand the following two distributions:

$$p_1(x) = 2x$$
$$p_2(x) = 2(1-x)$$

   where x is defined from zero to one. (b) What is the probability that these two distributions were drawn from the same original distribution? Hint: find the area of the overlapping area.
5) The histograms in Figure 5.3.1 were calculated from Poisson distributions. In panel (a), the expected number of events are μ_1=5 and μ_2=8. Show how you would find the probability that these two distributions are significantly different. Do we really need to sum from 0 to infinity, or can we stop at a smaller value of k? Why or why not? Hint: Set up the problem as shown for a discrete distribution in Equation 5.3.2.

Appendix A: Dirichlet Distribution

The **Dirichlet distribution** is a continuous distribution that can be employed with systems that have many outcomes. One way to think of the Dirichlet distribution is as a continuous counterpart of the multinomial distribution. The Dirichlet distribution is commonly employed as a prior distribution in Bayesian statistics as it yields analytically tractable expressions for an arbitrary number of variables.

The Dirichlet distribution for K possible outcomes has the following form

$$\frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K}\theta_i^{\alpha_i-1} \qquad \text{(A.1.1)}$$

where Γ is the gamma function, which is a continuous version of the factorial function. For any real positive number x, the gamma function is defined as

$$\Gamma(x+1)=x\Gamma(x) \qquad \text{(A.1.2)}$$

The constants α_i in Eqn. A.1.1 specify the particular shape of the distribution and must be real valued and positive. The parameter θ_i is the probability of outcome i. As is the case for probabilities they must be between 0 and 1 and their sum total must equal 1.

For the Dirichlet distribution, the mean is described by the normalized parameters. For example, the mean value of θ_i is

$$\langle\theta_i\rangle=\frac{\alpha_i}{\sum_{j=1}^{K}\alpha_j} \qquad \text{(A.1.3)}$$

Also, the tightness of the distribution is defined by the value of α, with larger values leading to sharper distributions.

In the limiting case of K=2, these distributions reduce down to a subclass of distributions known as the beta distribution which closely resembles the binomial distribution and is shown below

$$\beta(\alpha_1,\alpha_2)=\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\theta_1^{\alpha_1-1}(1-\theta_1)^{\alpha_2-1} \qquad \text{(A.1.4)}$$

A central difference between the beta distribution and the binomial distribution is not their functional forms, but is more how they are used. The binomial distribution is most often used to describe the probability of various configurations of outputs, while the beta distribution is used to describe the probability of the probability parameters, θ. Plots of the beta distribution for various values of α_1 and α_2 are shown in Figure A.1.1.

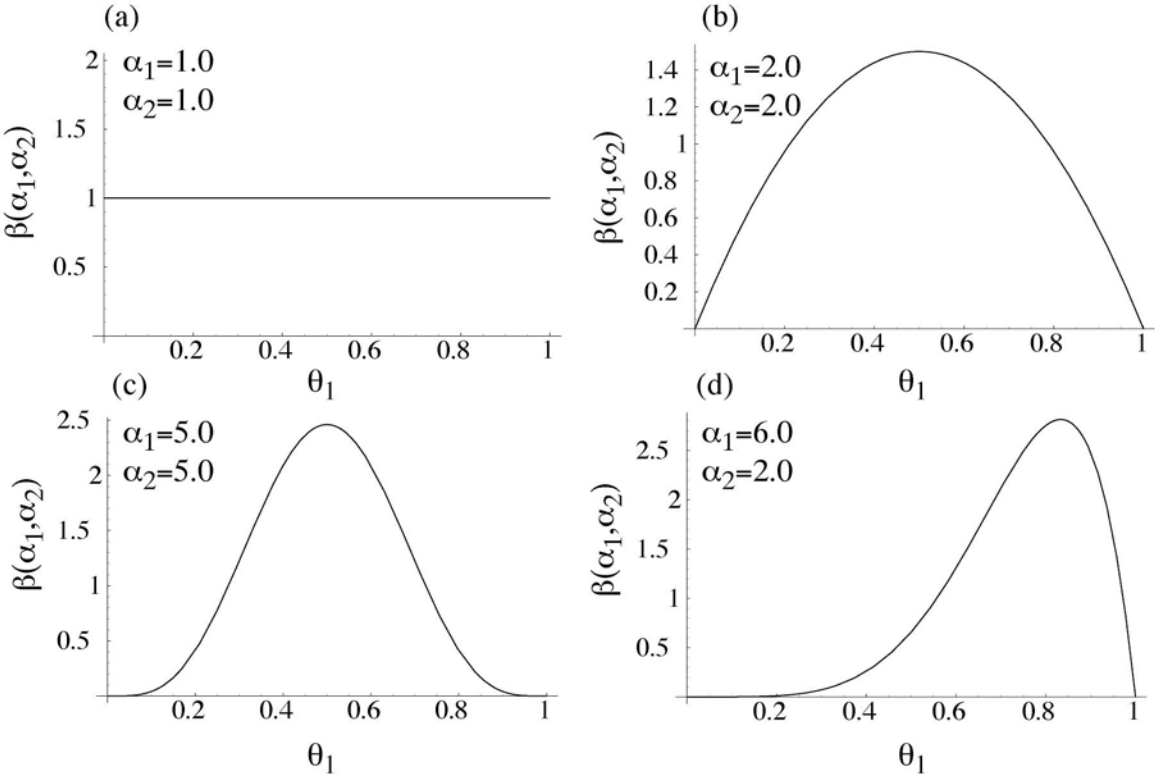


Figure A.1.1: Plots of the beta distribution for various values of α_1 and α_2. Note that equal α values produce a symmetric distribution, while unequal values α produce an asymmetric distribution. In addition, larger α values produce a tighter distribution.

Example A.1.1: Liquid handling robots

We are assessing two high-throughput liquid handling robots. To test the robots we have them dilute a fluorescently labeled solution to 50% of its original concentration. Both robots produce solutions that have an average of 50% less fluorescence with errors described by a beta distribution, but for robot A α_1= α_2=10, while for robot B, α_1=α_2=15. The cost per experiment with robot A is $2.25 while the cost per experiment with robot B is $3.00. If we assume a measurement error of more than ±5% means that the experiment must be repeated, then which robot is more economical?

Solution

Here we replace our probability measurement, θ_1, with the measured fraction of fluorescence in the resulting solution. This measure has the same properties as a probability in that it cannot be negative and all fractions must sum to one, so this is a valid replacement. First, we write out the Beta distribution for this system

$$\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\theta_1^{\alpha_1-1}(1-\theta_1)^{\alpha_2-1}$$

For robot A we have the distribution

$$\frac{\Gamma(20)}{\Gamma(10)\Gamma(10)}\theta_1^{10-1}(1-\theta_1)^{10-1}$$

For robot B we have

$$\frac{\Gamma(30)}{\Gamma(15)\Gamma(15)}\theta_1^{15-1}(1-\theta_1)^{15-1}$$

We would predict that given the higher alpha values of robot B, that robot B would produce more reliable data. We quantify this prediction by integrating these expressions over the fraction, θ_1, from 0.45 to 0.55 to determine what fraction of the experiments would be passable. For robot A we find

$$p_{A\ \ good}=\int_{0.45}^{0.55}\frac{\Gamma(20)}{\Gamma(10)\Gamma(10)}\theta_1^{10-1}(1-\theta_1)^{10-1}d\theta_1=0.342$$

For robot B we find

$$p_{B\ \ good}=\int_{0.45}^{0.55}\frac{\Gamma(30)}{\Gamma(15)\Gamma(15)}\theta_1^{15-1}(1-\theta_1)^{15-1}d\theta_1=0.414$$

To find which robot is more economical, we calculate the average cost of getting at least one passable result. All good results will require at least one experiment, so the initial cost is fixed. A second experiment will only need to be run if the first experiment does not work. Similarly a third experiment will be needed if the second does not work, and on and on. This cost can be represented by a weighted sum of probabilities

$$\langle \text{cost} \rangle = \sum_{i=0}^{\infty} C_o p_f^i = \frac{C_o}{1 - p_f}$$

where C_0 is the cost of each experiment and p_f is the probability of failure. The brackets surrounding cost denote the *expected* cost. Performing this calculation for both machines we find

$$\langle \text{cost}_A \rangle = \$6.58 \big/ \text{measurement}$$

$$\langle \text{cost}_B \rangle = \$7.25 \big/ \text{measurement}$$

Therefore, although machine B produces more accurate data, its additional cost is not justified.

A primary use of the Dirichlet distribution is in solving problems in Bayesian statistics. The reason for this is that Bayesian statistics requires that we state our uncertainty about a parameter value, for example, in the form of a probability distribution. Dirichlet distributions have the advantage that they can describe phenomena with an arbitrary number of outcomes and are relatively easy to integrate with multinomial distributions used in probability calculations. An example of how these distributions work together is shown in Example A.1.2.

Example A.1.2: Hair phenotypes

Imagine that we are interested in the mechanisms that govern hair phenotypes. Literature data has shown that mutations in a key gene responsible for hair development result in three distinct phenotypes: thick hair, thin hair, and no hair. However, no available data indicates the relative frequency of these three phenotypes, so we run our own experiment in mice. Our first litter yields four mice with the thick hair phenotype, and one with the thin hair phenotype and none with the no hair phenotype. Given this information, what is the expected fraction of mice with each phenotype?

Solution

This problem represents a common situation in biological research, as we have some background data plus a small body of experimental data. However, this situation presents a problem for traditional frequentist statistics, due to the small dataset size. For example, in this dataset, we do not see any mice with the no hair phenotype, although we know that they exist. It would be imprudent to assign a zero probability to this phenotype based on such a small sample size, but what other options do we have?

This conflict can be addressed using Bayesian statistics, multinomial distributions, and Dirichlet distributions. To begin, we introduce a number of variables to describe the problem. The unknown probability of each phenotype can be written as θ_1, θ_2, and θ_3. The sum of these probabilities must equal one by definition. Similarly, the counts for each phenotype can be written as n_1, n_2, and n_3. The probability of these counts can then be expressed as the following multinomial distribution

$$p(n_1,n_2,n_3 \mid \theta_1,\theta_2,\theta_3) = \frac{(n_1+n_2+n_3)!}{n_1!n_2!n_3!}\theta_1^{n_1}\theta_2^{n_2}\theta_3^{n_3}$$

However, our goal is not to find the probability of our data, but instead to find the probability of our probability parameters given our data. This rearrangement can be found using Bayes' rule as is shown below

$$p(\theta_1,\theta_2,\theta_3 \mid n_1,n_2,n_3) = \frac{p(n_1,n_2,n_3 \mid \theta_1,\theta_2,\theta_3)p(\theta_1,\theta_2,\theta_3)}{p(n_1,n_2,n_3)}$$

This expression can be simplified by dropping the denominator on the right hand side, as the probability of the data alone does not depend on the parameters and as such is a constant (we will come back to this at the end of the problem). Thus the probability that we are interested in reduces to

$$p(\theta_1,\theta_2,\theta_3 \mid n_1,n_2,n_3) \propto p(n_1,n_2,n_3 \mid \theta_1,\theta_2,\theta_3)p(\theta_1,\theta_2,\theta_3)$$

The first term in this expression is the multinomial distribution generated above. The second term is our prior belief about the parameters. This term can be simply defined as a Dirichlet distribution of the following form

$$p(\theta_1,\theta_2,\theta_3) = \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\theta_3^{\alpha_3-1}$$

One of the most frequent uses of Dirichlet distribution is in defining a prior for a Bayesian calculation involving multinomial distributions. To completely define this distribution, we have to choose values for the hyper parameters, α_1, α_2, and α_3. Because we have no particular information in this case, we can assign what is known as a uniform prior, meaning that any value of θ is equally likely. This uniform prior is assigned by setting $\alpha_1 = \alpha_2 = \alpha_3 = 1$, which generates a flat line as is shown in Figure A.1.1a. By substituting in these α values and merging both expressions we generate the following expression for the posterior probability of our parameters given data

$$p(\theta_1,\theta_2,\theta_3 \mid n_1,n_2,n_3) \propto \left(\frac{(n_1+n_2+n_3)!}{n_1!n_2!n_3!}\theta_1^{n_1}\theta_2^{n_2}\theta_3^{n_3}\right)\left(\frac{\Gamma(1+1+1)}{\Gamma(1)\Gamma(1)\Gamma(1)}\theta_1^{1-1}\theta_2^{1-1}\theta_3^{1-1}\right)$$

which simplifies to

$$p(\theta_1,\theta_2,\theta_3 \mid n_1,n_2,n_3) \propto \frac{(n_1+n_2+n_3)!\Gamma(3)}{n_1!n_2!n_3!\Gamma(1)\Gamma(1)\Gamma(1)}\theta_1^{n_1}\theta_2^{n_2}\theta_3^{n_3}$$

This final form is essentially a Dirichlet distribution also. When we enter our experimental data we obtain the following expression

$$p(\theta_1,\theta_2,\theta_3 \mid 4,1,0) \propto \frac{5!\Gamma(3)}{4!1!0!\Gamma(1)\Gamma(1)\Gamma(1)}\theta_1^4\theta_2^1\theta_3^0$$

This expression describes a probability density function that represents our certainty about each of the probabilitie parameters, θ_1, θ_2, and θ_3. Noting that the last probability is raised to the zero power, we can express this density in terms of only two variables

$$p(\theta_1,\theta_2 \mid 4,1,0) \propto \frac{5!\Gamma(3)}{4!1!0!\Gamma(1)\Gamma(1)\Gamma(1)}\theta_1^4\theta_2^1$$

A density plot of this function is shown in Figure 4.2.2.

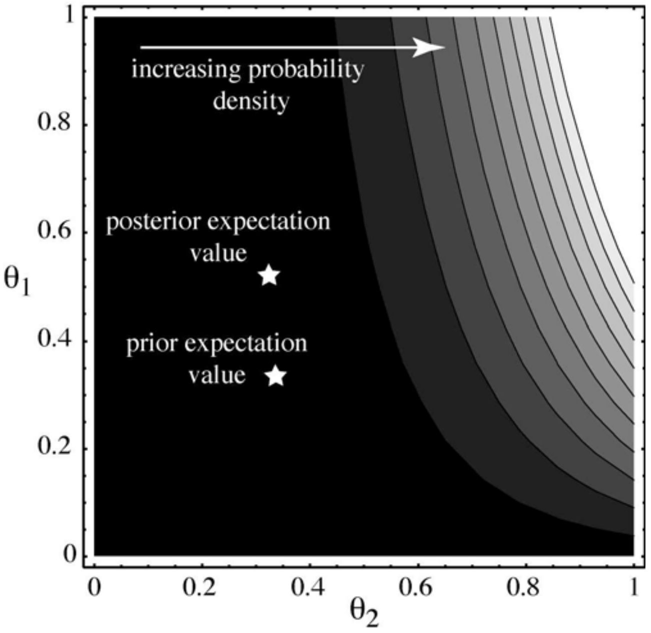


Figure 4.2.2: Plot of the posterior probability density given in Example A.2.2. Note that the probability density increases for both parameters as they approach one, however the posterior expectation value is not at the point of highest probability.

To find the expectation values for θ_1, θ_2, we integrate over all possible values of each variable

$$\langle \theta_1 \rangle = \int_0^1 \theta_1 \frac{5!\Gamma(3)}{4!1!0!\Gamma(1)\Gamma(1)\Gamma(1)} \theta_1^4 \theta_2^1 d\theta_1 = \frac{5\langle \theta_2 \rangle}{3}$$

$$\langle \theta_2 \rangle = \int_0^1 \theta_2 \frac{5!\Gamma(3)}{4!1!0!\Gamma(1)\Gamma(1)\Gamma(1)} \theta_1^4 \theta_2^1 d\theta_2 = \frac{10\langle \theta_1 \rangle^4}{3}$$

$$\langle \theta_3 \rangle = \int_0^1 \theta_3 \frac{5!\Gamma(3)}{4!1!0!\Gamma(1)\Gamma(1)\Gamma(1)} \theta_1^4 \theta_2^1 d\theta_3 = 5\langle \theta_1 \rangle^4 \langle \theta_2 \rangle$$

Solving for each of these values we obtain

$\langle \theta_1 \rangle = 0.564$

$\langle \theta_2 \rangle = 0.338$

$\langle \theta_3 \rangle = 0.172$

We are almost there. If we sum these values we note that they equal 1.074, not 1.00 as we require them. This is because we ignored the normalizing term in our original probability statement, $p(n_1,n_2,n_3)$. We can now correct for this omission by dividing all values by this normalizing constant 1.074 to obtain

$\langle \theta_1 \rangle' = 0.525$

$\langle \theta_2 \rangle' = 0.315$

$\langle \theta_3 \rangle' = 0.160$

Therefore from this small sample size of 5 observations, we are able to rigorously define the probability of various outcomes while accurately representing our uncertainty.

For more information, see

- Eric Weissteins's World of Mathematics at mathworld.wolfram.com
- Chapter 11 of "Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids" by R. Durbin *et al.*

- David Heckerman's "A Tutorial on Learning with Bayesian Networks" in Learning in Graphical Models, edited by Michael I Jordan.
- "Data Analysis: A Byesian Tutorial" by D. S. Silva
- "Bayesian Inference in Statistical Analysis" by G. E. Box and G. C. Tiao.

Problems

1) We observe that a protein has two distinct phosphorylation sites, A and B. We run a single experiment and find that only site A is phosphorylated, while B is not. Based on this single experiment, what can we say about the probability of finding A and B each phosphorylation configuration (e.g. AB, A_pB_p, A_pB, and AB_p, where the p subscript indicates that the site is phosphorylated)? How well would you expect a probabilistic model to describe this system? Why?
Hint: this problem is much like example A.1.2.

Appendix B: Bayesian Networks

Thus far we have introduced tools from probability theory to describe events that are independent and dependent. Although we have only dealt with examples that contain a small number of variables, the theory applies equally well to describe systems with an arbitrary number of variables. For example, a probabilistic model could be used to describe interactions between genes based on expression data containing thousands of variables.

However, when constructing a probability model with many variables, we encounter a number of problems. As an illustration, imagine that we have collected absent/present data for the expression level of 10 different genes over a variety of physiological conditions. Let us call each expression level a variable $x_1,x_2,\ldots,x_{10}$. Based on this information, we can then predict the value of any one variable assuming the values of the other 9 are known e.g.

$$P(x_1 \mid x_2,....,x_{10}) \qquad \text{(B.1.1.)}$$

However, to construct this probability distribution will take a large number of experiments for there are 2^9 (=512) different possible states that x_2-x_{10} could take on. Similarly, we might suspect that the expression level of some genes is more predictive of the expression level x_1, but this analysis does not tell us which relationships are more or less important.

These problems suggest that we might be able to break the probability distribution into independent and dependent parts. For example, we might know that the expression level of gene 2 is independent of the expression level of gene 1. With this information we can then rewrite the probability of finding gene one in any state as:

$$P(x_1 \mid x_3,....,x_{10}) \qquad \text{(B.1.2)}$$

Conceivably, we could have many interactions that are independent of one another, so for each of these we could construct a simplified conditional probability statement. As we simplify the conditional probability statement, we then need less data. For example, in removing the dependence on gene 1 from equation B.1.1 to B.1.2 we reduced the total number of parent states (conditioning states) from 2^9 to 2^8 , a change of 256 states.

A convenient way to organize and represent these groups of conditional probability relationships is as a **Bayesian network**. A Bayesian network is a directed acyclic graph representation of a joint probability statement. Taking the gene expression model introduced above, we could make many simplifications to the network to generate the following probability statement

$$\begin{aligned} P(x_1,x_2,....,x_{10}) = [&P(x_1 \mid x_2,x_3)P(x_2 \mid x_4,x_5,x_6)P(x_3 \mid x_2)P(x_4) \\ &P(x_5)P(x_6 \mid x_7,x_8,x_9)P(x_7 \mid x_{10})P(x_8)P(x_9)P(x_{10})] \end{aligned} \quad \text{(B.1.3)}$$

We can represent this as a Bayesian network by drawing each variable as a node, with arrows connecting variables that are conditionally dependent on each other. A graph of the statement in equation B.1.3 is shown in Figure B.1.1

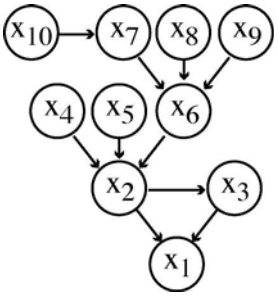


Figure B.1.3: A Bayesian network representation of the conditional dependency statements in equation B.1.3.

If we do not already know the graph structure of a Bayesian network, then it is also possible to derive this connectivity from data alone. In this way, we can rapidly identify strong relationships between variables that may not have been apparent using other tools for data analysis. For example, the status of a single gene could be dependent on the states of three other genes in a complex but reproducible way. Such relationships become immediately apparent from the Bayesian network structure, as shown in Figure B.1.3.

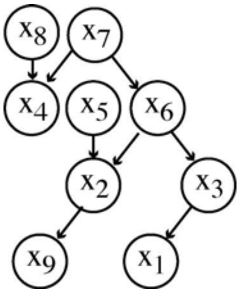
A classic example of a Bayesian network is a family tree that would be used in genetics or genetic counseling. In this tree, each person is a node and arrows connect parents to children. From the conditional probability standpoint, this graph structure makes sense, for if we are trying to predict the probability that a grandchild carries a genetic particular mutation, we know that the most direct predictor is the genetic state of the parents. Assuming no inbreeding, the genetic states of the mother and father should be independent, so there is no arrow directly connecting these two people.

For more information, see

- David Heckerman's "A Tutorial on Learning with Bayesian Networks" in Learning in Graphical Models, edited by Michael I Jordan (1998).
- "Bayesian Inference in Statistical Analysis" by G. E. Box and G. C. Tiao (1992).
- "An introduction to Bayesian networks" by Finn V. Jensen (1996)

Problems

1. Convert the following statements of conditional probabilities into a Bayesian network graph.
    a. $P(A \mid B)P(B \mid C,D,E)P(C \mid D)P(E \mid F)P(F)P(D)$
    b. $P(A \mid B,C)P(B \mid D)P(C \mid D)P(D)$
    c. $P(A \mid B,C)P(B \mid C)P(C)P(D \mid E)P(E)$
2. Convert the following Bayesian network graph into a formal statement of conditional probabilities.



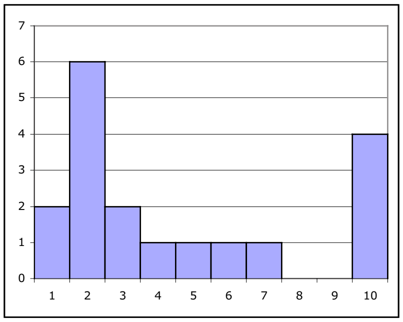
Appendix C: Solutions

Here we list the solutions to some of the exercises at the end of each chapter. Note that in some cases we provide only the numerical answer, as this section is meant as a self-check.

Chapter 1: Common Statistical Terms

1) mean=2.223, standard deviation=0.955

2)



mean=4.55, median=3, mode=2

3) (a) mean=2.9, standard deviation=1.79, standard error=0.566 (b) mean=3.05, standard deviation=1.73, standard error=0.39. (c) standard error would drop because the standard error is the standard deviation (~constant) divided by the square root of the number of experiments. Standard error would drop by a factor of 10 if we move from 20 to 2000 experiments.

Chapter 2: Probability

1) 0.625

2) 0.248

3) (a) 0.37 (b) 0.61

4) (a) 0.9083 (b) 0.9999

5) (a) 0.00083521 (b) 0.45

6) 0.2092

Chapter 3: Discrete Distributions

1) (a) Assuming a binomial distribution, 17.5 tosses out of 50 are expected to land point down (b) with a standard deviation of 3.37 tosses.

2) (a) Geometric distribution, p=0.236839 (b) Hypergeometric distribution for 1 to10 GPCRs, p= 0.236871. (Note this second calculation is non-trivial.)

3) Assuming a Poisson distribution,(a) p(0)= 0.00673795, (b) p(10)= 0.0181328.

4) Using a multinomial distribution, p=0.005467.

Chapter 4: Continuous Distributions

1) (a) Following the manufacturer literature, the standard deviation for the measurement should be (10)(0.30)=3.

(b) Given normally distributed errors around a mean of 10 with a standard deviation of 3, we can use the CDF in equation 4.1.3 to calculate the probability of making a measurement of between negative infinity and 11:

$$CDF_{Gaussian} = P(-\infty < x < 11) = \frac{1}{2}\left(Erf\left[\frac{11-10}{3\sqrt{2}}\right]+1\right) = 0.6305$$

Therefore the probability of measuring a value of 11 or greater is just one minus the above value, or 0.3694.

2) (a) Assuming a fair coin, the probability of finding 0,1,2, or 3 heads is

$$P(Heads \le 3{,}10\ \ tosses) = \sum_{k=0}^{3} \frac{10!}{k!(10-k)!}\left(\frac{1}{2}\right)^{k}\left(\frac{1}{2}\right)^{10-k} =$$

$$= \left[1+10+\frac{(10)(9)}{2}+\frac{(10)(9)(8)}{(2)(3)}\right]\left(\frac{1}{2}\right)^{10} = 0.1719$$

(b) We can calculate the mean and standard deviation of a binomial distribution directly using the equations 3.1.6:

$$\mu = Lf = (10)(0.5) = 5\ \ heads$$

$$\sigma = \sqrt{Lf(1-f)} = \sqrt{10(0.5)(0.5)} = 1.581\ \ heads$$

(c) We can use the CDF for a Gaussian distribution in equation 4.1.3 to calculate this quantity

$$CDF_{Gaussian} = P(-\infty < x < 3) = \frac{1}{2}\left(Erf\left[\frac{3-5}{1.581\sqrt{2}}\right]+1\right) = 0.1029$$

(d) They differ because the Gaussian distribution assumes that the viable is continuous, while the binomial distribution a discrete set of events. Because coin tosses are discrete, we expect that that the binomial distribution is a more accurate model.

(e) More experiments should reduce the difference between the Gaussian and binomial predictions. In the limit of an infinite number of experiments, the Gaussian and binomial distributions should converge.

3) For the 90% confidence interval, we use a z–score of 1.645, mean of 1000, and standard deviation of 50. Plugging into equation 4.2.1 and solving for x we obtain an upper limit of 1082.25. A lower limit is found by using a z–score of –1.645, yielding an x of 917.75. The 90% confidence interval is therefore 917.75 to 1082.25 units.

4) We start with the CDF in 4.1.3 and plug in our mean and standard deviation to obtain

$$\frac{1}{2}\left(Erf\left[\frac{k}{\sqrt{2}}\right]+1\right)=\frac{1-0.90}{2}=0.05$$

Our lower limit is the value k which we can solve numerically (using a calculator, Excel, or Mathematica for example) to yield k=-1.64485. Note that this is identical to the z–score that describes a 90% confidence interval, as our original distribution was a standard normal distribution.

5) Change units and scale,

$$\lambda=\frac{10\ \ events}{(day)(10{,}000\ \ cells)}\times\frac{1\ \ day}{24\ \ hours}\times(20{,}000\ \ cells)=\frac{0.833\ \ events}{hour}$$

Continuous variables and waiting times are exponentially distributed, so the probability of seeing a single cell differentiate in one or fewer hours is described by the exponential CDF and yields a probability of 0.565.

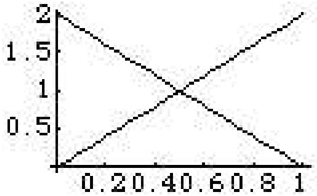
Chapter 5: Statistical Inference

1) (4!/3!1!+4!/2!2!+4!/3!1!)(1/2)^4=14/16=0.875 or 87.5%

2) The migration speed is slower. The p-value is determined by using the Gaussian CDF in equation 4.1.3 and using a standard error (versus standard distribution) of 0.15. The p-value is the probability that this measure actually reflects a speed equal to or greater than the average speed of 20 um/s. p-value=0.04779, which is below the 0.05 threshold and would often be considered significant.

3) (a) plot (b) By geometry or by calculus $\int_0^{0.5} 4x dx + \int_{0.5}^1 4(1-x) dx = 1$. (c) Expectation is defined as the average of the function. Because this distribution is symmetric around 0.5, the expectation is 0.5. This can also be shown by calculus: $\int_0^{0.5} (x) 4x dx + \int_{0.5}^1 (x) 4(1-x) dx = 0.5$ (d) Again, because the distribution is symmetric, we can do this calculation for half of the distribution. This could be done with calculus, or just algebra. Using algebra, we know that the total area under the left half of the distribution is (height*width)/2=((4*0.5)*0.5)/2=0.5. At the 99% confidence interval, we would leave an area on the left hand side of (1.0-0.99)*(0.5)=0.005. A similar triangle with an area of 0.005 can be described as (height*width)/2=area thus (4x*x)/2=0.005, thus x=0.05 at the lower bound of the 99% confidence interval. The upper bound is x=1.0-0.05=0.95.

4) (a) plot



(b) by integration or geometry, overlapping area=0.5.

5) Overlap area is

$$overlap = \sum_{k=0}^{\infty} \min \begin{cases} \dfrac{5^k e^{-5}}{k!} \\ \dfrac{8^k e^{-8}}{k!} \end{cases}$$

We can stop at a maximum k value of less than infinity, as both distributions have probabilities of near zero at k values of ~50.