# Math for Data Science

## Probability

Joanna Bieri DATA100

# Important Information

- Email: joanna_bieri@redlands.edu
- Office Hours take place in Duke 209 unless otherwise noted – Office Hours Schedule

# Today's Goals:

- Introduce the basics of probability
- Learn the mathematics of Joint and Union probabilities.
- Do an applied example

# (Review) Integral

$$\int G(t)\,dt = \lim_{N\to\infty} \sum_{i=1}^{N} G(t_i^*)\Delta t$$

1. The integral of a function is related to the area under the curve.
2. The solution to the integral is a family of functions where the $+c$ represents the starting value, or the value at some point.
3. The integral is the opposite of the derivative.
4. It is a type of fancy addition - we are adding up the area under the curve of a changing function.

# Probability

This lecture follows along with our book: Essential Math for Data Science, by Thomas Nield - Chapter 2

Probability is the study of uncertainty. It is the theoretical study of measuring certainty that an event will happen. People often think of examples like:

- probability of winning the lottery
- rolling two sixes on two dice
- predicting stock performance

It is foundational for the understanding of statistics, hypothesis testing, and machine learning.

Probability is related to how strongly we believe an event will happen and often we think of this as a percentage. "There is a 30% chance that the avocado I bought will be ripe"

# Notation

We express the idea of probability as

$$P(X)$$

where $X$ is the event of interest. P(X)=0.30 in my example above, we often use decimals instead of percents.

# Probability vs Likelihood vs Odds

In every day conversation probability and likelihood are basically used the same, but in mathematics things are more nuanced. Probability quantifies predictions on things yet to happen. Likelihood measures the frequency of events that already occurred.

"We use likelihood (the past) in the form of data to predict probability (the future)"

# Probability vs Likelihood vs Odds

Another thing to keep in mind is that probability is always between 0% and 100% for exclusive events, we can't be more than 100% certain that an event will occur! Mathematically, for our avocado example, we might write: If $P(X) = .3$ then $P(notX) = 1 - .3 = .7$, because either the avocado is ripe or it is not.

# Probability vs Likelihood vs Odds

Sometimes we talk about odds instead of probabilities. The odds that somethings happens is the ratio of the probability that it will happen to the probability that it won't.

$$O(X) = \frac{P(X)}{P(notX)} = \frac{P(X)}{1 - P(X)}$$

In my example above I have a 3/7 or 3:7 "three to seven" odds that my avocado is ripe.

# Probability vs Statistics

Probability and statistics are closely related but they each have their own distinctions. Probability is the purely theoretical study of how likely an event is to happen. We do not require any data to study probability. Statistics uses data to discover probabilities and provides tools to explore and describe data. Without data there are no statistics.

I strongly encourage you to take MATH 111 to learn more about statistics and hypothesis testing!

# Mathematics and Probability

In the simple case (marginal probabilities) you are working just with a single probability and the ideas are fairly straightforward. Either my avocado is ripe or not. But real life is rarely this simple! Isn't there so much more that goes into choosing an avocado?!?!? What are the chances that I get two or more ripe avocados in a row?

# Joint Probabilities

Joint probabilities help us understand events that occur together but are independent. For example what is the probability that I flip a fair coin and get two heads in a row?

- A coin has two sides and for a fair coin the probability of heads is $P(H) = \frac{1}{2}$ (marginal probability)
- The outcome of my first flip does not change the probability of my next flip (independent)

# Joint Probabilities

**Joint Probability**

$$P(A \; and \; B) = P(A) \times P(B)$$

So the probability that I flip two heads is
$P(H \; and \; H) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = 0.25$

# Joint Probabilities

Why does this make sense? Why multiply the probabilities together? It helps to think of all the possible combinations of outcomes:

- H1 T2
- H1 H2
- T1 T2
- T1 H2

We can review these possible outcomes to see there were four possibilities and only one of them was the outcome I wanted (H1 H2). So the multiplication (or **the product rule or probability**) is counting up these combinations for us.

# Union Probability

Sometimes we want to know what is the probability of getting one outcome OR another assuming those outcomes are independent. This is the **Union Probability**. We will explore the union probability from two examples depending on whether the events are mutually exclusive or not.

# Union Probability

**Example 1**, what is the probability of flipping heads OR tails in a single flip? Well since we can't flip a coin and get BOTH heads a tails we call these events **mutually exclusive**. This means we can just add up the probabilities to see the probability of getting one or the other:

$$P(H \text{ or } T) = P(H) + P(T) = \frac{1}{2} + \frac{1}{2} = 1$$

# Union Probability

Does this make sense? YES! Since there were only two outcomes, heads or tails, the possibility of getting one or the other is 100%

Now what happens if events are **nonmutually exclusive**?

# Union Probability

**Example 2**, say you have two fair six-sided die - one RED and one GREEN - and want to know what is the chance of rolling a 1 on the RED OR a 2 on the GREEN. If we just had ONE die then this would be two mutually exclusive events, since you can't roll both a 1 and a 2 on a single die. And we would just say:

$$P(1 \; or \; 2) = P(1) + P(2) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

# Union Probability

But, because we have two die, **now we can roll both a 1 and a 2 at the same time**. Lets write down all the possible combinations of this dice roll. There are a total of 36 possibilities (6 on each dice) and we can count up the ways we can achieve the 1 RED OR 2 GREEN:

# Union Probability

| RED | GREEN | |
|---|---|---|
| 1 | 1 | * |
| 1 | 2 | * |
| 1 | 3 | * |
| 1 | 4 | * |
| 1 | 5 | * |
| 1 | 6 | * |
| 2 | 1 | |
| 2 | 2 | * |
| 2 | 3 | |
| 2 | 4 | |
| 2 | 5 | |
| 2 | 6 | |

# Union Probability

| RED | GREEN | |
|---|---|---|
| 3 | 1 | |
| 3 | 2 | * |
| 3 | 3 | |
| 3 | 4 | |
| 3 | 5 | |
| 3 | 6 | |
| 4 | 1 | |
| 4 | 2 | * |
| 4 | 3 | |
| 4 | 4 | |
| 4 | 5 | |
| 4 | 6 | |

# Union Probability

| RED | GREEN | |
|---|---|---|
| 5 | 1 | |
| 5 | 2 | * |
| 5 | 3 | |
| 5 | 4 | |
| 5 | 5 | |
| 5 | 6 | |
| 6 | 1 | |
| 6 | 2 | * |
| 6 | 3 | |
| 6 | 4 | |
| 6 | 5 | |
| 6 | 6 | |

# Union Probability

There were 11 times (marked with *) in the table that we had a 1 RED or a 2 GREEN out of 36 possible combinations so the probability is $P(1\ or\ 2) = 11/36$.

Notice that our old calculation for the mutually exclusive case overcounts the possibilities

$$P(1\ or\ 2) = P(1) + P(2) = \frac{1}{3} \sim 0.33333 > \frac{11}{36} = 0.305556$$

# Union Probability

What happened? When we just add up the probabilities, we are double counting the case where 1 RED and 2 GREEN happened together. Thinking about this, what is $P(1) + P(2)$ saying?

- There is 1 out of 6 ways we could have rolled a 1 RED - and the other dice could be anything (six options)
- These is 1 out of 6 ways we could have rolled a 2 GREEN - and the other dice could be anything (six options)
- So we have a total of 12 ways we could get the 1 or the 2.
- AND $12/36 = 1/3$

# Union Probability

But look at how we actually counted the outcomes in the table above! When we counted the case where we rolled a 1 RED, we also already counted the case where we rolled BOTH 1 RED and 2 GREEN and we didn't count this one again when we looked for all the cases where we rolled a 2 GREEN. This is an example double counting the joint probability.

# Union Probability

To get the right answer we need to remove the join probability from the union calculation. So the real rule for the probability of $A$ OR $B$ is:

$$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$$

This is also called **the sum rule of probability**. In our case above:

$$P(1\ or\ 2) = P(1) + P(2) - P(1\ and\ 2) =$$
$$\frac{1}{6} + \frac{1}{6} - \frac{1}{6} \times \frac{1}{6} = 1/3 - 1/36 = 11/36$$

This matches the way we counted before. The general rule also works for mutually exclusive events - like the coin toss above, because the joint probability of getting heads and tails on a single coin is zero so $P(H\ and\ T) = 0$.

# You Try

Assume you have a single fair six-sided die. Answer the following questions:

- What is the probability of rolling a six?
- What is the probability of rolling a four?
- What is the probability of rolling a six and then a four?
- What is the probability of rolling either a six or a four?

# You Try

Assume you have THREE fair twenty-sided die, one RED, one BLUE, and one GREEN. Answer the following questions.

- What is the probability of rolling 10 RED and 10 BLUE?
- What is the probability of rolling 10 RED or 10 BLUE?
- What is the probability of rolling 10 RED and 10 BLUE and 10 GREEN?

# Using Statistics to find Probabilities

Our next goal is to use historical data to help us find some probabilities. Here is historical data that was scraped from the website:

https://www.laalmanac.com/weather/we08aa.php

Accessed on 3/11/25. This website records the amount of rainfall in Downtown Los Angeles from 1877-Present. The We will do some statistics on this data to generate probabilities and then ask questions about those probabilities.

# Using Statistics to find Probabilities - Data Cleaning

In order to prepare this data for you I made the following changes:

- I the original data the letter T was used to represent a trace amount of rain. For our purposes I read this to be zero rain.
- I removed one data outlier (—2) in the month of September that records the fact that this month had a lot of missing data. I reset this to 0 which means that our Sep probability might be a but low.
- I replaced the rainfall amounts with 0 to represent the case where there was no measurable rain and 1 for when there was some measurable rain.

# Generating Probabilities for Each Month

Based on this historical data. Let's find the probability that there will be rain for each month of the year.

**NOTE** We are making some strong assumptions about this data! For example we are assuming that each of these observations is independent, meaning that they are kindof like a flip of the coin. This is not really true if we think of the full complexity of weather over time.

# Using Statistics to find Probabilities

See the student notebook for a better look at the data and code!

| | Season | Jul | Aug | Sep | Oct | Nov | Dec | Jan | Feb | Mar | A |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2023-2024 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 2022-2023 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2021-2022 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 3 | 2020-2021 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| 4 | 2019-2020 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 142 | 1881-1882 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 143 | 1880-1881 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 144 | 1879-1880 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 145 | 1878-1879 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 146 | 1877-1878 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

# You Try

1. What is the probability of rain $P(R)$ and the probability that there is no rain $P(not\ R)$ for each month in the season? Fill in the table below:

| Month | $P(R)$ | $P(not\ R)$ |
|---|---|---|
| Jul | 0.1497 | 0.8503 |
| Aug | 0.2857 | 0.7143 |

2. Which of these numbers should sum to one?

# You Try

3. What is the probability that there is rain in both August and September?

$$P(R_{august} \; and \; R_{september})$$

4. What is the probability that there is not rain in both August and September? You calculated these numbers above, you can just use them.

$$P(not \; R_{august} \; and \; not \; R_{september})$$

# You Try

5. What is the probability that there is rain in August or September?

$$P(R_{august} \text{ or } R_{september})$$

6. Using the numbers you calculated above compare the results:

- $$P(\text{not } R_{august} \text{ and not } R_{september})$$
- $$1 - P(R_{august} \text{ or } R_{september})$$

does it make sense that the probability of no rain in August and no rain is September is the same as one minus the probability of rain in August or September?

# You Try

7. Write in words what these results mean.

*You can check your work with my results in the lecture notes!*