INTRODUCTION TO PROBABILITY AND STATISTICS

by

Tim Swartz

2014 revision by M. Lesperance, University of Victoria

Department of Statistics and Actuarial Science
Simon Fraser University

Contents

Chapter 1

Introduction

These course notes have been developed to assist you in the understanding of a first course in probability and mathematical statistics. Such courses are notoriously difficult and have gained reputations as "killer" courses on many university campuses. Despite the difficulty, such courses are amongst the most useful courses that a student ever takes. Having taken an introductory course, a student is better prepared to read scientific papers, assess experimental methodologies, interpret graphical displays, comprehend opinion polls, etc. Overall, the student becomes a better critical thinker in this age of information.

You may be wondering why these course notes were written. The primary reason is that some SFU faculty members were outraged to find that the previous textbook was selling for $238 plus tax at the SFU Bookstore. Furthermore, I think that you will find these relatively concise course notes less overwhelming than an 800-page textbook. If you happen to find a mistake or typo in the course notes, please email (mlespera@uvic.ca).

Many of the difficulties that students experience with introductory courses in probability and statistics are two-fold. First, there are many new concepts that are introduced. Second, these concepts tend to build on one another. Although the course uses some mathematics, and in particular some calculus, the focus is on statistics.

To succeed in this course, the most important thing is to avoid falling behind. It is almost impossible to cram in this course and do well.

Every chapter concludes with a set of exercises and an Appendix contains many additional exercises. It is *really* important that you attempt many of these problems to help you grasp the course material. The exercises that are marked with an asterisk have solutions that are provided in Appendix A. Unlike some textbooks, the solutions are presented in detail, to help you think through problems in a structured way. And there is a wrong way and a right way to attempt problems. The wrong way is to read the question and then immediately read the solution. When you do this, there is a tendency to develop a false reality that you understand things well. The right way to approach a problem is to give it a solid effort on your own. After giving a problem a good attempt, then read the solution. This is the best way to learn. When you are really stuck, it is usually more efficient to seek advice from your instructor or a teaching assistant. Hopefully, they will get you back on track quickly. This is better than muddling on your own for hours. If you are looking for additional practice problems, just go to your nearest library. There is no shortage of textbooks on introductory probability and statistics.

1.1 Course Overview

The course is divided into three components with each component subdivided into various chapters. A chapter corresponds to roughly two weeks of material.

The first component of the course is **descriptive statistics**. In a nutshell, descriptive statistics concerns the presentation of data (either numerical or graphical) in a way that makes it easier to digest the data. Sometimes well-presented data reveals patterns that are not obvious when looking at the data in raw form. Chapter 2 addresses descriptive statistics.

The second component of the course is **probability theory** which is addressed in Chapters 3, 4 and 5. In the real world, systems are rarely deterministic. That is, they do not always reproduce themselves. For example, the same amount of fertilizer and water given to two different seeds will produce two plants of different heights. In a situation like this, it is useful to describe the variation in the heights by a stochastic or probabilistic mechanism. Even when deterministic systems exist, it may be convenient to model them probabilistically. For example, if we knew all of the details about the inputs to tossing a coin (e.g. spin rates, height of spin, surface characteristics, etc.), we might be able to predict the outcome of heads or tails in this complex system. However, it is more convenient to simply say that the probability of a head is 1/2. In this component of the course, we gain a deeper understanding of the intuitive notion of probability.

The third and final component of the course is **statistical inference** which is addressed in Chapters 6 and 7. Inference is what most people think about when statistics is mentioned. Statistical inference is the process of learning about a population based on a sample from the population. At first thought, this may seem like an impossibility as the sample is a subset of the population, and the sampled units could be quite different from the non-sampled units. Therefore, we must be careful in making statements about the population, and this is where the probability theory from Chapters 3, 4 and 5 is relevant. Our statements regarding the population based on a sample are not made with 100% certainty. Instead they are qualified using probability. We consider statistical inference in the context of estimation and hypothesis testing for various probabilistic models.

1.2 Examples of Statistical Practice

To gain some appreciation of the relevance of statistical practice in the real world, let's quickly describe some problems where statistical theory has made an impact. The common thread amongst all of the problems is that they involve data. Note that none of the questions posed are answered as the discussion is merely intended to initiate “statistical thinking”.

1.2.1 Sample Surveys

Prior to elections, polling agencies sample the electorate and deliver statements such as “32.3% of the voters in British Columbia ± 4.2% support the Liberal Party 19 times out of 20”. There are a lot of statistical issues involved in a problem of this sort. For example, how many people should be sampled? How should we sample people? Not everyone has a phone and not everyone has a fixed address. How do we interpret the statement? In particular, what does the “19 times out of 20” mean? Although statements like these are common, my experience is that few people really understand what is being reported.

1.2.2 Business

I am sure that you have been on a flight where the airplane is overbooked. Naturally, the reason why this is done is that the airline does not expect everyone to show up, and profits are maximized by filling all the seats. On the other hand, passengers who sacrifice their seat on a full flight must be compensated by the airline. Therefore an airline must predict how many people will not show up on a given flight. The problem of **prediction** is a statistical problem, and depends on many factors including booking patterns on previous flights, the destination and time of the flight and the capacity of the airplane.

1.2.3 Agriculture

Farmers are always interested in increasing the yield of their crops. The relationship between yield and sunshine, fertilizer, temperature, soil and pesticides are no doubt relevant. Note that some of the factors can be controlled while others can not be controlled. How do we specify the relationship between yield and the factors? How do you experiment with different conditions when you can only plant one crop per year? When does the cost of a factor negate its benefits?

1.2.4 Medicine

Pharmaceutical companies run clinical trials to assess the efficacy of new drugs. There are endless statistical issues that come into play. Consider one of the simplest scenarios. Suppose that 100 patients are given drug A and another 100 patients are given drug B. If 53 get better using drug A and 46 get better using drug B, you might be tempted to say that drug A is better than drug B. However, is the difference $53-46 = 7$ a “real” difference or the sort of difference that might be expected due to the variation in the patients? How do you account for misclassification where some of the patients are actually sick but are recorded as being healthy, and vice versa? Apart from assessing the data, how do we incorporate our prior opinions concerning the drugs?

1.2.5 Sport

Although not huge in Canada, the game of cricket is a popular sport in many countries including Australia, England, India and Sri Lanka. As in baseball, each team in cricket has a batting order, and the question arises concerning the optimal batting order (i.e. the order which produces the most runs). With 11 batsmen, there are more than 40 million possible batting orders and it is therefore impossible to test each batting order in an actual match. This problem suggests simulation, and we note that the use of computers plays a major role in the practice of statistics.

1.3 About the Author

Tim Swartz received his B.Math. (1982) from the University of Waterloo and his M.Sc. (1983) and Ph.D. (1986) from the University of Toronto. He is a Full Professor in the Department of Statistics and Actuarial Science at SFU and has been at SFU since 1986. Together with Michael Evans, Swartz has written a book *Approximating Integrals via Monte Carlo and Deterministic Methods* (Oxford University Press, 2000) and has published over 60 research articles. His interests include statistical computing, Bayesian methods and applications, statistical inference and statistics in sport. In his spare time, Swartz likes to play and watch almost every sport imaginable. He also likes to goof around with his daughters and, to a lesser extent, his wife.

1.4 Some Thanks

I extend thanks to my daughter Philippa Swartz for her careful editing of an early version of the course notes. I am also appreciative of technical advice and for the many typos discovered by my colleague, Professor Joan Hu.

1.5 Final Words

Remember, the course is not easy. You need to work at it regularly. Good luck to you. $\ddot{\smile}$

Chapter 2

Descriptive Statistics

Presented with data (either a **sample** or a **population**), it may be difficult to see features in the data simply by staring at the numbers. This may be due to the sheer scope of the data (e.g. thousands of numbers) or the structure of the data (e.g. multiple observations recorded on different subjects).

The goal of descriptive statistics is to summarize data in a convenient form so that features of the data may be more easily revealed. The summarization may be either **numerical** or **graphical**, and the type of summarization that is done depends on the type of data and our intended goals.

There are a number of well-established descriptive statistics that are widely applicable. We look at some of these statistics in this chapter, and we maintain a critical eye in interpreting these statistics.

We also mention that special datasets may call for descriptive statistics that are specially designed for the particular application. For example, basketball players take shots during a basketball game, and the shots may either go in the basket or not. A convenient graphical descriptive statistic for this application is a diagram of the basketball court with circles placed at the locations where shots are taken by the player. Circles corresponding to shots that have gone in the basket are filled in. By looking at such a diagram, it is possible to see the locations where the player is experiencing success and the locations where the player is struggling. The diagram provides information for improvement, and for judiciously deciding when to take shots.

There is nothing too difficult about descriptive statistics and there is nothing to prevent you from developing your own descriptive statistics.

2.1 Dotplots

A **dotplot** is a graphical descriptive statistic that is applicable when we have **univariate** data. Univariate data consists of single measurements on subjects, and we typically denote univariate data with symbols such as:

$$x_1, x_2, \ldots, x_n.$$

Here we have a dataset of size n where x_1 is the first observation, x_2 is the second observation, and we continue until the last observation, x_n. For example, the data may be the weights of male students where x_i is the weight of the i-th student in kilograms, $i = 1, \ldots, n$. The dotplot places each data value on a horizontal scale giving a plot as shown in Figure 2.1. Some characteristics that we might be able to detect from the dotplot include:

- **outliers** (i.e. extreme observations; large or small)
- **centrality** (i.e. values in the middle portion of the dotplot)
- **dispersion** (i.e. spread or variation in the data)

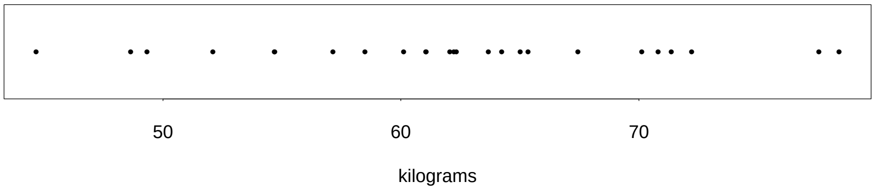

Figure 2.1: A dotplot of the weights in kilograms of 25 male students.

Now, frankly speaking, dotplots are not the most common graphical displays for univariate data. Their role is generally confined to initial investigations where you are able to quickly produce dotplots with the use of statistical software. Certainly, when the dataset is large (say $n \geq 50$), a more appropriate display is a histogram which is studied in Section 2.2. Finally, a general principle in building graphical displays is that they need to be labelled adequately. The sole purpose of graphs is to gain an understanding of the data, and the reader should not be puzzled about the details of a graph. Observe in Figure 2.1 that both a caption and a label on the horizontal axis are provided.

2.2 Histograms

Like the dotplot, the **histogram** is also a graphical descriptive statistic that is applicable when we have univariate data. In practice, histograms are typically constructed when the size of the dataset n is fairly large, and histograms are always produced with statistical software. However, for illustration, lets construct a histogram based on the following small dataset corresponding to the reaction time in seconds for $n = 14$ subjects:

4.7 5.5 7.9 6.3 6.4 6.7 5.4 5.9 5.8 8.4 7.0 6.1 6.5 5.9.

First, consecutive intervals of equal length are chosen to encompass all of the data. Frequencies (and perhaps relative frequencies) are then calculated as shown in Table 2.1. Observe that observations with the same value (e.g. 5.9) are treated as separate observations.

Intervals	Frequency	Relative Frequency
$[4.5, 5.5)$	2	$2/14 = 0.14$
$[5.5, 6.5)$	7	$7/14 = 0.50$
$[6.5, 7.5)$	3	$3/14 = 0.21$
$[7.5, 8.5)$	2	$2/14 = 0.14$

Table 2.1: Intervals, frequencies and relative frequencies for the reaction time data.

Note that the interval notation $[a, b)$ denotes all values greater than or equal to a, and strictly less than b. Similarly, $(a, b]$ denotes all values strictly greater than a, and less than or equal to b. Note also that the relative frequencies may not add exactly to 1.0 due to rounding. Based on Table 2.1, the **frequency histogram** in Figure 2.2 is constructed with frequency plotted on the vertical axis. Alternatively, a **relative frequency histogram** could have been constructed by replacing frequency on the vertical axis with relative frequency. Naturally, the relative frequency histogram has the same shape as the frequency histogram. Again, we are careful to annotate our graphs. For histograms, this means labelling the axes and providing either a title or a caption.

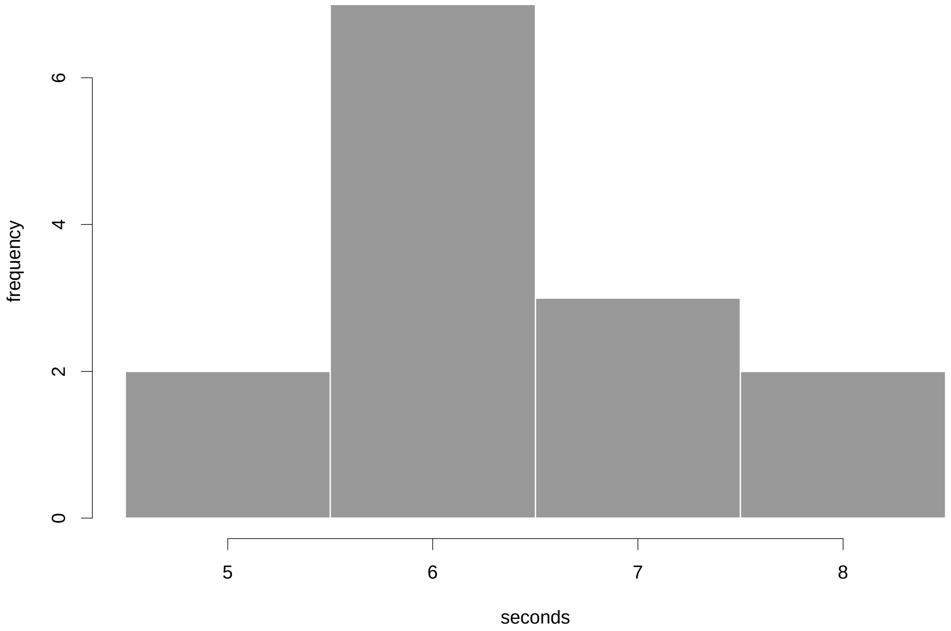

Figure 2.2: A frequency histogram for the reaction time data.

You may now be saying to yourself "big deal", and you are right. Histograms are commonplace, and are easy to interpret. However, lets look at a few aspects of histograms a little more deeply. First, what can be detected by studying a histogram? In addition to learning about outliers, centrality and dispersion, we may be able to comment on **modality**. A histogram with two distinct humps is referred to as **bimodal** and may reflect two underlying groups. For example, when plotting a large number of weights of students, a bimodal histogram may arise and be due to the presence of both males and females in the dataset. In studying a histogram, we may also be able to comment on **skewness** and **symmetry**. These are intuitive concepts where right-skewed histograms (i.e. those with a long right tail) are prevalent in many investigations. For example, think about the histogram produced based on the incomes of residents of Vancouver.

The next issue concerns the number of intervals chosen in producing a histogram. Although computer software typically makes a default choice for you, it is instructive to look at the extreme cases. Suppose that we choose one huge interval. In this case, the histogram has the shape of a rectangle since all observations fall in the interval. Very little can be learned by staring at the rectangle, and essentially, too much summarization is taking place. All of the n observations have been reduced to a single entity. At the other extreme, suppose that we have an enormous number of intervals of equal length and the lengths are all extremely short. Ignoring the possibility of observations with the same value, the intervals have frequencies of either 0 or 1. In this case, the histogram looks very much like the dotplot of Section 2.1, and we suggest that not enough summarization has been done. In the histogram, the n observations appear as n entities. Consequently, it is not easy to detect characteristics of the data. The upshot of this is that some compromise needs to be reached in terms of summarization (i.e. the number of intervals chosen). A rule of thumb (and I am not convinced that it is a good rule) is to choose the number of intervals roughly equal to $\sqrt{n}$ where n is the number of observations.

Up until this point, we have insisted that the intervals used in constructing histograms are of equal length. There is a reason for this. For example, consider the data $0, 1, \ldots, 10$ measured in feet with the intervals $[0, 2)$, $[2, 4)$, $[4, 6)$, $[6, 8)$ and $[8, 108)$. The first four intervals each have a frequency count of 2 and the last huge interval has a frequency count of 3. The frequency histogram appears in Figure 2.3. Our visual perception of the histogram is that there is a lot happening in the region $[8, 108)$ when in fact there are only three observations. When we look at two-dimensional presentations, we tend to perceive size relative to the area of the object. To obtain the correct visual interpretation when the intervals are not of equal length, we should instead plot relative frequency divided by interval length on the vertical axis. If we do so, then the area of each rectangle is its relative frequency and the total area of all rectangles is 1.0. In the case of the fictitious data $0, 1, \ldots, 10$, the values that we would instead plot on the vertical axis are 0.0909, 0.0909, 0.0909, 0.0909 and 0.0027 respectively. However, we stress again that it is better to stick with intervals of equal length.

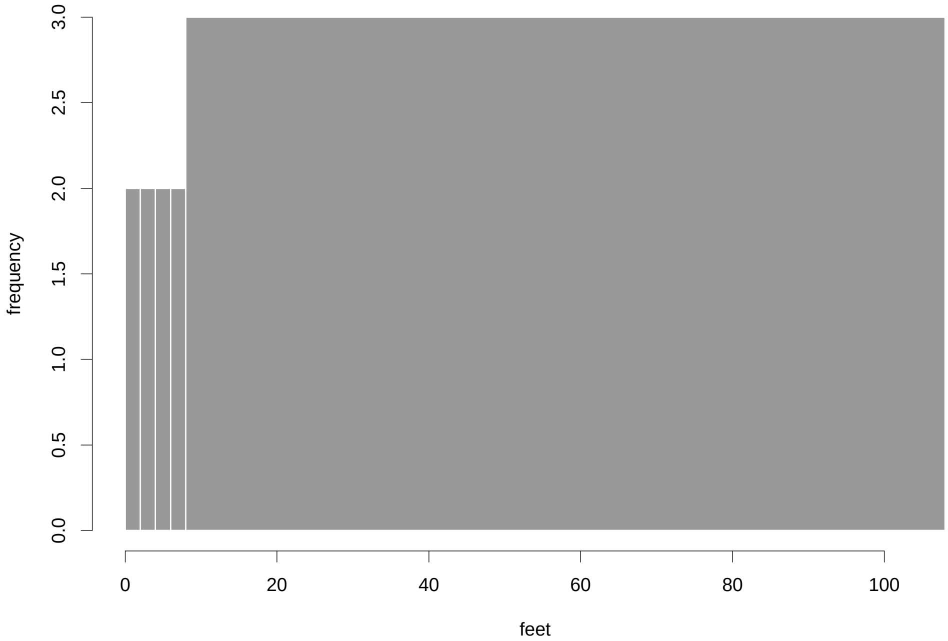

Figure 2.3: A frequency histogram for the fictitious data.

Let's now consider the effect of scaling the vertical axis of the histogram. We should be aware that if the vertical scale does not begin at 0.0, it is easy to obtain an incorrect visual interpretation. For example, construct for yourself a frequency histogram with frequencies 101, 102, 103 and 104 corresponding to four equally spaced intervals. If the vertical axis begins at 100.0, then it may appear that the frequencies are increasing dramatically over the four intervals. Conversely, if the scale begins at 0.0, then it is easier to interpret the increase more appropriately with a rate of roughly 1.0% per interval.

We also mention that it is not necessary to have numerical values on the horizontal axis. It is possible that **categorical** variables can be used on the horizontal axis. For example, we might have the years $1998, \ldots, 2006$ or the colours red, green, blue and yellow. With categorical variables, the histogram is sometimes referred to as a **barplot** or a **bar graph**.

As a final note concerning histograms, you may have noticed that many introductory statistics textbooks discuss **stem and leaf plots**. Stem and leaf plots are similar to histograms and have some advantages over histograms. However, I think it is fair to say that they have not caught on. I don't think that I have ever seen a stem and leaf plot except in a statistics textbook. You can ignore stem and leaf plots in this course.

2.3 Univariate Numerical Statistics

You have probably heard the expression "a picture is worth a thousand words". I share a similar feeling in that I prefer graphical descriptive statistics to numerical descriptive statistics. However, the world likes numbers and we now present a few numerical descriptive statistics that are appropriate for univariate data.

2.3.1 Measures of Location

When describing the centrality or middleness of univariate data $x_1, \ldots, x_n$, there are two statistics that are widely used. The first of these is the **sample mean** or **mean** which is defined by

$$\bar{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \;=\; (x_1 + x_2 + \cdots + x_n)/n.$$

As with a lot of statistical notation, sometimes the bounds of summation are understood, and we instead write $\bar{x} = \sum x_i/n$. The mean gives equal weight to each of the n observations.

The second statistic is the **sample median** or **median** which is defined by

$$\boxed{\tilde{x} \;=\; \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{when } n \text{ is odd} \\ \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)/2 & \text{when } n \text{ is even} \end{cases}}$$

where $x_{(i)}$ is referred to as the i-th **order statistic** and is the i-th data value when the data are listed in the order from smallest to largest.

An important thing to remember when reporting statistics is to include the units. For example, if the data are measured in kilometres, then the mean and median should be expressed in the units of kilometres.

To get a sense of the difference between the mean and median, consider the data:

$$1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9.$$

In this case, $n = 9$ and there is no difference in the mean and median as $\bar{x} = \tilde{x} = 5.0$. However, suppose that the last data value 9 is changed to 54. In this case, the mean becomes $\bar{x} = 10.0$ but the median $\tilde{x} = 5.0$ remains unchanged. Therefore, we observe that the median is more **robust** to outliers in the sense that extreme observations do not affect the median as greatly as the mean.

As a more relevant example involving the use of means and medians, consider the 404 homes listed for sale in Coquitlam in February 2011. It turns out that the mean list price was \$890,000 and the median list price was \$729,000. When the data form a histogram that is skewed to the right, we expect the mean to exceed the median. And when the data form a histogram that is skewed to the left, we expect the median to exceed the mean. When the data are nearly symmetric, the mean and median have comparable values.

In the case of the prices of houses, there are some very expensive houses that cause the mean price to exceed the median price.

To finish off the section on the mean and median, here is a little problem for you to think about. Suppose that there is some data that has a frequency histogram with intervals $[20, 40)$, $[40, 60)$, $[60, 80)$ and $[80, 100)$ with frequencies 8, 6, 10 and 7 respectively. Try to approximate $\bar{x}$ and $\tilde{x}$ for the data.

2.3.2 Measures of Dispersion

Let's begin by looking at two simplistic datasets:

Dataset 1	-2.0	-1.0	0.0	1.0	2.0
Dataset 2	-300.0	-100.0	0.0	100.0	300.0

where the measurements are in grams. Both datasets have the same mean and median, $\bar{x} = \tilde{x} = 0.0$ grams. Therefore, if we summarize the data only by using measures of centrality, the two datasets appear similar. However, they are obviously very different, and the big difference is that Dataset 2 is more spread out than Dataset 1.

To capture the notion of spread or dispersion, there are three commonly used statistics. The first one is the **range** R and is defined as the difference between the largest and smallest values. For the two datasets above, verify that the range is 4.0 grams and 600.0 grams respectively. Like the median, the range is said to be **inefficient** since its calculation relies on only two values from the entire dataset. As a measure of spread, the range is easy to calculate and was popular in the pre-computer days. Today it is taught more as a historical curiousity.

A better measure of spread is the **sample variance** or **variance** which is defined as

$$\boxed{s^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.} \tag{2.1}$$

There are several things to observe about the variance. First, it is non-negative since it consists of the sum of squared quantities. Its minimum value is 0.0 and this only occurs when all of the data values are the same (i.e. $x_1 = x_2 = \cdots = x_n$). To appreciate that the variance is a measure of spread, note that the quantity $(x_i - \bar{x})^2$ represents the squared distance of x_i from the middle of the data, and the distance from the middle is a measure of spread. Also, it may appear strange that the denominator $n-1$ is chosen rather than n. However, we leave the explanation until Chapter 6, and comment that if the dataset is substantial (i.e. n is large), then the difference between using $n-1$ and n is minor. Finally, note that the variance is measured in squared units since each of the data values are squared. Check for yourself that the variance in Dataset 1 is $s^2 = 2.5$ grams^2.

Although s^2 as defined in (2.1) gives us an understanding as to why variance is a measure of spread, it is more convenient to calculate variance via the formula

$$\boxed{s^2 \;=\; \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right).} \tag{2.2}$$

Verify for yourself that (2.1) and (2.2) are equivalent. Do this by squaring and expanding the contents of the parentheses in (2.1). It is good practice with summation notation. The expression (2.1) is known as a two-pass formula as it requires two loops of computer code to evaluate. The first loop adds the x_i to produce $\bar{x}$, and the second loop adds the squared differences $(x_i - \bar{x})^2$. On the other hand, (2.2) is a one-pass formula.

The third measure of spread is the **sample standard deviation** or **standard deviation** s which is simply the square root of the variance. It is reported in the same units as the original data. An interesting rule of thumb known as the **3-sigma rule** is that roughly 99% of the data $x_1, x_2, \ldots, x_n$ are contained in the interval $(\bar{x} - 3s, \bar{x} + 3s)$. People tend to prefer reporting the standard deviation instead of the variance.

2.4 Boxplots

A **boxplot** is another graphical descriptive statistic that is useful in describing univariate data. However, the boxplot is most appropriately used when the data are divided into groups. For example, suppose that we have univariate data on children, women and men, and one of our goals is to compare the three groups. Rather than construct three separate histograms, a boxplot allows us to compare the three groups in a single graphical display. There are slight variations in the construction of boxplots and the subtle differences are not important. What is important is to be able to interpret boxplots.

Referring to the boxplot in Figure 2.4, we consider the points scored in the 2010/2011 season up to February 15/2011 by the top 14 point scorers from each of the Vancouver Canucks and the Toronto Maple Leafs. Interpreting the boxplot, we note that the bolded line corresponds to the sample median of the data, and therefore, the median Canuck has more points than the median Leaf. The top edge of a box corresponds to the median of the largest 50% of the data values and the bottom edge of a box corresponds to the median of the smallest 50% of the data values. Therefore roughly half of each dataset is contained within its box. Since the Canucks box is higher than the Leafs box, this suggests that the Canucks are a higher scoring team. The **whiskers** (i.e. vertical dashed lines) extend to the outer limits of the data and circles correspond to outliers. For the Canucks, Daniel Sedin (73 points) and Henrik Sedin (68 points) are the top point scorers. Can you also see that there is skewness in the datasets where the upper tail is longer than the lower tail? When one box is longer than another box, this suggests greater variation in the data corresponding to the longer box.

2.5 Paired Data

Sometimes data arise in pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and we refer to the structure as **bivariate**. For example, we might measure red wine consumption and cholesterol level on each of n individuals. When we have bivariate

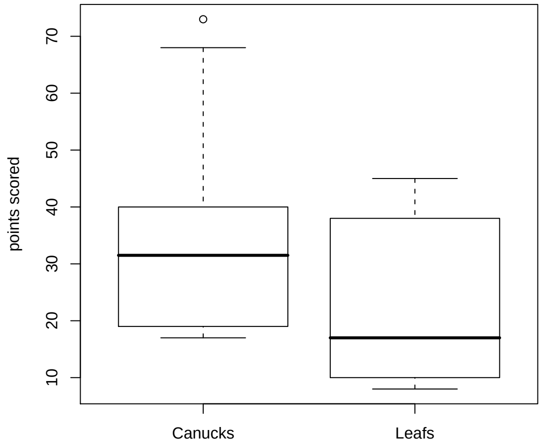

Figure 2.4: A boxplot of points scored as of February 16/2011 by the 14 top scorers on the Vancouver Canucks and the Toronto Maple Leafs.

data, our interest extends beyond the individual characteristics of the x data and the y data. In the example, we are also interested in the relationship between red wine consumption and cholesterol level. Note that the relationship can be nonexistent, one of **association** or **causal**.

2.5.1 Scatterplots

A scatterplot is a familiar graphical descriptive statistic that is appropriate for paired data. An example of a scatterplot is given in Figure 2.5 where the final exam score y is plotted against the term score x based on a sample of $n = 15$ students who previously took Introductory Statistics. There is a general increasing trend to the points which suggests that those students who did well during the term also tended to do well in the final exam. In fact, the scatterplot suggests that we might be able to predict the final exam score given the term score. For example, we might roughly predict $y = 25.0$ based on $x = 30.0$ by drawing a vertical line up from $x = 30.0$. If the points had appeared in the shape of a horizontal band, this would indicate no relationship between x and y.

Prediction is an important topic which receives greater attention in subsequent statistics courses. However, I want to make a general point concerning prediction which applies to scatterplots. When predicting, you should be cautious about predictions based on **extrapolated data**. For example, suppose that the x's represent caffeine consumption and the y's represent some performance score involving an individual's reaction time. You may expect to observe an increasing relationship between x and y. However, there is obviously some limit in terms of the relationship. If you were to consume five energy drinks, I doubt that your reaction score would be very good; you would likely be a mess and your reaction score would be poor. Therefore, you cannot look at the increasing relationship between x and y for reasonable values of x and expect that the same relationship exists for extreme (i.e. extrapolated) values of x.

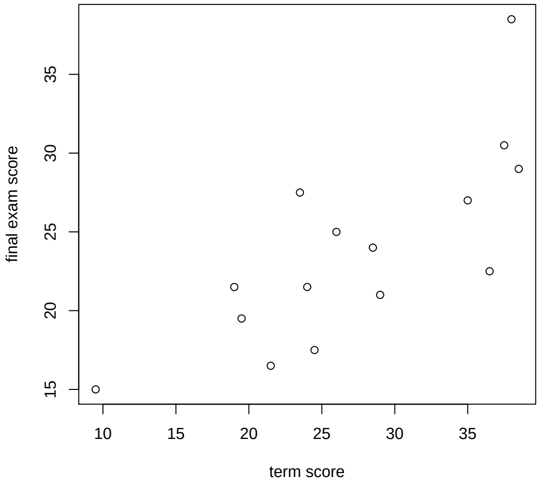

Figure 2.5: A scatterplot of final exam scores versus term scores in Introductory Statistics.

2.5.2 Correlation Coefficient

A numerical descriptive statistic for investigating paired data is the **sample correlation** or **correlation** or **correlation coefficient** r defined by

$$\boxed{r \;=\; \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}} \tag{2.3}$$

provided that not all of the x_i's are the same and not all of the y_i's are the same.

The correlation coefficient is **dimensionless** which means that it is not reported in any units. The reason for this is that the units in which x and y are measured are cancelled in the numerator and the denominator of expression (2.3). The correlation coefficient has a number of properties including:

- $-1 \leq r \leq 1$
- when $r \approx 1$, the points are clustered about a line with positive slope
- when $r \approx -1$, the points are clustered about a line with negative slope

Because of the three properties listed above, the correlation coefficient is a statistic that is used to measure the degree of linear association between x and y. As an exercise, suppose that there exists an exact linear relationship $y_i = a + bx_i$, $i = 1, \ldots, n$. By substituting into (2.3), establish that $r = 1$ when $b > 0$ and $r = -1$ when $b < 0$.

To better appreciate expression (2.3), suppose that the points are clustered about a line with positive slope. This means that large x's tend to occur with large y's and small x's tend to occur with small y's. From (2.3), a large x_i and a large y_i results in positive values of $(x_i - \bar{x})$ and $(y_i - \bar{y})$ leading to $(x_i - \bar{x})(y_i - \bar{y}) > 0$. Similarly, a small x_i and a small y_i results in negative values of $(x_i - \bar{x})$ and $(y_i - \bar{y})$ also leading to $(x_i - \bar{x})(y_i - \bar{y}) > 0$. For all datasets, the denominator is non-negative. While (2.3) is instructive in understanding the properties of the correlation coefficient, the following expression is more convenient for calculation

$$\boxed{r \;=\; \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}.} \tag{2.4}$$

Note that we have omitted the indices of summation in (2.4) to simplify the notation.

A couple more important points are in order. First, a weak correlation (i.e. $r \approx 0$) implies a lack of a linear relationship between x and y. It does not, however, imply a lack of a relationship. For example, x and y may have an exact quadratic relationship yet r may be close to zero. Second, when x and y are correlated (i.e. r is not close to zero), this does not necessarily mean that x causes y or that y causes x. It merely denotes the presence of a linear association. For example, weight and height are positively correlated yet it is obviously wrong to state that one causes the other.

Example 2.1 The headline "Prayer can Lower Blood Pressure" appeared in the August 11/1998 issue of USA Today, and the article explains that people who attended a religious service once a week, and prayed or studied the Bible were 40% less likely to have high blood pressure. As statisticians in the making, I hope that you notice that a cause and effect statement was made, and that the purported cause and effect relationship may be fallacious. The setting described here is referred to as an **observational study** where data are observed (usually **retrospectively**) and the data collector does not have control of auxiliary conditions. A possible explanation for the data is that those who attend church once a week are relatively healthy and do not suffer from high blood pressure. Shut-ins who are older and less healthy, may have high blood pressure and are unable to attend church. Therefore, the observational study was unable to account for auxiliary variables such as age and health. It may have been that age and health have a causal impact on attendance, and attendance is correlated with high blood pressure. Now, let me be clear (I don't want to receive any threatening emails); I am not saying that prayer, church, etc. is not beneficial. In fact, I think it is. What I am saying is that the data (as collected) do not address the existence of a cause and effect relationship.

We contrast observational studies with **controlled studies** where the data collector has control over the relevant conditions. For example, suppose that you are baking cakes. You have control over the number of eggs, the amount of sugar, the temperature of the oven, etc. In a controlled study, cause and effect relationships may be established. For example, you may find that increasing the number of eggs from two to three results in a better tasting cake.

2.6 Exercises

2.01* At the end of a course, Professor Swartz tallies up the assignment scores, the midterm scores and the exam score to produce a final numerical score for each student. When he has a small class of students (say less than 20 students), he then typically produces a dotplot of the final scores before converting the final scores to letter grades. Explain why this may be a reasonable thing to do and why he does not do this for large classes.

2.02 A car owner is interested in selling his 2005 Ford Escape with an odometer reading of 102,000 km. Going to the autotrader.ca website, he observes that there are 16 Ford Escapes manufactured from 2004 to 2006, with 80,000 to 125,000 km. The asking prices for these cars are: \$11850, \$13888, \$13900, \$13988, \$13990, \$13995, \$14500, \$14995, \$15850, \$15988, \$15988, \$15995, \$16987, \$16988, \$16995 and \$16995. All of these cars are for sale by dealers. Produce a dotplot of the prices and comment on features of the dotplot. Suggest a reasonable asking price for the 2005 Escape. Explain your reasoning.

2.03 The following are the heights of tomato plants measured in cm. Note that the dataset is fairly small ($n = 25$). Construct a histogram using five intervals each of length 5.0 cm beginning with $[40.0, 45.0)$. Comment on features of the histogram including outliers, centrality, dispersion and skewness.

46.6	43.8	43.4	57.2	55.1
56.7	57.6	49.7	44.0	47.2
51.0	61.5	55.9	48.7	43.0
54.9	51.9	49.1	51.3	49.3
57.8	49.8	61.7	58.5	42.6

Using the same data, construct a second histogram using five intervals each of length 5.0 cm beginning with $[42.0, 47.0)$. What does the difference in appearance of the two histograms tell you about the general interpretation of histograms?

2.04 Refer to the data of Exercise 2.03. Although I think there is no good reason for doing so, suppose that you were insistent on using the following intervals: $[40.0, 44.0)$, $[44.0, 48.0)$, $[48.0, 52.0)$, $[52.0, 58.0)$ and $[58.0, 65.0)$. Produce the appropriate histogram using these unequal length intervals.

2.05 On a long road vacation, my children spent an hour counting the colours of cars that passed by on the opposite side of the road. The counts were as follows: silver/grey (497), white (226), black (228), blue (173), red (146), green (58), other (172). (a) Produce a barplot corresponding to the count data. (b) If we change the order of the colours, the shape of the barplot changes. What does this tell you about the interpretation given to histograms and barplots as it relates to the type of variable on the horizontal axis?

2.06 The final scores for a statistics course of 35 students were as follows:

71.5	66.4	56.3	42.1	34.9	79.4	54.0
48.2	54.5	45.1	63.2	64.1	27.0	41.2
49.7	66.1	54.9	42.0	63.6	61.4	79.7
47.5	80.3	50.2	51.7	53.5	51.7	42.8
74.4	75.3	65.5	57.2	44.1	59.4	52.3

Produce a histogram, and draw some insights from the histogram.

2.07* Suppose that we take a sample of n people of all ages, and we record their times $x_1, \ldots, x_n$ in seconds to run 100 metres. Do you think $\bar{x}$ would exceed $\tilde{x}$? Explain.

2.08 Describe a dataset where the median exceeds the mean. Explain your reasoning. Left-skewed datasets are less common than right-skewed datasets.

2.09* Consider the following $n = 7$ measurements in kg: 24, 16, 21, 27, 28, 23, 19. A recording mistake was made and exactly one of the measurements was transposed (e.g. 42 was incorrectly recorded as 24). (a) Which of the measurements when transposed back to its correct value leads to a different median than when calculated with the incorrect value? (b) Which of the measurements do you believe was incorrectly transposed? Explain.

2.10 Twenty air samples were obtained and the carbon monoxide concentrations were recorded. The results in ppm (parts per million) were:

9.3	10.7	8.5	9.6	12.2	15.6	9.2	10.5	9.0	13.2
11.0	8.8	13.7	12.1	9.8	10.5	11.2	9.9	12.4	10.9

(a) Calculate the mean, the median and the standard deviation. (b) Obtain the mean, the median and the standard deviation in the units parts per thousand without recalculating everything from scratch.

2.11 A large number of authors ($n = 1074$) were classified according to the number of papers that they published in the past year.

# Papers	0	1	2	3	4	5	6	7	8
Frequency	210	365	253	186	45	12	0	2	1

Calculate the mean, median and standard deviation for the number of published papers.

2.12* Construct a dataset where the mean is $\bar{x} = 50.0$ and the standard deviation is $s = 50.0$.

2.13 Is it possible to construct a dataset where the mean is $\bar{x} = 30.0$, the range is $R = 10.0$ and the variance is $s^2 = 40$?

2.14 The lengths of power failures, in minutes, are recorded as follows:

22	18	135	15	90	78	69	98	102
83	55	28	121	120	13	22	124	112
70	66	74	89	103	24	21	112	21
40	98	87	132	115	21	28	43	37
50	96	118	158	74	78	83	93	95

Calculate the mean, the median and the standard deviation.

2.15 Suppose that in a sample of 73 observations, $s^2 = 10.8$ mm^2 and $\sum x_i^2 = 815.6$ mm^2. (a) Obtain the sample mean. (b) If 2.0 mm is added to every observation and then the resulting quantity is multiplied by 3.0, a new dataset is formed. Obtain the sample mean of the new dataset.

2.16 Consider data measured in kg on $n = 10$ subjects where the 10-th measurement is missing but the remaining measurements are: 65.6, 52.3, 85.9, 70.0, 58.4, 73.1, 69.6, 81.5 and 75.2. (a) If the sample mean of all 10 measurements is $\bar{x} = 70.9$ kg, calculate the missing measurement. (b) If the sample median of all 10 measurments is $\tilde{x} = 69.8$ kg, what can you conclude about the missing measurement? (c) If the range of all 10 measurments is $R = 34.3$ kg, calculate the missing measurement. (d) If the standard deviation of all 10 measurments is $s = 10.0$ kg, calculate the missing measurement.

2.17 Boxplots should be produced using statistical software. However, suppose that you have a sample of $n = 20$ body fat measurements in percentages for the university track and field team, and that you also have body fat measurements for a sample of $n = 50$ third year students from the Faculty of Arts. Carefully sketch what you think the relevant boxplot would look like.

2.18 Suppose that you have a dataset consisting of the average number of goals scored per game during the season for soccer teams. This data has been collected for 7 year old girls house league teams in the Lower Mainland, for 13 year old elite boys teams in the Lower Mainland and for the English Premier League. Carefully sketch what you think the relevant boxplot would look like.

2.19 Consider the construction of a dotplot, a histogram and a boxplot from a single sample of data. From the point of view of the original data, which of the three graphical displays contains the most information (least amount of summarization) and which contains the least information (greatest amount of summarization)? Explain.

2.20 The following data represent the chemistry grades for a sample of first year students along with their scores on an intelligence test taken while they were in high school.

Intelligence Score	65	50	55	65	55	70	65	70	55	70	50	55
Chemistry Score	85	74	76	90	85	87	94	98	81	91	76	74

(a) Sketch the appropriate scatterplot. (b) Calculate and interpret the correlation coefficient.

2.21 Below is a set of $n = 10$ randomly generated numbers x and another set of randomly generated numbers y. The numbers have been arbitrarily paired.

x	21.4	12.4	11.5	11.0	12.4	15.2	18.2	14.7	7.9	9.5
y	3.6	9.4	7.5	8.1	7.4	3.9	7.1	3.7	5.8	6.9

Sketch the scatterplot of y versus x. What sort of conclusion are you tempted to make from the scatterplot? What lesson concerning sample size can be taken from this example?

2.22* Suppose that data are collected on the time in seconds to perform a task. There are 8 x's corresponding to individuals who trained according to method 1. There are 8 y's corresponding to individuals who trained according to method 2. Explain whether a scatterplot would be a good graphical display for the data. If not, suggest alternative summary statistics for the data.

2.23 As of March 1/11, the points scored per game and the points allowed per game (PPG, APG) for each of the 15 teams in the Eastern Conference of the National Basketball Association (NBA) were recorded as follows: NY (107.3, 105.6), Mia (102.0, 94.4), Orl (100.0, 94.0), Ind (99.7, 99.9), Tor (98.8, 105.0), Phi (98.6, 96.7), Chi (98.4, 92.2), Bos (98.0, 91.3), Was (96.6, 103.8), Atl (96.2, 95.3), Cle (95.8, 105.9), Det (95.4, 99.6), Cha (94.3, 96.7), NJ (92.9, 99.2) and Mil (91.4, 92.9). Produce a scatterplot and calculate the correlation coefficient. Is there anything to be learned about scoring patterns in the NBA or about particular teams?

2.24 In a recent exercise, I asked students to list the country where they were born, where their mother was born and where their father was born. Propose a statistic that would help interpret the data.

Chapter 3

Probability

The word "probability" is thrown around in everyday conversation. For example, people talk about the probability that it is going to rain today. However, if you press people as to what they really mean, they will often rephrase the statement by referring to "chance", "likelihood" or "plausibility". These are also vague concepts and may be regarded as synonyms of probability. In this chapter, we attempt to gain a deeper understanding of the illusive concept of probability and we develop some tools for calculating probabilities.

3.1 Framework

Let's begin with some terminology that is useful in the study of probability. First, we think of an **experiment** in a rather broad setting. For us, an experiment does not refer to scientists in white coats. Instead, an experiment is any action that produces data. Corresponding to an experiment is the **sample space** which is the set of all possible outcomes of the experiment. Typically, we denote the sample space by S. Then an **event** is defined as a subset of the sample space. The probabilities of events are of interest to us.

Example 3.1 Consider the experiment of flipping a coin three times. This is obviously not the sort of experiment involving guys in white coats but it does satisfy our definition of an action that produces data. Ignoring the possibility that the coin lands on its side, the corresponding sample space is $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$, where for example, HTH refers to a head on the first flip, a tail on the second flip and a head on the third flip. Note that the set S is both **discrete** and **finite**. There are $2^8 = 256$ possible events that can be defined relative to this sample space (think about this). An example is the event A of two or more heads. Note that A is a subset of S and is given by $A = \{HHH, HHT, HTH, THH\}$.

Example 3.2 Consider the experiment of the number of automobile accidents in British Columbia in a future year. Here the sample space $S = \{0, 1, \ldots\}$ is a discrete but infinite set. Let A be the event of 1000 or more accidents which is given by the subset $A = \{1000, 1001, \ldots\}$.

Example 3.3 Consider the experiment of the life span in hours of two electronic components which form a system. Then $S = \{(x_1, x_2) : x_1 \geq 0, x_2 \geq 0\}$ and $A = \{(x_1, x_2) : 0 \leq x_1 < 10 \text{ or } 0 \leq x_2 < 10\}$ is the event that the system fails in under 10 hours. The system fails if any of its components fail. Note that S is bivariate and **continuous**.

3.2 Set Theory

We now introduce some notation, rules and diagrams that are useful in the study of events. Recall that our eventual goal is to assign probabilities to events.

When we speak of the **intersection** of the events A and B, we mean those elements in the sample space that are common to both A and B. We denote the intersection as AB or $A \cap B$, and we sometimes say "A and B". The intersection of A and B is conveniently shown in the **Venn diagram** of Figure 3.1. Venn diagrams are graphical displays of events where the sample space is depicted by the entire Venn diagram.

The **complement** of the event A is denoted by $\overline{A}$ and sometimes by A^c and sometimes by A'. The complement of A consists of elements in the

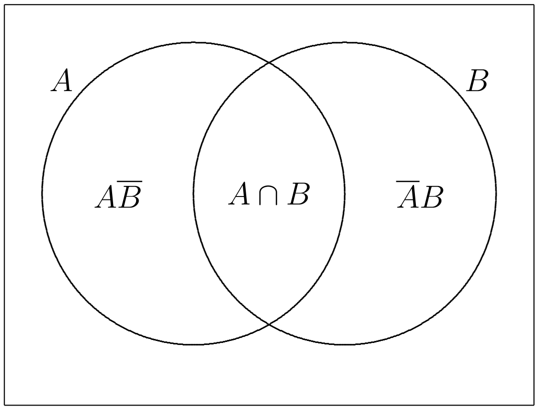

Figure 3.1: A Venn diagram highlighting different events.

sample space that do not belong to A. The events $A\overline{B}$ (i.e. the intersection of A and $\overline{B}$) and $\overline{A}B$ (i.e. the intersection of $\overline{A}$ and B) are both shown in Figure 3.1.

We denote the **union** of the events A and B by $A \cup B$. The union of A and B consists of all elements in the sample space that either belong to A or belong to B. Sometimes we say "A or B" when we refer to the union of A and B. In Figure 3.1, $A \cup B$ consists of all elements lying in any of $A\overline{B}$, $A \cap B$ or $\overline{A}B$.

The **empty set** is denoted by ϕ and we are careful in noting that zero is not the same as the empty set. We say that events A and B are **mutually exclusive** or **disjoint** if $A \cap B = \phi$. In other words, mutually exclusive or disjoint events have no common elements. Two disjoint events are shown in the Venn diagram of Figure 3.2.

Putting some of these ideas together, we note that **de Morgan's Laws** are derived as:

$$\boxed{\begin{aligned} \overline{A \cup B} &= \overline{A} \cap \overline{B} \\ \overline{A \cap B} &= \overline{A} \cup \overline{B}. \end{aligned}}$$

Fiddle around with a Venn diagram and convince yourself that de Morgan's

Laws are correct.

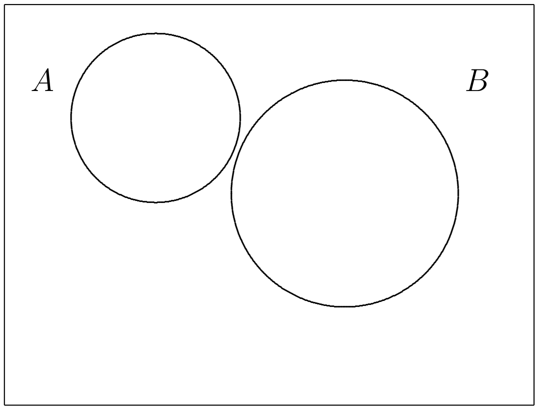

Figure 3.2: A Venn diagram where the events A and B are disjoint.

3.3 Three Definitions of Probability

I mentioned earlier that probability is an illusive concept. Let's present three definitions of probability (there are actually more!) and then criticize the definitions. The first definition is the **axiomatic definition** that is due to the great Russian probabilist Kolmogorov in 1933. He stated that a probability measure P satisfies three axioms:

1. for any event A, $P(A) \geq 0$

2. $P(S) = 1$

3. if events $A_1, A_2, \ldots$ are mutually exclusive, then $P(\cup A_i) = \sum P(A_i)$

These are simple axioms that have intuitive appeal. For example, axiom 1 states that probabilities are non-negative. Well, who ever heard of a negative probability? Axiom 1 is certainly acceptable. Axiom 2 states that the probability of the sample space is 1.0. Well, something has to happen in the experiment and the sample space is the set of all possible outcomes of the experiment. Finally, axiom 3 is referred to as **countable additivity** and it states that if events are disjoint, then the probability of their union is equal to the sum of their respective probabilities. Most people also find axiom 3 acceptable.

The wonderful thing about the axiomatic definition is that many useful properties concerning probability can be proved from the three simple axioms. For example,

$$1 \;=\; P(S) \;=\; P(A \cup \overline{A}) \;=\; P(A) + P(\overline{A})$$

where the first equality is due to axiom 2 and the third equality is due to axiom 3. We therefore obtain the result $P(\overline{A}) = 1 - P(A)$.

Another useful result that you should commit to memory concerns the probability of the union of two events

$$\boxed{P(A \cup B) = P(A) + P(B) - P(AB).} \tag{3.1}$$

It is derived as follows:

$$\begin{aligned}
P(A \cup B) &= P(A\overline{B} \cup AB \cup \overline{A}B) \\
&= P(A\overline{B}) + P(AB) + P(\overline{A}B) \\
&= [P(A\overline{B}) + P(AB)] + [P(\overline{A}B) + P(AB)] - P(AB) \\
&= P(A\overline{B} \cup AB) + P(\overline{A}B \cup AB) - P(AB) \\
&= P(A) + P(B) - P(AB).
\end{aligned}$$

Example 3.4 If 85% of Canadians like either baseball or hockey, 63% like hockey and 52% like baseball, what is the probability that a randomly chosen Canadian likes both baseball and hockey?

Solution: Baseball and hockey, hockey and baseball - questions like this can be confusing. A general recommendation that I make throughout the course is that you introduce some notation. Therefore let A be the event that a Canadian likes hockey and let B be the event that a Canadian likes baseball. It is immediately apparent that we are asked to calculate $P(A \cap B)$ where

$P(A) = 0.63$, $P(B) = 0.52$ and $P(A \cup B) = 0.85$. Substituting into (3.1) gives $P(A \cap B) = 0.30$.

As an exercise, use the three axioms to establish the simple result $P(\phi) = 0$ and to establish the generalization of (3.1) concerning the union of three events

$$\boxed{\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &- P(AB) - P(AC) - P(BC) \\ &+ P(ABC). \end{aligned}}$$

Now, although the axiomatic definition is appealing, it has two major drawbacks in that it neither tells us how to interpret probabilities nor how to calculate probabilities. For example, the axioms do not tell us how to interpret a weather forecast of 70% chance of rain. Also, the axioms do not help us determine the probability that the Vancouver Canucks will win the Stanley Cup. The axiomatic definition simply describes the properties that a probability measure must satisfy.

Our second definition of probability is the **symmetry definition** which is applicable in problems of symmetry. More specifically, when we have a finite number of equally likely outcomes in the experiment, the symmetry definition of the probability of an event A is defined by

$$P(A) = \frac{\text{number of outcomes leading to } A}{\text{total number of outcomes}}.$$

Example 3.5 Suppose that we roll two dice and we are interested in the probability of the event A that the sum of the two dice is 10.
Solution: In this case, the sample space $S = \{(1,1),(1,2),\ldots,(6,6)\}$ has 36 outcomes and $A = \{(4,6),(5,5),(6,4)\}$ has 3 outcomes. The symmetry definition therefore gives $P(A) = 3/36 = 1/12$.

Now there are serious difficulties with the symmetry definition. Foremost, a condition of the definition is that we have a finite number of equally likely outcomes. The definition is therefore somewhat circular in that we are defining probability in terms of outcomes that are “equally likely”. Also, the definition is not widely applicable as it requires a *finite* number of outcomes. For example, how does an individual calculate the probability of having a heart attack by age 40? Nevertheless, the symmetry definition is extremely useful when calculating probabilities with respect to games of chance (e.g. dice, coins, cards, etc.). We also note that the symmetry definition can sometimes be used in spatial problems. For example, ignoring bullseyes, the probability of obtaining an even number on a dart board is 1/2 given a randomly thrown dart. The reasoning is based on symmetry by noting that the area on the dart board corresponding to even numbers equals the area on the dart board corresponding to odd numbers.

The third definition of probability that we consider is the **frequency definition**. Under N hypothetical and identical trials, the frequency definition of the probability of an event A is given by

$$P(A) \;=\; \lim_{N\to\infty}\left(\frac{\text{number of occurrences of } A}{N}\right).$$

Suppose again that we are interested in the probability that the sum of the two dice is 10. We have already seen that the symmetry definition is applicable in this problem. Using the frequency definition, we imagine that the two dice could be thrown again and again. This is where the “hypothetical” part is involved. We do not throw the dice an infinite number of times. In fact, it is impossible to do anything infinitely often. But the frequency definition provides a more general framework as to how to think about probability. In this case we think of the probability as

$$P(\text{sum is 10}) \;=\; \lim_{N\to\infty}\left(\frac{\text{number of times the sum is 10 in } N \text{ trials}}{N}\right).$$

Apart from the fact that the frequency definition does not allow us to calculate probabilities, it has further limitations. For one, the probability is expressed as a limit, and there is no mathematical reason why the limit must exist. Secondly, although the frequency definition is more general than the symmetry definition, it does not address the interpretation of probabilities where events cannot be repeated (even hypothetically). For example, the definition needs to be modified to handle probabilities such as the probability that I will die tomorrow. Ignoring these criticisms, the frequency definition is our default definition for thinking about probability and is applicable in the majority of situations that we encounter.

Let me conclude this section with a few remarks. First, I hope that you now appreciate that probability is a tricky concept, and it is curious that so many people use probabilistic language despite its nebulous nature. Next, this is admittedly an abstract topic and therefore I don't want you to spend too much time on this section.

3.4 Conditional Probability

A very important type of probability is **conditional probability**. Conditional probability is relevant when partial information is revealed. For example, suppose that I am considering a gall bladder operation and I am told that the probability of survival for this operation exceeds 0.99. As good as it sounds, I am not completely satisfied with the information. I would rather know the probability of survival for a non-smoking, moderate drinking man in his 50's which more specific to my situation.

Let's suppose then that we are interested in event A but event B has been revealed to us. That is, we know that B is true. Then the relevant probability is no longer $P(A)$ but rather the conditional probability of event A given event B which we denote by $P(A \mid B)$. The **conditional probability** is defined by

$$\boxed{P(A \mid B) \; = \; \frac{P(A \cap B)}{P(B)}} \tag{3.2}$$

provided that $P(B) \neq 0$.

To motivate the definition of the conditional probability formula (3.2), recall that the unconditional probability of the event A is $P(A) = P(A)/P(S)$ where the probability of the sample space is $P(S) = 1$. In the case of the conditional probability, if you think of the Venn diagram involving the events A and B, then knowing that B has occurred, restricts the relevant sample space to B and restricts A to the region $A \cap B$. By analogy, this gives the ratio in (3.2).

Example 3.6 Suppose that I roll a die, conceal the result but tell you that the die is even. What is the probability that a six was rolled?
Solution: In this problem, partial information is revealed. Therefore, defining the event A that a six has occurred and defining the event B that the die is even, we obtain

$$P(A \mid B) \;=\; \frac{P(AB)}{P(B)} \;=\; \frac{P(A)}{P(B)} \;=\; \frac{1/6}{3/6} \;=\; 1/3$$

where the first equality is due to (3.2), the second inequality recognizes that the intersection of a six and an even number is a six, and the third equality is based on the symmetry definition of probability.

Example 3.7 The probability of surviving an operation involving an organ donation is 0.65. If a patient survives, the probability that the body rejects the organ within 60 days is 0.2. What is the probability of surviving both critical stages?
Solution: This strikes me as the sort of information that I would want to know as a patient. It does not thrill me if I survive the operation but die within 60 days. I want to survive both stages. When a problem has the sort of structure when one thing follows another, then it is often a sign that conditional probability is involved. Let O be the event of surviving the operation and let S be the event of surviving 60 days. We have $P(O) = 0.65$ and $P(S \mid O) = 0.8$ since surviving 60 days is the complement of dying within 60 days. Rearranging the equation $P(S \mid O) = P(SO)/P(O)$ gives $P(SO) = P(S \mid O)P(O) = (0.8)(0.65) = 0.52$.

Related to conditional probability is a formula sometimes referred to as **the law of total probability**. The law of total probability is simply a straightforward extension of conditional probability. To derive the law, consider an event of interest A and events $B_1, B_2, \ldots$ which form a **partition**.

A partition means that the B_i are disjoint and that $S = \cup_{i=1}^{\infty} B_i$ where S is the sample space. In other words, exactly one of the B_i is true. Note that the number of B_i can also be finite. Then by referring to the corresponding Venn diagram and the third axiom,

$$P(A) = P(\cup_{i=1}^{\infty} AB_i) = \sum_{i=1}^{\infty} P(AB_i).$$

Using conditional probability, we then obtain the **law of total probability**

$$\boxed{P(A) = \sum_{i=1}^{\infty} P(A \mid B_i)\ P(B_i)}$$

which sometimes provides an easier calculation for $P(A)$. We can use the law of total probability to obtain the probability of one of the events $B_r, r = 1, 2, \ldots$ conditional on A as follows:

$$\boxed{P(B_r|A) = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{\infty} P(A \mid B_i)\ P(B_i)}.}$$

This is known as **Bayes' Rule**.

3.4.1 Independence

Another important concept related to conditional probability is **independence**. Informally, two events are independent if the occurrence of either event does not affect the probability of the other. For example, if I conduct an experiment where I toss a coin and roll a die, the outcome of the coin toss does not affect the die and vice-versa. At some point (either a midterm or the final exam), I am likely to ask you whether or not two events are independent. What I do not want you to say is something like “the events are independent because they do not affect one another”. You must prove or disprove independence using the following formal definition of independence:

Events A and B are independent **iff** (if and only if)

$$P(AB) \;=\; P(A)P(B). \tag{3.3}$$

I am sometimes asked by students about the use of “iff” above. Sometimes they think that I actually mean “if” but that I can't spell. In fact, what is meant by the statement is that both independence implies expression (3.3) and expression (3.3) implies independence. When two events are not independent, we say that they are **dependent**.

Example 3.8 Suppose that a coin is tossed and a die is rolled. What is the probability of getting both a tail and a six?
Solution: There are at least a couple of approaches to this problem. You may consider the compound experiment of tossing the coin and rolling the die leading to the compound sample space $S = \{(H1), (H2), \ldots, (T6)\}$ consisting of 12 elements which are self-explanatory. In this case, using the symmetry definition of probability,

$$P(\text{tail} \cap \text{six}) \;=\; \frac{\#\{(T6)\}}{\#S} \;=\; \frac{1}{12}$$

where $\#A$ denotes the cardinality of the set A. Alternatively, we can appeal to the independence of the coin toss and the rolling of the die, and obtain

$$P(\text{tail} \cap \text{six}) \;=\; P(\text{tail})P(\text{six}) = (1/2)(1/6) \;=\; 1/12.$$

In some examples, it is helpful to list the elements of the sample space systematically using a **tree diagram**. Figure 3.3 is a tree diagram for Example 3.8. The paths along the branches of the tree yield distinct sample points. The branches are labelled with corresponding probabilities.

Here is a little tip. Consider a word problem where you are asked to calculate a probability. When you see phrases such as “**if** so and so happened” or “**given** that so and so happened”, these are often clues that conditional probability is involved.

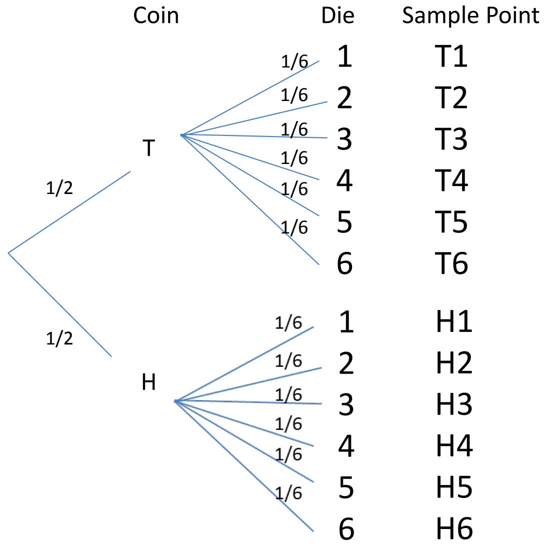

Figure 3.3: A tree diagram for Example 3.8.

Example 3.9 We now look at one of the most famous problems in probability. It is known as the **birthday problem** and is famous since the solution is counter-intuitive to many people. Suppose that you have $n = 30$ people in a room. What is the probability that at least two people share the same birthday?

Solution: First a few caveats: Let's assume that birthdays are spread evenly across days of the year and we ignore leap years. When we say that two people share the same birthday, an example of this is that they both have a June 2nd birthday. Naturally, as we are studying conditional probability, the solution uses conditional probability. Let A_i be the event that the first i people have different birthdays for $i = 1, \ldots, n$. We note that $A_{i+1} \subseteq A_i$ such that $A_{i+1}A_i = A_{i+1}$. In other words, if the first $i + 1$ people have different birthdays, then for sure, the first i people have different birthdays. We then calculate the probability P that at least two people share a common birthday where

$$\begin{aligned}
P &= P(\overline{A}_n) \\
&= 1 - P(A_n) \\
&= 1 - P(A_1 A_2 \cdots A_n) \\
&= 1 - P(A_n \mid A_1 A_2 \cdots A_{n-1}) P(A_1 A_2 \cdots A_{n-1}) \\
&= 1 - P(A_n \mid A_{n-1}) P(A_1 A_2 \cdots A_{n-1}) \\
&= 1 - P(A_n \mid A_{n-1}) P(A_{n-1} \mid A_{n-2}) \cdots P(A_2 \mid A_1) P(A_1) \\
&= 1 - \left(\tfrac{365-n+1}{365}\right)\left(\tfrac{365-n+2}{365}\right) \cdots \left(\tfrac{365-1}{365}\right) \cdot 1.
\end{aligned}$$

To appreciate the last equality, consider $P(A_n \mid A_{n-1})$. If the first $n-1$ people have different birthdays, then the nth person can have a birthday on any of the $365 - n + 1$ remaining days for A_n to be true.

When $n = 30$, the probability $P = 0.71$ is larger than many people anticipate. The reason why the probability is fairly large is that although it is improbable that someone in a group of 30 shares the same birthday as you, it is probable that amongst all of the many pairs of people, at least two share a common birthday. For interest, when $n = 40$, the probability of two people sharing a common birthday is $P = 0.89$. Did you find this surprising?

You may have anticipated that there is a connection between independence and conditional probability. It is expressed in the following proposition.

Proposition 3.1 Suppose that $P(B) \neq 0$. Then the events A and B are independent iff $P(A \mid B) = P(A)$.

Proof: From the definition of conditional probability, it follows immediately that

$$P(A \mid B) \;=\; P(AB)/P(B) \;=\; P(A)P(B)/P(B) \;=\; P(A).$$

The converse result is obtained by going backwards in the proof.

Proposition 3.1 is intuitive since $P(A \mid B) = P(A)$ suggests that there is no information in B relative to A, and therefore A and B do not affect one another (i.e. independence), and vice-versa.

3.5 Some Counting Rules

In order to calculate probabilities using the symmetry definition, it is useful to have some counting or **combinatorial** rules at our disposal.

Counting Rule 3.1 The number of **permutations** or arrangements of n distinct objects is $\boxed{n! = n(n-1)\cdots 1}$ where we define $0! = 1$. The notation $n!$ is expressed as "n factorial".

For example, the letters A, B and C can be arranged in the following $3! = 6$ ways:

$$ABC \quad ACB \quad BAC \quad BCA \quad CAB \quad CBA$$

Counting Rule 3.2 The number of permutations of r objects chosen from n distinct objects is $n^{(r)} = n!/(n-r)!$ where $r \leq n$. When we use the special notation $n^{(r)}$, we say "n to r factors".

For example, we can permute 2 of the 5 letters A, B, C, D, E in $5^{(2)} = 5!/(5-2)! = 120/6 = 20$ ways. The 20 arrangements are:

$$\begin{array}{cccccccccc} AB & AC & AD & AE & BA & BC & BD & BE & CA & CB \\ CD & CE & DA & DB & DC & DE & EA & EB & EC & ED. \end{array}$$

Note for example, that AB is counted as distinct from BA. To see why Counting Rule 3.2 is correct, the number of arrangements is

$$n \cdot (n-1) \cdots (n-r+1) \tag{3.4}$$

where there are n symbols that can be placed in the first position, any of $n-1$ symbols can then be placed in the second position, and so on, until

$(n-r+1)$ symbols are available for the r-th position. We can then rewrite (3.4) as

$$n \cdot (n-1) \cdots (n-r+1) \cdot \frac{(n-r) \cdot (n-r-1) \cdots 1}{(n-r) \cdot (n-r-1) \cdots 1} \;=\; \frac{n!}{(n-r)!}.$$

Counting Rule 3.3 The number of **combinations** of r objects chosen from n distinct objects is

$$\binom{n}{r} \;=\; \frac{n^{(r)}}{r!} \;=\; \frac{n!}{(n-r)!\; r!}. \tag{3.5}$$

Combinations are different than permutations in that order is not important. For example, we can choose 2 of the 5 letters A, B, C, D, E in $\binom{5}{2} = 5!/(3!\;2!) = 10$ ways. The 10 combinations are:

$$AB \quad AC \quad AD \quad AE \quad BC \quad BD \quad BE \quad CD \quad CE \quad DE\ .$$

To see why Counting Rule 3.3 is correct, there are $n^{(r)}$ ways of choosing r objects from n objects when order is important. We therefore divide by the $r!$ ways in which the r objects can be arranged. When we use the special notation $\binom{n}{r}$, we say "n choose r".

Note that we must be clever when we calculate $\binom{n}{r}$. For example, if we blindly attempt to calculate $\binom{30}{4}$ by evaluating both the numerator and the denominator according to (3.5), we may run into overflow problems since 30! and 26! are enormous numbers. Instead, we write

$$\binom{30}{4} \;=\; \frac{30!}{26!\;4!} \;=\; \frac{30 \cdot 29 \cdot 28 \cdot 27 \cdot 26!}{26! \cdot 4 \cdot 3 \cdot 2 \cdot 1}$$

and observe that 26! cancels in the numerator and the denominator. We then quickly obtain $\binom{30}{4} = 27405$. Also, you will know that you have made a mistake if the numerator is not divisible by the denominator since combinations are positive integers. Make sure that you are able to evaluate combinatorial quantities.

As a final note on combinations and permutations, people often refer to "combination locks", as in, "Hey Joe, what's your combination?" Such locks are typically opened by a three-number sequence such as 22-6-15. However, the sequence 6-15-22 does not open the lock. Therefore, such locks ought to be called “permutation locks”. So the next time, you better yell, “Hey Joe, what’s your permutation?”

3.6 Examples

We have introduced a lot of different concepts in this chapter. We now run through a series of examples where we calculate probabilities. What makes this difficult is that it may not be immediately apparent which principles you should apply in calculating the probabilities. I recommend that you try a good number of the Exercise problems in addition to the problems presented here.

Example 3.10 In a class of 100 students, 20 are female. If we randomly draw 5 students to form a committee, what is the probability that at least two committee members are female?
Solution: Denote the probability P. Then the symmetry definition of probability gives

$$\begin{aligned} P &= 1 - P(0 \text{ female}) - P(1 \text{ female}) \\ &= 1 - \tbinom{80}{5}/\tbinom{100}{5} - \tbinom{80}{4}\tbinom{20}{1}/\tbinom{100}{5} \end{aligned}$$

where for example, $\binom{80}{5}\binom{20}{0} = \binom{80}{5}$ is the number of ways that 5 students can be drawn from a class of 100 students such that 0 are female and 5 are male, and $\binom{100}{5}$ is the number of ways that 5 students can be drawn from a class of 100 students. Note that this is a problem where the order of selection to the committee is not important. Having Brenda and Bob selected to the committee is no different than having Bob and Brenda selected to the committee. Make sure that you are able to reduce the final expression to a fraction.

Example 3.11 In a row of four seats, two couples randomly sit down. What is the probability that nobody sits beside his or her partner?

Solution: In some problems, our combinatorial rules of Section 3.5 are not that useful. This is one of those problems, where instead, we do the counting by brute force. If we denote the first couple A_1 and A_2, and the second couple B_1 and B_2, we note that the 8 orderings of interest are:

$$\begin{array}{llll} A_1B_1A_2B_2 & A_2B_1A_1B_2 & A_1B_2A_2B_1 & A_2B_2A_1B_1 \\ B_1A_1B_2A_2 & B_1A_2B_2A_1 & B_2A_1B_1A_2 & B_2A_2B_1A_1. \end{array}$$

With $4! = 24$ total possible orderings, the probability is $8/24 = 1/3$.

Example 3.12 We roll a die. If we obtain a six, we choose a ball from box A where three balls are white and two are black. If the die is not a six, we choose a ball from box B where two balls are white and four are black.
(a) What is the probability of obtaining a white ball?
(b) If a white ball is chosen, what is the probability that it was drawn from box A?
Solution: Word problems like this can be overwhelming without notation. We therefore let W denote a white ball, B_A denote box A, and B_B denote box B. Noting that you can only draw a ball from either box A or box B, the probability in part (a) is

$$\begin{aligned} P(W) &= P(WB_A \cup WB_B) \\ &= P(WB_A) + P(WB_B) \\ &= P(W \mid B_A)P(B_A) + P(W \mid B_B)P(B_B) \\ &= \left(\tfrac{3}{5}\right)\left(\tfrac{1}{6}\right) + \left(\tfrac{2}{6}\right)\left(\tfrac{5}{6}\right) \\ &= 17/45. \end{aligned}$$

Note that we could have jumped immediately to the third line in the calculation by invoking the law of total probability. In part (b), we pay special attention to the phrase "If a white ball is chosen" which suggests conditional probability. The probability is therefore

$$P(B_A \mid W) \;=\; \frac{P(WB_A)}{P(W)} \;=\; \frac{(3/5)(1/6)}{(17/45)} \;=\; \frac{9}{34}.$$

Example 3.13 Five cards are dealt from a deck of playing cards. What is the probability of getting

(a) three of a kind?
(b) two pair?
(c) a straight flush?

Solution: These problems are especially tricky and there are various expressions that lead to the same correct solution. More frightening is the fact that there are an infinite number of incorrect solutions.

All that you need to know about cards is that an ordinary deck of 52 playing cards consists of 13 **denominations** in each of four **suits**. The 13 denominations are ace, $2, 3, \ldots, 10$, jack, queen and king. The four suits are diamonds and hearts (which are both red), and clubs and spades (which are both black). For example, one card in the deck is a queen of diamonds. There is also a queen of hearts, a queen of clubs and a queen of spades.

A **hand** known as **three of a kind** consists of three cards of one denomination, a fourth card of a different denomination and a fifth card of a denomination different from the first two denominations. For example, the 3 of hearts, 3 of spades, 3 of diamonds, 7 of clubs and jack of hearts comprise three of a kind. With this understanding, we use the symmetry definition of probability to obtain

$$P(\text{three of a kind}) \quad = \quad \tfrac{\binom{13}{1}\binom{4}{3}\binom{12}{2}\binom{4}{1}\binom{4}{1}}{\binom{52}{5}} \quad = \quad 0.02113. \tag{3.6}$$

The denominator in (3.6) is simply the number of ways that 5 cards can be dealt from a deck of 52 cards where we are not concerned with order. The numerator is the number of ways that 5 cards can be dealt from a deck of 52 cards such that the cards comprise a three of a kind hand. The numerator is broken into logical components. The first term $\binom{13}{1}$ refers to choosing the denomination that is shared by three of the cards. Once the denomination is chosen, then $\binom{4}{3}$ is the number of ways that three cards of that denomination can be chosen. Recall that there are only four cards for each denomination. At this point, three of the cards have been chosen, and we need the remaining two. We choose the two remaining denominations via $\binom{12}{2}$ and specify the two cards from these two denominations by $\binom{4}{1}\binom{4}{1}$. Here is an equivalent expression for the numerator based on alternative reasoning

$\rightarrow \binom{13}{3}\binom{3}{1}\binom{4}{3}\binom{4}{1}\binom{4}{1}$. See if you can rationalize the expression. Observe that the probability of three of a kind is small. Such a hand occurs roughly one time out of 50 hands. Card players who are new to the game of **poker** often overestimate the probability of hands such as three of a kind. Of course, if you are playing poker with **wild cards**, the probabilities increase.

A hand known as **two pair** consists of two cards of one denomination, two more cards of a different denomination, and a fifth card of a denomination different from the first two denominations. The corresponding probability is

$$P(\text{two pair}) \;=\; \frac{\binom{13}{2}\binom{4}{2}\binom{4}{2}\binom{11}{1}\binom{4}{1}}{\binom{52}{5}} \;=\; 0.04754. \tag{3.7}$$

With respect to the numerator in (3.7), $\binom{13}{2}$ represents the number of ways that two denominations are chosen from the 13 denominations. Two cards are then chosen from each of these denominations and this is done in $\binom{4}{2}\binom{4}{2}$ ways. Having chosen these four cards, a fifth card is required and it must be chosen from one of the remaining 11 denominations, and this is done in $\binom{11}{1}\binom{4}{1}$ ways. Note that a different expression is available by writing $\binom{44}{1}$ instead of $\binom{11}{1}\binom{4}{1}$, and this is reasoned by noting that the fifth card may be any one of the $52 - 8 = 44$ cards whose denominations are different from the first four cards. As an exercise, verify for yourself (both logically and via calculations) that the numerator in (3.7) is also equivalent to $\binom{13}{3}\binom{3}{2}\binom{4}{2}\binom{4}{2}\binom{4}{1}$.

For the straight flush in part (c), we need to know that a **straight** corresponds to a consecutive sequence of denominations, and that an ace has the dual role of both the lowest card and the highest card. Therefore, there are 10 possible straights:

$$(\text{ace}, 2, 3, 4, 5), \quad (2, 3, 4, 5, 6), \quad \ldots, \quad (10, \text{jack}, \text{queen}, \text{king}, \text{ace}).$$

Furthermore, a **flush** corresponds to cards all of the same suit. Therefore, an example of a straight flush is the 3 of diamonds, 4 of diamonds, 5 of diamonds, 6 of diamonds and 7 of diamonds. A royal flush which consists of a 10, jack, queen, king and ace of the same denomination is not considered a straight flush. Since flushes can occur in any of the four suits, the probability of a straight flush is therefore a miniscule $(10-1)(4)/\binom{52}{5} = 0.000014$.

Example 3.14 Ignoring some details, in the 649 lottery, six balls are drawn from an urn containing 49 balls marked between 1 and 49. Participants in the lottery purchase tickets where they select 6 different numbers between 1 and 49. Monetary prizes are given according to whether enough of the participant's numbers match the numbers that are drawn from the urn. Calculate the probabilities corresponding to

(a) winning the jackpot (i.e. all six numbers match)

(b) two matching numbers (no prize is given in this case).

Solution: In part (a), the probability of winning the jackpot is $1/\binom{49}{6} = 0.00000007$ which suggests that none of us is ever going to win the jackpot. However, if you insist on playing, there is a little bit of strategy that you may employ. Recall that if you win the jackpot, you are required to split it with all participants having winning tickets. Therefore, it would be better if you selected "unpopular numbers" so that in the event that you do win, you would not need to share the jackpot. Try then to stay away from "popular numbers" such as 7. Did you know that Tim Horton (who was a defenceman on the Toronto Maple Leafs in the 1960's) had jersey number 7? People also have "lucky numbers" that correspond to birth dates. Therefore you might consider picking most of your numbers between 32 and 49.

In part (b), the probability of matching two numbers is $\binom{6}{2}\binom{43}{4}/\binom{49}{6} = 0.132$ where the numerator is obtained by noting that the 49 numbers are logically divided into 6 winning numbers and 43 losing numbers. What is interesting here is that the probability 0.132 is quite large. Therefore, in a lot of cases, ticket holders are "close" to winning (i.e. two numbers match). As an extension of this, you may also be close to winning, if for example, a 34 is drawn when your ticket contains 43, or you may have a 5 but need a 6, and so on. Perhaps being close to winning is part of the reason why people keep playing lotteries.

Example 3.15 Of 300 woodpeckers, 30 have damage to the beak but not the crest, 50 have damage to the crest but not the beak, and 10 have damage to both the beak and the crest. Suppose that a woodpecker is randomly drawn from this population.

(a) Are beak damage and crest damage independent?

(b) Are beak damage and crest damage mutually exclusive?

Solution: Letting B denote beak damage and C denote crest damage, we immediately have $P(B\overline{C}) = 30/300 = 1/10$, $P(C\overline{B}) = 50/300 = 1/6$ and $P(BC) = 10/300 = 1/30$. Using a Venn diagram, it is immediately obvious that $P(B) = 40/300 = 2/15$ and $P(C) = 60/300 = 1/5$. Therefore $P(BC) \neq P(B)P(C)$ and we conclude that beak damage and crest damage are dependent. Since $P(BC) \neq 0$, this implies $BC \neq \phi$, and therefore events B and C are not mutually exclusive.

Example 3.16 If we scramble the letters ROTTNOO, what is the probability that the word TORONTO is spelled?

Solution: Using the symmetry definition of probability, the denominator consists of the 7! ways in which the 7 letters can be permuted. In the numerator, we want the number of these permutations that lead to the spelling of TORONTO. We note that the letters R and N must be fixed in the third and fifth positions respectively. We then note that the three O's can be permuted in 3! ways and the two T's can be permuted in 2! ways. This gives the probability $3!\ 2!/7! = 1/420$.

Example 3.17 Consider a bag containing four red marbles and six black marbles. Three marbles are drawn from the bag. What is the probability that all three marbles are red when the marbles are drawn

(a) with replacement?

(b) without replacement?

Solution: When a marble is chosen **with replacement**, this means that after the marble is drawn, it is then replaced into the bag. Since the composition of the bag of marbles remains the same on each of the three draws, the outcome of each draw is independent of the other draws, and the probability of three red marbles is therefore $(4/10)(4/10)(4/10) = 0.064$. Sampling with replacement is sometimes done in fisheries management. Fish are caught, characteristics are measured and the fish are returned to the lake.

It is possible to catch the same fish on several occasions. These are the less intelligent fish. There is an entire area of statistical research known as **capture-recapture** which involves the tagging and subsequent recording of observations with respect to various animal species. At SFU, Professor Carl Schwarz (who many people think resembles R.A. Fisher) is an expert in capture-recapture.

When a marble is chosen **without replacement**, this means that after the marble is drawn, it is not returned to the bag. The probability of obtaining three red marbles under this sampling scheme is therefore $\binom{4}{3}/\binom{10}{3} = 0.03\overline{3}$ where $\binom{4}{3}$ is the number of ways that three marbles can be chosen from the four red marbles. Note that the probability under sampling without replacement is roughly half the probability under sampling with replacement. Sampling without replacement is common in destructive testing. For example, suppose that you obtain a sample of steel rods and subject them to a stress test. You apply an increasing level of stress until the rods break. You then record the stress levels. Obviously, you cannot put broken rods back into the population of rods.

Example 3.18 Repeat Example 3.17 but change the composition of the bag from 10 marbles to 100 marbles where 40 marbles are red and 60 marbles are black.

Solution: The interesting aspect of the followup question is that the ratio of red balls to black balls remains the same. Using the same logic as in Example 3.17, the probability under sampling with replacement is $(40/100)(40/100)(40/100) = 0.064$. The probability under sampling without replacement is $\binom{40}{3}/\binom{100}{3} = 0.061$. In going from Example 3.17 to Example 3.18, the probabilities under sampling with replacement did not change. This makes sense as the composition of the bag is always the same from draw to draw. However, there was a dramatic change in probabilities under sampling without replacement. Moreover, sampling without replacement nearly resembles sampling with replacement in Example 3.18. This is because the sample size 3 is small relative to the population size 100, and therefore, removing a few marbles from the bag does not greatly affect the overall composition of marbles.

Typically, the probability calculations in sampling with replacement are simpler than the calculations in sampling without replacement. Accordingly, sampling with replacement calculations are sometimes carried out in situations involving sampling without replacement. The approximations are generally acceptable when the population size is large and the sample size is moderate.

Example 3.19 Coincidences are often misunderstood. This is not so much an example but more of a discussion on misperceptions involving coincidences. Consider the following example taken from Plous (1993): *The Psychology of Judgment and Decision Making* (McGraw Hill). Richard Baker left a shopping mall, found what he thought was his car, and drove away. Later, he realized that it was the wrong car and he returned to the parking lot. The car belonged to another Mr. Baker who had the same car, with an identical key! The police estimated the odds of the event at one million to one.

The first comment that I have regarding this remarkable example is that the calculation by the police is suspect since the event in question is typically ill-defined. Rather than calculating the probability of the exact event happening (which is not easy to do), one should instead consider similar events happening sometime, somewhere to someone since we would be equally astonished if a similar event happened anytime, anywhere to anyone.

Consider a second example where twins separated at birth are united as adults, and are amazed to find that they have striking characteristics in common (e.g. children have the same names, they use the same toothpaste, they have the same job, etc.). My question is whether they should be so amazed at these coincidences. Suppose that any characteristic has probability 1/50 of a match. Then, in the course of a discussion over several hours where 100 characteristics are mentioned, we would expect the two individuals to share roughly two characteristics.

One of the messages here is that coincidences are not as improbable as they may appear. Coincidences happen all of the time. I am sure that you have your own personal experiences involving coincidence. Send me an email (tim@stat.sfu.ca) if you have an amusing anecdote which you care to share.

3.7 Exercises

3.01 In 2002, the first round of the NBA playoffs consisted of a best-of-five playoff series. In a best-of-five series, the series concludes when one of the two teams has won three games, and each game results in a win for one of the two teams. In a best-of-five series involving a hypothetical team, write down the relevant sample space. Explain whether the symmetry definition of probability is appropriate with respect to the outcomes in the sample space.

3.02 Referring to Exercise 3.01, what is the cardinality (size) of the sample space in a best-of-seven playoff series?

3.03* Let y be a woman's age and let x be the age of her biological daughter (both measured in years). Write down the corresponding sample space.

3.04 Draw a Venn diagram involving sets A, B and C where A and B are mutually exclusive, B and C are disjoint, and $A \cap C \neq \phi$.

3.05 Provide a simplification of $(A \cup B \cup C) \cap (\overline{A}B \cup \overline{B}C \cup A\overline{C})'$.

3.06 If A and C are mutually exclusive, draw the Venn diagram that contains the sets A, B and C.

3.07* Suppose $A \subseteq B$. Using the three axioms of probability, establish that $P(B) \geq P(A)$.

3.08 Suppose that A and B are disjoint and $P(\overline{A}) = 0.63$. What are the minimum and maximum values of $P(B)$?

3.09 If $P(ABC) = 0.2$, are A and C mutually exclusive? Explain.

3.10 Consider the statement "the probability that an individual wins the lottery is one in a million". (a) Provide an interpretation for the statement in terms of the symmetry definition of probability. If you are unable to do so, explain. (b) Provide an interpretation for the statement in terms of the frequency definition of probability. If you are unable to do so, explain.

3.11 Consider the statement "the probability that the Dow Jones Industrial average exceeds the 13,000 barrier by year end is 0.4. (a) Interpret the statement using the symmetry definition of probability. If you are unable to do so, explain. (b) Provide an interpretation for the statement in terms of the frequency definition of probability. If you are unable to do so, explain.

3.12 Suppose $P(\overline{A} \cap \overline{B}) = 0.2$ and $P(\overline{A}) = 0.3$. Calculate $P(A \mid A \cup B)$.

3.13 Suppose A and B are disjoint, B and C are disjoint, $P(A) = 0.3$, $P(B) = 0.2$, $P(C) = 0.4$ and $P(A \cup C) = 0.5$. Calculate $P(A \mid B \cup C)$.

3.14* Suppose $P(A) \neq 0$, $P(B) \neq 0$ and $P(A \mid B) = P(A)$. Prove that $P(B \mid A) = P(B)$.

3.15 An automobile manufacturer is concerned about a possible recall of its four-door sedan. If there were a recall, there is probability 0.25 of a defect in the brake system, 0.18 of a defect in the transmission, 0.17 of a defect in the fuel system, and 0.40 of a defect elsewhere. (a) What is the probability that the defect is in the brakes or in the fueling system if the probability of defects in both systems simultaneously is 0.15? (b) What is the probability that there are no defects in either the brakes or the fueling system?

3.16* A town has two fire engines operating independently. The probability that a specific engine is available when needed is 0.96. What is the probability that neither is available when needed?

3.17 An allergist claims that 50% of her patients are allergic to weeds. What is the probability that (a) exactly three of her next four patients are allergic to weeds? (b) none of her next four patients are allergic to weeds?

3.18 There is a 35% chance that the queen carries the gene of hemophilia. If she is a carrier, then each prince independently has a 50-50 chance of having hemophilia. If the queen is not a carrier, her offspring will not have the disease. Suppose that the queen has three princes, none of whom has the disease. What is the probability that the queen is a carrier?

3.19* Referring to Example 3.13, consider a deck of playing cards where the face cards have been removed (i.e. no kings, queens or jacks). Four cards are dealt. Calculate the probability of obtaining a pair.

3.20 Referring to Example 3.13, consider a deck of playing cards where the 2's, 3's, 4's and 5's have been removed. If five cards are drawn, calculate the probability of obtaining a **full house** (i.e. three cards of one denomination and two cards of another denomination).

3.21 A machine has six switches. The probability that any particular switch works properly is 0.98. Assuming independent operation of the switches, calculate the probability that at least one switch fails to work properly.

3.22 A bag contains five blue marbles, three white marbles and two red marbles. If six marbles are randomly drawn without replacement, what is the probability of obtaining three blue, two white and one red marble?

3.23 When a show comes to town, there are friday, saturday and sunday performances. From historical records, the probability that the three performances are sold-out are 0.16, 0.13 and 0.09 respectively. Assuming independence of the shows, what is the probability that there is ticket availability for at least one show?

3.24* A particular disease has an incidence rate of 1 in 1000 adults. A test shows a positive result 99% of the time when an individual has the disease and a positive result 4% of the time when an individual is healthy. For a randomly selected adult, what is the probability of a negative result?

3.25 It is believed that one percent of children have autism. A test for autism has been developed whereby 90% of autistic children are correctly identified as having autism but 3% of non-autistic children are incorrectly identified as having autism. A child is tested at two independent clinics. What is the probability that the two clinics have the same diagnosis?

3.26 A clinical cohort consists of seven sets of girl/boy twins (i.e. each twin pair has a brother and a sister). Five of the fourteen subjects are randomly selected. (a) What is the probability that the selected group is neither all-male nor all-female? (b) What is the probability that the selected group contains no siblings?

3.27 Five teachers have taught for 3, 6, 7, 10 and 14 years respectively. If two of the teachers are randomly chosen, what is the probability that the pair have at least 15 years of total teaching experience?

3.28 Referring to the birthday problem in Example 3.9, suppose that there are five people in a room. Provide an upper bound for the probability that at least two people share a birthday or that their birthday is separated by at most one day. Don't give an upper bound of 1.0.

3.29 Referring to the birthday problem in Example 3.9, assume that it is equiprobable to have a birthday in any month of the year. Suppose that there are six people in a room. (a) What is the probability that at least two people have a birthday in the same month? (b) What is the probability that three of the people have birthdays in one month, and the other three have birthdays in another month?

3.30 Suppose that you roll two dice. What is the probability that the two dice are equal or differ by at most one?

3.31 A coin is flipped until a head appears. What is the probability that a head appears on an odd-numbered flip?

3.32 Consider a game where I roll a die. If it is a 1, I win. If it is a 2 or 3, I lose. Otherwise, I continue to roll the die until I either obtain the original result from the first roll or I roll a 1 or 2 or 3. If it is the original result, I win. If it is a 1 or 2 or 3, I lose. What is the probability that I win the game?

3.33 In a city, the probability that a person owns a dog is 0.28. Calculate the probability that the tenth person randomly interviewed in the city is the fifth to own a dog.

3.34 When I had just graduated from university, I went on a European vacation for one month, travelling to many of the major sites. I was amazed to bump into someone that I knew from my university. Using a reasoned argument, try to assign a ballpark probability to this coincidence.

Chapter 4

Discrete Distributions

In this course, we are interested in distributions that are either discrete or continuous. In fact, there are also distributions that are of a mixed type. In this chapter, we look at the simpler discrete distributions and we study topics that are related to discrete distributions. It is a good idea to pay close attention to these topics, as they carry over to the continuous distributions studied in Chapter 5.

4.1 Random Variables

Random variables are intrinsic to statistics and have the following simple definition: A random variable is a function of the sample space.

We denote random variables by capital letters such as X, Y and Z, and we make sure not to confuse them with **algebraic variables** which you see, for example, in the quadratic equation $ax^2 + bx + c = 0$. Sometimes the word random variable is abbreviated as "rv". However, for our purposes, an rv is not a recreational vehicle.

Example 4.1 Suppose that a coin is flipped three times and we define the random variable X as the number of heads. As in Example 3.1, the sample space is $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. We note that the random variable X is a function of the sample space since it defines a mapping from the set S to the set $\{0, 1, 2, 3\}$. For example, $X(HHH) = 3$, $X(HHT) = 2$, and so on.

As in Example 4.1, when the **range** of the random variable is discrete, then the random variable is said to be **discrete**. In particular, when the range is $\{0, 1\}$, then the random variable is said to be **Bernoulli**. Bernoulli random variables have an important role in applied statistics. For example, whether or not a patient recovers, whether or not a customer likes a new product and whether or not a a questionnaire is returned are all examples of situations that involve Bernoulli random variables.

Example 4.2 Consider the maximum temperature in degrees celsius in Vancouver on February 15, 2006. The sample space for the experiment is the set $S = \{s : -270 \leq s < \infty\}$. Note that even though S is continuous, we may define a discrete random variable on S. For example,

$$Y = \begin{cases} 1 & \text{if } s \leq 0 \\ 0 & \text{if } s > 0 \end{cases}$$

defines a Bernoulli random variable Y corresponding to whether the temperature is freezing. Note that Y is a function of the sample space S.

The following definition is a bit abstract and is a little difficult to digest. However, it gives you a better framework for understanding random variables. At a certain point, you can forget the definition.

The **probability mass function** or **pmf** of a discrete random variable X is

$$p_X(x) \;=\; P(s \in S : X(s) = x). \tag{4.1}$$

Let's try to explain the beast (4.1). First, to denote that the pmf p_X corresponds to the random variable X, the subscript X is introduced. However, it is often understood that we are discussing the rv X, and therefore p is sometimes used instead of p_X. An interpretation of (4.1) is that p_X is "the probability" that the random variable X equals x, and this is why x is an argument of the function p_X. Note however that the probability $p_X(x)$ is "induced" from the original probability measure P corresponding to the sample space. In words, $p_X(x)$ is the totality of all probability P given to elements s of the sample space such that $X(s) = x$.

Example 4.3 We now return to Example 4.1 which involves flipping a coin three times and let the random variable X denote the number of heads. Assume that the probability measure P on the sample space is the symmetric probability which gives equal weight to each of the 8 outcomes. Then using definition (4.1), the probability mass function is $p(x)$ where

$$p(1) \;=\; P(s \in S : X(s) = 1) \;=\; P(\{HTT, THT, TTH\}) = 3/8.$$

Verify for yourself that $p(0) = 1/8$, $p(2) = 3/8$ and $p(3) = 1/8$.

It turns out that a probability mass function $p(x)$ satisfies two intuitive properties:

1. $p(x) \geq 0$ for all x
2. $\sum_x p(x) = 1$

where the sum in the second property is taken over the range of X. Therefore, if you are given a function $p(x)$ and it does not satisfy both of the above properties, then $p(x)$ is not a probability mass function. Also, there is a converse result: if a function $p(x)$ satisfies the above two properties, then it is a probability mass function. Note that the converse result mentions neither the sample space nor the original probability measure. In fact, this is how we are going to treat pmf's. Given a pmf $p(x)$, we don't spend much time thinking about definiton (4.1).

Example 4.4 This is a continuation of Example 3.5. Let the random variable X be the sum of two dice. We obtain the probability mass function $p(x)$ as shown

x	2	3	4	5	6	7	8	9	10	11	12
p(x)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

and note that the two properties of pmfs are satisfied. I am sure that the results of this exercise coincide with your experiences from playing games with dice. The outcome $X = 7$ is the most common, and the probabilities tail off symmetrically with outcomes $X = 2$ and $X = 12$ being the least common.

Example 4.5 Consider a batter in baseball with a 300 average. Let's assume that this means that the batter gets a hit with probability 0.3. Clearly, this is not correct since the probability of getting a hit depends on many factors including the pitcher, the ballpark, etc. However, lets stick with the simplification and note that it is often useful to approximate complex systems with simple models that are slightly wrong. We define the random variable X as the number of "at bats" until the batter gets a hit. In addition, we make the independence assumption. It follows that the pmf of X is given by

$$p(x) \;=\; (.7)^{x-1}(.3) \quad \text{for } x = 1, 2, \ldots.$$

We easily verify that the two properties of pmfs hold. Note that $p(x) \geq 0$ for all $x = 1, 2, \ldots$ and

$$\sum_x p(x) \;=\; \sum_{x=1}^{\infty}(.7)^{x-1}(.3) \;=\; (.3)\sum_{x=0}^{\infty}(.7)^x \;=\; (.3)\left(\frac{1}{1-.7}\right) \;=\; 1$$

using the summation rule for infinite sums.

We remind ourselves one more time that pmfs are used to describe probability with respect to random variables. Having introduced the pmf $p_X(x)$ for the discrete random variable X, we now introduce the corresponding cumulative distribution function. The **cumulative distribution function** or **cdf** of the discrete random variable X is given by

$$\boxed{F_X(x) \;=\; P(X \leq x) \;=\; \sum_{y \leq x} p_X(y)} \tag{4.2}$$

where again, we often drop the subscript X when it is understood. The cdf $F_X(x)$ in (4.2) represents the probability of realizations of the rv X up to and including x.

Example 4.6 We return to the coin flipping experiment of Example 4.3 where the random variable X is the number of heads. Here, the cdf of X is given by

$$F(x) \;=\; \begin{cases} 0 & x \in (-\infty, 0) \\ 1/8 & x \in [0, 1) \\ 4/8 & x \in [1, 2) \\ 7/8 & x \in [2, 3) \\ 1 & x \in [3, \infty) \end{cases} \quad .$$

Note that the cdf in Example 4.6 is not only defined for the integer values $x = 0, 1, 2, 3$ but for all $x \in (-\infty, \infty)$. In fact, all cdfs $F(x)$ possess the following properties:

- $F(x)$ is normed (i.e. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$)
- $F(x)$ is a non-decreasing function
- $F(x)$ is right continuous (i.e. $\lim_{x \to x_0^+} F(x) = x_0$)

As an exercise, try to construct a plot of the cdf $F(x)$ in Example 4.6.

4.2 Expectations

Sometimes it is convenient to report characteristics of random variables. Expectations are useful in this regard. The **expected value** or **expectation** of a discrete random variable X is defined as

$$\boxed{E(X) \;=\; \sum_x x p_X(x)}$$

where the summation is over the range of X. Sometimes we use the notation μ_X for expectation or the simpler notation μ when X is understood. The interpretation of $E(X)$ is based on the frequency definition of probability. Consider hypothetical repetitions of the underlying experiment where each experiment gives rise to a realized value of X. In the long run (i.e. over many repetitions of the experiment), the average of the realized values is $E(X)$. To motivate the interpretation, imagine repeating the experiment

N times and denote the range of X by $x_1, x_2, \ldots$. Then we would observe approximately $Np(x_1)$ values of x_1, approximately $Np(x_2)$ values of x_2, and so on. Therefore, the average of the realized values over the N replications of the experiment is roughly

$$\frac{x_1 Np(x_1) + x_2 Np(x_2) + \cdots}{N} \;=\; x_1 p(x_1) + x_2 p(x_2) + \cdots \;=\; E(X).$$

Accordingly, we sometimes refer to the expected value $E(X)$ as the **mean** or **population mean** of X.

Example 4.7 This is a continuation of Example 4.3. Calculate the expected value of the rv X where X is the number of heads in three flips of a coin.
Solution: The expected value is

$$\begin{aligned} E(X) \;=\; \textstyle\sum_{x=0}^{3} xp(x) &= 0p(0) + 1p(1) + 2p(2) + 3p(3) \\ &= 0(\tfrac{1}{8}) + 1(\tfrac{3}{8}) + 2(\tfrac{3}{8}) + 3(\tfrac{1}{8}) \\ &= 1.5. \end{aligned}$$

Note that in this example, the long run average $\mu = E(X) = 1.5$ is not a value in the range of X. Remember that μ represents the long run average of X under hypothetical repetitions of the experiment. The interpretation here is that if you flipped three coins many times, the average number of heads that you would observe on a triplet of flips is 1.5.

The concept of expectation can be extended to functions of random variables. For a discrete random variable X, the expectation of the function $g(X)$ is given by

$$\boxed{E(g(X)) \;=\; \sum_x g(x) p_X(x)}$$

where the summation is taken over the range of X. Note that a function of a random variable is itself a random variable.

Example 4.8 Suppose that we roll a die and define X as the value appearing on the die. Then if $g(X) = X^2$,

$$E(g(X)) \;=\; \sum_{x=1}^{6} x^2 p_X(x) \;=\; 1^2(1/6) + 2^2(1/6) + \cdots + 6^2(1/6) \;=\; 91/6.$$

The interpretation refers to hypothetical rolls of the die where X^2 is recorded on each roll. In the long run (i.e. after an infinite number of repetitions), the average value of X^2 is 91/6.

It turns out that linear combinations of random variables are often of interest. It is worth committing the following formula to memory where a and b are constants:

$$\boxed{E(aX+b) = aE(X)+b.} \tag{4.3}$$

The result (4.3) is derived via

$$\begin{aligned} E(aX+b) &= \textstyle\sum_x (ax+b)p(x) \\ &= a\textstyle\sum_x xp(x) + b\sum p(x) \\ &= aE(X)+b. \end{aligned}$$

Another useful formula concerns the expectation of a sum of random variables. Referring to first principles (as we did in establishing (4.3)), obtain the following as an exercise:

$$\begin{aligned} & E(g_1(X)+g_2(X)+\cdots+g_k(X)) \\ = \; & E(g_1(X))+E(g_2(X))+\cdots+E(g_k(X)). \end{aligned} \tag{4.4}$$

I think it is time for a break; things are beginning to get a bit dull. So, a guy goes in to see his doctor. The doctor evaluates the patient and says, “I have bad news - you have Alzheimer’s disease and you have cancer”. The guy looks back at his doctor and says, “At least I don’t have Alzheimer’s.”

Ok, lets get back to business. The previous discussion leads to a prominent expectation $E(g(X))$ which is obtained when we choose $g(X) = (X - E(X))^2$. The **variance** or **population variance** of a discrete random variable X is defined as

$$\boxed{Var(X) \;=\; E((X-E(X))^2).} \tag{4.5}$$

Sometimes we use the notation σ_X^2 for variance or the simpler notation σ^2 when X is understood. Whereas $E(X)$ is a characteristic of the random variable X describing centrality, $Var(X)$ describes spread or dispersion in the random variable X. The quantity σ_X is referred to as the **population standard deviation** of X. The expression (4.5) is instructive as to why variance is a measure of spread since expectations are taken with respect to squared distances between X and its mean $E(X)$. However, it is often more convenient to use the following result when calculating variances:

$$\boxed{Var(X) = E(X^2) - (E(X))^2.} \tag{4.6}$$

We obtain (4.6) by expanding (4.5) via

$$\begin{aligned} Var(X) &= E((X - E(X))^2) \\ &= E(X^2 + (E(X))^2 - 2XE(X)) \\ &= E(X^2) + E((E(X))^2) - E(2XE(X)) \\ &= E(X^2) + (E(X))^2 - 2E(X)E(X) \\ &= E(X^2) + (E(X))^2 - 2(E(X))^2 \\ &= E(X^2) - (E(X))^2. \end{aligned}$$

Make sure that you are comfortable with all of the operations in the above derivation. Some of the steps rely on (4.3) and (4.4). Remember also that $E(X)$ is a constant.

Example 4.9 We return to the coin flipping experiment of Example 4.3 where the random variable X is the number of heads. From Example 4.7, $E(X) = 1.5$ and $E(X^2) = 0^2(1/8) + 1^2(3/8) + 2^2(3/8) + 3^2(1/8) = 3.0$. Therefore $\sigma^2 = 3.0 - (1.5)^2 = 0.75$ and $\sigma = 0.866$.

Like expression (4.3) concerning the mean of a linear combination, there is an expression for the variance of a linear combination:

$$\boxed{Var(aX + b) = a^2 Var(X).} \tag{4.7}$$

The result (4.7) is derived via

$$\begin{aligned} Var(aX + b) &= E((aX + b - E(aX + b))^2) \\ &= E((aX + b - aE(X) - b)^2) \\ &= E((aX - aE(X))^2) \\ &= a^2 E((X - E(X))^2) \\ &= a^2 Var(X). \end{aligned}$$

Note that the right side of (4.7) does not depend on b. The intuition behind this is that the constant b moves the probability mass function left or right depending on whether b is negative or positive. Shifting the pmf changes the mean but does not change the spread.

Going back to the study of descriptive statistics in Chapter 2, we see that there is an analogy between the sample mean $\bar{x}$ of the dataset $x_1, x_2, \ldots, x_n$ and the population mean $E(X)$ of the random variable X. Although both are measures of centrality, the former describes a finite dataset whereas the latter describes a hypothetical population. Similarly, there is an analogy between the sample variance s^2 (sample standard deviation s) of the dataset $x_1, x_2, \ldots, x_n$ and the population variance σ^2 (population standard deviation σ) of the random variable X.

Example 4.10 This example uses a few of the recently developed expressions. Consider the random variable X with pmf

x	4	8	10
p(x)	0.2	0.7	0.1

Verify for yourself that $E(X) = 7.4$, $E(X^2) = 58.0$, $\sigma = 1.8$ and $E(3X + 4X^2) = 254.2$.

Example 4.11 Consider a game of chance where you bet x dollars. With probability $p < 1$, you win y dollars. What is the value of x for this to be a fair game?

Solution: Hmmm, interesting question. As if often the case, introducing notation is helpful. Let the random variable W denote winnings where W has the pmf $p(w)$ given by

w	$-x$	y
p(w)	$1-p$	p

We then need an understanding of what it means for a game to be **fair**. Using the frequency definition of probability, we have a notion of hypothetical repetitions of an experiment. A game is fair when the long term average winnings under hypothetical repetitions of the experiment is zero (i.e. neither a long term average profit nor a long term average loss). In other words, a fair game requires $E(W) = 0$. Therefore

$$0 \;=\; E(W) \;=\; -x(1-p) + yp$$

which gives $x = yp/(1-p)$.

4.3 The Binomial Distribution

In a first course in probability and statistics, it turns out that **distribution** is not so easy to define. Without being precise, the distribution of a random variable X refers to the probabilities associated with X. For example, the probability mass function of a discrete random variable X determines its distribution.

Perhaps the most important discrete distribution is the **binomial distribution**. We say that a discrete random variable X has the binomial(n, p) distribution if it has pmf

$$\boxed{p(x) \;=\; \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}} \tag{4.8}$$

where n is a positive integer and $0 < p < 1$. It is immediate that (4.8) satisfies the two properties of pmfs since $p(x) > 0$ for $x = 0, 1, \ldots, n$ and $\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p + (1-p))^n = 1$ by the binomial theorem. Note that the binomial distribution is not a single distribution, but rather, it is a **family** of distributions where the family is indexed by the **parameters** n and p. Choosing a particular n and a particular p determines a member distribution of the binomial family. In the simplest case, when $n = 1$, the binomial distribution is referred to as a **Bernoulli distribution** where the corresponding Bernoulli random variable X has only two outcomes (i.e. $X = 0$ and $X = 1$) with probabilities $p(0) = 1 - p$ and $p(1) = p$.

We now motivate the binomial distribution given by (4.8). First, do not confuse the parameter n in the binomial distribution with n used to denote sample size in Chapter 2. Consider n trials where the probability of success in each trial is p and the n trials are independent of one another. We are interested in the probability of obtaining x successes. If we denote a success by S and a failure by F, then the probability of getting the specific sequence

$$\underbrace{SS\cdots S}_{x \text{ successes}} \qquad \underbrace{FF\cdots F}_{n-x \text{ failures}}$$

is $p^x(1-p)^{n-x}$ since the trials are independent. However, we are not concerned with order. Since there are $\binom{n}{x}$ ways that the trial outcomes can be ordered, we obtain the pmf in (4.8).

Example 4.12 You roll a die 10 times and you are interested in the outcomes 5 and 6. What is the probability that x of the rolls result in 5's or 6's?
Solution: There are $n = 10$ trials (i.e. tosses of the die) that are independent of one another where each trial has probability 1/3 of a success (i.e. either a 5 or a 6). Therefore the binomial$(10, 1/3)$ distribution is appropriate and the probability is

$$p(x) \;=\; \binom{10}{x}(1/3)^x(2/3)^{10-x}$$

where $x = 0, 1, \ldots, 10$.

When we write $X \sim$ binomial(n, p), this states that the random variable X is distributed according to the binomial(n, p) distribution. The mean and the variance of the binomial(n, p) distribution are given by $\boxed{E(X) = np}$ and $\boxed{Var(X) = np(1-p)}$. You should commit these expressions to memory. The derivation of the mean is given by

$$\begin{aligned}
E(X) &= \textstyle\sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x} \\
&= \textstyle\sum_{x=1}^{n} x\binom{n}{x}p^x(1-p)^{n-x} \\
&= \textstyle\sum_{x=1}^{n} \frac{n!}{(n-x)!(x-1)!}\, p^x(1-p)^{n-x} \\
&= np\textstyle\sum_{x=1}^{n} \frac{(n-1)!}{(n-x)!(x-1)!}\, p^{x-1}(1-p)^{n-x} \\
&= np\textstyle\sum_{y=0}^{n-1} \binom{n-1}{y}p^y(1-p)^{n-1-y} \\
&= np
\end{aligned}$$

where we change variables $y = x - 1$ in the fifth equality and we recognize that the term inside the summation sign in the fifth equality is the pmf of the binomial$(n-1, p)$ distribution. Try to derive the variance expression. As a hint, first obtain $E(X^2) = np((n-1)p+1)$.

Something that you *must* be able to do is determine when a binomial(n, p) distribution is appropriate for modelling. This is done by assessing whether the three assumptions of the binomial distribution are satisfied or approximately satisfied, as is more often the case. Since the three assumptions are so important, we repeat them here:

1. there are n trials each resulting in either success or failure
2. the trials are independent of one another
3. each trial has the same probability of success p

As mentioned earlier, the binomial distribution is an important distribution which has widespread use in statistical applications. We now consider a few examples involving the binomial distribution.

Example 4.13 Suppose that only 20% of all drivers come to a complete stop at an intersection having flashing red lights in all directions when no other cars are visible. Of the next 20 drivers that arrive at the intersection, how many do you expect to come to a complete stop?
Solution: The $n = 20$ trials correspond to the drivers, and we note that the drivers are independent of one another. Since the 20 drivers are a random sample from the population of drivers, we assume that each has probability $p = 0.2$ of coming to a complete stop. Therefore, letting X be the number of the 20 drivers that come to a complete stop, we have $X \sim$ binomial$(20, 0.2)$ and we expect $E(X) = np = 20(0.2) = 4$ to come to a complete stop.

Example 4.14 This problem is closely related to Example 4.5. A baseball player with a 300 average has 600 at bats in a season. What is the probability that he has at least 200 hits? Achieving 200 hits in a season is a benchmark of great achievement in Major League Baseball.

Solution: Let Y be the number of hits and lets unrealistically assume that the batter never walks. In this problem, we have $n = 600$ trials each resulting in success (i.e. a hit) or failure (i.e. an out). However, as discussed in Example 4.5, the assumption of the constant probability of success $p = 0.3$ is doubtful since game conditions affect the probability of a hit. The independence assumption may also be questionable as many people believe in **streakiness**. A streaky batter has "hot" periods where the probability of getting a hit following periods of success is higher than the probability of getting a hit following periods of failure. Putting these serious reservations aside, we have $Y \sim \text{binomial}(600, 0.3)$ and the probability of interest is

$$P(Y \geq 200) = \sum_{y=200}^{600} \tbinom{600}{y} (0.3)^y (0.7)^{600-y}. \tag{4.9}$$

The calculation of the probability in (4.9) brings up another difficulty which we address in Section 4.4 and in Section 5.2.1. For now, we simply note that the terms $\tbinom{600}{y}$ may be huge and cause our calculator to overflow. Similarly, the terms $(0.3)^y(0.7)^{600-y}$ are miniscule and cause an underflow. However, the product of the two terms often gives a reasonable number. Finally, even if the overflow/underflow problems are avoidable, there are 401 terms in the sum, too many to add by hand.

4.4 The Poisson Distribution

I suggested earlier that the binomial distribution is the most important of the discrete distributions. Running a close second on the list is the **Poisson distribution**. By saying that a distribution is important, I mean that the distribution has wide applicability in real world problems. In addition to being important, the Poisson is quite a fishy distribution.

A random variable X has a Poisson(λ) distribution if it has the pmf

$$\boxed{p(x) \; = \; \begin{cases} \lambda^x \exp\{-\lambda\}/x! & x = 0, 1, \ldots \\ 0 & \text{otherwise} \end{cases}} \tag{4.10}$$

where $\lambda > 0$. Like the binomial, the Poisson is not a single distribution, but rather, it is a family of distributions. The Poisson family is indexed (or characterized) by the parameter λ which is a constant term in (4.10). Unlike the binomial which is defined on a finite set, the Poisson is defined on the non-negative integers which is an infinite set. The Poisson distribution is especially good at modelling rare events (more later).

You should verify for yourself that $p(x)$ in (4.10) is indeed a pmf (see Section 4.1). Also, for $X \sim \text{Poisson}(\lambda)$, you should commit to memory the mean and variance expressions $\boxed{E(X) = Var(X) = \lambda}$. The derivation of the mean is given by

$$\begin{aligned}
E(X) &= \textstyle\sum_{x=0}^{\infty} x\lambda^x \exp\{-\lambda\}/x! \\
&= \lambda \exp\{-\lambda\} \textstyle\sum_{x=1}^{\infty} x\lambda^{x-1}/x! \\
&= \lambda \exp\{-\lambda\} \textstyle\sum_{x=1}^{\infty} \lambda^{x-1}/(x-1)! \\
&= \lambda \exp\{-\lambda\} \textstyle\sum_{y=0}^{\infty} \lambda^{y}/y! \\
&= \lambda \exp\{-\lambda\} \exp\{\lambda\} \\
&= \lambda
\end{aligned}$$

where we change variables $y = x - 1$ in the fourth equality and we recognize the Taylor series expansion of $\exp\{\lambda\}$ in the fourth equality.

Recall from Example 4.14 that binomial probabilities can sometimes be difficult to calculate. We now demonstrate that Poisson probabilities can often provide good approximations to binomial probabilities. Let's first state the result simply: **the binomial(n, p) pmf is approximately equal to the Poisson(np) pmf when n is much larger than np.** This is why we stated earlier that the Poisson distribution is particular good at modelling rare events. To establish the result, let $p_X(x)$ be the pmf of $X \sim \text{binomial}(n, p)$ and let $p_Y(x)$ be the pmf of $Y \sim \text{Poisson}(np)$. Consider the limit of $p_X(x)$ as $n \to \infty$ with np fixed. This may seem like a strange type of limit; it essentially means that n is getting big and p is getting small in such a way that np remains constant. Then

$$\begin{aligned}
\lim p_X(x) &= \lim \tbinom{n}{x} p^x (1-p)^{n-x} \\
&= \lim \tfrac{n^{(x)}}{x!}\tfrac{n^x}{n^x}\, p^x \left(1-\tfrac{np}{n}\right)^{n-x} \\
&= \lim (np)^x \tfrac{n^{(x)}}{n^x}\; \left(1-\tfrac{np}{n}\right)^n \left(1-\tfrac{np}{n}\right)^{-x}/x! \\
&= (np)^x \exp\{-np\}/x! \\
&= p_Y(x)
\end{aligned}$$

where the fifth equality uses $\lim_{n\to\infty} n^{(x)}/n^x \to 1$, the definition of the exponential function and the fact that np is fixed in the limit. Perhaps it is not surprising that $E(Y) = E(X) = np$ and that $Var(Y) = np \approx Var(X) = np(1-p)$ when p is small.

Example 4.15 A rare type of blood occurs in a population with frequency 0.001. If n people are tested, what is the probability that at least two people have this rare blood type?

Solution: Let X be the number of people with the rare blood type where $X \sim \text{binomial}(n, 0.001)$. The probability of interest is

$$\begin{aligned}
P(X \geq 2) &= 1 - P(X=0) - P(X=1) \\
&= 1 - (0.999)^n - n(0.001)(0.999)^{n-1}.
\end{aligned} \tag{4.11}$$

We compare the binomial probability $P(X \geq 2)$ in (4.11) with the Poisson approximation $P(Y \geq 2)$ where $Y \sim \text{Poission}(n(0.001))$ and

$$\begin{aligned}
P(Y \geq 2) &= 1 - P(Y=0) - P(Y=1) \\
&= 1 - \exp\{-0.001n\} - (0.001n)\exp\{-0.001n\}.
\end{aligned} \tag{4.12}$$

The comparison of the binomial probability (4.11) with the Poisson probability (4.12) is given in Table 4.1 for $n = 100$ and $n = 1000$. For both of these cases, n is large relative to np where $p = 0.001$, and we therefore expect an adequate approximation. In fact, we have agreement in the probabilities to 7 decimal places when $n = 1000$.

	$P(X \geq 2)$	$P(Y \geq 2)$
$n = 100$	0.0046381	0.0046788
$n = 1000$	0.2642411	0.2642411

Table 4.1: Comparison of the binomial probability in (4.11) with the Poisson probability in (4.12).

4.4.1 The Poisson Process

We have seen that the binomial distribution is applicable when there are n independent trials with a constant probability of success. This is a common situation. However, when we look at the "weird" pmf of the Poisson distribution (4.10), it is not obvious what sorts of real world situations lead to a Poisson distribution. We now look at some conditions under which a Poisson distribution may be appropriate. The conditions are not as easily assessed as in the binomial case but they should at least provide some guidance for modelling. As the topic is somewhat more difficult, I will not test you in depth on this subsection.

For our development of the Poisson process, imagine events occurring in time. For example, there may be cars that pass a section of highway at different times. Let $X_{(t_1,t_2)}$ denote the number of successes (i.e. the number of passing cars) in the interval (t_1, t_2) with the corresponding probability $p(x, t_1, t_2)$. In our example, $p(x, t_1, t_2)$ is the probability that x cars pass the section of highway in the time interval (t_1, t_2). We now consider the three assumptions of the Poisson process:

1. The events $X_{(t_1,t_2)}$ and $X_{(t'_1,t'_2)}$ are independent if (t_1, t_2) and (t'_1, t'_2) are non-overlapping intervals (i.e. $t'_1 > t_2$ or $t_1 > t'_2$). The idea is that events occurring in separate periods of time do not affect one another.

2. The events $X_{(t_1,t_2)}$ are stationary for all intervals (t_1, t_2). Stationarity means that the probability of the event $X_{(t_1,t_2)}$ depends on the interval (t_1, t_2) only through its length $t = t_2 - t_1$. Therefore, under stationarity,

we can write $p(x, t_1, t_2) = p(x, t_2 - t_1)$. Intuitively, stationarity states that probabilities associated with the Poisson process do not change over the time continuum.

3. In a small interval of time Δt, the probability of exactly one event occurring is $p(x, \Delta t) = \lambda(\Delta t)$ and the probability of zero events occurring is approximately $1 - p(x, \Delta t)$. The assumption states that in a small time period Δt, the probability of more than one event occurring is negligible and the probability of exactly one event occurring is proportional to Δt.

Under the three assumptions of the Poisson process, we obtain

$$\begin{aligned} p(x, t + \Delta t) &\approx P(x \text{ successes in } (0, t) \text{ and } 0 \text{ successes in } (t, t + \Delta t)) \\ &+ P(x - 1 \text{ successes in } (0, t) \text{ and } 1 \text{ success in } (t, t + \Delta t)) \\ &\approx p(x, t)(1 - \lambda(\Delta t)) + p(x - 1, t)(\lambda(\Delta t)). \end{aligned}$$

Rearranging the equation, we obtain

$$\frac{p(x, t + \Delta t) - p(x, t)}{\Delta t} \approx \frac{\lambda(\Delta t)(p(x - 1, t) - p(x, t))}{\Delta t}$$

and taking the limit as $\Delta t \to 0$, gives

$$\frac{\partial p(x, t)}{\partial t} \approx \lambda(p(x - 1, t) - p(x, t)) \qquad (4.13)$$

which holds for all non-negative integers x. A unique solution to the differential equation (4.13) is

$$p(x, t) = (\lambda t)^x \exp\{-\lambda t\}/x!$$

which is the pmf corresponding to the Poisson(λt) distribution.

Example 4.16 A switchboard receives calls at a rate of three per minute during a busy period. What is the probability of receiving more than two calls in a two minute interval during the busy period?

Solution: Your first thought might be that this is an impossible question since a probability is requested yet there is no probability distribution. However, this is a situation where the three assumptions of the Poisson process are likely to hold, at least approximately. For example, since there is a huge population of potential callers, what happens in one time interval should not affect what happens in a separate time interval. Also, since we are only considering the busy period, the stationarity assumption may be reasonable. Finally, it is conceivable that in a very short period of time, the probability of receiving a call should be proportional to the length of the time period. Therefore, it may be reasonable to assume that the number of calls X in a two minute interval is Poisson(6) where $6 = 2$ minutes times 3 calls/minute. The probability of interest is then

$$\begin{aligned} P(X > 2) &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - 6^0 e^{-6}/0! - 6^1 e^{-6}/1! - 6^2 e^{-6}/2! \\ &= 0.938. \end{aligned}$$

4.5 Exercises

4.01 Let W be a random variable giving the number of heads minus the number of tails in four tosses of a coin. List the elements of the sample space S for the four tosses of the coin, and to each sample point, determine the corresponding value of W.

4.02* Consider a weird coin where $P(\text{head}) = 0.6$ and $P(\text{tail}) = 0.4$. An experiment involves flipping the coin until a total of two tails is observed. Let X be the number of coin flips. (a) Write down the pmf for X. (b) Provide the values x such that $F(x) = 0.663$ where $F(x)$ is the cdf of X.

4.03* Consider rolling two four-sided dice with faces 1, 2, 3 and 4. (a) Obtain the pmf of X where X is the sum of the two resultant faces. (b) Suppose the two dice were rolled many times. Approximately, what would be the average of X? (c) Calculate the standard deviation of X.

4.04 Determine c such that $p(x) = c\binom{2}{x}\binom{3}{3-x}$ for $x = 0, 1, 2$ is a pmf.

4.05 Plot the cdf corresponding to the random variable with pmf $p(x) = (1/2)^{x-3}$ for $x = 4, 5, 6, \ldots$.

4.06 An investment firm offers bonds that mature after varying numbers of years. Let T denote the years to maturity of a randomly selected bond, and suppose that the cdf of T is

$$F(t) = \begin{cases} 0.00 & t < 1 \\ 0.10 & 1 \leq t < 3 \\ 0.50 & 3 \leq t < 5 \\ 0.55 & 5 \leq t < 7 \\ 1.00 & t \geq 7 \end{cases}$$

Calculate $P(T = 5)$, $P(T > 3)$, $P(1.4 < T < 6)$ and $P(T \leq 4 \mid T \geq 2)$.

4.07* The admission fee to a summer fair is \$1 for children and seniors, and \$3 for adults. Based on previous years, 65% of all visitors to the fair are adults. If 40 entrance tickets to the fair are purchased during a particular period, let X be the corresponding number of children and seniors. (a) What is the distribution of X? (b) What is the expected total ticket revenue during the period? (c) What is the variance of the total ticket revenue during the period?

4.08 Consider Y with pmf $p(y) = \exp(-3.8)(3.8)^{y-2}/(y-2)!$ defined on some set. (a) Determine a valid set. (b) Calculate $P(Y \geq 3)$.

4.09 An investor has \$100,000 to invest this year. One option is to purchase a guaranteed investment certificate with a 1.8% yield. Another option is to purchase a risky stock. The investor believes that the stock will increase by 5% with probability 0.7, and will decrease by 3% with probability 0.3. Based on these assumptions, what should the investor do? Do you have any practical caveats for the investor?

4.10 Suppose that Ian Bercovitz decides to go fishing every day in June at his favourite lake near Merritt, BC. Let X be the number of days where he catches fish during the month. He says that when he goes fishing on a given day in June, the probability that he gets skunked (i.e. doesn't catch fish) is 0.20. Discuss the adequacy of the model $X \sim$ binomial$(30, 0.8)$.

4.11 Suppose that a small grocery store purchases six cartons of skim milk each week at the wholesale price of \$1.20 per carton and retails the milk at \$1.65 per carton. At the end of the week, the unsold milk is removed from the shelf and the grocer receives a credit from the distributor equal to three-fourths of the wholesale price. Given the pmf $p(x)$ of the number X of cartons that are sold, calculate the expected weekly profit.

x	0	1	2	3	4	5	6
$p(x)$	1/15	2/15	2/15	3/15	4/15	3/15	0/15

Based on the pmf, what would you do if you were the grocer? Do you find the pmf believable? Explain.

4.12* A prominent physician claims that 70% of patients with lung cancer are chain smokers. (a) Suppose that 10 patients with lung cancer are admitted to a hospital. Calculate the probability that fewer than half of the patients are chain smokers. (b) Suppose that 20 patients with lung cancer are admitted to a hospital. Calculate the probability that fewer than half of the patients are chain smokers.

4.13 In a particular region, the probability that a randomly chosen individual is a smoker is 0.32. A random sample of 25 people is chosen from the region, and we define $X_i = 1$ if the i-th person is a non-smoker and $X_i = 0$ if the i-th person is a smoker. What is the distribution of $X_1 + X_2 + \cdots + X_{25}$?

4.14 In the 1996-1997 NBA regular season, the Chicago Bulls won 72 games and lost 10. Their winning percentage may therefore be estimated at $p = 72/82 = 0.88$. Using the corresponding binomial distribution, obtain the probability that the Bulls win their initial best-of-seven playoff series. Discuss the adequacy of the binomial distribution.

4.15 The probabilities of blood types O, A, B and AB in western populations are 0.46, 0.39, 0.12 and 0.03 respectively. Suppose that a clinic is seeking either type O or type B blood. From six randomly selected individuals, what is the probability that at least two people have the desired blood types?

4.16 Consider an experiment where three coins are tossed. The first coin has a "1" and a "2" marked on its two sides. The second coin has a "2" and a "3" marked on its two sides. The third coin has a "3" and a "4" marked on its two sides. Let X denote the product of the three numbers that are face up. (a) Obtain the pmf of X. (b) Calculate the expected value of $X/2$.

4.17 Suppose that, on average, 1 person in 1000 makes a numerical error in preparing their income tax return. If 10,000 returns are selected at random, find the probability that 6 or 7 or 8 of them contain a numerical error.

4.18* During a particular time of the year, beetles are found in the grass. The number of beetles is thought to follow a Poisson distribution with 3.0 beetles expected per square metre. Suppose that a circular trap is constructed such that the probability of capturing at least one beetle in the trap is 0.9. What is the radius of the trap?

4.19 Service calls arrive at a maintenance centre according to a Poisson process, with an average of 2.7 calls per minute. (a) Obtain the probability that no more than four calls arrive in a minute. (b) Obtain the probability that fewer than two calls arrive in a minute. (c) Obtain the probability that more than 10 calls arrive in a five-minute interval.

4.20 A local drugstore owner knows that an average of 100 people enter his store each hour. (a) Obtain the probability that nobody enters the store in a given three-minute period. (b) Obtain the probability that more than five people enter the store in a given three-minute period.

4.21 During the 2010 World Cup group stage, my friend Mahmood wanted me to pay him $5 if Brazil beat Portugal. At the time, I believed that Brazil would beat Portugal with probability 0.8. For the bet to be fair, how much should Mahmood have paid me if Brazil did not beat Portugal?

4.22 Suppose I repeat the following experiment 10 times: I toss a coin. If the coin lands heads, I roll a regular die. If the coin lands tails, I roll a die with faces 1, 1, 2, 2, 3, 3. Let X be the number of 5's observed. Either write down the binomial distribution for X or explain why X is not binomial.

4.23 Obtain k and b with respect to the rv X with pmf $p(x)$ and $E(X) = 0.6$.

x	-4	0	1	3
$p(x)$	k	$2k+b$	$k+3b$	$4k-2b$

4.24 Four passports are randomly returned to four students. Let X be the number of students who receive their own passport. Obtain the pmf of X.

4.25 A random variable X takes on the values 1, 2 and 4 with probabilities p_1, p_2 and p_4 respectively. Obtain p_1, p_2 and p_4 if $E(X) = 2.0$ and $Var(X) = 1.5$.

4.26 Two dice are thrown. Let M be the maximum of the two dice. (a) Obtain the pmf of M. (b) Evaluate the cdf at $M = 3.4$.

4.27 Let Y be the total number of goals scored by the Vancouver Canucks in the month of March. Is Y binomial? Explain.

4.28 A basketball player has an established practice routine. He does not leave the gym until he has made five free throws in a row. The probability that he makes a free throw is 0.85. What is the probability that he leaves after a total of 10 free throws?

Chapter 5

Continuous Distributions

A random variable X is **continuous** if it takes on values in an interval. The study of continuous random variables and their distributions is more difficult than in the discrete case in the sense that the associated theory relies on calculus. You may then wonder why do we bother with continuous distributions since we are only able to measure phenomena with a limited precision. In other words, everything that we measure is actually discrete. One answer is that inferential techniques (see Chapters 6 and 7) are more fully developed for continuous distributions, especially in the case of jointly distributed random variables.

In this chapter, you will need to hang tough. It is the last technical chapter where we build our arsenal of results before we address truly applied problems of inference. If you were diligent in Chapter 4 (Discrete Distributions), you may not find this chapter too onerous as many of the concepts translate directly from the discrete case to the continuous case.

With a discrete random variable, a probability mass function describes its distribution. With a continuous random variable, its distribution is described by a **probability density function**. Let X be a continuous random variable. Then the probability density function or **pdf** $f_X(x) \geq 0$ is such that

$$\boxed{P(a \leq X \leq b) \ = \ \int_a^b f_X(x)\ dx} \tag{5.1}$$

for all $a < b$. We often drop the subscript X from the pdf when the random variable is understood. Sometimes, we use the term **density** instead of pdf. From the definition of the pdf in (5.1), we observe that there is zero probability assigned to points (i.e. $P(X = c) = 0$ for any $c \in \mathcal{R}$).

Analogous to pmfs, a pdf $f(x)$ satisfies two properties:

1. $f(x) \geq 0$ for all x
2. $\int f(x)\ dx \ = \ 1$

where the integral is taken over the range of the random variable.

Example 5.1 Consider the pdf

$$f(x) \ = \ \begin{cases} 0 & x \leq 0 \\ x & 0 < x \leq 1 \\ 1/2 & 1 < x \leq 2 \\ 0 & 2 < x \end{cases} .$$

We first verify that $f(x)$ is a pdf by noting that $f(x) \geq 0$ for all x and

$$\begin{aligned} \textstyle\int f(x)\ dx \ &= \ \textstyle\int_0^1 x\ dx \ + \ \int_1^2 (1/2)\ dx \\ &= \ (1/2)(1)^2 \ + \ (1/2)(2-1) \\ &= \ 1. \end{aligned}$$

To calculate the probability $P(1 \leq X \leq 3/2)$, we write

$$P(1 \leq X \leq 3/2) \ = \ \int_1^{3/2} f(x)\ dx \ = \ (3/2-1)(1/2) \ = \ 1/4$$

where in this case, we are able to forego calculus and note that the probability corresponds to the area of a rectangle with base $(3/2 - 1)$ and height $1/2$.

A useful distribution and the simplest continuous distribution is the **uniform distribution**. A continuous random variable X has the uniform(a, b) distribution if it has the pdf

$$f(x) \ = \ \begin{cases} 1/(b-a) & a < x < b \\ 0 & \text{otherwise} \end{cases} . \tag{5.2}$$

The uniform is a family of distributions indexed by the parameters a and b where $a < b$. Note that the pdf (5.2) is flat on the interval (a, b) and may be applicable when we have "little knowledge" concerning a random variable other than it is constrained on (a, b). Note also that with continuous distributions, it makes no difference if the intervals are (a, b), $(a, b]$, $[a, b)$ or $[a, b]$ since there is zero probability assigned to points. The uniform$(0, 1)$ distribution is known as **standard uniform**.

5.1 Cdfs and Expectation for Continuous RVs

5.1.1 Cdfs

The definition of the cumulative distribution function (cdf) does not change when going from discrete to continuous distributions. A continuous random variable X with pdf $f(x)$ has cdf $F(x) = P(X \leq x)$. However, the evaluation of the cdf for a continuous distribution requires integration and the cdf is given by

$$\boxed{F(x) \;=\; \int_{-\infty}^{x} f(y)\; dy.}$$

Pdfs and cdfs are both functions, and often it is instructive to plot them to better understand the probability distribution of a random variable. For me, and I think for most statisticians, it is easier to interpret the graph of a pdf than the graph of a cdf.

Having introduced cdfs, it is easy to define percentiles. When we say that a test score is in the 80-th percentile, we mean that 80% of scores are below the test score, or equivalently, 20% of scores exceed the test score. The $100p$-th **percentile** of a continuous distribution with cdf $F(x)$ is a value $\eta(p)$ such that

$$\boxed{p \;=\; F(\eta(p)) = P(X \leq \eta(p)).}$$

Note that there is a slight technical problem with the definition of percentiles as percentiles are not necessarily unique. This may happen with "weird" distributions where there are gaps, as for example, $f(x) = 1/2$ where $x \in (0,1)$ and $x \in (2,3)$. Don't worry about the technicality; all of the distributions that we consider lead to unique percentiles. The **median** of a continuous distribution which we typically denote by $\tilde{\mu}$ is defined as the 50-th percentile.

Example 5.2 Consider $X \sim \text{uniform}(a,b)$. The corresponding cdf is given by $F(x) = \int_a^x (1/(b-a))\ dy = (x-a)/(b-a)$ and therefore the $100p$-th percentile is $\eta(p) = a + p(b-a)$. Note that the median $\tilde{\mu} = (a+b)/2$ is halfway between a and b as our intuition suggests.

Example 5.3 This is a continuation of Example 5.1. The corresponding cdf is given by

$$F(x) \;=\; \begin{cases} 0 & x \le 0 \\ \int_0^x y\ dy & 0 < x \le 1 \\ \int_0^1 y\ dy + \int_1^x (1/2)\ dy & 1 < x \le 2 \\ \int_0^1 y\ dy + \int_1^2 (1/2)\ dy & 2 < x \end{cases}$$

$$= \begin{cases} 0 & x \le 0 \\ x^2/2 & 0 < x \le 1 \\ x/2 & 1 < x \le 2 \\ 1 & 2 < x \end{cases}$$

and the $100p$-th percentile is given by

$$\eta(p) \;=\; \begin{cases} \sqrt{2p} & 0 < p \le 1/2 \\ 2p & 1/2 < p \le 1 \end{cases} .$$

5.1.2 Expectation

The concept of expectation is the same for both continuous random variables and discrete random variables. The difference lies in the calculations where we replace sums in the discrete case with integrals in the continuous case.

The expected value or expectation of a continuous random variable X with pdf $f(x)$ is

$$\boxed{\mu \;=\; E(X) \;=\; \int x f(x)\;dx}$$

where the integration is taken over the range of the random variable X. The corresponding expectation of a function $g(X)$ is given by

$$\boxed{E(g(X)) \;=\; \int g(x) f(x)\;dx.}$$

Again, the spread or dispersion of a random variable is an informative characteristic. Dispersion is often calibrated by the variance which is the expectation of $g(X) = (X - E(X))^2$. Therefore, the variance of the continuous random variable X is

$$\boxed{\sigma^2 \;=\; Var(X) \;=\; \int (x - E(X))^2 f(x)\;dx}$$

where σ is referred to as the standard deviation. It is often useful to remember the identity $\boxed{Var(X) = E(X^2) - (E(X))^2}$ which also holds for continuous distributions.

Example 5.4 Consider $X \sim \text{uniform}(a, b)$. We obtain $E(X) = \int_a^b x/(b - a)\;dx = (a + b)/2$ and $E(X^2) = \int_a^b x^2/(b - a)\;dx = (a^2 + b^2 + ab)/3$. This leads to the variance $\sigma^2 = (a^2 + b^2 + ab)/3 - (a + b)^2/4 = (b - a)^2/12$.

5.2 The Normal Distribution

The **normal distribution** is celebrated for its wide applicability in problems of statistical science and for its mathematical elegance. It is no doubt the most important distribution in all of statistics. A distribution which is normal often goes by other names including **Gaussian** and **bell shaped**.

A random variable X has a normal(μ, σ^2) distribution if it has the pdf

$$\boxed{f(x) \; = \; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\}} \tag{5.3}$$

where $x \in \mathcal{R}$ and the parameters $-\infty < \mu < \infty$ and $\sigma > 0$ index the family of normal distributions. Looking at (5.3), we see that the normal pdf is symmetric about the parameter μ (i.e. $f(\mu + \Delta) = f(\mu - \Delta)$ for all Δ). A little analysis also shows that $f(x)$ is maximized at $x = \mu$ and that the pdf $f(x)$ tails off exponentially as $x \to -\infty$ and $x \to \infty$. However, there is no need to memorize (5.3) since normal pdfs are **intractable**. By intractable, we mean that probabilities (i.e. integrals $\int_a^b f(x)\; dx$) can not be evaluated in a closed form. Normal probabilities can however be approximated using numerical methods (more later).

A curiosity of the normal(μ, σ^2) distribution is that we have chosen the parameters μ and σ to characterize the distribution even though the symbols μ and σ have traditionally been reserved to denote the mean and the standard deviation of distributions. It turns out that there is no conflict. If $X \sim$ normal(μ, σ^2), then $E(X) = \mu$ and $Var(X) = \sigma^2$ (the derivations require integration by parts). Pay particular attention to the second parameter in the normal(μ, σ^2) formulation; it is the variance and not the standard deviation. In Figure 5.1, we plot the densities of the normal$(4, 4)$ and the normal$(6, 1)$. We see that the effect of increasing the mean from 4 to 6 shifts the pdf to the right, and the effect of decreasing the variance from 4 to 1 decreases the spread in the pdf. As mentioned previously, note that the densities are symmetric.

A recurring issue from this point forward is the calculation of normal probabilities. However, we previously mentioned that the related integrals are intractable. The way that we obtain normal probabilities is through the use of normal tables. At first thought, this may seem impossible since there are an infinite number of normal distributions corresponding to an infinite number of choices of μ and σ. Does this imply the need for an infinite number of normal tables?

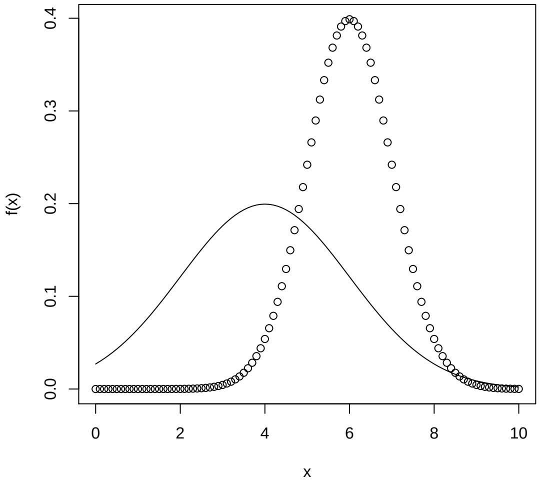

Figure 5.1: The normal$(4, 4)$ density (solid line) and the normal$(6, 1)$ density (dots).

Of course, the answer is no, and the way that we calculate a normal probability for a specific choice of μ and σ is to convert the problem to a probability involving the normal(0,1) distribution. Therefore, we require only a single normal table. A random variable $Z \sim$ normal$(0, 1)$ is referred to as **standard normal** and it has the simplified pdf

$$f(z) \;=\; \frac{1}{\sqrt{2\pi}}\exp\{-z^2/2\}.$$

A standard normal table is given in Table B.2 of Appendix B. It provides cumulative probabilities for Z. You must become adept at using the standard normal table.

Example 5.5 Here are some calculations to check your proficiency with the standard normal table: Look up $P(Z \leq 1.07)$ giving 0.8577 and note that this is the same as $P(Z < 1.07)$. To get $P(Z > 0.34)$, note that this equals $1 - P(Z \leq 0.34) = 1 - 0.6331 = 0.3669$. The argument of $P(Z < 1.246)$ has an extra digit, and this suggests **interpolation** giving $P(Z < 1.246) \approx P(Z < 1.24) + (6/10)[P(Z < 1.25) - P(Z < 1.24)] = 0.8936$. This being said, it is a rare statistical application where we require probabilities to a high degree of accuracy, say beyond two decimal points. Therefore there is rarely a need to interpolate. Recalling that the standard normal distribution is symmetric about $Z = 0$, $P(Z > -0.95) = P(Z < 0.95) = 0.8289$. Finally, "inverse" problems are important. Find the value z such that 43.25% of the standard normal population exceeds z. The answer is 0.17 since $P(Z \leq 0.17) = 0.5675$ and therefore $P(Z > 0.17) = 0.4325$. Finally, observe that the tails of the standard normal distribution die out quickly; beyond ±3.0, there is very little probability.

Now, some of you may own a scientific calculator that can calculate normal probabilities. Whereas this is convenient, make sure that you understand the basics of using a standard normal table as outlined in Example 5.5. You may also come across more comprehensive standard normal tables where table entries are available for negative z-values. Although we can interpolate from Table B.2, there are a few z-values for which higher precision is sometimes required. We list those z-values here and their corresponding cdf $F(z)$ for convenience.

z	$F(z)$
1.282	0.900
1.645	0.950
1.960	0.975
2.326	0.990
2.576	0.995

The relationship between an arbitrary normal random variable $X \sim \text{normal}(\mu, \sigma^2)$ and the standard normal distribution is expressed via

$$\frac{X - \mu}{\sigma} \sim \text{normal}(0, 1) \tag{5.4}$$

where the proof is based on the change of variables theorem in calculus.

Example 5.6 To my horror, a subset of Canadians watch an average of 6.9 hours of television per day. Assume that the viewing times are normally distributed with a standard deviation of 3.8 hours. What is the probability that a randomly selected Canadian from the subset watches more than 8.0 hours of television per day?

Solution: Let X denote the number of hours of television watched by a randomly selected Canadian. We have $X \sim \text{normal}(6.9, 3.8^2)$ and we calculate the probability

$$\begin{aligned} P(X > 8.0) &= P\left(\tfrac{X-6.9}{3.8} > \tfrac{8.0-6.9}{3.8}\right) \\ &= P(Z > 0.29) \\ &= 1 - P(Z \leq 0.29) \\ &= 0.39 \end{aligned}$$

where $Z \sim \text{normal}(0, 1)$, the second equality follows from (5.4), and the fourth equality is obtained from Table B.2.

Example 5.7 It is suggested that the length of a particular insect in cm is normally distributed with $\mu = 0.30$ cm and $\sigma = 0.06$ cm. How would you characterize the largest 5% of all insect lengths?
Solution: Let X be the length of a randomly chosen insect where $X \sim \text{normal}(0.30, 0.06^2)$. We are trying to find the value a such that $P(X > a) = 0.05$. This implies $P(Z > (a - 0.30)/(0.06)) = 0.05$ where $Z \sim \text{normal}(0, 1)$. From the standard normal tables, we obtain the z-value $(a - 0.3)/(0.06) = 1.645$ which gives $a = 0.40$ cm.

Example 5.8 Verbal SAT scores follow the normal(430, 100) distribution. What is the middle range of scores encompassing 50% of the population?
Solution: Let $X \sim \text{normal}(430, 100)$ denote an SAT score. We want y such that $P(430 - y \leq X \leq 430 + y) = 0.5$. This implies

$$P\left(\frac{430 - y - 430}{10} \leq Z \leq \frac{430 + y - 430}{10}\right) = 0.5$$

where Z is standard normal. Therefore $P(-y/10 \leq Z \leq y/10) = 0.5$ which gives $y/10 = 0.675$ interpolated from the standard normal table B.2. Therefore $y = 6.75$ and the required interval is $(423.3, 436.7)$.

Example 5.9 The maximum daily June temperature in degrees celsius (C) in a computer room has a normal distribution with mean 22.0 C and standard deviation 0.6 C. Suppose that computer equipment damage occurs when the temperature exceeds 23.44 C. What is the probability that there is computer damage in the month of June in at least one of five independent computer rooms?

Solution: Let $X \sim$ normal$(22.0, 0.6^2)$ be the maximum temperature in degrees C in a computer room in the month of June. The probability of damage in the computer room is given by

$$\begin{aligned} P(X > 23.44) &= P(Z > (23.44 - 22.0)/0.6) \\ &= P(Z > 2.40) \\ &= 0.0082. \end{aligned}$$

Therefore, by independence, computer damage in at least one of five rooms is $1 - (0.9918)^5 = 0.0403$.

Example 5.10 The volume placed in a bottle by a bottling machine follows a normal(μ, σ^2) distribution. Over a long period of time, it is observed that 5% of the bottles contain less than 31.5 ounces and 15% of the bottles contain more than 32.3 ounces. Find the probability that out of 10 bottles purchased, exactly three contain more than 32.2 ounces.

Solution: Let $X \sim$ normal(μ, σ^2) be the amount dispensed in ounces. We are given that $P(X < 31.5) = 0.05$ and $P(X > 32.3) = 0.15$. Standardizing the two equalities and obtaining z-values from the standard normal table gives $(31.5 - \mu)/\sigma = -1.645$ and $(32.3 - \mu)/\sigma = 1.036$ respectively. Therefore, we have two equations in the two unknowns μ and σ. Solving gives $\mu = 31.99$ and $\sigma = 0.2984$. The next step involves calculating the probability p that a bottle contains more than 32.2 ounces. This is given by

$$\begin{aligned} p &= P(X > 32.2) \\ &= P(Z > (32.2 - 31.99)/0.2984) \\ &= P(Z > 0.7037) \\ &= 0.241 \end{aligned}$$

where Z is standard normal. We finish the problem by calculating $P(Y = 3)$ where $Y \sim$ binomial$(10, p)$ is the number of bottles out of 10 that have more than 32.2 ounces. Therefore $P(Y = 3) = \binom{10}{3}(0.241)^3(0.759)^7 = 0.244$.

Returning to percentiles, there is a convenient formula which relates percentiles of the normal(μ, σ^2) distribution to percentiles of the standard normal distribution. Suppose that we are interested in the $100p$-th percentile $\eta(p)$ corresponding to $X \sim$ normal(μ, σ^2) and let $\eta_z(p)$ be the $100p$-th percentile corresponding to $Z \sim$ normal(01). Then $\eta(p) = \mu + \sigma\eta_z(p)$ since

$$\begin{aligned} P(X \leq \mu + \sigma\eta_z(p)) &= P\left(\tfrac{X-\mu}{\sigma} \leq \tfrac{\mu+\sigma\eta_z(p)-\mu}{\sigma}\right) \\ &= P(Z \leq \eta_z(p)) \\ &= p. \end{aligned}$$

Example 5.11 Find the 25.78-th percentile of the normal$(5, 100)$.
Solution: From the standard normal table, $P(Z \leq -0.65) = 0.2578$. Therefore $\eta_z(0.2578) = -0.65$ and the 25.78-th percentile is $\eta(0.2578) = 5 + 10(-.65) = -1.5$.

5.2.1 The Normal Approximation to the Binomial

In Example 4.14 we saw that binomial(n, p) probabilities can sometimes be difficult to calculate. We also saw in Section 4.4 that the Poisson(np) distribution can provide an adequate approximation to the binomial when n is much larger than np. In this section, we discuss the normal approximation to the binomial which is applicable under different circumstances.

Let $X \sim$ binomial(n, p) and suppose that p is neither too small nor too large (i.e. p is neither close to 0.0 nor 1.0). We express the condition by the **"rule of thumb"** which requires $\boxed{np \geq 5 \text{ and } n(1-p) \geq 5}$. The normal approximation to the binomial states that the distribution of X is "close" to the distribution of Y where $\boxed{Y \sim \text{normal}(np, np(1-p))}$. Note that we have not been rigorous in this statement as the notion of closeness of distributions has not been defined. The proof of the approximation is a special case of the Central Limit Theorem (Section 5.6). Although the approximation may seem surprising (a finite discrete distribution is approximated by a continuous distribution), we observe that the mean and variance of X equal the mean and variance of Y. Also note that whereas the normal distribution is symmetric, the binomial is only symmetric if $p = 1/2$ and becomes more skewed as $p \rightarrow 0$ and $p \rightarrow 1$; this provides some explanation why we require that p be neither too small nor too large.

Example 5.12 As an illustration of the normal approximation to the binomial, consider $X \sim$ binomial$(10, 1/2)$ and calculate

$$\begin{aligned} P(X \geq 8) &= \tbinom{10}{8}(1/2)^{10} + \tbinom{10}{9}(1/2)^{10} + \tbinom{10}{10}(1/2)^{10} \\ &= 0.0547. \end{aligned} \tag{5.5}$$

We define $Y \sim$ normal$(5, 2.5)$ and compare the binomial probability (5.5) with its normal approximation

$$\begin{aligned} P(Y \geq 8) &= P\left(\tfrac{Y-5}{\sqrt{2.5}} \geq \tfrac{8-5}{\sqrt{2.5}}\right) \\ &= P(Z \geq 1.90) \\ &= 0.0287 \end{aligned} \tag{5.6}$$

where $Z \sim$ normal$(0, 1)$. As an approximation, the probability in (5.6) may not seem too good. To improve the approximation, we introduce a **continuity correction** via

$$\begin{aligned} P(X \geq 8) &= 1 - P(X \leq 7) \\ &\approx 1 - P(Y \leq 7.5) \\ &= 1 - P\left(\tfrac{Y-5}{\sqrt{2.5}} \leq \tfrac{7.5-5}{\sqrt{2.5}}\right) \\ &= 1 - P(Z \leq 1.58) \\ &= 0.0571. \end{aligned} \tag{5.7}$$

The continuity correction used in (5.7) improves the approximation dramatically, and from this point on, we use a continuity correction whenever we approximate a discrete probability with a continuous probability. The continuity correction is based on the realization that the bars of pmfs have width. A general formula for computing approximate binomial probabilities is given below where $Z \sim$ normal$(0, 1)$:

$$\boxed{P(X \leq x) \approx P\left[Z \leq \frac{x + .5 - np}{\sqrt{np(1-p)}}\right].}$$

To check that you understand this concept, the continuity correction applied to $P(X < 7)$ in the above problem yields $P(X < 7) = P(X \leq 6) \approx P(Y \leq 6.5)$.

Example 5.13 A college would like 1200 students to enroll. Since not all admitted students enroll, the college admits more than 1200 students. Past experience shows that 70% of students who are offered admission actually enroll. The college offers admission to 1500 students. Find the probability that at least 1200 students enroll.

Solution: Enrollment may be viewed as a success/failure situation where there are 1500 trials. Furthermore, enrollment decisions are likely independent of one another and the probability of enrollment is constant across students as the admitted students represent a random sample from a hypothetical population of students. These considerations suggest that the number of students who enroll X may be adequately modelled using a binomial$(1500, 0.7)$ distribution. The probability of interest is therefore $P(X \geq 1200)$. The probability is difficult to calculate and we therefore approximate the random variable X with $Y \sim \text{normal}(1500(0.7), 1500(0.7)(0.3)) \sim \text{normal}(1050, 315)$. Our rule of thumb for the normal approximation to the binomial is satisfied since $np = 1050 \geq 5$ and $n(1-p) = 450 \geq 5$. Using the continuity correction, we calculate

$$\begin{aligned} P(X \geq 1200) &\approx P(Y \geq 1199.5) \\ &= P\left(Z \geq \tfrac{1199.5-1050}{\sqrt{315}}\right) \\ &= P(Z \geq 8.45) \\ &\approx 0 \end{aligned}$$

where Z is standard normal.

Now, there is a caveat to this subsection concerning the continuity correction. One of my colleagues, Derek Bingham, came up to me and essentially said, “Why are you teaching this continuity correction business? This is not in keeping with the modern practice of statistics where we use a computer for everything. If you want to calculate a binomial probability in a small sample size problem, then simply use a computer or a calculator to obtain the probability. In a large sample size problem, the continuity correction does not make a big difference." After punching him in the nose for his insolence, I think he may have a point. However, I have decided to retain the discussion of the continuity correction, even if he is right. It is not such a bad thing to reinforce our understanding of discrete versus continuous distributions.

5.3 The Gamma Distribution

Although the normal distribution is the pre-eminent distribution in statistics, its symmetry provides a drawback in data modelling. For right-skewed continuous data, the **gamma distribution** provides an alternative to the normal. A random variable X has a gamma(α, β) distribution if it has the pdf

$$f(x) \;=\; \frac{x^{\alpha-1}\exp\{-x/\beta\}}{\beta^{\alpha}\,\Gamma(\alpha)}$$

where $x > 0$, $\alpha > 0$ and $\beta > 0$ are parameters that index the gamma family of distributions, and the **gamma function** is a constant defined by $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\exp\{-x\}\;dx$.

Although we will not spend much time on the gamma distribution, there are a few relevant facts that you should know:

- $E(X) = \alpha\beta$ and $Var(X) = \alpha\beta^2$ (prove this as an exercise)

- except for particular parameter values, the gamma family is intractable

- the gamma function is also generally intractable but satisfies

  - $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$
  - $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$
  - $\Gamma(n) = (n-1)!$ for positive integers n

5.3.1 The Exponential Distribution

We have seen that the gamma distribution is a two-parameter family of continuous distributions. It turns out that the **exponential distribution** is a 1-parameter sub-family of the gamma family where the exponential(λ) distribution is obtained by setting $\alpha = 1$ and $\beta = 1/\lambda$. Therefore a random variable X is exponential(λ) if it has the pdf

$$\boxed{f(x) \; = \; \lambda \exp\{-\lambda x\}} \tag{5.8}$$

where $x > 0$ and the parameter $\lambda > 0$ indexes the exponential family. By referring to the mean and variance expressions for the gamma, $\boxed{E(X) = 1/\lambda}$ and $\boxed{Var(X) = 1/\lambda^2.}$ When plotted, the pdf (5.8) decreases exponentially to the right from its maximum $f(0) = \lambda$.

We note that the exponential(λ) distribution is tractable and its cdf is given by

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \int_0^x \lambda \exp\{-\lambda y\} \; dy \\ &= 1 - \exp\{-\lambda x\} \end{aligned}$$

for $x > 0$.

A fascinating property of the exponential distribution is the **memoryless property**. The memoryless property is sometimes a characteristic of electronic components such as light bulbs, and it essentially states that a used component is as good as a new component. Putting the problem in context, let $X \sim$ exponential(λ) be the life span in hours of a light bulb. Given that a light bulb has lasted a hours, the probability that it lasts an additional b hours is

$$\begin{aligned} P(X > a + b \mid X > a) &= P(X > a + b \; \cap \; X > a) \; / \; P(X > a) \\ &= P(X > a + b) \; / \; P(X > a) \\ &= \exp\{-\lambda(a + b)\} \; / \; \exp\{-\lambda a\} \\ &= \exp\{-\lambda b\} \\ &= P(X > b) \end{aligned}$$

where the final term is the probability that a new light bulb lasts b hours. Therefore, there is no deterioration in age; the old light bulb is as good as the new light bulb and the light bulb's performance is memoryless.

Over the years, a number of students have indicated their objection to the memoryless property. If you are convinced that it does not apply to the phenomena that interests you, then this is a signal that you should not use the exponential distribution in your application.

Another interesting property involving the exponential distribution is its relationship to the Poisson distribution. Suppose that the random variable N_T is the number of events that occur in the time interval $(0, T)$ and we assume that $N_T \sim \text{Poisson}(\lambda T)$. Define a second random variable X as the waiting time until the first event. Then the cdf of X is

$$\begin{aligned} F(x) &= 1 - P(X > x) \\ &= 1 - P(\text{zero events in } (0, x)) \\ &= 1 - P(N_x = 0) \\ &= 1 - (\lambda x)^0 \exp\{-\lambda x\}/0! \\ &= 1 - \exp\{-\lambda x\} \end{aligned}$$

where we recognize the last term as the cdf of the exponential(λ) distribution. Therefore, when the number of events in time is Poisson, then the waiting time until the first event is exponential.

Example 5.14 Let X denote the life span in weeks of a cheap MP4 player where $X \sim \text{exponential}(0.01386)$. What is the probability that a newly purchased MP4 player exceeds its mean life expectancy by more than two standard deviations?

Solution: The mean and standard deviation are equal for the exponential distribution. Here $E(X) = 1/(0.01386) = 72.15$ weeks. Referring to the cdf of the exponential distribution, the probability of interest is

$$\begin{aligned} P(X \geq 72.15 + 2(72.15)) &= 1 - P(X \leq 216.45) \\ &= \exp\{-0.01386(216.45)\} \\ &= 0.05. \end{aligned}$$

5.4 Jointly Distributed Random Variables

We are now into the "dog days" of the course where you need to continue to buckle down and endure some additional theoretical and abstract topics. I promise you that you will see more value in the course when we get to the last two chapters on inference. The heavens will open and you will be enlightened.

Up until this point, we have only considered univariate random variables (both discrete and continuous). However, there are many instances where **multivariate** data are collected on subjects. For example, one might measure cholesterol level, coronary calcium score and fitness on a group of potential heart attack victims. The variables are not always independent and it is sometimes useful to study the variables in tandem. For modelling multivariate phenomena, we need jointly distributed random variables and their associated multivariate distributions.

In the discrete setting, the joint probability mass function for the discrete random variables $X_1, X_2, \ldots X_m$ is the function $p(x_1, x_2, \ldots, x_m)$ where

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) = p(x_1, x_2, \ldots, x_m).$$

As a starting point, consider the discrete trivariate distribution given in Table 5.1 where the 12 cell entries are probabilities. For example, $P(X = 1, Y = 2, Z = 6) = 0.05$. The probabilities of events are obtained by summing over the appropriate cells. For example,

$$\begin{aligned} P(X \geq 2, Y \geq 1, Z = 5) &= p(2,1,5) + p(2,2,5) + p(3,1,5) + p(3,2,5) \\ &= 0.20 + 0.05 + 0.00 + 0.05 \\ &= 0.30 \end{aligned}$$

and note that the sum over all cells is 1.0. It is also possible to obtain **marginal distributions** by "summing out" the **nuisance** variables. For example, the marginal distribution of Z is given by

$$\boxed{P(Z = z) \;=\; \sum_{x,y} P(X = x, Y = y, Z = z)}$$

	$Z = 5$			$Z = 6$		
$Y = 1$	0.10	0.20	0.00	0.00	0.30	0.10
$Y = 2$	0.00	0.05	0.05	0.05	0.05	0.10
	$X = 1$	$X = 2$	$X = 3$	$X = 1$	$X = 2$	$X = 3$

Table 5.1: A trivariate pmf defined on (X, Y, Z).

which leads to $P(Z = 5) = 0.4$ and $P(Z = 6) = 0.6$.

In the continuous case, many of the topics that we have discussed in the univariate case generalize directly to the multivariate case. For example, the joint pdf $f(x_1, x_2, \ldots, x_m)$ defined on the m random variables $X_1, X_2, \ldots, X_m$ satisfies two properties:

1. $f(x_1, x_2, \ldots, x_m) \geq 0$ for all $x_1, x_2, \ldots, x_m$
2. $\int \int \cdots \int f(x_1, x_2, \ldots, x_m)\; dx_1\; dx_2 \cdots dx_m \;=\; 1$

where the multivariate integral is taken over the range of $X_1, X_2, \ldots, X_m$.

The probability of an event A (which corresponds to a set defined on $X_1, X_2, \ldots, X_m$) is specified by

$$P((X_1, X_2, \ldots, X_m) \in A) \;=\; \int \int \cdots \int_A f(x_1, x_2, \ldots, x_m)\; dx_1\; dx_2 \cdots dx_m.$$

Marginal pdfs are defined in an analogous way to the discrete case where instead of summation, the nuisance variables are integrated out. We emphasize that although probabilities may be easy to specify as integrals, their evaluation may sometimes be difficult or impossible.

Example 5.15 Consider the bivariate pdf

$$f(x, y) \;=\; \begin{cases} \frac{2}{5}(2x + 3y) & 0 < x < 1,\ 0 < y < 1 \\ \quad 0 & \text{otherwise} \end{cases}.$$

When you integrate a multivariate integral, you first integrate with respect to one of the variables where the other variables are treated as constants.

After the integrated variable is eliminated, you continue integrating in the same fashion. For example,

$$\begin{array}{rcl} P(X > 1/2, Y < 1/2) & = & \int_{x=1/2}^{1} \int_{y=0}^{1/2} \frac{2}{5}(2x+3y)\ dy\ dx \\ & = & \int_{x=1/2}^{1} \left[\frac{4}{5}xy + \frac{3}{5}y^2\right]_{y=0}^{1/2}\ dx \\ & = & \int_{x=1/2}^{1} \left(\frac{2}{5}x + \frac{3}{20}\right)\ dx \\ & = & \left[\frac{1}{5}x^2 + \frac{3}{20}x\right]_{x=1/2}^{1} \\ & = & \frac{1}{5} + \frac{3}{20} - \frac{1}{20} - \frac{3}{40} \\ & = & \frac{9}{40}. \end{array}$$

The marginal pdf of X is obtained by "integrating out Y" giving

$$\begin{array}{rcl} f(x) & = & \int_{y=0}^{1} \frac{2}{5}(2x+3y)\ dy \\ & = & \left[\frac{4}{5}xy + \frac{3}{5}y^2\right]_{y=0}^{1} \\ & = & \frac{4}{5}x + \frac{3}{5} \end{array}$$

where we must be careful to specify the range $0 < x < 1$. In general, when asked to derive a distribution, you will not receive full marks should you fail to specify the range.

Example 5.16 Here we consider a bivariate continuous distribution that is complicated by the fact that the range of X depends on Y. The question involves calculus in two variables. Obtain $P(X < 0.2)$ where the joint density of X and Y is given by

$$f(x,y) = 24xy \quad 0 < x < 1,\ 0 < y < 1,\ x+y < 1.$$

Solution: The probability is given by

$$\begin{array}{rcl} P(X < 0.2) & = & \int_{x=0}^{0.2} \int_{y=0}^{1-x} 24xy\ dy\ dx \\ & = & \int_{x=0}^{0.2} \left[_{y=0}^{1-x}\ 12xy^2\right]\ dx \\ & = & \int_{x=0}^{0.2} 12x(1-x)^2\ dx \\ & = & [_{x=0}^{0.2}\ 6x^2 - 8x^3 + 3x^4] \\ & = & 0.1808. \end{array}$$

You can also imagine joint distributions which consist of some discrete and some continuous random variables. When doing probability calculations, you need to sum and integrate accordingly.

Independence of events was discussed in Chapter 3. The concept of independence also extends to random variables. We say that jointly distributed discrete **random variables are independent** if their joint pmf factors into the product of their marginal pmfs. In the continuous case, independence requires the factorization of the joint pdf into the product of the marginal pdfs.

Example 5.17 Let X and Y be jointly distributed continuous random variables having bivariate pdf $f(x,y)$ defined on $(x,y) \in \mathcal{R}^2$ where

$$\begin{aligned} f(x,y) &= \tfrac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\tfrac{1}{2}\left(\tfrac{x^2}{\sigma_1^2} + \tfrac{y^2}{\sigma_2^2}\right)\right\} \\ &= \tfrac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\tfrac{1}{2}\left(\tfrac{x-0}{\sigma_1}\right)^2\right\} \times \tfrac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left\{-\tfrac{1}{2}\left(\tfrac{y-0}{\sigma_2}\right)^2\right\}. \end{aligned}$$

We observe that X and Y are independent since $f(x,y)$ factors according to $X \sim \text{normal}(0, \sigma_1^2)$ and $Y \sim \text{normal}(0, \sigma_2^2)$. The pdf $f(x,y)$ is referred to as a **bivariate normal density**.

Example 5.18 Consider discrete random variables X and Y with the joint pmf $p(x,y)$ where $p(1,1) = 0.4$, $p(1,2) = 0.1$, $p(2,1) = 0.2$ and $p(2,2) = 0.3$. Are X and Y independent?

Solution: By summing out Y, the marginal pmf of X is $p_X(x)$ where $p_X(1) = 0.5$ and $p_X(2) = 0.5$. Similarly, by summing out X, the marginal pmf of Y is $p_Y(y)$ where $p_Y(1) = 0.6$ and $p_Y(2) = 0.4$. In order for X and Y to be independent, we require $p(x,y) = p_X(x)p_Y(y)$ for all x and y. Noting that the factorization does not hold in the instance $p(1,1) = 0.4 \neq p_X(1)p_Y(1) = (0.5)(0.6) = 0.3$ establishes that X and Y are dependent. It is worth emphasizing that independence requires the factorization of the pdf/pmf to hold for all x and y.

The conditional probability of events was discussed in Chapter 3. The concept of conditional probability also extends to random variables. Suppose that we are interested in the continuous random variable X_1 given the values of the continuous random variables $X_2, X_3, \ldots, X_m$. Then the **conditional**

density or **conditional pdf** of X_1 given $X_2 = x_2, X_3 = x_3, \ldots, X_m = x_m$ is

$$f_{X_1 | X_2, X_3, \ldots, X_m}(x_1) \; = \; \frac{f_1(x_1, x_2, \ldots, x_m)}{f_2(x_2, x_3, \ldots, x_m)} \tag{5.9}$$

where f_1 and f_2 are the joint pdfs of $X_1, X_2, \ldots, X_m$ and $X_2, X_3, \ldots, X_m$ respectively. The use of subscripts in (5.9) can become a little burdensome, and therefore, notation is often slightly abused by dropping the subscripts.

In (5.9), we have the conditional density of a single random variable given the remaining random variables. It is also possible to derive conditional densities for a subset of random variables given another subset of random variables. For example, $f_{X_2, X_5 | X_7, X_8, X_{10}}(x_2, x_5)$ is a conditional density and we note that it is solely a function of x_2 and x_5. In the case of discrete random variables, **conditional pmfs** are obtained by appropriately replacing pdfs with pmfs.

Example 5.19 This is a continuation of Example 5.15. Suppose that we are interested in the distribution of Y given $X = 0.2$. Using the expressions derived in Example 5.15 and reintroducing subscripts,

$$\begin{aligned} f_{Y|X=0.2}(y) &= \tfrac{f_{XY}(0.2,y)}{f_X(0.2)} \\ &= \tfrac{\frac{2}{5}(2(0.2)+3y)}{\frac{4}{5}(0.2)+\frac{3}{5}} \\ &= \tfrac{0.4+3y}{1.9} \end{aligned}$$

where we are again careful to specify the range $0 < y < 1$. Note that conditional pmfs and pdfs satisfy the properties of pmfs and pdfs respectively. In this example, $f_{Y|X=0.2}(y) \geq 0$ for all $0 < y < 1$ and $\int_0^1 f_{Y|X=0.2}(y) \; dy = 1$.

Expectations also generalize to the multivariate setting. For example, the expectation of a function $g(X_1, X_2, \ldots, X_m)$ where $X_1, X_2, \ldots, X_m$ are discrete random variables with joint pmf $p(x_1, x_2, \ldots, x_m)$ is given by

$$E(g(X_1, X_2, \ldots, X_m)) \; = \; \sum_{x_1} \sum_{x_2} \cdots \sum_{x_m} g(x_1, x_2, \ldots, x_m) p(x_1, x_2, \ldots, x_m).$$

In the continuous setting, pdfs replace pmfs, and integrals replace sums. The interpretation of expectations is the same as before. Imagine a large number of hypothetical repetitions of the experiment where each experiment yields a realization of the vector $(X_1, X_2, \ldots, X_m)$. The expectation $E(g)$ is the average of the realized values of $g(X_1, X_2, \ldots, X_m)$.

Example 5.20 Refer to Table 5.1 and let $g(x, y, z) = xz$. We calculate $E(g) = E(g(X, Y, Z))$ via

$$\begin{aligned} E(g) &= g(1,1,5)p(1,1,5) + g(1,1,6)p(1,1,6) + \cdots + g(3,2,6)p(3,2,6) \\ &= (1)(5)(0.1) + (1)(6)(0.0) + \cdots + (3)(6)(0.1) \\ &= 11.85. \end{aligned}$$

As a related exercise, verify that $E(\max(x, y)) = 2.15$.

Example 5.21 Calculate $E(|X - Y|)$ where X and Y are independent with pdfs $f_X(x) = 3x^2$ for $0 < x < 1$ and $f_Y(y) = 2y$ for $0 < y < 1$ respectively.
Solution: Using the independence of X and Y, the expectation can be written as the double integral

$$E(|X - Y|) \; = \; \int_{y=0}^{1} \int_{x=0}^{1} |x - y| \; (6x^2 y) \; dx \; dy. \tag{5.10}$$

Due to the absolute value appearing in (5.10), the integral is a bit tricky. We therefore split the range of integration into two sub-regions where the absolute value expression simplifies in each sub-region. This leads to

$$\begin{aligned} E(|X - Y|) &= \textstyle\int_{y=0}^{1} \int_{x=y}^{1} (x - y)(6x^2 y) \; dx \; dy \\ &+ \textstyle\int_{y=0}^{1} \int_{x=0}^{y} (y - x)(6x^2 y) \; dx \; dy. \end{aligned}$$

After some tedious calculations (which you should verify), $E(|X - Y|) = 1/4$.

In a bivariate setting involving random variables X and Y, there is a particular expectation that is often of interest. It is called **covariance** and is given by

$$\boxed{Cov(X, Y) \; = \; E(\; (X - E(X))(Y - E(Y)) \;)} \tag{5.11}$$

where the expectation is taken over the bivariate distribution of X and Y. From (5.11), it is evident that covariances are positive when large X's tend to occur with large Y's and when small X's tend to occur with small Y's. Similarly, covariances are negative when large X's tend to occur with small Y's and when small X's tend to occur with large Y's. As an exercise, expand (5.11) and establish that $Cov(X,Y)$ can be expressed alternatively by

$$\boxed{Cov(X,Y) \;=\; E(XY) - E(X)E(Y).}$$

A scaled version of covariance is the **correlation** ρ which is given by

$$\boxed{\rho \;=\; Corr(X,Y) \;=\; \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}.}$$

It can be established that $-1 \leq \rho \leq 1$; this allows convenient calibration of the strength of the linear relationship between X and Y. The correlation ρ is the population analogue of the sample correlation coefficient r in (2.3) that is used to describe the degree of linear relationship involving paired data.

Example 5.22 Covariances and correlations are most meaningful in the continuous setting. However, nothing prevents us from doing covariance calculations with discrete distributions. Refer again to Table 5.1. In Table 5.2, we obtain the marginal distribution of (X,Y) by summing out Z. From Table 5.2, it is easy to calculate $E(X) = 2.1$, $E(Y) = 1.3$ and

$$E(XY) \;=\; (1)(1)(0.10) + 1(2)(0.05) + \cdots + (3)(2)(0.15) \;=\; 2.8.$$

We then obtain $Cov(X,Y) = 2.8 - (2.1)(1.3) = 0.07$.

	$X = 1$	$X = 2$	$X = 3$
$Y = 1$	0.10	0.50	0.10
$Y = 2$	0.05	0.10	0.15

Table 5.2: The bivariate pmf of (X,Y) obtained from Table 5.1.

Before moving on, I want to make a few more remarks concerning covariance and correlation. First, if X and Y are independent, then the covariance equals zero, and it follows that the correlation is also zero provided that both $Var(X)$ and $Var(Y)$ are nonzero. To see that the covariance is zero, without loss of generality, consider the continuous case where $f_{X,Y}$, f_X and f_Y are the joint pdf, the pdf for X and the pdf for Y respectively. We have

$$\begin{aligned} Cov(X,Y) &= E(XY) - E(X)E(Y) \\ &= \int\int xy\ f_{X,Y}(x,y)\ dx\ dy - E(X)E(Y) \\ &= \int\int xy\ f_X(x)f_Y(y)\ dx\ dy - E(X)E(Y) \\ &= \int y\ f_Y(y) \int\ xf_X(x)\ dx\ dy - E(X)E(Y) \\ &= \int y\ f_Y(y)E(X)\ dy - E(X)E(Y) \\ &= E(X) \int y\ f_Y(y)\ dy - E(X)E(Y) \\ &= E(X)E(Y) - E(X)E(Y) \\ &= 0. \end{aligned}$$

Second, as mentioned previously with respect to the correlation coefficient, the presence of a strong correlation (i.e. ρ near -1 or 1) does not necessarily imply a causal relationship between X and Y. It merely denotes the presence of **linear association**. For example, if X is the number of words in an individual's vocabulary and Y is the number of cavities, there is likely a negative correlation (i.e. ρ near -1) between X and Y. However, it is obvious that X does not cause Y nor does Y cause X. Here, the relationship between X and Y may be explained by a "lurking" variable age which we denote by Z. In this case, smaller values of Z (i.e. young children) are associated with smaller values of X (i.e. smaller vocabularies), and smaller values of Z are associated with larger values of Y (i.e. more cavities).

Linear combinations of random variables arise frequently in statistics. Let a_1, a_2 and b be constants. Using some previous results involving expectations, an expression for the mean of a linear combination of X and Y is derived as follows:

$$\begin{aligned} E(a_1X + a_2Y + b) &= E(a_1X) + E(a_2Y) + E(b) \\ &= a_1E(X) + a_2E(Y) + b \end{aligned}$$

and an expression for the variance of a linear combination of X and Y is derived as follows:

$$\begin{aligned}
Var(a_1X + a_2Y + b) &= E(((a_1X + a_2Y + b) - E(a_1X + a_2Y + b))^2) \\
&= E((a_1X + a_2Y - a_1E(X) - a_2E(Y))^2) \\
&= E((a_1(X - E(X)) + a_2(Y - E(Y)))^2) \\
&= E(a_1^2(X - E(X))^2) + E(a_2^2(Y - E(Y))^2) \\
&+ E(2a_1a_2(X - E(X))(Y - E(Y))) \\
&= a_1^2Var(X) + a_2^2Var(Y) + 2a_1a_2\ Cov(X, Y).
\end{aligned}$$

In the case of m variables, we have the more general expressions:

$$\boxed{E(\sum_{i=1}^{m} a_iX_i + b) \ = \ \sum_{i=1}^{m} a_iE(X_i) + b} \tag{5.12}$$

and

$$\boxed{Var(\sum_{i=1}^{m} a_iX_i + b) \ = \ \sum_{i=1}^{m} a_i^2Var(X_i) + 2\sum_{i<j} a_ia_jCov(X_i, X_j)} \tag{5.13}$$

Make sure that you are able to derive (5.12) and (5.13), and commit these expressions to memory.

5.5 Statistics and their Distributions

It is a little weird (perhaps bad planning on my part) that we get to this stage in a course on probability and statistics, and I have yet to define the term **statistic**. A statistic is simply a function of the data. For example, given observed data $x_1, x_2, \ldots, x_n$, the mean $\bar{x}$ and the standard deviation s are both statistics. Note that statistics are quantities that we can calculate, and therefore do not depend on unknown parameters. In the same way that data $X_1, \ldots, X_n$ are random (e.g. x is a realization of the random variable X), statistics $Q(X_1, \ldots, X_n)$ are random. Moreover, statistics have associated probability distributions, and we are sometimes interested in the distributions of statistics.

Example 5.23 Consider the bivariate pmf given in Table 5.2 and define the statistic $Q = X + Y$. The distribution of Q can be obtained by considering the Q-table shown in Table 5.3. The Q-table gives the entries Q as a function of X and Y. It is then a simple matter to obtain the pmf of Q by summing the probabilities in Table 5.2 corresponding to the Q values in Table 5.3. This leads to the pmf p_Q where $p_Q(2) = 0.1$, $p_Q(3) = 0.55$, $p_Q(4) = 0.2$ and $p_Q(5) = 0.15$.

	$X = 1$	$X = 2$	$X = 3$
$Y = 1$	2	3	4
$Y = 2$	3	4	5

Table 5.3: Values $Q = X + Y$ obtained from Table 5.2.

Let's now formalize what was done in Example 5.23 in more generality. Consider a statistic $Q = Q(X_1, X_2, \ldots, X_m)$. In a discrete setting, the pmf of Q is given by

$$p_Q(q) \;=\; \sum_{x_1} \sum_{x_2} \cdots \sum_{x_m} \, p(x_1, x_2, \ldots, x_m) \tag{5.14}$$

and in a continuous setting, the pdf of Q is given by

$$f_Q(q) \;=\; \int_{x_1} \int_{x_2} \cdots \int_{x_m} f(x_1, x_2, \ldots, x_m) \; dx_1 \; dx_2 \cdots dx_m \tag{5.15}$$

where the sum in (5.14) and the integral in (5.15) are taken over the m-tuples $(x_1, x_2, \ldots, x_m)$ such that $Q(x_1, x_2, \ldots, x_m) = q$.

Clearly, the calculations required in (5.14) and (5.15) can be overwhelming. Fortunately, the use of computer **simulation** can be extremely beneficial in this regard. Although you don't see it in this course, the practice of statistics is dominated by computation. Here, we provide a flavour of the powerful things that can be accomplished via computation. Suppose then that you are interested in the calculation of (5.15). Repeat the following two steps N times where N is large (say, $N = 1,000,000$):

1. generate a realized value $(x_1, x_2, \ldots, x_m)$ according to $f(x_1, x_2, \ldots, x_m)$ (we assume that this is not too difficult to do - for example, statistical software packages provide routines for generating values from standard distributions such as the normal and the gamma)

2. calculate $Q = Q(x_1, x_2, \ldots, x_m)$

At this stage, N realized values $Q_1, Q_2, \ldots, Q_N$ have been generated. A relative frequency histogram based on $Q_1, Q_2, \ldots, Q_N$ can then be constructed which approximates the pmf of Q.

Another term that has previously been used, yet needs to be defined is **a random sample**. A random sample $x_1, x_2, \ldots, x_n$ corresponds to the realized values of the random variables $X_1, X_2, \ldots, X_n$ where the X's are independent of one another and arise from the same probability distribution. Another way of saying this is that $X_1, X_2, \ldots, X_n$ are **independent and identically distributed** or **iid**.

The mean $\bar{x}$ of a random sample is a statistic that is often of interest. Suppose that the random sample arises from a population with mean μ and variance σ^2. In other words, the X's are such that $E(X_i) = \mu$, $Var(X_i) = \sigma^2$ and $Cov(X_i, X_j) = 0$ for all i, j. Then recognizing that the sample mean is a linear combination

$$\bar{x} \;=\; \frac{1}{n}x_1 + \frac{1}{n}x_2 + \cdots + \frac{1}{n}x_n$$

it follows immediately from (5.12) and (5.13) that

$$E(\bar{X}) \;=\; \mu$$

and

$$Var(\bar{X}) \;=\; \sigma^2/n.$$

Therefore the mean (or expected value) of $\bar{X}$ is the same as the mean (or expected value) of the underlying X's and the variance of $\bar{X}$, $Var(\bar{X})$, is smaller than the variance of the underlying X's, $Var(X)$, by a factor of $1/n$.

I think the latter comment corresponds to most people's intuition; there is less variation in $\bar{X}$ than in the X's.

In discussing the mean $\bar{X}$ corresponding to a random sample, we have established a couple of nice results concerning the expectation and variance. However, more can be said when the underlying X's are normal. It is an important fact (which we will not prove) that linear combinations of normal random variables are normal. Therefore, in a normal random sample, since the corresponding mean $\bar{X}$ is a linear combination, it follows that

$$\bar{X} \sim \text{normal}(\mu, \sigma^2/n) \tag{5.16}$$

or alternatively,

$$\boxed{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \text{normal}(0,1).} \tag{5.17}$$

Example 5.24 Consider $X_1 \sim \text{normal}(4,3)$, $X_2 \sim \text{normal}(5,7)$ and $X_3 \sim \text{normal}(6,4)$ where X_1, X_2 and X_3 are independent. Obtain the distribution of $W = 2X_1 - X_2 + 3X_3 + 3$.
Solution: Noting that W is a linear combination of normals, from (5.12) and (5.13) we obtain $W \sim \text{normal}(\mu_W, \sigma_W^2)$ where $\mu_W = (2)(4) + (-1)(5) + (3)(6) + 3 = 24$ and $\sigma_W^2 = (2^2)(3) + (-1)^2(7) + (3)^2(4) = 55$.

Example 5.25 Consider $X_1 \sim \text{normal}(5,10)$ and $X_2 \sim \text{normal}(3,8)$ where $Cov(X_1, X_2) = 2$. Obtain the distribution of $Y = X_1 - X_2$.
Solution: Noting that Y is a linear combination of normals, from (5.12) and (5.13) we obtain $Y \sim \text{normal}(\mu_Y, \sigma_Y^2)$ where $\mu_Y = (1)(5) + (-1)(3) = 2$ and $\sigma_Y^2 = (1^2)(10) + (-1)^2(8) + 2(1)(-1)(2) = 14$.

5.6 The Central Limit Theorem

The **Central Limit Theorem** (CLT) is known as one of the most beautiful theorems in mathematics. It is easy to state and has far reaching implications. It is remarkable in the sense that assuming very little, the theorem tells us quite a lot.

For the CLT, we assume that the random variables $X_1, X_2, \ldots, X_n$ are iid from a population with mean μ and variance σ^2. Although we really have not assumed that much, there exist weaker versions of the CLT where even less is assumed. **The CLT states that as $n \to \infty$, the distribution of $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ converges to the distribution of a standard normal random variable.**

Let's try to digest what was stated in the previous paragraph. The CLT says that as $n \to \infty$ (i.e. as we collect more and more data), convergence takes place. However, this is not the type of convergence that you see in your first calculus course where convergence refers to the convergence of a function as its argument approaches a fixed value. Rather, we are speaking about convergence of the distribution of the random variable $(\bar{X}-\mu)/(\sigma/\sqrt{n})$. Since the random variable $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ is a function of random quantities (i.e. the X's), it must have a probability distribution. But it is certainly not obvious that the distribution should be "close" to a normal distribution as n gets large. Moreover, it should be noted that nothing was assumed about the underlying distribution of the X's. The underlying distribution might be skewed, bimodal or even discrete! Under such a wide variety of possibilities, the convergence result is remarkable.

As it is written, the CLT is not practical since it is a result based on a limit, and in practice, we never have an infinite number of observations. Instead, our practical application of the CLT requires n large, and for our purposes, this is satisfied by the $\boxed{\text{rule of thumb, } n \geq 30}$. Therefore, when $n \geq 30$, we have the approximate distribution

$$\boxed{\bar{X} \sim \text{normal}(\mu, \sigma^2/n).} \tag{5.18}$$

We note that (5.18) appears to be the same result as obtained in (5.16). However, in (5.16), we assumed that the X's are normal. In the CLT, the X's can arise from any underlying distribution.

The proof of the CLT is beyond the scope of a first course in probability and statistics. However, it is possible to appreciate the theorem, and to understand it better, we motivate it by looking at a simple example. Consider then the random variable X with pmf $p(1) = 1/4$, $p(2) = 1/4$ and $p(3) = 1/2$. The pmf of X is plotted in Figure 5.2. The distribution is about as non-normal as you can get; it is discrete and is based on only three values.

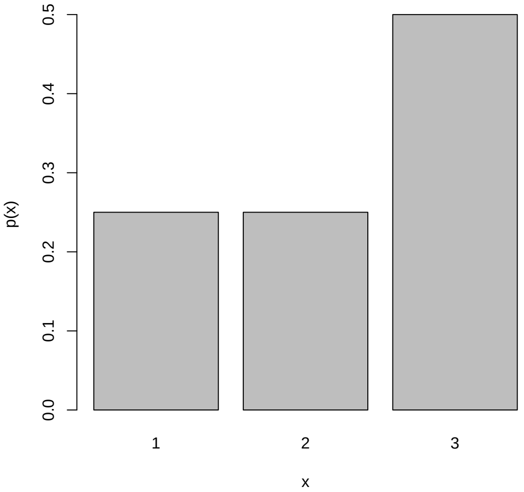

Figure 5.2: The pmf $p(x)$ based on $n = 1$.

We now consider random samples of size $n = 2$ taken from the distribution with pmf given in Figure 5.2. There are 9 possible sample pairs (X_1, X_2) with three possibilities for X_1 and three possibilities for X_2. Each pair (X_1, X_2) leads to a value $\bar{X}$ and we can calculate $P(\bar{x})$ for the possible values $\bar{x} = 1.0, 1.5, 2.0, 2.5, 3.0$. The calculation is done in a similar manner to the calculation in Example 5.23. The resulting pmf of $\bar{X}$ is given in Figure 5.3. Whereas the pmf in Figure 5.3 is far from normal, it at least has a single mode in the interior.

The process is now repeated for random samples of size $n = 3$. There are $3^3 = 27$ possible triples (X_1, X_2, X_3) each leading to a value $\bar{x}$. The pmf of $\bar{x}$ is obtained and is plotted in Figure 5.4. Even at this very early stage (i.e. $n = 3$) we begin to see some resemblance to a normal distribution. Of course, the proximity to normality improves as n gets larger. This is the essence of the Central Limit Theorem.

Example 5.26 Suppose that 500 apples are ordered and you know from pre-

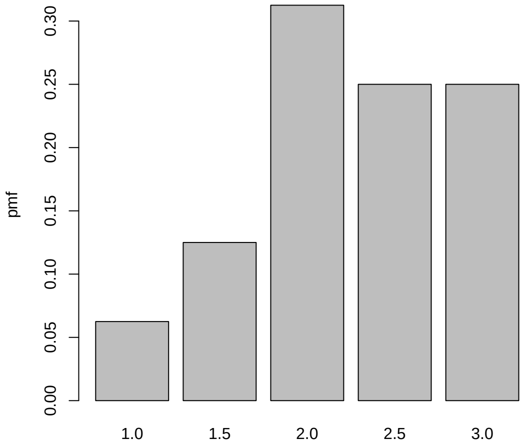

Figure 5.3: The pmf of $\bar{X}$ based on $n = 2$.

vious orders that the mean weight of an apple is 0.2 kilograms with standard deviation 0.1 kilograms. What is the probability that the total weight of the shipment is less than 98 kilograms?

Solution: When you read this for this first time, you may think that there is no way that the question can be answered since a probability is required yet no probability distribution is given. We assume that the weights of the apples $X_1, X_2, \ldots, X_{500}$ are independent. We also assume that the underlying distribution of the X's has mean $\mu = 0.2$ and standard deviation $\sigma = 0.1$. These conditions suggest the use of the CLT where we use the approximate distribution $\bar{X} \sim \text{normal}(0.2, (0.1)^2/500)$. Letting Z denote a standard normal random variable, the probabability of interest is

$$\begin{aligned} P(X_1 + X_2 + \cdots + X_{500} < 98) &= P\left(\tfrac{X_1+X_2+\cdots+X_{500}}{500} < \tfrac{98}{500}\right) \\ &= P(\bar{X} < 0.196) \\ &= P\left(\tfrac{\bar{X}-0.2}{0.1/\sqrt{500}} < \tfrac{0.196-0.2}{0.1/\sqrt{500}}\right) \\ &\approx P(Z < -0.894) \\ &= 0.185. \end{aligned}$$

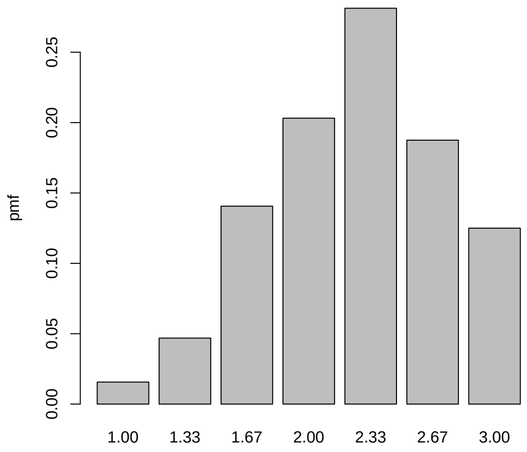

Figure 5.4: The pmf of $\bar{X}$ based on $n = 3$.

5.7 Exercises

5.01 (a) Evaluate k with respect to the pdf

$$f(x) \;=\; \begin{cases} k\exp(-2x) & x \geq 0 \\ k\exp(x) & x < 0 \end{cases} .$$

(b) Obtain the corresponding cumulative distribution function. (c) Obtain the corresponding 80-th percentile.

5.02 The shelf life, in days, for bottles of a certain prescribed medication is a random variable having the density function $f(x) = 20000/(x + 100)^3$ for $x > 0$. (a) Calculate the probability that a randomly chosen bottle has a shelf life of at least 200 days. (b) Calculate the probability that a randomly chosen bottle has a shelf life anywhere from 80 to 120 days.

5.03* The total number of hours, measured in units of 100 hours, that a family runs a vacuum cleaner during the year is a rv X with pdf

$$f(x) \;=\; \begin{cases} x & 0 < x \leq 1 \\ 2 - x & 1 < x \leq 2 \\ 0 & \text{otherwise} \end{cases} .$$

Obtain the probability that a randomly selected family runs their vacuum between 50 and 120 hours during the year.

5.04 The waiting time, in hours, between successive speeders spotted by a radar unit is a continuous random variable with cdf $F(x) = 1 - \exp(-8x)$ for $x > 0$. Obtain the probability of waiting less than 12 minutes between successive speeders.

5.05* Consider a rv X with density function $f(x) = k\sqrt{x}$ for $0 < x < 1$. (a) Evaluate k. (b) Obtain the corresponding cdf and use it to evaluate $P(0.3 < X < 0.6)$.

5.06 In a laboratory experiment, the pdf of the observed outcome X is $f(x) = 2(1 - x)$ for $0 < x < 1$. (a) Calculate $P(X \leq 1/3)$. (b) What is the probability that X exceeds 0.5? (c) Given that X exceeds 0.5, what is the probability that X is less than 0.75? (d) Obtain the 20-th percentile corresponding to X.

5.07 Suppose $X \sim \text{uniform}(1.0, 5.0)$. Obtain $P(X > 2.4 \mid X \leq 4)$.

5.08 The daily amount of coffee, in litres, dispensed by a machine is a rv X having a uniform(7.0,10.0) distribution. (a) Obtain the probability that at most 8.8 litres is dispensed. (b) Obtain the probability that at least 8.5 litres is dispensed. (c) Discuss whether you think the specified distribution is realistic.

5.09 Find the proportion X of individuals who are expected to respond to a mail-order solicitation if X has the density function $f(x) = 2(x + 2)/5$ for $0 < x < 1$.

5.10 A rv X has the density function $f(x) = \exp(-x)$ for $x > 0$. Find the expected value of $g(X) = \exp(2X/3)$.

5.11 The length of time, in minutes, for an airplane to obtain clearance for takeoff is a random variable $Y = 3X - 2$ where X has pdf $f(x) = \exp(-x/4)/4$ for $x > 0$. Find the mean and variance of Y.

5.12 Referring to Exercise 5.06, calculate the standard deviation of X.

5.13* Consider $X \sim \text{normal}(\mu, 100)$. Calculate $P(\mid X - \mu \mid \leq 5.57)$.

5.14 If $X \sim \text{normal}(3, 20)$, calculate $P((X-1)^2 < 4)$.

5.15 Loaves of rye bread from a bakery have an average length of 30.0 cm and standard deviation of 2.0 cm. Assume that the lengths of loaves are normally distributed. (a) What percentage of loaves are longer than 31.7 cm? (b) What percentage of loaves are between 29.3 cm and 33.5 cm in length? (c) What percentage of loaves are shorter than 25.5 cm?

5.16 If observations from an experiment are normally distributed, what percentage of the observations differ from the mean by more than 1.3σ where σ is the standard deviation?

5.17 A company pays its employees an average wage of \$15.90 an hour with standard deviation \$1.50 per hour. Assume that the wages are approximately normally distributed. (a) What proportion of employees receive hourly wages between \$13.75 and \$16.22? (b) What is the hourly wage that is exceeded by only 5% of the employees?

5.18 A pair of dice is rolled 180 times. (a) What is the probability that a total of 7 occurs at least 25 times? (b) What is the probability that a total of 7 occurs between 31 and 41 times inclusive?

5.19* The serum cholesterol level in 14-year old boys is approximately normally distributed with mean 170 and standard deviation 30. (a) For a randomly chosen 14-year old boy, what is the probability that his serum cholesterol level exceeds 230? (b) In a particular middle school, there are 300 14-year old boys. What is the probability that at least 8 of the boys have serum cholesterol levels that exceed 230?

5.20* Consider a discrete rv X defined on $X = 0, 1, 2, \ldots$ which is approximated by a continuous rv Y. In terms of Y, provide an expression for the approximating probability with continuity correction for $P(X > 7)$, $P(X \geq 10)$, $P(X < 4)$ and $P(X \leq 6)$.

5.21 Consider a discrete rv X defined on $X = 0.0, 0.5, 1.0, \ldots$ which is approximated by a continuous rv Y. In terms of Y, provide an expression for the approximating probability with continuity correction for $P(X > 7)$, $P(X \geq 10)$, $P(X < 4)$ and $P(X \leq 6)$.

5.22 The length of time to be served in a cafeteria is exponentially distributed with mean 4.0 minutes. What is the probability that a person is served in less than 3.0 minutes on at least four of the next six days?

5.23 If the number of calls received per hour by an answering service is a Poisson random variable with rate of 6 calls per hour, what is the probability of waiting more than 15 minutes between two successive calls?

5.24 Derive $E(X^2)$ corresponding to the rv X with pdf

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \, x^{\alpha-1} \exp(-x/\beta) \quad x > 0$$

where $\alpha > 0$ and $\beta > 0$.

5.25 Each tire on an airplane is supposed to be filled to 40 pounds per square inch. Let X and Y denote the actual pressures on the left and right tires respectively. The joint density of X and Y is $f(x, y) = k(x^2 + y^2)$ where $30 \leq x < 50$ and $30 \leq y < 50$. (a) Find k. (b) Obtain the probability that both tires are underfilled.

5.26 Obtain $P(2 < Y < 3 \mid X = 1)$ where the joint pdf of X and Y is $f(x, y) = (6 - x - y)/8$ where $0 < x < 2$ and $2 < y < 4$.

5.27 The joint pdf of X, Y and Z is $f(x, y, z) = 4xyz^2/9$ where $0 < x < 1$, $0 < y < 1$ and $0 < z < 3$. (a) Obtain the joint pdf of Y and Z. (b) Obtain the pdf of Y. (c) Obtain $P(1/4 < X < 1/2, Y > 1/3, 1 < Z < 2)$. (d) Obtain $P(0 < X < 1/2 \mid Y = 1/4, Z = 2)$.

5.28 If X and Y are independent random variables with variances 5.0 and 3.0 respectively, find the variance of $W = -2X + 4Y - 3$.

5.29 The joint density of X and Y is given by $f(x, y) = 2(x + 2y)/7$ for $0 < x < 1$ and $1 < y < 2$. Calculate the expected value of $g(X, Y) = X/Y^3 + X^2Y$.

5.30 From the joint pmf of X and Y, obtain the conditional pmf of Y given that $X \geq 2$.

	$x = 0$	$x = 1$	$x = 2$	$x = 3$
$y = 2$	0.10	0.05	0.00	0.05
$y = 3$	0.15	0.25	0.10	0.30

5.31 Assume that the random variables X and Y are uniformly distributed on a circle with radius r. This implies that the joint density is given by $f(x,y) = 1/(\pi r^2)$ where $x^2 + y^2 \leq r^2$. (a) Find the expected value of X. (b) Are X and Y independent? Explain.

5.32 Let $a \neq 0$ and $b \neq 0$ be constants. (a) Prove that $Cov(aX, bY) = ab\, Cov(X,Y)$. (b) Prove that $Corr(aX, bY) = Corr(X,Y)$.

5.33 Refer to Exercise 5.27 and calculate the covariance between X and Y.

5.34 Obtain the correlation coefficient for the random variables X and Y having joint density $f(x,y) = 2$ for $0 < x \leq y < 1$.

5.35 Obtain the correlation coefficient for the random variables X and Y having joint density $f(x,y) = 16y/x^3$ for $x > 2$ and $0 < y < 1$.

5.36 Sketch a contour plot for the joint pdf of X and Y where $X \sim$ normal$(3,1)$ independent of $Y \sim$ normal$(0,4)$.

5.37 Consider $X \sim$ binomial$(1, 1/3)$ independent of $Y \sim$ binomial$(2, 1/2)$. Obtain the pmf of $W = XY + 1$.

5.38 Suppose that the daily consumption of Pepsi in ounces by a high school student is normally distributed with $\mu = 13.0$ ounces and $\sigma = 2.0$ ounces. The daily amount consumed is independent of other days except adjacent days where the covariance is -1.0. If the student has two six-packs of 16 ounce Pepsi bottles, what is the probability that some Pepsi remains at the end of two weeks?

5.39 Evaluate k and sketch the cdf corresponding to the pdf

$$f(x) \;=\; \begin{cases} kx & 0 < x < 1 \\ \exp(-3x) & x \geq 1 \end{cases} .$$

5.40 Obtain the 80-th percentile of the normal$(100, 25)$ distribution.

5.41 Consider a shipment of 100 independent bags of fertilizer where the expected weight of each bag is 2.0 kg and the variance is 0.3 kg^2. What is the probability that the total shipment exceeds 202 kg?

5.42 The operating times in minutes of three machines are normally distributed with mean times 15, 30 and 20 minutes respectively, and standard deviations 1, 2 and 5 minutes respectively. The machines are independent except for machines 2 and 3 which have operating times that are correlated with $r = 0.1$. What is the probability that the total operating time of the three machines is at most one hour?

5.43 Suppose that $X_1, \ldots, X_n$ are iid from a distribution with mean μ and variance σ^2, n is large and $T = \sum X_i$. Obtain an approximation to $P(T > n\mu + \sqrt{n}\sigma)$

5.44 Refer to the motivation of the Central Limit Theorem in Section 5.6. Consider a discrete distribution with pmf $p(1) = 0.3$ and $p(2) = 0.7$. Obtain the distribution of $\bar{X}$ corresponding to a random sample of size $n = 5$ from the underlying distribution. Plot the resulting pmf.

Chapter 6

Inference: Single Sample

Statistical inference or **inference** addresses the following question: Given a random sample, what can be learned about the population? At first thought, this may seem like an impossibility since the sampled units from the population may be quite different from the non-sampled units in the population. Our inferential statements regarding the population are therefore couched in a language of uncertainty, and these statements utilize the probability theory developed in Chapters 3, 4 and 5.

One may think of inference as the reverse procedure to that which is used in mathematical reasoning. In mathematics, there are theorems. Theorems hold true under wide conditions and are then applied to particular cases. Therefore mathematical reasoning proceeds from the general to the specific. However, in statistical inference, we proceed from the specific (i.e. the sample) to the general (i.e. the population).

In Chapters 6 and 7, we consider statistical inference for some of the probabilistic models encountered earlier. There are three main inferential problems: **estimation**, **testing** and **prediction**. In this course, we consider only estimation and testing, and in this chapter, we restrict ourselves to single sample problems under random sampling.

6.1 Estimation

We have encountered a number of statistical models in Chapters 4 and 5. For example, $X_1, X_2, \ldots, X_n$ may be iid normal(μ, σ^2) and $Y_1, Y_2, \ldots, Y_n$ may be iid Poisson(λ). In these models, the probability distributions are characterized by parameters which are typically unknown. For example, μ and σ are parameters in the normal case and λ is a parameter in the Poisson case. One of the fundamental inferential problems is the estimation of unknown parameters or functions of parameters.

One way to estimate a parameter is to simply provide a number and this procedure is referred to as **point estimation**. For example, you might say that $\hat{\lambda} = 6.3$ is an estimate of λ. I personally think that point estimation is a bad idea since it does not provide any insight as to whether a point estimate is close to the parameter of interest. Therefore, we will not focus on the topic of point estimation. However, before moving on, I want you to at least appreciate that point estimates are not chosen willy nilly. For example, consider the case $X_1, X_2, \ldots, X_n$ iid normal(μ, σ^2) and suppose that we want to estimate μ. One thing that we ought to do is make use of the observed data $x_1, x_2, \ldots, x_n$ such that the estimate $\hat{\mu} = \hat{\mu}(x_1, x_2, \ldots, x_n)$ is a function of the data. In this example, the normal theory from Section 5.5 gives $\bar{X} \sim$ normal$(\mu, \sigma^2/n)$. If n is sufficiently large, then σ^2/n is small, and therefore the distribution of $\bar{X}$ is concentrated and centred about μ. In other words, for most datasets, $\bar{x}$ is close to μ. This suggests the point estimate $\hat{\mu} = \bar{x}$.

Instead of point estimation, our focus is on **interval estimation** where an interval of the form (a, b) is provided for an unknown parameter. The idea is that we have confidence that the unknown parameter lies in the specified interval. As with point estimation, it is sensible that the interval depend on the observed data.

6.1.1 The Normal and Large Sample Cases

Consider $X_1, X_2, \ldots, X_n$ iid normal(μ, σ^2) where our interest concerns the unknown parameter μ and we assume initially (for ease of development) that σ is known. We remark that this setting is unrealistic since it is difficult to imagine real world scenarios where you don't know μ but you happen to know σ. The normal theory from Section 5.5 gives $\bar{X} \sim$ normal$(\mu, \sigma^2/n)$ which allows us to write

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \;=\; 0.95. \tag{6.1}$$

Rearranging terms, the probability statement (6.1) is logically equivalent to

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) \;=\; 0.95 \tag{6.2}$$

where we recognize that whereas the interval in (6.1) is fixed, the interval in (6.2) is random. Using (6.2), we plug in the observed value $\bar{x}_{\text{obs}}$, and obtain a **95% confidence interval** $\bar{x}_{\text{obs}} \pm 1.96\sigma/\sqrt{n}$ for μ. More generally, letting $z_{\alpha/2}$ denote the $(1 - \alpha/2)100$-th percentile for the standard normal distribution, **a** $(1 - \alpha)100\%$ **confidence interval for** μ is given by

$$\boxed{\bar{x} \pm z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}.} \tag{6.3}$$

Note that in (6.3), we are using the observed value $\bar{x}_{\text{obs}}$, but for the sake of simplified notation, we write $\bar{x}$. From this point forward, it is understood that confidence intervals are functions of *observed* statistics.

Consider now the case where $X_1, X_2, \ldots, X_n$ corresponds to a random sample from some unspecified distribution. In other words, $X_1, X_2, \ldots, X_n$ are independent and identically distributed but we make no assumptions about the underlying distribution. We denote the mean of the underlying distribution by μ and the standard deviation by σ. Furthermore, lets assume that n is large (i.e. rule of thumb $n \geq 30$) so that the Central Limit Theorem kicks in. This implies the approximate distribution $\bar{X} \sim$ normal$(\mu, \sigma^2/n)$.

This brings us back to the distribution that allowed us to specify (6.1). Therefore, in the case when the sample size n is large, a $(1-\alpha)100\%$ confidence interval for μ is also given by (6.3).

Example 6.1 Consider heat measurements in degrees celsius where $\mu = 5$ degrees celsius and $\sigma = 4$ degrees celsius. A change is made in the process such that μ changes but σ remains the same. We observe $\bar{x} = 6.1$ degrees celsius based on $n = 100$ observations. Construct a 90% confidence interval for μ and determine the size of a random sample such that a 90% confidence interval is less that 0.6 degrees wide.
Solution: For a 90% confidence interval, $1 - \alpha = 0.90$ and hence $\alpha/2 = 0.05$. The percentile required is $z_{0.05} = 1.645$ and therefore a 90% confidence interval is $6.1 \pm 1.645(4/10) = 6.1 \pm 0.66$. The interval can also be written as $(5.44, 6.76)$. In the second part of the problem, we require that the width of the 90% confidence interval

$$(\bar{x} + 1.645(4/\sqrt{n})) - (\bar{x} - 1.645(4/\sqrt{n})) \;=\; 13.16/\sqrt{n} \;<\; 0.6.$$

Solving for n gives $n > 481.1$ and therefore the sample size n must be at least 482. This problem is an example of **statistical design** or **design**. We have used statistical theory to tell us how to conduct our experiment (i.e. determine the sample size) *before* the data are collected.

There is a further nicety in the large sample case. When n is large, it turns out that the sample standard deviation s provides a good estimate of σ. Therefore, we are able to provide a confidence interval for μ in the more realistic case where σ is unknown. We simply replace σ in (6.3) with s. It is worth reflecting on this remarkable result. Here we have a situation with a random sample $x_1, \ldots, x_n$ from any distribution imaginable, and we want to know something about the mean $E(X) = \mu$ from the population. We have assumed very little, yet $\boxed{\bar{x} \pm z_{\alpha/2} s/\sqrt{n}}$ provides us with a $(1-\alpha)100\%$ confidence interval for μ.

Let's go back from the large sample case and return to the initial premise of normal data. More specifically, assume that $X_1, X_2, \ldots, X_n$ correspond to a random sample from the normal(μ, σ^2) distribution where we are interested in μ. This time, we are going to be more realistic and assume that σ is unknown. Referring to (5.17), we replace σ with the sample standard deviation s. A result from mathematical statistics (which we will not prove) is that

$$\boxed{t_{n-1} \;=\; \frac{\bar{X} - \mu}{s/\sqrt{n}} \;\sim\; \text{Student}(n-1)} \tag{6.4}$$

and we refer to this as the **Student distribution** on $n - 1$ degrees of freedom or the $\boldsymbol{t}$ distribution on $n - 1$ degrees of freedom. The Student pdf corresponding to (6.4) is messy, and need not be memorized. It takes the form

$$f(x) = \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \frac{1}{\sqrt{\pi(n-1)}} \left(1 + x^2/(n-1)\right)^{-n/2} \qquad x \in \mathcal{R}$$

Suffice to say that the Student pdf resembles the standard normal pdf but has longer tails. In fact, as the degrees of freedom increase, the standard normal is the limiting distribution of the Student. As the Student pdf is intractable, selected probability points for the Student distribution are given in Table B.1 of Appendix B. Note that the rows correspond to the degrees of freedom, and you need to become comfortable with the Student table. Since the **pivotal quantity** $(\bar{X} - \mu)/(s/\sqrt{n})$ has the same form as the pivotal quantity $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ used in deriving (6.3), we can obtain a confidence interval for μ under the above conditions using similar steps. We obtain the interval

$$\boxed{\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}} \tag{6.5}$$

where $t_{n-1,\alpha/2}$ is the $(1 - \alpha/2)100$-th percentile of the Student$(n - 1)$ distribution. Verify your competence with the Student table B.1 by checking that $t_{15,0.01} = 2.6035$.

Let's conclude the discussion of the interval (6.5) with an amusing anecdote. Typically, when someone makes a great discovery, the phenomena is named after the discoverer (e.g. Halley's comet). In the case of the Student distribution, the distribution was discovered in 1908 by William Gosset who worked at the Guinness Brewery in Dublin, Ireland. The distribution was coined Student as Gosset worked and wrote under the pen name "Student" to conceal his identity.

At this point, I want to explain something that is subtle. I may ask you about this on the exam and many of you will get it wrong. It concerns the interpretation of confidence intervals. Suppose that you obtain a $(1-\alpha)100\%$ confidence interval $(7.2, 9.6)$ for μ. It is tempting but *incorrect* to say that with probability $1-\alpha$, μ lies in $(7.2, 9.6)$. The reason why the statement is incorrect is that in our framework, μ is a *fixed* parameter and does not have an associated probability distribution. Either μ does not lie in the interval $(7.2, 9.6)$ or it does lie in the interval. The probability is either 0 or 1.

The correct interpretation of confidence intervals relies on our frequency definition of probability discussed in Chapter 2. We can imagine many hypothetical experiments, each giving rise to a dataset where each dataset leads to a confidence interval. The parameter μ lies in some of these hypothetical confidence intervals and does not lie in some of these hypothetical confidence intervals. In the long run, μ lies in $(1-\alpha)100\%$ of the intervals. Therefore, our interpretation of confidence intervals does not concern the particular confidence interval, but rather, the process involving confidence intervals arising from repeated hypothetical experiments. We gain confidence in the particular interval by having confidence in the process. I told you that the interpretation was subtle.

There are a few remaining comments concerning the confidence interval (6.3), and these comments apply more generally to other confidence intervals. First, it can be seen algebraically from (6.3) that when n increases, the width of the confidence interval decreases. This corresponds with our intuition since more data (i.e. larger n) tells us more about the parameters and therefore we expect the interval to shrink. Second, as our confidence increases (i.e. $1-\alpha$ becomes larger), $z_{\alpha/2}$ increases and the width of the confidence interval increases. This is also intuitive since a wider interval increases our confidence that the interval captures the parameter. Third, the reason that I used the terminology "a confidence interval" is because confidence intervals are not unique. Return to the probability statement (6.1), and note that we could have instead written

$$P\left(-z_{0.01} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{0.04}\right) \;=\; 0.95.$$

leading to the 95% confidence interval $(\bar{x} - z_{0.04}\sigma/\sqrt{n}, \bar{x} + z_{0.01}\sigma/\sqrt{n})$ which is an asymmetric interval. In fact, there are an infinite number of alternative 95% confidence intervals. The **symmetric confidence intervals** are the shortest, and short intervals are preferable. From this point forward, we will only concern ourselves with symmetric confidence intervals.

6.1.2 The Binomial Case

In Chapter 4, we devoted a lot of attention to the binomial(n, p) distribution. It is applicable when we have n independent trials where each trial has a constant probability p of success. Typically, n is known, and the only unknown parameter in the binomial family is the success probability p.

Let's assume that the total number of successes $X \sim$ binomial(n, p) with $np \geq 5$ and $n(1-p) \geq 5$ so that the normal approximation to the binomial is reasonable (see Section 5.2.1). In practice, p is unknown and therefore the conditions are impossible to check. We might therefore guess the value of p, and use the guess to check the conditions. Under the normal approximation, we have $X \sim$ normal$(np, np(1-p))$ and we define $\hat{p} = X/n$ as the proportion of successes. Since $\hat{p}$ is a linear combination of a normal random variable, it follows from Section 5.6 that $\hat{p} \sim$ normal$(p, p(1-p)/n)$. We then write the probability statement

$$P\left(-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) \;=\; 0.95$$

and rearrange terms to give the logically equivalent expression

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{p(1-p)/n} < p < \hat{p} + z_{\alpha/2}\sqrt{p(1-p)/n}\right) \;=\; 0.95. \qquad (6.6)$$

From (6.6), we substitute the observed proportion of success $\hat{p}$ to obtain an approximate $(1-\alpha)100\%$ confidence interval

$$\boxed{\hat{p} \pm z_{\alpha/2} \left(\frac{\hat{p}(1-\hat{p})}{n} \right)^{1/2}} \tag{6.7}$$

for the binomial parameter p. We refer to (6.7) as an approximate confidence interval since it is based on the normal approximation to the binomial, and the substitution of p with $\hat{p}$.

Example 6.2 Suppose that 1250 voters in British Columbia are randomly selected and 420 of them claim to support the Liberal Party. Obtain a 95% confidence interval for the proportion of BC voters who support the Liberal Party.

Solution: This is an inferential problem since we are more interested in the population of BC voters than the sample of BC voters. Let X be the number of sampled voters who support the Liberals where it is reasonable to assume that $X \sim \text{binomial}(1250, p)$. We are interested in p, and using (6.7), we obtain a 95% confidence interval

$$(420/1250) \pm 1.96 \left(\frac{(420/1250)(1-420/1250)}{1250} \right)^{1/2} \;=\; 0.336 \pm 0.026.$$

We remark that we can now interpret reports from polling that you commonly hear in the news. In this example, it may be stated that 33.6% of BC voters support the Liberal Party, plus and minus 2.6%, 19 times out of 20. For us, 33.6% refers to the point estimate $\hat{p} = 0.336$, 2.6% refers to the half width of the confidence interval, and 19 times out of 20 refers to 95% confidence.

6.2 Hypothesis Testing

When you look at the world in the presence of random variation, questions arise at to whether something is true? For example, does reducing the speed limit from 100 kilometres per hour to 90 kilometres per hour on a particular stretch of highway decrease the number of motor vehicle accidents? Obviously, not everyone who travels at a higher speed will be involved in an accident. Our interest is not in observations corresponding to sampled individuals. Rather, the question that we pose is population based.

To address scientific questions in the presence of random variation, we describe a method of **hypothesis testing**. As presented here, hypothesis testing consists of the following three steps:

1. The experimenter begins by forming a hypothesis H_0 referred to as the **null hypothesis** to be tested against an **alternative hypothesis** H_1. In our framework, it is assumed that H_0 and H_1 are disjoint and are the only possible states of nature; exactly one of H_0 and H_1 must be true. In addition, H_0 and H_1 are not interchangeable. The null hypothesis H_0 represents the status quo (or no effect) and the alternative hypothesis H_1 is the state that the experimenter attempts to establish. In other words, testing begins by assuming that H_0 is true, and data is collected in an attempt to establish the truth of H_1. Since the hypotheses describe the population, and we characterize populations with statistical models, the hypotheses are statements involving parameters.

2. The second step in hypothesis testing is the data collection step. We do not concentrate on the art of data collection in this course. However, we note that it is an important step and falls under the guise of **experimental design** which is taught in subsequent courses in statistics. Obviously, you can make better decisions regarding the truth/falsity of H_0 and H_1 if you collect "good" data.

3. The third step is the inferential step which is the trickiest step. Essentially, the following question is asked "Are the data compatible with respect to H_0?" If the answer is "yes", then H_0 is not rejected. If the answer is "no", this implies that the data could not have occurred had H_0 been true. Therefore the decision is to reject H_0 in favour of H_1. In this step, you can see the elevated status that H_0 has over H_1. Testing

begins by assuming the truth of H_0, and H_0 is only overturned if the data and H_0 are incompatible.

As you study the following examples, it is instructive to review the three steps of hypothesis testing described above.

Example 6.3 Let's begin with an informal problem without any symbols and calculations. Suppose that you are playing cards with a friend and you are concerned that the so-called friend is cheating. You form the null hypothesis H_0 that the friend is playing fairly and the alternative hypothesis H_1 that the friend is cheating. Note that H_0 and H_1 are the only possibilities and that H_0 is what you would typically expect (i.e. H_0 represents the status quo). In the data collection step, you observe that the friend receive a royal flush three hands in a row. Returning to Example 3.13, you can see that the event is extremely improbable. The question is then asked, "Are the data compatible with respect to H_0?" The answer is certainly "no". It is nearly impossible for the friend to have played fairly and received a royal flush three hands in a row. The decision is to reject H_0 and conclude H_1 that the player is cheating.

From Example 6.3, I hope that you have a foretaste of what is to come. To answer the question, "Is the data compatible with respect to H_0?" requires some sort of calculation which involves assessing the probability of the data with respect to H_0. This is where all of our probability theory and statistical modelling is used to help address practical questions of interest.

Did you notice that our conclusions in hypothesis testing are either "reject H_0" or "do not reject H_0? The second option "do not reject H_0 involves curious language, and you might think that the use of a double negative is simply the case of poor grammar on the part of your instructor. However, there is reason behind my madness since "do not reject H_0" is not the same as "accept H_0" as demonstrated in the following silly example.

Example 6.4 Consider the null hypothesis H_0 that John Smith is poor, and the alternative hypothesis H_1 that John Smith is wealthy. For the sake of illustration, we do not allow the possibility that John Smith is of average income. With respect to the data, we simply have the observation that John Smith walked down the street. Is the data compatible with respect to H_0? Clearly, the answer is "yes". If John Smith is poor, he might walk down the street. And therefore, our conclusion is to not reject H_0. This is quite different from accepting H_0 which states that John Smith is poor. John Smith could very well be a wealthy man who owns an expensive car but walks to get some exercise. This is a case where the data are not sufficiently strong to reject H_0; in such cases, we can only conclude by stating that we do not reject H_0.

6.3 Examples

We now look at a series of examples that involve formal calculations to address the question "Are the data compatible with respect to H_0?" I will be careful to describe the peculiarities of each problem; the differences may involve the hypotheses considered and the corresponding distribution theory. You may see questions like these on the final exam.

Example 6.5 A shop sells coffee where the number of pounds of coffee sold in a week is normal$(320, 40^2)$. After some advertising, 350 pounds of coffee is sold in the following week. Has advertising improved business?

Solution: Since the amount of coffee sold increased from an average of 320 pounds to 350 pounds, it may seem that advertising helped. However, we don't expect exactly the same amount of coffee sold each week, and therefore the real question is whether the increase exceeded ordinary weekly variation. Of course, you should also keep in mind the comments (Section 2.5.2) about causal relationships and controlled studies where auxiliary variables should be held fixed. In a hypothesis testing problem, it is paramount that you state the assumptions that are made, and here, these are given by the statistical model $X \sim$ normal$(\mu, 40^2)$ where X is the number of pounds of coffee sold in the week following advertising. The null hypothesis $H_0 : \mu = 320$ represents the status quo (i.e. no change due to advertising) and the alternative hypothesis is $H_1 : \mu > 320$. Observe that H_0 and H_1 are statements regarding the population parameter μ. Furthermore, the hypotheses are disjoint and are assumed to be the only possibilities describing the state of the world. In the data collection step, we have $x = 350$. In the inference step, we calculate a ***p*-value** which is defined as the probability of observing data as extreme or more extreme (in the direction of H_1) than what we observed given that H_0 is true. Under H_0, we expect X to be 320, and therefore $X \geq 350$ refers to data as extreme or more extreme than what we observed. Therefore

$$\begin{aligned} p\text{−value} &= P(X \geq 350 \mid H_0 \text{ true}) \\ &= P(Z \geq (350 - 320)/40) \\ &= P(Z \geq 0.75) \\ &= 0.2266 \end{aligned}$$

where Z is standard normal. The calculation of the p-value allows us to address the question "Are the data compatible with respect to H_0?". Since the p-value is the probability of data as extreme or more extreme than our data under H_0, a small p-value suggests that the observed data are incompatible with H_0. Therefore, we reject H_0 if the p-value is small, and the rule of thumb compares the p-value with the 0.05 **level of significance** . In this case, the $p-$value $= 0.2266 > 0.05$, and the data are compatible with H_0. Therefore we do not reject H_0, and we state that there is insufficient evidence to conclude that advertising improves sales.

Example 6.6 A soup company makes soup in 10 ounce cans. A random sample of 48 cans has mean volume 9.82 ounces and $s = 0.8$ ounces. Can we conclude that the company is guilty of false advertising? Assess at 0.01 level of significance.

Solution: Let $X_1, X_2, \ldots, X_{48}$ be the iid volumes of soup in ounces arising from a population with mean μ and variance σ^2. This is the statistical model and it is important to state the assumptions and the symbols as they are used in the following calculations. To form the hypotheses, we look at the question being posed, and write $H_0 : \mu = 10$ and $H_1 : \mu < 10$. Note that false advertising corresponds to the alternative hypothesis. The data are summarized by the statistics $\bar{x} = 9.82$ ounces and $s = 0.8$ ounces. Under H_0, $E(\bar{X}) = 10.0$, and therefore, according to H_1, $\bar{x} = 9.82$ represents an extreme observation in the small sense. The p-value is obtained by recognizing that the large sample conditions for the Central Limit Theorem are satisfied which also allows us to replace the unknown σ with s. Denoting Z as a standard normal variate, we obtain

$$\begin{aligned} p\text{−value} &= P(\bar{X} \leq 9.82 \mid H_0 \text{ true}) \\ &= P(Z \leq (9.82 - 10)/(0.8/\sqrt{48})) \\ &= P(Z \leq -1.56) \\ &= 0.0594. \end{aligned}$$

Since the $p-$value $= 0.0594 > 0.01$, we do not reject H_0 and state that there is insufficient evidence to conclude that the company is guilty of false advertising.

Example 6.7 A coin is flipped 10 times and 8 heads appear. Is the coin fair?

Solution: Let X denote the number of heads in 10 flips of a coin. It is appropriate to assume $X \sim \text{binomial}(10, p)$ where p denotes the probability of a head on a single flip of the coin. The hypotheses are $H_0 : p = 1/2$ and $H_1 : p \neq 1/2$. Unlike Examples 6.5 and 6.6, we note that the alternative hypothesis has a different form which leads to a **two-tailed test**. To determine whether a test is two-tailed requires careful inspection of the question being posed. In this example, a coin that is not fair refers to both p large and p small. The observed data is $x = 8$. Under H_0, we expect $X = 5$, and according to H_1, $x = 8$ is extreme in the large sense. Therefore

$$\begin{aligned} p\text{−value} &= 2\ P(X \geq 8 \mid H_0 \text{ true}) \\ &= 2\ (P(X = 8) + P(X = 9) + P(X = 10)) \\ &= 2\left(\tbinom{10}{8}\left(\tfrac{1}{2}\right)^{10} + \tbinom{10}{9}\left(\tfrac{1}{2}\right)^{10} + \tbinom{10}{10}\left(\tfrac{1}{2}\right)^{10}\right) \\ &= 0.109 \end{aligned}$$

where the factor 2 was introduced to account for the two-tailed situation corresponding to small values of X which are also extreme. Again, the p-

value is not sufficiently small, and therefore we do not reject H_0 and state that there is insufficient evidence to conclude that the coin is not fair.

Example 6.8 A coin is flipped 100 times and 70 heads appear. Is the coin fair?
Solution: This problem is similar to Example 6.7 except that we have a larger sample which facilitates the normal approximation to the binomial (Section 5.2.1). Let $X \sim \text{binomial}(100, p)$ be the number of heads in 100 flips of a coin where $x = 70$ is observed. The hypotheses are the same as in Example 6.7 and the p-value is

$$\begin{aligned} p\text{−value} &= 2\ P(X \geq 70 \mid H_0 \text{ true}) \\ &= 2\ P(Z \geq (69.5 - 50)/\sqrt{25}) \\ &= 2\ P(Z \geq 3.9) \\ &= 0.000048 \end{aligned}$$

where a continuity correction has been introduced. In this problem, the $p-$value is much less than 0.05 and we strongly reject H_0. We conclude by stating that there is strong evidence that the coin is not fair. It is interesting to compare Examples 6.7 and 6.8. In the former, the percentage of heads is 0.8 which is more extreme than in the latter where the percentage of heads is 0.7. Yet, the evidence against H_0 was weaker in the former than the latter. This shows the obvious importance of sample size in assessing evidence.

Example 6.9 Paint is applied to tin panels and the panels are baked for one hour such that the mean index of durability is 35.2. Suppose that 20 test panels are baked for three hours and the sample mean index of durability is 37.2 with $s = 1.4$. Does baking for three hours strengthen panels? Assume normal data.
Solution: Let $X_1, X_2, \ldots, X_{20}$ denote the durability indices which we assume are iid from a normal(μ, σ^2) distribution where σ is unknown. The hypotheses are $H_0 : \mu = 35.2$ and $H_1 : \mu > 35.2$, and the data are summarized by the statistics $\bar{x} = 37.2$ and $s = 1.4$. With a sample size $n = 20$, this is not a large sample problem. However, the Student distribution is appropriate.

The p-value is given by

$$\begin{aligned} p\text{−value} &= P(\bar{X} \geq 37.2 \mid H_0 \text{ true}) \\ &= P(t_{19} \geq (37.2 - 35.2)/(1.4/\sqrt{20})) \\ &= P(t_{19} \geq 6.39) \\ &\approx 0.0. \end{aligned}$$

We therefore strongly reject H_0 and conclude that baking for three hours strengthens tin panels.

6.4 More on Hypothesis Testing

In hypothesis testing, we make statements concerning populations based on random samples. Therefore, it is possible to make mistakes. In Table 6.1, we illustrate the four possible outcomes in hypothesis testing where there are two hypotheses and two possible decisions. We observe that correct decisions correspond to the off diagonal entries. That is, we make a correct decision if we reject H_0 when H_1 is true, or if we do not reject H_0 when H_0 is true. Incorrect decisions correspond to the diagonal entries, and they are referred to as **Type I error** and **Type II error** respectively.

Since good tests make few errors, we first speculate whether it is possible to have a perfect test (i.e. a test with no errors). To see that perfect tests do not exist, imagine a test that always rejects H_0. Such a test is absurd because it does not even consider the data in its decision making process. However, such a test avoids Type II error completely and is perfect in that sense. The difficulty is that the test always makes a Type I error when H_0 is true. Similarly, a test that never rejects H_0 avoids Type I error completely but always makes a Type II error when H_1 is true.

The lack of a perfect test suggests that a compromise is required where some amount of Type I error and some amount of Type II error is tolerated. The standard approach is to fix the probability of Type I error at some acceptable level and allow the probability of Type II error to be a function of the test characteristics. The reason why Type I error is fixed is that it

	Parameter Space	
	H_0 true	H_1 true
Reject H_0	Type I Error	No Error
Do not reject H_0	No Error	Type II Error

Table 6.1: The four possible outcomes in hypothesis testing.

is generally seen to be more serious than Type II error as illustrated in the following example.

Example 6.10 Consider a criminal trial where H_0 corresponds to innocence and H_1 corresponds to guilt. After hearing the evidence (i.e. the data), the jury either issues a verdict of guilty (i.e. reject H_0) or not guilty (i.e. do not reject H_0). From the point of view of western societies, a Type I error is very serious as it corresponds to sending an innocent person to jail. A Type II error is not quite as bad as it corresponds to letting a guilty person go free. Finally, this example also highlights the curious language "do not reject H_0" since a verdict of not guilty does not necessarily imply that the accused is innocent. Remember, in 1995, OJ Simpson was found not guilty.

In Table 6.2, we present again the four possible outcomes of hypothesis testing and define the probabilities corresponding to the outcomes of interest. Following the previous discussion, a good test is one where α and β are small. In other words, a good test is one where the **significance level** α is small and the **power** $1 - \beta$ is large. At this stage, it is instructive to note that our null hypotheses have been typically **simple** which means that H_0 is completely specified (e.g. $H_0 : \mu = 5$ specifies μ completely). However, alternative hypotheses are often **composite** which means that H_1 is not completely specified (e.g. $H_1 : \mu > 54$ offers a range of values for μ). The consequence of this is that the power and the Type II error probability can only be calculated for specified values of H_1.

Example 6.11 Consider iid random variables $X_1, X_2, \ldots, X_{100}$ from a population with mean μ, standard deviation $\sigma = 1.8$ and where $H_0 : \mu = 3.0$

	Parameter Space	
	H_0 true	H_1 true
Reject H_0	α	Power $= 1 - \beta$
Do not reject H_0		β

Table 6.2: The probabilities of the outcomes of interest in hypothesis testing.

and $H_1 : \mu > 3.0$. Calculate the power of the test at $\mu = 3.2$ where the significance level $\alpha = 0.05$.

Solution: The first thing to notice is that this is a design question where properties of the test procedure are addressed prior to data collection. If we are unhappy with the properties, then we may tinker with the sample size, initially set at $n = 100$. Before calculating the power, we need to find the **critical region** or **rejection region** which is the subset of the sample space for which H_0 is rejected. Recognizing that this is a large sample problem and noting the form of the hypotheses, we reject H_0 for large values of $\bar{x}$. Therefore, we need to find a such that

$$P(\bar{X} \geq a \mid H_0 \text{ true}) \;=\; 0.05. \tag{6.8}$$

Letting Z denote a standard normal variate, (6.8) is equivalent to

$$P(Z \geq (a - 3.0)/(1.8/\sqrt{100})) \;=\; 0.05.$$

From the standard normal table, we obtain $1.645 = (a - 3.0)/(1.8/\sqrt{100})$ which yields $a = 3.296$. Therefore the critical region is given by the subset of the sample space $\{X_1, X_2, \ldots, X_{100} : \bar{X} > 3.296\}$. Since power is the probability of rejecting H_0 when H_1 is true, we calculate

$$\begin{aligned} \text{Power} &= P(\bar{X} \geq 3.296 \mid \mu = 3.2) \\ &= P(Z \geq (3.296 - 3.2)/(1.8/\sqrt{100})) \\ &= P(Z \geq 0.53) \\ &= 0.30. \end{aligned}$$

This is not a very powerful test since it only makes the correct decision 30% of the time when $\mu = 3.2$.

Example 6.12 Repeat Example 6.11 but calculate the power at $\mu = 3.5$.
Solution: Intuitively, we expect greater power in this example since it is easier to distinguish $\mu = 3.5$ from $\mu = 3.0$ than it is to distinguish $\mu = 3.2$ from $\mu = 3.0$. In this case,

$$\begin{aligned} \text{Power} &= P(\bar{X} \geq 3.296 \mid \mu = 3.5) \\ &= P(Z \geq (3.296 - 3.5)/(1.8/\sqrt{100})) \\ &= P(Z \geq -1.13) \\ &= 0.87 \end{aligned}$$

which is indeed greater than in Example 6.11.

Example 6.13 As a continuation of Example 6.11, what is the power of the test at $\mu = 3.2$ when the sample size changes from 100 to 400?
Solution: Our intuition suggests that power should increase since we have greater discernment with larger sample sizes. Following Example 6.11, we revise the calculation $1.645 = (a^* - 3.0)/(1.8/\sqrt{400})$ for the critical region which yields $a^* = 3.148$. We then calculate

$$\begin{aligned} \text{Power} &= P(\bar{X} \geq 3.148 \mid \mu = 3.2) \\ &= P(Z \geq (3.148 - 3.2)/(1.8/\sqrt{400})) \\ &= P(Z \geq -0.58) \\ &= 0.72 \end{aligned}$$

which is greater than in Example 6.11.

Having introduced a number of different inferential procedures, it is helpful to review and reflect on what has been done. In Table 6.3, we list the various estimation and testing scenarios. When you come across a single sample word problem, assess the stated assumptions and then determine which of the six scenarios is applicable. The corresponding pivotal will then help you proceed.

Our final topic in this chapter concerns significance. When the null hypothesis is rejected, it is often said that the result is **statistically significant**. I now make three points concerning the significance of significance. These are good things to keep in mind when reading articles with statistical content.

Sample Data	Pivotal	Comments
normal, σ known	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \text{normal}(0,1)$	unrealistic
normal, σ unknown	$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim \text{Student}(n-1)$	approx normal(0,1) when $n \geq 30$
non-specified, σ known, $n \geq 30$	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \text{normal}(0,1)$	unrealistic based on CLT
non-specified, σ unknown, $n \geq 30$	$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim \text{normal}(0,1)$	based on CLT
binomial	binomial	
binomial, $np \geq 5$, $n(1-p) \geq 5$	$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim \text{normal}(0,1)$	replace p with H_0 value or $\hat{p}$ in denominator for CI

Table 6.3: Summary of single sample inference based on random samples.

1. Sometimes experimenters simply report that a result is significant at significance level α. This means that the $p-$value $< \alpha$. However, it is better to report the p-value since p-values calibrate the degree of evidence against H_0. Certainly, a p-value of 0.00001 provides much stronger evidence than a p-value of 0.04 even though both are statistically significant at $\alpha = 0.05$.

2. Although it has a long history, keep in mind that the significance level $\alpha = 0.05$ has been arbitrarily chosen. There is very little difference in the strength of evidence between a p-value of 0.53 and a p-value of 0.48 even though the former is not significant and the latter is significant. There is nothing magical about $\alpha = 0.05$.

3. Statistical significance does not necessarily mean importance. For example, it was reported in Yahoo News, February 18, 1998 that in a study of 507,125 military recruits, the average height for those born in the spring is 1/4 inch greater than those born in the fall. Yes, there is a statistical difference in heights between those born in the spring and those born in the fall. However, the result cannot be viewed as important. The result makes no difference in terms of military clothing, military activities, etc. The difference of 1/4 inch is simply too small to matter. The reason why the difference is detectable is due to the large sample size. Moreover, the result is not surprising since military recruits are typically young men in their late teens some of whom may not have completed their growing. It is somewhat natural that those born two seasons earlier may be a little taller. As a final comment, sometimes other factors should also be considered when weighing importance. For example, suppose that in the comparison of two drugs, there is statistical significance indicating that drug A is slightly better than drug B. However price may be an issue, for example, if drug A is five times more expensive than drug B.

6.5 Exercises

6.01* A tire company claims that its tires can withstand 85 pounds of pressure per square inch. From past data, it is assumed that the pressure capacity of tires is normally distributed with variance 12 pounds^2. Five randomly chosen tires are slowly inflated until they burst. Their bursting pressures in pounds are 89.2, 87.0, 79.1, 84.5 and 86.7. (a) Give the statistical model for the experiment. (b) State the relevant hypotheses. (c) Calculate the p-value. (d) Provide a concluding sentence. (e) If the variance were unknown, how would the p-value change?

6.02 Consider $X_1, \ldots, X_{100}$ iid. Obtain a 90% confidence interval for the mean response of X having observed $\bar{x} = 3.5$ and $s^2 = 1.44$.

6.03 A mill produces logs whose weights (measured in kg) are normally distributed with $\sigma = 120$ kg. The milling process is modified, but the modification does not affect the normality assumption nor σ. A random sample of 25 logs yields an average weight of 8439 kg. Calculate a 92% confidence interval for the mean weight of a log produced under the modified milling process.

6.04 Let t_η denote a rv having the Student(η) distribution. (a) What happens to $P(t_\eta \leq 2)$ as η increases? Explain. (b) What happens to $P(t_\eta > 0)$ as η increases? Explain.

6.05 Based on a random sample of size $n = 40$, the 95% confidence interval for the true mean weight in mg for beans of a certain type is (229.7,233.5). Obtain the 99% confidence interval.

6.06* From a random sample of n Canadians, 30% support the Liberal Party. The 95% confidence interval (0.238,0.362) is obtained for the true proportion of Canadians who support the Liberal Party. Determine n.

6.07 A random sample of 10 chocolate energy bars has an average of 230 calories per bar with a sample standard deviation of 15 calories. Construct a 99% confidence interval for the true mean calorie content of energy bars. Assume that the calorie contents are approximately normally distributed.

6.08* A fridge is maintained at 8 degrees celsius. In a test at 11 degrees, spoilage was detected in 11 out of 55 salad dishes. Construct a 90% confidence interval for the true proportion of spoilage in salad dishes at 11 degrees celsius.

6.09 Consider a large lot of electronic components where there is interest in the proportion θ of working components. Suppose it is known for sure that $0.7 < \theta < 0.9$. How many components ought to be randomly sampled to provide an 80% confidence interval with length less than 0.01?

6.10 The following is a true story. While teaching in Indonesia, Professor Swartz met a man having four wives who gave birth to 40 children. Of these children, 38 were boys! It is conjectured that girls are born with probability 0.53. Test the conjecture. State the statistical model, the hypotheses and provide a p-value. What do you think was really going on?

6.11 The following $n = 16$ measurements were recorded for the drying times in hours for a brand of latex paint:

3.4	2.5	4.8	2.9	3.6	2.8	3.3	5.6
3.7	2.8	4.4	4.0	5.2	3.0	4.8	4.5

Assuming normality of the measurements, obtain a 95% confidence interval for the mean drying time in hours.

6.12 An allergist wishes to test the hypothesis that at least 30% of the public is allergic to some cheese products. (a) State the null and alternative hypotheses. (b) What is the implication of a Type I error? (c) What is the implication of a Type II error?

6.13 A random sample of 64 bags of white cheddar popcorn has an average weight of 5.23 ounces with sample standard deviation 0.24 ounces. Test the hypothesis $H_0 : \mu = 5.5$ versus $H_1 : \mu < 5.5$ where μ is the mean weight in ounces of a bag of white cheddar popcorn.

6.14* A sailor is thinking about sailing his boat across the Pacific Ocean and wants to make sure that his boat is seaworthy. Some tests are carried out on the boat. Without statistical notation, state the null hypothesis, the Type I error and the Type II error. Typically, Type I error is more important than Type II error. Discuss whether this is the case in this example.

6.15 A new radar device is being considered for a missile defence system. The system is checked by experimenting with aircraft in which a kill or a no kill is simulated. There is interest in whether the kill percentage exceeds 80%. In 300 trials, 250 kills are observed. (a) State the statistical model and hypotheses. (b) Test at the 0.04 level of significance and provide a conclusion.

6.16 Suppose $X = 6$ is observed where $X \sim \text{binomial}(36, p)$ in a test of $H_0 : p = 1/4$ versus $H_1 : p \neq 1/4$. Consider a second experiment where $Y \sim \text{binomial}(100, p)$. When compared to the first experiment, find the value y such that $Y = y$ provides less evidence against H_0 and $Y = y + 1$ provides more evidence against H_0.

6.17 The pmf of a random variable X is given below.

x	1	2	3	4	5
$p(x)$	$.1 + \theta$	$.2 + \theta$	$.2 + 3\theta$	$.1 - \theta$	$.4 - 4\theta$

In a test of $H_0 : \theta = 0$ versus $H_1 : \theta = 0.1$, H_0 is rejected if $X = 1$, $X = 2$ or $X = 3$. (a) What is the probability of a Type I error? (b) Is this a powerful test?

6.18 Consider a random sample of size $n = 4$ from a normal(μ, σ^2) distribution where we wish to test $H_0 : \mu = 1$ versus $H_1 : \mu < 1$. The critical region for rejecting H_0 is $\{X_1, X_2, X_3, X_4 : \bar{X} < -1.0\}$. Although the data have not yet been collected, we are confident that $s \leq 1.96$. Give an upper bound for the power of the test when $\mu = 0.60524$.

6.19 A board game enthusiast wonders whether the die provided in her board game is fair. She rolls the die 25 times and observes "6" only one time. State the relevant statistical model and hypotheses. Calculate the p-value and provide a conclusion. If she was only going to roll the die 25 times, what might she have done better?

6.20 A curing process for cement results in a mean compressive strength of 5000 kg per cm^2 with standard deviation 120 kg per cm^2. A new curing process is proposed, and a random sample of 50 pieces of cement is tested. To test $H_0 : \mu = 5000$ versus $H_1 : \mu < 5000$ with respect to the new curing process, the critical region $\bar{X} < 4970$ is determined. (a) Provide the statistical model stating all assumptions. (b) What is the probability of committing a Type I error? (c) Evaluate the power of the test when $\mu = 4960$.

6.21 Consider a test with critical region $\{X_1, \ldots, X_n : Q(\underline{X}) \leq a_1\}$ and a second test with critical region $\{X_1, \ldots, X_n : Q(\underline{X}) \leq a_2\}$ where $a_2 < a_1$. (a) What can you say about the significance level of test 1 compared to the signficance level of test 2? (b) What can you say about the power of test 1 compared to the power of test 2?

6.22 Let $X \sim \text{binomial}(5, \theta)$, and consider a test of $H_0 : \theta = 0.5$ versus $H_1 : \theta < 0.5$. If H_0 is rejected when $X = 0$, calculate the power when $\theta = 0.1$.

Chapter 7

Inference: Two Samples

In this chapter, we continue with the inferential themes presented in Chapter 6 where our focus is now on two-sample problems. In two-sample problems, there are data from two groups and our primary objective is to investigate differences between the groups. In the following sections, we consider different cases of two-sample problems. We will not go over every possible type of problem (e.g. estimation, testing, design, power calculations) that can arise for every possible case; you will need to figure things out for yourselves based on the distribution theory and the principles that have been developed.

7.1 The Normal Case

We begin with a case that is not so realistic but provides a good starting point for ease of instruction. Assume that $X_1, X_2, \ldots, X_m$ are iid observations from a normal(μ_1, σ_1^2) population and that $Y_1, Y_2, \ldots, Y_n$ are iid observations from a normal(μ_2, σ_2^2) population. Assume further that the X's are independent of the Y's. Our interest is in the difference $\mu_1 - \mu_2$, and the unrealistic part of the problem is that σ_1 and σ_2 are assumed known. Note that the sample sizes m and n in the two groups do not need to be the same.

In order to carry out inference, we require some distribution theory. Based on the normal theory from Section 5.5, $\bar{X} \sim$ normal$(\mu_1, \sigma_1^2/m)$ which is independent of $\bar{Y} \sim \text{normal}(\mu_2, \sigma_2^2/n)$. And, since $\bar{X} - \bar{Y}$ is a linear combination of independent normals, it again follows from Section 5.5 that $\bar{X} - \bar{Y} \sim \text{normal}(\mu_1 - \mu_2, \sigma_1^2/m + \sigma_2^2/n)$. After standardizing, this leads to the pivotal quantity

$$\boxed{\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sim \text{normal}(0,1)} \tag{7.1}$$

which is used for inference in the normal two-sample problem when σ_1 and σ_2 are known.

Example 7.1 Suppose that two types of teaching instruction (A and B) are being investigated. Two classes are formed where fifty test scores from class A are independent of 70 test scores from class B. Suppose that the scores are normal and the population variances are $\sigma_1^2 = \sigma_2^2 = 84.0$ (the unrealistic part). The observed data are summarized by the sample mean test scores 73.0 for class A and 59.0 for class B. Is there a difference between the two methods of instruction?

Solution: We interpret a difference between the two methods of instruction as a difference in the population means of the test scores. We let the X's denote the scores from class A where $m = 50$, and we let the Y's denote the scores from class B where $n = 70$. In addition, we make all of the assumptions stated at the beginning of the section. The relevant hypotheses are $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$. Note that both hypotheses are composite and that the form of the alternative hypothesis suggests a two-tailed test. It may be helpful to note that the hypotheses can be written as $H_0 : \mu_1 - \mu_2 = 0$ and $H_1 : \mu_1 - \mu_2 \neq 0$. Recall that the p-value is the probability of observing a result as extreme or more extreme than what we observed in the direction of H_1 assuming that H_0 is true, and note that $\bar{x} - \bar{y} = 73.0 - 59.0 = 14.0$ is extreme in the large sense. Therefore, denoting Z as standard normal and using (7.1) in a two-tailed test, the p-value is given by

$$\begin{aligned} p\text{−value} &= 2\ P(\bar{X} - \bar{Y} \geq 14.0 \mid H_0 \text{ true}) \\ &= 2\ P\left(Z \geq \tfrac{14.0-0}{\sqrt{84.0/50+84.0/70}}\right) \\ &= 2\ P(Z \geq 8.25) \\ &\approx 0.0. \end{aligned}$$

Therefore we strongly reject H_0 and conclude that there is a difference between the two methods of instruction.

Example 7.2 In a continuation of Example 7.1, a $(1-\alpha)100\%$ confidence interval for the mean difference $\mu_1 - \mu_2$ is obtained by referring to (7.1) and rearranging the probability statement

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

as was done in (6.2). This leads to the general $(1-\alpha)100\%$ confidence interval

$$\boxed{\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\sigma_1^2/m + \sigma_2^2/n}}$$

for the mean difference $\mu_1 - \mu_2$. In the particular example, the 95% confidence interval is $14.0 \pm 1.96(84.0/50 + 84.0/70)^{1/2} = 14.0 \pm 3.3$.

Example 7.3 In Example 7.1, investigate whether method A is better than method B.

Solution: This is a hypothesis testing problem with the same setup as before except that it corresponds to a one-tailed test where $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 > \mu_2$. In this case, we have

$$\begin{aligned} p\text{−value} &= P(\bar{X} - \bar{Y} \geq 14.0 \mid H_0 \text{ true}) \\ &= P\left(Z \geq \tfrac{14.0-0}{\sqrt{84.0/50+84.0/70}}\right) \\ &= P(Z \geq 8.25) \\ &\approx 0.0 \end{aligned}$$

and we strongly reject H_0 and state that method A is better than method B.

Example 7.4 Consider a further modification to Example 7.1 where we now ask whether method A is more than 8 marks better than method B.
Solution: We have the same setup as before except that the hypotheses become $H_0 : \mu_1 = \mu_2 + 8$ and $H_1 : \mu_1 > \mu_2 + 8$. The p-value is therefore

$$\begin{aligned} p\text{−value} &= P(\bar{X} - \bar{Y} \geq 14.0 \mid H_0 \text{ true}) \\ &= P\left(Z \geq \tfrac{14.0-8.0}{\sqrt{84.0/50+84.0/70}}\right) \\ &= P(Z \geq 3.54) \\ &\approx 0.0002 \end{aligned}$$

where the last equality was obtained via statistical software. We emphasize that 8.0 was subtracted in the second equality since it is the value of $\mu_1 - \mu_2$ under H_0. Since the p-value is small, we reject H_0 and conclude that method A has mean test scores that are more than 8 marks better than the mean test scores of method B.

Example 7.5 In a continuation of Example 7.1, suppose that $\sigma_1^2 = \sigma_2^2 = 84.0$, and the hypotheses are $H_0 : \mu_1 - \mu_2 = 3$ and $H_1 : \mu_1 - \mu_2 > 3$. Determine $m = n$ such that the test has level $\alpha = 0.01$ and Type II error probability $\beta = 0.05$ at $\mu_1 - \mu_2 = 5$.
Solution: Note that this is a design problem where characteristics of the test procedure are investigated prior to collecting the data. The Type I and Type II error probabilities translate into

$$P(\bar{X} - \bar{Y} \geq a \mid H_0 : \mu_1 - \mu_2 = 3) \;=\; 0.01 \tag{7.2}$$

and

$$P(\bar{X} - \bar{Y} \geq a \mid \mu_1 - \mu_2 = 5) \;=\; 0.95 \tag{7.3}$$

respectively where a determines the critical region. Therefore, (7.2) and (7.3) are two equations in the two unknowns a and m. After standardizing and obtaining z-values from the standard normal table, (7.2) and (7.3) can be expressed as $(a-3)/\sqrt{168/m} = 2.326$ and $(a-5)/\sqrt{168/m} = -1.645$ respectively. Solving the equations gives $m = 662.3$. Therefore, in order for the test to approximately satisfy the specified characteristics, we should choose a sample size of $m = n = 662$.

We have forcefully remarked that in most applications it is unrealistic to assume that σ_1 and σ_2 are both known. Suppose that we maintain the same assumptions described at the beginning of the section but assume that $\sigma_1 = \sigma_2$ and denote the common unknown standard deviation as σ. This is more realistic as it is often the case that the X's and the Y's are similar measurements and have approximately the same variances. In this case, instead of (7.1), the distribution that can be used for estimation, testing, design and power calculations is

$$\boxed{\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{(1/m + 1/n)s_p^2}} \sim \text{Student}(m+n-2)} \qquad (7.4)$$

where $\boxed{s_p^2 = ((m-1)s_1^2 + (n-1)s_2^2)/(m+n-2)}$ is referred to as the **pooled variance**. To have an appreciation as to why the pooled variance may be appropriate, note that the sample variances s_1^2 and s_2^2 are both estimators of the common variance σ^2. Therefore $s_p^2 \approx ((m-1)\sigma^2 + (n-1)\sigma^2)/(m+n-2) = \sigma^2$. The derivation of the distribution theory associated with (7.4) is taught in subsequent courses in statistics.

As has been done previously, a probability statement based on (7.4) can be constructed and rearranged to give a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$:

$$\boxed{\bar{x} - \bar{y} \pm t_{m+n-2,\alpha/2}\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right) s_p^2}}$$

Example 7.6 The Chapin Social Insight test was given to $m = 18$ males who recorded an average score of 25.34 and a standard deviation 13.36. The test was given independently to $n = 23$ females who recorded an average score of 24.94 and a standard deviation 14.39. Assuming normal data, does the mean score of males exceed the mean score of females?

Solution: The question concerns population characteristics, and we define the male test scores as iid observations $X_1, X_2, \ldots, X_{18}$ from a normal(μ_1, σ_1^2) distribution. Similarly, we define the female test scores as iid observations $Y_1, Y_2, \ldots, Y_{23}$ from a normal(μ_2, σ_2^2) distribution. Since s_1 and s_2 are close, it may be reasonable to assume a common variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$. The relevant hypotheses are $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 > \mu_2$, the pooled variance is $s_p^2 = ((18-1)(13.36)^2 + (23-1)(14.39)^2)/(18+23-2) = 194.61$, and it follows that the p-value is

$$\begin{aligned} p\text{−value} &= P(\bar{X} - \bar{Y} \geq 25.34 - 24.94 \mid H_0 \text{ true}) \\ &= P\left(t_{18+23-2} \geq \tfrac{0.4-0}{\sqrt{(1/18+1/23)(194.61)}}\right) \\ &= P(t_{39} \geq 0.09) \\ &\approx 0.46. \end{aligned}$$

Note that $P(t_{39} \geq 0.09)$ was obtained (Table B.2) by recognizing that when ν is large (i.e. $\nu \geq 30$), the Student distribution with ν degrees of freedom is well approximated by the standard normal distribution. Based on the p-value, we therefore conclude that there is no evidence to indicate that the mean test score of males exceeds the mean test score of females.

A formal test of the hypothesis that $\sigma_1 = \sigma_2$ is beyond the scope of this course, however, we provide a simple rule of thumb for deciding whether sample variances can be pooled to generate the pooled variance. The rule is if:

$$\frac{\max\{s_1, s_2\}}{\min\{s_1, s_2\}} \begin{cases} \leq 1.4 & \text{assume } \sigma_1 = \sigma_2, \\ > 1.4 & \text{do NOT assume that } \sigma_1 = \sigma_2. \end{cases}$$

In the case of small normal samples when the variances are unknown, we can use the pivotal given in (7.1) replacing σ_1^2 with s_1^2 and σ_2^2 with s_2^2. This statistic no longer has a normal(0,1) distribution, but rather a Student(ν) distribution, where ν is the complicated expression given below.

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{s_1^2/m+s_2^2/n}} \sim \text{Student}(\nu)$$

$$\nu = \text{integer part}\left[\frac{(s_1^2/n_1+s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1-1}+\dfrac{(s_2^2/n_2)^2}{n_2-1}}\right]$$

7.2 The Large Sample Case

A very common experiment involves iid data $X_1, X_2, \ldots, X_m$ which is independent of iid data $Y_1, Y_2, \ldots, Y_n$. Denote the mean and standard deviation of the X's as μ_1 and σ_1 respectively. Similarly, denote the mean and standard deviation of the Y's as μ_2 and σ_2 respectively. If m and n are both large (i.e. $m, n \geq 30$), then the Central Limit Theorem can be used to obtain the approximate distribution

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/m+\sigma_2^2/n}} \sim \text{normal}(0,1) \tag{7.5}$$

which is exactly the same as (7.1). However, because we are in the large sample case, s_1 well approximates σ_1, and s_2 well approximates σ_2. Therefore, we can replace the σ's in (7.5) with the s's and obtain the approximate distribution

$$\boxed{\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{s_1^2/m+s_2^2/n}} \sim \text{normal}(0,1).} \tag{7.6}$$

The distribution in (7.6) is very practical and can be used for estimation, testing, design and power calculations.

Example 7.7 A college interviewed 1296 students concerning summer employment. Amongst the 675 upper year students interviewed, the mean summer salary was \$5884.52 with standard deviation \$3368.37. Amongst the 621 lower year students interviewed, the mean summer salary was \$4360.39

with standard deviation \$2037.46. Test whether there is a mean difference in earnings between lower and upper year students.

Solution: The conditions for the large sample problem are satisfied where the X's represent the summer incomes of upper year students, and the Y's represent the summer incomes of lower year students. The hypotheses are $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$, and using (7.6), the p-value is given by

$$\begin{aligned} p\text{−value} &= 2\ P(\bar{X} - \bar{Y} \geq 5884.52 - 4360.39 \mid H_0 \text{ true}) \\ &\approx 2\ P\left(Z \geq \frac{1524.13-0}{\sqrt{3368.37^2/675+2037.46^2/621}}\right) \\ &= 2\ P(Z \geq 9.94) \\ &\approx 0.0 \end{aligned}$$

where Z is a standard normal variate. We therefore strongly reject H_0 and conclude that there is a difference in mean summer incomes between upper year and lower year students.

Example 7.8 The test scores of first year students admitted to college directly out of high school historically exceed the test scores of first year students with working experience by 10%. A random sample of 50 first year students admitted to college directly out of high school gave an average score of 74.1% with standard deviation 3.8%. An independent random sample of 50 first year students with working experience gave an average score of 66.5% with standard deviation 4.1%. Test whether the gap between the two groups of students has narrowed.

Solution: The conditions for the large sample problem are satisfied where the X's represent the test scores of first year students admitted directly out of high school, and the Y's represent the test scores of first year students with working experience. Note that the hypotheses $H_0 : \mu_1 = \mu_2 + 10$ and $H_1 : \mu_1 < \mu_2 + 10$ are a little unusual. Since $\bar{x} - \bar{y} = 74.1 - 66.5 = 7.6$ is small with respect to H_0 in the direction of H_1, the p-value based on (7.6) is given by

$$\begin{aligned} p\text{−value} &= P(\bar{X} - \bar{Y} \le 7.6 \mid H_0 \text{ true}) \\ &\approx P\left(Z \le \tfrac{7.6-10.0}{\sqrt{3.8^2/50+4.1^2/50}}\right) \\ &= P(Z \le -3.04) \\ &\approx 0.0012 \end{aligned}$$

where Z is a standard normal variate. This leads us to reject H_0 and conclude that the gap between the two groups of students has narrowed.

Example 7.9 As a continuation of Example 7.8, calculate the power of the $\alpha = 0.05$ level test corresponding to $\mu_1 = \mu_2 + 8$. Assume that $\sigma_1 = \sigma_2 = 4.0$.
Solution: When $\mu_1 = \mu_2 + 8$, we would like our test to reject $H_0 : \mu_1 = \mu_2 + 10$. The power tells us how frequently this is done. Using (7.5), the critical region is $\{\bar{X}, \bar{Y} : \bar{X} - \bar{Y} < 8.68\}$ since

$$P\left(\frac{\bar{X} - \bar{Y} - 10}{\sqrt{4^2/50 + 4^2/50}} \le -1.645\right) = 0.05.$$

The power is therefore

$$\begin{aligned} \text{Power} &= P(\bar{X} - \bar{Y} < 8.68 \mid \mu_1 - \mu_2 = 8) \\ &= P\left(Z < \tfrac{8.68-8.0}{\sqrt{4^2/50+4^2/50}}\right) \\ &= P(Z < 0.85) \\ &= 0.80. \end{aligned}$$

7.3 The Binomial Case

The two-sample binomial is one of the most common experimental designs. It consists of two independent binomial experiments where there is interest in the difference between the success rates.

Recall the conditions of a binomial experiment where there are m trials each resulting in either a success or failure. The trials are independent, and each trial has probability p_1 of success. In this case, the total number of successes X has the binomial(m, p_1) distribution. Following Section 5.2.1, when m is large and p_1 is moderate (i.e. $mp_1 \geq 5$ and $m(1-p_1) \geq 5$), then it is approximately the case that $X \sim \text{normal}(mp_1, mp_1(1-p_1))$. Since linear combinations of normal random variables are themselves normal (see Section 5.5), it then follows that the proportion of successes in the binomial experiment $\hat{p}_1 = X/m$ has the distribution

$$\hat{p}_1 \sim \text{normal}(p_1, p_1(1-p_1)/m). \tag{7.7}$$

Now consider a second binomial experiment which is independent of the first experiment. The second experiment is based on n trials where the probability of success on each trial is p_2. With n large and p_2 moderate, it follows that the proportion of successes $\hat{p}_2$ in the second experiment has the distribution

$$\hat{p}_2 \sim \text{normal}(p_2, p_2(1-p_2)/n). \tag{7.8}$$

With $\hat{p}_1$ independent of $\hat{p}_2$, we use (7.7), (7.8) and normal theory to obtain the pivotal quantity

$$\boxed{\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/m + p_2(1-p_2)/n}} \sim \text{normal}(0,1).} \tag{7.9}$$

As highlighted in the following example, the distribution in (7.9) is useful for problems of design, interval estimation and testing involving two-sample binomial experiments.

Example 7.10 Consider a two-sample binomial problem presented below in the form of a **two by two contingency table**. Here we have two groups of patients; the first group receives therapy 1 and the second group receives therapy 2. The patients are then assessed to see whether there is improvement from their previous conditions. We are interested in the relative efficacies of the two therapies.

	Therapy 1	Therapy 2
Improvement	35	88
No Improvement	65	112
	100	200

Solution: In a problem like this, one should first verify that the conditions leading to the derivation of (7.9) are approximately true. For example, do the patients consist of two random samples? If they are not random samples, then it may be unreasonable to assume that all individuals from a therapy group have the same probability of improvement. Furthermore, it would be best if all 300 patients comprised a random sample from the same population. Otherwise, a difference in the two groups could be due to some factor other than a difference due to therapies.

To address the question of whether there is a difference in the improvement rates between the two therapy groups, we test $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$. The observed data is given by $\hat{p}_1 = 0.35$ and $\hat{p}_2 = 0.44$. The denominator in (7.9) is problematic as it is based on an unknown p_1 and an unknown p_2. Under H_0, we therefore estimate the common $p = p_1 = p_2$ with $\tilde{p} = (X+Y)/(m+n) = (35+88)/(100+200) = 0.41$. A two-tailed test based on (7.9) yields

$$\begin{aligned} p\text{−value} &= 2\ P(\hat{p}_1 - \hat{p}_2 \leq 0.35 - 0.44) \\ &= 2\ P\left(Z \leq \frac{0.35-0.44-(p_1-p_2)}{\sqrt{\tilde{p}(1-\tilde{p})/m+\tilde{p}(1-\tilde{p})/n}}\right) \\ &= 2\ P\left(Z \leq \frac{0.35-0.44-0}{\sqrt{(0.41)(0.59)(1/100+1/200)}}\right) \\ &= 2\ P(Z \leq -1.494) \\ &= 0.135 \end{aligned}$$

which suggests no statistical difference between the two therapies. Note that leading into the experiment, there may have been some inkling that therapy 2 was the preferred therapy. If the investigation had instead focused on whether therapy 2 was superior, then a one-tail test yields the $p-$value 0.068 which is significant at the 0.10 level of significance.

To construct an 80% confidence interval for the difference $p_1 - p_2$ in the improvement rates, we invert (7.9) as was done in Section 6.1.1 and we substitute $\hat{p}_1$ and $\hat{p}_2$ in the denominator giving

$$\begin{aligned}
& \hat{p}_1 - \hat{p}_2 \pm z_{0.1}\sqrt{\hat{p}_1(1-\hat{p}_1)/m + \hat{p}_2(1-\hat{p}_2)/n} \\
= \;& 0.35 - 0.44 \pm 1.282\sqrt{(0.35)(0.65)/100 + (0.44)(0.56)/200} \\
= \;& -0.09 \pm 0.076 \\
\to \;& (-0.166, -0.014).
\end{aligned}$$

7.4 The Paired Case

In Section 2.5, we discussed paired data in the context of descriptive statistics. Paired random variables arise in the form $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ where there is a natural relationship between each X and Y in a pair. For example, a pair may correspond to two measurements on the same subject. In this section, we consider inference for paired data.

The analysis of paired data is actually quite simple. For each pair (X_i, Y_i), $i = 1, \ldots, n$, we reduce the pair to a difference $D_i = X_i - Y_i$ and then proceed with the differences as though they arose in a single sample problem. We illustrate the approach in the following four examples.

Example 7.11 Suppose scores of jitteriness (is that a word?) are taken on a group of 8 subjects where a high score indicates jitteriness. The scores are taken twice; once before drinking coffee and once after drinking coffee. The data are presented in Table 7.1. Assuming that the scores are normal, test at significance level $\alpha = 0.1$ whether there is an increase in jitteriness due to drinking coffee.

Solution: In this problem, we assume that the scores before drinking coffee $X_1, X_2, \ldots, X_8$ are iid normal(μ_1, σ_1^2) and the scores after drinking coffee $Y_1, Y_2, \ldots, Y_8$ are iid normal(μ_2, σ_2^2). However, unlike Sections 7.1 and 7.2, the X's are not independent of the Y's. There is a natural pairing between X_i and Y_i for $i = 1, \ldots, 8$ since the measurements correspond to the same individual. Since linear combinations of normal random variables are normal

x (Before Coffee)	y (After Coffee)	d (Difference)
50	56	-6
60	70	-10
55	60	-5
72	70	2
85	82	3
78	84	-6
65	68	-3
90	88	2

Table 7.1: Scores taken on subjects before and after drinking coffee.

(Section 5.5), $D_1, D_2, \ldots, D_8$ are iid normal$(\mu_1 - \mu_2, \sigma^2)$ where $D_i = X_i - Y_i$, $i = 1, \ldots, 8$ and $\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2Cov(X, Y)$. Under these conditions (i.e. normal data with an unknown σ), the relevant pivotal is given by (6.4). The hypotheses are $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 < \mu_2$, and the data are summarized by $\bar{d} = -2.875$ and $s_d = 4.734$ (the sample standard deviation of the d's). Therefore, the p-value is given by

$$\begin{aligned} p\text{−value} &= P(\bar{D} \leq -2.875 \mid H_0 \text{ true}) \\ &= P\left(t_7 \leq \tfrac{-2.875-0}{\sqrt{4.734^2/8}}\right) \\ &= P(t_7 \leq -1.72). \end{aligned}$$

From Table B.1, we observe that the p-value lies somewhere between 0.05 and 0.10. Since the problem asked to test at level $\alpha = 0.1$ significance, we reject H_0 and conclude that drinking coffee increases mean jitteriness scores.

Example 7.12 Referring to Example 7.11, obtain a 95% confidence interval for the mean difference in jitteriness scores.

Solution: In this problem, the pivotal (6.4) translates to

$$\frac{\bar{D} - (\mu_1 - \mu_2)}{s_d/\sqrt{8}} \sim \text{Student}(8-1).$$

Therefore, writing down the probability statement and rearranging terms, the 95% confidence interval for $\mu_1 - \mu_2$ is given by

$$\begin{aligned} \bar{d} \pm t_{7,0.025}(s_d/\sqrt{8}) &= -2.875 \pm (2.365)(4.734)/\sqrt{8} \\ &= -2.875 \pm 3.958. \end{aligned}$$

Example 7.13 Suppose that instead of 8 subjects with before and after measurements, 16 subjects were enlisted where 8 of them had uncaffeinated coffee and the remaining 8 had regular coffee. Using the same data as in Table 7.1, how does the analysis change?

Solution: In this case, there is no natural pairing, and the X's are independent of the Y's. This is a two-sample problem with normal data. Assuming common variances in the two groups, the relevant pivotal is given in (7.4). With pooled variance $s_p^2 = 168.813$, the p-value is given by

$$\begin{aligned} p\text{−value} &= P(\bar{X} - \bar{Y} \leq -2.875 \mid H_0 \text{ true}) \\ &\approx P\left(t_{8+8-2} \leq \frac{-2.875-0}{\sqrt{(1/8+1/8)(168.813)}}\right) \\ &= P(t_{14} \leq -0.44) \\ &= 0.33 \end{aligned}$$

where the last equality was obtained from statistical software. In this case, we do not reject H_0 and state that there is no evidence that drinking regular coffee increases jitteriness. As a side issue, we note that half of the subjects consumed uncaffeinated coffee. In the example, uncaffeinated coffee served as a **placebo** where subjects did not know which treatment had been given to them. Placebos act as controls against psychological factors that sometimes influence test scores.

The comparison of the p-values in Example 7.11 and Example 7.13 are noteworthy because the same data were used but the conclusions are different. It appears as though the paired test (Example 7.11) is more sensitive than the non-paired test (Example 7.13). That is, we are better able to detect differences between the coffee drinkers and the non-coffee drinkers when the data are paired. The reason for the increased sensitivity is that the paired test removes differences due to the individuals. Looking at Table 7.1 in the context of Example 7.11, there are great differences between the individuals as scores range from the 50's to the 90's. However, there is less variation in the differences d which range from -10 to 3. Five of the eight individuals had scores that increased.

The phenomenon of pairing is a specific case of the more general statistical technique of **blocking**. Blocking attempts to remove variation by grouping observations that are similar, and this sometimes leads to more sensitive tests. In Example 7.11, the blocks correspond to the 8 "groups" of individuals each of whom has two measurements. As an investigator, sometimes you have the choice to block. If you believe that differences between individuals are large, then pairing may be useful. Do note however that when pairing (or blocking), there is a tradeoff to consider as degrees of freedom are lost (e.g. t_{14} in Example 7.13 versus t_7 in Example 7.11). Recall that a Student distribution with lower degrees of freedom has longer tails. An intuitive way of understanding the tradeoff is by recognizing that when pairing, you essentially cut your number of observations in half.

Example 7.14 Finally, return to Example 7.11 and suppose that the observations were taken on twins. For example, X_1 may be the score of Charles Keizer who did not consume coffee, and Y_1 may be the score of his twin Angus Keizer who had consumed coffee. What is the approach in this problem?

Solution: Although we do not have repeated measurements on the same subject, there is a natural pairing and therefore the analysis of Example 7.11 is appropriate. However, if you believe that twins don't necessarily share jitteriness properties, then you may be better off proceeding with the analysis in Example 7.13.

Having introduced a number of different inferential procedures for two-sample problems, it is helpful to review and reflect on what has been done. In Table 7.2, we list the various estimation and testing scenarios. When you come across a two-sample word problem, assess the stated assumptions and then determine which of the five scenarios is applicable. The corresponding pivotal will then help you proceed.

Sample Data	Pivotal	Comments
paired data, $m = n$	take $D_i = X_i - Y_i$ and refer to single sample case	
non-paired, m, n large	$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/m+\sigma_2^2/n}} \sim \text{normal}(0,1)$	replace σ_i's with s_i's if σ_i's unknown
non-paired, m, n not large, data normal, σ_i's known	$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/m+\sigma_2^2/n}} \sim \text{normal}(0,1)$	unrealistic
non-paired, m, n not large, data normal, $\sigma_1 \approx \sigma_2$ unknown	$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right)s_p^2}}$ $\sim \text{Student}(m+n-2)$	$s_p^2 = \frac{(m-1)s_1^2+(n-1)s_2^2}{m+n-2}$ $\frac{\max\{s_1,s_2\}}{\min\{s_1,s_2\}} \leq 1.4$
non-paired, m, n not large, data normal, $\sigma_1 \neq \sigma_2$ but unknown	$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\left(\frac{s_1^2}{m}+\frac{s_2^2}{n}\right)}} \sim \text{Student}(\nu)$	$\nu =$ integer part $\frac{(s_1^2/n_1+s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}}$
binomial data, m, n large, p_1, p_2 moderate	$\frac{\hat{p}_1-\hat{p}_2-(p_1-p_2)}{\sqrt{p_1(1-p_1)/m+p_2(1-p_2)/n}}$ $\sim \text{normal}(0,1)$	replace p_i's with H_0 estimates or with $\hat{p}_i$'s in denominator for CI

Table 7.2: Summary of two-sample inference where $X_1, \ldots, X_m$ are iid with mean μ_1 and standard deviation σ_1, and $Y_1, \ldots, Y_n$ are iid with mean μ_2 and standard deviation σ_2.

7.5 Exercises

7.01 Two kinds of thread are compared for strength. Fifty pieces of each type of thread are tested under similar conditions. Brand A has an average tensile strength of 78.3 kg with a sample standard deviation of 5.6 kg, while Brand B has an average tensile strength of 87.2 kg with a sample standard deviation of 6.3 kg. Construct a 95% confidence interval for the difference in population means.

7.02* It is conjectured that driver's training improves one's chances of passing their driver's test. From a school where driver's training is offered, 89 of 100 students who took driver's training passed their test. From a group of 100 individuals who did not take driver's training, only 74 passed their test. (a) Write down the statistical model and relevant hypotheses. (b) Calculate the corresponding p-value and provide conclusions. (c) Is there any cause for worry in this experiment that assesses driver's training?

7.03 Before and after measurements with respect to an exercise program were taken on 8 subjects. It is assumed that the measurements (heartbeats per minute) are normally distributed. We are interested in whether the program decreased mean counts by at least 5.0 beats per minute.

Subject	1	2	3	4	5	6	7	8
Before	73	81	55	62	69	71	85	52
After	70	73	49	51	59	64	79	48

(a) Write down an appropriate statistical model stating all assumptions and introducing relevant notation. (b) Write down the null and alternative hypotheses. (c) Obtain the corresponding p-value and give a concluding statement.

7.04* An experiment was carried out to investigate the amount of residual chemicals in the soil due to two different treatments. Treatment 1 had six measurements: 28.5, 24.7, 26.2, 23.9, 29.6, 29.1. Treatment 2 had five measurements: 38.7, 41.6, 35.9, 41.8, 42.0. Assume normality of the measurements. (a) Write down an appropriate statistical model stating all assumptions and introducing relevant notation. (b) Write down the null and alternative hypotheses corresponding to the question "Is there a difference in the two treatments regarding the amount of residual chemicals?" (c) Obtain the corresponding p-value and give a concluding statement.

7.05 In an experiment at Virginia Tech, 20 northern oak seedlings were exposed to a fungus. Half of the seedlings received 368 ppm of nitrogen (treatment) during planting, and the other half received no nitrogen (control). The stem weights in grams, at the end of 140 days were recorded as follows:

Treatment	.32	.53	.28	.37	.47	.43	.36	.42	.38	.43
Control	.26	.43	.47	.49	.52	.75	.79	.86	.62	.46

Construct a 95% confidence interval for the difference in the mean stem weights between the treatment and control groups. Assume that the underlying populations are normally distributed with equal variances.

7.06 Amstat News 2011 reports on a survey of salaries in the United States of full professors with 6-10 years of service at the given rank. The average salary of 59 statisticians is \$168,500 with an estimated standard deviation of \$41,480. The average salary of 37 biostatisticians is \$188,100 with an estimated standard deviation of \$41,185. Test whether the mean yearly income of biostatisticians is at least \$10,000 greater than the mean income of statisticians. State the statistical model and hypotheses. (P.S. I am a full professor in statistics with 6-10 years of service, and let me tell you, I don't make anywhere near this kind of money.)

7.07 A university is considering the proposal of a new policy. Not only is it important that the policy have widespread support, the university wants to have a good understanding of the difference in support between the teaching staff and the non-teaching staff. It is proposed that n teaching staff and n non-teaching staff are randomly selected and asked whether they support the new policy. Determine n so that the length of the 95% confidence interval for the difference in the percentage support between teaching staff and non-teaching staff is at most 5%.

7.08 Ten engineering schools were surveyed. The first random sample of 250 electrical engineers had 80 women, and the second random sample of 175 chemical engineers had 40 women. Obtain an 80% confidence interval for the difference in proportions of women in these two fields of engineering.

7.09 Consider iid random variables $X_1, \ldots, X_n$ of mother's ages in years arising from a population with mean μ_1 and variance σ_1^2. Corresponding to X_i, is the age Y_i in years of the mother's eldest child. The ages of eldest children who have mothers that are still alive arise from a distribution with mean μ_2 and variance σ_2^2. There is interest in testing $H_0 : \mu_1 - \mu_2 = 20$ versus $H_1 : \mu_1 - \mu_2 > 20$. Approximately how large should n be such that the power of the test is 0.90 when $\mu_1 - \mu_2 = 25$ years? You will need to make an assumption concerning variances; try to make it realistic.

7.10* A group of 100 students are randomly assigned to classes of size 50 and 50. Before a test, the students in the first group are instructed to pull an all-nighter (i.e. no sleep the previous night) whereas the second group has their normal amount of sleep. Suppose we know that the population variance for test scores is 98.2 (likely an unrealistic assumption) and that the sample mean test scores from the two groups are 75.1 and 82.3 respectively. Assume further that test scores are normally distributed. (a) Write down an appropriate statistical model stating all assumptions and introducing relevant notation. (b) Write down the null and alternative hypotheses corresponding to the question "Does sleep deprivation hinder performance in test scores by more than 3%?" (c) Obtain the corresponding p-value and give a concluding statement. (d) Provide a 99% confidence interval for the mean improvement in test scores based on normal sleep versus deprived sleep.

7.11 There are plans to carry out a survey concerning the planned tolling of the Port Mann Bridge, We denote the mean support for the tolling initiative as p_1 in the Fraser Valley and p_2 in the Lower Mainland. Let X be the number of positive responses from n respondents randomly chosen from the Fraser Valley and let Y be the number of positive responses from n respondents randomly chosen from the Lower Mainland. We want to design a test of the hypothesis $H_0 : p_1 = p_2$ versus $H_1 : p_1 < p_2$ where H_0 is rejected if $X - Y < a$. Determine $a(n)$ (i.e. a is a function of n) such that the test makes a Type I error with probability no greater than 0.05.

7.12 The average points scored per game for each NBA team for the seasons concluding in 2009 and 2010 are given in the following table:

	Pho	GS	Den	Ut	Tor	Mem
2009	109.4	108.6	104.9	103.2	99.0	93.9
2010	109.7	108.8	106.6	104.5	104.1	102.5
	Hou	NY	Orl	Cle	LAL	Dal
2009	97.5	105.2	100.1	99.9	105.9	101.7
2010	102.4	102.1	101.9	101.9	101.6	101.4
	OKC	SA	Ind	Atl	NO	Sac
2009	97.0	96.7	105.1	96.5	95.1	100.6
2010	101.0	100.8	100.8	100.2	100.2	100.0
	Min	Bos	Por	Phi	Chi	Mil
2009	97.8	101.1	98.6	96.8	102.7	99.3
2010	98.2	98.0	97.8	97.7	97.4	97.1
	Was	Mia	LAC	Cha	Det	NJ
2009	96.1	97.5	95.1	93.6	93.5	98.1
2010	96.2	96.0	95.7	94.8	94.0	92.4

There is some speculation that the scoring rate has changed between seasons. (a) Write down an appropriate statistical model stating all assumptions and introducing relevant notation. (b) Write down the null and alternative hypotheses. (c) Obtain the corresponding p-value and give a concluding statement. (d) If the comparison had involved the years 2010 and 1995, would your method of analysis be the same? Explain.

7.13 During the regular season of the NBA, the great Larry Bird made 3,960 free throws out of 4,471 attempts for a career free throw percentage of 88.6%. As of March 21/11, British Columbia's own Steve Nash is the current all-time NBA leader with a success percentage of 90.4% based on 2,794 made free throws out of 3,091 attempts. We are interested in whether Nash is really better than Bird, or whether the percentage discrepancy is due to natural variation. (a) Propose a statistical model, and discuss whether it is adequate. (b) State relevant hypotheses, calculate a p-value and provide a conclusion.

Index

Appendix A: Solutions to Selected Exercises

2.01* Unfortunately, the process of marking is subject to some variability. Markers may not be consistent with part marks, tired markers may give lower grades, etc. It is therefore desirable that the final letter grade not be subject to these variabilities. Suppose a professor imposed a strict rule whereby 80% or higher in the numerical score receives an A grade. This would be unfortunate for the student who scored 79.6% as there is clearly not much difference between this student and the student who scored 80.3%. Professor Swartz lets the data determine the grades. When he looks at a dotplot with large gaps between scores, he uses the gaps to distinguish letter grades. Therefore his grades are not as susceptible to variability in marking. In larger classes, the numerical scores are more closely clustered together, and gaps are less apparent.

2.07* As of March 2011, the world record for 100 metres was 9.58 seconds held by Usain Bolt of Jamaica. This could be regarded as a lower bound for the data. We would expect most healthy people to be able to run 100 metres somewhere between 15 and 25 seconds. However, there would be some infirm individuals who are unable to run, and would take a very long time walking 100 metres. Therefore we expect a right-skewed distribution where the mean exceeds the median.

2.09* The median of the dataset retaining the incorrect value is 23 kg. (a)

There are three values which lead to a change in the median. If 16 $\rightarrow$ 61, then the median changes to 24 kg. If 19 $\rightarrow$ 91, then the median changes to 24 kg. If 23 $\rightarrow$ 32, then the median changes to 24 kg. (b) My instinct is that 23 kg was incorrectly transposed from 32 kg. Transposing any of the other values would give a measurement which would be a more extreme outlier with respect to the dataset.

2.12* There are an infinite number of solutions to the problem, and probably an infinite number of ways of arriving at a solution. Here is a straightforward approach. Lets keep things really simple, and have symmetric data with n even, half of the x_i's equal to $50.0 - a$ and half of the x_i's equal to $50.0 + a$. This gives $\bar{x} = 50.0$. We have some flexibility left, and to make things as simple as possible, set $n = 2$. To force $s = 50.0$, and referring to (2.1), this means $(x_i - \bar{x})^2 = 1250$ for $i = 1, 2$. This implies $\mid x_i - 50.0 \mid = \sqrt{1250} = 25\sqrt{2}$ which implies $a = 25\sqrt{2}$. Therefore $x_1 = 50.0 - 25\sqrt{2}$ and $x_2 = 50.0 + 25\sqrt{2}$ form a dataset which gives $\bar{x} = s = 50.0$.

2.22* There is no inherent pairing between the x's and the y's. Hence the scatterplot is inappropriate. A boxplot with two groups may be better, but the size of the dataset is likely too small. It is probably best to simply report some numerical descriptive statistics for each group.

3.03* Since mothers are older than daughters, $S = \{(x, y) : y > x > 0\}$. The sample space may be further restricted by considering $y_{\min}$, the minimum age for a woman to give birth and $y_{\max}$, the maximum age for human beings. Unbelievably (see http://youngest_mother.tripod.com), $y_{\min}$ may be close to 5.7 years, and $y_{\max}$ may be close to 125 years. In this case, $S = \{(x, y) : y_{\min} < y < y_{\max},\ 0 < x < y - y_{\min}\}$.

3.07* From a Venn diagram, $B = A \cup \overline{A}B$. Using axiom (3), it follows that $P(B) = P(A) + P(\overline{A}B)$. From axiom (1), $P(\overline{A}B) \geq 0$ which establishes $P(B) \geq P(A)$.

3.14* We have $P(A \mid B) = P(A)$ and by the formula for conditional probability, this implies $P(AB)/P(B) = P(A)$. Rearranging gives $P(AB)/P(A) =$

$P(B)$ which implies $P(B \mid A) = P(B)$ by conditional probability.

3.16* Denote the availability of the two engines by E_1 and E_2. Then the probability of interest is $1 - P(E_1 \cup E_2) = 1 - (P(E_1) + P(E_2) - P(E_1 E_2)) = 1 - (0.96 + 0.96 - 0.96(0.96)) = 0.0016$ using independence.

3.19* The probability is given by $\binom{10}{1}\binom{4}{2}\binom{9}{2}\binom{4}{1}\binom{4}{1}/\binom{40}{4} = 0.37816$ where we note that there are 40 cards in the deck and there are 10 denominations. The first term refers to choosing the denomination which forms the pair, the second term refers to choosing the two cards of the selected denomination, the third term refers to choosing the denominations of the remaining two cards, the fourth and fifth terms give us the cards of the denominations corresponding to the non-pair, and the sixth term is the number of ways of choosing four cards from the deck.

3.24* Let N denote a negative result and S denote sick. Then the probability of interest is $P(N) = P(NS \cup N\overline{S}) = P(NS) + P(N\overline{S}) = P(N \mid S)P(S) + P(N \mid \overline{S})P(\overline{S}) = 0.01(1/1000) + 0.96(999/1000) = 0.96$.

4.02* (a) We note that the last flip is a tail and that there is one other tail imbedded in the remaining $x - 1$ flips. This gives the pmf $p(x) = \binom{x-1}{1}(0.6)^{x-2}(0.4)^2$ for $x = 2, 3, 4, \ldots$. (b) For $5 \leq x < 6$, $F(x) = p(2) + p(3) + p(4) + p(5) = 0.663$.

4.03* (a) Enumerate the 16 outcomes and let $p(x)$ denote the pmf. Then $p(2) = 1/16$, $p(3) = 2/16$, $p(4) = 3/16$, $p(5) = 4/16$, $p(6) = 3/16$, $p(7) = 2/16$ and $p(8) = 1/16$. (b) In the limit as the number of rolls becomes large, the average is the expectation $E(X) = 2(1/16) + 3(2/16) + \cdots + 8(1/16) = 5$. (c) We calculate $E(X^2) = 2^2(1/16) + 3^2(2/16) + \cdots + 8^2(1/16) = 27.5$, $Var(X) = E(X^2) - E^2(X) = 27.5 - 5^2 = 2.5$, and therefore $SD(X) = \sqrt{2.5} = 1.58$.

4.07* (a) Since the tickets are iid over the period, $X \sim \text{binomial}(40, 0.35)$. (b) The expected ticket revenue is therefore $E(g(X))$ where $g(x) = x + 3(40 - x)$ dollars. The expectation simplifies to $E(X) + 3(40 - E(X)) = 40(0.35) + 3(40 - 40(0.35)) = 92.00$ dollars. (c) The variance is $Var(X + 3(40 - X)) =$

$Var(120 - 2X) = 2^2 Var(X) = 4(40)(0.35)(0.65) = 36.40$ dollars^2.

4.12* Let X be the number of chain smokers admitted to the hospital. (a) Then $X \sim$ binomial$(10, 0.7)$ and $P(X < 5) = 0.047$ using a hand calculator. (b) Similarly, $X \sim$ binomial$(20, 0.7)$ and $P(X < 10) = 0.017$.

4.18* Let r be the radius of the trap and let X be the number of beetles in the trap such that $X \sim$ Poisson$(3\pi r^2)$ with pmf $p(x)$. Then the probability that at least one beetle is in the trap is $1-p(0) = 1-(3\pi r^2)^0 \exp(-3\pi r^2)/0! = 0.9$. Solving $\exp(-3\pi r^2) = 0.1$ gives $r = 0.494$ metres.

5.03* The probability is given by the integral $\int_{0.5}^{1.2} f(x)\ dx = \int_{0.5}^{1.0} x\ dx + \int_{1.0}^{1.2} (2-x)\ dx = [{}_{0.5}^{1.0}\ x^2/2\ ] + [{}_{1.0}^{1.2}\ -(2-x)^2/2\ ] = (0.5 - 0.125) + (0.5 - 0.32) = 0.555$.

5.05* (a) Since $f(x)$ is a pdf, $1 = \int_0^1 f(x)\ dx = [{}_0^1\ (2/3)kx^{3/2}\ ] = k(2/3)$ which gives $k = 1.5$. (b) The cdf is given by $F(x) = \int_0^x f(x)\ dx = x^{3/2}$ for $0 < x < 1$ and the probability of interest is $F(0.6) - F(0.3) = 0.30$.

5.13* Letting Z denote a standard normal rv, the probability is $P(\mu - 5.57 \leq X \leq \mu + 5.57) = P(-0.557 \leq Z \leq 0.557) = 1 - 2P(Z \geq 0.557) = 0.42$.

5.19* (a) Letting X denote the serum cholesterol level and $Z \sim$ normal$(0, 1)$, then the probability of interest is $P(X > 230) = P(Z > (230 - 170)/30) = P(Z > 2) = 0.0228$. (b) Letting Y denote the number of boys in the school with levels exceeding 230, $Y \sim$ binomial$(300, 0.0228)$ which may be approximated by $W \sim$ normal$(6.840, 6.684)$. Then $P(Y \geq 8) \approx P(W > 7.5) = P(Z > (7.5 - 6.840)/\sqrt{6.684}) = 0.40$.

5.20* We first note that $P(X > 7) = P(X \geq 8)$ and $P(X < 4) = P(X \leq 3)$ since X is discrete. Then, if you sketch a pmf for X and superimposes the pdf for Y, it is clear that the four approximating probabilities are $P(Y \geq 7.5)$, $P(Y \geq 9.5)$, $P(Y \leq 3.5)$ and $P(Y \leq 6.5)$ respectively.

6.01* (a) Let X_i denote the bursting pressure in pounds of the i-th tire. It is assumed that $X_1, \ldots, X_5$ are iid from the normal(μ, σ^2) distribution. (b) The null hypothesis is $H_0 : \mu = 85$ which denotes the status quo, and the alternative hypothesis is $H_1 : \mu < 85$. (c) The $p-$value is given by $P(\bar{X} < 85.3) = P(Z < (85.3 - 85)/\sqrt{12/5}) = P(Z < 0.1936) = 0.58$. (d) There is no reason to reject the company's claim that the tires can withstand 85 pounds of pressure. (e) We have $s^2 = 14.785$ and the test is based on the Student distribution giving $P(t_4 < (85.3 - 85)/\sqrt{s^2/n}) = P(t_4 < 0.1744)$ which is less than 0.58 due to the thicker tails of the Student and the fact that $0.1744 < 0.1936$.

6.06* Assume that the sample size n is large such that the normal approximation to the binomial is appropriate. Then the confidence interval is $\hat{p} \pm z_{0.05/2}\sqrt{\hat{p}(1-\hat{p})/n} \rightarrow 0.30 \pm 1.96\sqrt{(0.3)(0.7)/n}$. Therefore the half-length of the interval $0.062 = 1.96\sqrt{(0.3)(0.7)/n}$ which gives $n = 209.9$ which rounds to the integer $n = 210$.

6.08* Let X be the number of spoiled dishes where $X \sim \text{binomial}(55, p)$. With the estimate $\hat{p} = 11/55$, we note that $n\hat{p} = 11 \geq 5$ and $n(1-\hat{p}) = 44 \geq 5$ which suggests the adequacy of the normal approximation to the binomial. The 90% confidence interval is therefore given by $\hat{p} \pm z_{.05}(\hat{p}(1-\hat{p})/n)^{1/2} = 11/55 \pm 1.645((11/55)(44/55)/55)^{1/2} = 0.20 \pm 0.09$.

6.14* The null hypothesis is that the boat is seaworthy. The Type I error is that the boat is declared unseaworthy when it is seaworthy. The Type II error is that the boat is declared seaworthy when it is unseaworthy. Here a Type II error is possibly more important as the sailor may die in an unseaworthy boat. A Type I error may simply mean that the sailor does some unnecessary repairs on the boat.

7.02* (a) Let X and Y be the numbers of students who passed their driver's test from each of the two groups. We assume $X \sim \text{binomial}(100, p_1)$ which is independent of $Y \sim \text{binomial}(100, p_2)$. We are interested in testing $H_0 : p_1 = p_2$ versus $H_1 : p_1 > p_2$. (b) We have $\hat{p}_1 = 0.89$, $\hat{p}_2 = 0.74$ and $\tilde{p} = 0.815$ leading to the $p-$value given by $P(\hat{p}_1 - \hat{p}_2 \geq 0.15) = P(Z \geq (0.15 - 0)/\sqrt{(0.815)(0.185)(2/100)}) = P(Z \geq 2.73) = 0.003$. The small p-value suggests that there is an advantage in taking driver's training with respect to passing the test. (c) One should be cautious as there may be something different about the students who take driver's training. For example, perhaps they have a more serious personality, and it is this trait which allows them to excel in the test rather than the training component.

7.04* (a) Let $X_1, \ldots, X_6$ denote the measurements corresponding to treatment 1 which are assumed iid normal(μ_1, σ_1^2). Let $Y_1, \ldots, Y_5$ denote the measurements corresponding to treatment 2 which are assumed independent of the X's and are iid normal(μ_2, σ_2^2). In addition, the assumption $\sigma = \sigma_1 = \sigma_2$ is required. (b) The null hypothesis is $H_0 : \mu_1 = \mu_2$ and the alternative is $H_1 : \mu_1 \neq \mu_2$. (c) The p-value corresponds to the probability under H_0 of observing a result as extreme or more extreme than what was observed and is given by $p-$value $= 2\,P(\bar{X} - \bar{Y} \leq 27.0 - 40.0) = 2\,P(t_9 \leq -13.0/(s_p\sqrt{(1/m + 1/n)}) = 2\,P(t_9 \leq -13.0/(2.522\sqrt{(1/6 + 1/5)}) = 2\,P(t_9 \leq -8.51) \approx 0.000$. We therefore strongly conclude that there is a difference between the two treatments in terms of the mean amount of residual chemicals.

7.10* (a) Let $X_1, \ldots, X_{50}$ denote the test scores from the sleep-deprived students which are assumed iid normal$(\mu_1, 98.2)$. Let $Y_1, \ldots, Y_{50}$ denote the test scores from the students receiving normal sleep which are assumed independent of the X's and are iid normal$(\mu_2, 98.2)$. (b) The null hypothesis is $H_0 : \mu_1 + 3 = \mu_2$ and the alternative is $H_1 : \mu_1 + 3 < \mu_2$. (c) The p-value corresponds to the probability under H_0 of observing a result as extreme or more extreme than what was observed and is given by $p-$value $= P(\bar{X} - \bar{Y} \leq -7.2) = P(Z \leq (-7.2 + 3)/\sqrt{2(98.2)/50}) = 0.017$. Based on the 0.05 level of significance, we therefore conclude that sleep deprivation does decrease mean test scores by more than 3%. (d) The corresponding 99% confidence interval for $\mu_2 - \mu_1$ is $7.2 \pm z_{0.005}\sqrt{(98.2)/50 + (98.2)/50} \rightarrow 7.2 \pm 5.1 \rightarrow (2.1, 12.3)$.

Appendix B: Statistical Tables

			α		
ν	0.100	0.050	0.025	0.010	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
12	1.356	1.782	2.179	2.681	3.055
14	1.345	1.761	2.145	2.624	2.977
16	1.337	1.746	2.120	2.583	2.921
18	1.330	1.734	2.101	2.552	2.878
20	1.325	1.725	2.086	2.528	2.845
22	1.321	1.717	2.074	2.508	2.819
24	1.318	1.711	2.064	2.492	2.797
26	1.315	1.706	2.056	2.479	2.779
28	1.313	1.701	2.048	2.467	2.763
30	1.310	1.697	2.042	2.457	2.750
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
∞	1.282	1.645	1.960	2.326	2.576

Table B.1: Probability points $t_{\nu,\alpha}$ corresponding to $t_\nu \sim \text{Student}(\nu)$ where $\text{Prob}(t_\nu > t_{\nu,\alpha}) = \alpha$.

z	$F(z)$	z	$F(z)$	z	$F(z)$	z	$F(z)$	z	$F(z)$
0.00	0.5000	0.40	0.6554	0.80	0.7881	1.20	0.8849	1.60	0.9452
0.01	0.5040	0.41	0.6591	0.81	0.7910	1.21	0.8869	1.64	0.9495
0.02	0.5080	0.42	0.6628	0.82	0.7939	1.22	0.8888	1.68	0.9535
0.03	0.5120	0.43	0.6664	0.83	0.7967	1.23	0.8907	1.72	0.9573
0.04	0.5160	0.44	0.6700	0.84	0.7995	1.24	0.8925	1.76	0.9608
0.05	0.5199	0.45	0.6736	0.85	0.8023	1.25	0.8944	1.80	0.9641
0.06	0.5239	0.46	0.6772	0.86	0.8051	1.26	0.8962	1.84	0.9671
0.07	0.5279	0.47	0.6808	0.87	0.8078	1.27	0.8980	1.88	0.9699
0.08	0.5319	0.48	0.6844	0.88	0.8106	1.28	0.8997	1.92	0.9726
0.09	0.5359	0.49	0.6879	0.89	0.8133	1.29	0.9015	1.96	0.9750
0.10	0.5398	0.50	0.6915	0.90	0.8159	1.30	0.9032	2.00	0.9772
0.11	0.5438	0.51	0.6950	0.91	0.8186	1.31	0.9049	2.04	0.9793
0.12	0.5478	0.52	0.6985	0.92	0.8212	1.32	0.9066	2.08	0.9812
0.13	0.5517	0.53	0.7019	0.93	0.8238	1.33	0.9082	2.12	0.9830
0.14	0.5557	0.54	0.7054	0.94	0.8264	1.34	0.9099	2.16	0.9846
0.15	0.5596	0.55	0.7088	0.95	0.8289	1.35	0.9115	2.20	0.9861
0.16	0.5636	0.56	0.7123	0.96	0.8315	1.36	0.9131	2.24	0.9875
0.17	0.5675	0.57	0.7157	0.97	0.8340	1.37	0.9147	2.28	0.9887
0.18	0.5714	0.58	0.7190	0.98	0.8365	1.38	0.9162	2.32	0.9898
0.19	0.5753	0.59	0.7224	0.99	0.8389	1.39	0.9177	2.36	0.9909
0.20	0.5793	0.60	0.7257	1.00	0.8413	1.40	0.9192	2.40	0.9918
0.21	0.5832	0.61	0.7291	1.01	0.8438	1.41	0.9207	2.44	0.9927
0.22	0.5871	0.62	0.7324	1.02	0.8461	1.42	0.9222	2.48	0.9934
0.23	0.5910	0.63	0.7357	1.03	0.8485	1.43	0.9236	2.52	0.9941
0.24	0.5948	0.64	0.7389	1.04	0.8508	1.44	0.9251	2.56	0.9948
0.25	0.5987	0.65	0.7422	1.05	0.8531	1.45	0.9265	2.60	0.9953
0.26	0.6026	0.66	0.7454	1.06	0.8554	1.46	0.9279	2.64	0.9959
0.27	0.6064	0.67	0.7486	1.07	0.8577	1.47	0.9292	2.68	0.9963
0.28	0.6103	0.68	0.7517	1.08	0.8599	1.48	0.9306	2.72	0.9967
0.29	0.6141	0.69	0.7549	1.09	0.8621	1.49	0.9319	2.76	0.9971
0.30	0.6179	0.70	0.7580	1.10	0.8643	1.50	0.9332	2.80	0.9974
0.31	0.6217	0.71	0.7611	1.11	0.8665	1.51	0.9345	2.84	0.9977
0.32	0.6255	0.72	0.7642	1.12	0.8686	1.52	0.9357	2.88	0.9980
0.33	0.6293	0.73	0.7673	1.13	0.8708	1.53	0.9370	2.92	0.9982
0.34	0.6331	0.74	0.7704	1.14	0.8729	1.54	0.9382	2.96	0.9985
0.35	0.6368	0.75	0.7734	1.15	0.8749	1.55	0.9394	3.00	0.9987
0.36	0.6406	0.76	0.7764	1.16	0.8770	1.56	0.9406	3.04	0.9988
0.37	0.6443	0.77	0.7794	1.17	0.8790	1.57	0.9418	3.08	0.9990
0.38	0.6480	0.78	0.7823	1.18	0.8810	1.58	0.9429	3.12	0.9991
0.39	0.6517	0.79	0.7852	1.19	0.8830	1.59	0.9441	3.16	0.9992

Table B.2: Table entries for the cumulative distribution function $F(z)$ corresponding to $Z \sim \text{normal}(0, 1)$.