Probability Fundamentals

Thomas Nield
O'Reilly Media

What to Expect

1. Introduction
2. Probability Math
3. Bayes Theorem
4. Normal Distribution

Section I

Why Learn Probability?

Why Learn Probability?

Probability is the building block to statistics, machine learning, data science, engineering, and several other disciplines.

It is common to have to make decisions with limited information; as a matter of act, this is most decisions we make.

Probability enables us to measure how certain we are in an outcome occurring and while this is an imperfect art, it can be immensely valuable.

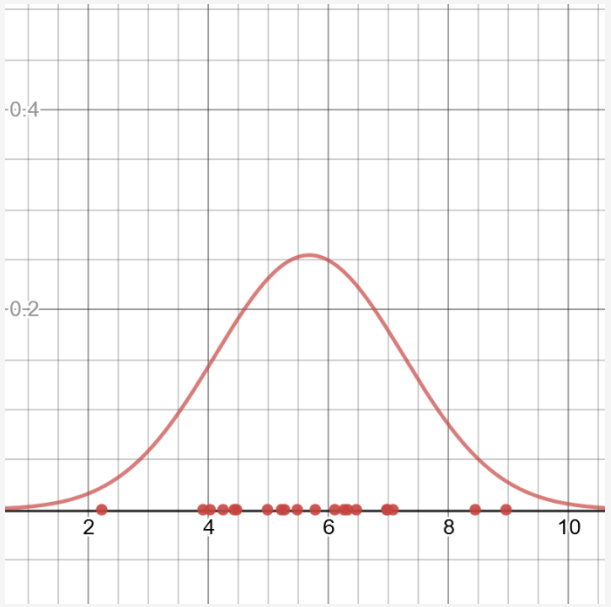
Probability versus Statistics

Probability and statistics often get confused and said interchangeably, but there is a distinction.

- **Probability** is solely about studying likelihood.
- **Statistics** utilizes data to discover likelihood.

In practicality, these two things are going to be tightly tied together, as one can argue it is hard to have probability without data.

We will put more emphasis on probability in this class, but we will work with data and statistics where it makes sense.



A normal distribution is a probability tool, but it becomes statistical when it is fit to data points.

What is Probability?

Probability is how likely an event will happen, based on observations or belief.

- How likely is it I will get 7 heads in 10 fair coin flips?
- What are my chances in winning an election?
- What is the likelihood my flight will be late?
- How certain am I a product is defective?

Probability is often expressed in two ways:

- As a percentage: 60% chance my flight will be late
- As an odds: 3:2 odds my flight will be late

However, there are two philosophies on probability.

Frequentist Statistics

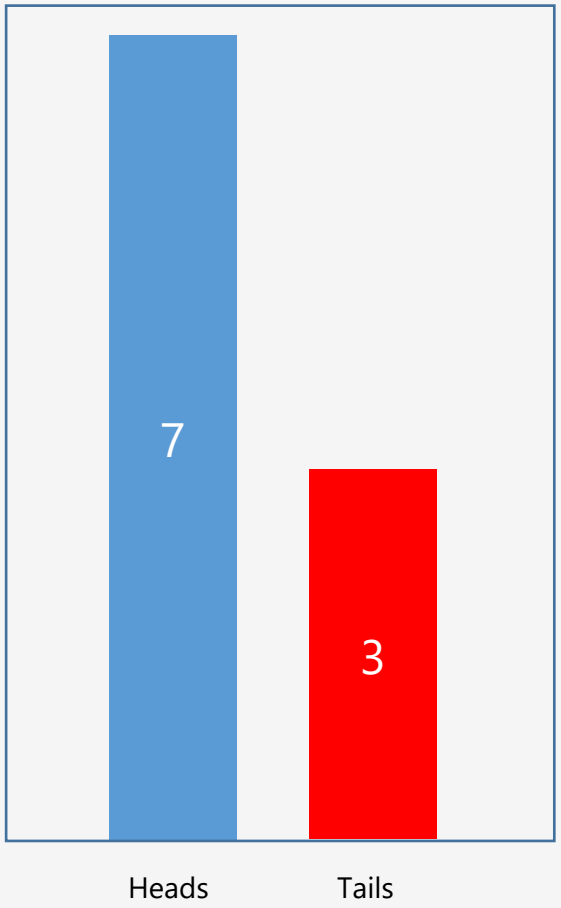
Frequentist statistics is the most popularly understood approach to deriving probability, believing that frequency of an event provides hard evidence of the probability.

EXAMPLE: I flip a coin 10 times and I get 7 heads, so I suspect the probability of heads is 70%.

This definition might be a little simplistic, as frequentists believe gathering more data will increase confidence in the probability.

Frequentism tends to work best when a lot of data is available, reliable, and complete.

Tools frequentists use include p-values and confidence intervals.



Bernoulli distribution of coin flip outcomes

Bayesian Statistics

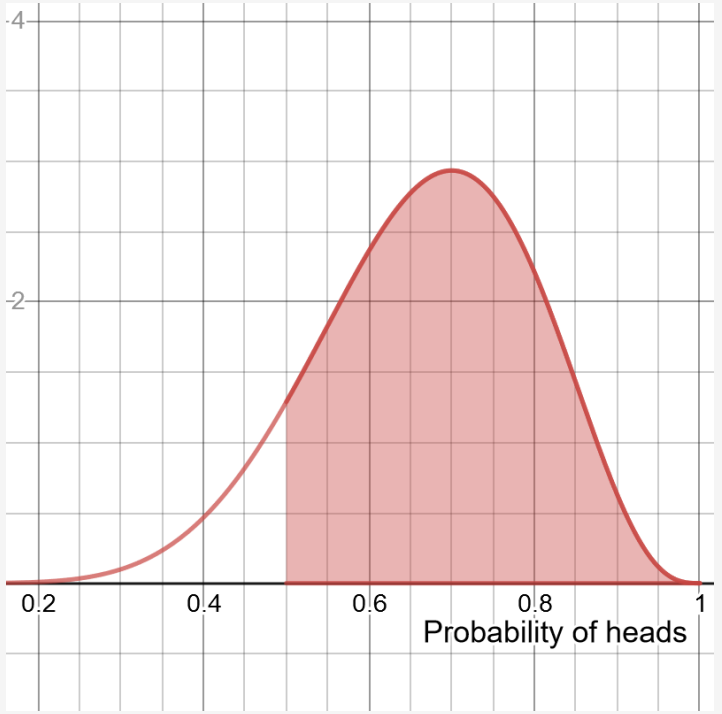
Bayesian statistics is much more abstract in that it assigns subjective beliefs in a probability and not just data.

An arbitrary probability can be assigned based on subjective beliefs, and then data can be used to gradually update that belief.

EXAMPLE: I believe a coin has 50% probability of heads. I flip it 10 times and I get 7 heads. I update a beta distribution (to the right) and see there are greater likelihoods of heads being more than 50%.

Bayesian methods tend to work well when data is limited, a large amount of domain knowledge is present, or uncertainty is hard to eliminate.

Bayesian tools include the Bayes factor and credible intervals.



Beta distribution showing the probability of probabilities of heads, given 7 heads and 3 tails.

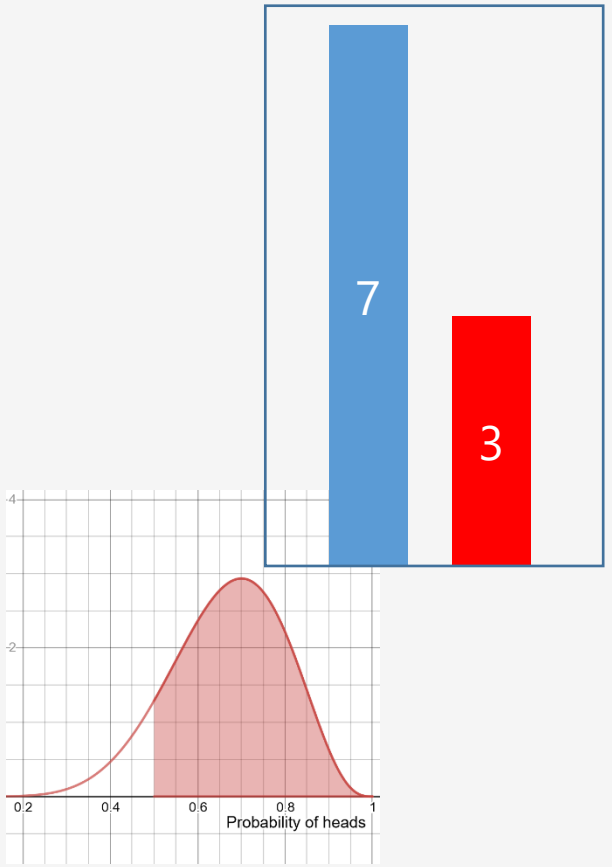
Bayes versus Frequentist? Which is right?

Frequentists believe probability is absolute, and Bayesians believe probability is a fuzzy construct.

Which of these philosophies are right? Both!

- Some situations warrant Frequentism, while others warrant Bayes.
- Frequentism is better if you can converge on a single probability based on enough data, but Bayes is more suited if you want to entertain larger ranges of possibility for a limited amount of data.

Keep in mind that all models are wrong but some are useful, and you should always seek the right tool for the job.



Section II

Probability Fundamentals

Probability Basics

Hopefully the concept of a **probability** is familiar, which measures how likely an outcome x is and typically is represented as a number $P(x)$ between 0.0 and 1.0.

A probability is typically represented as a decimal between 0.0 and 1.0 but is also represented as a percentage between 0% and 100%.

The probability of an event $P(x)$ ***NOT*** occurring can be calculated by $1.0 - P(x)$, which indicates both outcomes must add to 1.0.

When we work with a single simple probability, it is known as a **marginal probability**.

Where Does Probability Come From?

Probability can be based off data, a belief, or both!

Probability based on data: If we sample 10 products from a factory line and find 4 items are defective, that would be a 40% defective rate.

Probability based on belief: An engineer realizes an inferior material was used and guesses the defective rate for the product will be 50%.

Probability based on data + belief: we can quantify the engineer's belief and the data, merge them together, and find a 44.44% probability is most likely.

Expressing Probability as Odds

You may sometimes see probability expressed as an **odds**, which expresses how many times we believe in something being true versus not being true.

If we believe out of 10 products, 6 will be defective and 4 will not, the odds would be $6:4$ or $\frac{6}{4}$

This would reduce to 1.5, which means we believe a product is 1.5 times more likely to be defective than not defective.

Quantifying Belief with Odds

oddss are also a helpful way to quantify subjective beliefs by means of "betting."

If my friend is willing to pay me $200 if the Dallas Cowboys win the Superbowl, but I must pay him $50 if they do not, that means he believes the Dallas Cowboys are 4x more likely to lose rather than win ($\frac{200}{50} = 4.0$).

To hyperbolize, if he pays $200 for them winning but I must pay him $1 if they lose, that means he *REALLY* believes the Dallas Cowboys are going to lose: 200x more likely ($\frac{200}{1} = 200$) and feels no need to hedge his bet.

Putting money on something (horse races, sports teams, stock markets) forces you to quantify how strongly you believe in an event, and can be effective measure to turn subjective beliefs into something quantitative.

Turning Odds into Probabilities

If you need to turn an odds $O(X)$ into a probability, you can do so with this formula:

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

```
def prob_to_odds(p):
  return p / (1.0 - p)

def odds_to_prob(o):
  return o / (1.0 + o)


p_x = .75
o_x = prob_to_odds(p_x)

# prints ODDS: 3.0
print("ODDS: {}".format(o_x))
```

So if you believe something is 3x likely to happen versus not happen (3:1), the probability for that event is 75%:

$$P(X) = \frac{3}{1+3} = .75$$

Joint Probabilities

Probability gets interesting when we think about how multiple probabilities interact with each other.

Let's look at the probability of flipping a coin and rolling a die and getting a *heads* (Event A) and a *six* (Event B).

This is known as a **joint probability**, the probability of two events *A* and *B* occurring simultaneously.

Since *A* and B are **independent** in this case, meaning they do not affect each others' outcomes, their joint probability is as simple as multiplying them together:

$$P(A \text{ and } B) = P(A) * P(B)$$

$$P(Heads \text{ and } 6) = \frac{1}{2} * \frac{1}{6} = \frac{1}{12} = .0833$$

$P(A \cap B)$

$\mathrm{P}(A \text{ and } B)$

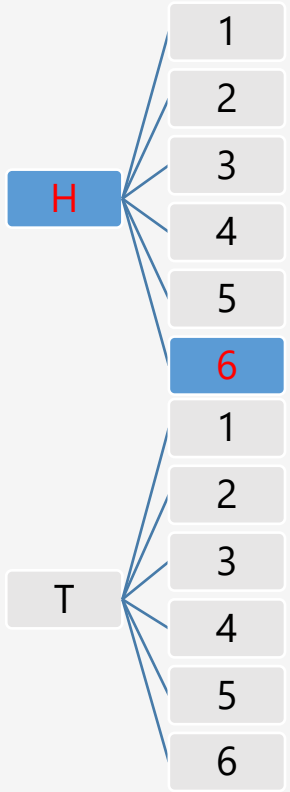
$\mathrm{P}(A, B)$

and

```
p_heads = 1.0 / 2.0
p_six = 1.0 / 6.0

p_heads_and_six = p_heads * p_six

print(p_heads_and_six) # 0.08333333333333333
```

Joint Probabilities

Why does joint probability work like this for independent events? Let's consider all possible outcomes in the tree to the right:



Notice that "heads" has a 50% probability, and "six" has a 16.66% probability. We have 12 possible outcomes:

H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6

Only one of these 12 possible outcomes meets our criteria, so $\frac{1}{12} = .08333$ or 8.333%

The multiplication allows us to avoid generating and counting all these combinations and is known as the **product rule**.

```
# Declare possible outcomes for coin and die
coin_outcomes = ["H", "T"]
die_outcomes = [1, 2, 3, 4, 5, 6]

# Combine each outcome between coin and die
all_combinations = [(c,d) for c in coin_outcomes for d in die_outcomes]

# Find only outcomes for Heads and 6 (should only be one)
head_and_6 = [t for t in all_combinations if t[0] == "H" and t[1] == 6]

# 1/12 = .083333
print(float(len(head_and_6)) / float(len(all_combinations)))
```

$$P(Heads\ and\ 6) = \frac{1}{2} * \frac{1}{6} = \frac{1}{12}$$

Joint Probabilities

You can use the product rule to combine as many probabilities as you want.

For example, the probability of rolling a “six” 10 times in a row:

$$\frac{1}{6} * \frac{1}{6} * \frac{1}{6} * \frac{1}{6} * \frac{1}{6} * \frac{1}{6} * \frac{1}{6} * \frac{1}{6} * \frac{1}{6} * \frac{1}{6} = \frac{1}{60466176} = 0.0000000165382$$

Of course, since this multiplication is repetitive, we can use an exponent:

$$(\frac{1}{6})^{10} = \frac{1}{60466176} = 0.0000000165382$$

```
# Probability of rolling a "six"
p = 1.0 / 6.0

# Probability of rolling a "six" ten times in a row
p_10_sixes = p ** 10

print(p_10_sixes)
```

Exercise #1

There is a 30% chance of rain today, and a 40% chance your umbrella order will arrive on time. You are eager to walk in the rain today and cannot do so without either!

What is the probability it will rain ***AND*** your umbrella will arrive?

Exercise #1

There is a 30% chance of rain today, and a 40% chance your umbrella order will arrive on time. You are eager to walk in the rain today and cannot do so without either!

What is the probability it will rain ***AND*** your umbrella will arrive?

$$P(A \text{ and } B) = P(A) * P(B)$$
$$P(Rain \text{ and } umbrella \text{ } arrives) = .30 * 40 = \ .12$$

12% chance it will rain and your umbrella will arrive.

Union Probability – Mutually Exclusive

Things get a little more nuanced when we work with **union probabilities**, or the probability that at least one of multiple events will occur.

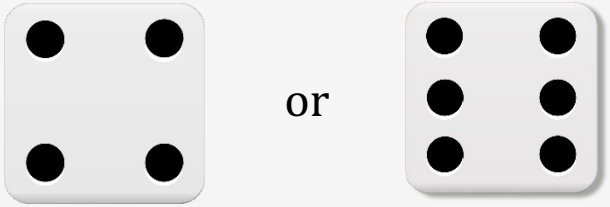
When two events are **mutually exclusive**, meaning that only one of the events can occur but not both, then it is as easy as adding their probabilities together.

For example, what is the probability of getting a “4” or “6” on a die roll? Since we cannot get “4” and a “6” simultaneously, we just add these probabilities together.

$$\frac{1}{6} + \frac{1}{6} = \frac{2}{6} = .333$$

$$P(A \cup B)$$

$$\mathrm{P}(A \text{ or } B)$$



Union Probability – Non-Mutually Exclusive

When two events are **non-mutually exclusive**, meaning that two events can occur simultaneously, their union probability gets tricky.

$$P(A \cup B)$$
$$\mathrm{P}(A \text{ or } B)$$

What is the probability of getting a "heads" ***OR*** a "six" with a coin flip and a die roll? Before we try anything let's look at the possible outcomes:

H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6

or

The outcomes in red are the ones that meet our condition, and counting them would lead us to a correct answer:

$$\frac{7}{12} = .5833$$

But if we added those two probabilities together, we would get a wrong answer! So why does this not work?

$$\frac{1}{2} + \frac{1}{6} = \frac{2}{3} = .6666$$

Union Probability – Non-Mutually Exclusive

H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6

The problem with adding non-mutually exclusive events is it causes double-counting of joint probability events as highlighted above.

It's subtle but notice above that if we add the probability of "heads" (1/2) with the probability of getting a "six" (1/6), we have counted that "six" probability of 1/6 twice!

We can remedy this by subtracting the joint probability for non-mutually exclusive events:

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

$$P(heads\ or\ six) = \frac{1}{2} + \frac{1}{6} - \left(\frac{1}{2} * \frac{1}{6}\right) = \frac{7}{12}$$

This is known as the **sum rule**, where we can find the OR probability between two or more events by summing them and then subtracting their joint probability.

$$P(A \cup B)$$

$$\mathrm{P}(A \text{ or } B)$$

or

The Sum Rule

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

$$P(A \cup B)$$

$$\mathrm{P}(A\ \mathrm{or}\ B)$$

Notice that the sum rule applies to both mutually exclusive and non-mutually exclusive probabilities.

When events A and B are mutually exclusive, the joint probability is 0 since they both cannot occur simultaneously.

But when events A and B are not mutually exclusive, the joint probability plays a role in subtracting the double-counting caused by both events occurring.

or

Exercise #2

There is a 30% chance of rain today, and a 40% chance your umbrella order will arrive on time.

You will only be able to run errands if it does not rain or your umbrella arrives.

What is the probability it will not rain ***OR*** your umbrella arrives?

Exercise #2

There is a 30% chance of rain today, and a 40% chance your umbrella order will arrive on time.

You will only be able to run errands if it does not rain or your umbrella arrives.

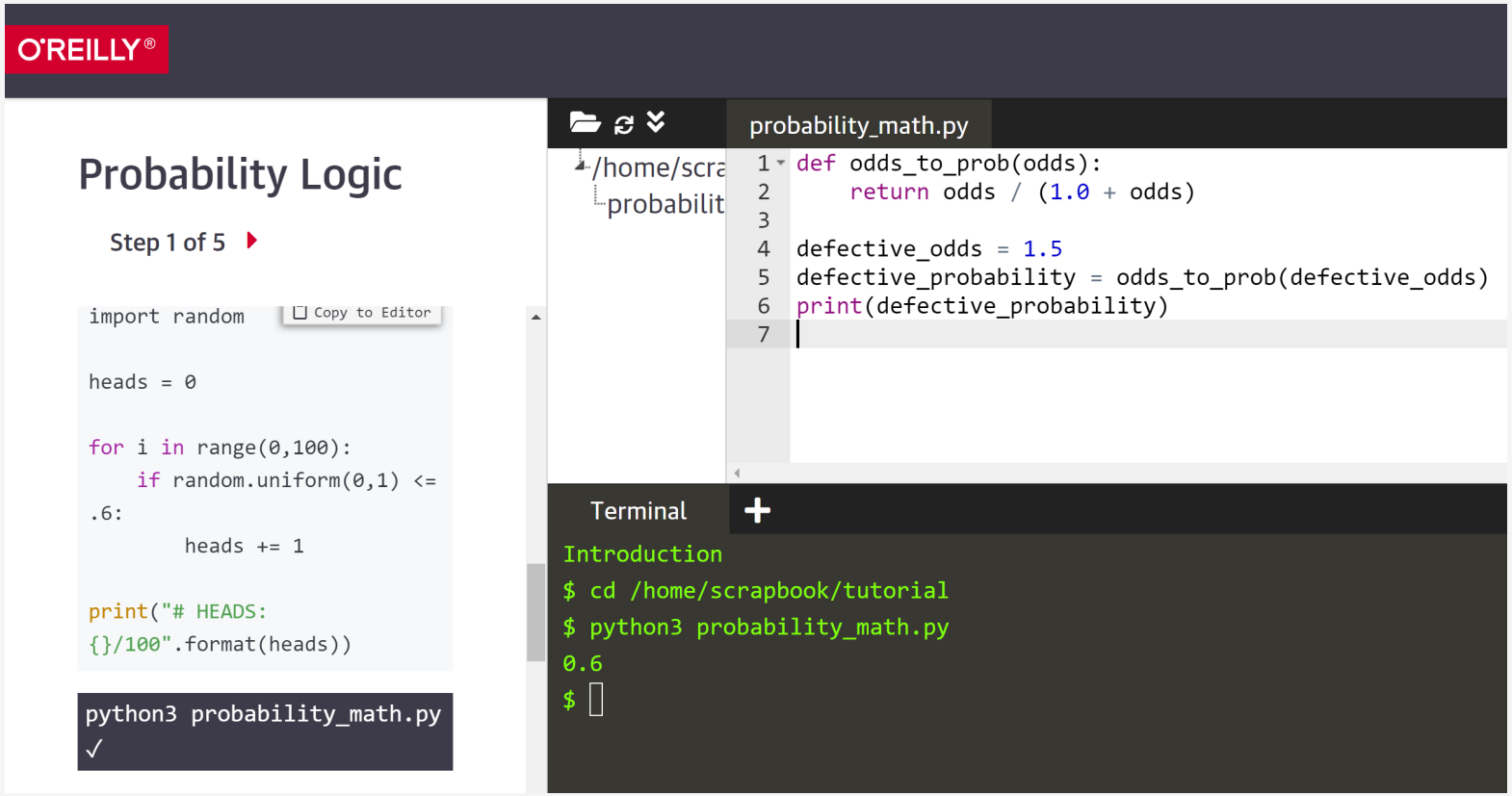
What is the probability it will not rain ***OR*** your umbrella arrives?

$$P(rain) = .30$$
$$P(no\ rain) = 1.0 - P(rain) = 1.0 - .30 = .70$$
$$P(umbrella\ arrives) = .40$$

$$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$$
$$P(no\ rain\ or\ umbrella\ arrives\ ) = .70 + .40 - (.70 * .40) = .82$$

82% chance it will not rain or your umbrella arrives.

Katacoda Interactive Scenario – Probability Logic



https://learning.oreilly.com/scenarios/probability-from-scratch/9781492080541/

Conditional Probability

Another nuanced idea in probability is **conditional probability**, or the probability of *A* given *B* has occurred.

If *B* has no impact on whether *A* occurs, then $P(A) = P(A\ given\ B)$.

But if *B* does increase or decrease the probability of *A* occurring, then the $P(A)$ is going to be different than $P(A\ given\ B)$.

The probability of someone being colorblind is 4.25%, but the probability of a male being colorblind is 8%.

$$P(colorblind) = .0425$$

$$P(colorblind\ given\ male) = .08$$

Does this mean any colorblind person is 8% likely to be male? Or that any male is 8% likely to be colorblind?

$P(A|B)$

P(A given B)

Conditional Probability

Direction of the conditional probability matters! The *probability of a male given they are colorblind* is not the same as the *probability of being colorblind given they are male*!

We will learn how to flip this probability with Bayes Theorem shortly, but first let's see why conditional probability plays a critical role in **AND** and **OR** probabilities.

Let's say we draw a random person from the population, and the probability of them being male $P(male)$ is 50%.

If we want to calculate the probability they are male and colorblind, do we multiply $P(male)$ with $P(colorblind)$ or $P(colorblind\ given\ male)$?

$P(male) = .5$
$P(colorblind) = .0425$
$P(colorblind\ given\ male) = .08$

$P(male\ and\ colorblind) = P(male) \quad * \quad ?$

$P(A|B)$

P(A given B)

Conditional Probability

$P(male) = .5$

~~$P(colorblind) = .0425$~~

$P(colorblind\ given\ male) = .08$

$P(male\ and\ colorblind) = P(male) * P(colorblind\ given\ male)$

$P(male\ and\ colorblind) = .5 * .08$

$P(male\ and\ colorblind) = .004$

With joint probability we use the conditional probability when it is available! If we want to calculate the probability of someone being male and colorblind, use the conditional probability that accounts for them being male!

If we do not use the conditional probability, then the probability of them being male and colorblind will be no different than being female and colorblind!

$P(A|B)$

P(A given B)

Conditional Probability

We can update our joint probability and union probability formulas to account for conditional probability:

$$\boldsymbol{P(A\ and\ B) = P(A) * P(B\ given\ A)}$$

$$\boldsymbol{P(A\ or\ B) = P(A) + P(B) - P(A) * P(B\ given\ A)}$$

Remember that if A and B are independent and have no impact on each other, then these formulas still work because $P(B) = P(B\ given\ A)$.

It can be hard to determine if two events are conditional and related, and therefore it is common to assume in statistics they are independent.

We will see this assumption in machine learning applications like Naïve Bayes.

$P(A|B)$

P(A given B)

Exercise

There is a 30% chance of rain today, and a 40% chance your umbrella order will arrive on time.

However, you found out if it rains there is only 20% chance your umbrella will arrive on time.

What is the probability it will rain and your umbrella will arrive on time?

Exercise

There is a 30% chance of rain today, and a 40% chance your umbrella order will arrive on time.

However, you found out if it rains there is only 20% chance your umbrella will arrive on time.

What is the probability it will rain and your umbrella will arrive on time?

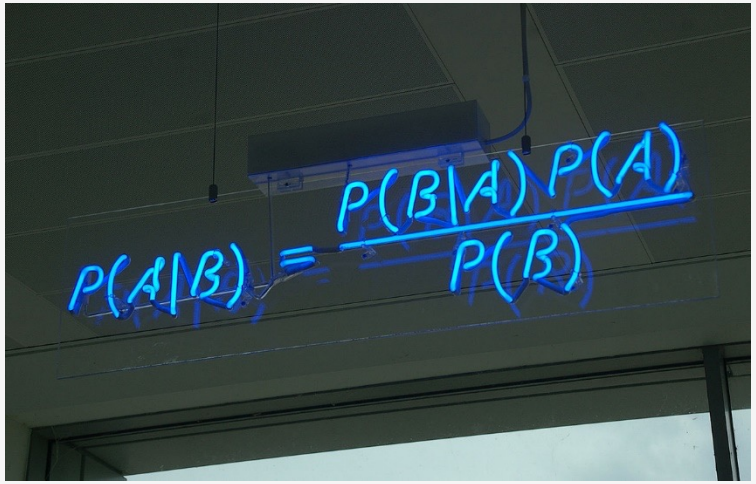
$$P(A \text{ and } B) = P(A) * P(B \text{ given } A)$$
$$P(\text{rain and umbrella arrives}) = .30 * .20 = .06$$

6% chance it will rain and your umbrella will arrive.

Bayes Theorem

It is worth bringing up Bayes Theorem it allows us to flip a conditional probability, turning $P(A\ given\ B)$ into $P(B\ given\ A)$.

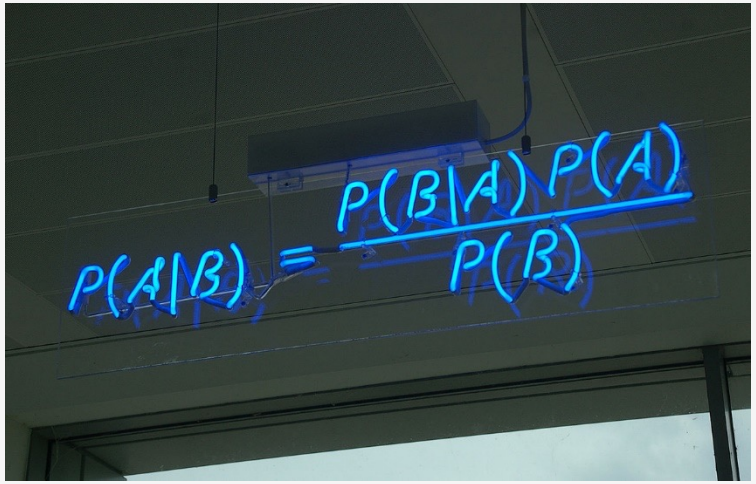
$$P(A\ given\ B) = \frac{P(B\ given\ A) * P(A)}{P(B)}$$



Bayes Theorem

Let's look at the color blindness problem again. We have these data points:

$$P(color\ blind) = .0425$$
$$P(male) = .50$$
$$P(color\ blind\ given\ male) = .08$$



What is the probability of a *person being male given they are color blind*?

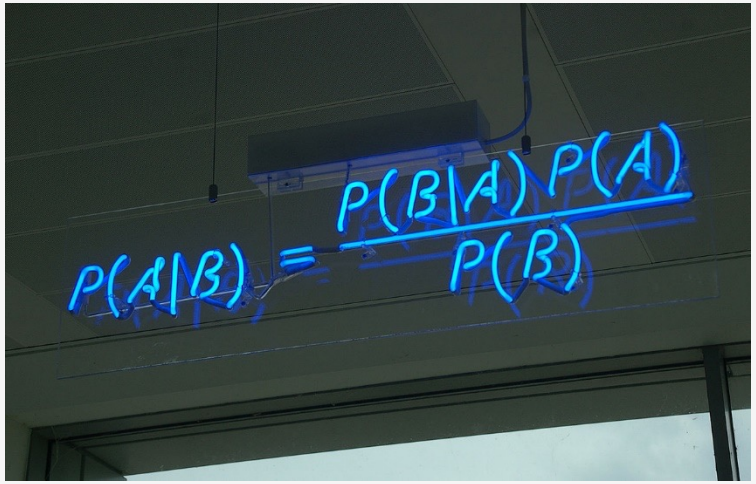
$$P(A\ given\ B) = \frac{P(B\ given\ A) * P(A)}{P(B)}$$

$$P(male\ given\ color\ blind) = \frac{.08*.5}{.0425} = .9411$$

Bayes Theorem

The probability of a person *being male given they are color blind* is 94.11%, which is much different than the probability of *being color blind given they are a male*, which is 8%.

Bayes Theorem is the core of conditional probability, as it allows us to flip inferences based on conditions.



Section IV

Discovering the Binomial and Beta Distribution

Discussion: Measuring Uncertainty

You and an engineer colleague are testing a system to figure out its rate of failure, and you want at least 90% success rate.

Each test is expensive and time-consuming, but you were able to complete 10 tests.

8/10 of the tests succeeded but you are dissatisfied that just 10 tests already yielded two failures, and thus conclude an 80% failure rate.

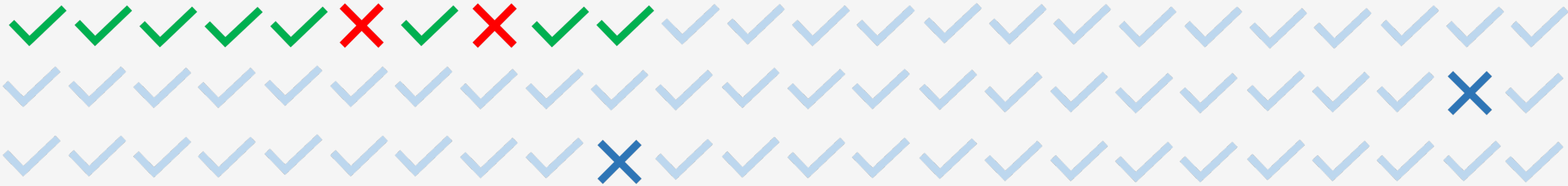
TEST OUTCOME:

Discussion: Measuring Uncertainty

You propose going back to the drawing board, but the engineer wants to run more tests arguing this could be a fluke. The only way we will know for sure is to run more tests.

He argues “what if more tests might yield 90% or greater success? Consider the possible scenario below:”

POSSIBLE FUTURE TEST OUTCOME:



Discussion: Measuring Uncertainty

You *really* do not want to do more tests, but you consider the engineer's argument. After all, if you flip a coin 10 times and get 8 heads, it does not mean the coin is fixed at 80%.

DISCUSSION

Is the success rate really 80%?

How do we determine whether 10 tests are enough to discover the underlying probability of success?

If the true success rate is 90% or more, could we see 8 out of 10 successes by chance?

Why or why not?

Discussion: Measuring Uncertainty

Getting more data is always helpful, but sometimes expense and availability prevents us from doing so.

In situations where data is limited, a Bayesian approach can be advantageous over a Frequentist one.

Think about the probabilities of probabilities… if an event inherently has a 90% probability of occurring, what is the likelihood I would get 8/10 successes?

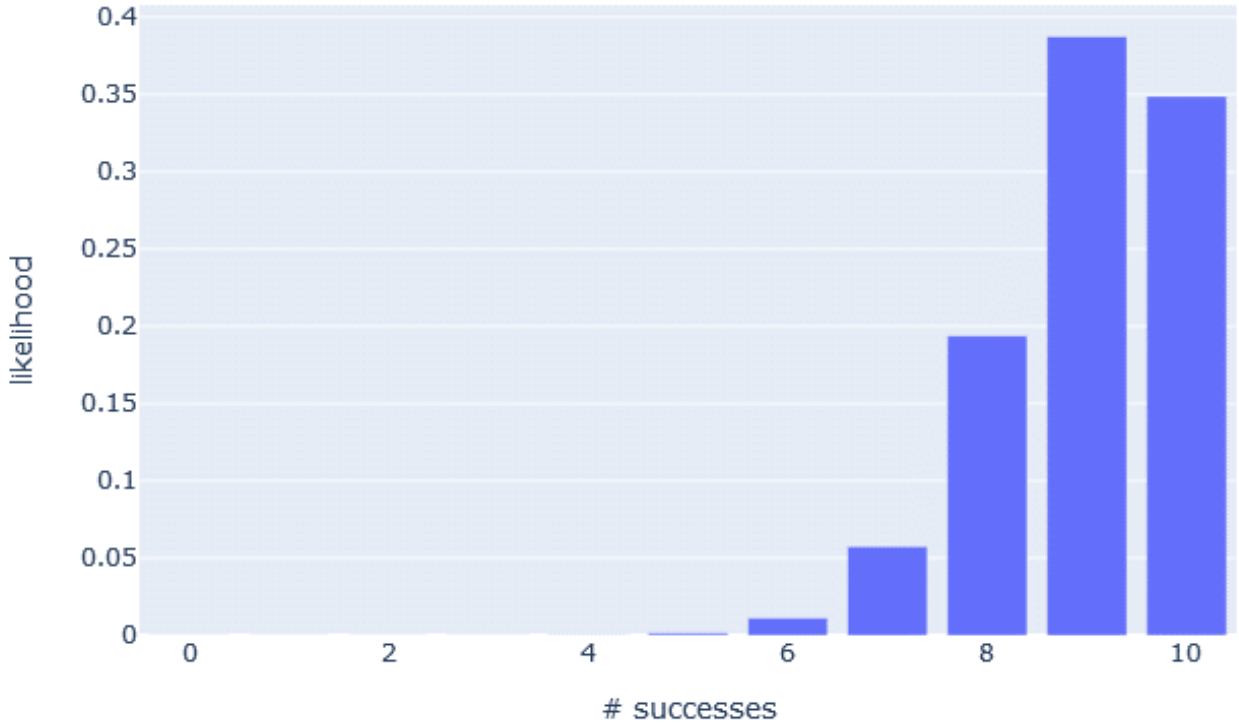
The Binomial Distribution

Let's first consider the **binomial distribution** to the right, which shows the *y likelihood* of *x successes* when the probability of success is 90% and there are 10 trials.

Each bar represents the probability of getting x successes.

If the probability of success is 90%, there is a 19.37% probability I will get 8 successes out of 10 trials.

Notice that all the outcomes add up to 1.0, or 100%, so this is a **probability distribution**.

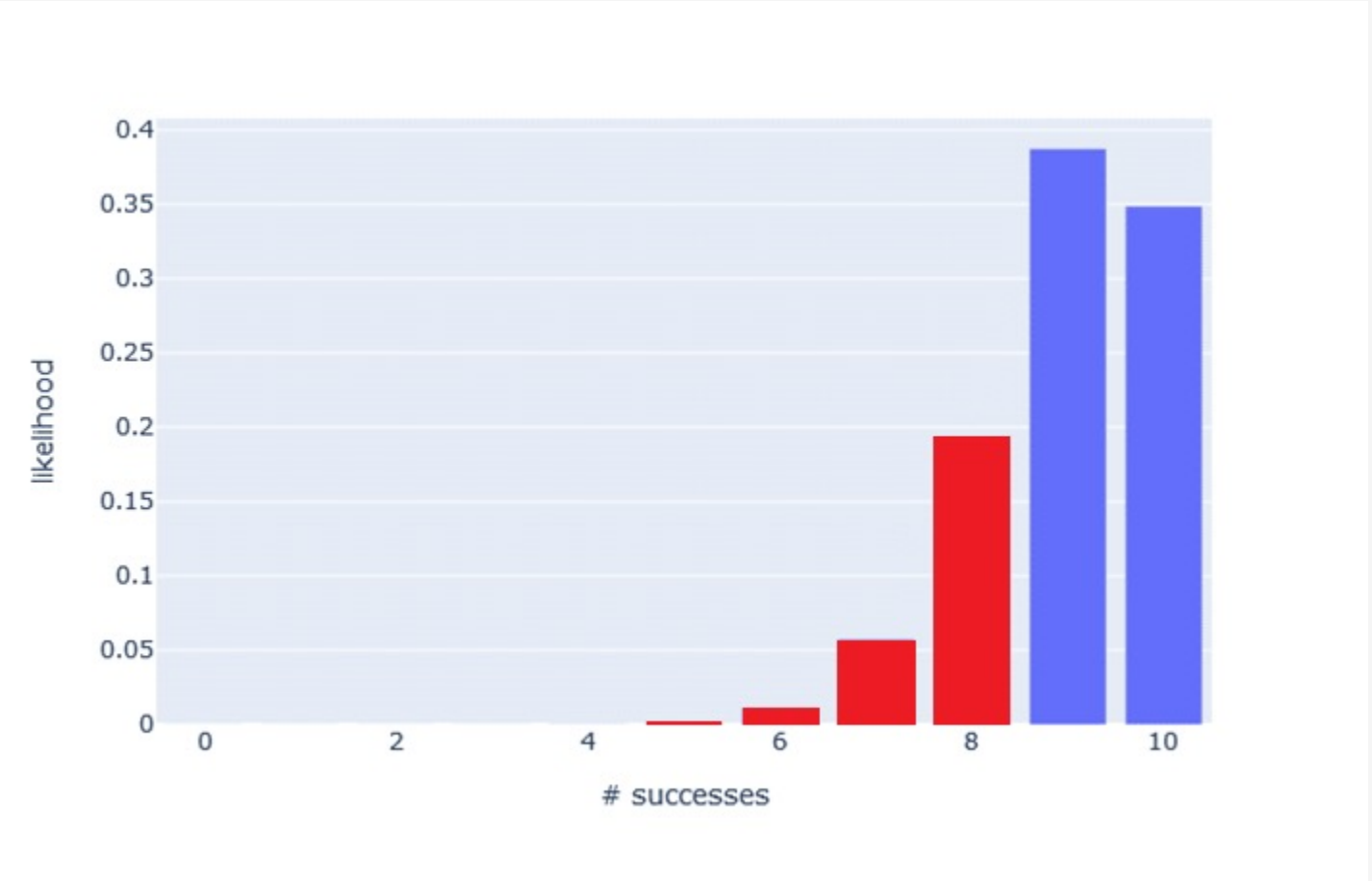


The Binomial Distribution

How would I calculate the probability of 8 or less successes?

Sum the probabilities (the bars in red) and you will find the probability of 8 or less successes to be 26.39%

If we are interested in having 90% or greater success, then the probability of seeing less than 90% success is 26.39% **EVEN IF** each outcome has a 90% success rate.



Binomial Distribution in Python

```
from scipy.stats import binom

n = 10
p = 0.9

for x in range(n + 1):
    probability = binom.pmf(x, n, p)
    print("{0} - {1}".format(x, probability))
```

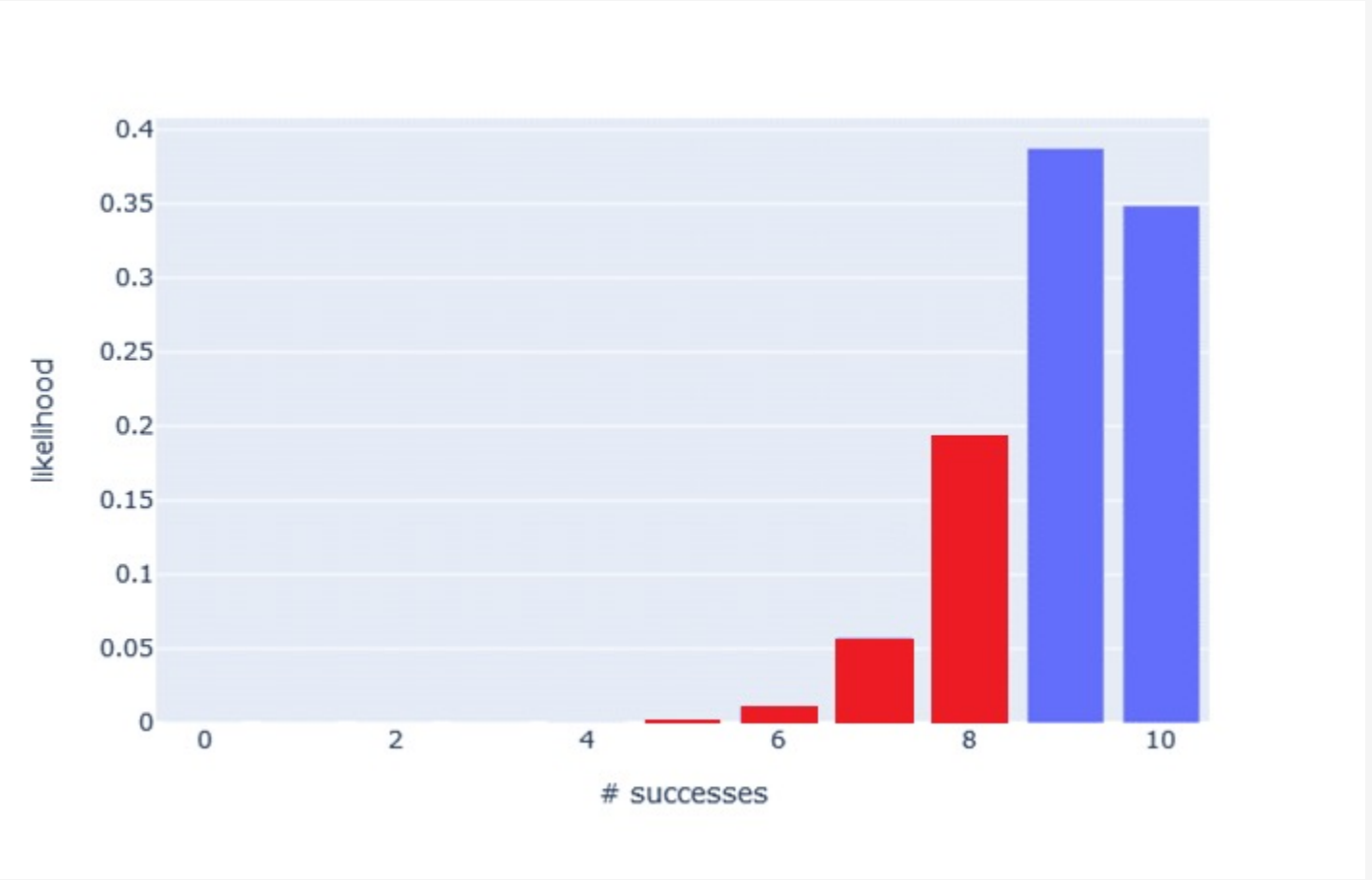
0 - 9.99999999999996e-11
1 - 8.99999999999996e-09
2 - 3.64499999999996e-07
3 - 8.74800000000003e-06
4 - 0.000137780999999999
5 - 0.00148803479999999 88
6 - 0.0111602609999999 96
7 - 0.057395628000000 01
8 - 0.193710244499999 93
9 - 0.387420489000000 37
10 - 0.348678440100000 04

The Binomial Distribution

So there is a 26.39% chance the engineer is correct, ***ASSUMING*** each success has a 90% success rate.

This is informative showing it's 26.39% possible to see 8 or less successes with an underlying 90% probability.

But notice that our model ***ASSUMED*** the probability of success is 90%. What if we flipped the question and explored other probabilities besides 90%, and how likely each would produce 8/10 successes?



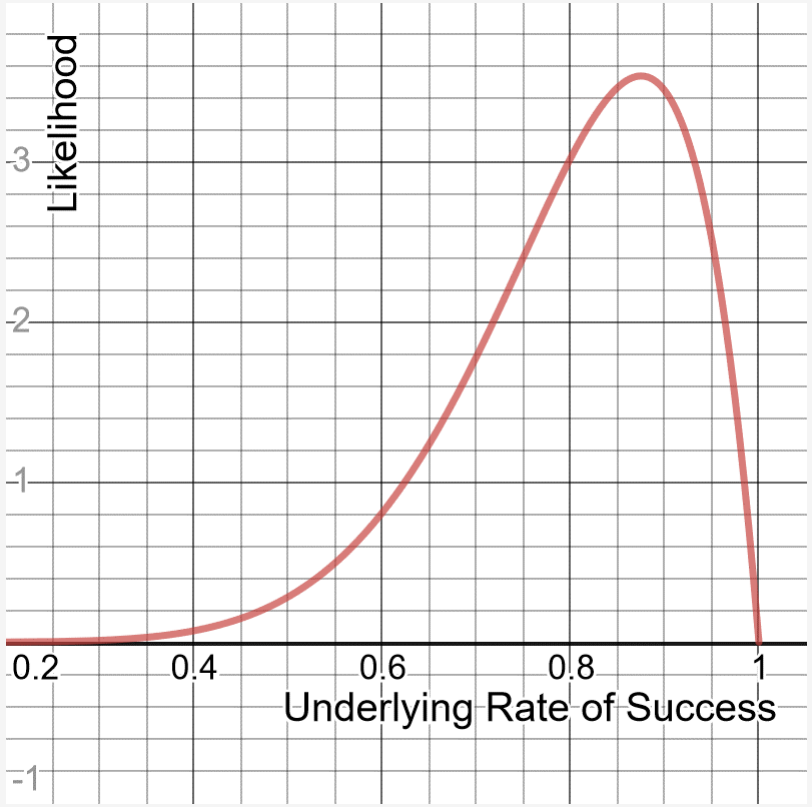
The Beta Distribution

Here's another way to look at this: consider all underlying rates of success and how likely each is to produce 8/10 successes.

This is the **beta distribution**, which allows us to see the probabilities of probabilities given so many successes and failures.

To the right shows the range of likelihoods for each actual rate of success, when 10 trials yields 8 successes.

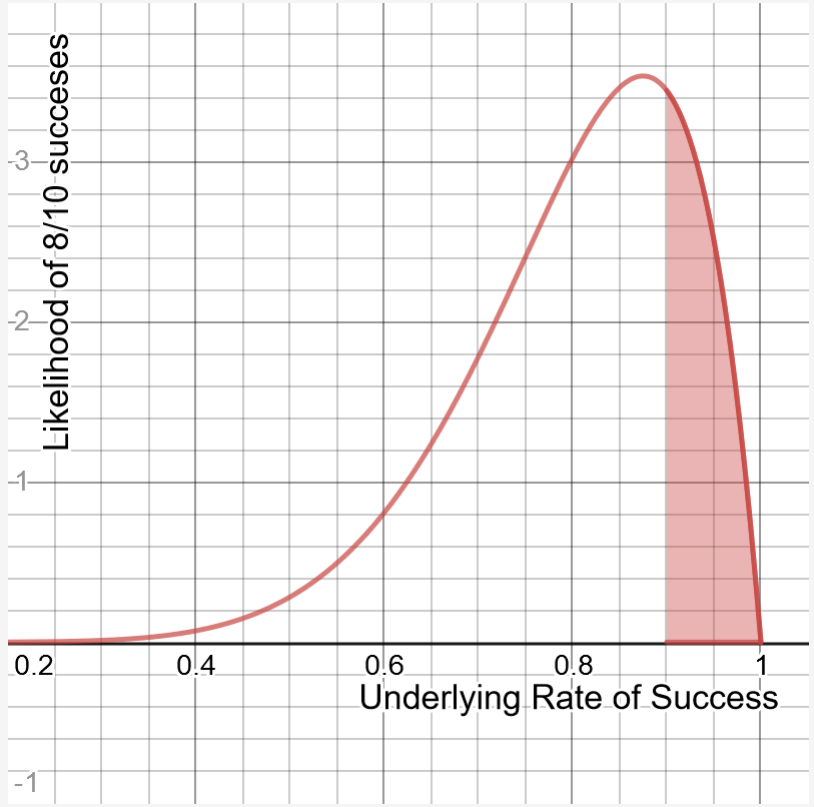
The beta distribution accepts two arguments, the **alpha** which is the number of successes, and the **beta** which is the number of failures.



https://www.desmos.com/calculator/zyranflxo4

The Beta Distribution

Given 8/10 successes, the probability that the underlying rate of success is 90% or higher is 22.5%.

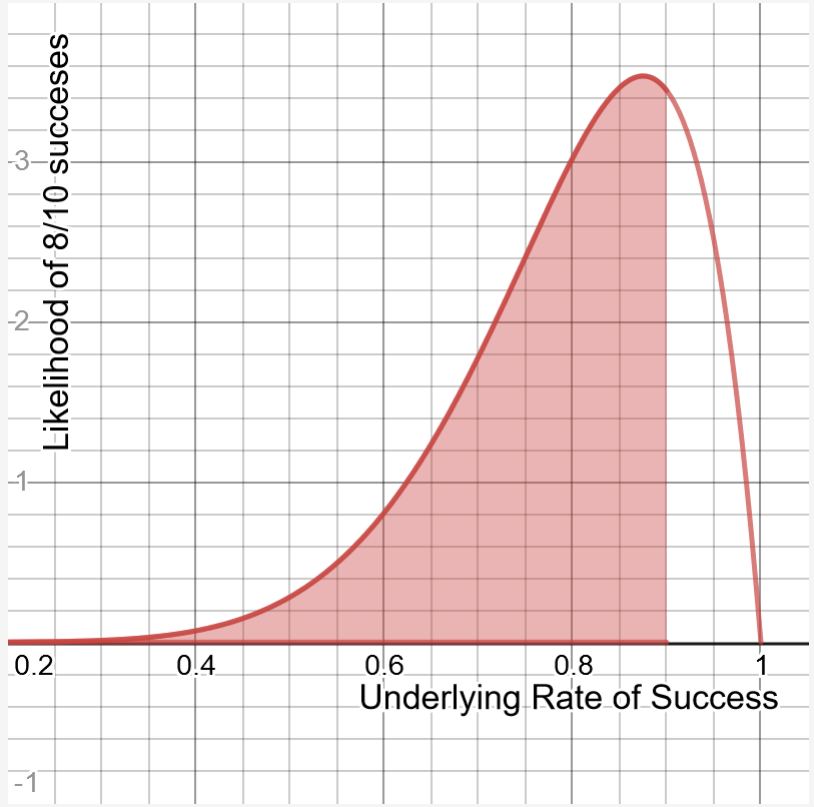


https://www.desmos.com/calculator/cakj42wlie

The Beta Distribution

Given 8/10 successes, the probability that the underlying rate of success is 90% or higher is 22.5%.

But the probability of the underlying rate being less than 90% is 77.5%.



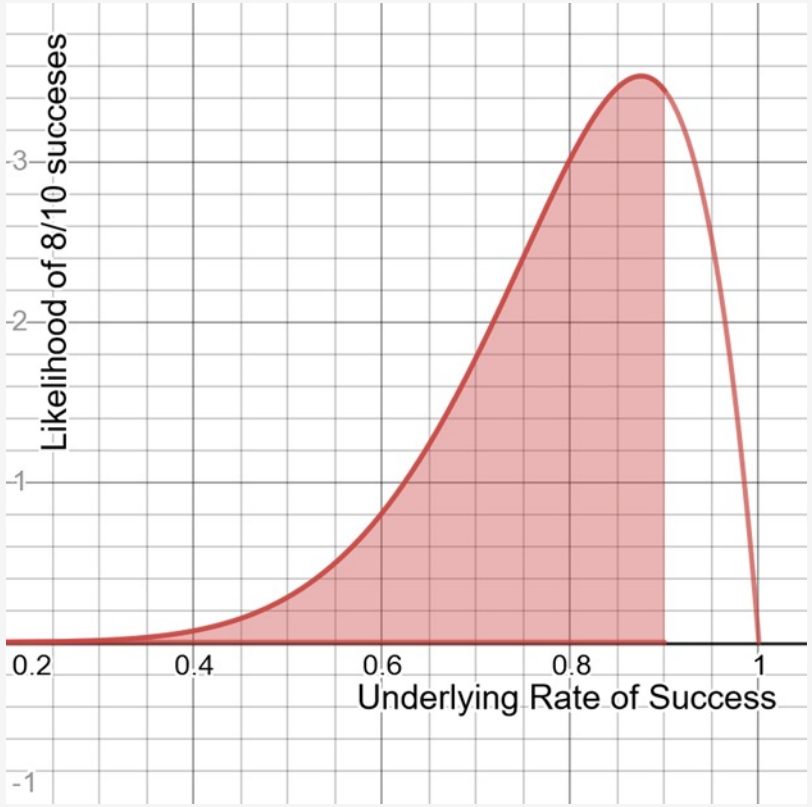
https://www.desmos.com/calculator/cakj42wlie

The Beta Distribution

Given 8/10 successes, the probability that the underlying rate of success is 90% or higher is 22.5%.

But the probability of the underlying rate being less than 90% is 77.5%.

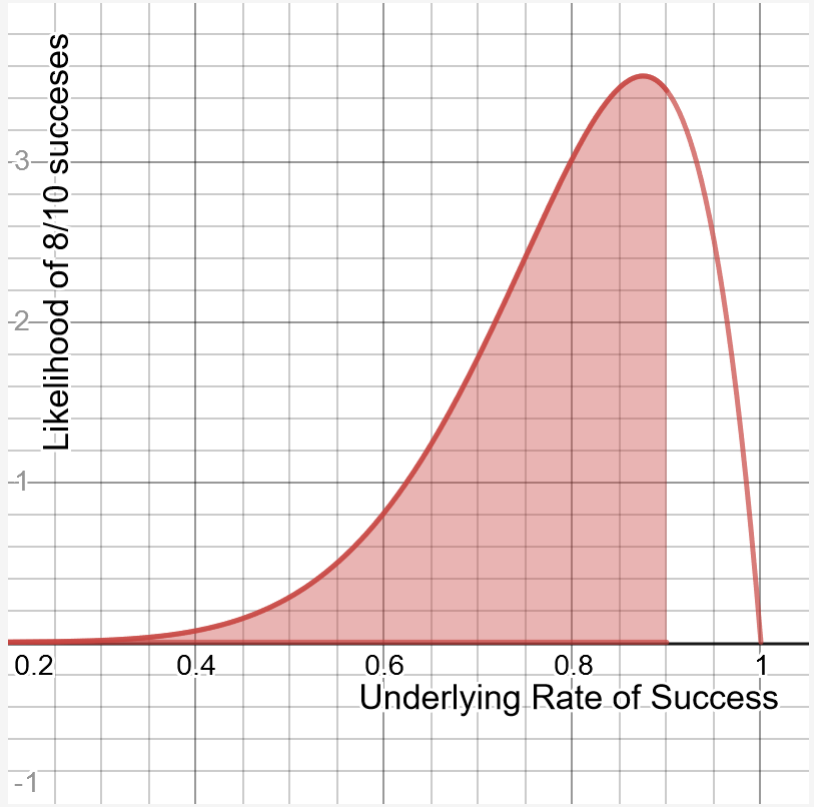
Given these numbers, is it worth doing more tests?



https://www.desmos.com/calculator/cakj42wlie

The Beta Distribution

What we essentially have found is there is a 77.5% chance the underlying success rate is below the required threshold of 90%.

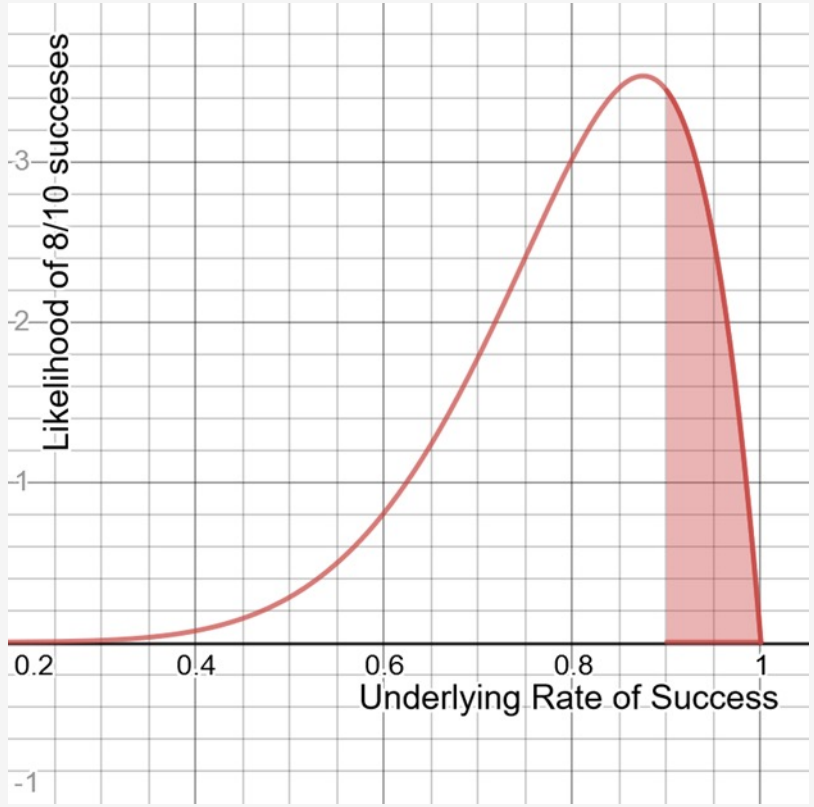


https://www.desmos.com/calculator/cakj42wlie

The Beta Distribution

What we essentially have found is there is a 77.5% chance the underlying success rate is below the required threshold of 90%.

But if we want to gamble on that other 22.5% and order a few more tests to confirm, we can make that call too.



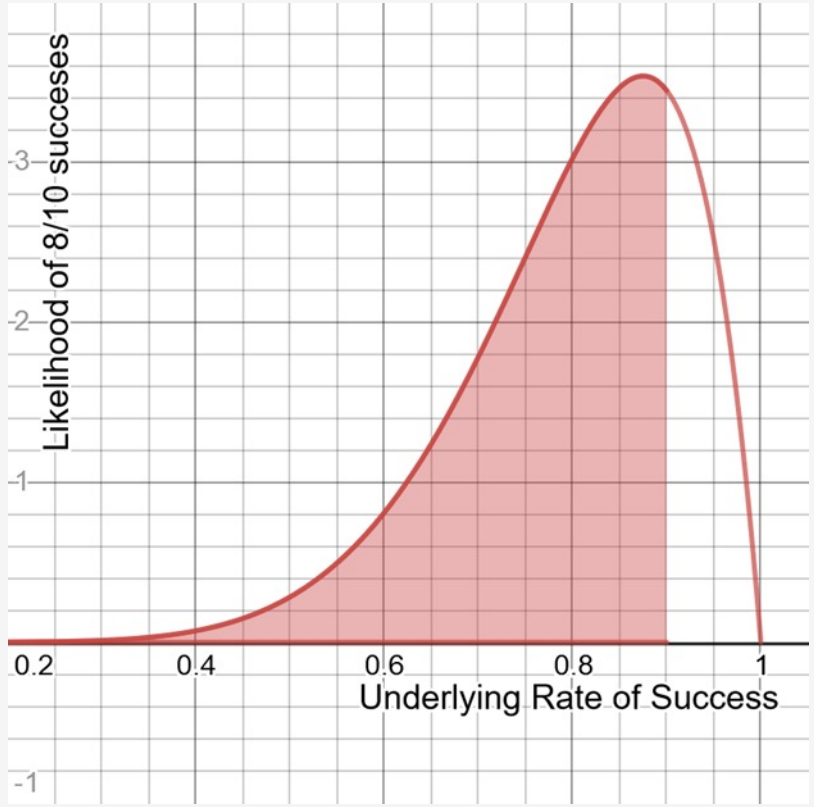
https://www.desmos.com/calculator/cakj42wlie

The Beta Distribution

What we essentially have found is there is a 77.5% chance the underlying success rate is below the required threshold of 90%.

But if we want to gamble on that other 22.5% and order a few more tests to confirm, we can make that call too.

But it may be a risk! Judging just by the numbers, we may just want to abandon testing because there's a 77.5% chance the test truly failed to meet the 90% success rate.



https://www.desmos.com/calculator/cakj42wlie

Using Beta Distribution to Update Predictions

Let's say our CFO granted funding for 26 more tests and wanted to take that risk.

Accumulating on top of the 8 successes and 2 failures, we now have 30 successes and 6 failures.

Notice our beta distribution got a lot narrower, giving us a stronger **credible interval**.

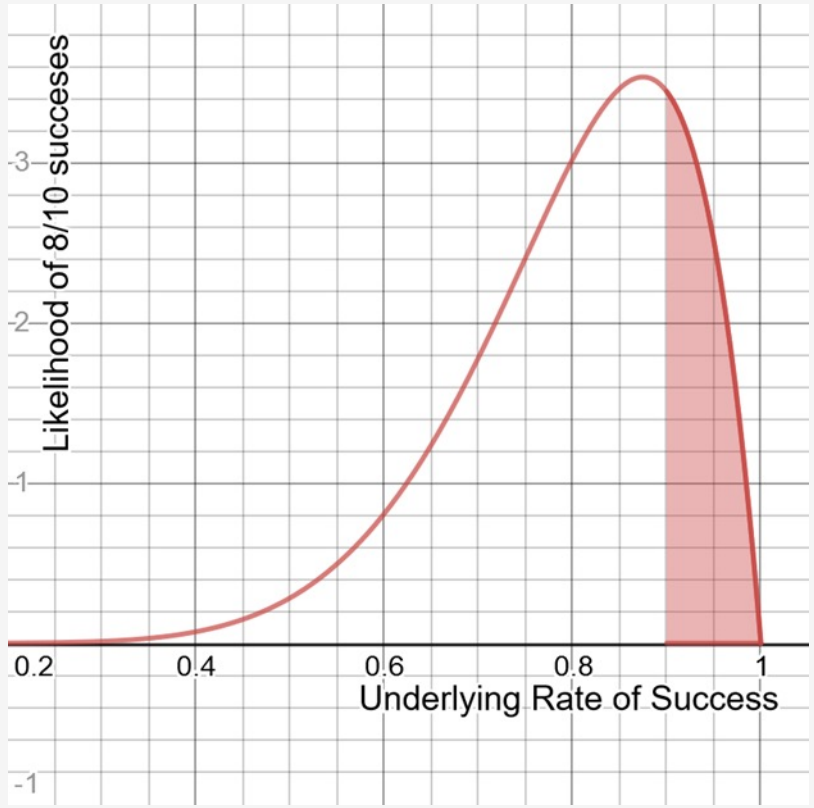
The probability of the underlying success rate being less than 90% is now 86.8%, leaving only 13.2% possibility that our system meets requirements.

At this point, it may be time to abandon testing and go back to the drawing board, unless you want to continue gambling on that 13.2%!

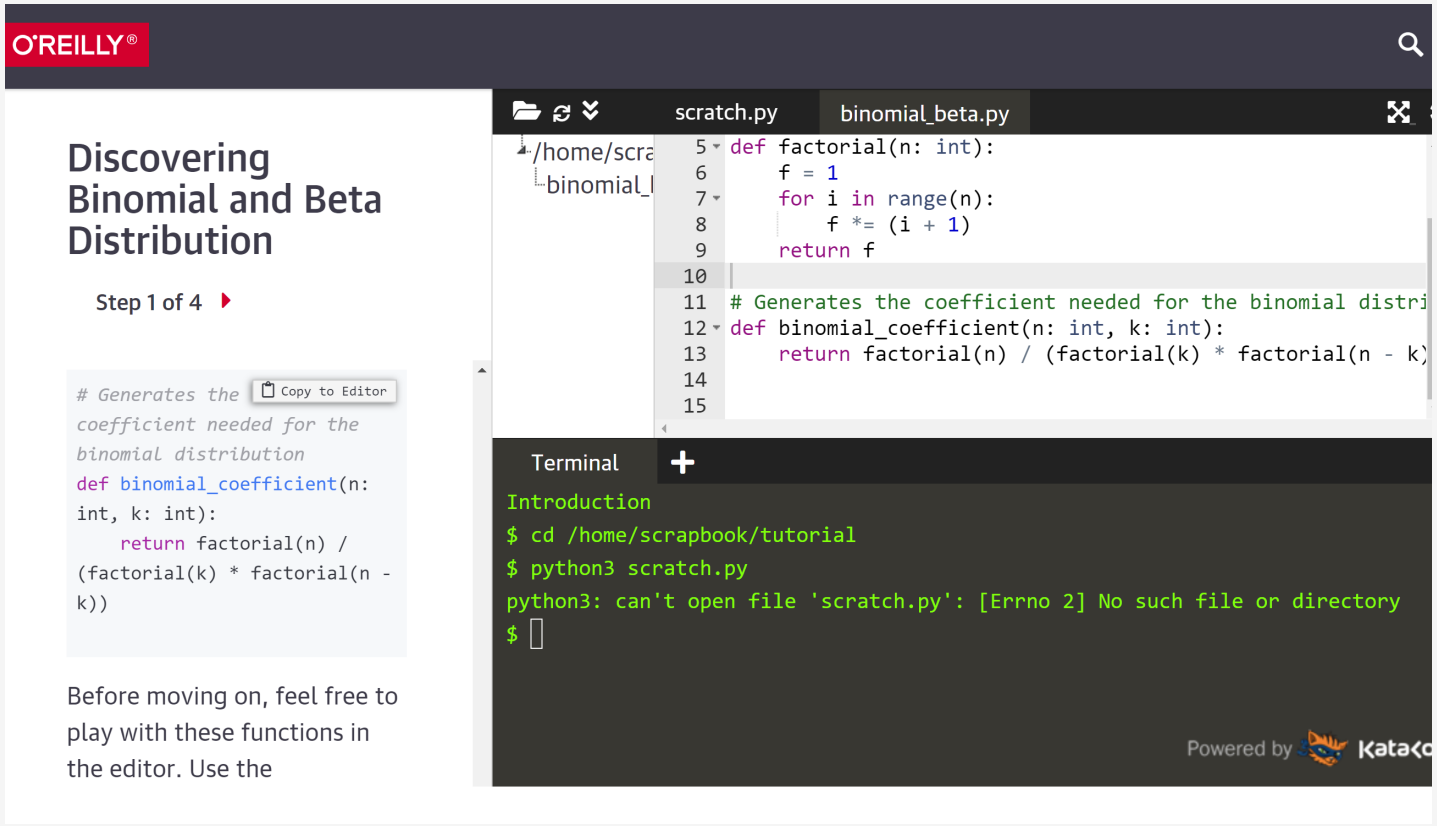
https://www.desmos.com/calculator/xnpjwqclnc

Beta Distribution in Python

```
from scipy.stats import beta

a = 8
b = 2

p = beta.cdf(1.0, a, b) - beta.cdf(.90, a, b)

print(p)
```



Katacoda Interactive Scenario – Binomial and Beta Distribution

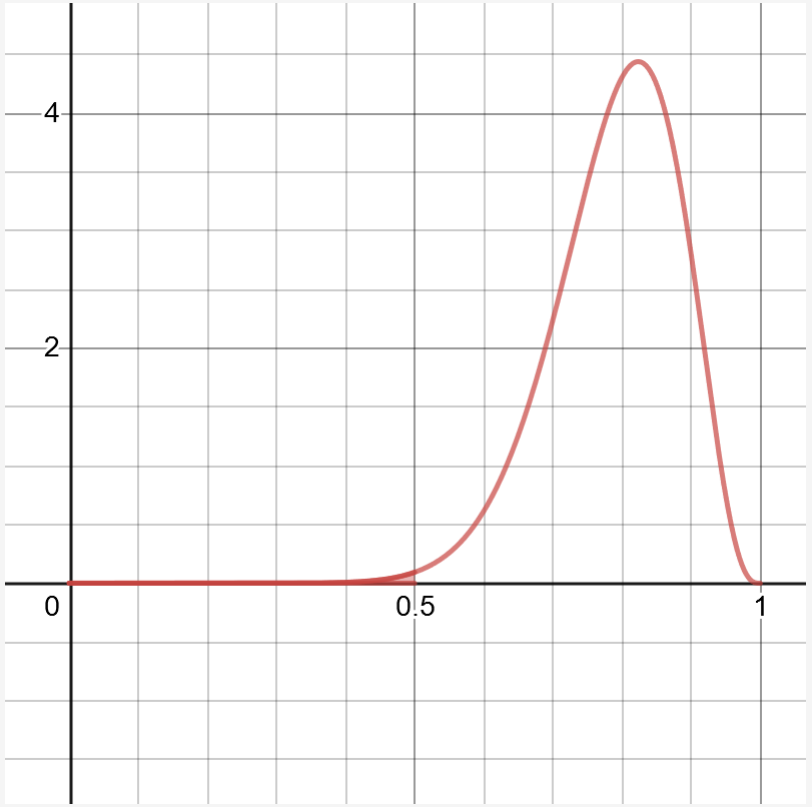


https://learning.oreilly.com/scenarios/probability-from-scratch/9781492080565/

Exercise

You flipped a coin 19 times and got heads 15 times and tails 4 times.

Do you think this coin has any good probability of being fair?



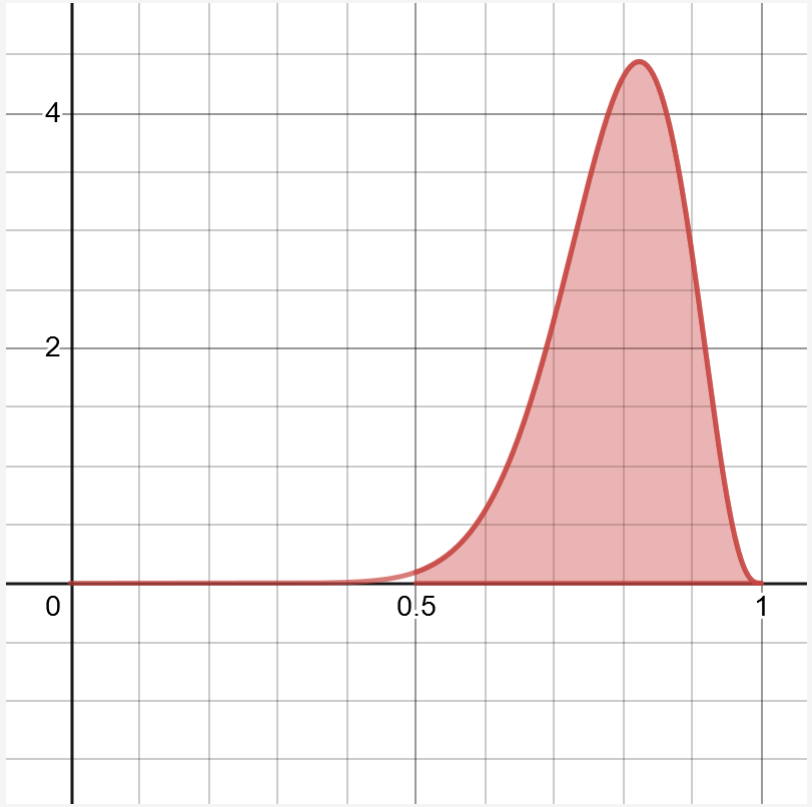
Exercise

You flipped a coin 19 times and got heads 15 times and tails 4 times.

Do you think this coin has any good probability of being fair?

Considering 99.62% of the density is above 50%, this coin is highly unlikely to be fair.

```
from scipy.stats import beta

a = 15
b = 4

x = beta.cdf(1.0, a, b) - beta.cdf(.50, a, b)

print(x)
```



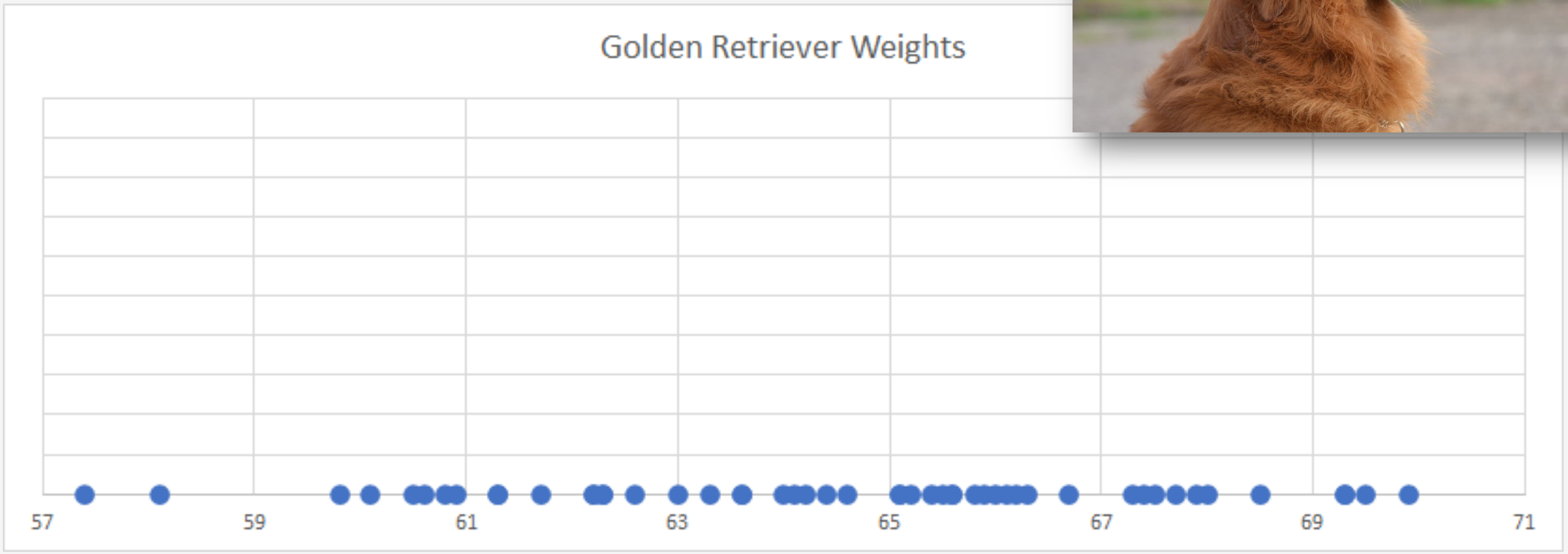
Section V

Discovering the Normal Distribution

Discovering the Normal Distribution

Below is a number line showing recorded weights (in pounds) for 50 adult golden retrievers.

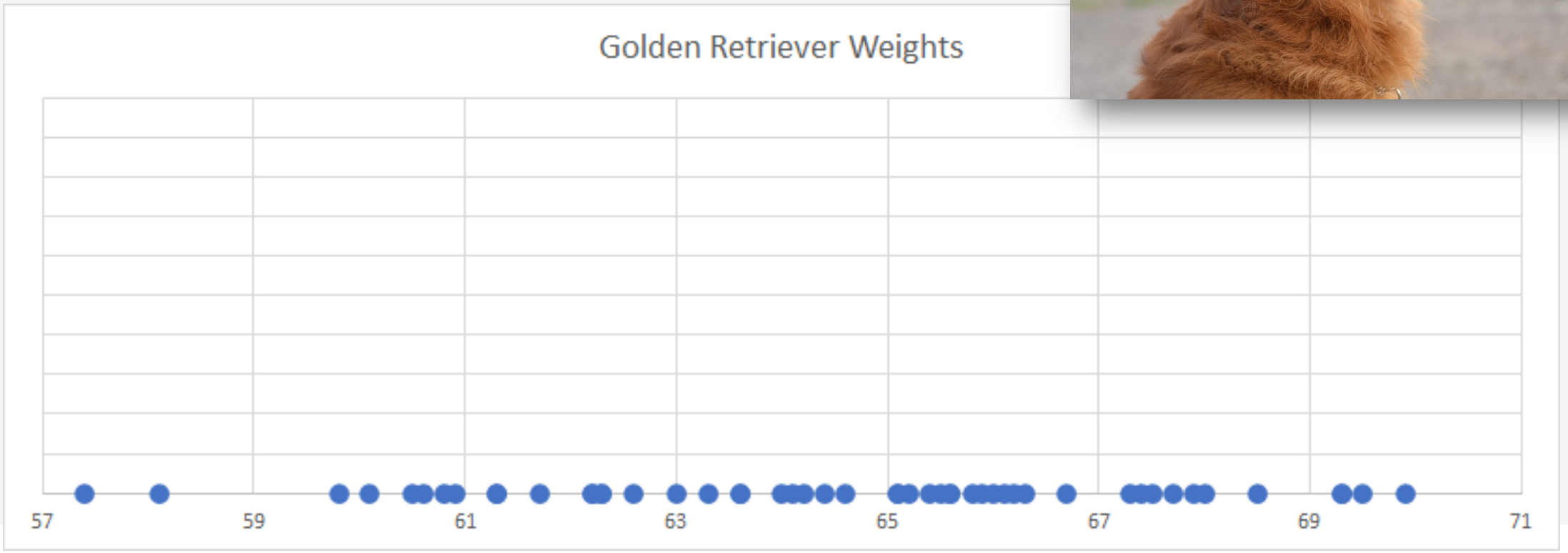
What do you notice about this data?



Discovering the Normal Distribution

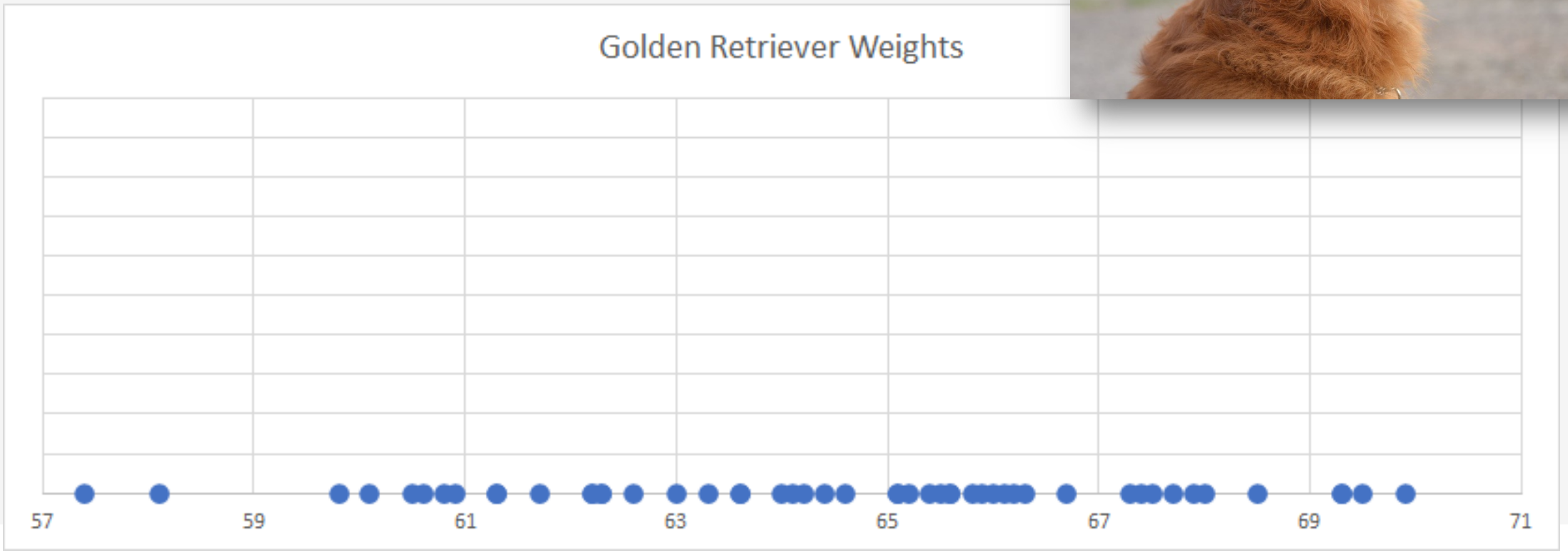
Notice that towards the “center” there are more points clustered together, while the “tails” on either end have less points.

We can infer that we are more likely to see more golden retrievers in the 61-67 lb range as opposed to lower or higher weights like 57 and 71.



Discovering the Normal Distribution

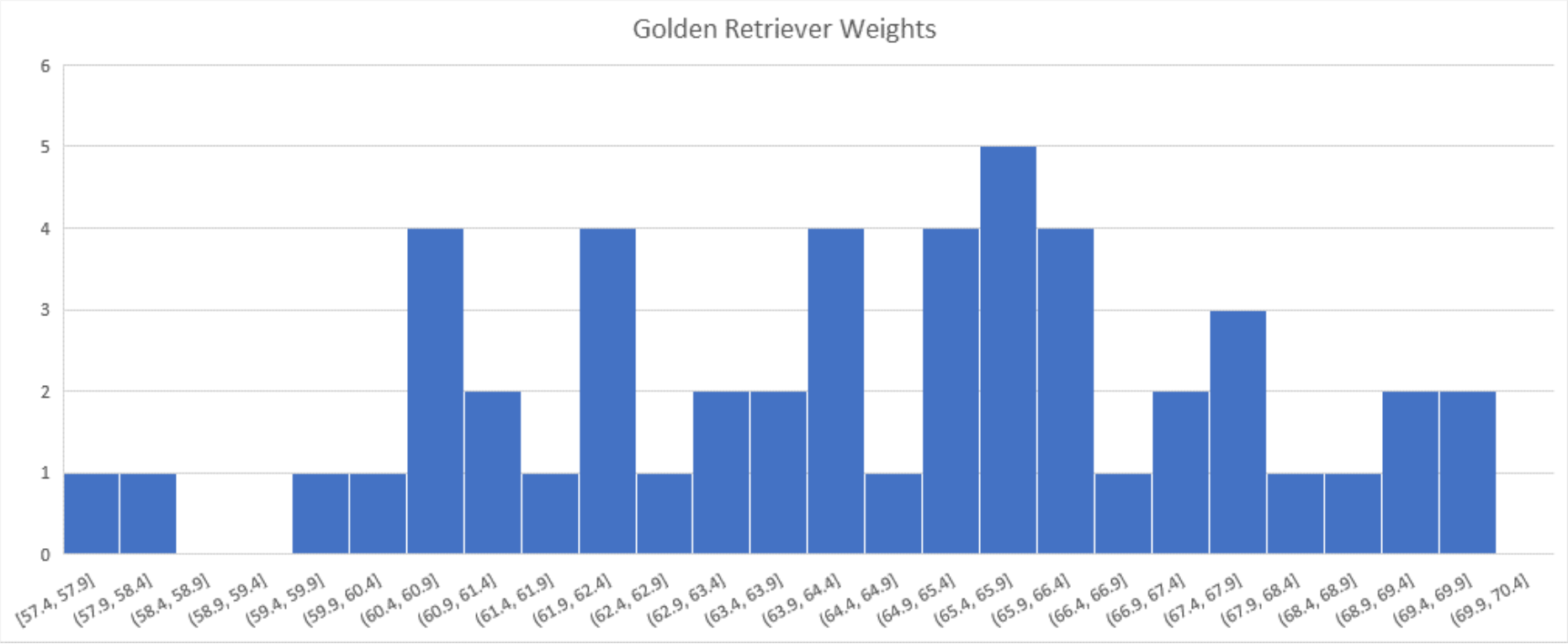
Is there a better way we can visualize this data? And get an idea where we would see more golden retrievers in terms of their weight?



Discovering the Normal Distribution

We can try to **bin** up points on equally-sized ranges, and then count the number of instances as a bar chart to create a **histogram**.

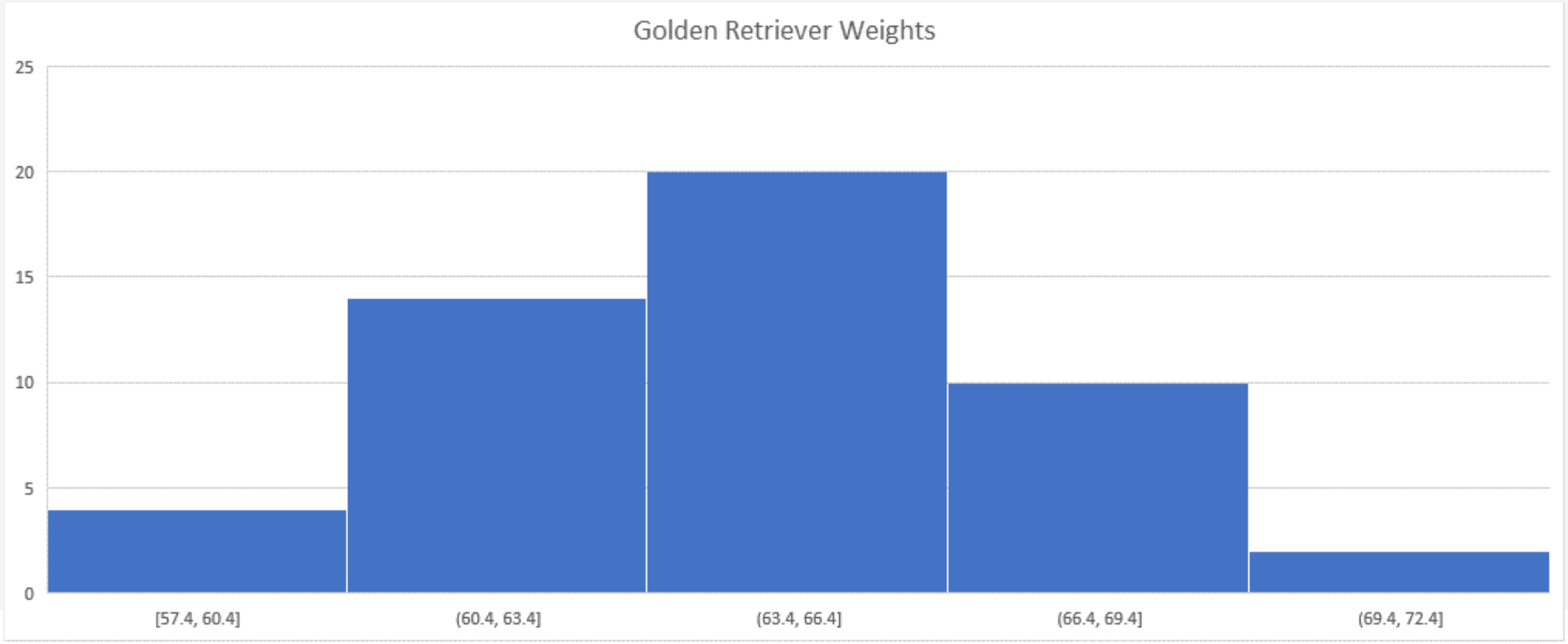
Unfortunately, we don't have an infinite amount of data so making the bin ranges too small will not reveal much about the underlying distribution.



Discovering the Normal Distribution

But if we make the bin sizes just right, we might be able to see a shape resembling a probability distribution.

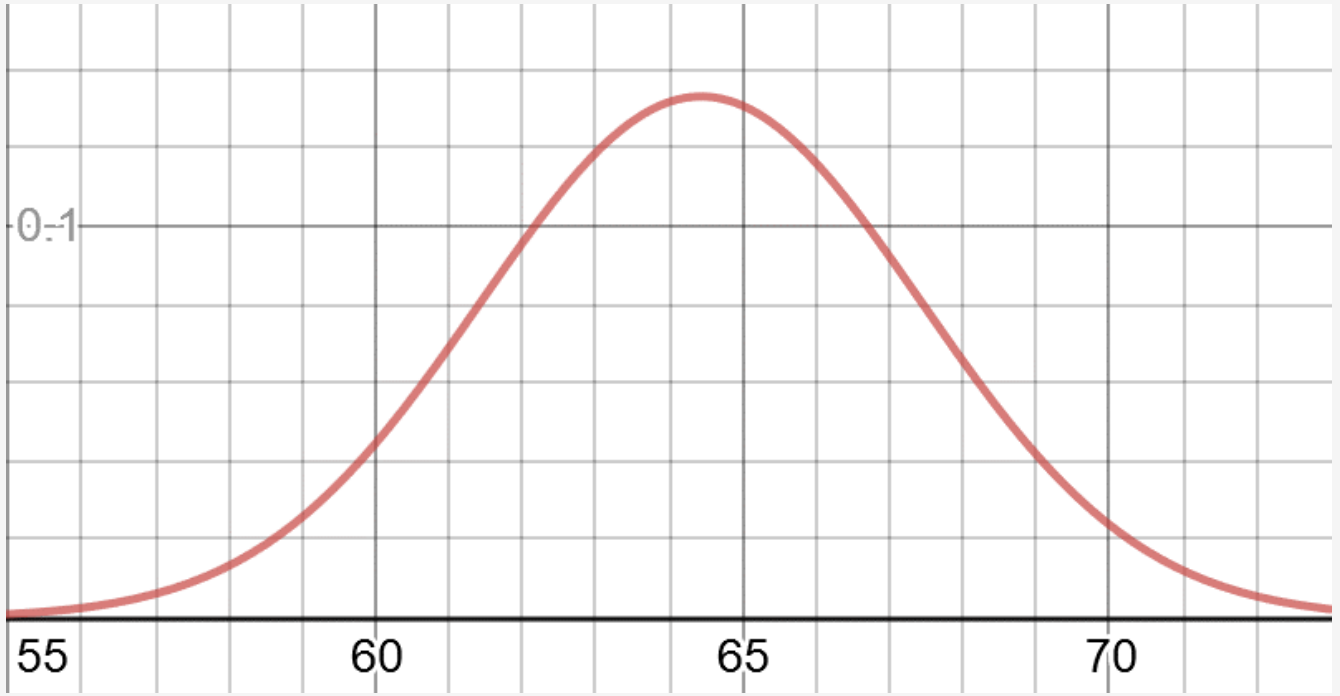
There are many types of distributions, but in this case we can see a nice bell curve known as the **normal distribution**, also called the **Gaussian Distribution**.



Discovering the Normal Distribution

We can fit this curve onto a histogram to make inferences about the entire population, as it's unlikely I will be able to weigh every golden retriever in existence.

There are ways to measure uncertainty whether a bell curve resembles the population's bell curve, but we will steer clear of statistics in this course.



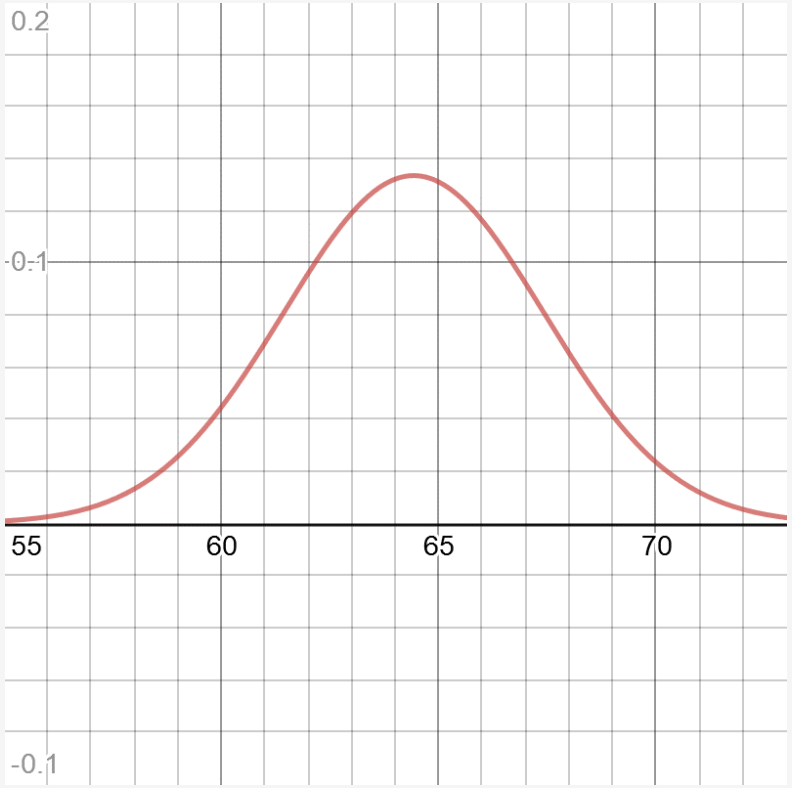
https://www.desmos.com/calculator/yftpag0tse

Discovering the Normal Distribution

The normal distribution has several important properties that make it useful:

- Symmetrical
- Most mass is at the center around the **mean**
- Has a spread (being narrow or wide) that is specified by **standard deviation**.
- The "tails" are the least likely outcomes, and approach zero infinitely but never touch zero.
- It resembles a lot of phenomena in nature and daily life, and even generalizes non-normal problems because of the central limit theorem.

We can expect any golden retriever to have a weight most likely around 64.43 (the mean), but highly unlikely around 55 or 73.



https://www.desmos.com/calculator/yftpag0tse

The Normal Distribution Formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\mu = mean$$
$$\sigma = standard\ deviation$$
$$x = observed\ value$$

$$f(x) = probability\ density\ function$$

Python function:

```
# normal distribution, returns likelihood
def normal_pdf(x: float, mean: float, std_dev: float) -> float:
    return (1.0 / (2.0 * math.pi * std_dev ** 2) ** 0.5) * math.exp(-1.0 * ((x - mean) ** 2 / (2.0 * std_dev ** 2)))
```

Discovering the Normal Distribution

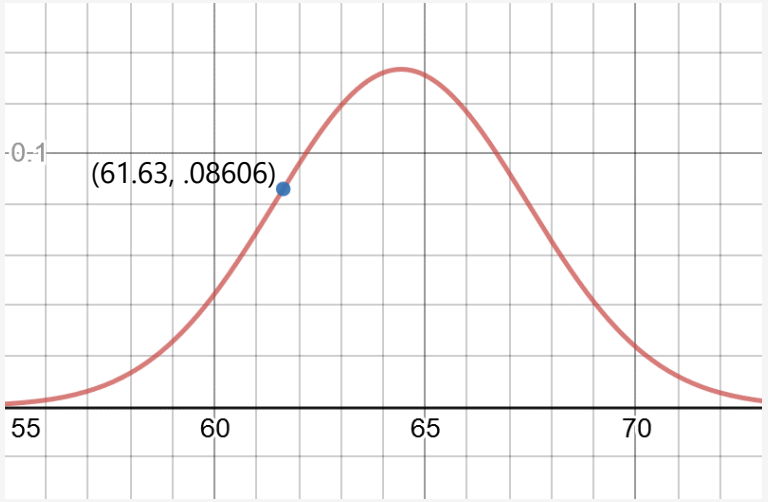
Trick question: What is the probability any golden retriever chosen at random will have a weight of exactly 61.63 lbs?

The answer may surprise you: It is 0%!

This is one of the great paradoxes of continuous distributions like the normal distribution and the beta distribution, as decimals can become so infinitely small there is an infinite number of possibilities on the curve.

The probability of getting a specific value is virtually impossible ***unless*** it is already observed.

The y-axis represents probability density, and to find a probability you need to use the area under a curve for a range of values.



Calculating Probabilities

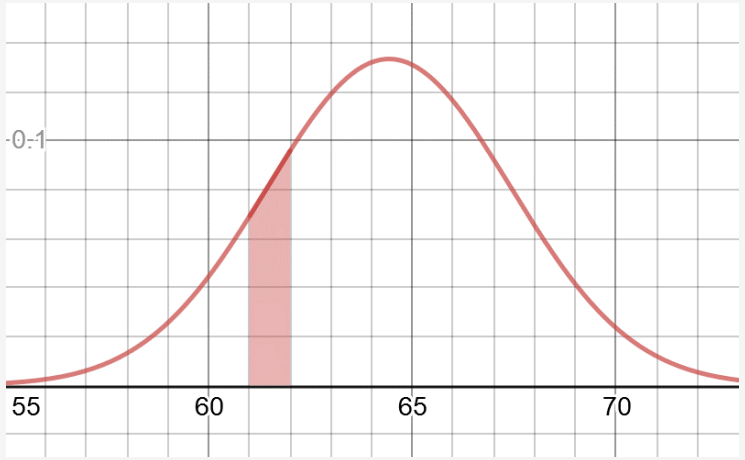
Rather than ask:

"What is the probability a golden retriever will be exactly 61.63 pounds?"

A more productive question would be:

"What is the probability a golden retriever will be ***between*** 61 and 62 pounds?"

The answer is 8.25%, and we will learn how to calculate this using integration just like we did with the beta distribution.



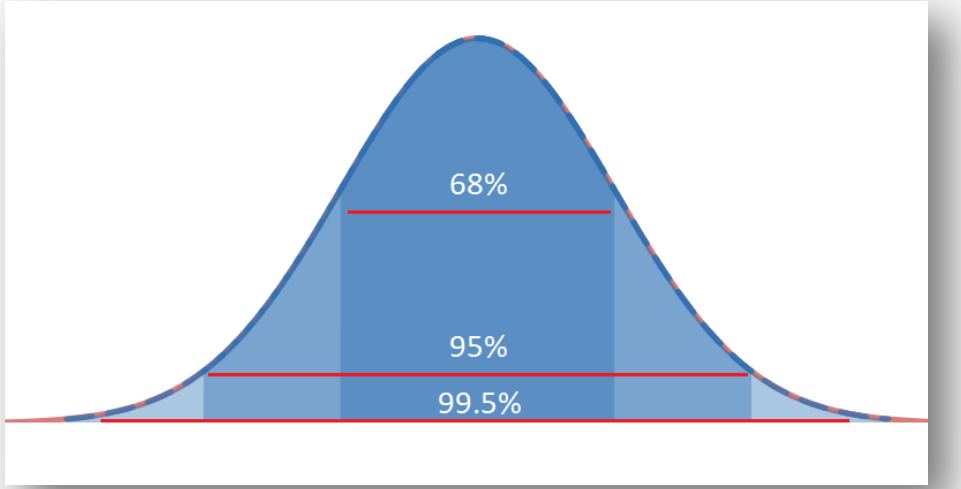
68-95-99.5 Rule

One of the nice features of the normal distribution is the **68-95-99.5 rule**, which states the area/probability within 1, 2, and 3 standard deviations from the mean respectively.

This means that if I have a mean of 10 and standard deviation of 3, then…

- 68% of the probability will be between 7 and 13
- 95% will be between 4 and 16
- 99.5% will be between 1 and 19

You can use this rule to quickly calculate probabilities when using standard deviations.



Cumulative Density Function (CDF)

The way we used integration to find the probability between a range is perfectly valid and intuitive.

Conventionally however, the textbook approach is to use a **cumulative density function** when available, which is a function that returns the area up to a value “x”.

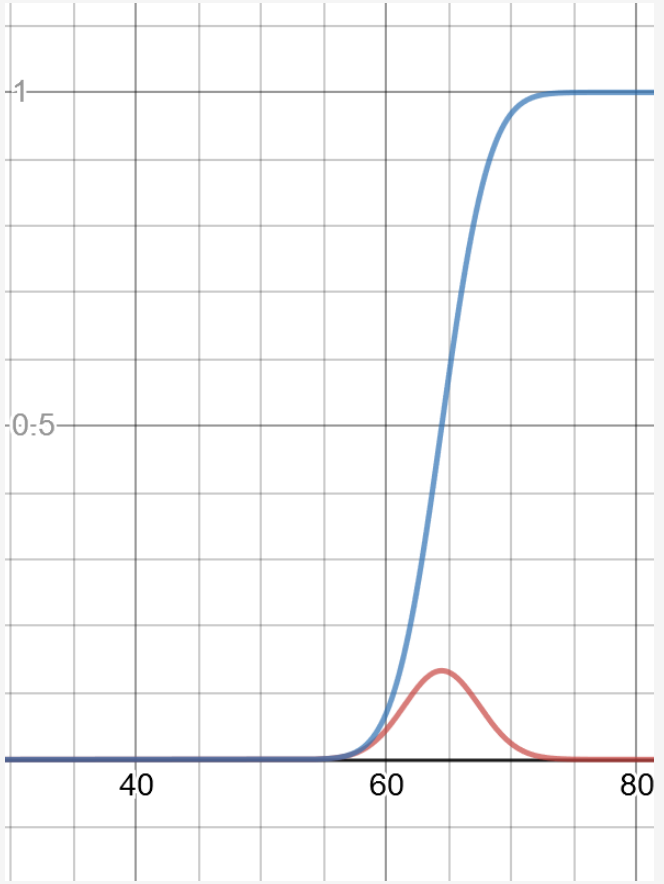
$$\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

The normal distribution’s probability density function (PDF) is shown alongside its cumulative density function (CDF), and the CDF is projecting the area captured up to that x-value on the PDF.

Here is the CDF expressed in Python:

```
def normal_cdf(x:float,mean:float=0.0,std_dev:float=1.0):
    return 0.5*(1+math.erf((x-mean)/((std_dev*2)**0.5)))
```

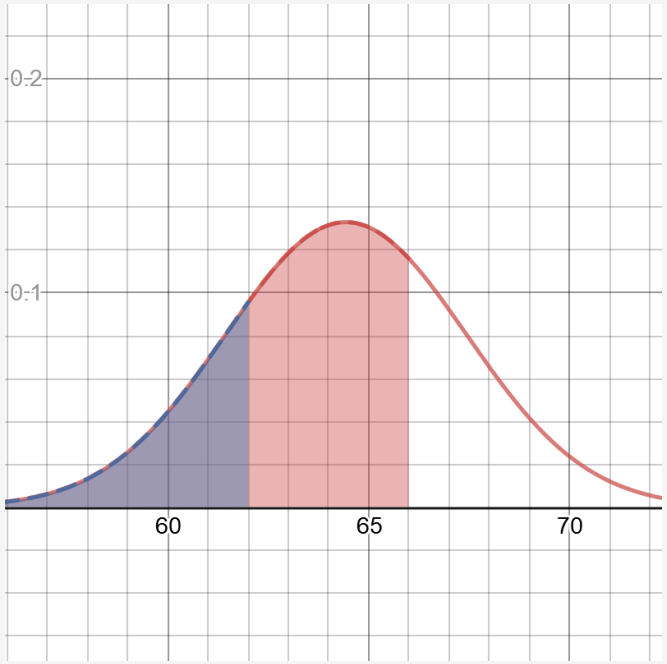
We do cheat a little here and use the `erf` function, and we will leave it a black box as it is beyond the scope of this class.



Cumulative Density Function (CDF)

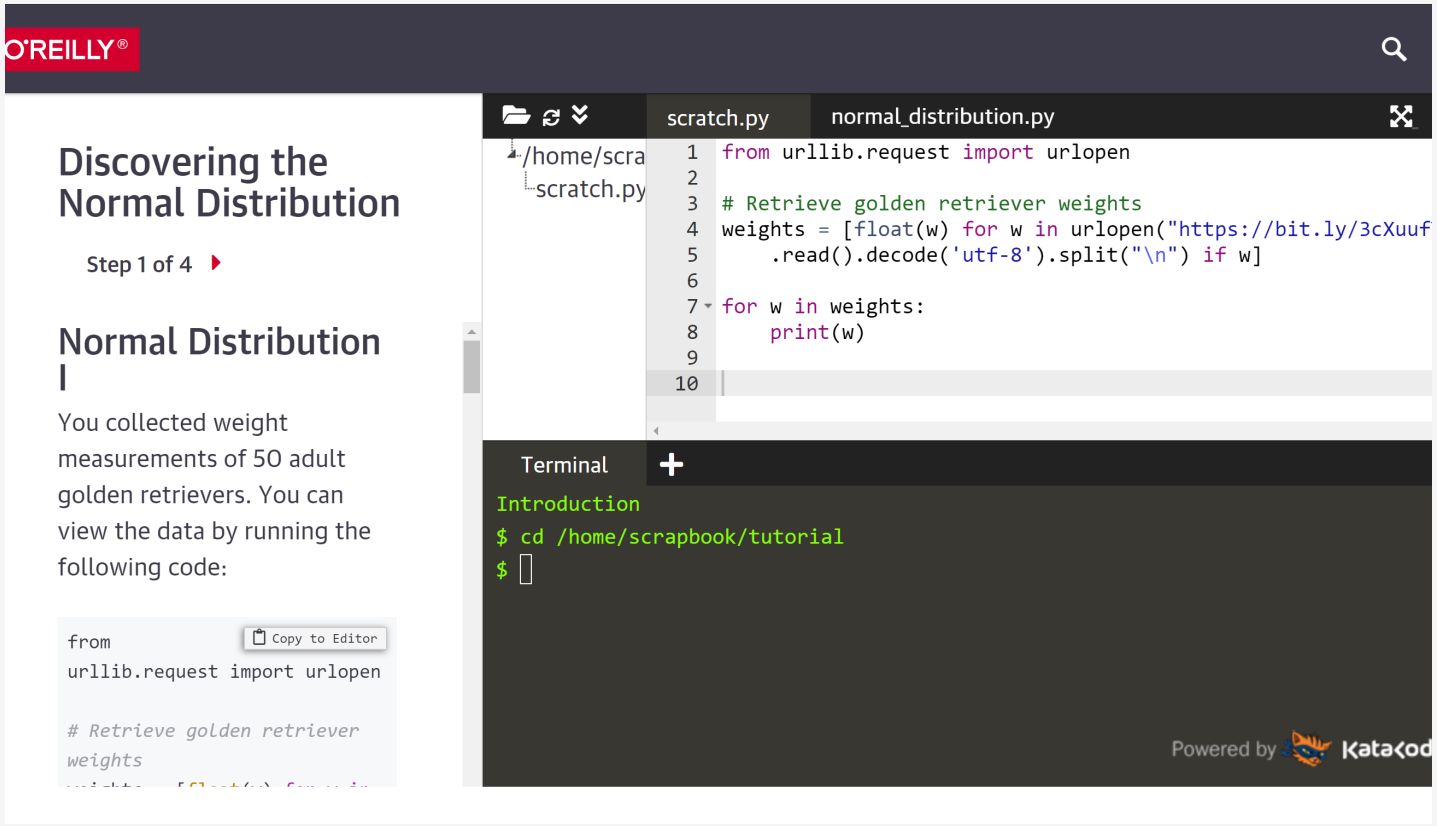
I can use the CDF to find the probability of a golden retriever having a weight between 62 and 66 pounds, without integrating several rectangles and summing their areas.

I first find the area up to 66 (highlighted in red to the right) and then subtract the area up to 62 (highlighted in blue), and we should get 49.2%.




```
from scipy.stats import norm

mean = 64.43
std_dev = 2.99

x = norm.cdf(66, mean, std_dev) - norm.cdf(62, mean, std_dev)

print(x)
```

Katacoda Interactive Scenario – Normal Distribution



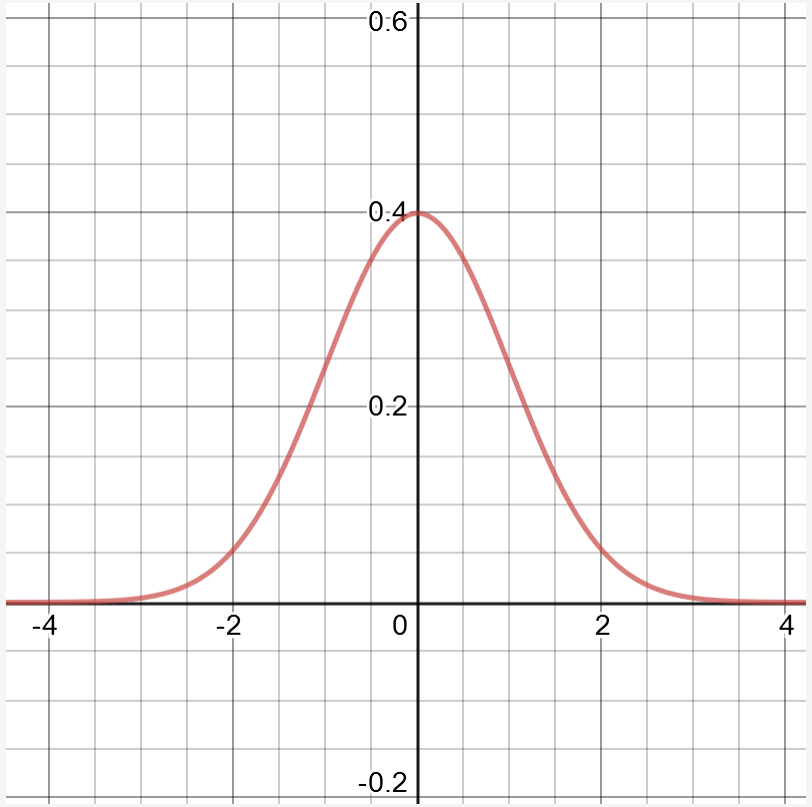
https://learning.oreilly.com/scenarios/probability-from-scratch/9781492080572/

Standard Normal Distribution

A special type of normal distribution is the **standard normal distribution**, which has a mean of 0 and standard deviation of 1.

Sometimes a normal distribution will be converted into a standard normal distribution.

- This creates a convenient way to express all values in terms of the standard deviation, known as **z scores**.
- Turning several normal distributions into standard normal distributions also makes them easier to compare, as comparisons can be made relative to their means rather than absolute values.



Central Limit Theorem

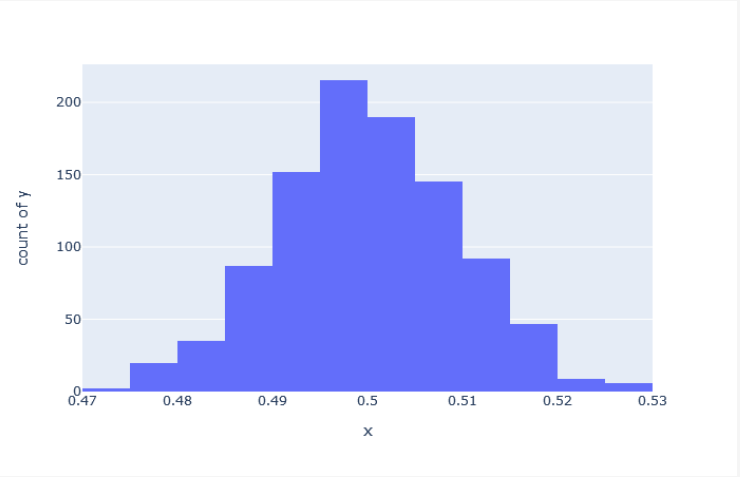
The normal distribution shows up in an important phenomena called the **central limit theorem**, which states that the means of different samples will form a normal distribution no matter what the underlying distribution is.

Say I generated 1000 samples, each containing 1000 random numbers coming from a uniform distribution between 0.0 and 1.0.

If I take the average of each sample and plot them in a histogram, we will see they form a normal distribution!

Even if the data is completely random or has a distribution different from a normal distribution, the averages of the samples will form a normal distribution.

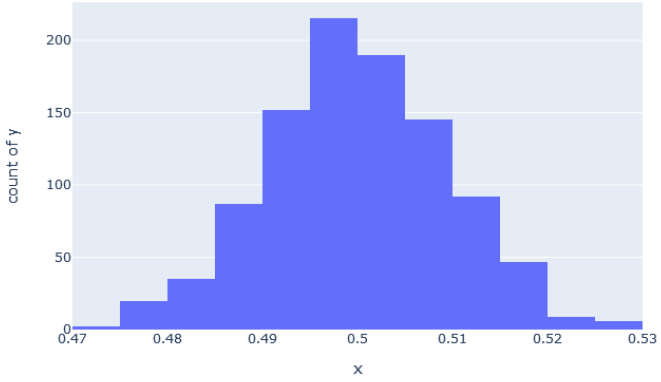
This is another reason why the normal distribution is useful, as it shows up even in data that is not normally distributed.



The above data came from a uniform distribution, and yet the mean of each sample of 1000 data points shows a normal distribution!

Central Limit Theorem

```
# Samples of the uniform distribution will average out to a normal distribution.

import random

import plotly.express as px

# Central limit theorem, 1000 samples each with 1000 random numbers between 0.0 and 1.0
x_values = [(sum([random.uniform(0.0, 1.0) for i in range(1000)]) / 1000.0) for _ in range(1000)]
y_values = [1 for _ in range(1000)]

px.histogram(x=x_values, y = y_values, nbins=20).show()
```



The above data came from a uniform distribution, and yet the mean of each sample of 1000 data points shows a normal distribution!

Other Distribution Types

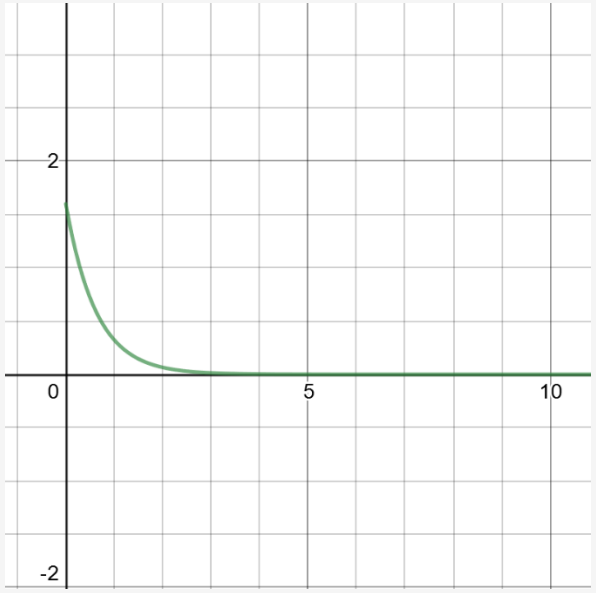
There are many types of probability distributions, some much more niche than others, but here are a few other notable ones:

Poisson – Discrete distribution of probabilities for **x** number of events in a fixed amount of time.

Exponential – Distribution of time between events in a Poisson process

Gamma – A generalization of a two-parameter distribution that includes exponential, Erlang, and chi-squared distributions

Weibull – Often used for survival analysis, failure analysis, and reliability engineering.

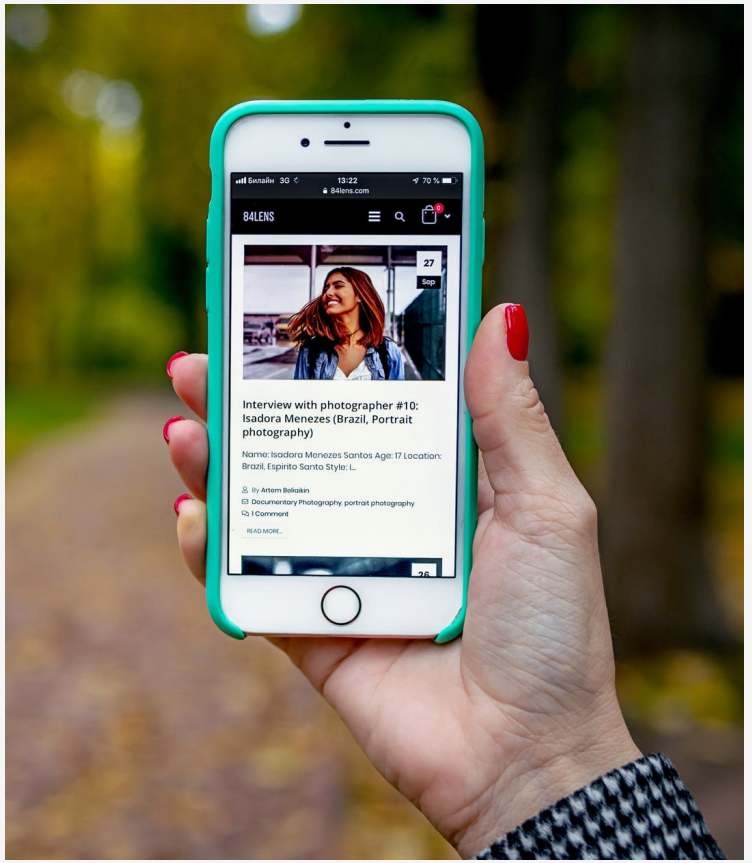


Exponential Distribution

Exercise

A market researcher estimates that the Z-Phone smart phone has a mean consumer life of 42 months with a standard deviation of 8 months.

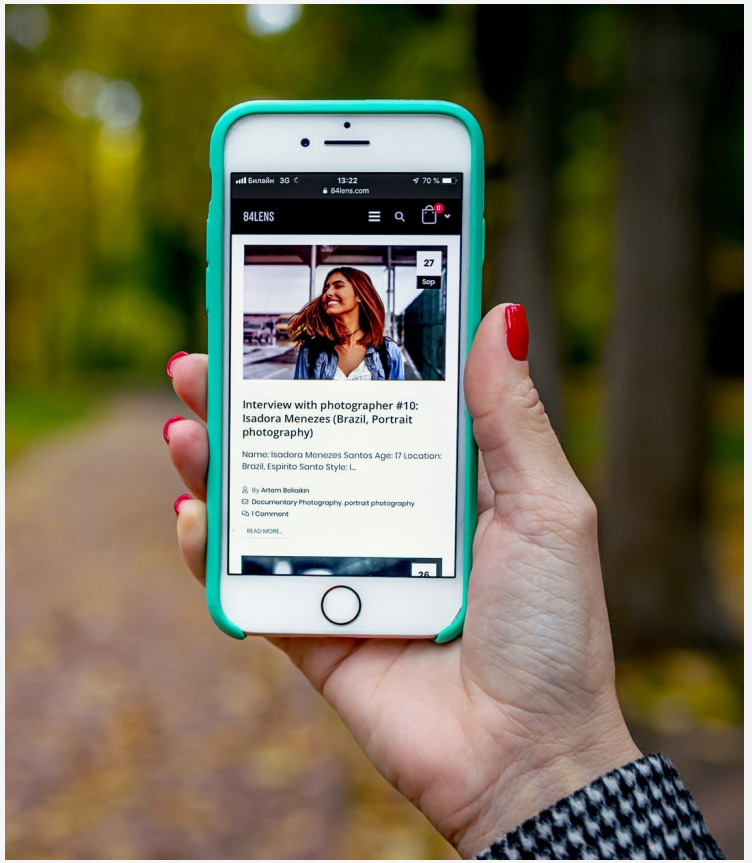
Assuming a normal distribution, what is the probability a given random Z-Phone will last between 20 and 30 months?



Exercise

A market researcher estimates that the Z-Phone smart phone has a mean consumer life of 42 months with a standard deviation of 8 months.

Assuming a normal distribution, what is the probability a given random Z-Phone will last between 20 and 30 months?




```
from scipy.stats import norm

mean = 42
std_dev = 8

x = norm.cdf(30, mean, std_dev) - norm.cdf(20, mean, std_dev)

print(x)
```

There is a 6.3827% probability the Z-Phone will last between 20 and 30 months.