Theory Meets Data

A Data Scientist's Handbook to Statistics

Editors: Ani Adhikari and Dibya Jyoti Ghosh

Authors

Ani Adhikari, 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.

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

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.

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 $1 \leq i \leq n$.

   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

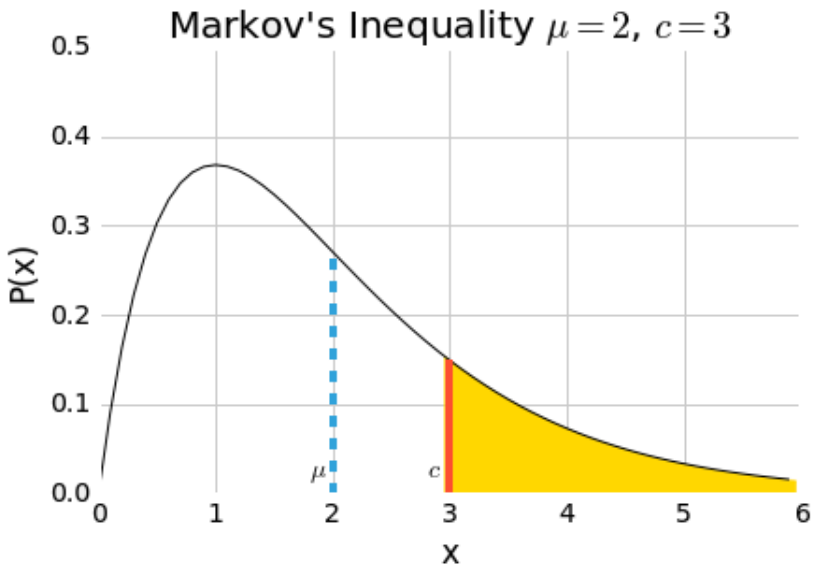
Suppose that a list of non-negative numbers $x_1, x_2, \ldots, x_n$ has average $\bar{x}$. **Markov's Inequality** gives an upper bound on the proportion of entries that are greater than some positive integer c:

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 5 ***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 6 ***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 7 ***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 8 ***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 9 ***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 $= \underset{\nearrow}{\frac{5}{6}} \times \underset{\uparrow}{\frac{5}{6}} \times \underset{\nwarrow}{\frac{5}{6}} = (\frac{5}{6})^3$

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

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:

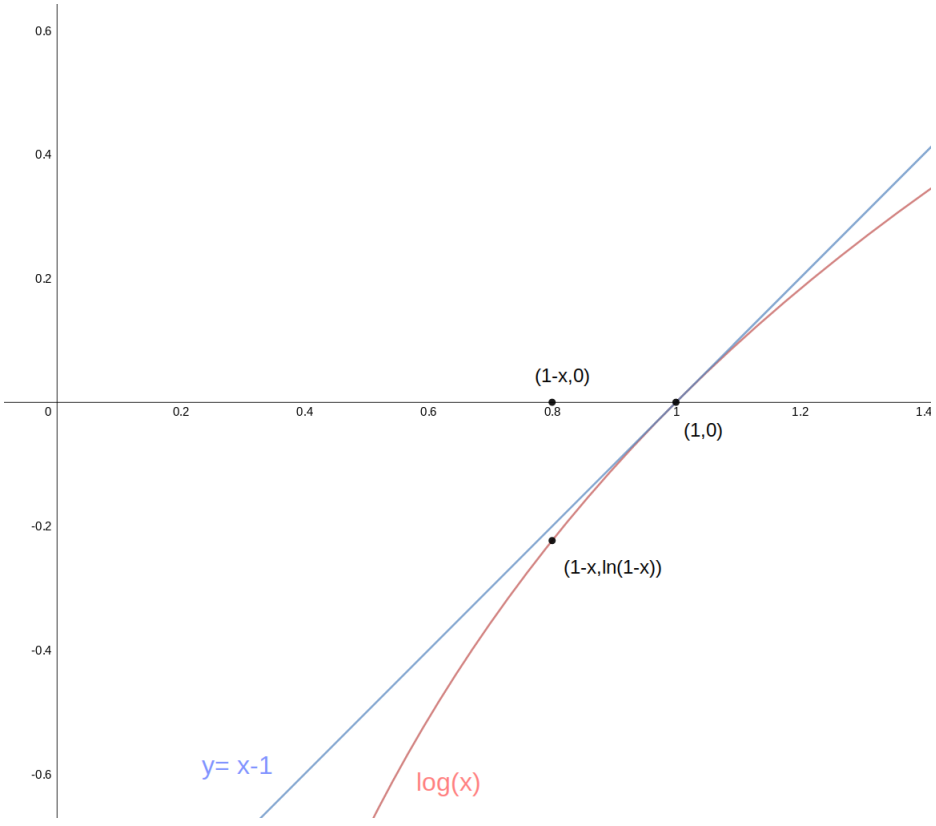
$$n \geq \frac{\log \frac{1}{2}}{\log(1 - \frac{1}{N})}$$

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

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

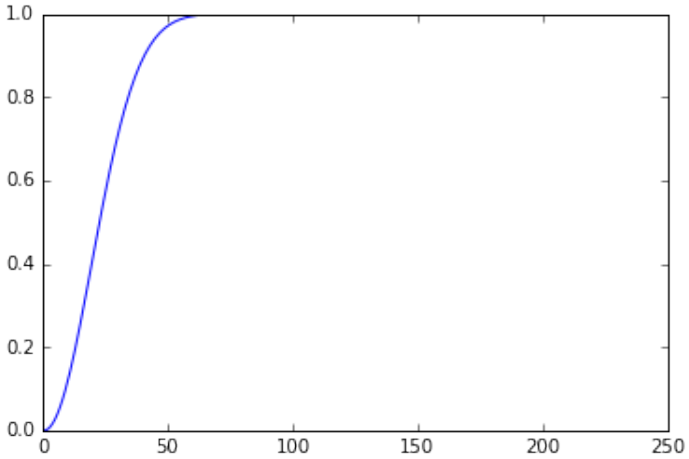
When there are three people, each of the three birthdays has to be unique, and so the chance that all three have different birthdays is

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

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

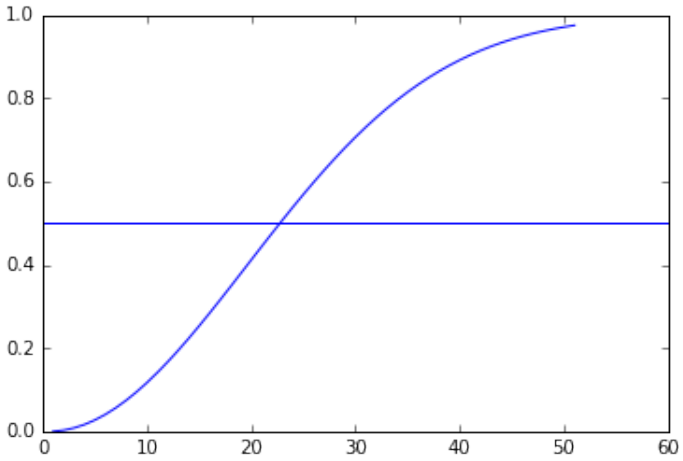
Birthday Problem - Probability Table		
Class Size (n)	Chance that all birthdays are different	Chance that at least 2 or more people in the class have the same birthday
1	0	1
2	$\frac{365}{365} \times \frac{364}{365}$	1 - ($\frac{365}{365} \times \frac{364}{365}$)
3	$\frac{365}{365} \times \frac{364}{365}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

6. 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 $0 \leq k \leq n$. 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?

7. 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?

Chapter 5

Appendix

5.1 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 10 ***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.
$$\sum_{i=1}^{n} 0 = 0$$