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

Part I

Probability

Chapter 1

An Introduction to 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, whether the chance of rain on the weather app, batting averages for our local baseball teams, or even the success rate of a medical treatment.

Through the language of statistics, one 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 these 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. There will be important methods of application and cautionary tips.

Probability

We can use probability to measure the likelihood of an event occurring. We define the probability of an event as a proportion of the number of favorable outcomes and the total number of outcomes. This is possible only under the assumption that all outcomes are equally likely. (The word favorable for statisticians refers to the event you are studying, and is not necessarily a "good" event.)
Because the number of favorable outcomes is always less than or equal to the total number of outcomes and it is never negative, probability is always between 0 (probability of an impossible event) and 1 (probability of a certain event).

Definition 1 ***Probability***

$$P(A) = \frac{Number\ of\ Favorable\ Outcomes}{Total\ Number\ of\ Outcomes}$$

Example 0. You are drawing items out of a box. In the box is one green tennis ball, one orange tennis ball, and two white golf balls. You are equally likely to pick any of the balls at every draw. 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 tennis 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, of the four possible outcomes. 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, the event of 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, the event of picking a red golf ball is an impossible event. Thus, the probability = 0

Probability of Events Occuring on a Single Trial

Let's say you toss a coin: what's the probability that you get both heads and tails on the same toss? As you very well know, it is not possible for the result of a coin toss to be both heads and tails. These events are called mutually exclusive events. Two or more events are said to be mutually exclusive if they cannot occur at the same time. For mutually exclusive events, the probability of either occurring is the sum of the probabilities of each occurring. The sum of the probabilities of all inclusive events is 1. For example, when you toss a coin, the sum of the probability of getting a head and the probability of getting a tail is 1.

Definition 2 ***Laws of Probability***

$$\sum_{A \in \omega} P(A) = 1 \tag{1.1}$$

$$P(A \text{ or } B) = P(A) + P(B) \tag{1.2}$$

Probabilities of Events Occuring on Multiple Trials

Example 1. What's the probability that you get 2 heads when you flip two coins?
Solution

The process of calculating the probability of, for example, two events requires counting the favorable outcomes divided by the total outcomes. Observe in Figure 1.1 that the favorable outcome of getting two heads is 1 possibility out of 4 equally likely possibilities. When probabilities become more complex, we can use simple computations instead. Note that these computations are representations of the possible probabilities used in the table.

$$P(2\ Heads) = P(Heads\ on\ Coin\ 1) * P(Heads\ on\ Coin\ 2)$$
$$P(2\ Heads) = \frac{1}{2} * \frac{1}{2}$$
$$P(2\ Heads) = \frac{1}{4}$$

	Heads	Tails
Heads	1	1
Tails	1	1

Example 2. What's the probability that you get one head and one tail when you flip two coins?
Solution

The solution to this question is very similar to the solution to Example 1, despite the question asking for a different outcome. Refer back to Figure 1.1. One head and one tail is one possibility of four. The calculation here, (we will be using calculations in the textbook from here on out) is 1/2 out of 1/2, which is equivalent to multiplying 1/2 and 1/2.

Replacement

When something is done *with* replacement, that means the probability of getting a certain outcome remains unchanged despite having performed the event already. A common example of this phenomenon, which also appears later in the section, is a fair 6-sided die. Even if the die is rolled once and the outcome is a 5, the next roll can still yield a 5 with an equal probability.

Polling with Replacement

Polling with replacement means that every time when an event happens, it does NOT affect the probability of other independent events. Therefore, the probability of a particular independent event is fixed regardless of previous occurrence of any event. This implies that the same event could possibly occur more than once in one setting.

Example 2. Rolling A Die A fair 6-sided die has numbers from 1 to 6. Each time when a die is rolled, the outcome will be a number from 1 to 6. The probability of getting any of the 6 numbers is the same, which is 1/6. Each roll is independent from other rolls because the previous roll does not affect the outcome of the next roll. Let the probability of getting a 1 be P(A), (P(A) = 1/6) and the probability of getting a 2 be P(B), (P(B) = 1/6).

(i) What is the probability of rolling a 1 and a 2 on the same roll?
(ii) What is the probability of rolling a 1 or a 2 on the same roll?
(iii) 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 impossible since the outcome could only be a single number, therefore P(A and B) = 0.

(ii) The chance of getting either 1 or 2 on the same roll is the sum of P(A) and P(B) because it includes the possibilities of two outcomes out of 6 outcomes in total, which is 1/6 + 1/6 = 2/6. Therefore, P(A or B) = P(A) + P(B).

(iii) The chance of getting 1 on the first roll(event A) then getting a 2 on the second roll(event B) equals to (1/6) × (1/6), because in order for event B to occur after event A, there is only one possible way, therefore we multiply P(A) and P(B). The reason why P(B) stays the same after Event A has occurred is that every roll is a poll with replacement–in this case, it can be interpreted as a die has 6 sides or 6 outcomes on every roll regardless of previous results.

Based on the observations we just made in the previous example, we can establish a set of formulas for dealing with the probabilities of two independent events, A and B.

Definition 3 *The probability of event B occurring given the fact that Event A has already occurred is equal to the product of the probabilities of the individual events.*

Example 3. 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, 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 in subsequent rolls.
The probability that 1 does not appear in any of 3 rolls $= \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} = (\frac{5}{6})^3$

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

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

Solution Chance that either $2, 4$ or 6 appear in a single roll $= \frac{3}{6}$

Since we are rolling a die, the chance that either $2, 4$ or 6 appear 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 5. 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 equally favorable. In the second roll, the face should be anything but first roll and thus, it can be any of 5 different faces.

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

Example 6. There are 20 students in a class. A computer program selects a random sample of student by drawing 5 students at random with replacement. This implies that every student has the same chance to be selected and may be picked on more than one draws. What is the chance that a particular student is among the 5 selected students?

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

Since we are sampling with replacement, probability that a particular student is not selected in a single draw $= (\frac{20-1}{20}) = \frac{19}{20}$
Probability that a particular student is not selected in all 5 draws (which is the entire sample) $= (\frac{19}{20})^5$

Probability of a particular student getting selected in the sample = 1 - Probability that a particular 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 $= 1 - (1 - \frac{1}{N})^n$

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.You might find it strange that, in a room of only perhaps 30-40 people, you share a birthday with someone else(after all there are 365 possible birthdays), but statistics can prove otherwise. 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 a fair chance that two 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' i.e. Each child's birthday is equally likely to be on any of the days regardless of others' birthday. For instance, there are no twins in the room.

Calculating the Probability

The first step in figuring out our problem, is to find the probability of two people sharing a birthday when there are N people.

Let P(A) be the probability that at least two children in the room have the same birthday. In order to compute P(A), it is often easier to compute the complement, that is, P(A') the probability that they do not have the same birthday.

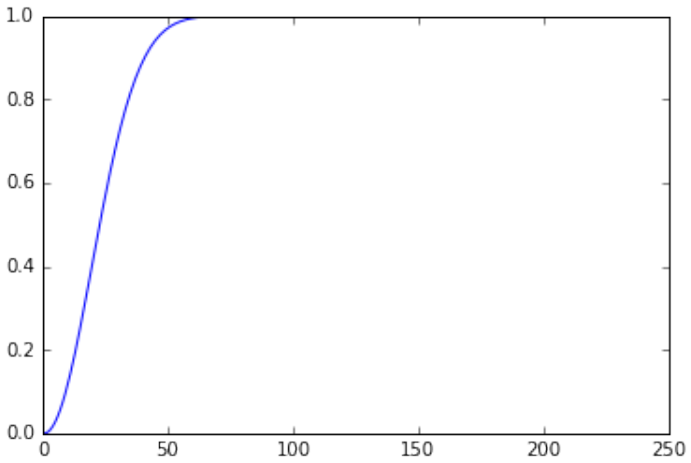
When there are two person in the room, the probability that everyone has different birthdays will be $\frac{364}{365}$. When there are three people, the probability can be defined by the previous probability (with 2 people) and the probability that the third birthday is unique. Thus, the probability can be given by

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

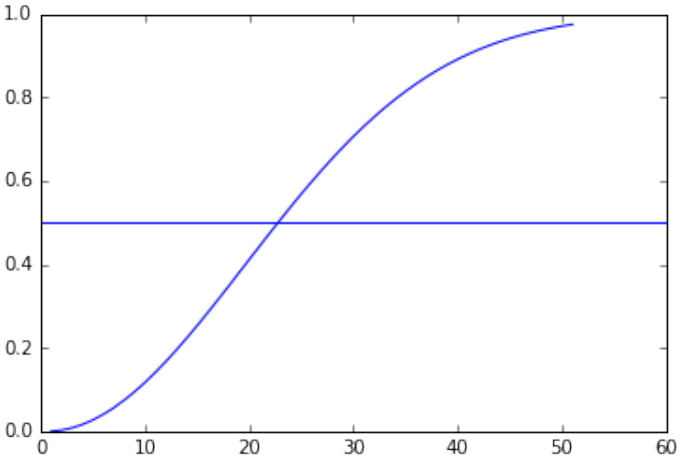
Let's extend the logic to n people, with a table:

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 (n > 3)	$\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times ... \times \frac{(366-n)}{365}$	1 - $(\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times ... \times \frac{(366-n)}{365})$
n ($n \leq 365$)	$\frac{365!}{(365-n)!365^n}$	1 - $(\frac{365!}{(365-n)!365^n})$

Now that we've found a formula for the probability, let's graph it:



Wow! The probability spikes up very quickly, and when n is greater than 100 people, the probability is near 1. The reason for this spike-up is the involvement of combinatorics, which causes the probability to increase very fast compared to the increase in the number of people. 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.

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? 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

For this application, we will need to define our conditions. For what follows, N and n are integers greater than 1. When you are playing this gambling game n times, you have a chance of winning $\frac{1}{N}$ each time you play. The chance of winning a certain bet is the same no matter what the outcomes of any previous bets were. These are the only conditions needed; they will remain constant.

$$P(Winning\ Bet) = \frac{1}{N}$$

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

$$P(Losing\ 1\ Bet) = 1 - \frac{1}{N}$$

Knowing the probability we can lose one bet brings us to the next thing we want to find: What is the chance that we will lose n times straight? Let's first figure out the chance we can win $n = 2$ bets. We can figure this out by applying simple rules of probability we have just learned. The probability of getting both results is their probabilities multiplied together, essentially returning a probability of a probability. Our assumptions above tell us that the probability of winning a bet is the same for each bet. No differently, our probability of losing the second bet is the same as our probability of losing the last bet. This gives us:

$$P(Losing\ 2\ bets) = (1 - \frac{1}{N}) * (1 - \frac{1}{N})$$

Now, finding the chance that we lose n times straight is simple. If we think of each succeeding bet as a probability of its preceding bet's probability, the probability of losing the next bet is simply the probability of losing the next bet times the probability of losing the previous bets. With our assumptions, we can conclude that:

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

The final chance we have to find is the chance of winning at least 1 bet out of n bets. 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 is simple. The opposite of losing all of the bets is winning at least 1 bet. That is it!

$$P(Winning\ at\ least\ 1\ bet)\ =\ 1-(1-\frac{1}{N})^n$$

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? Come up with a guess and keep it with you for the end, when we solve the problem. In order to solve for this, we will need to solve for the smallest n for which the chance that you win at least one of the n times is greater than $1/2$. Mathematically, we say:

$$\frac{1}{2} < 1-(1-\frac{1}{N})^n$$
$$\frac{1}{2} >\ (1-\frac{1}{N})^n$$

In order to isolate n, let's take the logarithm of both sides

$$log\frac{1}{2} > nlog(1-\frac{1}{N})$$

Since the logarithm is a strictly increasing function, it preserves the inequality

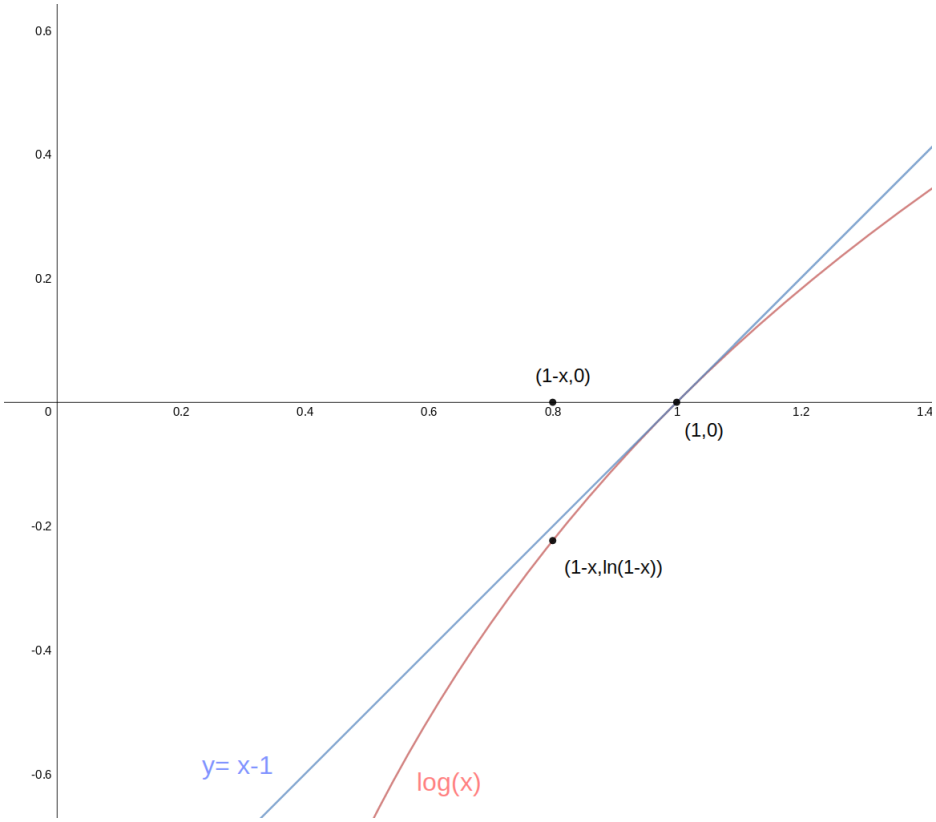
$$\frac{log\frac{1}{2}}{log(1-\frac{1}{N})}\ <\ n$$

Remember that we must flip the inequality because $log(1-\frac{1}{N})$ is negative!

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

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. To help us solve our problem, we'll find 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). We will take a small side track and prove this before we move on.

Let us draw a graph of the function f(x)= log(x) along with its tangent line at x = 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 plots? They produce a triangle! Not just any triangle, but a 45-45 right triangle. Using our knowledge of 45-45 right triangles, solve for the distance from 1 to point A. The two sides of the triangle are both x. Therefore, the distance from log(1) to point A, which is log(1-x) closely approximates to x (Recall that distances are always positive). This means that:

$$log(1-x) \approx x$$

Now that we know that the approximation for log(1-x) is about -x itself for small, positive values of x, we can proceed to our proof.

Let's plug that into our equation:

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

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

$$\frac{log\frac{1}{2}}{-\frac{1}{N}} < n$$

All that remains now is simplification

$$\begin{aligned} -Nlog\frac{1}{2} &< n \\ -N(log(1)-log(2)) &< n \\ -N(-log(2)) &< n \\ Nlog2 &< n \end{aligned}$$

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

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

Gamblers have know for centuries that the answer to the question we posed is about 2/3 of N. Plug large numbers into this equation. You will soon realize that you need an absurdly large number to get a fair chance. You need 2/3 of a million bets as a matter of fact! Was this close to your guess?

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.

2 Probability With and Without Replacement

Sampling With Replacement

A *population* is a list consisting of N distinct individuals $i_1, i_2, , ..., i_N$. The method of *sampling with replacement* occurs when a drawn element of a population is placed back into the population, in which it may be redrawn for another sample.

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

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

Example 1. If a random sample of size n is drawn with replacement from the list, how many possible samples are there?

Solution: Suppose that k represents the number of times that an element is drawn from N. If $k = 1$, then the number of possible samples is simply N. If $k = 2$, then the number of possible samples amounts to N^2, since there are N possible second draws for each of the N possible first draws.

Hence, the following formula can be used to derive the number of possible samples drawn *with replacement*,

$$N^k$$

where the variable N is the number of elements in the total population and k is the number of elements in the sample.

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

Solution: Suppose that N represents the population size and n represents the sample size in all the following examples in this section.

In this experiment, the probability

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

represents the probability of you not being chosen at all, since $1 - \frac{1}{N}$ represents the chance that you are not chosen in any given draw, and there are n draws within the sample. Thus, using the complement rule, the probability

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

represents the chance of being chosen at least once in the sample.

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

Solution:

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

The answer to Example 3 is the same as the answer to Example 2 because the identity of the individual in question does not matter. By definition, a random sample yields the same probability for any individual being chosen. Hence, it does matter if a question asks about you, your best friend, or Batman in all his caped glory; the random sample does not care. As long as they are all unique individuals within the population, the probability of any given individual being chosen is the same.

Example 4. Assume that N is very large in comparison to n. Give a simple approximation of the chance in Question 2. This approximation should be easy to understand without a calculator.

Solution: We've completed the task of solving Question 2. Now we must perform operations toward making this complicated answer more understandable.

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

For the purpose of this exercise, let's write this as follows:

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

Try to put this equation in a form so that we can use the approximation we just recalled. Taking the natural log of both sides and applying log properties yields

$$log(1 - P) = n * [log(1 - \frac{1}{N})]$$

Now we apply the approximation, (your x in this case equals $(-1/N)$), which leaves us with

$$log(1 - P) \approx n * [\frac{-1}{N}] \Rightarrow log(1 - P) \approx \frac{-n}{N}$$

We can escape the log by raising e to the power of both sides. This gives us the probability of the complement and, therefore, the probability of the event.

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

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

Before the gambler's approximation was formulated, gamblers would employ a very simple approximation to estimate probabilities. They found that for the probability of at least one success to be 1/2 in this sort of game, you would have to draw roughly (2/3)N samples of size 1. That is to say, in order for P to be $\approx \frac{1}{2}$, n would have to be $\approx \frac{2}{3}N$.

Sampling Without Replacement

Recall that a *population* is a list consisting of N distinct individuals i_1, i_2,,...,i_N. A *random sample without replacement* from the population is defined by the following sampling scheme:

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

A random sample of size n is drawn without replacement from the list.

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

Example 1. How many possible samples are there?

Solution: Let's first consider a sample size of one ($n = 1$). The total number of samples would be N, as we could pick any one of the N numbers. Now, let's consider a sample size of two. The total number of samples becomes $N * (N - 1)$. This is because we have any choice in N for our first selection, then any choice in $N - 1$ for our next choice. We can continue this relationship up to a sample size of n:

$$N - (n - 1)!$$

The $n - 1$ is because we start off with just $N!$ when $n = 1$. To simplify this quantity, let's multiply the numerator and denominator by $(N - n)!$.

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

Simplify:

$$\frac{(N - n + 1)! * (N - n)!}{(N - n)!}$$

By combining the factorials in the numerator, we get:

$$\frac{N!}{(N - n)!}$$

But wait! We're not done yet. Suppose we were choosing two kinds of fruits out of a population consisting of apples, bananas, and peaches. Our formula says we'd get 6 possible outcomes, but let's list them out.

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

The problem we face here is that many of these combinations are contain the same two fruits, but are arranged in a different order. In order to account for this, we divide the number of possible combinations one more time by $(n!)$. This way, we're eliminating all of the repeat cases and obtaining the number of unique combinations. We now can write our final formula.

When drawing a sample without replacement from a population, the total number of samples possible is written:

$$\frac{N!}{n!(N-n)!}$$

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

Because of the prevalence of this quantity in probability, another way to write $\frac{N!}{n!(N-n)!}$ is:

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

where once again the variable N is the number in the total population and n is the sample size. We read this notation as N *choose* n.

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

Solution: To begin, we know the total number of samples is $\binom{N}{n}$, all of which are equally likely to result. We'll divide the number of samples that include you by the total number of samples possible to find the chance of you being chosen. The samples in which you are already chosen represent choosing a sample size of $(n-1)$ out of $(N-1)$ samples. Putting these two quantities together, we get:

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

Let's expand it out to what we got in Example 1 and simplify a little bit.

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

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

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

By simplifying the factorials out we get our final answer:

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

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

Solution: The chance that any particular individual is chosen in the sample is the same for all individuals in the sample because all individuals in the sample are treated identically in the sampling process. Therefore, the answer to this example is the same as the answer to Exercise 2.

Example 4. What is the chance that you and your best friend are both chosen in the sample? Would this chance be different for any other pair of individuals on the list?

Solution: Since it's more intuitive to find the probability that you and your best friend are not chosen in the sample, we'll first find the probability of that occurring and then apply the complement rule to find the chance that you and your friend are both chosen.

The chance that you or your friend aren't chosen in the first draw of the sample is:

$$\frac{N-2}{N}$$

and in the second draw is:

$$\frac{N-3}{N-1}$$

And in subsequent draws will be:

$$\frac{N-4}{N-2}, \frac{N-5}{N-3}, ..., \frac{N-(n+1)}{N-(n-1)}$$

Thus, the probability of you and your friend not being chosen could be represented as:

$$\frac{N-2}{N} * \frac{N-3}{N-1} * \frac{N-4}{N-2} ... * \frac{N-(n+1)}{N-(n-1)}$$

which is equivalent to:

$$\frac{\binom{N-2}{n}}{\binom{N}{n}}$$

after applying the complement rule, you arrive at the answer:

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

Random Permutations

A random permutation is a random ordering of a set of elements. A practical example of what a random permutation is would be the combinations produced by shuffling a deck of cards so that they are randomly distributed. These combinations are essentially random permutations of a set of 52 elements i.e. the number of cards in a deck.

Have a look at the following examples to further clarify this concept:

Let us assume that there are N cards in a deck. A random permutation is the sequence of cards obtained by drawing a random sample of n without replacement from the deck.

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

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

Solution: Mathematically, there are $n!$ ways that a set of unique elements can be arranged in a straight line. Thus, there are $n!$ ways that the set can be permuted, and all are equally likely.

To find the chance that the top card has the gold star, we have to count the number of permutations in which this occurs. In this case, we assume that the card with the gold star occupies a fixed position on that line i.e., it is fixed at the top of the deck. This implies that there are $(n-1)!$ ways that the other cards can be arranged on that line, or in other words, $(n-1)!$ possible deck combinations with the the gold star on the top card.

Hence, the probability that the card with the gold star is on the top can be calculated in the following way:

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

$$P = \frac{1}{N}$$

Let us look at another example:

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

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

Suppose that we were calculating the probability that the card with the gold star was located at any single position k within the deck. The calculation for this chance, based on our work in the previous example, would look like:

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

Therefore, for any position k, including the position below the top card in the deck, the chance the the card with the gold star occupies the position k is

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

.

Key Takeaway: *The probability that a certain card occupies any particular position k in a standard deck of cards is* $\frac{1}{n}$

Note: For examples 3 to 6, a *standard deck* is represented by a set of cards with faces given by the set of integers from 2 to 10, jacks, queens, kings and aces; all with suits:

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

Example 3. What is the chance that the last card dealt is the ace of spades?

Solution: To solve this problem, consider we have a simpler problem that asks us to find the chance that the second card dealt is the ace of spades. The probability can be represented as

$$\begin{aligned} P &= P(1st\ card\ is\ not\ the\ ace\ of\ spades) * P(2nd\ card\ is\ the\ ace\ of\ spades) \\ &= \frac{51}{52} * \frac{1}{51} \\ &= \frac{1}{52} \end{aligned}$$

From this simpler example, we can see that the probability is $\frac{1}{52}$ or $\frac{1}{n}$. Therefore, we can conclude that the formula applies regardless of the position of the card drawn. The chance that the last card dealt is the ace of spades is simply

$$P = \frac{1}{52}$$

Example 4. What is the chance that the 37th dealt is black?

Solution: Recall from Example 7 that the position of the card drawn does not matter, the probability would simply be

$$P = \frac{26}{52} \ or \ \frac{1}{2}$$

because all cards with spade or clubs suite are considered black cards. There's a total of 26 black cards in a deck.

Example 5. What is the chance that the 28th card dealt is red given that the 50th card dealt is black?

Solution: Instead of having 52 cards, we treat the deck as one with only 51 cards, in which there are 26 red cards and 25 black cards, since we already know that the 50th card is occupied by a black card. Therefore:

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

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

Solution:

The chance that any ace occupies a given position within the standard deck is $\frac{4}{52}$ which is

the same regardless of whether the card is drawn with or without replacement. If an ace already occupying a given position within the deck, then the chance that another ace is occupying a second given position is $\frac{3}{51}$. Therefore, the probability of having four aces at the bottom of a uniformly shuffled standard deck is:

$$\frac{4*3*2*1}{52*51*50*49} = \frac{48!4!}{52!}$$

Note that this probability applies to the chance of dealing aces in any four fixed locations within the deck, whether at the top of the deck, bottom of the deck, or dispersed throughout.

A better way to imagine how this works would be to picture the deck being dealt in a circle so that there is no first or last card in the deck, or to calculate the chance of having four aces at the top of the deck - it's the same.

Therefore, the probability of drawing certain cards at predetermined locations within the deck does not depend on where the locations are, but how many locations are there.

3 Averages and Deviations

Sigma 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 4 ***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 $3n$. Representing numbers and expressions through sigma notation is a good segue to our next topics like averages.

Averages

What is an average?

The average is a single representational value for a list of numbers. For a list of numbers $x_1, x_2, \ldots, x_n$, we define the average $\bar{x}$ to be

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

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, and then equalize them into n equal numbers, still having the same sum. We can also define the average by

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

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

Errors in Averages

Say we're computing an average, and we accidentally mis-enter one of our values x_i, replacing it with a value k. The error E in our average (that is, the difference between the averages of $x_1, \ldots, x_i, \ldots, x_n$ and $x_1, \ldots, k, \ldots, x_n$) is given by

$$\boxed{E = \frac{x_i - k}{n}}$$

(Check this yourself!) Thus the only things we need to know to determine E are the correct value, the mistakenly entered value, and the total number of values being averaged. The equalizing behavior of the average is clear from this result: the more values we have in our list, the smaller the effect a single outlier can have. To quickly notice if there is an error in an average, it would be nice to establish lower and upper bounds on an 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

$$x_j \geq n$$

so

$$\frac{1}{n}\sum_{n} x_i \geq \frac{1}{n}\sum_{n} m$$

$$\bar{x} \geq \frac{nm}{n}$$

$$\bar{x} \geq m$$

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 $\bar{x}$,

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

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? One might expect that we could just add the two averages and divide by two, but a little calculation shows this doesn't quite work. Let's call the average of $x_1, x_2 \ldots x_n, y_1, y_2, \ldots y_m$ by A. Then we see

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

Rather than just directly joining the two averages, we first have to "weight" them according to their length. Similarly if we take the average of three lists with averages $\bar{x}, \bar{y}, \bar{z}$ and lengths n, m, p, the average of all $n+m+p$ elements is

$$A = \frac{n\bar{x} + m\bar{y} + p\bar{z}}{n+m+p}$$

and for k lists with averages a_i and lengths n_i, the average is

$$\boxed{A = \frac{\sum_{i=1}^{k} n_i a_i}{\sum_{i=1}^{k} n_i}}$$

Proofs of the last two statements are similar to the two-average case, and are left to the reader. One application of the last result is to finding 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 $n_1, n_2, \ldots n_k$. Then we can split the list into i sub-lists, each containing only one distinct value v_i. Each of these lists has average v_i and length n_i, so applying the above result, the average of the entire list is

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

Markov's Inequality

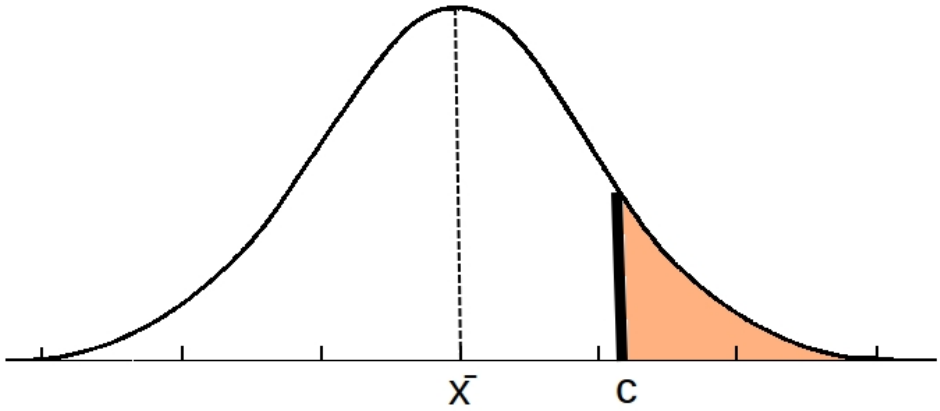
Introduction

We now know 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 is at most $\bar{x}/c$.

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

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

So if we have a set S with the following distribution, this is what Markov's Inequality looks like graphically:



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 $\frac{\bar{x}}{c}$.

Example 1: For a list of non-negative numbers, what can you say about the proportion of entries that are at least 10 times the mean?

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 $c = 10\bar{x}$, will give us 1/10. Therefore, **at most one-tenth** of all entries in the list are greater than ten times the mean.

Using Markov's Inequality

It's important to note that Markov's Inequality does not give the exact proportion of entries that are at least as large as c. Instead, it gives an *upper bound* for the proportion. How useful this bound is varies problem by problem. Let's look at another example.

Example 2: Now let $c = \bar{x}/2$. What does Markov's Inequality say for this value of c?

Applying Markov's Inequality, with $c = \bar{x}/2$, we get that the proportion of entries greater than c is 2. However, all proportions are already upper-bounded by 1, so while this bound is not wrong, it is trivial.

Looking back at the inequality, we can see that for the bound to be less than 1, c has to be greater than $\bar{x}$. Therefore, Markov's Inequality is only useful for $c > \bar{x}$.

Proof

From the examples above, we can construct a formal math statement of the inequality:

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

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

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.

Now let's construct a proof for Markov's Inequality.

Let S be the set consisting of $x_1, x_2, \ldots, x_n$ non-negative numbers with average $\bar{x}$, and $c > 0$. Then the sum of all x_i for i from 1 to n is n times $\bar{x}$. In math notation:

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

(Equation 1)

We can split the right hand side of this equation into two groups: the sum of all x_i's that are less than c, and those that are greater than or equal to c.

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

(Equation 2)

Let's look at the first summation on the right side of Equation 2:

$$\sum_{i:x_i<c} x_i$$

This is the sum of all entries in S that are less than c. To get to Markov's Inequality, we want to say that this summation is *greater than* something. Since the tightest lower bound we know of the entries is 0, we can say that

$$\sum_{i:x_i<c} x_i \geq \sum_{i:x_i<c} 0$$

(Equation 3)

Now looking at the second summation:

$$\sum_{i:x_i\geq c} x_i$$

This is the sum of all entries in S that are at least c. We want to do the same thing with this summation as we did with the last one: find a lower bound, and state that this summation is greater than said bound. We know that every x_i is at least c, so

$$\sum_{i:x_i\geq c} x_i \geq \sum_{i:x_i\geq c} c$$

(Equation 4)

Substituting in Equations 3 and 4 into Equation 4 and simplifying, we get

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

This concludes our proof for Markov's Inequality.

Standard Deviation and Variance

What is Standard Deviation?

The word "deviation" naturally suggests a notion of "distance". For example, how could we calculate the "distance" between A and B? Usually, we do subtraction, using $A - B$ or $B - A$ to denote the "distance" between any of two numbers. If our measure of how spread out a set of data values are to be the average deviation from mean, we might want to sum up all "distances" first and then divide the sum by the number of terms.

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

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, and therefore it cannot be used to describe the spread of a variable. To avoid cancellation, we have to ensure all "distances" to be non-negative, so we square all "distances" before taking the average. Moreover, to

"fix" the unit which has also been squared, we then have to extract the square root of the average.

Definition 6 ***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

Example: 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.

Solutions:

Step 1: Calculate the Mean ($\bar{x}$)

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

Step 2: Calculate $(x_i - \bar{x})^2$

For each value, subtract the mean and square the result. For x_1, the result would be

$$(x_1 - \bar{x})^2 = (82 - 80)^2 = 4$$

Step 3: Calculate $\sum_{i=1}^{n}(x_i - \bar{x})^2$

Add up all squared differences

$$(82-80)^2 + (63-80)^2 + (81-80)^2 + ... + (89-80)^2 + (84-80)^2 = 1228$$

Step 4: Calculate the Standard Deviation

Divide the sum by n (the number of values) and extract the root

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

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 extracting the square root. 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. Is there a formula that allows us to compute variance by hand?
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.

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

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

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

Simplify the expression and combine like terms

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

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

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

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.

Chebyshev's Inequality

Chebyshev as an extension of Markov

Chebyshev's inequality is an extension of Markov's inequality that allows us to get a bound for the proportion of x_i in the tails of a distribution. Chebyshev is at most $\frac{1}{k^2}$, where k is a number of standard deviations away from the mean.

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.

4 Random Variables

In this chapter, we will put together the ideas we have developed for probability and descriptive statistics, to build tools that will help us understand the averages of random samples.

What is a Random Variable?

The story begins with an *outcome space*, that is, the set of all possible outcomes of an experiment. Standard notation for this space is Ω, the upper case Greek letter Omega. **For mathematical simplicity, we will assume that the outcome space Ω is finite.** Each outcome ω (that's lower case omega) is assigned a probability; the total probability of all the outcomes is 1. Here is an example where Ω consists of 8 equally likely outcomes. You can think of Ω as the possible outcomes of three tosses of a coin, as an aid to comprehension.

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

A *random variable* X is a real-valued function defined on Ω. That is, the domain of X is Ω and the range of X is the real line. In the example above, X could be the number of times the letter H appears in an outcome; you can think of X as the number of heads in three tosses of a coin.

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

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

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

Probability distribution.

The *probability distribution* of X is a distribution on the range of X. It specifies all the possible values of X along with all their probabilities. For example, for X defined in the paragraph above, the probability distribution is given by

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

All the probabilities in a distribution must add up to 1. Note that the probabilities on the range of X are determined by the probabilities on the domain. In our example, had all 8 elements in the domain not been equally likely, the possible values of X would still have been the same but the probabilities might have been different.

Characteristics of Random Variables

Expectation.

The *expectation* of X is the average of the possible values of X, weighted by their probabilities. This can be computed in two equivalent ways, one defined on the domain of X and one on the range:

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

In our numerical example, the first form of the calculation is

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

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

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

The form

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

is most commonly used in calculation, and shows that $E(X)$ is the balance point of the histogram of the probability distribution of X. Therfore it is just an ordinary average and has all the familiar properties of averages. For example, it preserves linear transformations:

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

There is a trivial case that is worth noting, and arises if $a = 0$ in the transformation above. If the random variable X is a constant, that is, if there is a constant c such that $P(X = c) = 1$, then $E(X) = c$.

Expectation is often denoted by μ. That's the lower case Greek letter mu; it stands for "mean".

Standard Deviation.

We will be interested in how far a random variable is likely to be from its expectation, just as we were interested in seeing how far numbers in a list are from their average. Define the *deviation* of X as the random variable $D = X - \mu$, where $\mu = E(X)$.

To find the rough size of D, suppose we calculate $E(D)$. Since D is a linear transformation of X,

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

The expected deviation is 0, no matter what the distribution of X is. This is the parallel to the result that the average of a list of deviations from average is 0 no matter what the list is. The problem, as before, is cancellation: the negative deviations cancel out the positive ones. So, as before, we square the deviations to get rid of the cancellation. This leads to the definition of the *variance* of X, denoted $Var(X)$.

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

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

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

This can be thought of as the SD of a list of numbers, computed using the distribution table. As such, it is an ordinary SD and has all the properties of SDs. For example, only the multiplicative factor of a linear transformation affects the SD:

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

If a random variable X is a constant, then $SD(X) = 0$.

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

Bounding Random Variables

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

Markov's Inequality.

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

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

Chebychev's Inequality.

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

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

More formally,

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

These are exactly the same as the bounds obtained earlier for distributions of data, and they are true for the same reasons. If you would like to prove them afresh, you can follow the steps in the proof in Chapter 3 but write them in random variable notation.

Multiple Random Variables

Additivity of expectation.

Thus far, we have worked with one random variable and functions of it. We will now look at functions of several random variables defined on the same space. Among the simplest and most powerful such functions is the sum. Let X and Y be two random variables defined on the same space Ω. For every $\omega \in \Omega$, define the sum $S(\omega) = X(\omega) + Y(\omega)$. In random variable notation, $S = X + Y$. Then, no matter what the relation between X and Y,

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

This is a result of fundamental importance, as you will see in the examples below. But first, let us prove it. We will use the first definition of expectation, where the sum is over all the elements in the domain of the random variable.

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

Example 1: Computational formula for Variance. Let $E(X) = \mu$ and $SD(X) = \sigma$. Then

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

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

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

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

We will now develop some tools that allow us to derive the expectation and SD of the mean of a random sample.

Independence.

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

Formally, events A and B are *independent* if the conditional chance of B given that A has occurred is the same as the unconditional chance of B. This, along with the multiplication rule, motivates the formal definition of independence: Events A and B are independent if $P(AB) = P(A)P(B)$. Here $P(AB)$ is the chance that A and B both occur. Notice that according to the multiplication rule, the second factor on the right hand side should be the conditional chance of B given that A has happened; but independence means that the value of this conditional chance is the same as $P(B)$.

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

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

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

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

Proof.

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

Variance of a sum of independent random variables. We have seen that $E(X+Y) = E(X)+E(Y)$ no matter what the relation between X and Y. We will now show that

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

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

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

The expectation and SD of a random sample sum. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables. For example, they could be the results of draws made at random with replacement from a population. Then each X_i has the same expectation μ and the same SD σ as all the others. Let $S_n = X_1 + X_2 + \cdots + X_n$. Then for all $n \geq 1$,

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

Proof. By additivity of expectation,

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

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

$$\begin{aligned} Var(S_n) &= Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) \quad \text{by independence} \\ &= n\sigma^2 \end{aligned}$$

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

The expectation and SD of a random sample mean. Let $X_1, X_2, \ldots, X_n$ be i.i.d. as above, and let $A_n = S_n/n$ be the sample average. Then for all $n \geq 1$,

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

Proof. A_n is just a linear transformation of S_n:

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

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

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

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

Part II

Regression

5 Correlation

Formal Definition of R

Up to this point, we have only dealt with the analysis of single variables; but often, in statistics we are curious about the relationship between two or more variables. A classical way of visualizing the relationship between two lists of values, say heights and weights of data scientists, is to plot the data by way of a scatter plot. A scatter plot takes both lists, and treats them as coordinates in two dimensional space, where the axes are defined by the categories of the variables (in this case heights and weights comprise the x and y axes, or vice versa). We are left with some visualization of the relationship between these two lists, a shape in form of a cloud, but without numerical estimate of this relationship. This is where the correlation, helps us interpret our scatter plot:

Correlation, often expressed by the symbol r, is a number ranging from -1 to 1, which expresses the degree of clustering around a straight line. The correlation will give us a numerical relationship between these lists; the formal definition of this numerical relation being the mean of the sum of the products of two lists, $\mathbf{x}$ and $\mathbf{y}$. If $\mathbf{x}$ and $\mathbf{y}$ have been transformed into standard units, their correlation, r, would look like this in notation:

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

This will measure, roughly, the degree of clustering around a straight line. Furthermore, the correlation coefficient is the slope of the line of best fit for a regression line for predicting y when given x if both are in standard units.

We can simplify this definition of r to a more computational form, by simply expanding

and simplifying the product.

$$
\begin{aligned}
r &= \frac{1}{\sigma_x \sigma_y}\left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \bar{y} - \frac{1}{n}\sum_{i=1}^{n} y_i \bar{x} + \frac{1}{n}\sum_{i=1}^{n} \bar{x}\bar{y}\right] \\
r &= \frac{1}{\sigma_x \sigma_y}\left[\overline{xy} - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y}\right] \\
r &= \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y}
\end{aligned}
$$

We'll call this simplified formula the "computational r", since it is much more straightforward to compute. The numerator, $xy - \overline{xy}$ is a statistical metric called covariance, which is the measure of how much variables change with respect to each other, of the lists $\mathbf{x}$ and $\mathbf{y}$. When the covariance between the two lists is zero, the correlation is unsurprisingly zero.

When $x_i = y_i$ for every i, the two lists would have the same mean and variances, since they are the same list, so it makes intuitive sense that they would have a correlation of 1. As such, the proof is in recognizing the computational variance when $\mathbf{x} = \mathbf{y}$ is the mean of squares minus the square mean of x. By simplification:

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

Linear Transformations and Correlation

Let's look at a case where we have to find the correlation between y and another linear transformation of x. Say we have another set: $z = ax + b$ where a and b are constants; we'd like to find the correlation between

$$
\begin{aligned}
r(ax+b,y) &= \frac{\frac{1}{n}\sum_{i=1}^{n}(ax+b)y - (a\bar{x}+b)y}{(|a|\sigma_x)\sigma_y} \\
&= \frac{\frac{a}{n}\sum_{i=1}^{n}(xy) + \frac{b}{n}\sum_{i=1}^{n}(y) - ax\bar{y} - b\bar{y}}{|a|\sigma_x\sigma_y} \\
&= \frac{a\overline{xy} + b\bar{y} - ax\bar{y} - b\bar{y}}{|a|\sigma_x\sigma_y} \\
&= \frac{a(\overline{xy} - \bar{x}\bar{y})}{|a|\sigma_x\sigma_y} \\
&= \pm\frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x\sigma_y} \\
&= \pm r(x,y)
\end{aligned}
$$

After simplifying the initial formula we end up getting the computational formula for r, the value for which ranges from -1 to 1. This is an interesting result, and it tells us that the magnitude of correlation is preserved across linear transformations. This result will play a big part in proving some interesting relationships in Chapter 7, and allows us to put mathematical rigor behind the intuitions that we develop.

We can use this intuition to prove a related result: what happens to the correlation when **x** and **y** are linearly related, as in the list of **y** is a linear transformation of **x**: $y_i = ax_i + b$.

Recall that the correlation between a variable and itself is 1:

$$r(x,x) = 1$$

Since a linear transformation preserves the correlation between variables,

$$r(x,ax+b) = \pm 1$$

$$r(x,y) = \pm 1$$

So, depending on the sign of a defined through the linear transformation, the correlation between x and a linear transformation is 1 if $a \geq 0$ and -1 if $a \leq 0$. This seems to be an example of a perfect correlation between two variables, but is ± 1 the largest correlation we can get between two variables?

Bounds on Correlation

As it turns out, the correlation coefficient is always bounded between ± 1, a fact we'll attempt to prove mathematically in the following section.

Theorem 1 ***Bounds on Regression***
The correlation coefficient defined by $r = \frac{1}{n}\sum_{i=0}(\frac{x_i-\bar{x}}{\sigma_x})(\frac{y_i-\bar{y}}{\sigma_y})$ *is always between -1 and 1.*

Before we proceed to the proof, let's define some variables to reduce complexity and simplify calculations. Since we're considering all possible sets of x and y, we'd like to work with these sets on the same scale, so we normalize them. Specifically, we normalize x and y to x' and y' respectively, so now:

$$x' = (\frac{x_i - \bar{x}}{\sigma_x}) \qquad y' = (\frac{y_i - \bar{y}}{\sigma_y}) \qquad r = \frac{1}{n}\sum_{i=0} x'y'$$

Notice that we've now cleverly rewritten the formula for correlation into terms of x' and y', which represent the sets x and y converted to standard units. Normalizing to x' and y' does something interesting to the means and standard deviations, which we shall describe here. When we convert the set x to standard units, first we subtract the mean of the original set, $\bar{x}$ from every member of that set. This is a linear transformation, so the distribution, or SD (σ_x), of the set is unaltered, but the mean of the new set is now 0. Next, every member of the new set is divided by the standard deviation of the original set. This squeezes or stretches the set such that the new standard deviation is 1, in standard units. By this explanation, $\bar{x'}$ and $\bar{y'}$ equal 0, and $\sigma_{x'}$ and $\sigma_{y'}$ should equal 1.

$$\sigma_{x'}^2 = 1 = \frac{1}{n}\sum(x' - \bar{x'})^2 \qquad \text{since } \sigma_{x'} = 1$$
$$1 = \frac{1}{n}\sum(x')^2 \qquad \text{since } \bar{x'} = 0$$

The final step in preparing for our proof is to recall the formula for the variance: $\sigma_{x'}^2 = \frac{1}{n}\sum(x' - \bar{x'})^2$, so the mean sum of squares for the normalized variables is $\frac{1}{n}\sum(x')^2 = 1$ as seen above.

Our goal in this proof is to place bounds on the values of $r = \frac{1}{n}\sum x'y'$. Realize that $(x' + y')^2$ and $(x' - y')^2$, when expanded, have the term $x'y'$ (one positive and one negative) in them. Also, conveniently, both of these terms have a lower bound of zero, since any real squared value must be greater than or equal to zero. So if we look at $\frac{1}{n}\sum(x' + y')^2$ and $\frac{1}{n}\sum(x' - y')^2$, they should simplify to a positive and negative bound for $r = \frac{1}{n}\sum x'y'$, and the computation follows.

$$
\begin{aligned}
\forall i: \quad & 0 \le (x_i' + y_i')^2 \\
& 0 \le \sum_{i=0}^{n} (x_i' + y_i')^2 \\
& 0 \le \frac{1}{n} \sum_{i=0}^{n} (x_i' + y_i')^2 \\
& 0 \le \frac{1}{n} \sum_{i=0} ((x')^2 + 2x'y' + (y')^2) \\
& 0 \le \frac{1}{n} \sum_{i=0} (x')^2 + \frac{1}{n} \sum_{i=0} 2x'y' + \frac{1}{n} \sum_{i=0} (y')^2 \\
& 0 \le 1 + 2\frac{1}{n} \sum_{i=0} x'y' + 1 \\
& -2 \le 2\frac{1}{n} \sum_{i=0} x'y' \\
& -1 \le \frac{1}{n} \sum_{i=0} x'y' \\
& -1 \le r
\end{aligned}
$$

$$
\begin{aligned}
\forall i: \quad & 0 \le (x_i' - y_i')^2 \\
& 0 \le \sum_{i=0}^{n} (x_i' - y_i')^2 \\
& 0 \le \frac{1}{n} \sum_{i=0}^{n} (x_i' - y_i')^2 \\
& 0 \le \frac{1}{n} \sum_{i=0} ((x')^2 - 2x'y' + (y')^2) \\
& 0 \le \frac{1}{n} \sum_{i=0} (x')^2 - \frac{1}{n} \sum_{i=0} 2x'y' + \frac{1}{n} \sum_{i=0} (y')^2 \\
& 0 \le 1 - 2\frac{1}{n} \sum_{i=0} x'y' + 1 \\
& -2 \le -2\frac{1}{n} \sum_{i=0} x'y' \\
& 1 \ge \frac{1}{n} \sum_{i=0} x'y' \\
& 1 \ge r
\end{aligned}
$$

$$\therefore -1 \le r \le 1$$

6 Linear Regression

Deriving the Formula of the Regression Line

Suppose we have a data set in two variables $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$, not necessarily in standard units, and we want to fit a linear equation to the data. We know that since this regression is linear, it will have the equation $y = ax + b$ for some real a, b, and we want to find a and b in terms of what we already know about our x's and y's. To make our line fit our data as closely as possible, we'd like to minimize the distance between the actual values in our data set and their estimates on the line. We call these differences residuals, and they will be discussed at length in the next chapter. Specifically, our choice of a and b should minimize the mean square of the residuals (we square to eliminate signs). This minimization is the same as that in seen in any first semester calculus class, though it may be seem more intimidating because of all the sums and variables. First, we will pose our mean square residuals as a function $f_a(b)$ of the y-intercept b, holding the slope a constant. Once we've obtained an optimal b in terms of a, we'll plug that value back into our definition of the residuals and make another function $f(a)$, which we will also minimize. Consider $f_a(b)$, the mean square error of the residuals with respect to b for any given a:

$$\begin{aligned} f_a(b) &= \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - (ax_i + b)\Big)^2 \\ \frac{d}{db}f_a(b) &= \sum_{i=1}^{n}\frac{-2}{n}\Big(y_i - (ax_i + b)\Big) \\ &= \frac{-2}{n}\sum_{i=1}^{n}\Big(y_i - (ax_i + b)\Big) \end{aligned}$$

This derivative likely seems very intimidating, but it absolutely isn't! Remembering that every term in a sum can be differentiated separately and that coefficients can be factored out of sums, you should take a moment and convince yourself that this calculation is correct. You should also convince yourself that $\frac{d^2 f_a}{db^2} = \frac{2}{n} > 0$ Thus any critical points will be

minima, and to find the minima, we set the derivative equal to 0:

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \Big(y_i - (ax_i + b_a)\Big) \\ \sum_{i=1}^{n} b_a &= \sum_{i=1}^{n} (y_i - ax_i) \\ b_a &= \frac{1}{n} \sum_{i=1}^{n} (y_i - ax_i) \\ &= \bar{y} - a\bar{x} \end{aligned}$$

Thus for a given a, the optimal y intercept will be $b_a = \bar{y} - a\bar{x}$.

Now, to find the slope of the regression line that minimizes error, we will substitute our value for b_a into our original equation. We will first set up the original equation as:

$$f(a) = \sum_{i=1}^{n} (y_i - (ax_i + b_a))^2$$

We know from our differentiation above that for the fixed slope a, the optimal y intercept $b_a = \bar{y} - a\bar{x}$ We can make a substitution in the original equation for b_a within our original equation. With the substitution in, we will now have:

$$f(a) = \sum_{i=1}^{n} (y_i - (ax_i + \bar{y} - a\bar{x}))^2$$

Notice that we can separate the x's and y's in our equation. Our next step will be to separate them and to keep them in that form without any further simplification. After this step, our equation will look like:

$$f(a) = \sum_{i=1}^{n} ((y_i - \bar{y}) + a(x_i - \bar{x}))^2$$

Now that our set up is finally finished, we will begin our simplification. Notice that the whole equation is being squared. We know from basic algebra that $(a+b)^2 = a^2 + 2ab + b^2$ by using FOIL. After simplification, we will end up with:

$$f(a) = \sum_{i=1}^{n} [(y_i - \bar{y})^2 - 2a(y_i - \bar{y})(x_i - \bar{x}) + a^2(x_i - \bar{x})^2]$$

In order to solve for the a that minimizes the error, we will need to set the derivative of our equation to 0. As such, the a^* we find will give us our minimal error. We need to have an a^* as the variable a can take up many values, while the a^* can only take up one value, which in this case is the minimum error. Therefore, after differentiating with respect to a, we have the following: at the point $a = a^*$,

$$\frac{d}{da}f(a) = 0 = \sum_{i=1}^{n}[-2(y_i - \bar{y})(x_i - \bar{x}) + 2a(x_i - \bar{x})^2]$$

We will now split what we have into to different sums and bring out any constants we have (in this case that is 2 and a^*). It will look like:

$$0 = -2\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) + 2a^*\sum_{n=1}^{n}(x_i - \bar{x})^2$$

Separate the equation in order to isolate the a^*. We will now have:

$$2\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = 2a^*\sum_{i=1}^{n}(x_i - \bar{x})^2$$

By dividing through by $2\sum_{i=1}^{n}(y_i - \bar{y})(x\_\bar{x})$, we will end up with our final answer:

$$a^* = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

You may realize we have assumed nothing about the values for x and y in our distribution. This is because it does not even matter whether x and y are linear, as there is going to still be one best fit line regardless of the shape of the distribution. So even if the data are not linear, there still will exist exactly one best-fitted line.

Also notice that our optimal y intercept looks like the equation we expected from class but the slope we found is much different than that of the

$$\frac{r * \sigma_y}{\sigma_x}$$

Equivalent Formulas for Regression Line

In the previous section, we prove that the slope of the "best-fit" regression line, in a sense that minimizes the mean squared error in prediction is as below:

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

However, this expression does not look the same as what we expect to see based on the football shaped scatter plots. What is the relationship between the slope we've learned ($r\frac{\sigma_y}{\sigma_x}$) and the one we just derived? If there is, how are they related? We have the intuition that these two are essentially the same though differ in forms. Therefore,we are going to transform its form into the one with which we are familiar.

According to the formal definition of correlation coefficient

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

Take out common factors: $\frac{1}{\sigma_x}$ and $\frac{1}{\sigma_y}$

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

Multiply both expressions by σ_x, σ_y, and n

$$nr\sigma_x\sigma_y = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{6.1}$$

Does the expression remind you of anything? It is the numerator of the slope! So what about the denominator ($\sum_{i=1}^{n}(x_i - \bar{x})^2$) ? Can you think of any formula that includes it and can be used for transformation?

Recall the formula of variance

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

By multiplying both sides by n, we now have the denominator expressed as below:

$$n{\sigma_x}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{6.2}$$

Then,we get the slope of the "best-fit" regression line.

$$slope = \frac{\text{r.h.s. of Equation (1.1)}}{\text{r.h.s. of Equation (1.2)}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{nr\sigma_x\sigma_y}{n{\sigma_x}^2} = r\frac{\sigma_y}{\sigma_x}$$

The slope of the regression line ($y = ax + b$) in original units equals to

$$a^* = r\frac{\sigma_y}{\sigma_x}$$

Finally, we reach the conclusion that though differ in mathematical forms, these slopes are essentially identical.

Definition 7 ***Equation of Regression Line in original units***

$$\hat{y}_i = \frac{r\sigma_y}{\sigma_x}(x_i - \bar{x}) + \bar{y} \tag{6.3}$$

or the point-slope form:

$$(\hat{y}_i - \bar{y}) = \frac{r\sigma_y}{\sigma_x}(x_i - \bar{x}) \tag{6.4}$$

From the formula above, we can derive the formula for regression line in standard units.

First, subtract $\bar{y}$ from both sides to get Equation (1.4), which is the point-slope form of Equation (1.3)

$$(\hat{y}_i - \bar{y}) = \frac{r\sigma_y}{\sigma_x}(x_i - \bar{x})$$

Multiply both sides by $\frac{1}{\sigma_y}$

$$\frac{\hat{y}_i - \bar{y}}{\sigma_y} = r\frac{x_i - \bar{x}}{\sigma_x} \tag{6.5}$$

Recall the formula that converts x_i from original scales to standard unit scales, where x* is the x value in standard units for any x_i

$$x_i^* = \frac{x_i - \bar{x}}{\sigma_x} \tag{6.6}$$

The same formula can also be used to convert $\hat{y}_i$ to $\hat{y}_i{}^*$, where $\hat{y}_i{}^*$ is the fitted value i in standard units for any $\hat{y}_i$

$$\hat{y}_i{}^* = \frac{\hat{y}_i - \bar{y}}{\sigma_y} \tag{6.7}$$

Then substitute Equation (1.6) and Equation (1.7) into Equation (1.5)

$$\hat{y}_i{}^* = rx_i^*$$

Definition 8 ***Equation of Regression Line in standard units****:*

$$\hat{y}^* = rx^*$$

$\hat{y}^*$ *= fitted value in standard units*
x^* *= x value in standard units.*

7 Properties of Regression

Residuals

The Average of Residuals

An important aspect of the regression line is that on average, its deviations (also called **residuals**) cancel out. If residuals did not have this property, then the regression line would not be a good method of approximation because it may slightly overestimate/underestimate the properties of a variable. A simple calculation shows why the residuals of regression lines exhibit this quality. We start with the definition of average applied to residuals:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (a^*x_i + b^*))$$

where $y_i - (a^*x_i + b^*)$ is the the definition of a residual. Since the fitted values are on the regression line, we can assume the following values for the slope and the intercept: $a^* = \frac{ro_y}{o_x}$ and $b^* = \bar{y} - a^*\bar{x}$. This will allow us to simply our equation further:

$$\begin{aligned}
&= \tfrac{1}{n}\sum_{i=1}^{n}(y_i - (a^*x_i + \bar{y} - a^*\bar{x})) \\
&= \tfrac{1}{n}\sum_{i=1}^{n}[(y_i - \bar{y}) - a^*(x_i - \bar{x})] \rightarrow ① \\
&= \tfrac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}) - \tfrac{a^*}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \;=\; 0
\end{aligned}$$

because the averages of the deviations of both **y** and **x** are 0.

The Standard Deviation of Residuals

Another important aspect of a regression line is the standard deviation its residuals. This is the quantity that tells you how accurate is your regression line and gives us a tool to measure the rough size of error. In order to find this value, we can start by finding the

variance of residuals, where:

Variance = Mean square of deviations from average = $\frac{1}{n}\sum_{i=1}^{n}[y_i-(a^*x_i+b^*)]^2$

Using ①,

$= \frac{1}{n}\sum_{i=1}^{n}[(y_i-\bar{y})-a^*(x_i-\bar{x})]^2$

$= \frac{1}{n}\sum_{i=1}^{n}[(y_i-\bar{y})^2-2a^*(y_i-\bar{y})(x_i-\bar{x})+(a^*)^2(x_i-\bar{x})^2]$

Applying summation to each term in square bracket,

$= \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2-\frac{1}{n}\sum_{i=1}^{n}2a^*(y_i-\bar{y})(x_i-\bar{x})+\frac{1}{n}\sum_{i=1}^{n}(a^*)^2(x_i-\bar{x})^2$

a^* is constant hence,

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2-\frac{2a^*}{n}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})+\frac{(a^*)^2}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= \quad A \quad - \quad (2a^*)\,B \quad + \quad (a^*)^2\,C \quad \rightarrow ②$$

By definitions of standard deviations and correlation, we get,

$Variance\ of\ y = \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 = A$

$Variance\ of\ x = \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = C$

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

Substituting the above values in ②

$= \sigma_y^2-(2a^*)(r\ \sigma_x\ \sigma_y)+(a^*)^2\sigma_x^2$

We have established that $a^* = \frac{r\sigma_y}{\sigma_x}$, thus,

$= \sigma_y^2-(2\ \frac{r\ \sigma_y}{\sigma_x})(r\ \sigma_x\ \sigma_y)+(\frac{r\ \sigma_y}{\sigma_x})^2\ \sigma_x^2$

$$= \sigma_y^2 - 2\, r^2\sigma_y^2 + r^2\sigma_y^2$$

$$= \sigma_y^2 - r^2\sigma_y^2$$

$$= (1 - r^2)\, \sigma_y^2$$

Thus, the variance of residuals = $(1 - r^2)\, \sigma_y^2$

$$\Rightarrow \sigma_{res} = \sqrt[2]{variance\ of\ residuals} = \sqrt[2]{1 - r^2}\ \sigma_y$$

From the above equation, we observe the following:

Case 1A: When there is no x
In such case, the estimate is the average of y and rough size of error is the SD of y.

$$\Rightarrow SD\ of\ errors = \sigma_y$$

Case 1B: When there is no correlation between x and y, i.e., r = 0
In this case, we can use the formula we derived, to find the rough size of error.

$$\Rightarrow Rough\ size\ of\ error\ =\ SD\ of\ residuals = \sqrt[2]{1 - r^2}\ \sigma_y\ =\ \sqrt[2]{1 - 0^2}\ \sigma_y\ =\ \sigma_y$$

We observe that SD of residuals in both Case 1A and the current case is the same, therefore, we conclude that having no **x** and having no correlation between **x** and **y**, mathematically, are same.

Case 2: When there is a perfect correlation between x and y, i.e., r = 1 or -1
In this case, we again plug-in the value of r in the formula we derived for the rough size of error.

$$\Rightarrow Rough\ size\ of\ error\ =\ SD\ of\ residuals = \sqrt[2]{1 - r^2}\ \sigma_y\ =\ \sqrt[2]{1 - 1^2}\ \sigma_y\ =\ 0 \times \sigma_y\ =\ 0$$

Thus, we see that the SD of residuals is zero. This implies that our prediction has perfect accuracy, which is what we expect since the scatter plot is a straight line.

Definition 9 ***Mean, Deviation and SD of Residuals***
Mean:

$$\bar{r} = 0 \tag{7.1}$$

Deviation:

$$r_i - \bar{r} = (y_i - \bar{y}) - a^*(x_i - \bar{x}) \tag{7.2}$$

SD:

$$\sigma_{res} = \sqrt[2]{1 - r^2}\ \sigma_y \tag{7.3}$$

Fitted Values

The Average of Fitted Values

The definition of residual is given as

$$r_i \ = \ y_i - \hat{y_i}$$

where y_i is the acutal value of y and $\hat{y_i}$ is the fitted value or the expected value of y calculated from the line of best fit.

We can rearange the residual equation to solve for the fitted value :

$$\hat{y_i} \ = \ y_i - r_i$$

Then the mean of $\hat{y_i}$ can be simply written as:

$$\begin{aligned}
\bar{\hat{y}}_i &= \frac{1}{n}\sum_{i=1}^{n}(y_i - r_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} r_i \\
&= \bar{y} - \bar{r} \\
&= \bar{y} - 0 \qquad \text{Since } \bar{r} \text{ is } 0 \\
\bar{\hat{y}}_i &= \bar{y}
\end{aligned}$$

The Standard Deviation of Fitted Values

The standard deviation of the fitted values will tell us how far we can expect our fitted values to be from the average of the fitted values. We can calculate the SD of the fitted values by starting with the equation of the fitted line we established earlier:

$$\hat{y_i} = a^* x_i + b^*, \text{ where } a^* = \frac{r\sigma_y}{\sigma_x} \text{ and } b^* = \bar{y} - a^*\bar{x}$$

We want to use this equation to derive the SD of $\hat{y}$, so we will turn the left side of the equation into the definition of standard deviations step by step. First, lets subtract $\bar{\hat{y}}$ from both sides of the equation, since the standard deviation is calculated using the distance between each value and the mean.

$$\begin{aligned}\hat{y}_i &= a^* x_i + b^* \\ \hat{y}_i - \bar{\hat{y}} &= a^* x_i + b^* - \bar{\hat{y}} \\ &= a^* x_i + b^* - \bar{y}\end{aligned}$$

In the last step, we substituted $\bar{y}$ for $\bar{\hat{y}}$ as we established in the previous section that they are equal. Now let's substitute in for b^*:

$$\begin{aligned}\hat{y}_i - \bar{\hat{y}} &= a^* x_i + b^* - \bar{\hat{y}} \\ &= a^* x_i + b^* - \bar{y} \\ &= a^* x_i + \bar{y} - a^* \bar{x} - \bar{y} \\ \hat{y}_i - \bar{\hat{y}} &= a^*(x_i - \bar{x})\end{aligned}$$

Equation 2.1 gives us the deviation from average. We want the *mean squared deviation*, so let's square both sides and take the mean:

$$\begin{aligned}\hat{y}_i - \bar{\hat{y}} &= a^*(x_i - \bar{x}) \\ (\hat{y}_i - \bar{\hat{y}})^2 &= a^{*2}(x_i - \bar{x})^2 \\ \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 &= \frac{1}{n}\sum_{i=1}^{n} a^{*2}(x_i - \bar{x})^2 \\ &= \frac{a^{*2}}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\end{aligned}$$

Now taking the square root of both sides to get to the *root mean squared deviation*, we get:

$$\begin{aligned}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2} &= \sqrt{\frac{a^{*2}}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= |a^*|\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \sigma_{\hat{y}} &= |a^*|\sigma_x\end{aligned}$$

In the last step, we substituted in the definition for standard deviation. Now substituting in our previous definition of a^*, we get to our value for the standard deviation of the fitted values:

$$\begin{aligned}\sigma_{\hat{y}} &= |a^*|\sigma_x \\ &= \frac{|r|\sigma_y}{\sigma_x}\sigma_x \\ \sigma_{\hat{y}} &= |r|\sigma_y\end{aligned}$$

We find that the standard deviation of $\hat{y}$, our fitted line, is r, the correlation coefficient, multiplied by the standard deviation of y, our actual values.

Definition 10 ***Mean, Deviation, SD of Fitted Values***

Mean:

$$\bar{\hat{y}}_i = \bar{y} \tag{7.4}$$

Deviation:

$$\hat{y} - \bar{\hat{y}} = a^*(x_i - \bar{x}) \tag{7.5}$$

SD:

$$\sigma_{\hat{y}} = |r|\sigma_y \tag{7.6}$$

Final note, the formula derived from the previous step can be used to derive R-squared because R-squared is always a positive value. We can begin by rewriting the formula to solve for |r|:

$$|r| = \frac{\sigma_{\hat{y}}}{\sigma_y}$$

then square both sides:

$$r^2 = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2}$$

We discover that the R-squared for multiple regression is the expected variance of y divided by the actual variance of y.

Correlation between Regression Statistics

In the previous sections, we defined and explored the properties of the original values x, fitted values y, and residual values r and we understand these three elements individually.

In regression models, it is important to understand the relationship between these three elements. For instance, x is related to y and we defined the relationship as we found the correlation in an earlier section. Also, the original values of y are related to the fitted values of y. Not only this, the correlation between the original and fitted values of y has to be positive as we want $\hat{y_i}$ to increase as y_i increases for our predictions to be accurate. Hence, in this section, we will examine the relationships between these three elements and their importance in regression models.

Here's what we've established in the previous sections so far:

- The correlation between x and y:

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

- The equation of the regression line

$$\begin{aligned}\hat{y_i} &= a^*x_i + b^*, where\\ a^* &= slope\ of\ regression\ line\\ &= \frac{r\sigma_y}{\sigma_x}\\ b^* &= intercept\ of\ regression\ line\\ &= \bar{y} - a^*\bar{x}\end{aligned}$$

- The average of the residuals, $\bar{r} = 0$
- The standard deviation of the residuals, $\sigma_{res} = \sqrt[2]{1-r^2}\sigma_y$
- The average of fitted values, $\bar{\hat{y}_i} = \bar{y}$
- The standard deviation of the fitted values, $\sigma_{\hat{y}} = |r|\sigma_y$
- The deviations of the residuals, $r_i - \bar{r} = (y_i - \bar{y}) - a^*(x_i - \bar{x})$
- The deviations of the fitted values, $\hat{y_i} - \bar{\hat{y}} = a^*(x_i - \bar{x})$

Correlation between x and the Residuals

Objective: Prove the correlation between **x** and residual values r_i is 0.

Proof: Now, let's begin with the definition of the correlation coefficient between r and x.

$$\begin{aligned}r(r_i,x) &= correlation\ coefficient\\ &= \frac{1}{n}\sum_{i}^{n}\frac{(r_i-\bar{r})}{\sigma_r}\frac{(x_i-\bar{x})}{\sigma_x}\end{aligned}$$

In order to prove that the correlation between the residuals and x is equal to zero, we only need to look at the numerator of the above equation. The following steps show that the numerator, r_{num} is equal to zero.

$$\begin{aligned} r_{num} &= \frac{1}{n}\sum_{i}^{n}(r_i - \bar{r})(x_i - \bar{x}) \\ &= \frac{1}{n}\sum_{i}^{n}(y_i - \bar{y})(x_i - \bar{x}) - a^*\frac{1}{n}\sum_{i}^{n}(x_i - \bar{x})^2 \end{aligned}$$

Notice, here the first term $\frac{1}{n}\sum_{i}^{n}(y_i - \bar{y})(x_i - \bar{x})$ is equivalent to $r\sigma_x\sigma_y$. Remember that $a^* = r\frac{\sigma_x}{\sigma_y}$ and the variance of x is ${\sigma_x}^2 = \frac{1}{n}\sum_{i}^{n}(x_i - \bar{x})^2$.

Now, using the identities we defined before, we find out that the difference of equivalent values is zero.

$$\begin{aligned} r_{num} &= r\sigma_x\sigma_y - r\frac{\sigma_y}{\sigma_x}\sigma_x^2 \\ &= 0 \end{aligned}$$

Because the numerator r_n is equal to zero, the entire value of $r(r_i, x)$ is equal to zero.

$$r(r_i, x) = 0$$

What this means intuitively is that for any given x, we cannot predict what the residual will be. If there is some sort of pattern for the residuals it means that your regression line is either miscalculated (there exists some line that fits the cloud better) or wrong (perhaps a quadratic would fit better than a straight line). You should expect your residual plot to be flat (correlation, slope of 0).

Correlation Between Fitted Values and Residuals

Just as there is no correlation between x and the set of residual values, there is also no correlation between the sets of fitted values and residual values. If a regression model presents a correlation between fitted values and residuals other than 0, then it is invalid because a given fitted value cannot possibly predict what its corresponding residual is.

Objective: Prove the correlation between the fitted values and the residuals is 0.

As in part 1) let's begin with the definition of the correlation coefficient but this time between the fitted values of y, $\hat{y}_i$ and the residual values, r_i.

$$r(\hat{y}_i, r_i) = \frac{1}{n}\sum_{i=1}^{n}(\frac{\hat{y}_i - \bar{\hat{y}}}{\sigma_{\hat{y}}})(\frac{r_i - \bar{r}}{\sigma_r})$$

Now, recall that $\hat{y}_i$ is a linear function of x_i,

$$\hat{y}_i = a^* x_i + b^*$$

In one of the previous sections, we saw that if any variable, say A is a linear function of another variable B, then the absolute value of the correlation between A and a third variable C is the same as the correlation between B and the third variable C. The sign of the correlation coefficient depends on the sign of a^*, the slope of the regression equation.

Since, $\hat{y}_i$ is a linear function of x_i, the correlation between $\hat{y}_i$ and r_i is the same as the correlation between x_i and r_i. In part 1), we proved that the correlation between x_i and r_i is 0. Hence, the correlation between $\hat{y}_i$ and r_i is also 0.
Thus,

$$r(\hat{y}_i, r_i) = 0$$

Correlation between the fitted values and the original values of y

So far, we have proved that the correlation between x and the residual values as well as between the fitted values of y and the residuals is 0.

Objective: To establish the correlation between $\hat{y}_i$, the fitted values of y and y, the original values of y.

We define the correlation coefficient between $\hat{y}_i$ and y as,

$$r(\hat{y}, y) = \frac{1}{n}\sum_{i}^{n}\frac{(y_i - \bar{y})}{\sigma_y}\frac{(\hat{y}_i - \bar{\hat{y}_i})}{\sigma_{\hat{y}}}$$

As discussed in part 2), the absolute value of the correlation between a linear function of x_i and a variable C, is the same as the correlation between x_i and C. Also, we know that $\hat{y}_i$ is a linear function of x_i. Hence, the absolute value of the correlation between $\hat{y}_i$ and y_i is the same as the correlation between x_i and y_i.

While defining the correlation coefficient, we proved that the correlation between x_i and y_i is $r(x, y)$. Hence, the absolute value of the correlation between $\hat{y}_i$ and y_i is also $r(x, y)$.

Something to keep in mind is that the correlation coefficient, $r(\hat{y}, y)$ is always positive because any increment in y corresponds to an increment in $\hat{y}$. Intuitively, the bigger the y, the bigger we want the estimate, $\hat{y}$ to be. Hence, the original values of y, y_i, and the fitted values of y, $\hat{y}_i$ must be positively correlated for the regression estimate to be appropriate.

Thus, the correlation between $\hat{y}_i$ and y_i is the absolute value of the correlation between x_i and y_i.

$$r(\hat{y}, y) = |r(x, y)|$$

8 Further Steps

g

Regression Lines without Intercepts

In mathematics, the y-intercept in a linear regression analysis appears to be a simple mathematical concept since it is the value at which the fitted line crosses the y-axis. Paradoxically, while the value is generally simple, it is crucial to comprehend a deeper level of understanding when the intercept is equal to zero with regression models. For example, when two variables are proportional, than a change in one of them is accompanied by a proportional in the other term. To express the statement that y is proportional to x, it can be written as:

$$y = ax$$

This is the equation of a straight line with no intercept. It passes through the origin.

Regression Line with Intercept Zero

In a linear regression model with a zero intercept, given the variables $(x_1, y_1), \ldots, (x_n, y_n)$, we must best fit $y = ax_1$. Dropping the y-intercept in a regression model forces the regression line to go straight through the origin. As you may recall, the original equation is as follows:

$$f(a) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

To find the slope of the regression line that minimizes error, we will substitute our value zero for b_a into our original equation. With the substitution in, we will now have:

$$f(a) = \sum_{i=1}^{n} (y_i - (ax_i + 0))^2$$

To minimize the error, we will now take the derivative on both sides of the equation.

$$\frac{d}{da} f(a) = \frac{d}{da} \sum_{i=1}^{n} (y_i - ax_i)^2 = \sum_{i=1}^{n} 2(y_i - ax_i)(-x_i)$$

In order to solve for the a^* that minimizes the error, we will need to set the derivative of our equation to 0. We do this because this will give us our minimal error. Therefore, after differentiating with respect to a, we have the following:

$$0 = \sum_{i=1}^{n} 2(y_i - a^* x_i)(-x_i)$$

We now combine the values to make the right hand side easier to visualize:

$$0 = \sum_{i=1}^{n} (-2y_i x_i + 2a^* x_i^2)$$

Since the summation is on the outside, we then distribute this to each term inside the function and bring out any constants we have, which in this case will be a^*.

$$0 = \sum_{i=1}^{n} (-2y_i x_i) + a^* \sum_{i=1}^{n} 2x_i^2$$

Separate the equation in order to isolate the variable a on on side. We will now have:

$$\sum_{i=1}^{n} 2y_i x_i = a^* \sum_{i=1}^{n} 2x_i^2$$

In order to isolate a^*, we divide $\sum_{i=1}^{n} 2y_i x_i$ by $\sum_{i=1}^{n} 2x_i^2$. Since the constant 2 is on both sides, it can be eliminated before division occurs. As a result, the final answer is:

$$a^* = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

The general method of minimizing the mean squared error that was used to find the equation of the regression line can also be applied in solving other problems. For example, you may want to fit a line that has a y-intercept of zero. In order to do this, adjusting the general method so that the line goes through the origin is essential in solving the problem.

R^2 in Multiple Linear Regression

Before we delve into the topic, let's begin with an example. Say you are a weightlifter, and you participate in multiple different excercises. For example, you participate in deadlifting, squatting, and bench pressing. With multiple linear regression, we want to see if there is a way to predict how much you could bench press given your ability to deadlift and squat. Moreover, we can include more and more excercises in order to see if we can get a better predictor of the amount that you can bench press.

We will step away from the weight lifting problem momentarily. From examining the relationships between the variances of the residuals, fitted variables, and actual values in single variable linear regression, we get the relationship:

$$\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_r^2 \tag{8.1}$$

We can now consider multiple linear regression, where the model is given by an equation of the form:

$$y = b_0 + b_1 x_1 + ... + b_k x_k \tag{8.2}$$

Now, for each variable included, we can calculate the correlation one variable at a time, however that will not tell us anything about the overall correlation between the model and the predicted values. Thus, we need to find a way to express multiple correlation, the total correlation of the model, and more importantly a method of determining how good our model predicts the data.

To do so, we will consider our first equation. What we see is that this relationship holds true no matter how many independent variables are included in the model. In other words:

Variance of y = Variance of residuals + Variance of fitted values

Note: There is a long mathematical process to prove this. Though it is valuable to understand, for the purposes of this discussion, it has been omitted from this text.

From (5.1), we know that $\sigma_{\hat{y}}^2$ must be less than σ_y^2, which means that $\frac{\sigma_{\hat{y}}^2}{\sigma_{\hat{y}}^2} < 1$. In linear regression, this fraction is defined as the R^2 value. Similarly, in multiple linear regression, this fraction is defined as the multiple-R^2 value. Going back to our first equation, we can do some algebra to find:

$$\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_r^2 \tag{8.3}$$

$$1 = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2} + \frac{\sigma_r^2}{\sigma_y^2} \tag{8.4}$$

$$1 = \text{multiple-R}^2 + \frac{\sigma_r^2}{\sigma_y^2} \tag{8.5}$$

$$\frac{\sigma_r^2}{\sigma_y^2} = 1 - \text{multiple-R}^2 \tag{8.6}$$

What we can conclude is that just as R^2 was an indication of how good the model fits the data in linear regression, so does multiple-R^2 indicate how good a model fits data in multiple linear regression. When multiple-R^2 is close to 1, σ_r^2 must be a small number, which means that the variance of the residuals is a small value and moreover that the model predicts well. On the other hand, when multiple-R^2 is close to 0, σ_r^2 must be a larger number, which means that the variance of the residuals is a large value and the model does not predict very well.

As an aside, there is a graphical analog that does exist such that the residuals are represented by orthogonal distances from predicted values in 3-space to a plane that is given by an equation of independent variables. This previous statement sounds overly complex, however we can tie it back into the weightlifting example to hopefully elucidate the idea.

Imagine an XYZ coordinate system such that the XZ plane is parallel to the floor (informally, the horizontal plane). On this plane, each point represents a coordinate corresponding to the amount you can deadlift and the amount you can squat. On the Y axis is the amount that you can bench press. Our goal is to use the points on the XZ plane to place a predicted point on the Y axis that is as close to the actual value.

It is important to note that when we say "on the Y axis", we are referring to a Y axis that intersects a point on the XZ plane. In other words, imagine a line perpendicular to the XZ plane that goes through any given point, and on that line is where the corresponding bench press value can be found. The residual in this example is the distance between the actual amount you can bench press, some point in the space, and the predicted amount you can bench press, a point collinear with the coordinate on the XZ plane used to predict the bench pressing value and the actual bench pressing weight that corresponds to that. By minimizing multiple-R^2, we are hoping to minimize all the distances between the predicted value and the actual value of bench pressing.

In summation, R^2, or multiple-R^2, is a prediction of how well a model fits the data regardless of how many independent variables are present. Furthermore, multiple-R^2 can be represented graphically as the distances between collinear points parallel to the Y axis in an XYZ coordinate space.

9 About the Authors

This is a book written for, and written by, students. This book is a product of the pilot "Probability and Mathematical Statistics in Data Science" class at UC Berkeley. Each chapter was co-written by groups of students, who consequently revised and finalized their sections. Authorships are mentioned below (in alphabetical order) :

Probability

Chapter 1: *An Introduction to Probability* Shreya Agarwal, Adith Balamurugan, Dibya Ghosh, Jiayi Huang, Andrew Linxie, Kyle Nguyen, Anthony Xian, Parth Singhal

Chapter 2: *Probability With and Without Replacement* Bryannie Bach, Aditya Gandhi, Edward Huang, Arvind Iyengar, Nishaad Navkal,Rohan Singh, Yu Xia

Chapter 3: *Averages and Variations* Thomas Anthony,Betty Chang, J. Weston Hughes, Rahil Mathur, Maxwell Weinstein, Ling Xie

Chapter 4: *Random Variables* Ani Adhikari

Regression

Chapter 5: *Correlation* Adith Balamurugan, Aditya Gandhi, Dibya Ghosh, Maxwell Weinstein

Chapter 6: *Linear Regression* Thomas Anthony, Jiayi Huang, Weston Hughes, Kyle Nguyen,Rohan Singh, Ling Xie

Chapter 7: *Properties of Regression* Shreya Agarwal, Betty Chang,Arvind Iyengar, Andrew Linxie, Nishaad Navkal, Parth Singhal, Yu Xia, Anthony Xian

Chapter 8: *Further Steps* Bryannie Bach, Rahil Mathur