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* 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*

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

"Mean" is another name for "average", so other sources may use μ to represent the average. That's the Greek letter m, read as "mu".

In other words, we add the values of all n x_i's and then divide that total into n even pieces. We take n unequal numbers, put them all in one big pot, and then split them into n equal pieces. Thus, we can consider the mean to be an "equalizer".

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, 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 $\mathbf{x}$: $(x_1, x_2, \dots, x_n)$ and our second dataset $\mathbf{y}$: $(k, x_2, \dots, x_n)$ 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 $x_1, x_2, \dots, x_n$. 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 $x_1, \dots, x_n$ 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 $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots y_m$, 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 $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots y_m$. 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 $x_1, x_2 \dots x_n, y_1, y_2, \dots y_m$ by A. Then we see

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

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

$$\frac{3}{5} \cdot 7 \; + \; \frac{2}{5} \cdot 8 \qquad (1.5)$$

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 $x_1, x_2, \dots x_n$ with lots of repeating values. Say there are k distinct values $v_1, v_2, \dots v_k$ in our list, appearing respectively with frequencies $n_1, n_2, \dots n_k$. 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 $1 \le i \le n$. 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 $\{x_1, x_2, \dots, x_n\}$ 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 $x_1, x_2, \dots, x_n$ 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 $\{x_1, x_2, \dots, x_n\}$ has average $\bar{x}$ and the list $\{y_1, y_2, \dots, y_m\}$ has average $\bar{y}$. Consider the combined list of $n + m$ entries $\{x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m\}$. 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 $\{x_1, x_2, \dots, x_n\}$ 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 $\{y_1, y_2, \dots, y_n\}$, 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 $\{2, 4, 6, \dots, 2n\}$. What is the average of this list? Prove your answer.

10. Let $\{x_1, x_2, \dots, x_n\}$ 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?

The word *deviation* naturally suggests a notion of distance. For example, how could we calculate the “distance” between two numbers A and B? Usually, we use the unsigned difference, e.g $|A-B|$, to denote the “distance” between any of two numbers. When we measure distance between points in space, we use the Euclidean distance ($\sqrt{(a_1-b_1)^2+\cdots+(a_n-b_n)^2}$). Similarly, we’d like to develop some measure of the distance inside a dataset as the standard deviation.

It might make sense to call the standard deviation the average of all the deviations from the mean. That is, we sum up all the "distances" first and then divide the sum by the number of terms. Let’s try it out on a sample dataset $x_1 \ldots x_n$ with mean $\bar{x}$

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

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.Therefore, we’ll have to find a way past this cancellation error.

To avoid cancellation, we have to ensure that all "distances" are non-negative, so we square all "distances" before taking the average. (Notice that the absolute value has the same property, but we prefer squaring numbers since it reduces complexity when solving extensive equations). When squaring the distances, we’ve artificially inflated the distances, and so to return to our "standard" distance, we’ll need to take the square root (when working with measurements, this can be thought of "fixing" the unit).

Definition 2 ***Standard Deviation***

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

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

This definition can look rather complicated, and so it's often better to recall the steps that we took to arrive at this formula. First, we took the *distances* to the mean, and *squared* them. Then, we found the *average* squared distance, and *normalized* it (took the square root) to revert to the correct units. Thus, the standard deviation is simply the **normalized average squared distance from the mean**

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 ($\bar{x}$)

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

2. Calculate the Average Squared Distance $(x_i - \bar{x})^2$

   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. Normalize

$$\sigma = \sqrt{\frac{1228}{18}}$$
$$= 8.260$$

2.2 Variance

Variance is defined as the *average of squared differences from mean*. It is calculated the same way as is the standard deviation but without taking the square root. In probability, the variance is a much more useful measure than the standard deviation, because it is much easier to manipulate (as you'll see in future sections).

Definition 3 ***Variance***

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.1}$$

$$Var(X) = \sigma^2 \tag{2.2}$$

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

As a result, the variance will soon become your best friend. However, to calculate 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

Recall the formal definition of variance

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

First, expand the square (Reminder:$(a-b)^2 = a^2 - 2ab + b^2$)

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

Then, remove the parenthesis by multiplying items in and out of the parenthesis respectively.

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

Next, take constants($\bar{x}$ and 2) out of $\sum_{i=1}^{n}$. The average of squared averages equals to the squared average($\frac{1}{n}\sum_{i=1}^{n}\bar{x}^2 = \bar{x}^2$).

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

Simplify the expression and combine like terms

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

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

Finally, we arrive at the computational formula of variance in terms of $\bar{x}^2$ and $\sum_{i=1}^{n} {x_i}^2$.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \bar{x}^2 \tag{2.3}$$

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

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \bar{x}^2 \tag{2.4}$$

n = the number of values x_i = each value in the set $\bar{x}$ = the mean of the values

If given any two of $\bar{x}$, σ^2, and $\sum_{i=1}^{n} {x_i}^2$; we can always figure out the third one. Later we will find this formula also useful in mathematical proof.

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

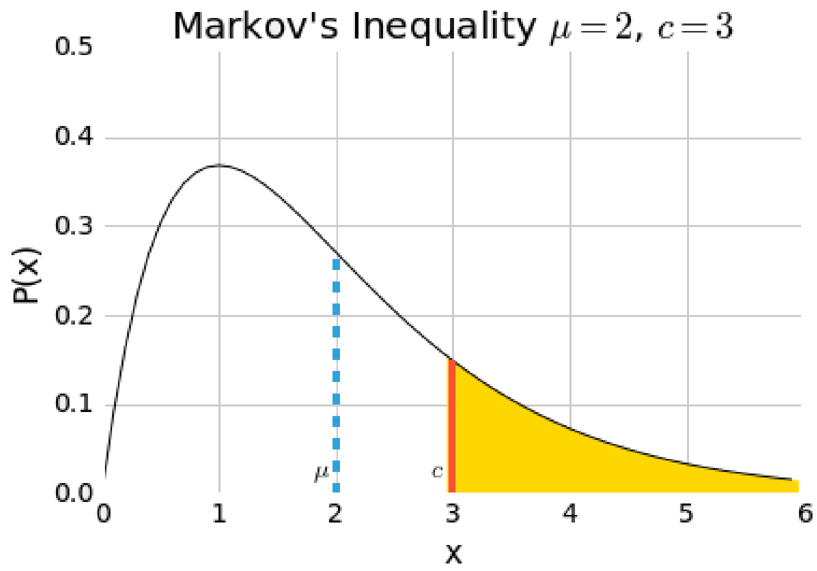
Suppose that a list of non-negative numbers $x_1, x_2, \ldots, x_n$ has average $\bar{x}$. **Markov's Inequality** gives an upper bound on the proportion of entries that are greater than some positive integer c:

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

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

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:

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

$$\tag{3.3}$$

3.2 Chebyshev's Inequality

Markov's inequality gave us a way to bound the tails of a nonnegative probability distribution, using only the mean. Can we tighten our bound any more if we know the standard deviation?

The Weatherman

Consider a weatherman in Northern Alaska interested in examining temperatures, where temperatures are cold and stay that way: after some investigation, we know that the average temperature $\bar{t}$ is -25 C, and that the standard deviation of the temperatures σ_w 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 which are further away, since a small standard deviation indicates closeness. As we saw with Markov, there's no way without looking at the numbers to calculate the exact probability, but we can bound the probability that the temperature is between -15C and -35 C.

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

Chebyshev's Bound

Chebyshev's inequality states that the proportion of items which are at least k standard deviations away from the mean is at most $\frac{1}{k^2}$. For 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 $\frac{1}{2^2} = \frac{1}{4}$.

There are a couple of interesting things to note about this example: firstly, Chebyshev's inequality works with all numbers, not just nonnegative numbers like Markov's inequality. Secondly, as we expand the range (increase k), the proportion scales by k^2, which means that the strength of our bound gets much stronger, the further out we go.

Definition 6 ***Chebyshev's Inequality***

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

σ = the standard deviation
$\bar{X}$ = the mean of the values

Notice that Chebyshev's inequality also requires the standard deviation to calculate, showing that in order to get a better bound, you need more information. This is a pattern that we'll see throughout this book. Notice that the left side of the inequality contains the term $\frac{X-\bar{X}}{\sigma}$, which

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

is often referred to as the z-score (or normalized version) of X. You can think of this term as the "number of standard deviations" is away from the mean.

Proof:

How do we find the proportion of observations in the tails of a distribution? Markov's inequality only applies to a set of non-negative numbers, so its application would require transforming all the numbers in the list x to non-negative values. Recall, by Markov:

$$\frac{\bar{x}}{c} \geq \text{proportion(i: } X_i \geq \text{c)}$$

Notice here, how the bound on the proportion greater than c is the average of the list x, divided by the bound c.

Say that we have a distribution of negative and positive numbers; we want to write out our question in a similar syntax so it can be answered in the same way. We want to know the bound on the proportion in our tails; which can be defined as entries outside k standard deviation away from a mean. In this case, we want to know the proportion of observations:

$$\text{proportion(i: } x_i \text{ is outside } \bar{x} \pm \text{ k}\sigma)$$

We can re-express the list as the squared difference between x_i and $\bar{x}$; transforming the left hand side by subtracting the mean and squaring, we obtain:

$$\text{proportion(i: } (x_i - \bar{x})^2 \geq k^2\sigma^2)$$

To apply Markov we express all x_i as a distance from some mean, squared, and we are looking through, one by one, to pick those entries that exceed the tail boundaries, or $k^2\sigma^2$.

Now, in this case, the proportion will not exceed the mean of the list, divided by the bound. It's easy to see that the mean of our list is by definition, the variance, and our bound is the variance multiplied by some squared scalar, k^2. By cancellation of these variances:

$$\frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2} \geq \text{proportion(i: } (X_i - \bar{X})^2 \geq k^2\sigma^2)$$

We were able to answer a much different question than Markov's inequality intends to answer for us, simply by reforming our question to fit Markov's environment. By using the basic truth of Markov's proof, we were able to see why the proportion of a suitably transformed set of numbers, answers Chebyshev's bounded proportion in the tails.

Applying this, say that we have any list and want to know how many observations are three standard deviations away from the mean (k=3). Chebyshev tells us that the bound on that proportion is 1/9, for any list.

That is the power of Chebyshev.

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

Appendix

4.1 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 7 ***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:

$$\sum_{i=1}^{n} \mathrm{a}_i = \mathrm{a}_1 + \mathrm{a}_2 + ... + \mathrm{a}_{n-1} + \mathrm{a}_n$$

The above statement is read: "The sum of the 1st term to the nth *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. $$\sum_{i=1}^{n} n = \frac{n(n+1)}{2}$$
3. $$\sum_{i=1}^{n} 0 = 0$$