

[Home](#)[About](#)[Contact](#)

[HOME](#) › [PROBABILITY THEORY](#) › [BAYES' THEOREM](#) › [BAYES' THEOREM: AN INFORMAL DERIVATION](#)

# Bayes' Theorem: An Informal Derivation

Posted on **FEBRUARY 28, 2016** Written by **THE CTHAEH**  **8 COMMENTS**

[g+ Share](#)[f Share](#)[t Tweet](#)[p](#)

0

[in](#)

0

[j](#)

0



If you're reading this post, I'll assume you are familiar with Bayes' theorem. If not, take a look at [my introductory post on the topic](#).

Here I'm going to explore the intuitive origins of the theorem. I'm sure that after reading this post you'll have a good feeling for where the theorem comes from. I'm also sure you will find the simplicity of its mathematical derivation impressive. For that, some familiarity with sample spaces (which I discussed [in this post](#)) would come in handy.

So, what does Bayes' theorem state again?

(BAYES' THEOREM)

$$P(\text{Event-1} \mid \text{Event-2}) = \frac{P(\text{Event-2} \mid \text{Event-1}) \times P(\text{Event-1})}{P(\text{Event-2})}$$

Here's a quick reminder on the terms of the equation:

- $P(\text{Event-1})$ : **Prior probability**
- $P(\text{Event-2})$ : **Evidence**
- $P(\text{Event-2} \mid \text{Event-1})$ : **Likelihood**
- $P(\text{Event-1} \mid \text{Event-2})$ : **Posterior probability**

Put simply, Bayes' theorem is used for updating prior probabilities into posterior probabilities after considering some piece of new information (that is, some piece of evidence). The exact way the updating process takes place is given by the relationship asserted by the theorem. Namely, the posterior probability is obtained after multiplying the prior probability by the likelihood and then dividing by the evidence.

But have you wondered where this exact relationship comes from? The easiest way to answer the question is by first defining joint probabilities and showing how the theorem naturally pops out.

### Table of Contents



1. Joint probabilities and joint sample spaces in the context of Bayes' theorem
  - 1.1. An alternative look at joint probabilities
2. The incredibly simple derivation of Bayes' theorem
3. The intuition behind the theorem
  - 3.1. The likelihood
  - 3.2. The evidence
4. Summary

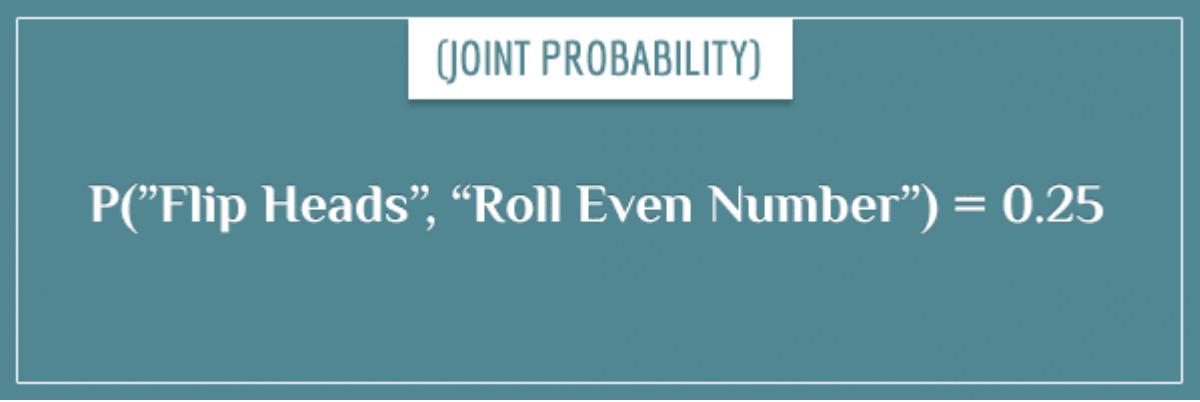
## Joint probabilities and joint sample spaces in the context of Bayes' theorem

The joint probability of two events is simply the probability that the two events will both occur. The most common mathematical notation for expressing a joint probability is:

$$P(\text{Event-1}, \text{Event-2})$$

Notice the difference with the notation for conditional probabilities where there is a vertical line between the two events (instead of a comma).

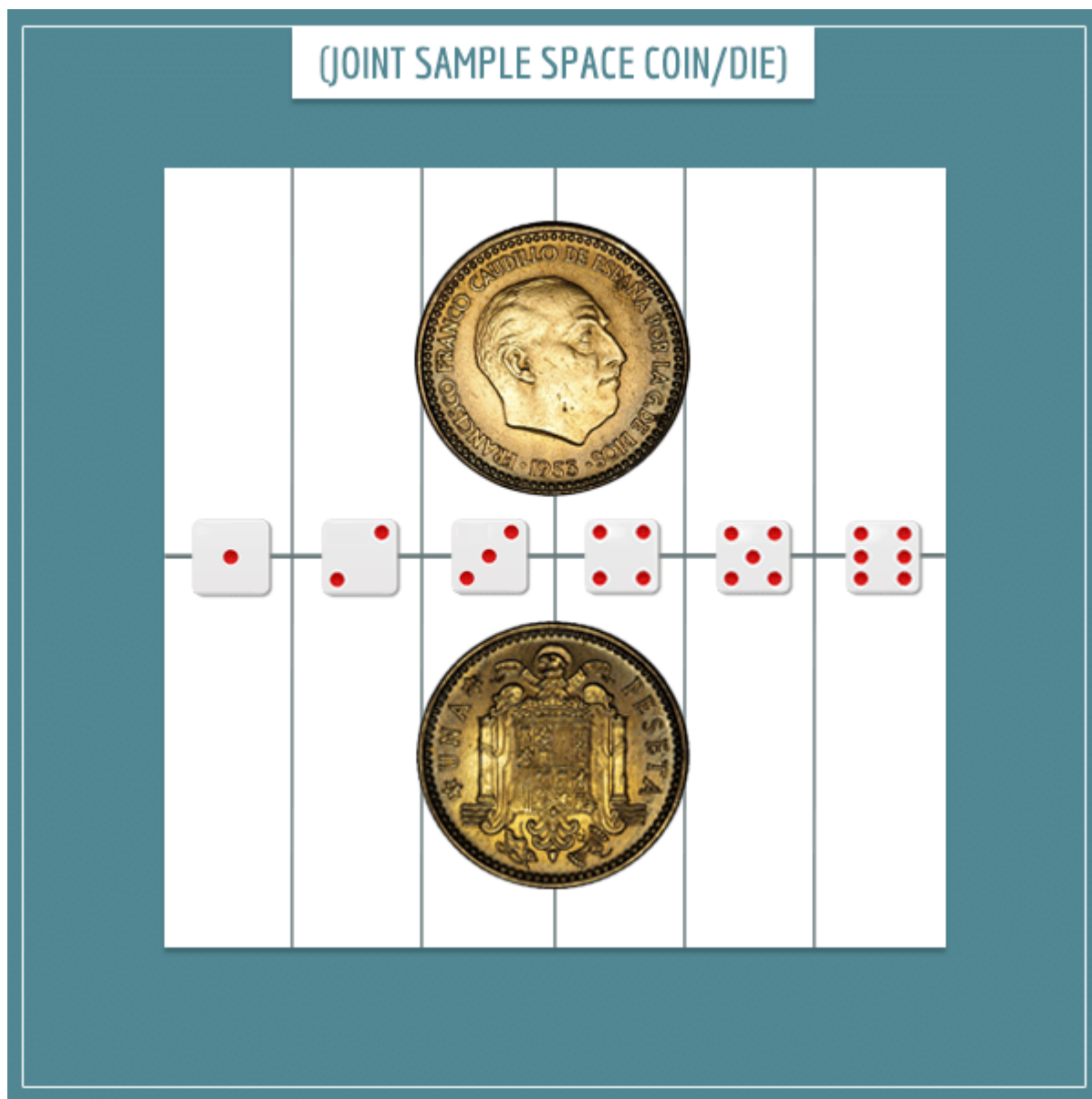
Imagine you're simultaneously flipping a coin and rolling a die. A joint probability example would be the probability of flipping heads and rolling an even number in a single repetition:



(JOINT PROBABILITY)

$$P(\text{"Flip Heads", "Roll Even Number"}) = 0.25$$

You can get to this probability from a graphical analysis of the sample space. As a reminder, the sample space of a random process is the set of all possible outcomes of that process. The joint sample space of two random processes is all possible combinations of outcomes of the two processes. Take a look:

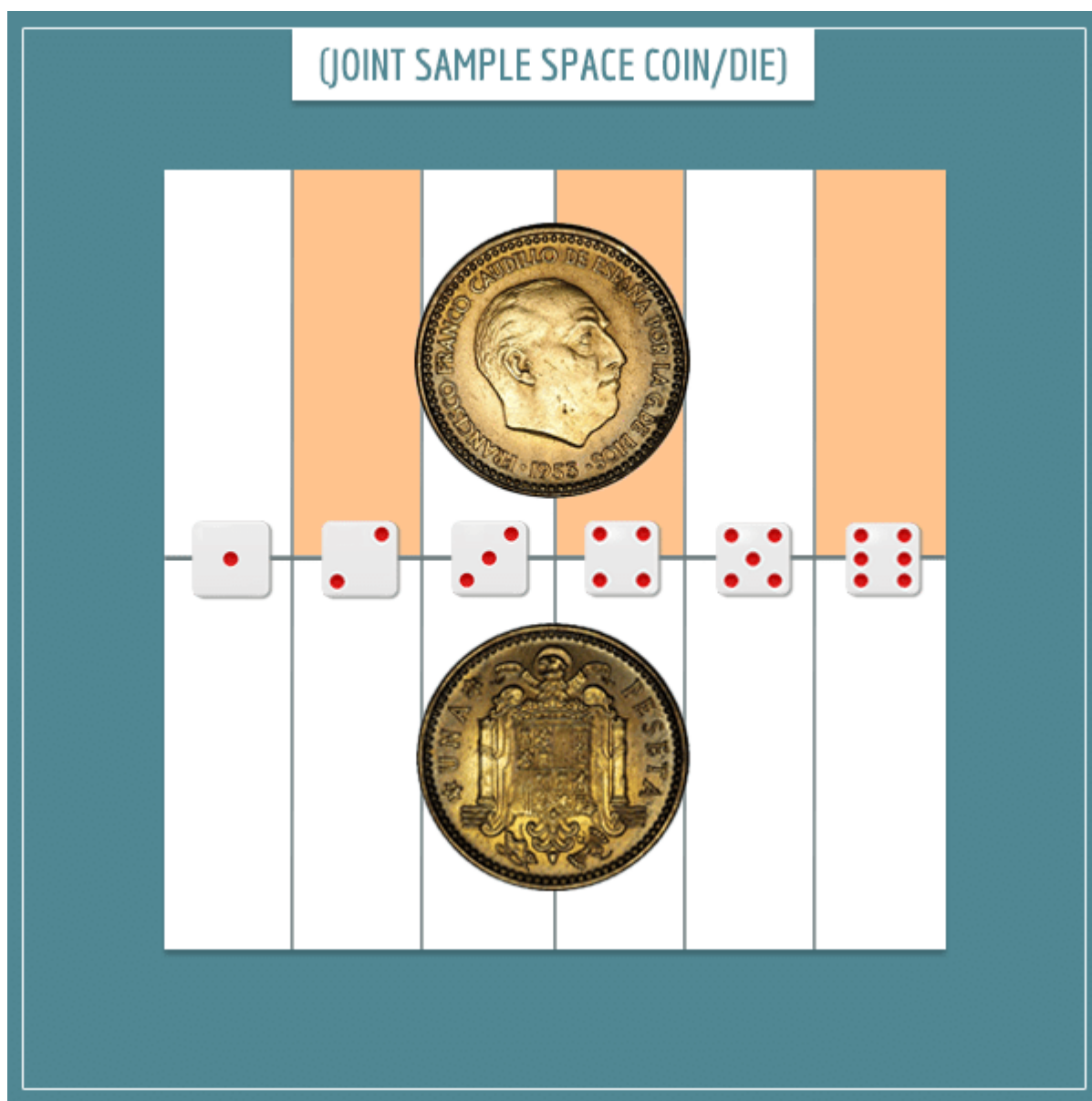


The horizontal line divides the square into 2 equal parts which represent the outcome of the coin flip.

The vertical lines further divide the square into 6 equal parts which represent the possible outcomes of rolling the die. You see that when the two sample spaces are superimposed on each other the square is divided into 12 equal rectangles. Each rectangle represents a possible combination, like:

- “Flip Heads”, “Roll 1”
- “Roll 4”, “Flip Heads”
- “Flip Tails”, “Roll 3”
- Etc.

The marked rectangles in the image below represent the event of flipping heads and rolling an even number:



First, the rectangles are from the upper half of the square which stand for the probability of flipping heads. Second, only those rectangles associated with rolling one of the three even numbers are marked. As you can see, 3 out of the 12 (or  $1/4$ ) of the rectangles satisfy the definition of the event.

The sum of the areas of the marked rectangles is the area of the intersection of the two events. You can calculate the joint probability by considering the fraction of the total area that it covers:

$$P(\text{"Heads"}, \text{"Even Number"}) = \frac{1}{4} = 0.25$$

## An alternative look at joint probabilities

Here's a very important relationship between probabilities, conditional probabilities, and joint probabilities:

(JOINT PROBABILITY)

$$P(\text{Event-1, Event-2}) = P(\text{Event-1}) \times P(\text{Event-2} \mid \text{Event-1})$$

The joint probability of two events is the probability of the first event times the conditional probability of the second event, given the first event. Why? This part is slightly tricky, so arm yourself with your abstract reasoning skills.

Remember, the joint probability of two events is the probability that both events will occur. For that, first *one of the events* needs to occur, which happens with probability  $P(\text{Event-1})$ .

Then, once you *know* Event-1 has occurred, what is the probability that Event-2 will *also* occur? Naturally, this is the conditional probability  $P(\text{Event-2} \mid \text{Event-1})$ : the probability of Event-2, given Event-1.

The conditional probability is simply the fraction of the probability of Event-1 which also allows Event-2 to occur. In other words, to get the joint probability of two events, you need to take the fraction of  $P(\text{Event-1})$  that is consistent with the occurrence of Event-2. To take the fraction of something really means to multiply that “something” by the fraction:

$$P(\text{Event-1, Event-2}) = P(\text{Event-1}) \cdot P(\text{Event-2} \mid \text{Event-1})$$

The joint probability from the previous example becomes:

$$P(\text{"Heads", "Even Number"}) = P(\text{"Heads"}) \cdot P(\text{"Even Number"} \mid \text{"Heads"})$$



You can infer the values of the two terms on the right-hand side from the joint sample space:

$$P(\text{"Heads"}) = 0.5$$

$$P(\text{"Even Number"} \mid \text{"Heads"}) = 0.5$$

Then:

$$P(\text{"Heads"}, \text{"Even Number"}) = 0.5 \cdot 0.5 = 0.25$$

The result is the same as with the previous approach. Pretty neat.

## The incredibly simple derivation of Bayes' theorem

Now that you've convinced yourself about the last relationship, let's get down to business.

First, notice that the relationship can be stated in two equivalent ways:

$$P(\text{Event-1}, \text{Event-2}) = P(\text{Event-1}) \cdot P(\text{Event-2} \mid \text{Event-1})$$

$$P(\text{Event-1}, \text{Event-2}) = P(\text{Event-2}) \cdot P(\text{Event-1} \mid \text{Event-2})$$

The intuition behind this symmetry is that the order of the events isn't a concern. You only care about their intersection in the sample space, which is the probability of both events occurring (for more intuition on this, check out [my post about compound event probabilities](#)).

However, that doesn't mean the terms are in any way interchangeable in general. Each term represents a different subset of the sample space and hence a different probability or conditional probability. However, if you combine the two equations, you can see that the expressions on the right-hand side are really equal to the same thing. And, therefore, they are equal to each other:



**(JOINT PROBABILITY)**

$$P(\text{Event-1}) \times P(\text{Event-2} \mid \text{Event-1}) = P(\text{Event-2}) \times P(\text{Event-1} \mid \text{Event-2})$$

To get to Bayes' theorem from here, you only have to divide both sides of the equation by  $P(\text{Event-1})$  or  $P(\text{Event-2})$ :

**(BAYES' THEOREM)**

$$P(\text{Event-2} \mid \text{Event-1}) = \frac{P(\text{Event-1} \mid \text{Event-2}) \times P(\text{Event-2})}{P(\text{Event-1})}$$

Yes, this is really it. Now you officially know the origin of Bayes' theorem! I told you it was simple.

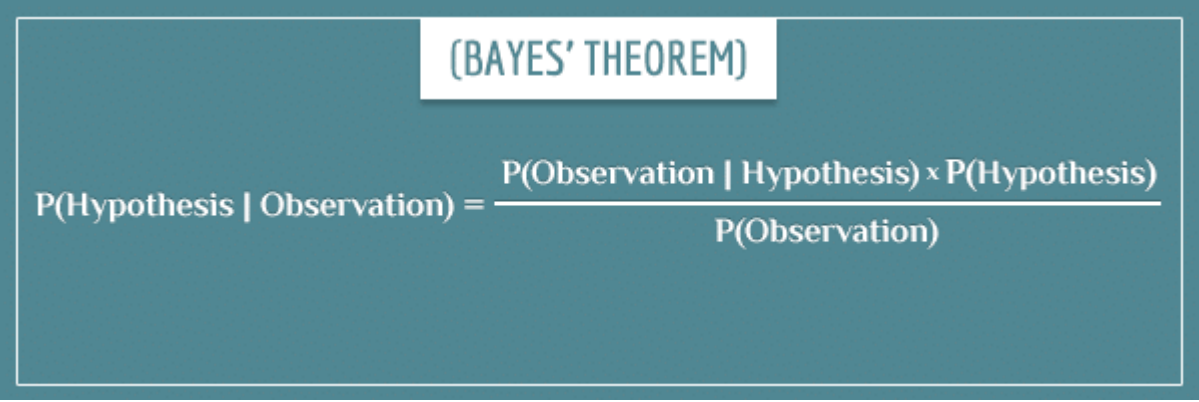
Okay, this is one way to mathematically derive it. It's still not quite enough to get a good feeling for it, however. Does Bayes' theorem make intuitive sense or is it some mathematical truth that you just have to accept? I'm dedicating the final section of this post to exploring this question.

## The intuition behind the theorem

Imagine you have a hypothesis about some phenomenon in the world. What is the probability that the hypothesis is true?

To answer the question, you make observations related to the hypothesis and use Bayes' theorem to update its probability. The hypothesis has some prior probability which is based on past knowledge (previous observations). To update it with a new observation,

you multiply the prior probability by the respective likelihood term and then divide by the evidence term. The updated prior probability is called the posterior probability.



(BAYES' THEOREM)

$$P(\text{Hypothesis} | \text{Observation}) = \frac{P(\text{Observation} | \text{Hypothesis}) \times P(\text{Hypothesis})}{P(\text{Observation})}$$

The posterior probability then becomes the next prior which you can update from another observation. And so the cycle continues.

Notice the dependence between the posterior probability and the three terms on the right-hand side of the equation. All else being equal:

1. A large prior probability term will make the posterior probability *larger* (compared to a small prior probability).
2. Similarly, a large likelihood term will make the posterior probability *larger*.
3. A large evidence term will make the posterior probability *smaller*.

These relationships come from the fact that the posterior is directly proportional to the prior and likelihood terms, but inversely proportional to the evidence term.

Now, the first relationship should be pretty intuitive. All else being equal, the posterior probability will be higher if you already had strong reasons to believe the hypothesis is true. But what about the other two relationships?

Let's explore this in some more detail.

## The likelihood

The likelihood reads as “the probability of the observation, given that the hypothesis is true”.

This term basically represents how strongly the hypothesis predicts the observation. If the observation is very consistent with your hypothesis, all else being equal, it will increase the posterior probability.

Similarly, if according to your hypothesis the observation is very unlikely and surprising, that will reduce the posterior probability.

Here's a simple example that makes this intuition explicit.

Say someone tells you that there's some living being on the street: I'll name it Creature. Your strongest hypothesis is that Creature is a human (because of the frequency with which you're used to seeing humans in your neighborhood, compared to other animals). That is, the prior probability  $P(\text{Human})$  is high.

Next, you have the following observation:

- Creature is wearing clothes.

Now the posterior probability (the “new prior”) is:

$$P(\text{Creature is a human} \mid \text{Creature wears clothes})$$

The corresponding likelihood term is:

$$P(\text{Creature wears clothes} \mid \text{Creature is a human})$$

The value of this likelihood will obviously be very high. It's not that it's impossible to come across a naked person on the street, but the probability is quite low.

On the other hand, if you know that Creature is currently barking, the following likelihood term will have a very low value:

$$P(\text{Creature is barking} \mid \text{Creature is a human})$$

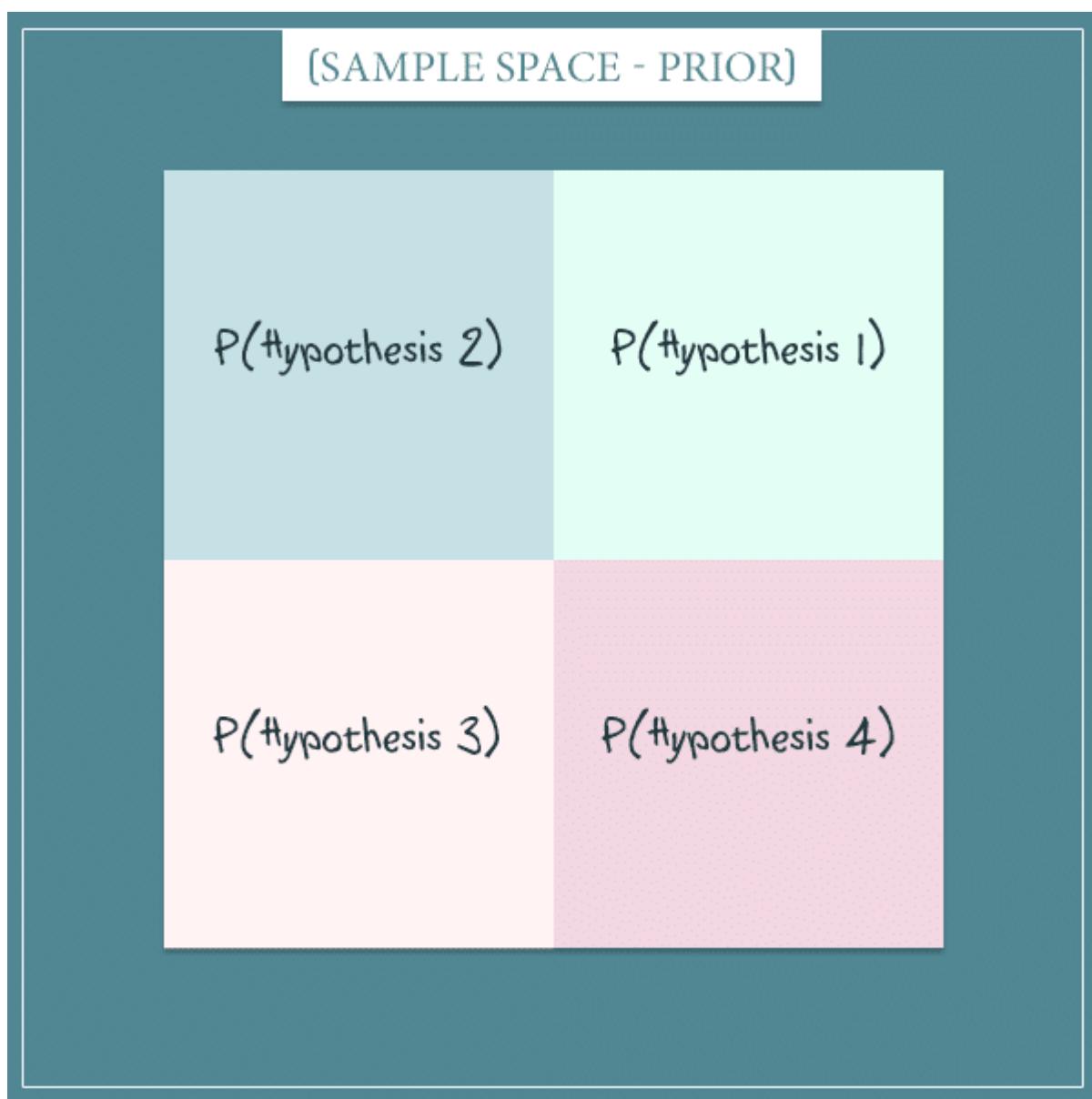
To sum up, the more consistent the observation is with the hypothesis, the more probable it is that the hypothesis is true. This aspect of Bayes' theorem should also have a strong intuitive appeal.

## The evidence

In the context of Bayes' theorem, the evidence is the probability of the observation used to update the prior. Examples of such observations are:

- Creature has 4 legs
- It is breathing
- It is inside a car

I'll change the example a little bit and say there are only 4 equiprobable hypotheses about Creature:

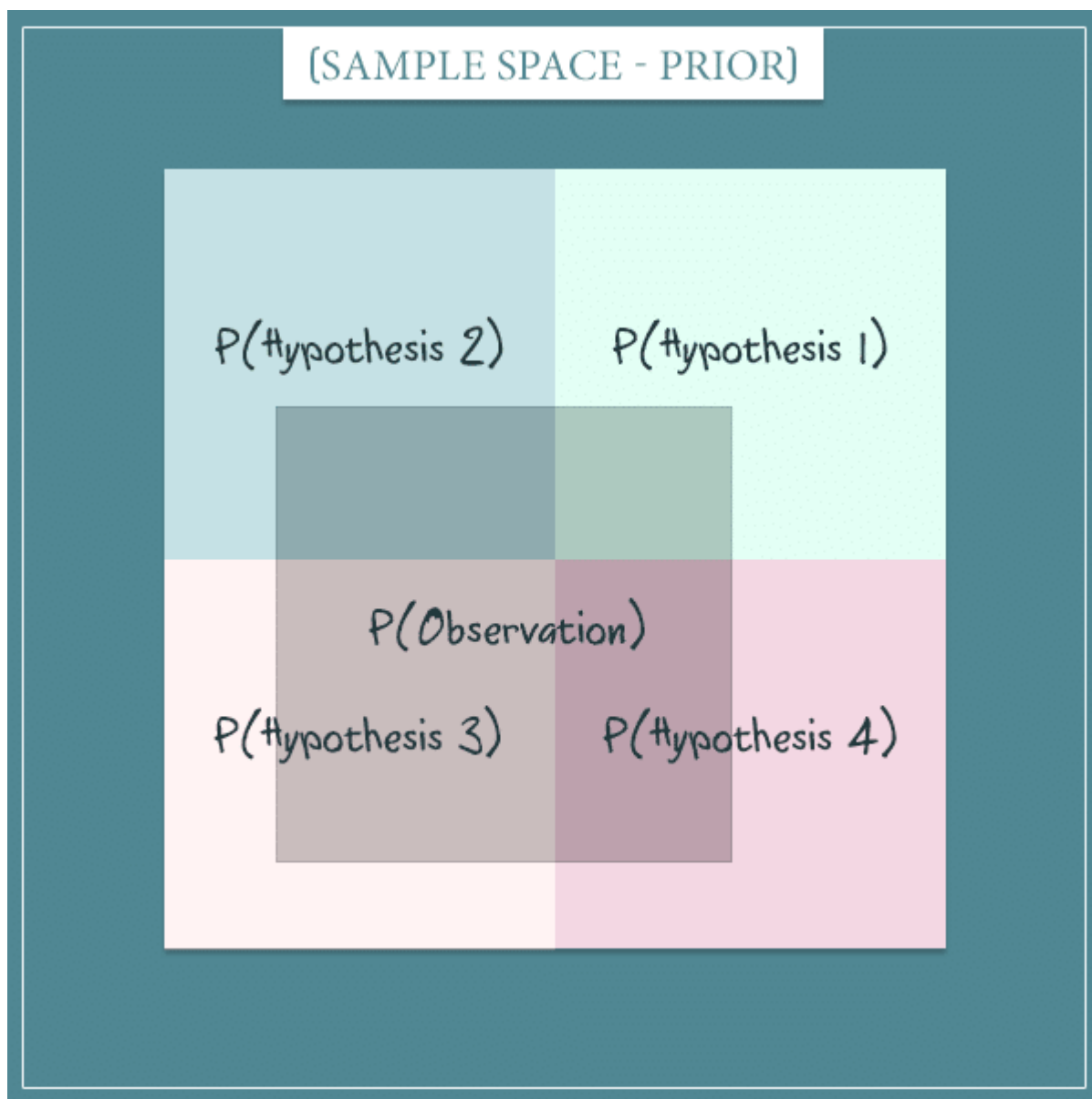


The four hypotheses can, for example, represent “human”, “dog”, “cat”, and “mouse”.

The uncertainty is currently too high and you decide to collect more information to update the probabilities. You're hoping to make some observations about Creature to help you narrow down the posterior and make it somewhat more informative.

### The effect of the evidence on the sample space

Let's say you make a particular observation. After you make it, you can represent the probability it had prior to making it in the same sample space like so:

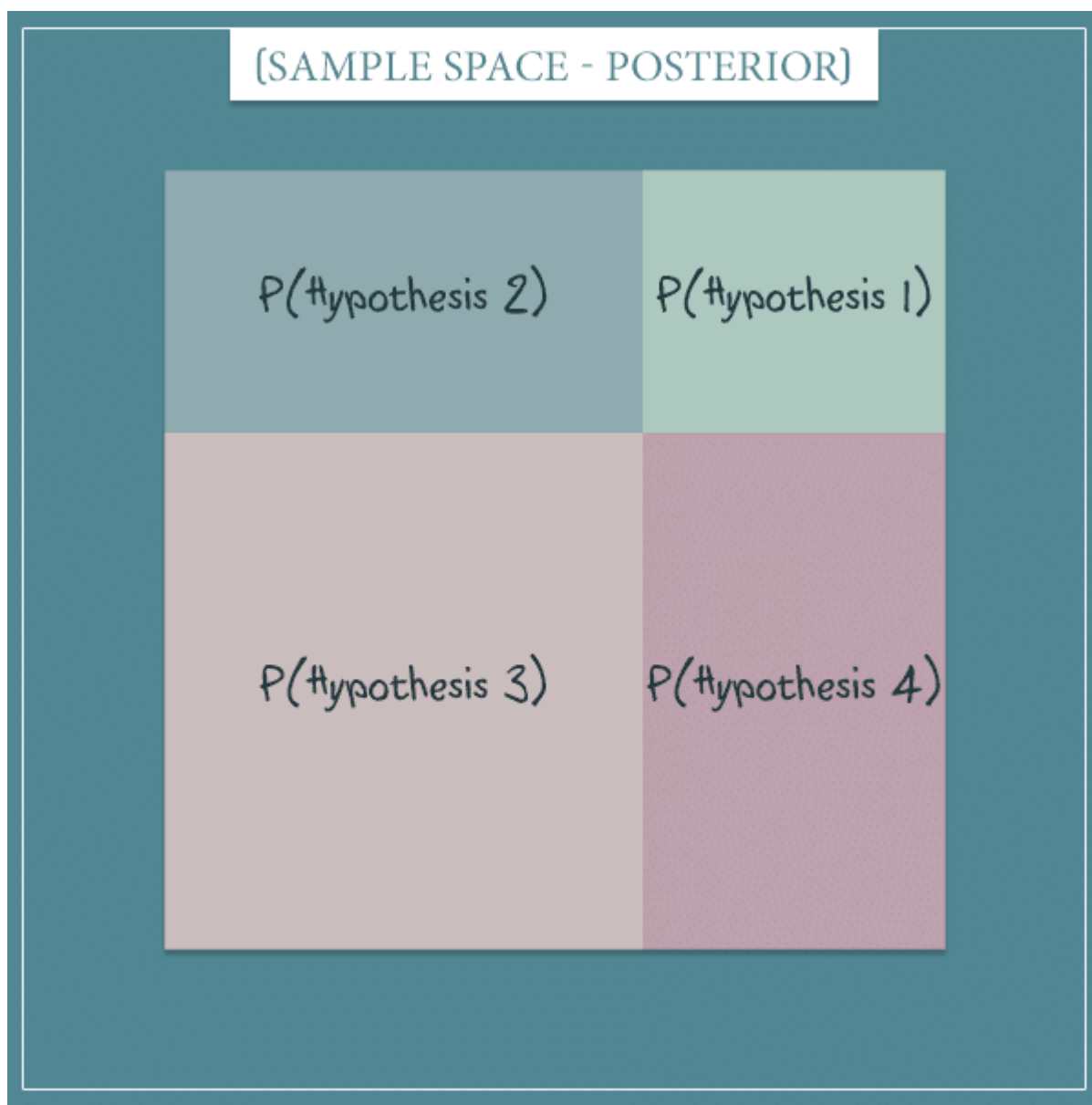


Think about the sample space as representing all possible worlds. Some of those worlds are consistent with the observation and are bound within the dark square depicting  $P(\text{Observation})$ .

Now this is the crucial part. The above is the old sample space. After you make the observation, you eliminate the worlds that are inconsistent with the observation from the sample space.

For example, if you learn that Creature has 4 legs, you will ignore the worlds in which Creature has any different number of legs. Those worlds would be the parts of the sample space outside  $P(\text{Observation})$ . In other words, the world you live in happens to be within the area of  $P(\text{Observation})$ .

Therefore, that area now becomes the new sample space:



Remember that the total probability of a sample space is always equal to 1. So, you redistribute this probability among the hypotheses proportionally to the percentage of

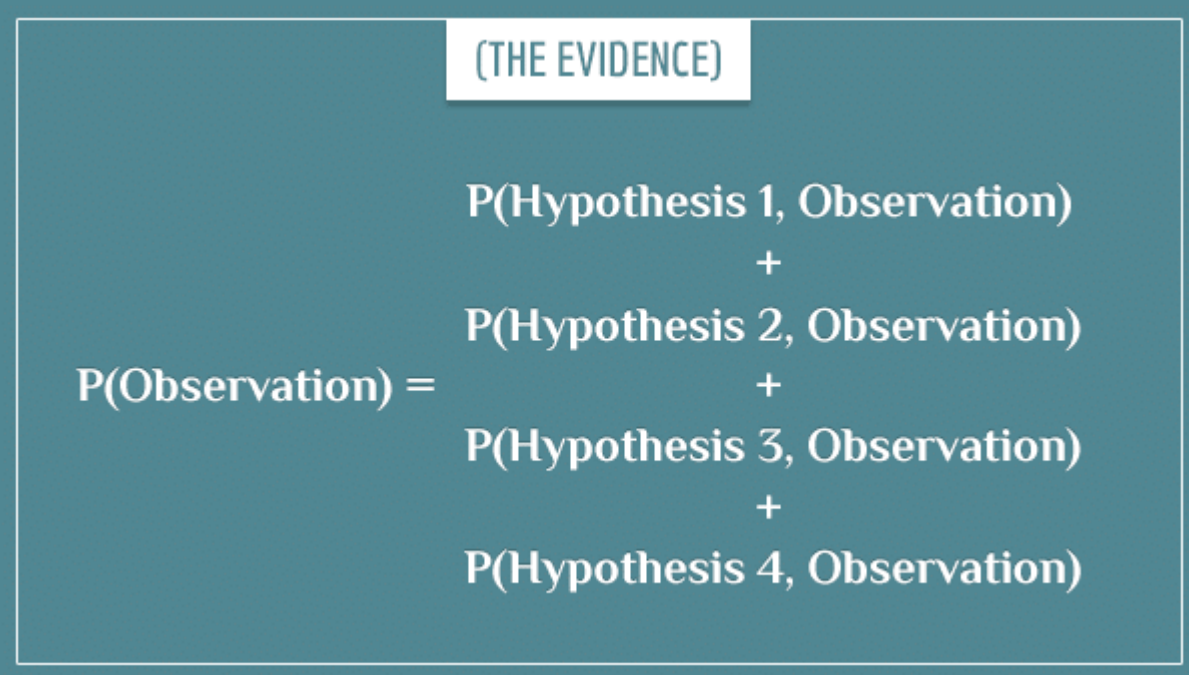
the new sample space they occupy.

Each hypothesis used to have a probability of 0.25 but this is no longer the case. For example, you can see that in the updated sample space Hypothesis 3 has gained a much larger portion, whereas Hypothesis 1's probability has shrunk significantly. That is because Hypothesis 1 covered a larger portion of  $P(\text{Observation})$ .

Well, this is Bayes' theorem in a nutshell! The “old” sample space represents the prior probability and the “new” sample space represents the posterior probability of each hypothesis.

### More intuition about the evidence

Take another look at the second-to-last image above. Before you make the observation, the overlap between the area of  $P(\text{Observation})$  and the area of a particular hypothesis is actually their joint probability. Notice that the 4 joint probabilities completely cover  $P(\text{Observation})$ . You can express this as their sum:


$$\begin{array}{l} \text{(THE EVIDENCE)} \\ \\ P(\text{Observation}) = \\ \quad P(\text{Hypothesis 1, Observation}) \\ \quad + \\ \quad P(\text{Hypothesis 2, Observation}) \\ \quad + \\ \quad P(\text{Hypothesis 3, Observation}) \\ \quad + \\ \quad P(\text{Hypothesis 4, Observation}) \end{array}$$

Remember that the numerator in Bayes' theorem is one of those joint probabilities. The denominator is the evidence: as I just showed, it's also the sum of all joint probabilities. So, the posterior probability  $P(\text{Hypothesis} \mid \text{Observation})$  is the fraction of the area of  $P(\text{Observation})$  covered by  $P(\text{Hypothesis, Observation})$ :



$$P(\text{Hypothesis} \mid \text{Observation}) = \frac{P(\text{Hypothesis}, \text{Observation})}{P(\text{Observation})}$$

You can think about it as the portion of the observation that your hypothesis explains (in relation to how much the other hypotheses explain it).

Since the evidence term is in the denominator, it's inversely proportional to the posterior probability. That is, the higher the evidence, the lower the posterior is going to be.

To understand this, think about what makes the evidence term grow in the first place.

If  $P(\text{Observation})$  is high, that means the observation wasn't a surprising one and it was probably strongly predicted not just by your hypothesis, but by some of the alternative hypotheses. For example, if your observation is that Creature is breathing, the posterior  $P(\text{Creature is a human} \mid \text{Creature breathes})$  won't be too different from the prior  $P(\text{Creature is a human})$ . The reason is that, even though this observation is consistent with the hypothesis, it is equally consistent with all alternative hypotheses and it will support them just as much.

On the other hand, if  $P(\text{Observation})$  is small, that means the observation was surprising and unexpected. A hypothesis which happens to strongly predict such an observation is going to get the largest boost of its probability (at the expense of the probabilities of the other hypotheses, of course).

## Summary

In this post I presented an intuitive derivation of Bayes' theorem. This means that now you know why the rule for updating probabilities from evidence is what it is and you don't have to take any part of it for granted. I also gave some insight on the relationship between the 4 terms of the equation.

Perhaps you already understood the theorem, but now you also *feel it*! So, go ahead and start confidently applying it in whatever areas interest you most.

---

To get more experience and even more intuition about Bayes' theorem, check out my posts on [Bayesian belief networks](#) and applying Bayes' theorem to solving the so-called

inverse problem.

Filed Under: **BAYES' THEOREM**

Tagged With: **COIN FLIP** , **CONDITIONAL PROBABILITY** , **SAMPLE SPACE**

---

## Comments



**KD** says

October 22, 2018 at 8:18 am

Super intuitive explanation. Thanks.

Reply



**FILIP SJÖSTRAND** says

February 26, 2019 at 9:23 pm

Really intuitive. Thanks! Big thumbs up for the blog in general.

Reply

**PSYCHSTUDENT** says

May 4, 2019 at 10:36 am

Thank you for this, great explanation!

Reply

**HAYDUM** says

September 11, 2019 at 11:55 am

Beautifully explained! Thanks!

Reply

**THE CTHAEH** says

October 24, 2019 at 5:29 pm

Thank you all for the positive feedback!

I received a question about this post through the contact form and I want to post the answer here, as other readers might be wondering the same thing. Here's the original question from John:

“Hi, in your article you say “Similarly, if according to your hypothesis the observation is very unlikely and surprising, that will reduce the posterior probability.”. That seems to go against the equation of Bayes' theorem and on later elaborations. It seems like a rare observation means a smaller number in the denominator and therefore a higher posterior probability. Am I misunderstanding or is this is a mistake?

And here's my response:

“  
Hi John,

Thanks for writing! Notice that there is a difference between an observation being unlikely according to your hypothesis, versus it being overall unlikely. If the observation is overall likely, this means that many other (alternative) hypotheses gave it a high probability. This doesn't suggest that your hypothesis will also give it a high probability, however.

The posterior probability would be really high in situations where the overall probability of the observation is low (since the denominator will be low) but *\*only\** your hypothesis gives the observation a high probability (since the likelihood term is in the numerator). The intuition here is that, if the overall probability is low, this observation isn't explained well by alternative hypotheses. But since your hypothesis gives it a high probability, it is explained well by it. Hence, your hypothesis gains “probability points” relative to its rivals.

Similarly, the posterior probability would be low if your hypothesis gives a low probability of the observation but the overall probability of the observation is high.

Reply

**ALEX COSTELLO** says

November 5, 2019 at 3:50 pm

Great post!

I was reading and had a question about the below paragraph:

“Then, once you know the first event has occurred, what is the probability that the second event will also occur? Naturally, this is the conditional probability  $P(\text{Event-2} \mid \text{Event-1})$ : the probability of the first event, given that the second has occurred.”

Isn't the notation  $P(A|B)$  referred to as the probability of event A, given B occurred? So therefore  $P(\text{Event-2} \mid \text{Event-1})$  is the probability of the second event given that the first event occurred?

Reply

**THE CTHAEH** says

November 5, 2019 at 5:02 pm

Hi, Alex! Thanks for writing.

“

So therefore  $P(\text{Event-2} \mid \text{Event-1})$  is the probability of the second event given that the first event occurred?

That's correct, it's a wording mistake on my part.

Also, I just realized that there is some ambiguity in me referring to the events as "the first event" and "the second event" in that bit, since it's not clear if it means Event-1 and Event-2 or the order of the events in the expression  $P(\text{Event-2} \mid \text{Event-1})$ .

My wording there is actually unnecessarily confusing. I fixed it to:

"Then, once you know Event-1 has occurred, what is the probability that Event-2 will also occur? Naturally, this is the conditional probability  $P(\text{Event-2} \mid \text{Event-1})$ : the probability of Event-2, given Event-1."

Thanks for pointing this out!

**BUMBLE BEE** says

June 24, 2020 at 9:24 am

Terrific!

Reply

## Leave a Reply

Your email address will not be published. Required fields are marked \*

## Comment

Name \*

Email \*

Website

POST COMMENT

Sign Up

## SIGN UP FOR THE PROBABILISTIC WORLD NEWSLETTER

Enter your email below to  
receive updates and be notified  
about new posts.



SIGN UP

## Follow Probabilistic World



## Recent posts

- [Arithmetic Properties: A Comprehensive Breakdown](#)
- [Numbers, Arithmetic, and the Physical World \(Series\)](#)
- [Binomial Distribution Mean and Variance Formulas \(Proof\)](#)
- [The Binomial Distribution \(and Theorem\): Intuitive Understanding](#)
- [The Bernoulli Distribution: Intuitive Understanding](#)

PROBABILISTIC WORLD