Theory Meets Data

A Data Scientist's Handbook to Statistics

Ani Adhikari

Editor: Dibya Jyoti Ghosh

Contributors:
Shreya Agarwal, Thomas Anthony, Bryannie Bach, Adith Balamurugan,
Betty Chang, Aditya Gandhi, Dibya Jyoti Ghosh, Edward Huang, Jiayi Huang,
J. Weston Hughes, Arvind Iyengar, Andrew Linxie, Rahil Mathur, Nishaad Navkal,
Kyle Nguyen, Christopher Sauceda, Rohan Singh, Parth Singhal, Maxwell Weinstein,
Yu Xia, Anthony Xian, Ling Xie

Contents

Chapter 1

Averages

When analyzing data, one of the first things we'd like to know about is the center of the data. The *average* or the *mean*[1] of a list of numbers, is a measure that is used to represent a "central" value of the dataset. As we will see, there is more than one reasonable definition of "central". The average is one of these.

1.1 What is an average?

For a list of numbers $x_1, x_2, \ldots, x_n$, we define the average $\bar{x}$ (x with a bar above it, read as "x bar").

Definition 1 ***Average***

$$\bar{x} = \frac{1}{n} \sum_{1}^{n} x_i \tag{1.1}$$

In other words, we take n unequal numbers, stitch them all together, and then split them into n equal pieces, each whose size is the mean. In this representation, we can consider the mean to be the "equalizing value" of the data.

Recall that constants can be moved through the sum, and so we can rearrange our definition of the average to the following:

$$\bar{x} = \sum_{1}^{n} \frac{x_i}{n} \tag{1.2}$$

illustrating that the equalization process can take place before we pool all the values together.

1.2 Perturbing the list

Understanding the equalizing or "smoothing" property of averages allows us to quickly see how the average changes when you change entries in the list. Suppose a list of dollar amounts has 100 entries in it, and one of the entries goes up by $500. When you take the average of the new list,

[1] "Mean" is another name for "average", so other sources may use μ to represent certain types of averages. That's the Greek letter m, read as "mu".

those additional 500 dollars will get split evenly 100 ways, and so the average will go up by $5. No algebra needed. All you need is the amount of change to the entry and the number of items in the list.

If you want to work through the algebra, of course you can. Say that the first value in our list, x_1, becomes k. The change c in our average (that is, the difference between the averages of our first dataset **x**: $(x_1, x_2, \ldots, x_n)$ and our second dataset **y**: $(k, x_2, \ldots, x_n)$ is given by

$$\begin{aligned} c &= |\bar{x} - \bar{y}| \\ &= \left| \sum_1^n \frac{x_i}{n} - \sum_1^n \frac{y_i}{n} \right| \\ &= \left| \frac{x_1}{n} + \sum_2^n \frac{x_i}{n} - \frac{k}{n} - \sum_2^n \frac{x_i}{n} \right| \\ &= \left| \frac{x_1}{n} - \frac{k}{n} \right| \\ c &= \left| \frac{x_1 - k}{n} \right| \end{aligned} \tag{1.3}$$

This calculation confirms that the only things we need to know to determine c are the change to the entry $(x_1 - k)$, and the total number of values being averaged (n).

As a consequence, we have the following observation: *the more values we have in our list, the smaller the effect the change to a single entry can have.*

1.3 Bounds on the Average

How big or small can the average be? A natural answer is that the average will be somewhere in between the smallest and largest value in the list.

You can formally establish these lower and upper bounds on the average. Let m be the minimum value of a list $x_1, x_2, \ldots, x_n$. Then by definition for all x_j's we have

$$\begin{aligned} x_j &\geq n \\ \frac{1}{n} \sum_n x_i &\geq \frac{1}{n} \sum_n m \\ \bar{x} &\geq \frac{nm}{n} \\ \bar{x} &\geq m \end{aligned}$$

A similar assertion can be made for the maximum M of a list of numbers, but we leave that proof to the reader. Thus for any list of numbers $x_1, \ldots, x_n$ with minimum m, maximum M, and average $\bar{x}$,

$$\boxed{m \leq \bar{x} \leq M}$$

1.4 Averaging averages

Say you have two lists of numbers $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots y_m$, and say they have averages of $\bar{x}$ and $\bar{y}$ respectively. How would we go about finding the average of a combined list of all $n + m$ entries together?

A common misconception is that we can just average the two averages(sum and divide by two), but a simple example shows this doesn't hold. The average marathon time at the Rio Olympics was 2 hours and 15 minutes, and the average marathon time for amateur runners is 4 hours and 19 minutes. If we create a group of all amateur runners as well as the Rio Olympics marathoners, do you think that the average time of that group would be 3 hours and 17 minutes, halfway between the two times? We hope not!

Clearly, we need to take into account the fact that there are many more amateur marathon runners than Rio Olympians.

To see how to do this, let us return to our two lists $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots y_m$. You can figure out the average of the overall list (often called the "pooled" list) by remembering that the average is an equalizer. The contribution of the x's to the total pot will be their sum, which is $n\bar{x}$. The contribution of all the y's will be $m\bar{y}$. So the average of the pooled list will be

$$\frac{n\bar{x} + m\bar{y}}{n+m}$$

Not formal enough for you? Ok, then let's denote the average of the entire list $x_1, x_2 \ldots x_n, y_1, y_2, \ldots y_m$ by A. Then we see

$$\begin{aligned} A &= \frac{1}{n+m}\left(\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i\right) \\ &= \frac{1}{n+m}\left(\frac{n}{n}\sum_{i=1}^{n} x_i + \frac{m}{m}\sum_{i=1}^{n} y_i\right) \\ &= \frac{n\bar{x} + m\bar{y}}{n+m} \\ &= \frac{n}{n+m}\bar{x} + \frac{m}{n+m}\bar{y} \end{aligned} \tag{1.4}$$

Rather than just just averaging the two averages, we first have to "weight" them according to the corresponding number of entries.

1.5 Another way to calculate the average

Now that we know how to put two lists together and find the average of the combined list, we have another way of finding the average of any list.

Consider the list 7, 7, 7, 8, 8. You can think of this as a pooled list, if you pool the list 7, 7, 7 and the list 8, 8. The average of the first list is 7, and the average of the second list is 8. So the average of the pooled list 7, 7, 7, 8, 8 is

$$\frac{3}{5} \cdot 7 \; + \; \frac{2}{5} \cdot 8 \tag{1.5}$$

What gets used in this calculation? We need the two distinct values in the list, namely 7 and 8. We also need the proportion of each of those values in the list. The average of the list can be thought of as the average of the distinct values weighted by their proportions.

You can extend this formally to find the average of a list $x_1, x_2, \ldots x_n$ with lots of repeating values. Say there are k distinct values $v_1, v_2, \ldots v_k$ in our list, appearing respectively with frequencies $n_1, n_2, \ldots n_k$. In the list in our numerical example above, $n = 5$ and there are two distinct values, so $k = 2$. The two distinct values are $v_1 = 7$ and $v_2 = 8$.

It should now be apparent that the average of the list is

$$\bar{x} = \sum_{i=1}^{k} \frac{n_i}{n} \cdot v_i \tag{1.6}$$

Try to do the math that proves this! Also note that for each i, the proportion of times v_i appears in the list is $p_i = \frac{n_i}{n}$. So the average can be expressed as

$$\bar{x} = \sum_{i=1}^{k} p_i \cdot v_i \tag{1.7}$$

As before, that's the average of the distinct values in the list, weighted by their proportions.

This way of expressing the average shows you that if 3/5 of a list consists of the value 7, and the other 2/5 consists of the value 8, then the average will be

$$\frac{3}{5} \cdot 7 \; + \; \frac{2}{5} \cdot 8 \tag{1.8}$$

regardless of whether the list has 5 entries or 500. In other words, the list consisting of 300 7's and 200 8's has the same average as the list 7, 7, 7, 8, 8.

1.6 Questions

1. Prove the following two simple but very useful facts about averages.

   a) If all the entries in a finite list of numbers are the same,then the average is equal to the common value of the entries.

   b) If a finite list of numbers consists only of 0's and 1's, then the average of the list is the proportion of 1's in the list.

2. Consider the list $\{1, 2, \ldots, n\}$, where n is a positive integer.

   a) Guess the average of the list and give an intuitive explanation for your guess.

   b) Prove that your guess in part **a** is correct.

   c) Let i be an element of the list; in other words, suppose i is an integer such that $1 \leq i \leq n$. Suppose the element i gets replaced by 0. By how much does the average change? If you followed what we did in class,you should be able to just write down this answer and explain it without calculation.

   d) Start with the original list $\{1, 2, \ldots, n\}$ and delete an element i. What is the average of the new list?

3. Suppose you are in a class that has the following grading scheme:

   - 70% of the grade comes evenly from two exams: a midterm and a final
   - 20% comes from homework
   - 10% comes from quizzes

You have received an average score of 93% on your homework and 75% on your quizzes. On the midterm, you scored 82%. Write down a formula for the minimal percentage score you need on the final to achieve an overall score of 90% in this course. You do not need to evaluate this expression.

4. A dataset consists of the list $\{x_1, x_2, \ldots, x_n\}$ and has average $\bar{x}$. Someone is going to pick an element of the list and I have to guess its value.I have decided that my guess will be a constant c, regardless of which element is picked. Therefore if the element picked is x_i, the error that I make will be $x_i - c$.

   Define the *mean squared error* of my guess to be

$$mse_c \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$$

   Show that the minimum value of mse_c over all c is attained when $c = \bar{x}$, in two different ways:

   a) In the definition of mse_c, replace $x_i - c$ by $(x_i - \bar{x}) + (\bar{x} - c)$ and use algebra.

   b) Use the definition of mse_c and calculus.

5. Suppose all the entries in a finite list of numbers are equal. Prove that the average of the list is equal to the common value of the entries.

6. Let $x_1, x_2, \ldots, x_n$ be a list of numbers, and let $\bar{x}$ be the average of the list.Which of the following statements **must** be true? There might be more than one such statement, or one, or none;

   a) At least half of the numbers on the list must be bigger than $\bar{x}$.

   b) Half of the numbers on the list must be bigger than $\bar{x}$.

   c) Some of the numbers on the list must be bigger than $\bar{x}$.

   d) Not all of the numbers on the list can be bigger than $\bar{x}$.

7. Suppose the list of numbers $\{x_1, x_2, \ldots, x_n\}$ has average $\bar{x}$ and the list $\{y_1, y_2, \ldots, y_m\}$ has average $\bar{y}$. Consider the combined list of $n + m$ entries $\{x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m\}$. Write a formula for the average of this combined list, in terms of $\bar{x}$, $\bar{y}$, n, and m. You do not have to prove your answer.

8. Let $\{x_1, x_2, \ldots, x_n\}$ be a list of numbers and let $\bar{x}$ denote the average of the list. Let a and b be two constants, and for each i such that $1 \le i \le n$, let $y_i = ax_i + b$. Consider the new list $\{y_1, y_2, \ldots, y_n\}$, and let the average of this list be $\bar{y}$. Prove a formula for $\bar{y}$ in terms of a, b, and $\bar{x}$.

9. Let n be a positive integer. Consider the list of even numbers $\{2, 4, 6, \ldots, 2n\}$. What is the average of this list? Prove your answer.

10. Let $\{x_1, x_2, \ldots, x_n\}$ be a list of numbers with average $\bar{x}$, and let c be a constant. Show that

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \;+\; (\bar{x} - c)^2$$

Chapter 2

Deviations

2.1 What is Standard Deviation?

Two lists of data with the same average can look quite different. For example, the lists 5, 5, 5, 5 and 3, 3, 7, 7 both have 5 as their average. But while all of the entries are equal to 5, none of the entries in the second list is 5. The second list is more "spread out" than the first. The list 1, 1, 9, 9 also has 5 as its average, and it is even more "spread out" than the 3, 3, 7, 7.

To see how far the numbers on a list are from their average, it is natural to look at distances. Suppose the list is $x_1, x_2, \ldots, x_n$ with average $\bar{x}$. For each index i in the range 1 through n define the ith *deviation from the mean* to be

$$d_i = x_i - \bar{x}$$

To see how big the deviations are, it is natural to take the average of all these deviations. Let's try it out.

$$\begin{aligned}
\text{Average Deviation} = \bar{d} &= \frac{1}{n}\sum_{i=1}^{n} d_i \\
&= \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}\right) \\
&= \frac{1}{n}(n\bar{x} - n\bar{x}) \\
&= \bar{x} - \bar{x} \\
&= 0
\end{aligned}$$

Oh no! Since all positive "distances" offset all negative ones when added together, the average deviation from mean for any data sets is always equal to 0. While that's correct, it's not helpful for our purpose, which is to find roughly how far off the numbers can be from the mean.

We have to find a way past this problem of cancellation. We have to ensure that all distances are non-negative. There are two time-honored ways of doing this. The first is to take the absolute value of each distance. But the absolute value function has some mathematical properties that make it complicated to work with – for example, it is not differentiable at 0.

The other way of getting rid of minus signs is to calculate squares. So let us find the average of the *squared deviations from mean* That is a non-negative number, but unfortunately it has different

units from the original list. For example if the numbers were money in dollars, then deviations would also be in dollars (though possibly negative), and squared deviations would be in squared dollars which is a difficult unit to interpret.

So, once we have found the average of the squared deviations, we must then take the square root to get back to the original units. This motivates the definition of the *standard deviation* of the list.

The standard deviation (SD) is the root mean square of the deviations from average. Read that definition backwards, and you'll see that it's a formula for how to calculate the SD.

Here is the definition using notation.

Definition 2 ***Standard Deviation***

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

s = the standard deviation
n = the number of values
x_i = each value in the list
$\bar{x}$ = the mean of the list

Example 2: Students' Scores

A class of 18 students took a maths test.Their scores are as below

82	63	81	95	79	90
80	75	64	74	88	72
87	77	82	78	89	84

Work out the standard deviation of students' scores.

Solution:

1. Calculate the Mean

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{(82+63+81+...+89+84)}{18} \\ &= \frac{1440}{18} \\ \bar{x} &= 80\end{aligned}$$

2. Calculate the Average Squared Distance from the Mean

For each value, subtract the mean and square the result. We then find the average of all these squared differences:

$$\begin{aligned}&\frac{1}{n}\sum_{1}^{n}(x_i-\bar{x})^2\\&=\frac{1}{18}((82-80)^2+(63-80)^2+(81-80)^2+...+(89-80)^2+(84-80)^2)\\&=\frac{1228}{18}\end{aligned}$$

3. Finally, take the square root:

$$\begin{aligned}s&=\sqrt{\frac{1228}{18}}\\&=8.260\end{aligned}$$

We say that the entries in the list are around 80, give or take about 8.3. Later in this chapter we will see precisely what that statement means.

2.2 Variance

The standard deviation is the root mean square (r.m.s.) of deviations from average. The quantity inside the square root is the *mean square of deviations from average* and is known as the **variance** of the list.

Variance has units that are hard to understand, as we have seen. But it has excellent mathematical properties. So if you want to find an SD, often a good move is to first find the variance and then take the square root.

Definition 3 ***variance*** *The variance of the list $x_1, x_2, \ldots, x_n$ is*

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

Calculating the variance based on its formal definition involves a great deal of computation which must be carried out with a calculator or computer. In this section, we'll develop a formula that allows us to compute variance much faster.

Start with the formal definition of variance, expand the square inside the sum, and then collect

terms.

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}({x_i}^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}{x_i}^2 - \frac{1}{n}\sum_{i=1}^{n}2x_i\bar{x} + \frac{1}{n}\sum_{i=1}^{n}\bar{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}{x_i}^2 - \frac{2\bar{x}}{n}\sum_{i=1}^{n}x_i + \bar{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}{x_i}^2 - 2\bar{x}^2 + \bar{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}{x_i}^2 - \bar{x}^2 \\
s^2 &= \frac{1}{n}\sum_{i=1}^{n}{x_i}^2 - \bar{x}^2
\end{aligned}
$$

Definition 4 ***Computational Formula for Variance***

$$Variance(x_1, x_2, ..., x_n) = \frac{1}{n}\sum_{i=1}^{n}{x_i}^2 - \bar{x}^2 \tag{2.1}$$

Thus the variance is **the average of the squares, minus the square of the average**.

This formula shows that given any two of $\bar{x}$, s^2, and $\sum_{i=1}^{n}{x_i}^2$; we can always figure out the third one. This turns out to be useful when combining datasets.

Linear Transformations A "linear transformation" is a function of the form $y = ax + b$. Its graph is a straight line. Linear transformations arise naturally when we change units of measurement. For example, lengths in centimeters are 2.5 times lengths in inches. Temperatures in degrees Fahrenheit are (9/5) times temperature in degrees Celsius, plus 32 degrees.

So it is useful to understand how averages and SDs behave under linear transformations of the variable.

Let the dataset $\{x_1, x_2, \ldots, x_n\}$ have average $\bar{x}$ and SD s_x. For constants a and b such that $a! = 0$, let $y_i = ax_i + b$ for all i.

Definition 5 ***Average and SD of a Linear Transformation*** $\bar{y} = a\bar{x} + b$ *and* $s_y = |a|s_x$.

The steps of the proof are outlined in Exercise 2.

2.3 Questions

1. Consider a list of numbers $x = \{x_1, x_2, \ldots, x_n\}$

a) If all the entries in x are the same, then what is the variance of this list?

b) Suppose some proportion p of the numbers in the list are 1 and the remaining $1-p$ proportion of the numbers are 0. For instance, if the list had 10 numbers and $p = 0.4$, then 4 of the numbers would be 1 and the remaining 6 would be 0. Show that the standard deviation of the list is $\sqrt{p(1-p)}$.

2. Suppose we have a list $x = \{x_1, x_2, \ldots, x_n\}$ and constants a and b. Let μ be the mean of the list, and σ the standard deviation. In what follows, we will be creating new lists by using x, a, and b. The notation $y = f(x)$ means that $y_i = f(x_i)$ for each i such that $1 \le i \le n$.

a) What is the standard deviation of $y = ax$, in terms of a, σ, and μ?

b) What is the standard deviation of $y = x + b$, in terms of b, σ, and μ?

c) What is the standard deviation of $y = ax + b$, in terms of a, b, σ, and μ?

3. Suppose we have a class consisting of n students. This class has two sections, A and B. Section A has m students and section B has $n - m$ students. In the two parts below, you will find the "computational" formula for variance to be quite useful.

a) Let $n = 100$ and suppose Section A had 70 students. Section A's students have an average score of 60 with a standard deviation of 10. Section B's students have an average score of 89 with a standard deviation of 6. Find the mean and standard deviation of student scores across the entire class. You do not have to simplify the arithmetic.

b) Suppose that section A has n students and B has $n-m$ students. The average of section A is μ_A and the standard deviation is σ_A. For Section B, the average and standard deviation are μ_B and σ_B. Find the mean and standard deviation of student scores across the entire class, in terms of n, m, μ_A, μ_B, σ_A, and σ_B.

4. Let $\{x_1, x_2, \ldots, x_n\}$ be a list of numbers with mean μ and standard deviation σ. True or false (if true, prove it; if false, explain why):

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

5. A population consists of n men and n women (yes, the same number of each). The heights of the men have an average of μ_m and an SD of σ_m. The heights of the women have an average of μ_w and an SD of σ_w. Find a formula for the SD of the heights of all $2n$ people, in terms of μ_m, μ_w, σ_m, and σ_w.

6. A list $\mathbf{x}$ consists only of 0's and 1's. A proportion p of the entries have the value 1 and the remaining proportion $(1-p)$ have the value 0.

Let a and b be two constants with $b > a$. Consider the list $\mathbf{y}$ defined by $\mathbf{y} = (b-a)\mathbf{x} + a$. This means that each entry of $\mathbf{y}$ is created by first multiplying the corresponding entry of $\mathbf{x}$ by $(b-a)$ and then adding a to the result.

a) What are the values in the list $\mathbf{y}$, and what are their proportions?

b) Find the simplest formula you can for the average of the list $\mathbf{y}$ in terms of a, b, and p.

c) Find the simplest formula you can for the SD of the list $\mathbf{y}$ in terms of a, b, and p.

Chapter 3

Bounds

3.1 Markov's Inequality

As data scientists, one question that we have to be able to answer is, "If we know the average of a dataset, what information are we gaining about that dataset?" In this section, we are going to see what we can say about a dataset if all we know is its average.

Is half of a dataset above average?

For example, suppose you know that you have scored above the average on a test. Does that mean you are in the top half of scores on the test?

Not necessarily, as we can see in a simple example with just four students in a class. Suppose the scores are 10, 70, 80, and 90. Then the average is 62.5, and 75% of the list is above average.

What proportion of the data are far above average?

Now suppose you have a set of rocks whose average weight is 2 pounds. Based on this information, what can use say about the proportion of rocks that weigh 10 pounds or more?

Of course you can't say what the proportion is exactly, because you don't have enough information. But it is natural to think that the proportion can't be large, since 10 pounds is bigger than the average 2 pounds.

While it is not possible to say exactly what the proportion is, or even approximately, it turns out that it is possible to say that it can't be very large.

In fact, a famous inequality due to the Russian mathematician Andrey Markov (1856-1922) says that the proportion can be no bigger than 1/5. Here is how it works.

Markov's Bound

A bound is an upper or lower limit on how large a value can be. A lower bound is a lower limit; the value can be no less than that. An upper bound is an upper limit; the value can be no more than than that.

Markov's bound says that if the data are non-negative, then for any positive number k, the proportion of the data that are at least as large as k times the average can be no more than $1/k$.

Thus Markov provides an upper bound on the proportion. We will prove the bound later in the section. For now, assume it is true and apply it to our list of weights of rocks.

The data are weights, which are non-negative. So Markov's inequality applies. The average weight is 2 pounds, and we are looking at the proportion that are 10 pounds or more. That is, we are looking at the proportion that are at least as large as 5 times the average.

Markov's bound is that the proportion can be no bigger than 1/5.

What proportion have weights greater than 23 pounds? To use Markov's bound, note that 23 pounds is 23/2 = 11.5 times the average. Thus Markov's bound says that the proportion of rocks that weigh more than 23 pounds can be no more than 1/11.5.

Here is a detail to note. The proportion of rocks that weigh more than 23 pounds is less than the proportion that weigh 23 pounds or more, because the second set includes those that weigh exactly 23 pounds as well. Markov gives an upper bound on the proportion in the second set. So it is also an upper bound on the first.

Another detail: What does Markov say about the proportion that is bigger than half the average? Plug in $k = 1/2$ to see that Markov's bound is 2. In other words, the bound says that the proportion of data that are greater than half the average is no more than 2.

While that is correct, it is also completely useless. Any proportion is no more than 1. We don't need a calculation to tell us that it can be no more than 2.

The lesson is the Markov's bound is not useful for small k, and especially for $k < 1$. It is only interesting when you are looking at data that are quite a bit larger than average.

Markov's Inequality: Formal Statement

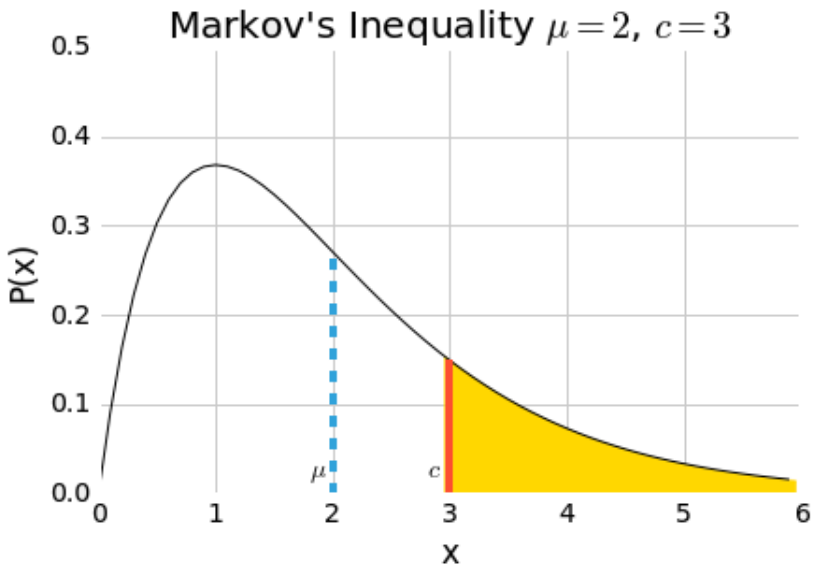
Suppose that a list of non-negative numbers $x_1, x_2, \ldots, x_n$ has average $\bar{x}$. **Markov's Inequality** gives an upper bound on the proportion of entries that are greater than some positive integer c:

For all positive values c, the proportion of entries that are at least as large as c can be no more than $\bar{x}/c$.

Definition 6 ***Markov's Inequality***
For any list of non-negative numbers with mean $\bar{x}$,

$$Proportion(x \geq c) \leq \frac{\bar{x}}{c}$$

This is what Markov's Inequality looks like graphically:



The graph shows the distribution of the data. Notice that the horizontal axis starts at 0; the data are non-negative. The shaded area is the proportion of entries that are greater than or equal to c. Markov's Inequality tells us that this area is at most $\frac{\bar{x}}{c}$.

Relation to our original statement of Markov's bound. For a list of non-negative numbers, what can you say about the proportion of entries that are at least 10 times the mean?

Our calculation using Markov's bound would say that the proportion can be no more than 1/10. To see that this also follows from the formal statement, let $\bar{x}$ denote the average of the list. We are looking for the proportion of entries greater than $10\bar{x}$.

Applying Markov's Inequality with $c = 10\bar{x}$, we get a bound of $\frac{\bar{x}}{10\bar{x}} = \frac{1}{10}$. Therefore, at most one-tenth of all entries in the list are greater than ten times the mean, which is exactly what we got by our old calculation.

Proof

To prove the statement, we will start by writing the proportion as a count divided by n:

$$\text{Proportion}(x \geq c) = \frac{\#\{i : x_i \geq c\}}{n} \tag{3.1}$$

The set $\{i : x_i \geq c\}$ consists of of all the entries that are greater than or equal to c. The # sign counts the number of items in that set, giving us the total number of entries that are at least c. That count divided by the number of total entries gives us the proportion of entries that are at least c.

Let $x_1, x_2, \ldots, x_n$ be non-negative numbers with average $\bar{x}$, and $c > 0$. We have to show that

$$\frac{\#\{i : x_i \geq c\}}{n} \leq \frac{\bar{x}}{c}$$

Ready? Here we go.

Step 1. We will start by splitting the sum of all the entries into two pieces: the sum of all the entries that are less than c, and the sum of all the entries that are at least c. Remember that the

sum of all the entries in the dataset is $n\bar{x}$.

$$\begin{aligned} n\bar{x} &= \sum_{i=1}^{n} x_i \\ &= \sum_{i:x_i<c} x_i + \sum_{i:x_i\geq c} x_i \end{aligned}$$

Step 2. In the first sum, all the entries are at least 0, since the dataset is non-negative. In the second sum, all the entries are at least c. So now our calculation becomes:

$$\begin{aligned} n\bar{x} &= \sum_{i=1}^{n} x_i \\ &= \sum_{i:x_i<c} x_i + \sum_{i:x_i\geq c} x_i \\ &\geq \sum_{i:x_i<c} 0 + \sum_{i:x_i\geq c} c \end{aligned}$$

Step 3. Almost done! The first sum above is 0. The second sum is just the constant c multiplied by the number of terms in the sum. The number of terms is the number of indices i for which $x_i \geq c$. In other words, the number of terms is the number of data points that are at least c.

$$\begin{aligned} n\bar{x} &= \sum_{i=1}^{n} x_i \\ &= \sum_{i:x_i<c} x_i + \sum_{i:x_i\geq c} x_i \\ &\geq \sum_{i:x_i<c} 0 + \sum_{i:x_i\geq c} c \\ &= \sum_{i:x_i\geq c} c \\ &= \#\{i : x_i \geq c\} * c \end{aligned}$$

Step 4. Finally, divide both sides by n and then by c. You're done!

$$\frac{\bar{x}}{c} \geq \frac{\#\{i : x_i \geq c\}}{n}$$

This is the same as what we are trying to prove:

$$\frac{\#\{i : x_i \geq c\}}{n} \leq \frac{\bar{x}}{c} \tag{3.2}$$

(3.3)

3.2 Chebychev's Inequality

Markov's inequality gave us a way to bound the tail non-negative distribution, using only the mean. "Tails" of lists are sets of entries that start far away from the center and go out even further.

The standard deviation of a list measures spread around the mean. Could we tighten our bound on the tail any more if we also knew the SD of the list?

The Weatherman

Consider a weatherman in Northern Alaska interested in examining temperatures, where temperatures are cold and stay that way. Suppose that after some investigation, we find that the average temperature $\bar{x}$ is -25 C, and that the SD of the temperatures s is 5 C. Northern Alaskans prefer temperatures between -15 C and -35 C, and we'd like to figure out a way to measure the proportion of days which lie in this zone.

Intuitively, it makes sense that we're less likely to see temperatures further away from the mean (as the mean is a measure of centrality). Furthermore, we would expect that the smaller the standard deviation is, the less likely we are to see temperatures that are far away, since a small standard deviation indicates closeness to the mean.

As we saw with Markov's bound, there's no way without looking at the numbers to calculate the exact proportion, but we can bound the proportion of days with temperatures between -15C and -35 C.

For inspiration, we look to Markov's mentor, Putnafy Chebychev [1], whose theorem claims that the proportion of days which **do not** fall between -15 and -35 is at most 1/4. Equivalently, at least 3/4 fall within the range. Here is how this works.

Chebychev's Bound

Chebychev's inequality states that the proportion of entries which are at least k standard deviations away from the mean is at most $\frac{1}{k^2}$. Here k is any positive number, and need not be an integer.

In our weather example, we were looking for items outside of -15 and -35. Both of these are 2 standard deviations away from our mean, -25 ($\frac{|-15-(-25)|}{5} = 2$ and $\frac{|-35-(-25)|}{5} = 2$). Thus, the proportion of temperatures which are not in the range $(-35, -15)$ is at most $\frac{1}{2^2} = \frac{1}{4}$.

It is important to note that Chebychev's inequality works for **all datasets**, not just non-negative datasets like Markov's inequality.

Also note that just as we observed with Markov's inequality, small values of k don't lead to interesting bounds. For example, Chebychev's inequality says that the proportion of entries that are at least half an SD away from the mean is at most $1/(1/2)^2 = 4$. Since we already know that the proportion is at most 1, the inequality isn't telling us anything. Chebychev's bound is interesting for tails, that is, entries that are far away from the mean. That is, Chebychev's bound is interesting when k is large.

[1] Chebychev is a transcription from Russian: you may see it as Chebyshev, Chebysheff, Chebyshov, Tchebychev, Tchebycheff,Tschebyschev, Tschebyschef, or Tschebyscheff

Definition 7 ***Chebychev's Inequality*** *Suppose the list $x_1, x_2, \ldots, x_n$ has average $\bar{x}$ and SD s. Let k be any positive number. Then the proportion of entries that are at least k SDs away from the mean is at most $1/k^2$.*

$$Proportion\{i : |x_i - \bar{x}| \geq ks\} \leq \frac{1}{k^2}$$

That is,

$$Proportion\{i : x_i \text{ is outside } \bar{x} \pm ks\} \leq \frac{1}{k^2}$$

We will prove the bound after making a few observations about its use.

First, note that in order to find Chebychev's bound, you need both the mean and the SD, whereas to use Markov's bound you just need the mean of a non-negative list. When both inequalities apply, Chebychev often provides tighter bounds than Markov. But Chebychev's bound requires more information than Markov's.

Also note that the condition

$$|x_i - \bar{x}| \geq ks$$

is equivalent to

$$\left|\frac{x_i - \bar{x}}{s}\right| \geq k$$

The quantity

$$\frac{x_i - \bar{x}}{s}$$

is often denoted z_i, and measures "how many SDs above average" the value x_i is. If z_i is negative, then x_i is a negative number of SDs above average; that means it is below average. If z_i is 0 then x_i is exactly at the average.

The number z_i is called x_i *in standard units*, or the *z-score* of x_i.

Proof of Chebychev's Bound:

The only bound we know so far is Markov's. It says that for a list of non-negative numbers,

$$\text{Proportion}(i : x_i \geq c) \leq \frac{\bar{x}}{c}$$

How can we use this to establish Chebychev's bound for all lists? Let's begin by rewriting the proportion in Chebychev's bound:

$$\text{Proportion(i: } x_i \text{ is outside } \bar{x} \pm ks)$$

$$= \text{Proportion(i: } |x_i - \bar{x}| \geq ks \text{)}$$

$$= \text{Proportion(i: } (x_i - \bar{x})^2 \geq k^2 s^2)$$

This works because $|x_i - \bar{x}|$ is a non-negative number, and therefore, by squaring both sides, $|x_i - \bar{x}| \geq ks$ is equivalent to $(x_i - \bar{x})^2 \geq k^2 s^2$.

The list $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \ldots, (x_n - \bar{x})^2$ is the list of squared deviations from mean, so its average is the variance s^2 (look up the definition of variance and you'll see that this is true).

Also, the list of squared deviations is non-negative, so Markov's inequality applies to it.

By applying Markov's inequality to the list of squared deviations, we get

$$= \text{Proportion(i: } (x_i - \bar{x})^2 \geq k^2 s^2) \leq \frac{s^2}{k^2 s^2} = \frac{1}{k^2}$$

That's Chebychev's bound.

The importance of Chebychev's bound is that it applies to all datasets. Thus for example we can say that no matter what the list looks like, the proportion of entries that are at least 3 SDs away from the mean is at most 1/9. The proportion that are at least 4 SDs away from the mean is at most 1/16, and so on.

In other words, *no matter what the list*, the bulk of the entries lie in the range "average plus or minus a few SDs."

That is the power of Chebychev.

3.3 Questions

1. Suppose a list of numbers $x = \{x_1, \ldots, x_n\}$ has mean μ_x and standard deviation σ_x. We say that a number y is within z standard deviations of the mean if $\mu_x - z\sigma_x < y < \mu_x + z\sigma_x$.

   a) Let c be smallest number of standard deviations away from μ_x we must go to ensure the range $(\mu_x - c\sigma_x, \mu_x + c\sigma_x)$ contains at least 50% of the data in x. What is c?

   b) Suppose that a BART ride from Berkeley to San Francisco takes a mean time of 38 minutes with a standard deviation of 4 minutes. If you want to make the claim "At least 90% of BART rides from Berkeley to San Francisco take between ___ and ___ minutes", what numbers should be used to fill in the blanks?

2. At an elementary school, 45 children are raising money for charity. The teacher has 20 candy bars, and has promised to give one candy bar to each child who raises $5 or more. The average amount raised by the children is $2. Does the teacher have enough candy bars to keep her promise? Why or why not?

3. A list of incomes has mean $75,000 and SD $25,000. Give the best upper bound you can for the proportion of incomes that are more than $150,000.

4. A list of incomes has an average of $\$60,000$ and an SD of $\$40,000$. Let p be the proportion of incomes that are over $\$200,000$.

   a) What, if anything, does Markov's inequality say about p?

   b) What, if anything, does Chebychev's inequality say about p?

   c) Is either of the answers to parts (a) and (b) more informative about p than the other? Explain your answer.

5. A list of test scores has an average of 55 and and SD of 10. What can you say about the proportion of scores that are in the interval $(30, 80)$?

6. A list of test scores has an average of 55 and an SD of 10. What can you say about the proportion of scores in the interval $(25, 95)$?

7. A class of 58 students takes a true-false quiz consisting of 20 questions. Each answer will get a score of 1 if it is correct and -1 otherwise; no other score is possible.

   The GSIs keep track of the number of answers each student gets correct. The average of these 58 numbers is 16.1 and the SD is 2.3.

In each of the following parts, find the quantity if it is possible to do so with the information given. If it is not possible, explain why not.

a) the average number of answers that were anything other than correc
b) the SD of the number of answers that were anything other than correct
c) the average score on the test
d) the SD of scores on the test

Chapter 4

Probability

Probability theory is a discipline rooted deeply in the real world and in mathematics. We use probabilities and statistics to represent integral parts of our lives, as diverse as the chance of rain on the weather app, batting averages for our local baseball teams, or the success rate of a medical treatment.

Through the language of statistics, we can concisely describe a situation and make predictions about what's to come. By building on the basic structures of probability laid out in this chapter, we will be able to understand how these probabilities combine. Understanding probability will make us better equipped to calculate likelihoods and make decisions without taking unnecessary risks.

A note to the reader: the examples in this text have been designed with the intention that readers follow along by doing the calculations. Please don't just read the text like a novel. Thanks!

4.1 Probability

We can use probability to measure the likelihood of an event occurring.

Suppose an experiment can result in exactly one of several possible outcomes. In data science, we will almost invariably be looking at a finite set of possible outcomes. In what follows, you can just assume that the set of all possible outcomes is finite.

One way to define the probability of an event is as a proportion of the number of favorable outcomes relative to the total number of outcomes. This definition makes sense only under the assumption that all outcomes are equally likely.

For example, if you are rolling a six-sided die and believe that all sides have the same chance of being rolled, then the set of all possible outcomes is $\{1, 2, 3, 4, 5, 6\}$ and the chance of the event "the number of spots is a multiple of 3" is

$$\frac{\#\{3, 6\}}{\#\{1, 2, 3, 4, 5, 6\}} = \frac{2}{6} = \frac{1}{3}$$

(The word *favorable* in this context refers to the event you are studying, and is not necessarily a "good" event.)

Probability is always between 0 (corresponding to an impossible event) and 1 (corresponding to a certain event).

In probability theory it is standard to denote events by the "early" letters of the alphabet, such as A, B, and so on.

Definition 8 ***Probability of A, assuming equally likely outcomes***

$$P(A) = \frac{\textit{Number of outcomes favorable to } A}{\textit{Total number of outcomes}}$$

Example 1. You are drawing an item out of a box. In the box there is one green tennis ball, one orange soccer ball, and two white golf balls. The box contains nothing else. You are equally likely to pick any of the balls. What's the probability that:

1. You pick an orange ball?
2. You pick a golf ball?
3. You pick a ball?
4. You pick a golf ball that is red?

Solution
Probability = $\frac{\text{Favorable Outcomes}}{\text{Total Outcomes}}$.

1. There is one orange ball in the box. So there is only one favorable outcome out of the four total possible outcomes. Therefore, the probability of picking an orange ball = $\frac{1}{4}$
2. Now, there are two favorable outcomes as there are two golf balls. Thus, the probability of picking a golf ball is $\frac{2}{4} = \frac{1}{2}$
3. We know that there are only balls in the box. Therefore, picking a ball is a certain event. So, the probability of picking a ball = 1.
4. There is no golf ball in the box that is red. Therefore, picking a red golf ball is an impossible event. Thus, the probability = 0.

Partitioning events

When computing probabilities, it is natural to break events up into simpler events and then combine the probabilities of the simpler events.

Example 2. suppose the distribution of age (measured in completed years) in a population is as follows:

Age	20-34	35-49	50-64	65-79	80-100
% of People	20	20	30	20	10

Suppose one person is picked at random. What is the chance that the person is a senior citizen (age 65 or older)?

Note on terminology: In this text, the term "at random" will mean "all outcomes are equally likely". In our example, an "outcome" is a person. We're assuming that all the people are equally likely to be picked.

Under this assumption it's quite natural to say that the answer is 30%, as that's the percent of senior citizens in the population from which the draw is made. That's correct, and we will now break down the argument into finer detail.

The event "the person picked is a senior citizen" partitions into two simpler events: the person's age is either in the range 65-79 or 80-100. In a partition, only one of the events can occur. When

age is measured in completed years, a person can't be in both age groups 65-70 and 80-100. We say that the two events in a partition are "mutually exclusive". Each excludes the other.

Now "senior citizen" partitions into "age 65-79 or 80-100". The chance of picking a senior citizen is the sum of the chances of the two groups: 20% + 10% = 30%.

Definition 9 ***Addition Rule.***

$$P(A \text{ or } B) = P(A) + P(B) \text{ if } A \text{ and } B \text{ are mutually exclusive}$$

Events that Satisfy Multiple Conditions

Example 3. Suppose you draw two times at random without replacement from a box that contains one ticket each of the colors Red, Blue, and Green. What is the chance that you get the Blue ticket first, and then the Red? **Solution.**

"At random without replacement" means that all tickets are equally likely to be drawn, and once you have drawn a ticket, you don't replace it in the box before you draw the next one.

Under these assumptions the possible pairs you can draw are RB, RG, BG, BR, GB, and GR. You can't get the same color twice.

The outcome we want is BR. So the chance is 1/6.

Easy enough. But one again, the answer merits further examination.

The chance of getting the Blue ticket on the first draw is 1/3. So if you imagine running this experiment over and over again, the Blue ticket will appear on the first draw about 1/3 of the time. **Among those times,** the Red ticket will appear on the next draw about 1/2 the time. So the chance of BR can be thought of as

$$\frac{1}{2} \text{ of } \frac{1}{3} = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$$

Thus the probability that two events both occur (that is, B on the first draw and R on the second) is a **fraction of a fraction**. The more conditions you place on an event, the smaller its chance becomes.

Definition 10 ***Multiplication Rule.***

$$P(A \text{ and } B) = P(A) \cdot P(B \text{ given that } A \text{ has happened})$$

Example 4. What's the probability that you get a head followed by a tail when you flip two coins? You can assume the coins are fair.

Solution

The chance of getting a head on the first toss is 1/2. Since the outcome of the first toss doesn't affect outcomes for the second (a natural assumption, that turns out to be fine in practice), the chance that the second toss is a tail is 1/2 no matter how the first toss came out. So the answer is

$$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

You can also solve this problem by enumerating all the outcomes:

$$\frac{\#\{HT\}}{\#\{HH, HT, TH, TT\}} = \frac{1}{4}$$

Example 5. What's the probability that you get a head and a tail when you flip two coins?
Solution

Notice the difference between this example and Example 3. In this one, the order in which the two faces appear isn't specified. So the event includes them appearing in any order.

So the event "a head and a tail" partitions into "HT or TH", which is a partition because the two coins can't show HT as well as TH on the same pair of tosses. So

$$P(\text{a head and a tail}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

4.2 Examples: Sampling with Replacement

Suppose you have a finite population from which you sample repeatedly. We will define *random* sampling to mean that at each stage, every element has the same chance of being selected. Formally, this is sometimes called *sampling uniformly at random*.

When you sample repeatedly, you have to specify whether or not the tickets that you have already drawn out continue to be part of the population. If they do, you are *sampling with replacement*.

A common way to visualize this is to imagine each member of the population being represented by one ticket in a box. When sampling with replacement, you shuffle all the tickets and draw one, then replace in the box and repeat the process.

One example where this is a good model is rolling a fair 6-sided die. A natural assumption is that if the die is rolled once and the outcome is a 5, the next roll can still yield any of the numbers 1 through 6 with equal probability. Hence rolling a die is like sampling at random with replacement from $\{1, 2, 3, 4, 5, 6\}$.

Example 1. Rolling A Die. A fair 6-sided die has numbers from 1 to 6. Each time it is rolled, the outcome will be a number from 1 to 6. The probability of getting any of the six numbers is the same, which is 1/6. No roll affects the outcome of any other roll.

(i) Suppose the die is rolled once. What is the probability of rolling a 1 and a 2?

(ii) If the die is rolled once, what is the probability of rolling a 1 or a 2?

(iii) If the die is rolled twice, what is the probability of rolling a 1 on the first roll and a 2 on the second roll?

Solution

(i) The chance of getting both 1 and 2 on the same roll is 0 since the outcome could only be one of the two numbers.

(ii) The chance of getting either 1 or 2 on the same roll is

$$\frac{\#\{2, 6\}}{\#\{1, 2, 3, 4, 5, 6\}} = \frac{2}{6}$$

Another way to solve this is to note that "the roll shows 1" and "the roll shows 2" are mutually exclusive, so by the addition rule, the chance that the roll shows 1 or 2 is

$$\frac{1}{6} + \frac{1}{6} = \frac{2}{6}$$

(iii) By the multiplication rule, the answer is the chance of getting a 1 on the first roll times the chance of getting a 2 on the second roll given that 1 appeared on the first roll. Since no roll affects any other, both chances are 1/6. So the answer is 1/36.

Example 2. A die is rolled 3 times. What is the probability that the face 1 never appears in any of the rolls?
Solution Let's break the question into simpler problems. What is the chance that 1 does not appear in a single roll?
The possible faces that can appear in a single roll, excluding 1, are $2, 3, 4, 5,$ and 6.
Therefore, the probability of not getting 1 in a single roll of die $= \frac{5}{6}$
Since we are rolling a die, the chance of not getting 1 is the same on each subsequent roll.
Since we want "not 1" to occur on each of the three rolls, the answer will be "a fraction of a fraction of a fraction" by the multiplication rule:
The probability that 1 does not appear in any of 3 rolls $= \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} = (\frac{5}{6})^3$

$1^{st} roll \quad 2^{nd} roll \quad 3^{rd} roll$

Example 3. A die is rolled n times. What is the chance that only faces $2, 4$ or 6 appear?

Solution The chance that either $2, 4,$ or 6 appears in a single roll $= \frac{3}{6}$

Since we are rolling a die, the chance that either $2, 4,$ or 6 appears in a single roll is the same in subsequent rolls.
Therefore, chance that only $2, 4,$ or 6 appear in n rolls $= (\frac{3}{6})^n = (\frac{1}{2})^n$

Example 4. A die is rolled two times. What is the probability that the two rolls had different faces?

Solution. To understand the problem, we can think in the following way:
The first roll can be any of $1, 2, 3, 4, 5$ or 6. Hence, we will accept any face for the first roll since all faces are favorable. In the second roll, the face should be anything but first roll and thus, it can be any of five different faces.

Probability of getting any of the six faces in the first roll $= \frac{6}{6} = 1$
On the second roll: Probability of getting any face but the face that appeared on the first roll $= \frac{5}{6}$
Probability that the two rolls had different faces $= \frac{6}{6} \times \frac{5}{6} = \frac{5}{6}$

Example 5. There are 20 students in a class. A computer program selects a random sample of students by drawing 5 students at random with replacement. What is the chance that a particular student is among the 5 selected students?

Solution. Since it is difficult to enumerate every possible case that includes a particular student, we look at its complement and see if it is simpler to work with.

Because we are sampling with replacement, the probability that the student is selected on any particular draw is not affected by what happened on other draws. So:
The probability that a particular student is not selected in a single draw $= (\frac{20-1}{20}) = \frac{19}{20}$
The probability that a particular student is not selected in all five draws (which is the entire sample) $= (\frac{19}{20})^5$

The probability of a particular student getting selected in the sample = 1 - Probability that the student is not selected in the sample $= 1 - (\frac{19}{20})^5$

Generalization:
Total number of students = N
Sample size = n

Probability that a particular student is not selected = $(\frac{N-1}{N})^n = (1 - \frac{1}{N})^n$

Probability of a particular student getting selected = 1 - probability that a particular student is not selected $= 1 - (1 - \frac{1}{N})^n$

4.3 The Gambler's Rule

So far, we've only applied probabilities to small games, finding the chances of events occurring in dice and coin games with a small number of events. Now, we'll combine all the ideas presented to examine the mechanics of a real world gambling scenario.

The Game. Say you are playing a game where N people put in a bet, and one person is chosen at random to win the whole pot. What is the chance that you will win if you play once? What is the chance that you will win at least once, if you play n times?

Using the concepts we have learned from probability with replacement, we can find a good strategy about how we can approach this game.

Placing Bets

We need to state our assumptions. For what follows, N and n are integers greater than 1. The main assumption is that when you are playing this gambling game n times, you have a chance of winning $\frac{1}{N}$ each time you play, regardless of the outcomes of all other times. These are the only assumptions needed.

If you play once,

$$P(\text{you win one bet}) = \frac{1}{N}$$

From this, we can already conclude that the chance you will lose one bet is $1 - \frac{1}{N}$ because the probability of your losing is the chance that you are not able to win.

$$P(\text{lose one bet}) = 1 - \frac{1}{N}$$

Knowing the probability we can lose one bet brings us to a scary question: What is the chance that you will lose n times straight?

In such situations it's always a good idea to start out with a small fixed value of n, and then see if you can generalize. If $n = 2$, we are trying to find the chance that you lose 2 bets. Two

conditions have to be satisfied: you have to lose the first bet, then you have to lose the second as well. Remember that bets remain unaffected by the results of other bets. So by the multiplication rule, the chance is

$$P(\text{lose 2 bets}) \;=\; (1-\frac{1}{N})*(1-\frac{1}{N})$$

Now, finding the chance that we lose n times straight is simple. Just repeat the reasoning above. Because of our assumptions, we can conclude that:

$$P(\text{lose all } n \text{ bets}) \;=\; (1-\frac{1}{N})^n$$

We now have the chance of losing all n bets. But the chance that we had originally set out to find was the chance of winning at least one bet out of the n bets. How do we go about finding that?

At this point, many students will be quite dumbfounded and try to use some other fancy probabilistic method involving combinations or what not, but the answer to this problem is simple. The complement of losing all of the bets is winning at least one bet. That's all that's needed!

$$P(\text{win at least one bet}) \;=\; 1-(1-\frac{1}{N})^n$$

How to Get a Fair Chance?

When you flip a coin, you get a 50% chance of landing heads and a 50% chance of landing tails. We say that this is a fair chance as there is no difference in chance between landing either outcome. How many bets do you think it will take to give you a fair chance of winning at least one out of the n bets?

Come up with a guess and save it for the end, when we have solved the problem. You'll be able to see how your intuition matches up with the answer!

We have to solve for the smallest n for which the chance that you win at least one of the n times is at least $1/2$. Remember that N is fixed. Mathematically:

$$1-(1-\frac{1}{N})^n \;\geq\; \frac{1}{2}$$

This is the same as

$$\frac{1}{2} \;\geq\; (1-\frac{1}{N})^n$$

In order to isolate n, let's take the logarithm of both sides. (Note that "logarithm" means "natural logarithm" here; we won't be taking logs to the base 10 at this level of math.) Since the logarithm is a strictly increasing function, it preserves the inequality.

$$\log\frac{1}{2} \;\geq\; n\log(1-\frac{1}{N})$$

Now, to isolate n, we have to divide both sides by $\log(1-\frac{1}{N})$. Remember that we must flip the inequality because $\log(1-\frac{1}{N})$ is negative!

$$\frac{\log\frac{1}{2}}{\log(1-\frac{1}{N})} \;\leq\; n$$

That gives us our bound:

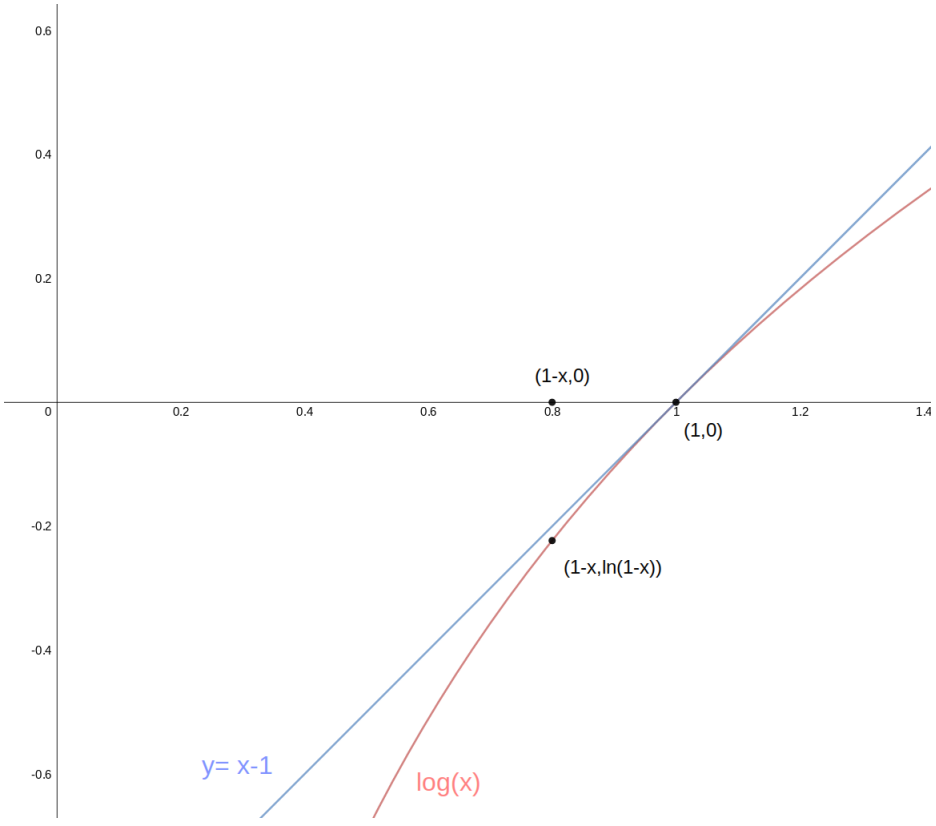
$$n \geq \frac{\log \frac{1}{2}}{\log(1 - \frac{1}{N})}$$

We've now come up with a solution to our original problem, although it doesn't really give us a good understanding of how large this value is. So, let's try to approximate it.

Approximating n

To find an approximation to the smallest n that satisfies the condition above, we'll start by examining the log function. Let's try to approximate the value of $\log(1-x)$ for small, positive x. We know that the log of numbers close to 1 is close to zero (recall that $\log(1) = 0$ and log is a continuous function).

Let us draw a graph of the function $f(s) = \log(s)$ along with its tangent line at $s = 1$. Now, for a small positive x, plot and label the three following points on this graph:
A: $((1-x), 0)$
B: $((1-x), \log(1-x)$
C: $(1, 0)$



Do you notice something about these points? They produce a triangle; not just any triangle, but approximately a 45-45-90 right triangle. That's because the derivative of the log function at $s = 0$ is 1, so the tangent line is a 45 degree line.

The two legs of the right triangle are equal, and one of them is clearly equal to $-x$. That's the distance between A and C. Therefore, the other leg is also approximately $-x$, and we already know that it's $\log(1-x)$. So

$$\log(1-x) \approx -x \quad \text{for small positive } x$$

This approximation gets used over and over again in probability theory, so it's a good idea to understand it well. As an exercise, draw the diagram that shows that

$$\log(1+x) \approx x \quad \text{for small positive } x$$

Let's plug our approximation into our inequality:

$$n \geq \frac{\log \frac{1}{2}}{\log(1-\frac{1}{N})}$$

Since we are assuming that N is large, we can likewise say that $\frac{1}{N}$ is very small, so we can use our approximation to make a substitution:

$$n \gtrsim \frac{\log \frac{1}{2}}{-\frac{1}{N}} = -N \log(\frac{1}{2}) = N \log(2)$$

Remember rules of logarithms: $\log(1/2) = -\log(2)$.

Now $\log(2)$ is approximately equal to $2/3$, so we can say:

$$n \gtrsim \frac{2}{3}N$$

Gamblers have known for centuries that the answer to the question we posed is about $2/3$ of N, and have used that as a rule of thumb. That is why the result is called the Gambler's Rule.

Plug large numbers into this bound for n. You will soon realize that n needs to be absurdly large number for you to get a fair chance of winning at least one bet. If N is a million, you need to bet at least $2/3$ of a million times as a matter of fact (and that only gives you about an even chance of winning at least once)! Was this close to your guess?

4.4 The Birthday Problem

Parties are great social events, and while mingling in the crowd, you might learn that you share the same favorite color, same car, or perhaps even the same birthday with another person. It may seem strange that, in a room of only perhaps 30 people, it is more likely than not that someone shares a birthday with someone else – after all there are 365 possible birthdays – but a calculation will actually tell us that it's not strange at all.

This situation is the premise of the birthday problem: What is the minimum number of people that need to be in a room so there is about a 50% chance that at least two of them share the same birthday?

Some common assumptions that we'll use to make our calculations simpler:

1. There are 365 days in every year (we're ignoring leap years).

2. There is no 'clumping'. That is, each person's birthday is equally likely to be on any of the 365 days regardless of others' birthday. For instance, there are no twins in the room.

3. Nobody's birthday affects the chance of anybody else being born on any particular day.

Calculating the Probability

Suppose that there are n people. Let A be the event that at least two of them share a birthday.

Then P(A) is the probability that at least two people in the room have the same birthday. This is a complicated event, because any two people could share any birthday, or there could be three common birthdays, or ... the event could happen in myriad ways.

Fortuantely, the complement is easier. Let $P(A^c)$ be the probability of the complement. Then $P(A^c)$ is the chance that all n people have different birthdays.

When there are two people in the room (that is, $n = 2$), the probability that the two have different birthdays is

$$\frac{365}{365} \cdot \frac{364}{365} = \frac{364}{365}$$

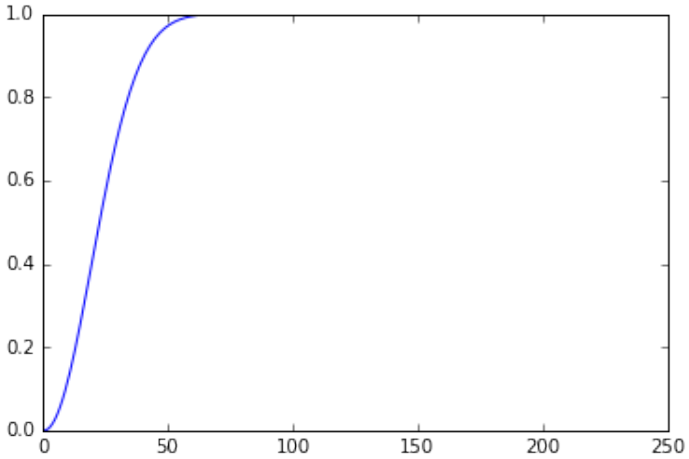
When there are three people, each of the three birthdays has to be unique, and so the chance that all three have different birthdays is

$$\frac{364}{365} \times \frac{363}{365}$$

Let's extend the logic to n people, with a table:

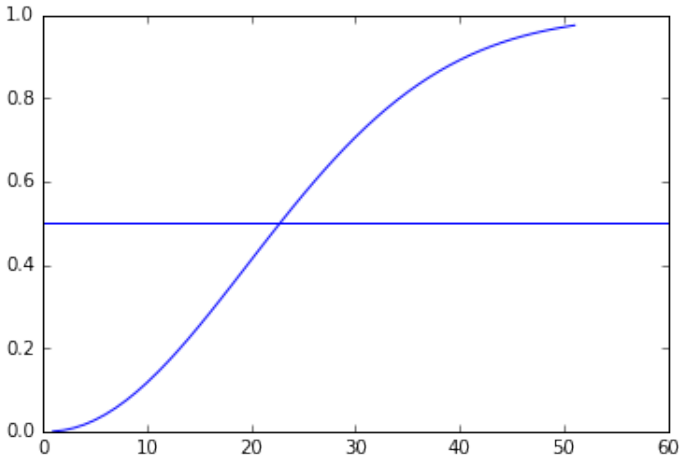
Birthday Problem - Probability Table		
Class Size (n)	Chance that all birthdays are different	Chance that at least 2 or more people in the class have the same birthday
1	0	1
2	$\frac{365}{365} \times \frac{364}{365}$	1 - ($\frac{365}{365} \times \frac{364}{365}$)
3	$\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365}$	1 - ($\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365}$)
4	$\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \frac{362}{365}$	1 - ($\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \frac{362}{365}$)
:	:	:
:	:	:
$n \geq 3$	$\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times ... \times \frac{(365-(n-1))}{365}$	1 - ($\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times ... \times \frac{(365-(n-1))}{365}$)

Now that we've found a formula for the probability, let's graph it. The horizontal axis shows n and the vertical axis shows the chance that at least two of the n people have the same birthday.



Wow! The probability spikes up very quickly, and when n is greater than 100 people, the probability is near 1.

Our original question was to find the point at which there was a 50% probablity that two people have the same birthday, so let's zoom in, and find where $P(A) = 50\%$.



If you look closely, you can notice that our graph hits halfway when $n = 23$. This is somewhat counterintuitive, but this interesting statistical example only goes to show how powerful probabilities can be when we combine them on a large scale.

Approximating the Probability

The log approximation we used in the Gambler's Rule also helps us approximate the birthday probability. The chance that at least two out of n people have the same birthday is

$$P(A) = 1 - (\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times ... \times \frac{(365-(n-1))}{365})$$

To approximate this, we have to approximate

$$P(A^c) = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times ... \times \frac{(365-(n-1))}{365}$$

and then subtract from 1.

Now $P(A^c)$ is a product, but we are much better at dealing with sums. So let's use log convert the product to a sum:

$$\log(P(A^c)) = \sum_{i=0}^{n-1} \log(\frac{365-i}{365}) = \sum_{i=1}^{n-1} \log(\frac{365-i}{365})$$

because $\log(1) = 0$. So

$$\log(P(A^c)) = \sum_{i=1}^{n-1} \log(\frac{365-i}{365}) = \sum_{i=1}^{n-1} \log(1 - \frac{i}{365})$$

Now use the approximation $\log(1-x) \approx -x$ for small positive x:

$$\log(P(A^c)) \approx \sum_{i=1}^{n-1} -\frac{i}{365} = -\frac{1}{365}\sum_{i=1}^{n-1} i = -\frac{1}{365}\cdot\frac{(n-1)n}{2}$$

using the result $1+2+\cdots+k = k(k+1)/2$ for every positive integer k.

Thus

$$P(A^c) \approx e^{-\frac{1}{730}(n-1)n}$$

and so

$$P(A) \approx 1 - e^{-\frac{1}{730}(n-1)n}$$

This exponential approximation to the probability in the birthday problem shows why the graph above rises so sharply. The probability of the complement is dropping very fast, on the order of e^{-n^2}.

Conclusion

As you can see, probability is not all about math and calculations. Knowing what that probability really means and being able to apply that knowledge in real life situations can keep you from ending up in high-risk, low-reward situations, such as in gambling.

Maybe you can try out the birthday problem at your next large family gathering or come up with a new magic trick using your new found knowledge of probability. Keep probability in mind when there is any uncertainty surrounding an outcome and maybe you can impress your family and friends when you make bold, but confident, predictions and they turn out to be true.

4.5 Questions

1. A die is rolled 8 times. What is the chance that the same face appears on all 8 rolls?

2. A roulette wheel has 38 pockets, 2 of which are green, 18 black, and 18 red. The wheel is spun 10 times.

   a) What is the chance that all of the winning pockets are red?

   b) What is the chance that at least one of the winning pockets is green?

3. The English alphabet consists of 26 letters. From this alphabet, 4 letters will be drawn at random with replacement.

   a) How many possible sequences of 4 letters can appear? Note that the sequence keeps track of the order in which the letters appear. For example, ABCD is different from BACD; AAAB is different from ABAA; etc.

   b) What is the chance that the first three letters all different and the fourth one is the same as one of the previous three that appeared?

4. There are 2,598,960 different poker hands. Suppose I play poker two times so that all hands are equally likely each time regardless of what appeared the other time. The chance that I get the same hand both times is equal to (pick one option and explain):

   (i) $\frac{1}{2{,}598{,}960} \times \frac{1}{2{,}598{,}960}$. (ii) $\frac{1}{2{,}598{,}960}$. (iii) neither (i) nor (ii).

5. A coin is tossed n times.

   a) What is the total number of ways the n tosses could come out?

   b) What is the number of ways the n tosses could come out so that there are exactly k heads among them? Here k is an integer in the range $0 \le k \le n$. Check that your answer makes sense for the boundary cases $k = 0$ and $k = n$.

   c) What is the chance that there are exactly k heads among the n tosses?

6. A monkey hits the keys of a typewriter at random, picking each of the 26 letters of the English alphabet each time regardless of what it has picked on all the other times.

   a) What is the chance that the first six letters form the word ORANGE, in that order?

   b) What is the chance that the first six letters form the word ORANGE by rearrangement if necessary?

7. There are 59 students enrolled in a class. Find the chance that at least one other student has the same birthday as the instructor's, assuming that the instructor's birthday is a fixed day and every student's birthday is equally likely to be any of 365 days of the year regardless of everyone else's birthday.

8. A population consists of 1000 people. A sample of n people is drawn at random with replacement from the population. For $n \le 1000$, write a formula for the chance that the sample consists of n different people.

9. Derive an exponential approximation to the answer in the problem above. You should recognize this as an approximation you've done before using different numbers.

10. A sample consists of n people, one of whom is called Special. A bootstrap sample (n draws at random with replacement) is drawn from the sample. Find the probability that Special is **not** in the sample, and show that if n is large, the chance is about $1/e$.

Chapter 5

Sampling

In the previous chapter, we learned about the basics of probability, as well as the various rules that determine how probabilities can be combined. The main examples were about sampling *with* replacement, which we will recap briefly here. Then we will move on to a different method of sampling that arises for example when shuffling cards or selecting people to answer a survey.

5.1 Sampling With Replacement

In what follows, *population* is a list consisting of N distinct individuals indexed by $1, 2, \ldots, N$.

A *random sample with replacement* from the population is defined by the following sampling scheme:

- Draw one element uniformly at random from the list. Replace the item into the list.
- Repeat.

Example 1. If you draw a sample of size n at random with replacement from the population, how many different sequences of individuals could you draw?

Solution. If $n = 1$, then you are drawing just one individual, which gives you a rather boring sequence of length 1. The number of different such sequences is N, one for each of the different individuals.

If $n = 2$, you have to start by picking one individual at random, which you can do in N different ways as noted above. Then, for each of these N choices, the next choice is also from N individuals because the first one is replaced in the list. So the total number of sequences of length 2 is $N \times N = N^2$.

By induction on n, the total number of possible sequences of length n is N^n.

In what follows, assume that a random sample of size n is drawn with replacement from a list of size N.

Example 2. You are one of the N individuals on the list. What is the chance that you are chosen in the sample?

Solution. There are many ways for you to enter the sample: early in the sequence of late, once or more than once. The complement is a much simpler event: you don't enter the sample at all.

For this to happen, all n people must be chosen from among the $N-1$ other people. The chance of that is

$$\frac{\#\{\text{sequences of length } n \text{ from } N-1 \text{ people}\}}{\#\{\text{sequences of length } n \text{ from } N \text{ people}\}} = \frac{(N-1)^n}{N^n}$$
$$= (1-\frac{1}{N})^n$$

By the complement rule, the chance that you *are* chosen is

$$1-(1-\frac{1}{N})^n$$

Notice that this is exactly the same as the probability in the Gambler's Rule.

Example 3. Your best friend is also an individual on the list. What is the chance that your best friend is chosen in the sample? Is it different from the answer to Example 2?

Solution. The answer is the same as the answer to Example 2 because the identity of the individual in question does not matter. It doesn't matter if a question asks about you, your best friend, or Batman in all his caped glory; the random sample does not care. As long as they are all unique individuals within the population, the probability of any given individual being chosen is the same.

Example 4. Assume that N is very large in comparison to n. Give an exponential approximation of the chance in Example 2.

Solution. The answer in Example 2 is

$$p = 1-(1-\frac{1}{N})^n$$

Therefore, what we must approximate is

$$1-p = (1-\frac{1}{N})^n$$

To do this, first take the log on both sides :

$$\log(1-p) = n \cdot \log(1-\frac{1}{N})$$

Now recall the exponential approximation from the last chapter:

$$\log(1-\frac{1}{N}) \approx -\frac{1}{N}$$

because N is large. So

$$\log(1-p) \approx n \cdot \frac{-1}{N} = \frac{-n}{N}$$

We can escape the log by raising e to the power of both sides.

$$1-p \approx e^{-\frac{n}{N}}$$

and so

$$p \approx 1-e^{-\frac{n}{N}}$$

5.2 Sampling Without Replacement

Recall that a *population* is a list consisting of N distinct individuals indexed by $1, 2, \ldots, N$. A *random sample without replacement* from the population is defined by the following sampling scheme:

- Draw one individual uniformly at random from the list. Cross that person's name off the list.
- Draw one individual uniformly at random from the reduced list. Cross that person's name off the list. Repeat.

NOTE: A random sample without replacement is known as a *Simple Random Sample*.

In what follows, assume that a simple random sample of size n is drawn from a population of size N.

Example 1. What is the number of different sequences that you can draw?

Solution. The first individual can be drawn in N ways. For each of these ways, the second individual can be drawn in $N-1$ ways, because the first individual drawn is not replaced. So there the number of distinct sequences of length 2 is $N(N-1)$. Continuing this argument by induction, the number of sequences of length n is

$$\begin{aligned} & N(N-1)(N-2)\cdots(N-(n-1)) \\ = \; & N(N-1)(N-2)\cdots(N-n+1) \\ = \; & \frac{N!}{(N-n)!} \end{aligned}$$

after multiplying by $1 = \frac{(N-n)!}{(N-n)!}$.

Example 2. What is the number of different samples you can draw, if a "sample" is just the subset drawn, regardless of the order in which the individuals appeared?

Solution. Suppose we were choosing two kinds of fruit out of a population consisting of apples, bananas, and peaches. From Example 1 we know that the number of ordered pairs we can choose is 6:

Apples and Bananas
Apples and Peaches
Bananas and Apples
Bananas and Peaches
Peaches and Apples
Peaches and Bananas

As you can see, each pair of fruit has been counted twice. For example, Apples and Bananas has been counted as different from Bananas and Apples. But our question wants us to consider these two outcomes to be the same. To account for this, we have to divide 6 by 2, to get the 3 distinct subsets:

Apples and Bananas
Apples and Peaches
Bananas and Peaches

In general, we have to divide the answer to Example 1 by the number of ways of rearranging n individuals in a row. That's $n!$, by applying Example 1 with $N = n$.

Hence, finally, the number of different subsets of n elements out of N is

$$\frac{N!}{(N-n)!} \cdot \frac{1}{n!} = \frac{N!}{n!(N-n)!}$$

where N is the size of the total population and n is the sample size.

This quantity is the number of ways of choosing n individuals out of N. It is therefore called "N choose n", and is denoted

$$\binom{N}{n}$$

Example 3. You are part of a group pool lottery. Among your group of N people, n tickets (one ticket per person) will be drawn to win prizes. What is the chance that you win a prize?

Solution. Assume that the sample of n winning tickets will be drawn at random without replacement. Then the total number of possible samples is

$$\binom{N}{n}$$

all of which are equally likely to result.

Now we have to count the number of samples in which you appear. How do we come up with this number though? The fog clears when we realize that if we know that you are going to be one of the n winning tickets, then there are $n-1$ other spots to fill from among the remaining $N-1$ individuals. Combinatorially then, your being picked is the same as choosing a sample of $(n-1)$ out of $(N-1)$ indviduals. So the number of samples in which you appear is

$$\binom{N-1}{n-1}$$

Therefore the *chance* that you appear in the sample is

$$\frac{\binom{N-1}{n-1}}{\binom{N}{n}}$$

We can simplify this expression to find an interesting result.

$$\begin{aligned}\frac{\binom{N-1}{n-1}}{\binom{N}{n}} &= \frac{(N-1!)}{(n-1)!((N-1)-(n-1))!} \cdot \frac{n!(N-n)!}{N!} \\ &= \frac{n}{N}\end{aligned}$$

We find that the probability of your being chosen is just the fraction of the number of spots relative to the number of people in the population. Why is this the case though? When an answer is as simple as this, it's good to find other ways of understanding it.

To understand the answer, it's important to notice that sampling without replacement can be done in another way. You can randomly shuffle all N individuals and take the first n. Therefore, you will appear in the sample if your ticket falls within the first n tickets among the N shuffled tickets. Since your ticket is equally likely to fall in any of the N spots, the chance that it falls in the first n is just

$$\frac{n}{N}$$

which is exactly what we got by simplification of our previous answer.

This sort of symmetry argument is very powerful in probability, and works very well in problems to do with sampling without replacement.

Example 4. Your best friend is also in your lottery pool. What is the chance that your best friend wins? Is it different from the answer to Example 3?

Solution. The chance that any particular individual is chosen in the sample is the same for all individuals in the sample because all individuals in the sample are treated identically in the sampling process. Therefore, the answer to this example is the same as the answer to Example 3.

Example 5. What is the chance that you and your best friend both win a prize? Would this chance be different for any other pair of individuals on the list?

Solution. We'll use the same logic as in Example 3, and solve the problem in two different ways.

First, we know that the total number of possible samples is $\binom{N}{n}$. The number of samples in which both you and your friend win, is equivalent to fixing two of the spots, and then choosing $n-2$ other people from the remaining pool of $N-2$ people. Thus, this number of samples is $\binom{N-2}{n-2}$. Thus the chance that you and your friend both appear is

$$\frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n}{N} \cdot \frac{n-1}{N-1}$$

after simplifying all the factorials. This answer would be the same no matter which pair of individuals the question was about.

The second method is to remember that the sampling method is equivalent to randomly shuffling all N people and taking the first n. The probability that you are in one of these n spots is $\frac{n}{N}$, and given that, the probability that your friend is also in a good spot is $\frac{n-1}{N-1}$ (since you take up one of the good spots, s/he has one fewer spot to choose from). Thus the overall probability is

$$\frac{n}{N} \cdot \frac{n-1}{N-1}$$

as before.

5.3 Random Permutations

A random permutation is a random ordering of a set of elements. Formally, if a set consists of N elements, then a random permutation is an arrangement of those elements such that all $N!$ arrangements are equally likely.

A practical example of a random permutation is shuffling a deck of cards so that they are randomly distributed. All 52! possible permuations are equally likely in a random shuffle.

In general, let us assume that there are N cards in a deck.

Assume that one of the cards in a deck has a gold star on it and that the deck has been permuted randomly.

Example 1. What is the chance that the card with the gold star is at the top of the deck?

Solution. There are $N!$ ways that the N cards can be arranged in a straight line. Thus, there are $N!$ ways that the deck can be permuted, and all are equally likely.

To find the chance that the top card has the gold star, we have to count the number of permutations in which this occurs. So put the card with the gold star at the top of the deck. Now there are $(N-1)!$ ways that the other cards can be arranged on that line, or in other words, $(N-1)!$ possible deck combinations with the the gold star on the top card.

Hence, the probability that the card with the gold star is on the top is

$$\frac{(N-1)!}{N!} = \frac{1}{N}$$

Another way to arrive at this simple answer is to remember that the card with the gold star is equally likely to be placed anywhere in the deck. Thus the chance that the gold star card is at the top of the deck is $\frac{1}{N}$.

Example 2. What is the chance that the card with the gold star is in the second spot, that is, one below the top of the deck?

Solution: This case also has a card with the gold star occupying a fixed physical position in the deck, i.e., the one below the top.

The argument is exactly the same as in Example 1: to count the number of arrangements that make the event happen, put the gold star card in position 2 and permute the remaining $N-1$ cards. As before, the answer is

$$\frac{(N-1)!}{N!} = \frac{1}{N}$$

Indeed, for any fixed position k in the deck, whether it be the top or the bottom or anywhere in between, the chance the the card with the gold star occupies position k is

$$\frac{1}{N}$$

. To understand this more visually, imagine the deck dealt in a circle. Would you know which was the first card dealt? No, you wouldn't. That is why the chance that the gold star card appears in any specific position is the same as the chance that it appears in Position 1.

Key Takeaway: No matter what k is, the probability that a specified card is in position k in a randomly shuffled deck of N cards is $\frac{1}{N}$.

A *standard deck* is a set of 52 cards with *ranks* given by the set of integers from 2 to 10, Jack, Queen, King, and Ace. There are 13 complete sets of ranks in each of 4 suits:

$$\{\spadesuit, \heartsuit, \clubsuit, \diamondsuit\}$$

These are called spades, hearts, clubs, and diamonds respectively. Spades and clubs are black. Hearts and Diamonds are red.

Example 3. Suppose you deal a poker hand, that is, 5 cards at random without replacement. What is the chance that the last card dealt is the ace of spades?

Solution. The power of the symmetry argument of Example 2 is that we can answer this new question without any further calculation. The chance that the ace of spades is in position 5 (or any other specified position) is simply 1/52.

Example 4. Suppose you deal all 52 cards at random. What is the chance that the 37th dealt is black?

Solution. For any particular black card, say the ace of spades, the chance that the card is in position 37 is 1/52. There are 26 black cards. So the chance that one of them is in position 37 is 26/52.

Notice that "37" doesn't appear anywhere in the answer. The only thing that matters is that the question is about one position.

Example 5. Suppose you shuffle the whole deck. What is the chance that the 28th card dealt is red given that the 43th card dealt is black?

Solution. Imagine the deck dealt in a circle again. We can see a black card in one spot. So each red card is equally likely to be in any of the 51 remaining spots. So, for example, the chance that the ace of hearts is in Spot 28 is 1/51. There are 26 red cards in all, so

$$P(\text{28th card is red}) = \frac{26}{51}$$

Again, notice that "28" and "43" don't appear anywhere in the answer on the right hand side. The only thing that matters is that the question is about one position and you are given information about another.

Example 6. What is the chance that the last four cards dealt are all aces?

Solution. By symmetry (imagine the deck dealt in a circle again), this is the same as the chance that the first four cards dealt are all aces. That's

$$\frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} \cdot \frac{1}{49}$$

Note that this probability would be correct as the chance of dealing aces in any four fixed locations within the deck, whether at the top of the deck, bottom of the deck, or dispersed in between.

These examples all show that the probability of drawing specified cards at predetermined locations does not depend on where the locations are, but on how many locations there are and what (if anything) is known about the cards in the other locations.

5.4 Questions

1. Let n and k be integers such that $0 < k < n$. Consider the following three quantities:

$$\binom{n-1}{k} \qquad \binom{n}{k} \qquad \binom{n-1}{k-1}$$

   One of them is equal to the sum of the other two. Which one is it? Justify your answer either by algebra or by counting subsets. For the latter approach, it might help to consider the following – if you are one of n students in a class, then how many subsets of k of the n students contain you? How many don't?

2. A standard deck consists of 52 cards. The two red aces are the ace of hearts and the ace of diamonds. The deck is well shuffled, so that all permutations are equally likely.

   a) What is the chance that the ace of hearts ends up at the top of the deck and the ace of diamonds at the bottom?

   b) What is the chance that one of the red aces ends up at the top of the deck and the other one at the bottom?

3. Jo, Bo, and Mo are in a class that has a total of N students. A sample of n students will be chosen at random without replacement from this class. The order in which the students are chosen doesn't matter.

   a) What is the chance that all three of Jo, Bo, and Mo are chosen?

   b) What is the chance that Jo and Bo are chosen and Mo is not?

4. A population consists of 20 people, of whom 8 are women and 12 are men. A simple random sample of 5 people will be drawn from the population. Which of the following is the chance that the sample contains exactly 3 women? Why?

$$(i) \quad \frac{\binom{8}{3}}{\binom{20}{5}} \qquad\qquad (ii) \quad \frac{\binom{8}{3}\binom{12}{2}}{\binom{20}{5}}$$

5. **The Urn Problem from Peter Norvig's Talk.** An urn contains 23 balls: 8 white, 6 blue, and 9 red. We select six balls at random (each possible selection is equally likely).

   Peter Norvig is assuming that you're selecting a set of six balls, that is, six distinct balls. In other words, the draws are made without replacement.

   Find the probability that:

   a) all the balls are red

   b) 3 are blue, 2 are white, and 1 is red

   c) exactly 4 balls are white

6. A fair six-sided die will be rolled 10 times. Consider the smallest of the 10 numbers rolled.

   a) For each k in the range $1, 2, 3, 4, 5, 6$, find the chance that the smallest number rolled is larger than k.

   b) For each k in the range $1, 2, 3, 4, 5, 6$, let p_k be the probability that the smallest number rolled is equal to k. Use your answers to part

   c) a to find p_k.

   d) Find the sum of your answers to part

   e) b). Do the sum and get a numerical answer; don't just explain what the sum ought to be.

7. A coin that lands heads with chance 1/1000 is tossed 1000 times. Find the chance that all the tosses show tails. Find a simple approximation for this chance in terms of e.

8. A coin that lands heads with chance $1/N$ is tossed n times. Find the chance that all the tosses show tails. Find a simple exponential approximation to this chance, assuming that $N \to \infty$ and $n \to \infty$ in such a way that $n/N \to \mu$ for some positive number μ.

Chapter 6

Random Variables

6.1 Random Variables

In this chapter, we will put together the ideas we have developed for probability and descriptive statistics, to build tools that will help us understand statistics such as the average of a random sample.

The story begins with an *outcome space*, that is, the set of all possible outcomes of an experiment that involves chance. Standard notation for this space is Ω, the upper case Greek letter Omega. **For mathematical simplicity, we will assume that the outcome space Ω is finite.** Each outcome ω (that's lower case omega) is assigned a probability, and the total probability of all the outcomes is 1.

Example 1. You flip a fair coin three times, and record what face the coin landed on. What is a reasonable outcome space Ω for this experiment, and what is the probability distribution on that space ?

You can think of Ω as consisting of 8 equally likely outcomes, since the coin is not biased. Here's a table representing Ω and the probabilities of all the outcomes.

ω	TTT	TTH	THT	HTT	THH	HTH	HHT	HHH
$P(\omega)$	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

A *random variable* X is a real-valued function defined on Ω. That is, the domain of X is Ω and the range of X is the real line.

Note. We are going to restrict attention to random variables that have a finite number of values.

In Example 1, X could be the number of times the letter H appears in an outcome. You can think of X as the number of heads in three tosses of a coin.

ω	TTT	TTH	THT	HTT	THH	HTH	HHT	HHH
$P(\omega)$	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
$X(\omega)$	0	1	1	1	2	2	2	3

The probability function on Ω determines probabilities for X. For example, the chance that X is 1 is defined as follows:

$$P(X = 1) \;=\; P(\{\omega : X(\omega) = 1\}) \;=\; P(\text{TTH, THT, HTT}) \;=\; 1/8 + 1/8 + 1/8 \;=\; 3/8$$

6.2 Probability distribution

The *probability distribution* of X is a distribution on the range of X. It specifies all the possible values of X along with all their probabilities.

For random variables that have a finite number of possible values, the probability distribution is also known as the *probability mass function*.

For example, for X defined in the example above, the probability distribution is given by

x	0	1	2	3
$P(X = x)$	1/8	3/8	3/8	1/8

All the probabilities in a distribution must add up to 1.

Note that the probabilities on the range of X are determined by the probabilities on the domain. In our example, had all 8 elements in the domain not been equally likely, the possible values of X would still have been the same but the probabilities might have been different.

Example 2: Uniform distribution A word is picked at random from the set science, computer. Let V be the number of distinct vowels in the word. What is the distribution of V?

Solution. Ω = {science, computer}. The probabililty distribution on Ω assigns probability 1/2 to each element. Now V(science) = 2 and V(computer) = 3.

So the distribution of V puts probability 1/2 on each of the values 2 and 3. We say that V has the *uniform* distribution on $\{2, 3\}$.

Example 3: Binomial distribution. Suppose a coin is tossed n times. Let X be the number of heads. Find the distribution of X.

Solution. Each possible outcome of n tosses is a sequence of n H's and T's. You have seen in an earlier exercise that there are 2^n equally likely such sequences.

When you are finding a distribution, an excellent idea is to start with the possible values rather than the probabillities. The possible number of heads in n tosses is 0 through n.

For any fixed k in the range 0 through n, the number of sequences that contain exactly k H's is $\binom{n}{k}$. So the distribution of X is

$$P(X = k) = \frac{\binom{n}{k}}{2^n}, \quad k = 0, 1, \ldots, n$$

This is called the *binomial distribution with parameters n and* 1/2. You should convince yourself that the formula gives a sensible answer in the two edge cases $k = 0$ and $k = n$.

If the coin is unfair, it is still true that there are $\binom{n}{k}$ sequences with exactly k H's, but all 2^n sequences are no longer equally likely. To find the probability of each sequence, let p be the chance that the coin lands heads on a single toss, and make the natural assumption that the outcome of any set of tosses does not affect chances for any other set.

Let's work out an example under these assumptions before finding the general formula. Suppose $n = 3$ as in Example 1, and let $k = 2$. Then the sequences corresponding to 2 heads in 3 tosses are HHT, HTH, and THH. Each of these has probability $p^2(1-p)$, because there are two factors of p and one of $(1-p)$ in different order. So the chance of 2 heads in 3 tosses is P(HHT or HTH or THH) = $3p^2(1-p)$.

Notice that the 3 in the answer above is the number of sequences that have 2 H's, that is, $3 = \binom{3}{2}$.

Now we are ready to extend the argument to the general case. Suppose the coin lands heads on a single toss with probability p. If X is the number of heads on n tosses, then the distribution of X is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

This is called the *binomial distribution with parameters n and p*. Notice that the fair coin is a special case:

$$P(X = k) = \frac{\binom{n}{k}}{2^n} = \binom{n}{k} (\frac{1}{2})^k (\frac{1}{2})^{n-k}$$

The name *binomial* comes from the fact that the probabilities in the binomial distribution are the terms in the *binomial expansion* $(a+b)^n$ for $a = p$ and $b = 1-p$. This also shows why the sum of the terms is equal to 1.

The distribution can be used to find probabilities of events, as in the following examples.

The chance of getting between 45 and 55 heads (inclusive) in 100 tosses of a fair coin is

$$\sum_{k=45}^{55} \frac{\binom{n}{k}}{2^n}$$

The chance of getting fewer than 10 sixes in 12 rolls of a die is

$$\sum_{k=0}^{9} \binom{12}{k} (\frac{1}{6})^k (\frac{5}{6})^{12-k} = 1 - \sum_{k=10}^{12} \binom{12}{k} (\frac{1}{6})^k (\frac{5}{6})^{12-k}$$

The second form is simpler to calculate, as the sum only has three terms.

Example 4. A 5-card poker hand is dealt from a standard deck. What is the distribution of the number of queens in the hand?

Solution. Let's give the variable a name: Q for "queens". We will start by listing the possible values of Q. We are dealing 5 cards but there are only 4 aces in the deck. So the possible values of Q are $0, 1, 2, 3, 4$.

The total number of possible hands is the total number of subsets of 5 that can be chosen from among 52 cards. That's $\binom{52}{5}$, and they are all equally likely.

Now fix k in the range 0 through 4. To find $P(Q = k)$, we will need the number of hands that contain exactly k queens. These k queens can be chosen in $\binom{4}{k}$ ways. For each of these ways, there are $\binom{48}{5-k}$ ways of choosing the remaining $5-k$ cards in the hand. So the number of 5-card hands that contain exactly k queens is $\binom{4}{k}\binom{48}{5-k}$.

So the distribution of Q is given by

$$P(Q = k) = \frac{\binom{4}{k}\binom{48}{5-k}}{\binom{52}{5}}, \quad k = 0, 1, 2, 3, 4$$

The chance of three or more queens in a 5-card hand is

$$P(Q \geq 3) = \frac{\binom{4}{3}\binom{48}{2}}{\binom{52}{5}} + \frac{\binom{4}{4}\binom{48}{1}}{\binom{52}{5}}$$

6.3 Functions of Random Variables

We will frequently be interested in quantities that can be calculated based on a random variable. In other words, we will be interested in functions of random variables. A function of a random variable is also a random variable, because a function of a function on Ω is also a function on Ω.

Example 5. Let X have the binomial $(4, 1/2)$ distribution. Let Y be the absolute deviation of X from 2, that is, $Y = |X - 2|$. What is the distribution of Y?

Solution. The table below gives the distribution of X along with the possible value y of Y computed from each possible value x of X:

y	2	1	0	1	2
x	0	1	2	3	4
$P(X = x)$	1/16	4/16	6/16	4/16	1/16

Collect terms get the distribution of Y:

y	0	1	2
$P(Y = y)$	6/16	8/16	2/16

Example 6: Net gain on red at roulette. In Nevada roulette, the bet on "red" pays 1 to 1 and you have 18 in 38 chances of winning. This means that if you bet a dollar on "red" and the winning color is not red, you lose your dollar; if the winning color is red, then your net winnings are a dollar. Suppose the roulette wheel is spun 10 times, and you bet a dollar on red each time. What is the chance that you make money?

Solution. Let X be the number of times you win. Then X has the binomial distribution with parameters $n = 10$ and $p = 18/38$. But the question is about your net gain being positive overall. So let your net gain be G, and let's try and work out what G has to be for you to make money.

For any possible value x of X, the amount of money you make is the total of \$1 for each of the x bets that you win and -\$1 for each of the $10 - x$ bets that you lose. This corresponds to making g dollars overall, where

$$g = 1 \cdot x + (-1) \cdot (10 - x)$$

So

$$g = 2x - 10$$

So your random net gain G is the following function of the random number of bets X that you win:

$$G = 2X - 10$$

The question asks for $P(G > 0)$.

$$P(G > 0) = P(2X - 10 > 0) = P(X > 5) = \sum_{x=6}^{10} \binom{10}{x} (\frac{18}{38})^x (\frac{20}{38})^{10-x}$$

6.4 Expectation

The *expectation* of X, denoted $E(X)$, is the average of the possible values of X, weighted by their probabilities.

Remember that we are assuming that X has finitely many values. So the expectation of X is finite.

The expectation can be computed in two equivalent ways, one defined on the domain of X and one on the range:

$$E(X) = \sum_{\omega} X(\omega)P(\omega) = \sum_{x} xP(X = x)$$

Let X be the number of heads in three tosses of a coin. The natural outcome space Ω, along with the probabilities of all the outcomes ω, is given by:

ω	TTT	TTH	THT	HTT	THH	HTH	HHT	HHH
$P(\omega)$	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

X is a function from Ω to $\{0, 1, 2, 3\}$, with distribution given by:

x	0	1	2	3
$P(X = x)$	1/8	3/8	3/8	1/8

So the first form of the calculation of the expectation of X is

$$E(X) = 0 \cdot 1/8 + 1 \cdot 1/8 + 1 \cdot 1/8 + 1 \cdot 1/8 + 2 \cdot 1/8 + 2 \cdot 1/8 + 2 \cdot 1/8 + 3 \cdot 1/8 = 1.5$$

The second form is based on the probability distribution of X:

$$E(X) = 0 \cdot 1/8 + 1 \cdot 3/8 + 2 \cdot 3/8 + 3 \cdot 1/8 = 1.5$$

Expectation is often denoted by μ, the lower case Greek letter mu. That's because expecations and averages are often called *means*.

The form

$$E(X) = \sum_{x} xP(X = x)$$

is most commonly used in calculations involving simple distributions. Go back and look at the section **Another Way to Calcuate the Average** at the end of the section on averages in Chapter 1. You will recognize that the formula there is analogous to the formula for $E(X)$ above.

This implies that $E(X)$ is just an ordinary average and thus has all the familiar properties of averages. For example, if the random variable X is a constant, that is, if there is a constant c such that $P(X = c) = 1$, then $E(X) = c$.

Linear transformations. Also, expectation transforms linearly when the random variable is transformed linearly:

$$E(aX + b) = aE(X) + b$$

The expectation is the balance point of the histogram of the probability distribution. And so on. Expecations have the properties of averages.

Expectation of a function of X. Any function of X is also a random variable, and thus has an expectation. Usually, the easiest way to find the expectation of a function of X is to do the calculation using the distribution of X. That is, if the random variable $Y = g(X)$, then

$$E(Y) = E(g(X)) = \sum_x g(x)P(X = x)$$

So if X is the number of heads in three tosses of a coin, and $Y = |X - 1.5|$, then

$$\begin{aligned} E(Y) &= E(|X - 1.5|) \\ &= |0 - 1.5| \cdot (1/8) + |1 - 1.5| \cdot (3/8) + |2 - 1.5| \cdot (3/8) + |3 - 1.5| \cdot (1/8) \end{aligned}$$

As you will see below, we will sometimes need to find the expectation of the square of a random variable. By the "function rule" above,

$$E(X^2) = \sum_x x^2 P(X = x)$$

6.5 Standard Deviation and Bounds

How far away from its expectation is a random variable likely to be? To answer this question, we need a measure of deviation away from the expectation. We will develop one that is analogous to the standard deviation of a list of numbers.

Define the *deviation* of X to be the random variable $D = X - \mu$, where $\mu = E(X)$.

To find the rough size of D, suppose we calculate $E(D)$. Notice that D is a linear transformation of X. By our rule about linear tranformations,

$$E(D) = E(X - \mu) = E(X) - \mu = \mu - \mu = 0$$

The expected deviation is 0, no matter what the distribution of X is. This is the parallel to the result that the average of a list of deviations from average is 0 no matter what the list is.

As with the deviations of values in a list, the problem is cancellation: the negative deviations cancel out the positive ones. So we can't learn anything meaningful about the spread of the variable if we just calculate the expected deviation.

So, just as we did with the deviations from average of a list of numbers, we'll square the deviations to ensure that they don't cancel each other out. This leads us to the definition of the *variance* of X, denoted $Var(X)$.

$$Var(X) = E[D^2] = E[(X - \mu)^2]$$

To correct the units of measurement, we have to take the square root of the variance. The *standard deviation* of X is then

$$SD(X) = \sqrt{Var(X)} = \sqrt{E(D^2)} = \sqrt{E[(X - \mu)^2]}$$

$SD(X)$ is often denoted by σ, the lower case Greek letter sigma.

The standard deviation defined here can also be thought of as the ordinary SD of a list of numbers, computed using the distribution table of the list. As such, it is an ordinary SD and has all the properties of SDs that we discovered in Chapter 2.

For example, if a random variable X is a constant, then $SD(X) = 0$.

Linear transformations. Only the mutiplicative factor of a linear transformation affects the SD:

$$SD(aX + b) = |a|SD(X)$$

Thus if X is a random temperature in degrees Celsius, and Y the corresponding temperature in degrees Fahrenheit, then $Y = (9/5)X + 32$. So

$$E(Y) = (9/5)E(X) + 32 \qquad SD(Y) = (9/5)SD(X)$$

Another consequence is that a constant shift simply slides the probability distribution along and doesn't affect the SD. For any constant c,

$$SD(X + c) = SD(X)$$

6.6 Bounding Tail Probabilities

The tail bounds that we established for distributions of data work for probability distributions of random variables as well.

Markov's Inequality

Let X be a non-negative random variable. That is, assume that $P(X \geq 0) = 1$. Let $E(X) = \mu$. Then for any constant $c > 0$

$$P(X \geq c) \leq \frac{\mu}{c}$$

Chebychev's Inequality

Let X be a random variable that has $E(X) = \mu$ and $SD(X) = \sigma$. Let k be any positive constant. Then

$$P(X \text{ is outside the range } \mu \pm k\sigma) \leq \frac{1}{k^2}$$

More formally, for all $k > 0$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

These are exactly the same as the bounds obtained earlier for distributions of data, and they are true for the same reasons.

If you would like to prove them afresh, you can follow the steps in the proof in Chapters 1 and 2 but write them in random variable notation.

Standard Units

Let X be a random variable with expecation μ_X and SD σ_X. The random variable Z defined by

$$Z = \frac{X - \mu_X}{\sigma_X}$$

is called X *in standard units*. Notice that Z is a linear transformation of X.

As you can see from the definition of Z, standard units measure the deviation of X relative to the SD of X. In other words, they measure *how many SDs above the expected value* the value of X is.

Conversion to standard units allows us to compare the distributions of random variables that have been measured on different scales.

Measuring a random variable in standard units is equivalent to setting the origin to be μ_X and measuring distances from the origin in units of SDs. By our results about linear transformations,

$$\mu_Z = E(Z) = \frac{E(X) - \mu_X}{\sigma_X} = 0$$

and

$$\sigma_Z = SD(Z) = \frac{SD(X)}{\sigma_X} = 1$$

We can re-write Chebychev's Inequality in terms of Z. For any $k > 0$,

$$P(|X - \mu_X| \geq k\sigma_X) = P(|\frac{X - \mu_X}{\sigma_X}| \geq k) = P(|Z| \geq k)$$

So if Z is X measured in standard units, Chebychev's Inequality becomes

$$P(|Z| \geq k) \leq \frac{1}{k^2}$$

The chance that a random variable is at least 4 in standard units (that is, at least 4 SDs away from its expected value) is 1/16, which is quite small. Thus, when any random variable is measured in standard units, the bulk of its probability distribution lies in the interval $(-4, 4)$.

6.7 Questions

1. Let X be a random variable, and for constants a and b let $Y = aX + b$. Derive a formula for $E(Y)$, the expectation of Y, in terms of a, b, and $E(X)$.

2. Suppose X represents the number of spots showing on one roll of a strange six-sided die. The die is strange in the following sense. It has probability p of showing 1 or 6 spots; the probability p is split evenly between those two faces. The remaining probability $1 - p$ is split evenly among the remaining four faces (the faces showing 2, 3, 4, and 5 spots).

   Find $E(X)$. Give a calculation or explanation.

3. The Statistics department used to have a "birthday cake" tradition: once a month, there would be a cake to celebrate the birthdays of all department members whose birthday was in that month. Suppose this tradition continues and suppose the department has n members. Assume that each person's birthday is equally likely to be in any of the 12 months, independently of everyone else's birthday.

   Consider one calendar year (January through December) starting in January.

   a) What is the chance that no birthday cake will be needed after September?

   b) What is the chance that there will be birthday cake in September but not after that?

4. Suppose X is the random variable in Question 2. Find a formula for the variance of X (in terms of p), and explain what happens to the variance as p increases.

5. An instructor is trying to set up office hours during RRR week. On one day there are 8 available slots: 10-11, 11-noon, noon-1, 1-2, 2-3, 3-4, 4-5, and 5-6. There are 6 GSIs, each of whom picks one slot. Suppose the GSIs pick the slots at random, independently of each other.

   a) Find a decimal answer for the expected number of slots that no GSI picks.

   b) Find the chance that six different slots are chosen.

   c) There are 100 prize tickets among 1000 tickets in a lottery. Suppose you buy 12 tickets; assume that the tickets are like 12 draws made at random without replacement from among the 1000 tickets.

      i. What is the expected number of prizes you get? Please provide a decimal answer.

      ii. What is the chance that you get at least one prize?

Chapter 7

Sums of Random Variables

7.1 Joint Distributions

In data science, we frequently study the relation between several random variables defined on the same space. For example, we might look at first two cards dealt at random from a deck. Or if we are tossing a coin 15 times, we might look at the relation between the number of heads in the first 10 tosses and the number of heads in the last 10 tosses.

The *joint distribution* of random variables X and Y defined on the same space is the set of probabilities $P(X = x, Y = y)$ for all posssible values x of X and y of Y. This is just an ordinary probability distribution over the set of all pairs (x, y). So

$$\sum_x \sum_y P(X = x, Y = y) \;=\; 1$$

Because the event $\{Y = y\}$ can be partitioned according to how X came out, we have

$$\{Y = y\} \;=\; \bigcup_x \{X = x, Y = y\}$$

where the right hand side is a disjoint union. So

$$P(Y = y) \;=\; \sum_x P(X = x, Y = y)$$

Note that y is a constant here; it is x that varies across all the possible values of X.

Example 1. Let C be uniform on $\{1, 2, 3\}$ and let H be the number of heads in C tosses of a coin. Find $P(H = 2)$.

Solution. To get 2 heads, you had to toss 2 or 3 coins. So

$$\begin{aligned} P(H = 2) \;&=\; P(C = 2, H = 2) + P(C = 3, H = 2) \\ &=\; \frac{1}{3} \cdot \binom{2}{2} \cdot \frac{1}{4} \;+\; \frac{1}{3} \cdot \binom{3}{2} \cdot \frac{1}{8} \end{aligned}$$

7.2 The Expectation of a Sum

We will now look at functions of several random variables defined on the same space. Among the simplest and most powerful such functions is the sum. Let X and Y be two random variables defined on the same space Ω. For every $\omega \in \Omega$, define the sum

$$S(\omega) = X(\omega) + Y(\omega)$$

In random variable notation, $S = X + Y$.

Addition Rule for Expectation

Let $S = X + Y$. Then, no matter what the relation between X and Y,

$$E(S) \;=\; E(X+Y) \;=\; E(X) \;+\; E(Y)$$

This is a result of fundamental importance, as you will see in the examples below. But first, let us prove it.

Proof. We will use the first definition of expectation, where the sum is over all the elements in the domain of the random variable.

$$\begin{aligned} E(S) &= \sum_{\omega} S(\omega)P(\omega) \\ &= \sum_{\omega} (X(\omega) + Y(\omega))P(\omega) \\ &= \sum_{\omega} X(\omega)P(\omega) \;+\; \sum_{\omega} Y(\omega)P(\omega) \\ &= E(X) \;+\; E(Y) \end{aligned}$$

Example 1. Two dice are rolled. What is the expected total number of spots? What would be the expected number of spots if 10 dice were rolled?

Solution. Let X_i be the number of spots on Roll i. Then for each i, X_i has the uniform distribution on $\{1, 2, 3, 4, 5, 6\}$, and so $E(X_i) = 3.5$. So by the addition rule, $E(X_1 + X_2) = 3.5 + 3.5 = 7$.

By induction on the number of random variables being added, the addition rule for expectation extends to any finite sum of random variables. So the expected total number of spots on 10 rolls of a die is 35.

Sum and Average of Identically Distributed Random Variables

Let $X_1, X_2, \ldots, X_n$ all have the same distribution, and let $E(X_1) = \mu$. Since all the X_i's have the same distribution, $E(X_i) = \mu$ for each i. Now let $S_n = X_1 + X_2 + \cdots + X_n$. Then by the addition rule,

$$E(S_n) \;=\; n\mu$$

Let A_n be the "sample average" defined by $A_n = S_n/n$. Then

$$E(A_n) \;=\; \frac{n\mu}{n} \;=\; \mu$$

In sampling language, this result says that the expected value of a random sample average is equal to the average of the population from which the sample is drawn.

Example 2. Suppose a population of incomes has mean \$70,000. Let X_1, X_2, X_3 be a random sample drawn at random without replacement from this population. Find $E(S_3$ and $E(A_3)$. How would your answers change if the sample were drawn with replacement?

Solution. The addition rule for expectation doesn't depend on the relation between the random variables. By the symmetry observed in the Random Permutations section, each X_i has the same distribution. So $E(S_3) = \$210,000$ and $E(A_3) = \$70,000$. It doesn't matter whether the sample was drawn with or without replacement.

The Method of Indicators

Counting is the same as adding 0's and 1's: 1 for each element that you want to count, and 0 for each of the others. To formalize this idea, let A be any event and define the random variable I_A by

$$\begin{aligned} I_A &= 1 \text{ if } A \text{ happens} \\ &= 0 \text{ if } A \text{ doesn't happen} \end{aligned}$$

Then I_A is called the *indicator of A*, and

$$E(I_A) = 1 \cdot P(A) + 0 \cdot P(A^c) = P(A)$$

Thus the expectation of the indicator of an event is the probability of the event.

Example 3: The Expectation of the Binomial. Let X be the number of heads in n tosses of a coin that lands heads with probability p. Then

$$X = I_1 + I_2 + \cdots + I_n$$

where I_j is the indicator of heads on toss j. That is,

$$\begin{aligned} I_j &= 1 \text{ if toss } j \text{ lands heads} \\ &= 0 \text{ if toss } j \text{ lands tails} \end{aligned}$$

Now $E(I_j) = p$ for all j, so

$$E(X) = np$$

Example 4. A five-card poker hand is dealt from a well-shuffled deck. Find the expected number of aces in the hand.

Solution. Let X be the number of aces. Then

$$X = I_1 + I_2 + I_3 + I_4 + I_5$$

where I_j is the indicator that card j is an ace. That is,

$$\begin{aligned} I_j &= 1 \text{ if card } j \text{ is an ace} \\ &= 0 \text{ if card } j \text{ is not an ace} \end{aligned}$$

By the symmetry of random permutations, $E(I_j) = 4/52$ for all j, so

$$E(X) = 5 \cdot \frac{4}{52}$$

Computational Formula for Variance

The addition rule for expectation can be used to derive a useful result about computing variance.

Let $E(X) = \mu$ and $SD(X) = \sigma$. Then

$$\sigma^2 \;=\; Var(X) \;=\; E(X^2) \;-\; \mu^2$$

Notice that this is the random variable analog of the computational formula for the variance of a list of numbers, derived in Chapter 3.

Proof.

$$\begin{aligned}
\sigma^2 \;=\; Var(X) &= E[(X-\mu)^2] \\
&= E(X^2 - 2X\mu + \mu^2) \\
&= E(X^2) \;-\; E(2X\mu) \;+\; E(\mu^2) \text{ by the addition rule} \\
&= E(X^2) \;-\; 2\mu E(X) \;+\; \mu^2 \\
&= E(X^2) \;-\; \mu^2
\end{aligned}$$

Thus the variance is the expectation of the square minus the square of the expectation.

You will have noticed that this is the random variable analog of the computational formula for the variance of a list of numbers, derived in Chapter 3. It shows that given any two of μ, σ, and $E(X^2)$, you can compute the third. For example,

$$E(X^2) \;=\; \sigma^2 \;+\; \mu^2$$

Note that in general, $E(X^2) \neq \mu^2$. That is, the expectation of the square is in general *not* the square of the expectation. Unlike a linear function, the square is not preserved by expectation.

7.3 The Variance of a Sum

We will now develop tools that will allow us to derive the expectation and SD of the sum of a random sample, and hence derive parallel results for the random sample mean.

Independence

We have used this concept frequently in calculating probabilities, but have not yet given it a formal name. For example, we have said that draws with replacement don't affect each other; or that given the result of one draw, chances for the other draws remain unchanged.

Formally, events A and B are *independent* if the conditional chance of B given that A has occurred is the same as the unconditional chance of B. That is, A and B are independent if

$$P(B \mid A) \;=\; P(B)$$

Along with the multiplication rule, this motivates the formal definition of independence:

Events A and B are independent if

$$P(AB) \;=\; P(A)P(B)$$

Notice that according to the multiplication rule, the second factor on the right hand side should be the conditional chance of B given that A has happened; but independence means that the value of this conditional chance is the same as $P(B)$.

Random variables X and Y are *independent* if for every x and y,

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Products of Independent Random Variables

Products of random variables can be hard to understand. But as you will soon see, they crop up when we calculate the variance of a sum, and so it is a good idea to examine properties of products. A useful and simple property is that the expectation of a product of random variables is the product of the expectations, if the random variables are independent:

$$E(XY) = E(X)E(Y) \quad \text{if } X \text{ and } Y \text{ are independent}$$

Proof.

$$\begin{aligned} E(XY) &= \sum_x \sum_y xyP(X = x, Y = y) \\ &= \sum_x \sum_y xyP(X = x)P(Y = y) \text{ by independence} \\ &= \sum_x xP(X = x) \sum_y yP(Y = y) \text{ by pulling all the } x \text{ terms out of the inner sum} \\ &= E(X)E(Y) \end{aligned}$$

Note that independence is an important assumption here. If X and Y are not independent, then $E(XY)$ need not be equal to $E(X)E(Y)$. For example, in the extreme case of dependence where $X = Y$, we have $E(XY) = E(X^2)$ which we know is in general *not* equal to $E(X)E(X) = (E(X))^2$.

Addition Rule for Variance

We have seen that $E(X + Y) = E(X) + E(Y)$ no matter what the relation between X and Y. We will now show that

$$Var(X + Y) = Var(X) + Var(Y) \text{ if } X \text{ and } Y \text{ are independent}$$

Proof. For notational convenience, let $E(X) = \mu_X$, $SD(X) = \sigma_X$, $E(Y) = \mu_Y$, and $SD(Y) = \sigma_Y$. Also let $D_X = X - \mu_X$ and $D_Y = Y - \mu_Y$ be the two deviations. Recall that $E(D_X) = 0 = E(D_Y)$. Finally, note that since X and Y are independent, so are D_X and D_Y.

$$\begin{aligned}
Var(X+Y) &= E[((X+Y)-E(X+Y))^2] \\
&= E[((X+Y)-(\mu_X+\mu_Y))^2] \quad \text{by additivity of expectation} \\
&= E[((X-\mu_X)+(Y-\mu_Y))^2] \\
&= E[(D_X+D_Y)^2] \\
&= E(D_X^2+2D_XD_Y+D_Y^2) \\
&= E(D_X^2)+2E(D_XD_Y)+E(D_Y^2) \quad \text{by additivity again} \\
&= Var(X)+2E(D_XD_Y)+Var(Y) \quad \text{by definition of variance} \\
&= Var(X)+2E(D_X)E(D_Y)+Var(Y) \quad \text{because } D_X \text{ and } D_Y \text{ are independent} \\
&= Var(X)+Var(Y) \quad \text{because } E(D_X)=0=E(D_Y)
\end{aligned}$$

The Expectation and SD of a Random Sample Sum

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables. That is one of the most commonly used acronyms in probability theory.

For example, $X_1, X_2, \ldots, X_n$ could be the results of draws made at random **with** replacement from a population. Let μ be the average of the population and σ the SD of the population. Then each X_i has the same expectation μ and the same SD σ as all the others.

Let $S_n = X_1 + X_2 + \cdots + X_n$.

Then for all $n \geq 1$,

$$E(S_n) = n\mu \qquad SD(S_n) = \sqrt{n}\sigma$$

Proof. As we saw in the previous section, the addition rule for expectation says that

$$E(S_n) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = n\mu$$

Because $X_1, X_2, \ldots, X_n$ are independent, we also have additivity of variance:

$$Var(S_n) = Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) \quad \text{by independence}$$

$$= n\sigma^2$$

Therefore $SD(S_n) = \sqrt{n}\sigma$.

You can see that as the sample size increases, the expectation of the sum of an i.i.d. random sample gets larger. The SD of the sum gets larger as well, but at a slower rate ($\sqrt{n}$ compared to n) than the expectation.

Example: The expecation and the SD of the binomial. Let X be the number of heads in n tosses of a coin that lands heads with proability p. Find $E(X)$ and $SD(X)$.

Solution. We wil use the method indicators, as we did in the previous section.

$$X = I_1 + I_2 + \cdots + I_n$$

where I_j is the indicator of heads on toss j. That is,

$$\begin{aligned} I_j &= 1 \text{ if toss } j \text{ lands heads} \\ &= 0 \text{ if toss } j \text{ lands tails} \end{aligned}$$

Now $E(I_j) = p$ for all j, so

$$E(X) = np$$

To find $SD(I_j)$, notice that $I_j^2 = I_j$ since the only possible values of I_j are 0 and 1. So

$$\begin{aligned} Var(I_j) &= E(I_j^2) - (E(I_j))^2 \quad \text{by the computational formula for variance} \\ &= E(I_j) - (E(I_j))^2 \quad \text{since } I_j^2 = I_j \\ &= p - p^2 \quad \text{since } E(I_j) = p \\ &= p(1-p) \end{aligned}$$

Thus

$$SD(I_j) = \sqrt{p(1-p)}$$

Now $I_1, I_2, \ldots, I_n$ are independent as they are indicators of heads on different tosses of a coin. So the number of heads

$$X = I_1 + I_2 + \cdots + I_n$$

is the sum of n i.i.d. indicators, and hence

$$SD(X) = \sqrt{n} \cdot \sqrt{p(1-p)} = \sqrt{np(1-p)}$$

The Expectation and SD of a Random Sample Mean

Let $X_1, X_2, \ldots, X_n$ be i.i.d. as above. Let $A_n = S_n/n$ be the sample average. Then for all $n \geq 1$,

$$E(A_n) = \mu \qquad SD(A_n) = \frac{\sigma}{\sqrt{n}}$$

Proof. A_n is a linear transformation of S_n:

$$A_n = \frac{S_n}{n}$$

Therefore, by properties of expectation and SD under linear transformations,

$$E(A_n) = \frac{E(S_n)}{n} = \frac{n\mu}{n} = \mu$$

$$SD(A_n) = \frac{SD(S_n)}{n} = \frac{\sqrt{n}\sigma}{n} = \frac{\sigma}{\sqrt{n}}$$

Thus, no matter what the sample size, the expectation of the average of an i.i.d. random sample is always equal to the population average. But as the sample size increases, the SD of the sample average decreases. Because of the factor of $\sqrt{n}$ in the denominator, we have the following law:

The Square Root Law. If you increase the sample size by a factor f, then the SD of the average of an i.i.d. random sample decreases by a factor of $\sqrt{f}$.

The expected value of the random sample average is the population average, no matter what the sample size. Because smaller SDs correspond to more accurate estimates of the population mean, we say that if you increase the sample size by a factor of f, then the accuracy of the i.i.d. random sample average increases by a factor of $\sqrt{f}$.

As a consequence, to increase accuracy by a factor of 10, you have to multiply the sample size by a factor of 100. Accuracy doesn't come cheap.

7.4 Questions

1. An office worker in a city takes a bus to work. Her journey consists of the following four random components.

   X_1: the time it takes her to walk from home to the bus stop; $E(X_1) = 7$ minutes

   X_2: the time she waits at the bus stop till she gets on the bus; $E(X_2) = 6$ minutes

   X_3: the time she spends on the bus till she gets out at the stop nearest her work; $E(X_3) = 12$ minutes

   X_4: the time it takes her to walk from the bus stop (where she got out) to her office; $E(X_4) = 5$ minutes

   a) Let X be the total time from her home to her office. Find $E(X)$, the expectation of X.

   b) The times X_i, $i = 1, 2, 3, 4$ are not independent of each other. For example, X_1 and X_4 are both affected by the office worker's state of health or whether it is raining. If X_1, X_2, X_3, and X_4 had been independent, would your answer to **a** have been different? Explain briefly.

2. A class has 7 GSIs. Each GSI rolls a die 4 times. Let X be the number of GSIs who get the face with six spots at least once. Find $E(X)$ and $SD(X)$.

3. Let X be a random variable such that $E(X) = \mu_X$ and $SD(X) = \sigma_X$. Let Y be independent of X, with $E(Y) = \mu_Y$ and $SD(Y) = \sigma_Y$.

   Let $W = 5Y - 2X + 4$. Find $E(W)$ and $SD(W)$ in terms of μ_X, μ_Y, σ_X, and σ_Y.

4. A random number generator draws at random with replacement from the 10 digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. Suppose the random number generator is run n times.

   For each i in the range $0 \le i \le 9$, let N_i be the number of times the digit i appears among the n draws.

   The odd digits are 1, 3, 5, 7, and 9. Let N_{odd} be the number of times odd digits appear among the n draws.

   a) If possible, use the identity $N_{odd} = N_1 + N_3 + N_5 + N_7 + N_9$ to find $E(N_{odd})$. If this is not possible, find $E(N_{odd})$ in some other way.

   b) If possible, use the identity $N_{odd} = N_1 + N_3 + N_5 + N_7 + N_9$ to find $Var(N_{odd})$. If this is not possible, find $Var(N_{odd})$ in some other way.

5. Let X be a random variable such that $E(X) = \mu_X$ and $SD(X) = \sigma_X$. Let Y be independent of X, with $E(Y) = \mu_Y$ and $SD(Y) = \sigma_Y$.

   Let $V = (X + 2Y - 3)(5X - Y)$. Find $E(V)$ in terms of μ_X, μ_Y, σ_X, and σ_Y.

   [Hint: First write $E(X^2)$ in terms of μ_X and σ_X.]

6. The average age of a group of 10 students is 19 years. I pick 5 of the students one by one at random without replacement. Let X_i be the age of the ith student picked. If possible, find $E(X_3)$. If this is not possible, explain why not.

7. A die is rolled 120 times. Find the expectation of

   a) the average number of spots on all 120 faces that appear

   b) the number of rolls on which the face with six spots appears

   c) the number of rolls on which the face that appears shows an even number of spots

   d) the proportion of rolls on which the face that appears shows more than four spots

8. The average age of a population is 32 years and the SD is 5 years. One percent of the population is over 80 years old. A random sample of 500 people is drawn at random with replacement from the population. Fill in the blanks below.

   The average age of the people in the sample is expected to be _____ years with an SD of _____ years.

9. A fair coin is tossed 100 times. Let H be the number of heads and T the number of tails. Let $D = H - T$. Find $E(D)$ and $SD(D)$.

10. A large population of incomes has average \$80,000 and SD \$50,000. A simple random sample of n incomes will be taken; you can assume that n is small enough relative to the population size that the probabilities are essentially the same as if the sampling had been carried out with replacement.

    About how large does n have to be so that there is at most a 5% chance that the sample average will be outside the range \$80,000 ± \$1,000? Answer this question by using Chebychev's inequality

Chapter 8

Correlation

Up to this point, we have only dealt with the analysis of single variables; but often, data scientists are interested in the relationship between two or more variables. A classical way of visualizing the relationship between two numerical variables measured on one set of individuals – for example, heights and weights of a group of data scientists – is to plot the data in a scatter plot. This provides a visualization of the relationship between these two variables, as a shape in form of a cloud.

To further quantify the relationship, we will define the *correlation coeffiecient* between the two variables. The correlation, often expressed by the symbol r, will turn out to be a number ranging from -1 to 1, which measures the degree of clustering of the scatter plot around a straight line.

8.1 The Correlation Coefficient

Suppose there are n individuals, on each of Whom two variables $\mathbf{x}$ and $\mathbf{y}$ have been measured, resulting in n points (x_i, y_i) for $1 \le i \le n$.

Definition of r. The correlation coefficient r between $\mathbf{x}$ and $\mathbf{y}$ is the mean of the product of the two variables in standard units:

$$r = r(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sigma_x} \frac{(y_i - \bar{y})}{\sigma_y} = \frac{1}{n} \sum_{i=1}^{n} x_i^* y_i^*$$

where for $1 \le i \le n$, x_i^* is x_i in standard units and y_i^* is y_i in standard units.

In this chapter we will establish the main properties of r and show how it can be used to estimate the value of y given a value of x. But first, let us examine the definition more closely. Notice that

$$r = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

The numerator is called the *covariance* between $\mathbf{x}$ and $\mathbf{y}$. Like variance, covariance has excellent mathematical properties and is useful for deriving properties of estimates. But also like variance, the units of covariance are hard to understand – they are the product of the units of $\mathbf{x}$ and $\mathbf{y}$. For example, covariance might be measured in "inch pounds", which is difficult to interpret.

That is where correlation comes in. In the formula for correlation, we divide the covariance by both the SDs, resulting in a quantity that is a pure number with no units. This number turns out to have clear practical interpretations, which we will develop in this chapter.

$$r = \frac{Cov(\mathbf{x}, \mathbf{y})}{\sigma_x \sigma_y}$$

Just as the calculation for the variance can be written as "the average of squares minus square of average", so also covariance is "the average of the products minus the product of the averages."

$$\begin{aligned}
Cov(\mathbf{x}, \mathbf{y}) &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\
&= \frac{1}{n} \sum_{i=1}^{n} (x_i y_i - x_i \bar{y} - y_i \bar{x} + \bar{x}\bar{y}) \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{y} \frac{1}{n} \sum_{i=1}^{n} x_i - \bar{x} \frac{1}{n} \sum_{i=1}^{n} y_i + \frac{1}{n} \sum_{i=1}^{n} \bar{x}\bar{y} \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y} \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \\
&= \overline{xy} - \bar{x}\bar{y}
\end{aligned}$$

Thus

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y}$$

We'll call this simplified formula the "computational formual for r".

8.2 Linear Transformations

Since r is based on standard units, changing units of measurement doesn't change r. In fact, we will prove a somewhat more general fact about r and linear transformations. For constants a and b where $a \neq 0$, let $\mathbf{z} = a\mathbf{y} + b$ be a linear transformation of $\mathbf{y}$.

Then for $1 \leq i \leq n$, the value z_i in standard units is

$$\begin{aligned}
z_i^* &= \frac{z_i - \bar{z}}{\sigma_z} \\
&= \frac{a y_i + b - (a\bar{y} + b)}{|a| \sigma_y} \\
&= \frac{a}{|a|} \cdot \frac{y_i - \bar{y}}{\sigma_y} \\
&= \frac{a}{|a|} y_i^*
\end{aligned}$$

That's equal to y_i^* if $a > 0$ and equal to $-y_i^*$ if $a < 0$.

Since correlation is the average of products of standard units, this implies that for a linear transformation $\mathbf{z} = a\mathbf{y} + b$,

$$r(\mathbf{x}, \mathbf{z}) = \frac{1}{n} \sum_{i=1}^{n} x_i^* z_i^* = \frac{1}{n} \sum_{i=1}^{n} x_i^* y_i^* = r(\mathbf{x}, \mathbf{y}) \quad \text{if } a > 0$$

and

$$r(\mathbf{x}, \mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} x_i^* z_i^* = \frac{1}{n}\sum_{i=1}^{n} x_i^*(-y_i^*) = -r(\mathbf{x}, \mathbf{y}) \quad \text{if } a < 0$$

Thus the magnitude of correlation is preserved across linear transformations. This result will play a big part in proving some interesting facts about regression estimates.

An important special case is when $\mathbf{y} = \mathbf{x}$. Then the scatter plot is the 45 degree line through the origin. Notice that by the definition of covariance,

$$Cov(\mathbf{x}, \mathbf{x}) = Var(\mathbf{x})$$

and so

$$r(\mathbf{x}, \mathbf{x}) = \frac{Cov(\mathbf{x}, \mathbf{x})}{\sigma_x \sigma_x} = 1$$

Since a linear transformation preserves the magnitude of the correlation between variables,

$$r(\mathbf{x}, \mathbf{a}\mathbf{x} + \mathbf{b}) = \pm 1 \quad \text{depending on whether } a > 0 \text{ or } a < 0$$

Thus if the scatter plot is a straight line, the correlation is +1 if the line is sloping upwards, and −1 if the line is sloping downwards.

But what could the value of r be if the scatter plot is not a straight line?

8.3 Bounds on Correlation

As it turns out, the correlation coefficient is always between −1 and +1. To prove this, recall that correlation is defined in terms of standard units:

$$r(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i^* y_i^*$$

Two important properties of standard units will be familiar from the corresponding properties of random variables converted to standard units:

When a list is converted to standard units, its average is 0 and its SD is 1.

The proof is left as an exercise. Just note that conversion to standard units is a linear transformation, and use the results of Exercise 2 in Chapter 2.

So we know that

$$\frac{1}{n}\sum_{i=1}^{n} x_i^* = 0 = \frac{1}{n}\sum_{i=1}^{n} y_i^*$$

Since variance is the square of the SD, we also know by the definition of variance that

$$\frac{1}{n}\sum_{i=1}^{n} (x_i^* - 0)^2 = 1 = \frac{1}{n}\sum_{i=1}^{n} (y_i^* - 0)^2$$

That is,

$$\frac{1}{n}\sum_{i=1}^{n} (x_i^*)^2 = 1 = \frac{1}{n}\sum_{i=1}^{n} (y_i^*)^2$$

We are now ready to derive the bounds on r. Now

$$r = \frac{1}{n}\sum_{i=1}^{n} x_i^* y_i^*$$

If we could connect this to

$$\frac{1}{n}\sum_{i=1}^{n}(x_i^*)^2 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i^*)^2$$

which we know are both equal to 1, we might be able to make some progress on the bounds.

Which results connect products and squares? The most familiar ones are $(a+b)^2 = a^2 + 2ab + b^2$ and $(a-b)^2 = a^2 - 2ab + b^2$.

$$\begin{aligned}
\forall i: \quad & 0 \le (x_i^* + y_i^*)^2 \\
& 0 \le \sum_{i=0}^{n}(x_i^* + y_i^*)^2 \\
& 0 \le \frac{1}{n}\sum_{i=0}^{n}(x_i^* + y_i^*)^2 \\
& 0 \le \frac{1}{n}\sum_{i=0}^{n}((x_i^*)^2 + 2x_i^* y_i^* + (y_i^*)^2) \\
& 0 \le \frac{1}{n}\sum_{i=0}^{n}(x_i^*)^2 + \frac{1}{n}\sum_{i=0}^{n} 2x_i^* y_i^* + \frac{1}{n}\sum_{i=0}^{n}(y_i^*)^2 \\
& 0 \le 1 + 2\frac{1}{n}\sum_{i=0}^{n} x_i^* y_i^* + 1 \\
& -2 \le 2\frac{1}{n}\sum_{i=0}^{n} x_i^* y_i^* \\
& -1 \le \frac{1}{n}\sum_{i=0}^{n} x_i^* y_i^* \\
& -1 \le r
\end{aligned}$$

$$\begin{aligned}
\forall i: \quad & 0 \le (x_i^* - y_i^*)^2 \\
& 0 \le \sum_{i=0}^{n}(x_i^* - y_i^*)^2 \\
& 0 \le \frac{1}{n}\sum_{i=0}^{n}(x_i^* - y_i^*)^2 \\
& 0 \le \frac{1}{n}\sum_{i=0}^{n}((x_i^*)^2 - 2x_i^* y_i^* + (y_i^*)^2) \\
& 0 \le \frac{1}{n}\sum_{i=0}^{n}(x_i^*)^2 - \frac{1}{n}\sum_{i=0}^{n} 2x_i^* y_i^* + \frac{1}{n}\sum_{i=0}^{n}(y_i^*)^2 \\
& 0 \le 1 - 2\frac{1}{n}\sum_{i=0}^{n} x_i^* y_i^* + 1 \\
& -2 \le -2\frac{1}{n}\sum_{i=0}^{n} x_i^* y_i^* \\
& 1 \ge \frac{1}{n}\sum_{i=0}^{n} x_i^* y_i^* \\
& 1 \ge r
\end{aligned}$$

$$\therefore -1 \le r \le 1$$

Chapter 9

The Regression Line

Suppose we have n individuals on each of whom we have measured two numerical variables. Then our data set will look like $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. And suppose that we would like to find a straight line that best fits these data, according to some reasonable definition of "best". How should we go about doing that?

9.1 Mean Squared Error

First, we have to define what "best" means. For this, we need some measure of the fit of the line. Suppose the line has the equation $y = ax + b$ for some slope a and y-intercept b. To make our line fit our data as closely as possible, we'd like to minimize the difference between the actual values in our data set and their estimates on the line. So we want some way of capturing the overall size of the errors

$$e_i \;=\; y_i - (ax_i + b), \quad i = 1, 2, \ldots, n$$

For many lines, some of the errors will be positive and others negative. To avoid cancellation, we won't just take the average of the errors; we'll square the errors first, and then take the average. That quantity is called the *mean squared error* of the line. Let's denote it by $mse_{a,b}$. Then

$$mse_{a,b} \;=\; \frac{1}{n}\sum_{i=1}^{n} e_i^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

Our choice of a and b should minimize this mean squared error over all possible a and b.

Minimizing the Mean Squared Error

The process of minimization is the same as that in seen in any first semester calculus class, though it may be seem more intimidating because of all the sums and variables.

First, we will fix the slope a and just think of $mse_{a,b}$ as a function of the y-intercept b. Once we've obtained an optimal b in terms of a, we'll plug that value back into our definition of mean squared error, and minimize it with respect to a. That will give us the minimizing values of both a and b.

9.2 The Best Intercept for a Fixed Slope

Fix a, and let $mse_a(b)$ be the mean squared error of the line with the fixed slope a and b as its intercept. We will try to find the best b for the given a.

Notice that in the formula for $mse_a(b)$, *everything except b is constant*. The only variable is b.

$$mse_a(b) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - (ax_i + b)\Big)^2$$

Our goal is to minimize this over all b. Since $1/n$ is a constant, that's the same as minimizing the *sum of squared errors*:

$$sse_a(b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\Big(y_i - (ax_i + b)\Big)^2$$

That's a quadratic function of b. To minimize it with respect to b, we'll take the derviative with respect to b and set it equal to 0. Remeber the chain rule!

$$\begin{aligned}
\frac{d}{db} mse_a(b) &= \sum_{i=1}^{n} -2\Big(y_i - (ax_i + b)\Big) \\
&= -2\sum_{i=1}^{n}\Big(y_i - (ax_i + b)\Big) \\
&= -2\Big(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} b\Big) \\
&= -2\Big(n\bar{y} - an\bar{x} - nb\Big)
\end{aligned}$$

Set this equal to 0 and solve for the mininizing value of b; we'll call the minimizing value b_a to remind us that it depends on a which is fixed for now.

$$0 = -2\Big(n\bar{y} - an\bar{x} - nb_a) \quad \text{and so} \quad b_a = \bar{y} - a\bar{x}$$

Thus among all lines with a given slope a, the optimal one has a y-intercept of $b_a = \bar{y} - a\bar{x}$.

9.3 The Best Slope

Now, to find the slope of the regression line that minimizes the mean squared error, we will substitute our value for b_a into our original equation and let a vary. Thus the minimization becomes

$$\begin{aligned} sse(a) &= \sum_{i=1}^{n} \Big(y_i - (ax_i + b_a)\Big)^2 \\ &= \sum_{i=1}^{n} \Big(y_i - (ax_i + \bar{y} - a\bar{x})\Big)^2 \\ &= \sum_{i=1}^{n} \Big((y_i - \bar{y}) - a(x_i - \bar{x})\Big)^2 \\ &= \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2a \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) \Big) + a^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= n\sigma_y^2 - 2aCov(\mathbf{x}, \mathbf{y}) + a^2\sigma_x^2 \end{aligned}$$

where in our usual notation, σ_x^2 is the variance of the x-variable, σ_y^2 is the variance of the y-variable, and $Cov(\mathbf{x}, \mathbf{y})$ is the covariance between the two variables.

Our job is to find the value of a that minimizes this function. It's a quadratic function of a, so differentiation is straightforward.

$$\frac{d}{da} sse(a) = -2Cov(\mathbf{x}, \mathbf{y}) + 2\sigma_x^2 a$$

For the minimizing value of a, the derivative is 0. Let a_{best} be that minimizing value. Then

$$0 = -2Cov(\mathbf{x}, \mathbf{y}) + 2\sigma_x^2 a_{best} \quad \text{and so} \quad a_{best} = \frac{Cov(\mathbf{x}, \mathbf{y})}{\sigma_x^2}$$

Remember that

$$Cov(\mathbf{x}, \mathbf{y}) = r\sigma_x\sigma_y$$

where r is the correlation between the two variables. Plug this into the formula for a_{best}:

$$a_{best} = \frac{r\sigma_x\sigma_y}{\sigma_x^2} = \frac{r\sigma_y}{\sigma_x}$$

The Equation of the Regression Line

We have found the slope and intercept of the line that minimizes the mean squared error among all lines. This "best" line has many names, including "least squares line" and "regression line."

Our calculation has proved that

$$\text{slope of the regression line} = \frac{r\sigma_y}{\sigma_x}$$

$$\text{intercept of the regression line} = \bar{y} - \text{slope} \cdot \bar{x}$$

Notes

- It should be clear that the calculations have led to a minimum, not a maximum. T You can make the mean squared error as large as you want by choosing truly horrible lines, such as lines that never get near the scatter plot. You can also check by finding the second derivative that the optimal values above result in a minimum.

- The minimizing line is unique: there is only one minimizing slope and one minimizing intercept.
- The calculations made no assumption about the shape of the scatter diagram other than $\sigma_x > 0$. We divided by σ_x^2 to get the best slope. If $\sigma_x = 0$ then the x-variable is a constant and there's no linearity to think about. But other than that case, no matter how ugly or complicated the scatter plot happens to be, there is a unique line that is the best among all lines in the sense of minimizing mean squared error, and its equation is given above.

9.4 Fitted Values

Let $a^* = r\sigma_y/\sigma_x$ be the slope of the regression line and $b^* = \bar{y} - a^*\bar{x}$ the intercept.

Then for each $i = 1, 2, \ldots, n$, the ith **fitted value** is the predicted value of the y-variable given that the value of the x-variable is x_i. That is, the ith fitted value is the height of the regression line at x_i.

Denote the fitted values by $\hat{y}$. Then the ith fitted value is

$$\begin{aligned} \hat{y}_i &= a^* x_i + b^* \\ &= a^* x_i + \bar{y} - a^* \bar{x} \\ &= \bar{y} + a^*(x_i - \bar{x}) \end{aligned}$$

Notice that the fitted values are a linear transformation of the deviations of x. You can use this observation to quickly work out the mean and SD of the fitted values.

It will help to recall that the average of the deviations is 0, and also that the SD of the deviations is equal to the SD of x, because the deviations are obtained by shifting all the values of x to the left by the constant amount $\bar{x}$.

Average of the Fitted Values

The average of the deviations of x is 0. So the average of the fitted values is $\bar{y}$, the average of the observed values of y.

Thus on average, the regression predictions are neither too high nor too low. Of course, the predicted values vary because the values of x vary. The SD of the fitted values measures the variability of our predictions.

SD of the Fitted Values

Once again, use the observation that the fitted values are a linear transformation of the deviations of x, and remember how variance behaves under linear transformations.

The variance of the fitted values $\hat{y}$ is

$$\begin{aligned} \sigma_{\hat{y}}^2 &= a^{*2}\sigma_x^2 \\ &= r^2 \frac{\sigma_y^2}{\sigma_x^2}\sigma_x^2 \\ &= r^2\sigma_y^2 \end{aligned}$$

The SD of the fitted values is

$$\sigma_{\hat{y}} = |r| \cdot \sigma_y$$

r^2 as a Ratio

Thus r^2 is the ratio of the variance of the fitted values and the variance of the observed values of y:

$$r^2 = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2} \qquad \text{and} \qquad |r| = \frac{\sigma_{\hat{y}}}{\sigma_y}$$

Unless the scatter diagram is a perfect straight line, these ratios are less than 1. This is consistent with the observation that the fitted values are all on the regression line and hence less variable than the observed values.

Multiple r^2

In more complex analyses you might encounter **multiple regression**, in which a linear combination of several different x-variables is used to predict y. It turns out that in that case also, the variance of the fitted values can be no larger than the variance of the observed values of y. One measure of the strength of the linear relation between y and all the x's is the *multiple* r^2 defined by

$$\text{multiple } r^2 = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2}$$

This is a number between 0 and 1. The larger it is, the stronger the predictive power of the regression.

Chapter 10

Residuals

The ith **residual** is the error in the ith fitted value. That is, the ith residual is the difference between the ith observed value of y and the corresponding fitted value.

Thus the ith residual e_i is given by

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - (\bar{y} + a^*(x_i - \bar{x})) \\ &= (y_i - \bar{y}) - a^*(x_i - \bar{x}) \end{aligned}$$

Notice how the residual only depends on the slope and the ith deviations of x and y. This will help us calculate the mean and SD of the residuals.

10.1 The Rough Size of the Residuals

For a prediction to be useful, you have to be able to quantify roughly how good it is. So let's see roughly how big the residuals are. That will tell us roughly how good our regression predictions will be.

Average of the Residuals

The deviations of x and y both average out to 0, so the average of the residuals is also 0:

$$\bar{e} = 0$$

SD of the Residuals

The variance of the residuals is

$$\begin{aligned}
\sigma_e^2 &= \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} e_i^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 - 2a^*\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + a^{*2}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sigma_y^2 - 2a^* Cov(\mathbf{x}, \mathbf{y}) + a^{*2}\sigma_x^2 \\
&= \sigma_y^2 - 2(r\frac{\sigma_y}{\sigma_x})(r\sigma_x\sigma_y) + r^2\frac{\sigma_y^2}{\sigma_x^2}\sigma_x^2 \\
&= \sigma_y^2 - 2r^2\sigma_y^2 + r^2\sigma_y^2 \\
&= (1 - r^2)\sigma_y^2
\end{aligned}$$

The SD of the residuals is

$$\sigma_e = \sqrt{1 - r^2} \cdot \sigma_y$$

In summary, the residuals average out to 0 and have an SD that is a fraction of the SD of y. When r is large (either positive or negative), that fraction $\sqrt{1 - r^2}$ is small, and so the residuals tend to be small. This is consistent with the notion that a large value of r corresponds to the scatter plot being tightly clustered about a line.

10.2 A Variance Decomposition

The variance of the observed values of y splits neatly into two recognizable pieces.

$$\begin{aligned}
\sigma_y^2 &= r^2\sigma_y^2 + (1 - r^2)\sigma_y^2 \\
&= \sigma_{\hat{y}}^2 + \sigma_e^2
\end{aligned}$$

In words, **the variance of y is equal to the variance of the fitted values plus the variance of the residuals**.

Which of the two terms $\sigma_{\hat{y}}^2$ and σ_e^2 is bigger depends on r^2.

Large r^2

If r^2 is large, then the variance of y is dominated by the variance of the fitted values. This is consistent with our notion of high r^2 corresponding to the scatter diagram being tightly clustered about a line. The variance of he residuals is correspondingly small. If the scatter is tightly clustered about a line, then using regression to predict y is a good idea.

Small r^2

If r^2 is small, then the variance of y is dominated by the variance of the residuals. Small r^2 implies large errors in regression. The variance of the fitted values is small, meaning that the points on the line are at pretty close to the same vertical level. In other words, the regression line is pretty flat, and we are not gaining much by using regression to predict y based on x.

The decomposition

$$\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_e^2$$

holds true in the case of multiple regression as well. As we have seen earlier, the fraction $\sigma_{\hat{y}}^2/\sigma_y^2$ is used as the defintion of multiple r^2.

10.3 A Residual Plot

Plotting the residuals can be used as a *regression diagnostic*, that is, a way to assess whether a regression is good. One way to create such a plot is to make a scatter plot of the residuals (on the vertical axis) against x (on the horizontal axis).

The residuals average out to 0, so this residual plot will be centered at the horizontal line at level 0, though it might have a variety of shapes. If the shape is a formless blob, the regression is good. If the shape shows a pattern, fitting a straight line might not have been a good idea.

It is important to note that while a residual plot might show a pattern, it cannot show a trend, either upwards or downwards. That is because no matter what the shape of the scatter plot, **the residuals and x are uncorrelated**.

Correlation between e and x

As we have done in previous sections, we will use the notation $r(\mathbf{v}, \mathbf{w})$ to denote the correlation between variables v and w.

We know already that $r(\mathbf{x}, \mathbf{y}) = r$, a number between -1 and 1. We will now show that $r(\mathbf{x}, \mathbf{e}) = 0$. Recall that

$$r(\mathbf{x}, \mathbf{e}) = \frac{Cov(\mathbf{x}, \mathbf{e})}{\sigma_x \sigma_e}$$

So it is enough to show that $Cov(\mathbf{x}, \mathbf{e}) = 0$. To do this we will use our earlier observation that the ith residual e_i can be found using the slope a^* of the regression line and the ith deviations of y and x, as follows:

$$e_i = (y_i - \bar{y}) - a^*(x_i - \bar{x})$$

Now

$$\begin{aligned} Cov(\mathbf{x},\mathbf{e}) &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(e_i-\bar{e}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})e_i \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) - a^*\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ &= Cov(\mathbf{x},\mathbf{y}) - r\frac{\sigma_y}{\sigma_x}\sigma_x^2 \\ &= Cov(\mathbf{x},\mathbf{y}) - r\sigma_x\sigma_y \\ &= Cov(\mathbf{x},\mathbf{y}) - Cov(\mathbf{x},\mathbf{y}) \\ &= 0 \end{aligned}$$

Another Residual Plot

The fitted values $\hat{y}$ are a linear transformation of x. As correlation is preserved (apart from sign) under linear transformations, **the residuals and the fitted values are uncorrelated**. That is,

$$r(\hat{\mathbf{y}},\mathbf{e}) = r(\mathbf{x},\mathbf{e}) = 0$$

So residual plots are often drawn with the residuals on the vertical axis and the fitted values on the horizontal axis. The main reason is that this way of drawing the residual plot extends to multiple regression. Even when we use a linear combination of several x-variables to predict y, the fitted values and residuals are uncorrelated, and the plot of those two variables can be used as a regression diagnostic just as in simple regression.

10.4 Some Questions for You

It's time to end this text. We leave you with some questions to think about.

- What is the value of $r(\mathbf{x},\hat{\mathbf{y}})$?
- Suppose $r \neq 0$. Can you explain, without calculation, why $r(\mathbf{y},\hat{\mathbf{y}})$ must be positive, regardless of whether r is positive or negative?
- Can you find the value of $r(\mathbf{y},\hat{\mathbf{y}})$? Remember that $\hat{y}$ is a linear function of x, and be careful about the case when r is negative.
- Can you find the value of $r(\mathbf{y},\mathbf{e})$? Follow the process we used to get $r(\mathbf{x},\mathbf{e})$.
- Can you explain why the value you got for $r(\mathbf{y},\mathbf{e})$ is consistent with our intuitive notions of what it means for r^2 to be near 0 or near 1?

Chapter 11

Appendix

11.1 Summation Notation

Expressing sums can be a lot of work, especially when you have a lot of terms. For example, the sum of all the numbers from 1 to 100 takes 100 terms. We need a way to express this sum in a much shorter way. For this, we use sigma notation:

$$1 + 2 + ... + 99 + 100 = \sum_{i=1}^{100} i$$

Definition 11 ***Sigma Notation***
Sigma notation allows us to express sums that are either finite or infinite. The general form of a finite summation is as follows:

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + ... + a_{n-1} + a_n$$

The above statement is read: "The sum of the 1st term to the n*th term of the series* a_n*."*

Breaking down the notation, we start off with an index. The *i=1* term specifies our first **index**, which determines the starting value of the iteration.

$$\sum_{i=a}^{n} a_i = a_a + a_{a+1} + ... + a_{n-1} + a_n$$

We next want to consider the ending value, which is represented in previous examples by the n above the sigma symbol. This value determines what the last term will be. In prior examples, the n means that the last term will be the nth term in the sequence.

We can consider other examples to see how changing either the bottom or top index of the summation can change the expression.

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + ... + a_{n-1} + a_n$$

$$\sum_{i=100}^{n} a_i = a_{100} + a_{101} + ... + a_{n-1} + a_n$$

$$\sum_{i=100}^{200} a_i = a_{100} + a_{101} + ... + a_{199} + a_{200}$$

We now will look at the last component, which is the **body** of the sigma. In the previous examples, the body has been the series a_n. Now, we can replace that with other expressions. The following are examples of what happens when you replace the body with other expressions:

$$\sum_{i=1}^{n} i = 1 + 2 + ... + (n-1) + n$$
$$\sum_{i=a}^{n} i^2 = 1 + 4 + ... + (n-1)^2 + n^2$$

We can also put in constant values:

$$\sum_{i=1}^{n} 3 = 3 + 3 + ... + 3 + 3 = 3n$$

Notice that there are n 3's in the above summation, which is why we can simplify the sigma expression to $3n$.

Manipulating Summations

There are a couple of ways we can manipulate summations, to simplify complicated expressions.

1. Splitting the body of a sum
$$\sum (a_i + b_i) = \sum a_i + \sum b_i$$
2. Moving constants through the summation symbol
$$c \sum a_i = \sum c * a_i$$
3. Splitting the sum by the index
$$\sum_{i=0}^{n} a_i = \sum_{i=0}^{j} a_i + \sum_{i=j+1}^{n} a_i$$
$$\sum_{i \in (A \cup B)} a_i = \sum_{i \in A} a_i + \sum_{i \in B - A} a_i$$

Common Summations

1.
$$\sum_{i=1}^{n} 1 = n$$
2.
$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$
3.
$$\sum_{i=1}^{n} 0 = 0$$