Applied Statistics for Data Analytics

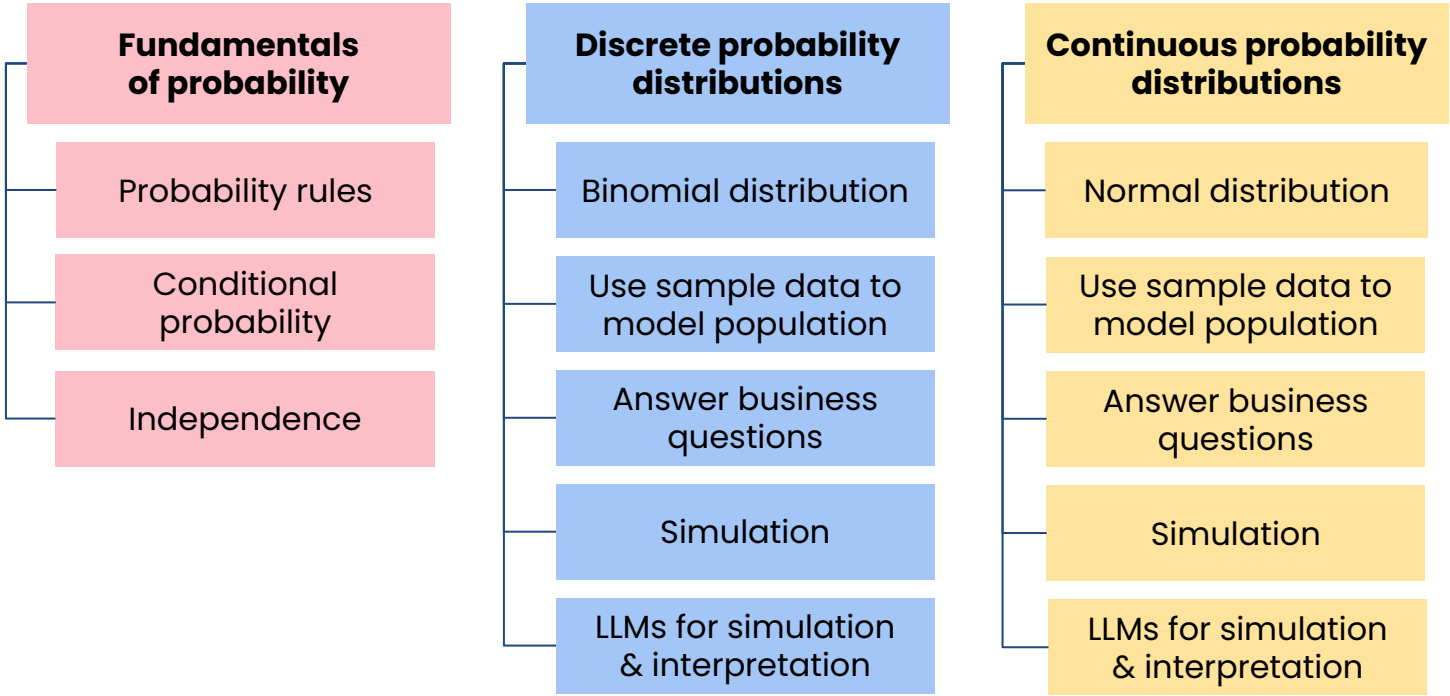
Module 2: Probability and simulation

Probability and simulation

Module 2 introduction

DeepLearning.AI

Module 2 outline



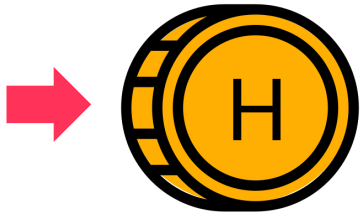
Probability and simulation

Randomness and uncertainty

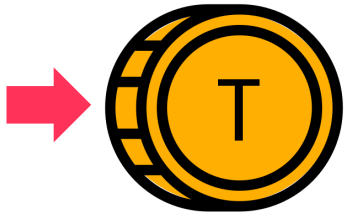
DeepLearning.AI

Randomness

A coin flip is **random**:



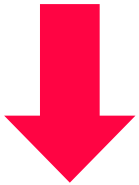
50% chance



50% chance

🧠 Each outcome has a known probability

⏳ Past coin flips don't influence future ones



1 2 3 4

Scenario

Whether friend will arrive on time for coffee:

On time

?% chance

Late

?% chance

Factors at play:

- 🚗 Traffic
- ⏰ Alarm clock
- 🤒 How they're feeling

Sources of randomness

1. Hidden features that still influence the outcome
2. Complex interactions between features
3. Measurement limitations introduced by imperfect tools
4. True unpredictability (subatomic level!)

The role of probability and statistics

Create models that **approximate** real-world randomness

⛔ Not trying to predict each individual event

✅ Understand overall distribution of events

✅ Make informed decisions in the face of uncertainty

What you will learn

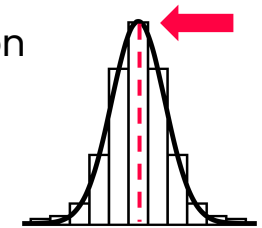
In this module:

1. Describing **probability distributions**

   Theoretical distributions that represent likelihood of all possible outcomes

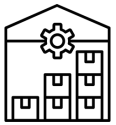
   Example: Normal distribution

   📝 Test scores 📏 Heights



2. **Simulation** of sampling from a distribution

Customer demand



Optimize inventory

In the next two modules:

3. **Confidence intervals**

   A range of values estimated to contain a true feature of a population

4. **Hypothesis testing**

   Technique to determine if result is likely to represent a true effect

Probability and simulation

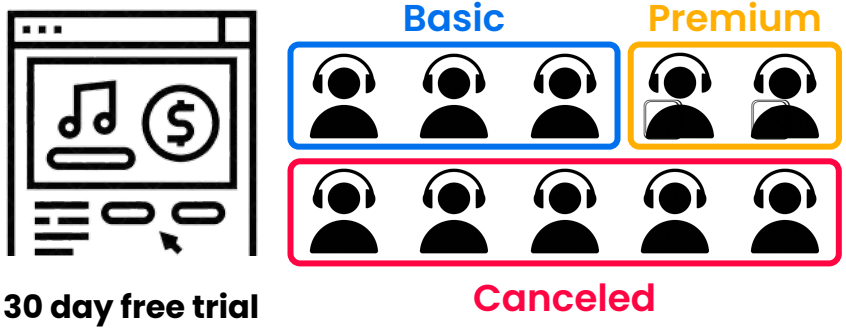
Probability and the addition rule

DeepLearning.AI

Scenario

Outcome - choosing customer 1

Outcome - choosing customer 2



30 day free trial

At the end of the free trial, customers can either:

- ☐ Continue with a basic subscription
- ☐ Upgrade to a premium subscription
- ☐ Cancel their subscription

Experiment - observing which outcome out of a set of outcomes actually occurred

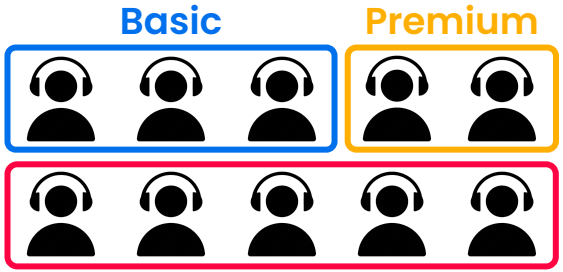
Choose a customer at random to interview

What is the likelihood that they chose the premium subscription?

Event - the specific set of outcomes you're interested in measuring

Scenario

30 day free trial



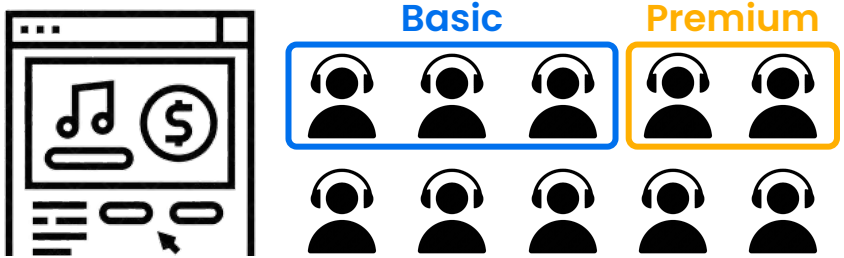
The probability of a user having the premium subscription is **2 out of 10** or **20%**.

$$P(\text{premium}) = \frac{\text{Number of favorable outcomes}}{\text{Number of possible outcomes}} = \frac{\text{Having premium}}{\underbrace{3 + 2 + 5}_{\text{Sample space}}} = \frac{2 \text{ outcomes}}{10 \text{ outcomes}} = 0.2$$

P stands for probability

Scenario

❌ **Not both**



30 day free trial

Probability of any given set can never be:

- ⛔ Less than 0%
- ⛔ Greater than 100%

$$P(\text{any subscription}) = \frac{\overbrace{\text{Total basic or premium subscribers}}^{\textbf{Mutually exclusive}}}{\text{Number of possible outcomes}} = \frac{\overbrace{3 + 2}^{\textbf{Addition rule}}}{10 \text{ outcomes}} = \textbf{5 out of 10 (or 50%)}$$

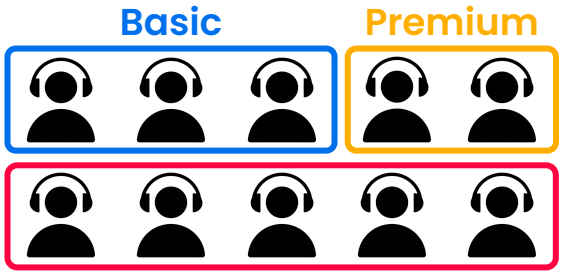
Probability and simulation

The multiplication and complement rules

DeepLearning.AI

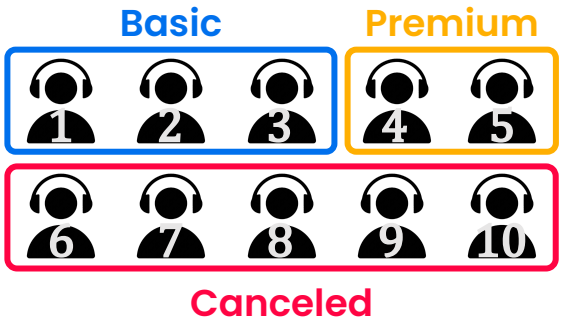
Scenario

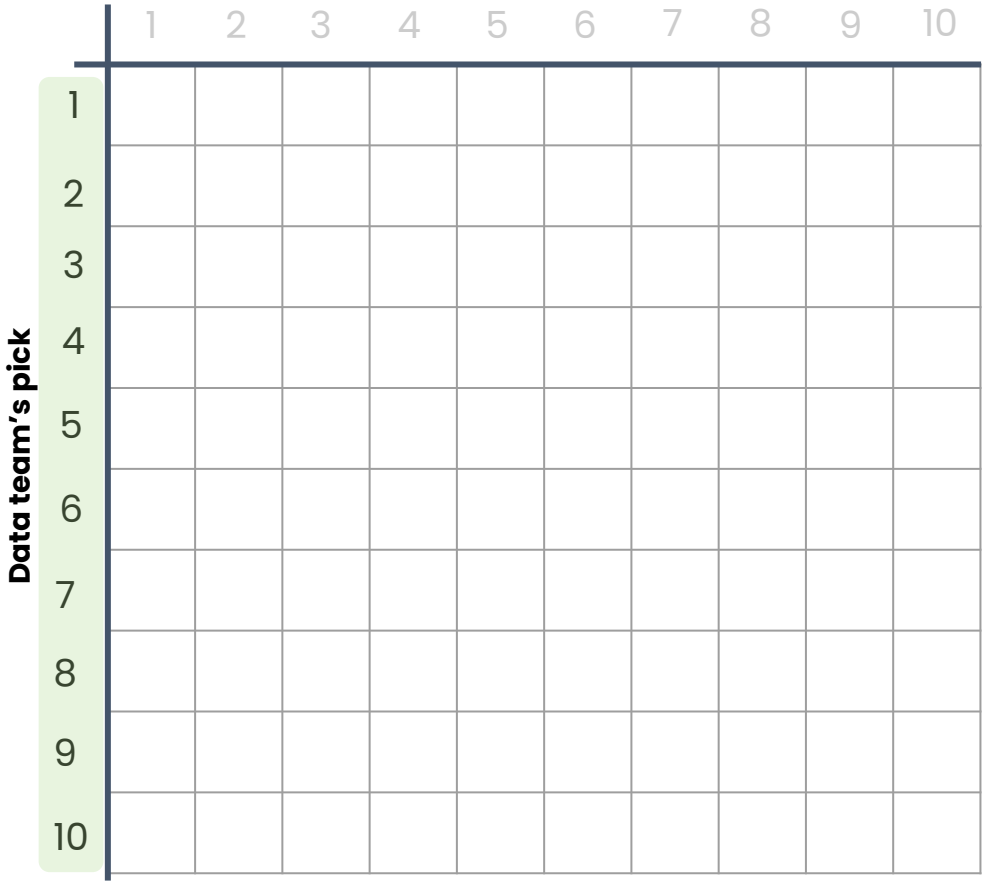
30 day free trial



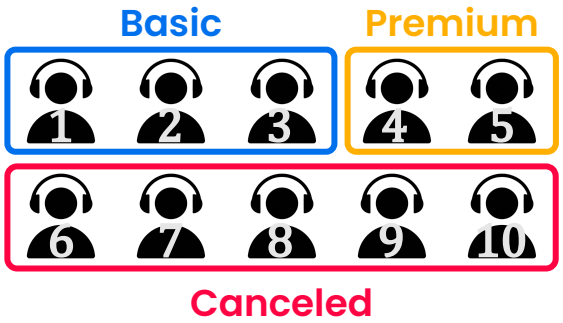
Probability that you and marketing team **both pick people with premium**?

Scenario


Basic
Premium
1
2
3
4
5
6
7
8
9
10
Canceled


1 2 3 4 5 6 7 8 9 10
Data team's pick
1
2
3
4
5
6
7
8
9
10

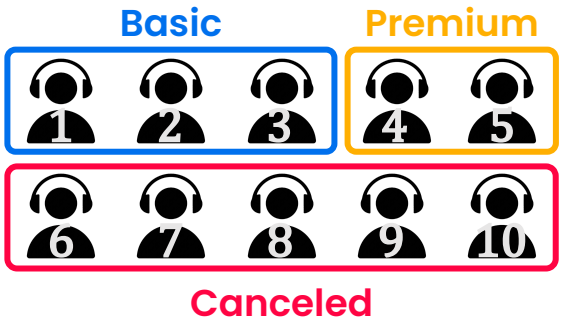
Scenario



Marketing team's pick

Data team's pick	1	2	3	4	5	6	7	8	9	10
1	B	B	B	P	P	C	C	C	C	C
2										
3										
4										
5										
6										
7										
8										
9										
10										

Scenario



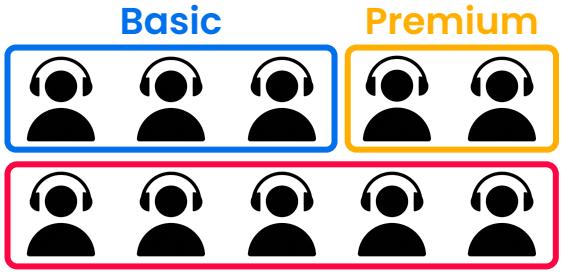
100
possible outcomes

Data team's pick \ Marketing team's pick	1	2	3	4	5	6	7	8	9	10
1	BB	BB	BB	PB	PB	CB	CB	CB	CB	CB
2	BB	BB	BB	PB	PB	CB	CB	CB	CB	CB
3	BB	BB	BB	PB	PB	CB	CB	CB	CB	CB
4	BP	BP	BP	PP	PP	CP	CP	CP	CP	CP
5	BP	BP	BP	PP	PP	CP	CP	CP	CP	CP
6	BC	BC	BC	PC	PC	CC	CC	CC	CC	CC
7	BC	BC	BC	PC	PC	CC	CC	CC	CC	CC
8	BC	BC	BC	PC	PC	CC	CC	CC	CC	CC
9	BC	BC	BC	PC	PC	CC	CC	CC	CC	CC
10	BC	BC	BC	PC	PC	CC	CC	CC	CC	CC

Sean Barnes

Scenario

30 day free trial



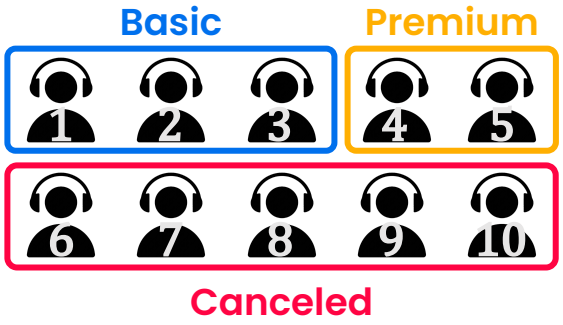
Probability that you and marketing team **both pick people with premium**?

Answer: 1 of every 25 experiments

Your pick: premium, Marketing's pick: premium

$$P(\text{premium}, \text{premium}) = \underbrace{\frac{2}{10} \times \frac{2}{10}}_{\text{Independent}} = \underbrace{\frac{4}{100} = \frac{2}{50} = \frac{1}{25}}_{\text{Multiplication rule: } P(A \text{ and } B) = P(A) \times P(B)}$$

Scenario



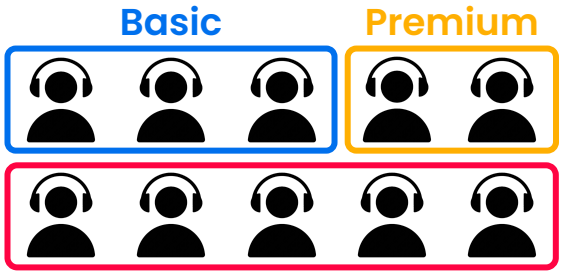
- Choosing premium was a relatively rare outcome
- So many outcomes involve one person without premium

Marketing team's sample space

Your sample space	1	2	3	4	5	6	7	8	9	10
1	BB	BB	BB	PB	PB	CB	CB	CB	CB	CB
2	BB	BB	BB	PB	PB	CB	CB	CB	CB	CB
3	BB	BB	BB	PB	PB	CB	CB	CB	CB	CB
4	BP	BP	BP	PP	PP	CP	CP	CP	CP	CP
5	BP	BP	BP	PP	PP	CP	CP	CP	CP	CP
6	BC	BC	BC	PC	PC	CC	CC	CC	CC	CC
7	BC	BC	BC	PC	PC	CC	CC	CC	CC	CC
8	BC	BC	BC	PC	PC	CC	CC	CC	CC	CC
9	BC	BC	BC	PC	PC	CC	CC	CC	CC	CC
10	BC	BC	BC	PC	PC	CC	CC	CC	CC	CC

Scenario

30 day free trial



Probability you and marketing team both pick people with premium, **without picking the same person?**

- ⛔ Can't use multiplication rule
- ⛔ The two events aren't independent

Scenario

1 Define the sample space

Count up outcomes that fit conditions

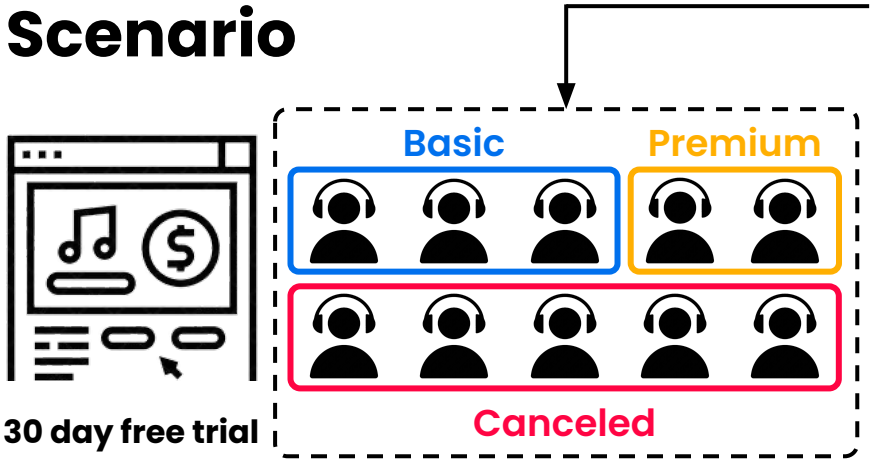
2 Find the favorable outcomes

$$P(\text{two } \textbf{different} \text{ users with premium}) = \frac{\textbf{2 outcomes}}{\textbf{90 outcomes}} = \textbf{2.2\%}$$

Marketing team's sample space

Your sample space	1	2	3	4	5	6	7	8	9	10
1	~~BB~~	BB	BB	PB	PB	CB	CB	CB	CB	CB
2	BB	~~BB~~	BB	PB	PB	CB	CB	CB	CB	CB
3	BB	BB	~~BB~~	PB	PB	CB	CB	CB	CB	CB
4	BP	BP	BP	~~PP~~	PP	CP	CP	CP	CP	CP
5	BP	BP	BP	PP	~~PP~~	CP	CP	CP	CP	CP
6	BC	BC	BC	PC	PC	~~CC~~	CC	CC	CC	CC
7	BC	BC	BC	PC	PC	CC	~~CC~~	CC	CC	CC
8	BC	BC	BC	PC	PC	CC	CC	~~CC~~	CC	CC
9	BC	BC	BC	PC	PC	CC	CC	CC	~~CC~~	CC
10	BC	BC	BC	PC	PC	CC	CC	CC	CC	~~CC~~

Scenario



Probability that person **does not have basic**?

Useful when:

- Don't know probabilities of every outcome
- Might take a lot of work to calculate them

Complement rule

$$P(\text{not basic}) = 1 - P(\text{basic})$$

All possible occurrences

$$= \frac{10}{10} - \frac{3}{10} = \frac{2+5}{10} = \frac{7}{10}$$

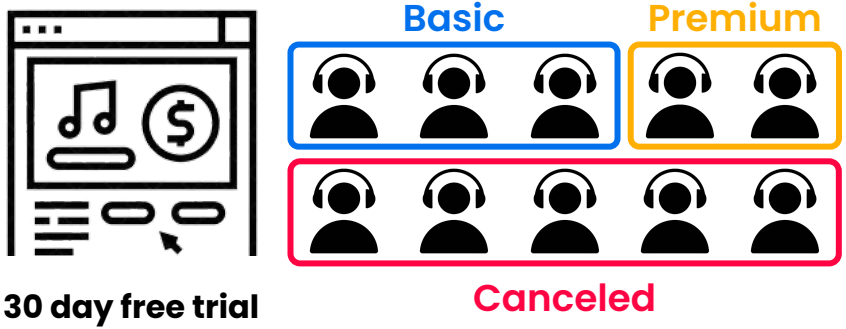
$$= 1 - 0.3 = 0.7 = 70\%$$

Probability and simulation

Conditional probability

DeepLearning.AI

Scenario



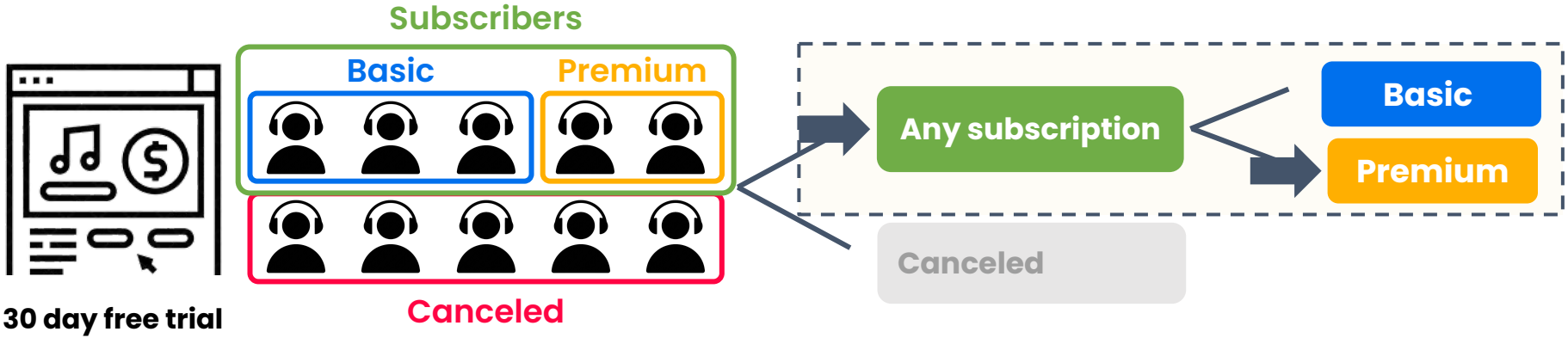
Probability that person chose **premium if they got a subscription at all**?

given

$$P\ (\text{premium}\ |\ \text{any subscription})$$

General form: $P\ (A\ |\ B)$

Scenario

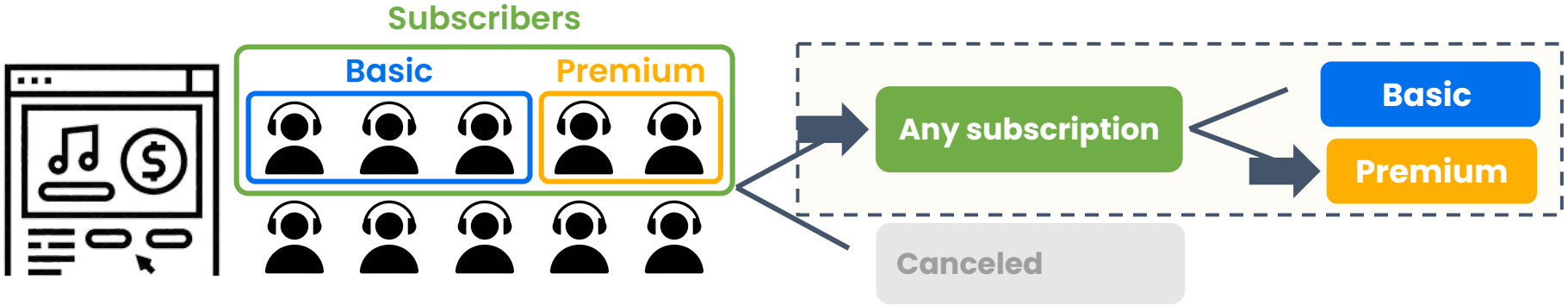


given

$$P\ (\ \text{premium}\ \mid\ \text{any subscription}\ )$$

General form: $P\ (\ A \mid B\ )$

Scenario



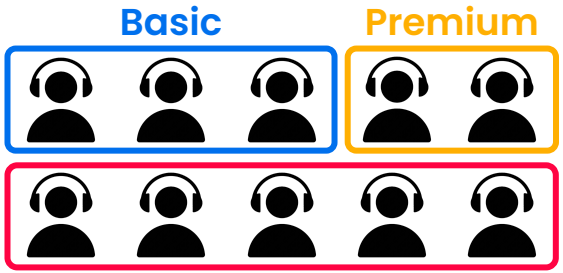
$$P(\text{premium} \mid \text{any subscription}) = \frac{2}{5} = 2 \text{ out of } 5 = \mathbf{40\%}$$

$$P(A \mid B) = \frac{P(A \textbf{ and } B)}{P(B)} = \frac{P(\text{premium } \textbf{and} \text{ any subscription})}{P(\text{any subscription})} = \frac{P(\text{premium})}{0.3 + 0.2} = \frac{0.2}{0.5} = \frac{2}{5}$$

40%

Scenario

30 day free trial



❌ Can't have premium without having a subscription

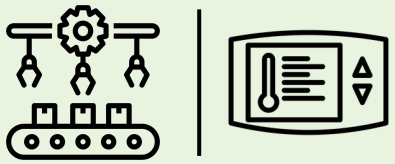
$$P(A \mid B) \neq P(B \mid A)$$

$$P(\text{premium} \mid \text{any subscription}) \neq P(\text{any subscription} \mid \text{premium})$$

100%

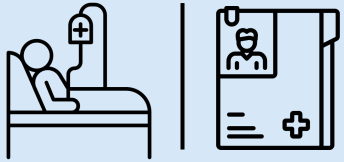
Use cases

Manufacturing



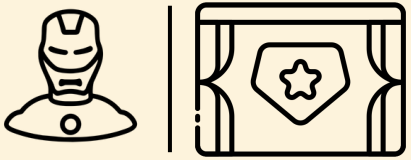
P(failure | temperature)

Healthcare



P(condition | symptoms)

Streaming



P(❤️ Ironman | ❤️ heroes)

Segmentation

Calculating statistics **given** a certain outcome is within the segment

Probability and simulation

Independence

DeepLearning.AI

What is independence?

- The occurrence of one event does not affect the probability of the other

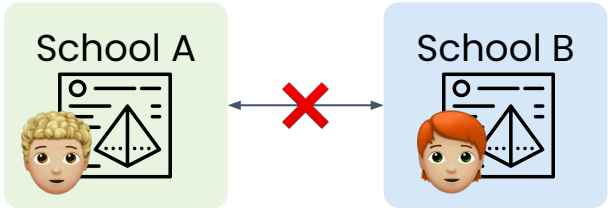
Examples:

Coin flips

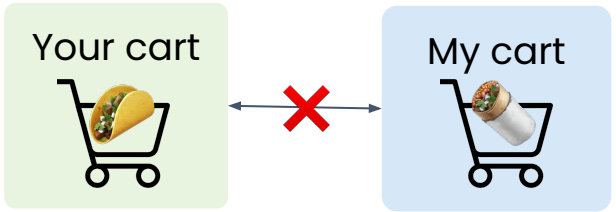
Dice rolls

- Many statistical methods assume two events are independent
- Events that seem independent at first glance often are **not**

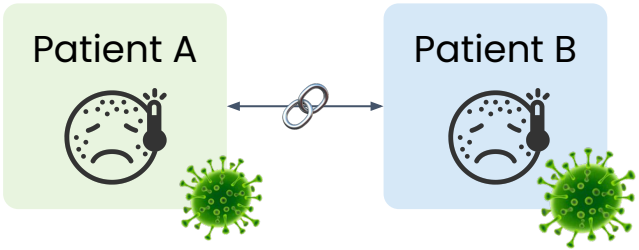
Independent



Independent



NOT independent



Scenario

Two events are independent if:

$$P(A \mid B) = P(A)$$

Event A

Always 1 in 6

$P(6) = 1/6$

Event B

$P(6 \mid 1) = 1/6$

Ways to determine if events are independent

1. Study nature of events to see if outcomes influence each other
2. Collect data on the events and see if probabilities satisfy rule

Non-independence

Affects many calculations, like the multiplication rule

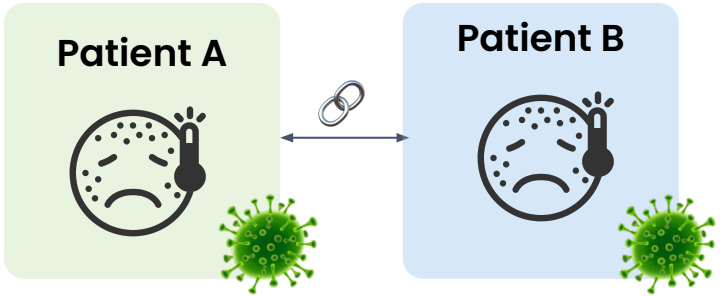
✅ **If two events are independent:**

Can multiply probabilities together

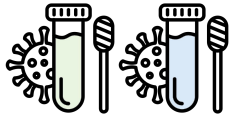
$$P(A \text{ and } B) = P(A) \times P(B)$$

❌ **If two events are not independent:**

Multiplication rule cannot be applied



Scenario



Assume independence

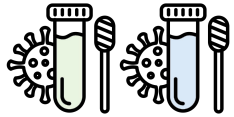
$$P(\text{Patient A}) \times P(\text{Patient B})$$

1% 1%

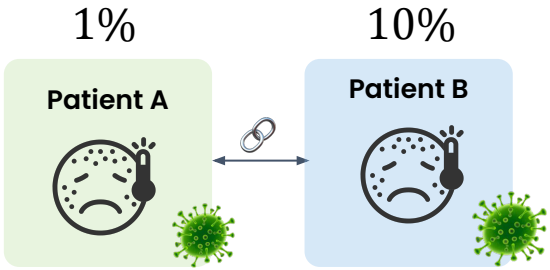
$$= 0.01 \times 0.01 = 0.0001$$

1 in **10,000**

Scenario



Measles is communicable:



Multiplication rule for dependent events

P (Both have measles) =

Conditional probability

$$P(\text{Patient A}) \times P(\text{Patient B} \mid \text{Patient A})$$

$$= 0.01 \times 0.10 = 0.001$$

1 in **1,000**

- Importance of decision
- Stronger evidence needed

Probability and simulation

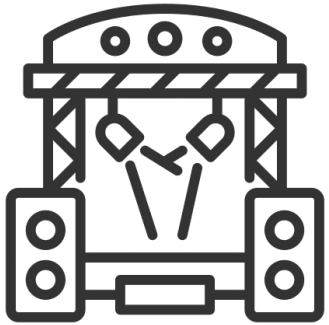
Random variables

DeepLearning.AI

What are random variables?

Represent all the possible outcomes of a random phenomenon

💬 **Problem**: Model number of rainy days in a week



8 possible values

$$X = \{ 0, 1, 2, 3, 4, 5, 6, 7 \}$$

Each value represents a possible outcome of X

What are random variables?

Different from traditional variables in math

Traditional variable

$$x + 5 = 30$$

x has one value: 25

- Only has one value at a time

Random variable

$$X = \{0, 1, 2, 3, 4, 5, 6, 7\}$$

- Can represent multiple values

Probability notation

Random variables make your probability notation much easier.

"The probability of 3 days of rain" → $P(\text{3 days of rain})$ → $P(X=3)$

The probability that X takes on the value 3

Random variables

Must be numbers

$$\mathbf{X} = \{ 0, 1, 2, 3, 4, 5, 6, 7 \}$$

A single day's rain ☔

$$\mathbf{Y} = \{ 0, 1 \}$$

- To represent a non-numerical outcome, create mapping between:
  - Numbers in random variable
  - What they mean in the real world

Number	Meaning
0	Not rainy
1	Rainy

Discrete random variables

- Both X and Y have a countable number of values
- X and Y are **discrete** random variables
- Same as discrete feature in a data set:
  - Can only have a **countable** set of values

8 values

$$\mathbf{X} = \{ 0, 1, 2, 3, 4, 5, 6, 7 \}$$

2 values

$$\mathbf{Y} = \{ 0, 1 \}$$

Continuous random variables

Discrete random variables

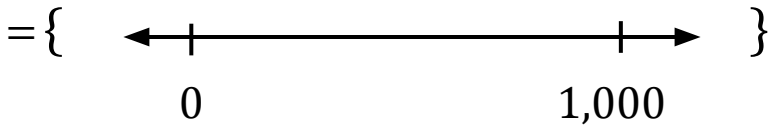
- Distinct values
- $X = \{ 0, 1 \}$
- $X = \{ 0, 1, 2, 3, 4, 5, 6, 7 \}$

Continuous random variables

- Represent ranges of values

Example: Total rainfall in a given week

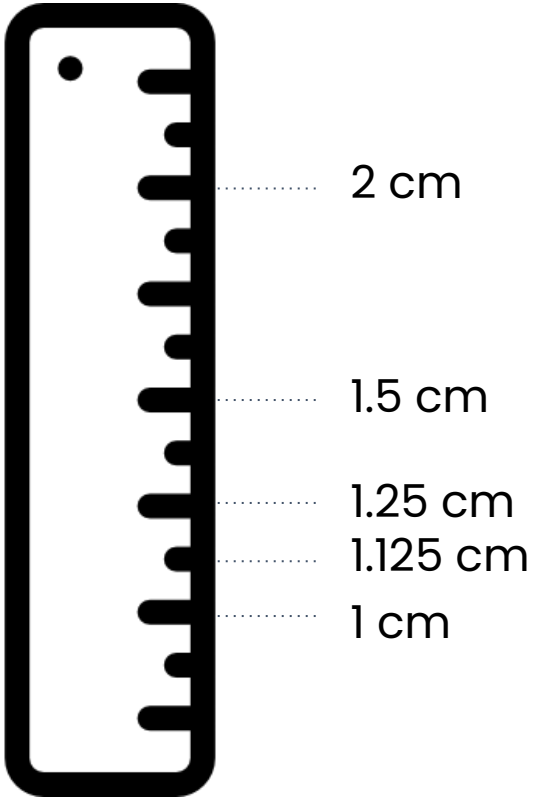
W - total centimeters of rain in a given week



Rainfall isn't a set of distinct values, but a range of measurements

Continuous random variables

- No matter how close two values are, there's always another value between
- This process can continue **indefinitely**
- Continuous random variable can take on **any real number** within its range



Test to determine random variable type

1. Try to list out the possible values
2. If you can:
   - ✅ List out those values
   - ✅ There aren't numbers in between
3. Your random variable is **discrete** ✅

S = the number of students in a given elementary school

$\{0, \quad 300, \quad 1700, \quad 1701 \ldots\}$

~~1700.5~~

Scenario

Random variable

$$\mathbf{X} = \{ 0, 1, 2, 3, 4, 5, 6, 7 \}$$

✅ Only contains values that represent the outcomes

⛔ Not how likely they are

Not contained within the random variable

💬 What are the probabilities of these values?

💬 How likely is it that it rains 0 or 6 days?

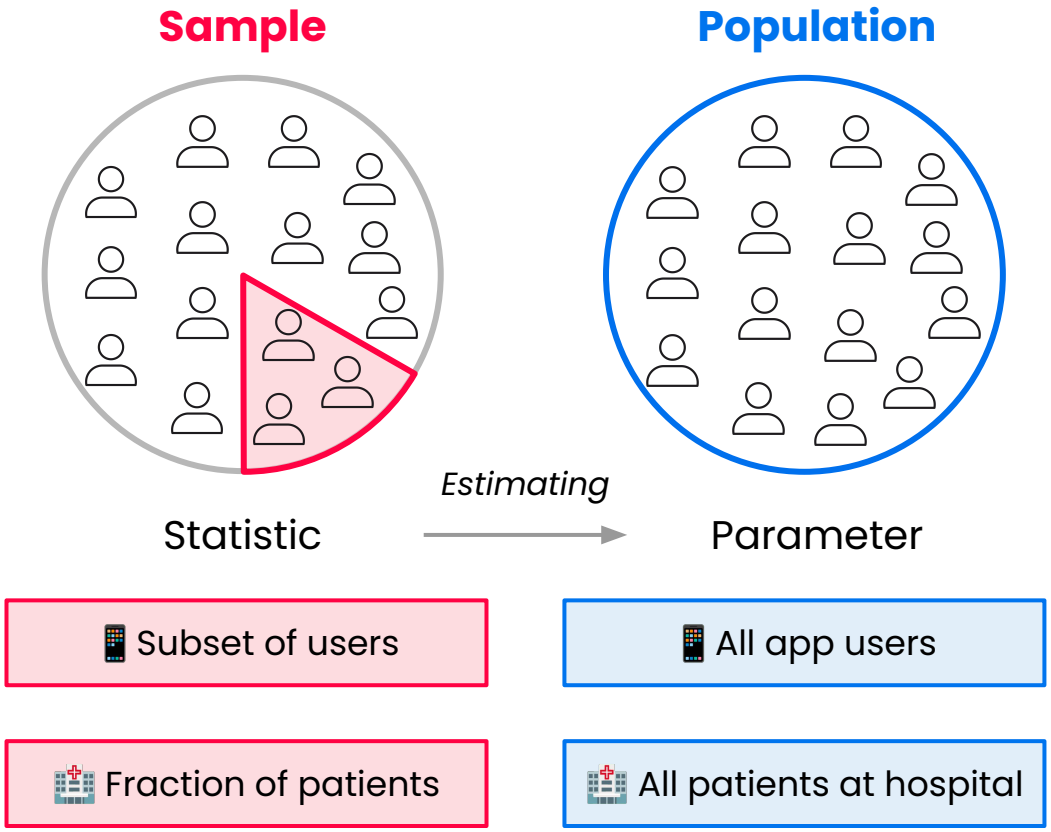
Probability and simulation

Estimation

DeepLearning.AI

Estimates

- An approximation for the truth
- Will probably not be exactly the same as the truth



Working with a population

✅ Have data from all 50 states

✅ Population

✅ Measured everyone's height

✅ Population

🔄 Constantly changing

- You don't need to measure everyone
- With large enough sample, you can get close to the true value

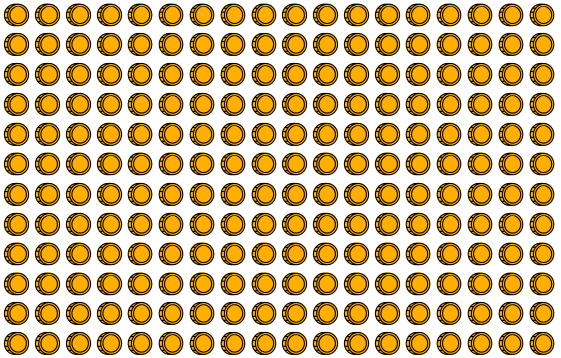
Working with a population

Coin flip

50% probability

⛔ Need to flip infinite number of times to prove

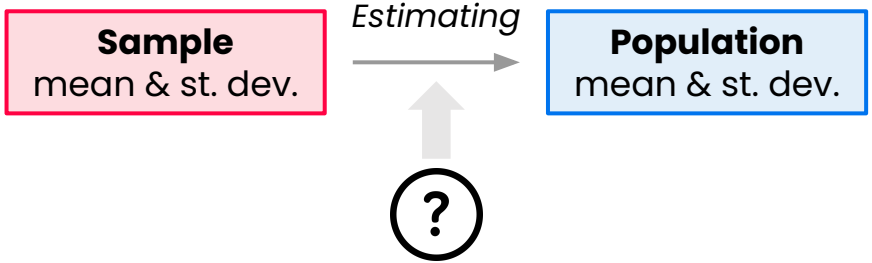
Large sample



✅ Close enough estimate to be useful

How this affects your approach

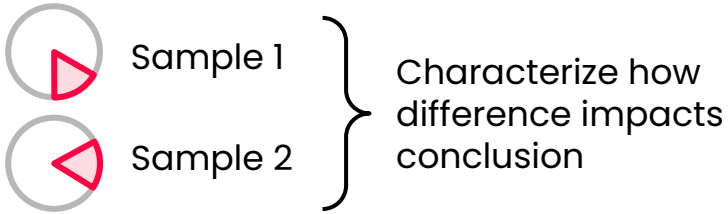
1 Calculating statistics & analytical approach



Formulas are sometimes different:

- Mean → same formula
- Standard deviation → two formulas

2 Sample is just one possible version



3 Sampling bias

- If sample isn't representative, estimates will be off

Probability and simulation

From sample distributions
to population distribution

DeepLearning.AI

Visualizing distributions

🌧️ Rainy days per week

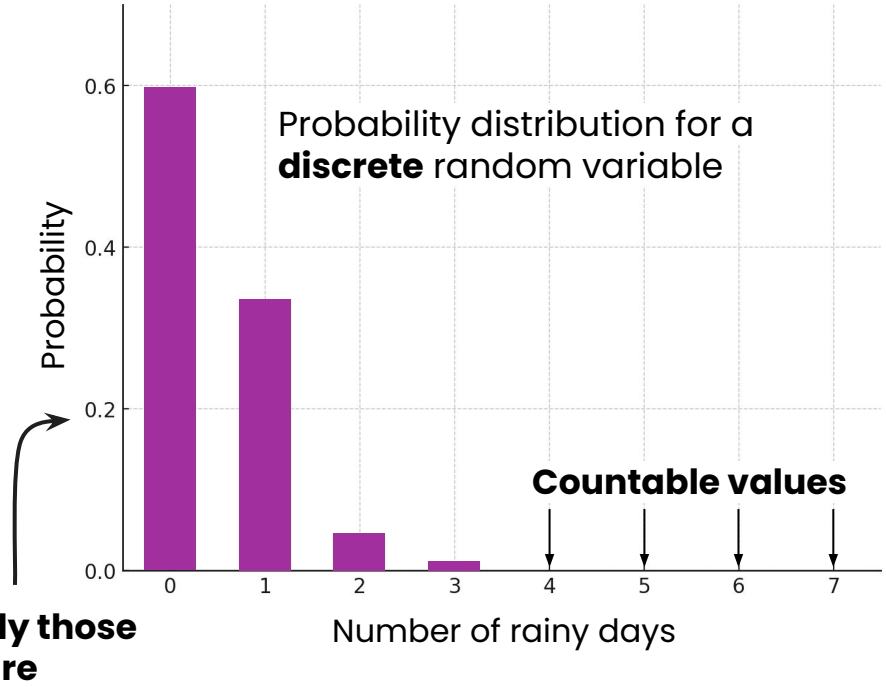
$$X = \{0, 1, 2, 3, 4, 5, 6, 7\}$$

Set of probabilities that correspond with each outcome: **probability distribution**

Probability mass function (PMF):

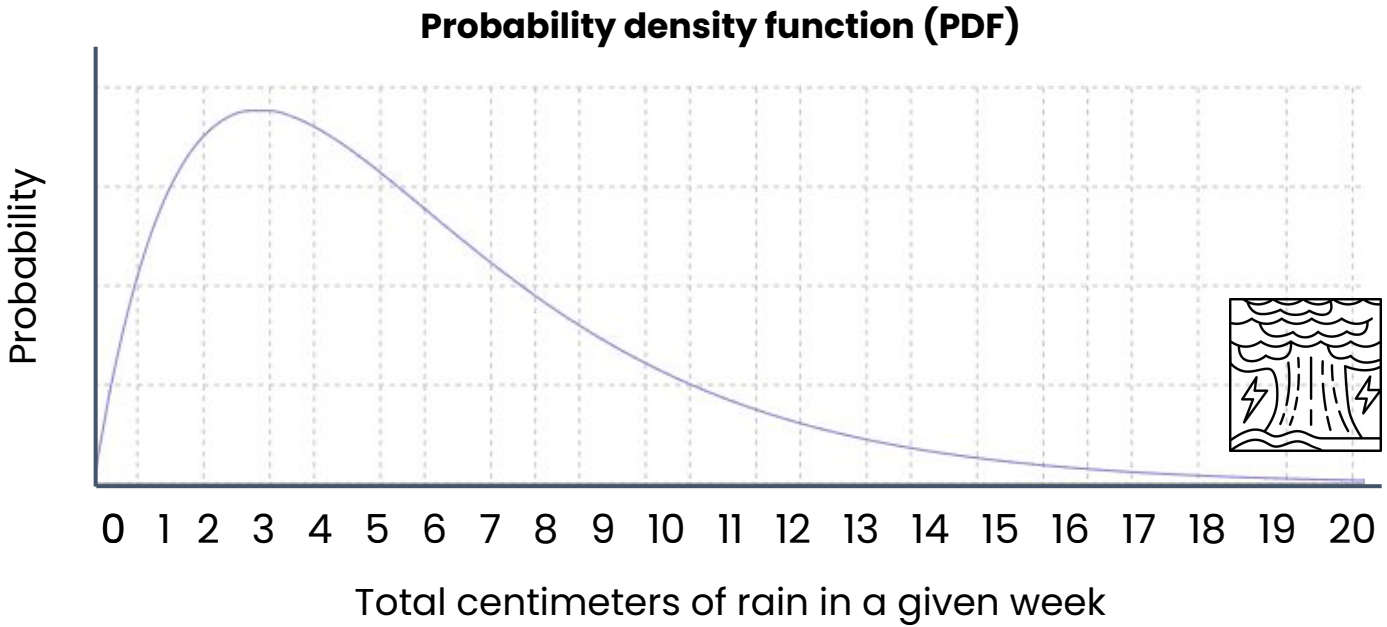
Defines probabilities of each event for **discrete** random variable

$$\begin{cases} P(X=0) = 0.6, \\ P(X=1) = 0.35, \\ P(X=2) = 0.04, \\ P(X=3) = 0.01, \\ P(X \geq 4) = 0 \end{cases}$$

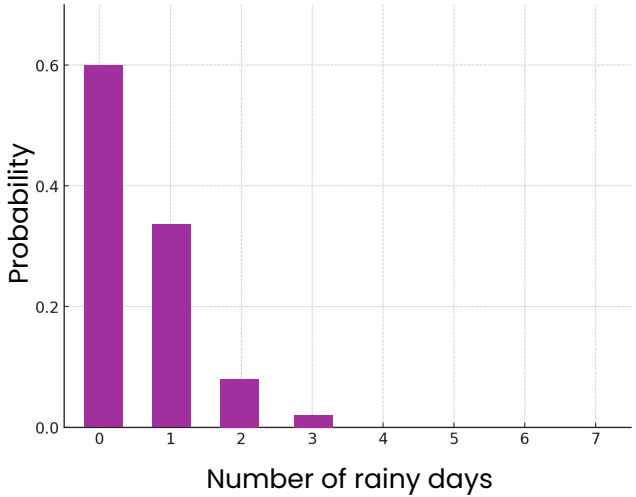


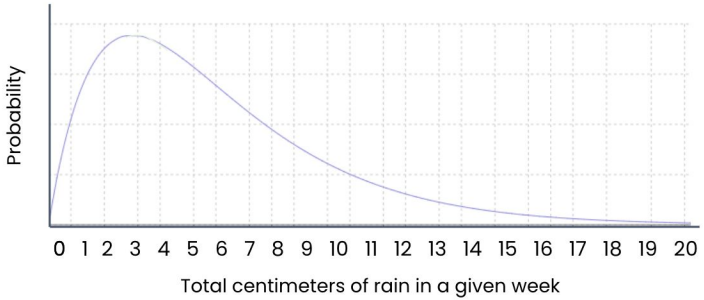
Visualizing distributions

W - total centimeters of rain in a given week



- Calculate the probability for a specific range of values
- Infinite number of possible values





These are **distributions**, but these are **not sample** distributions.

Population vs sample distribution

Sample distribution

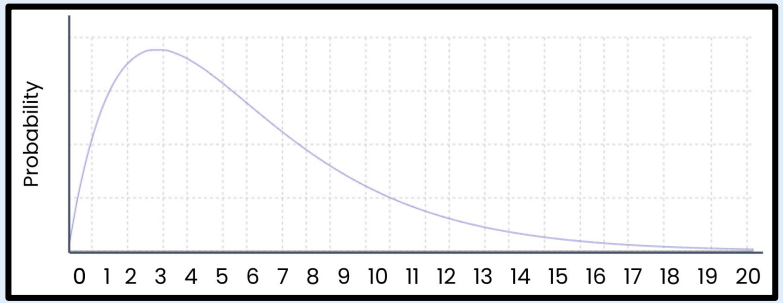
Go out and measure 30 days of rainfall

Record it in your spreadsheet

Graph it

"I recorded rainfall for 30 days, and on 1 of those days, the rainfall was greater than 3 centimeters."

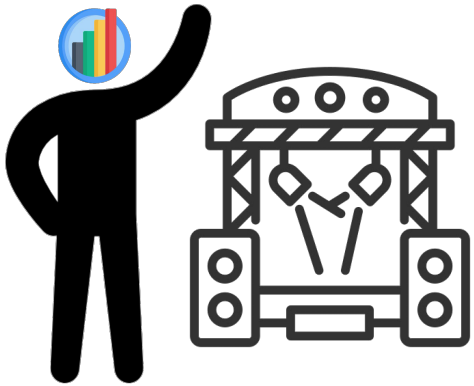
Population distribution



Use characteristics of sample to estimate population distribution

"**In general,** there is a 4% chance that there will be 3 or more centimeters of rain on any given day."

Scenario

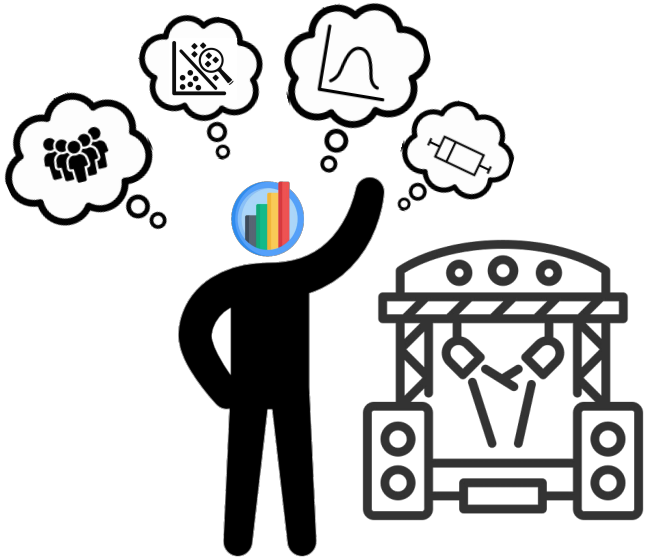


You
Data Analyst

💬 **Problem**: Characterize how much each customer spends on tickets in a year

- Sample: 100 customers
- Tally their spending for the year
- Characterize **sample distribution** with descriptive statistics:
  - **Mean**: $123
  - **Standard deviation**: $15.40
  - **Median**: $100

Scenario



You
Data Analyst

"The average customer spends $x in a year."

📍 The **average customer in this sample** spent $123 in a year.

🎯 Draw a conclusion about the **population distribution** of all customers based on this **sample distribution** of customers

- Sample: 100 customers ---- 🪟 Window of truth

Probability and simulation

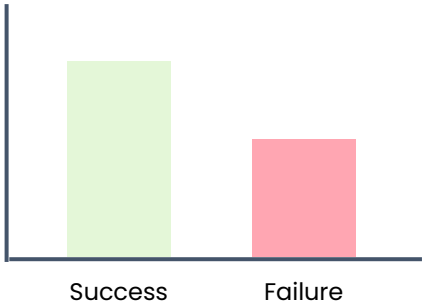
The Bernoulli distribution

DeepLearning.AI

What is the Bernoulli distribution?

- Models a random variable that has only **two** possible outcomes:

- Visualize probability of the two outcomes



$$X = \{ 1, 0 \}$$

Probability: p, $1 - p$

- Population mean
- Population variance
- Population standard deviation

Scenario

You
Data Analyst

🧬 Genetic predispositions

🐶 Breed

🔍 **Your role**: Understand distribution of invalid canine DNA samples

- Either a sample is: ✅ Valid ❌ Invalid
- Rate at which samples are valid: **70%**

💬 **Problem**: Can you model the distribution of each sample's validity?

Scenario

You
Data Analyst

- The Bernoulli distribution is appropriate because:

✅ Two possible outcomes:

$$\mathbf{X} = \{ \text{✅ Valid} , \text{❌ Invalid} \}$$

✅ Each sample has the **same** 70% chance of being valid

Scenario

You
Data Analyst

- Model Bernoulli distribution using one parameter:

  $\mathbf{p}$ - probability of success

- In this case:

  $\mathbf{p} = 0.7$

Tilde: "Is distributed as"

$$X \sim Bernoulli(0.7)$$

Random variable X is distributed as a Bernoulli with parameter 0.7

$$P(\text{✅ Valid}) = 0.7$$

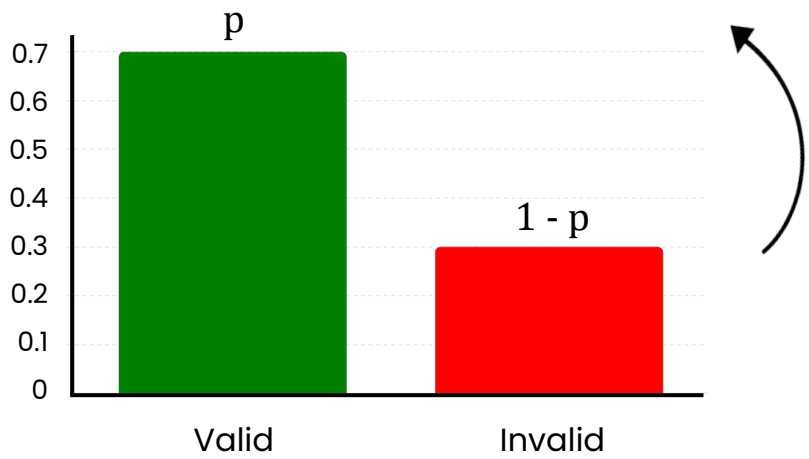
$$P(\text{❌ Invalid}) = 0.3$$

Complement rule:

$$1 - 0.7 = 0.3$$

Probability mass function (PMF)

$$P(X=x) = \begin{cases} 0.7, & \text{if } \mathbf{x} = \text{✅ Valid} \\ 0.3, & \text{if } \mathbf{x} = \text{❌ Invalid} \end{cases}$$



- Every discrete probability distribution has a PMF
- Probabilities sum up to 1
- When graphing the distribution, what you're graphing is the PMF.

Calculating Bernoulli population parameters

Parameter	Symbol	Calculation	Value
Mean	μ	p	0.7
Variance	σ^2	$p(1 - p)$	$0.7(0.3) = 0.21$
Standard deviation	σ	$\sqrt{p\,(1 - p)}$	$\sqrt{0.21} = 0.458$

Gives you a sense of how much outcomes can vary around the mean

Probability and simulation

The Binomial distribution

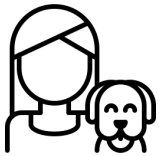
DeepLearning.AI

What is the binomial distribution?

- Models probability of a number of successes in a fixed number of trials
- Can only model a distribution with two outcomes
- Each trial must have the same probability of success

✅ Success ❌ Failure

Scenario

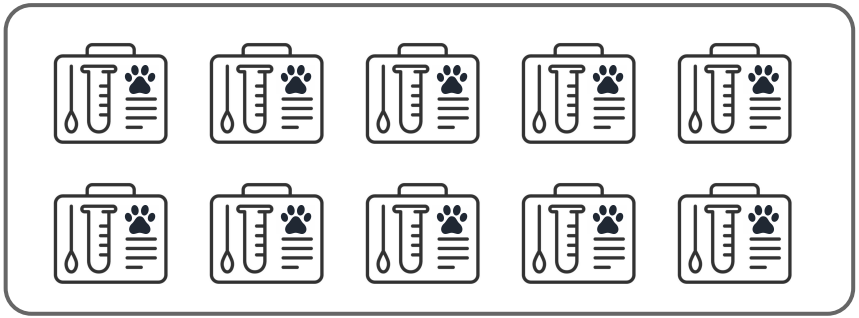


Each owner collects own sample

Shipments arrive in boxes of 10

70% Valid

💬 **Problem**: How can we understand how many valid samples in each box?



Scenario

You
Data Analyst

Modeling with a probability distribution will allow the lab to:

📦 Estimate how many valid samples in each box

📉 Determine probability of getting a critically low number

🧪 Set realistic expectations for testing process

Scenario

You
Data Analyst

You can model using the binomial distribution:

- ⭐ Two possible outcomes: ✅ Valid ❌ Invalid
- ⭐ Each sample has same chance of being valid: **70%** Valid
- Fixed number of trials: 10 samples in each box
- Trials are independent: Each sample collected by different owners

Binomial distribution

The binomial distribution is defined by two parameters:

1. **n** - the number of trials

number of samples in a box

2. **p** - the probability of success

0.7 Valid

$$X \sim \text{Binomial}(10, 0.7)$$

Random variable X is distributed as a binomial with parameters $n=10$ and $p=0.7$

Binomial distribution

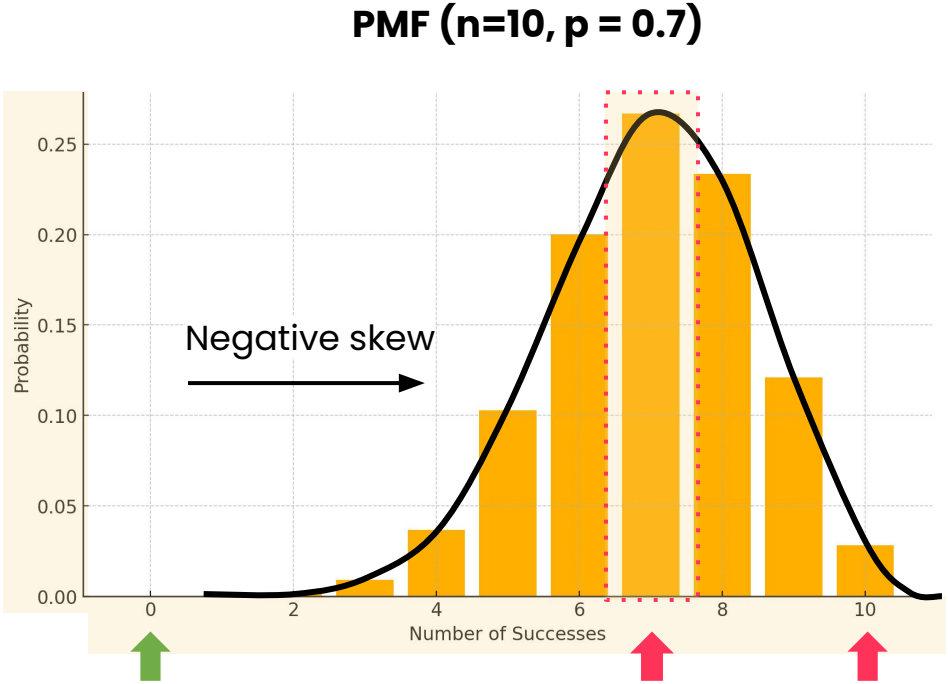
The binomial distribution is defined by two parameters:

1. **n** - the number of trials

number of samples in a box

2. **p** - the probability of success

0.7 Valid



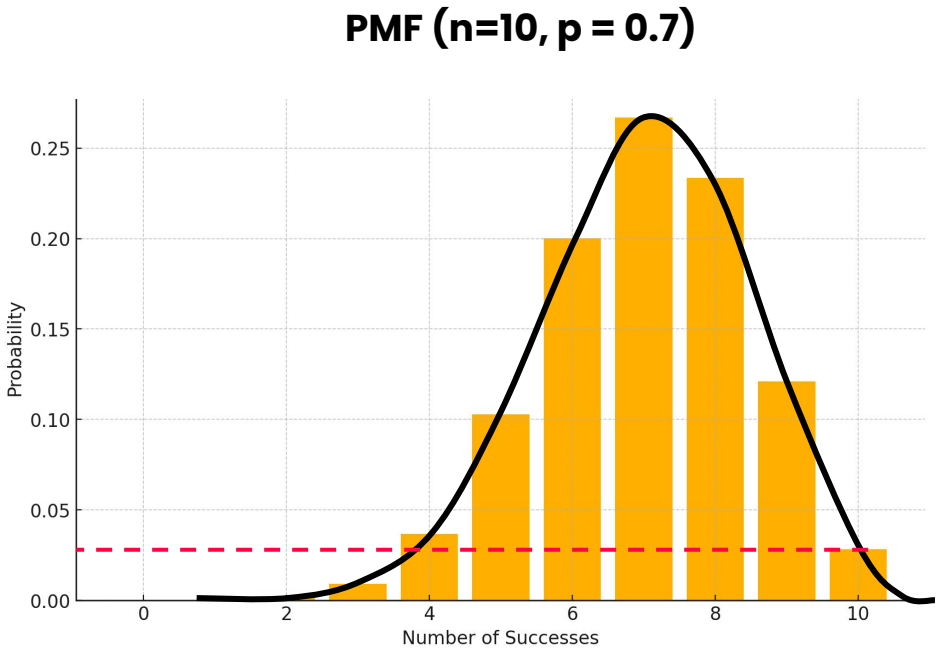
Increasingly unlikely to get a box of samples with greater or smaller numbers of valid samples

Binomial distribution

What are the chances of getting **10 valid samples** in a box?

$$0.7 \times 0.7 \times 0.7 \times 0.7 \times 0.7 \times 0.7 \times 0.7 \times 0.7 \times 0.7 \times 0.7$$

$$= \mathbf{0.7^{10}} \quad \approx \mathbf{0.025}$$

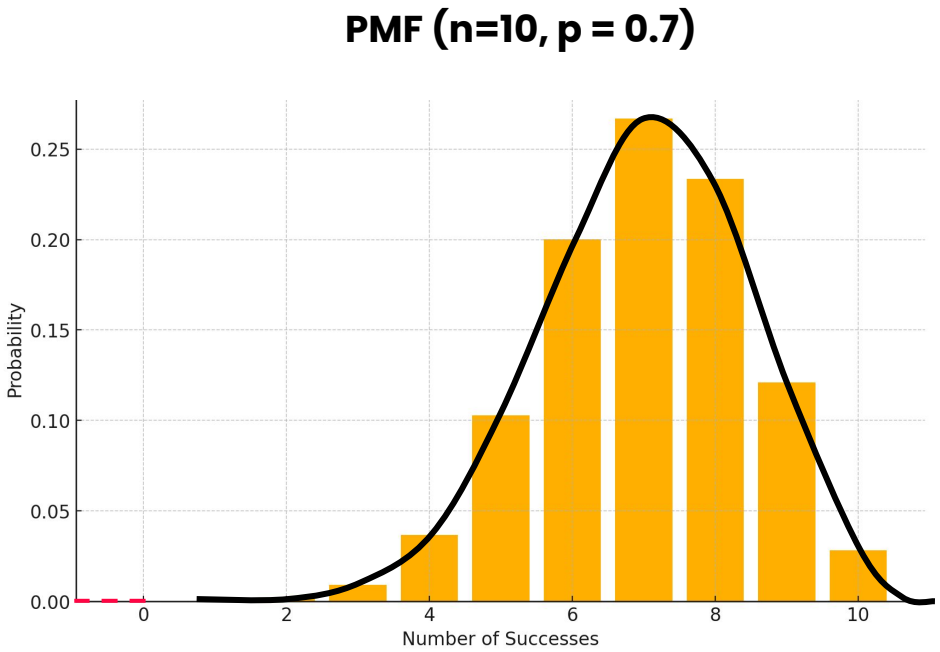


Binomial distribution

What are the chances of getting **10 valid samples** in a box?

$$0.3 \times 0.3 \times 0.3 \times 0.3 \times 0.3 \times 0.3 \times 0.3 \times 0.3 \times 0.3 \times 0.3$$

$$= \mathbf{0.3^{10}} \quad \approx \mathbf{0.0000059}$$

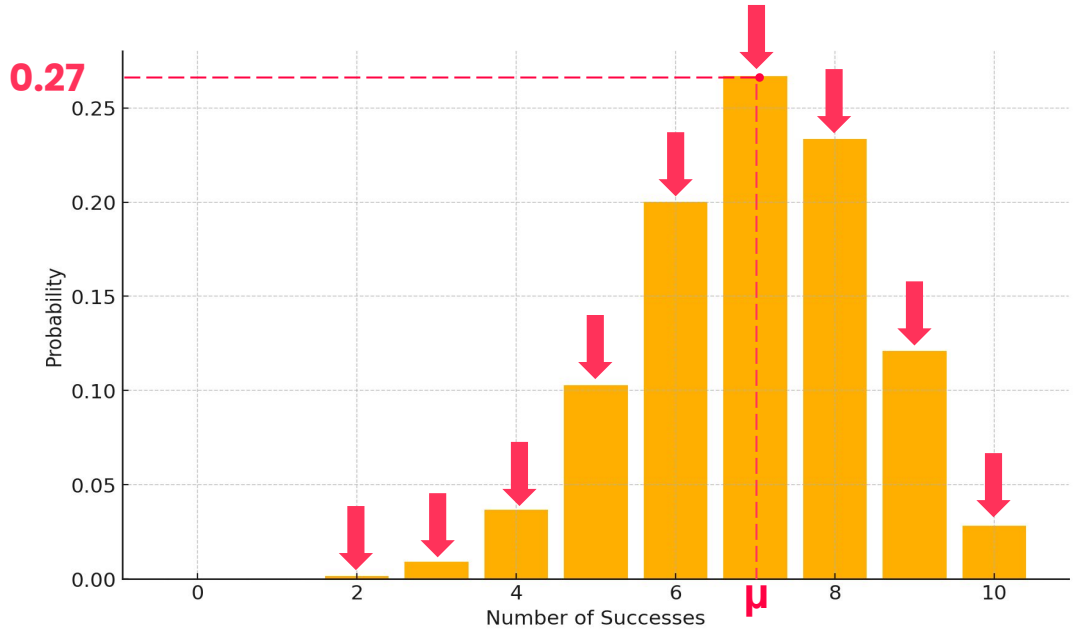


Calculating binomial population parameters

Statistic	Symbol	Calculation	Value
Mean	μ	$n * p$	$10 * 0.7 = 7$
Variance	σ^2	$n * p * (1 - p)$	$10 * 0.7 * (0.3) = 2.1$
Standard deviation	σ	$\sqrt{n * p * (1 - p)}$	$\sqrt{2.1} = 1.45$

"It's common to have between 5.5 and 8.5 samples"

Binomial distribution



- Probabilities can be derived from $\mathbf{n}$ and $\mathbf{p}$

PMF

$P(X=0) = <0.001$

$P(X=1) = <0.001$

$P(X=2) = 0.001$

$P(X=3) = 0.009$

$P(X=4) = 0.036$

$P(X=5) = 0.103$

$P(X=6) = 0.200$

$P(X=7) = 0.266$

$P(X=8) = 0.233$

$P(X=9) = 0.121$

$P(X=10) = 0.028$

$= \mathbf{1.0}$

Scenario

How often can the lab expect the mean value of **7** valid samples?

Answer: around 27%

How often can the lab expect **0** valid samples?

Answer: in < 1 in 1000 boxes

How often can the lab expect **10** valid samples?

Answer: 2.8% of all boxes

PMF

$P(X=0) = <0.001$

$P(X=1) = <0.001$

$P(X=2) = 0.001$

$P(X=3) = 0.009$

$P(X=4) = 0.036$

$P(X=5) = 0.103$

$P(X=6) = 0.200$

$P(X=7) = 0.266$

$P(X=8) = 0.233$

$P(X=9) = 0.121$

$P(X=10) = 0.028$

Probability and simulation

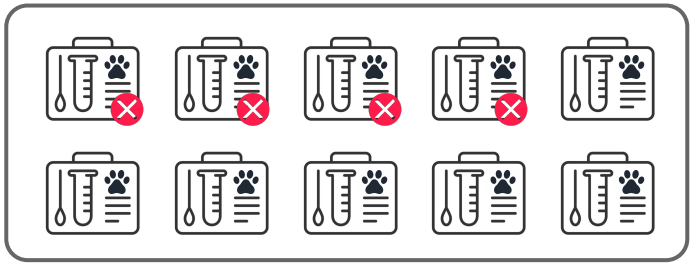
The cumulative distribution function

DeepLearning.AI

Scenario

🔍 The lab has a quality control trigger at 50% valid samples

❌ Fewer than 50% of the samples valid: box is unfit for testing



💬 **Problem**: How often can we expect to hit the quality control trigger?

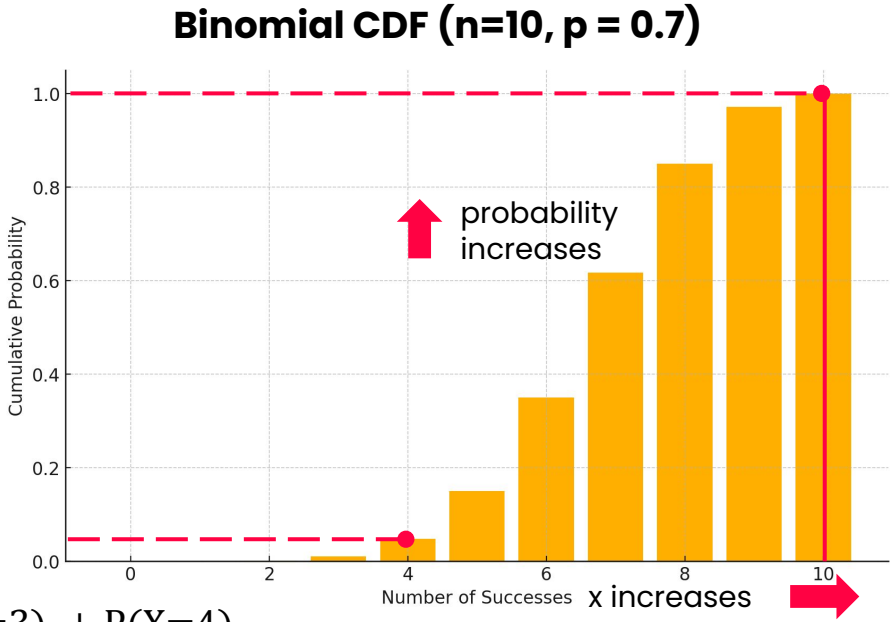
“Unfit” outcomes:

- Box with 4 valid samples
- Box with 3 valid samples
- Box with 2 valid samples
- Box with 1 valid sample
- Box with 0 valid samples

Cumulative distribution function (CDF)

- Models how likely is a value **less than or equal to** a given value
- CDF for discrete random variable X:

  $P(X <= x)$
- For discrete probability distributions, calculate using the **addition rule**



$$P(X<=4) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)$$

$$0.001 + 0.009 + 0.036 = 0.046 = 4.6\%$$

Scenario

How often does the lab get an **above average box?**

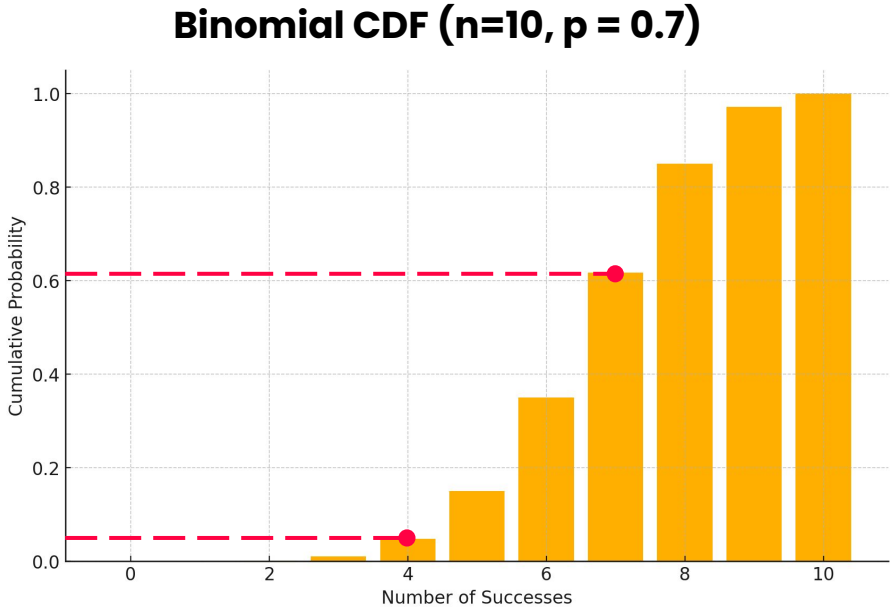
$$1 - P(X<=7) = 1 - 0.61 = 0.39 = \mathbf{39\%}$$

How often does the lab **not hit the quality control trigger?**

$$1 - P(X<=4) = 1 - 0.046 = 0.954 = \mathbf{95.4\%}$$

You answered a lot of useful questions:

- Where is the center of this distribution?
- What is the variability?
- How common are outcomes or ranges?



Use cases

Modeling other scenarios with:

- Yes/No outcomes
- Outcomes with success and failure conditions

Customer conversion rates in marketing campaigns

Defective product rates in quality control

Employee retention in HR analytics

Loan default rates in financial services

Probability and simulation

Random sampling – discrete

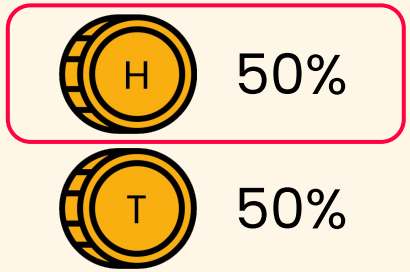
DeepLearning.AI

Developing a simulation model

Probability

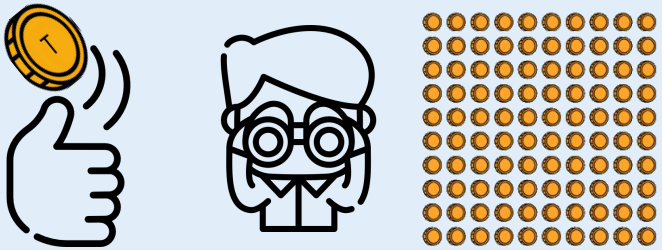
- **Random variable**: all possible values for a particular outcome
- **Probability distribution**: likelihood of each possible value in a random variable

Not enough!



Random sampling

- Generate a specific outcome
- Simulate outcomes according to a probability distribution

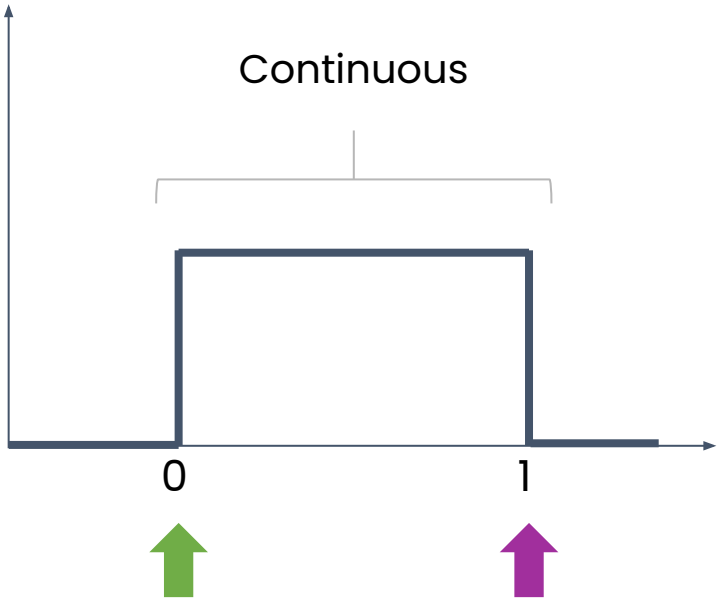


Standard uniform distribution

RAND

- Often used as a starting point to generate random samples from other distributions
- Random number generator between 0 and 1

$$U \sim \text{UNIFORM}(0, 1)$$



Scenario

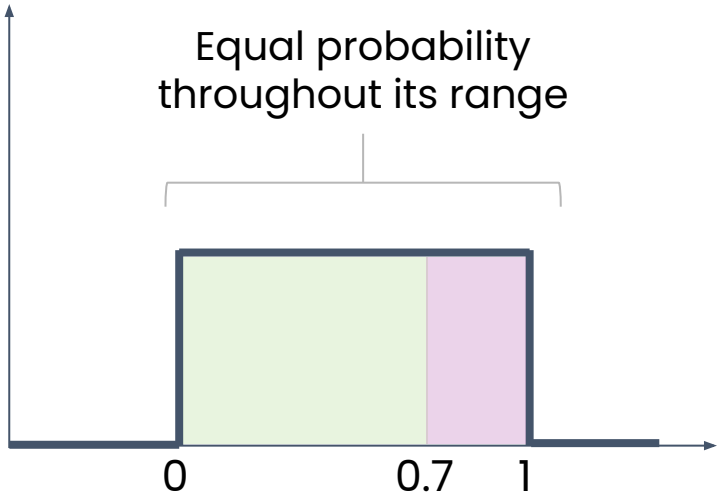
Goal: Simulate validity of one canine DNA test kit

- Generate random sample between 0 and 1 using **RAND**
- If number is <= 0.7, consider the test valid

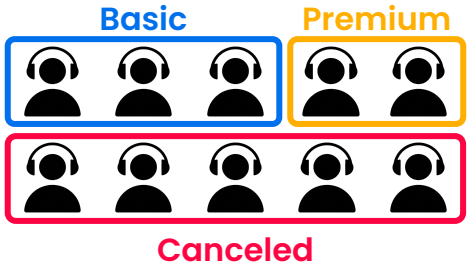
$$P(\text{random value} <= 0.7) = 0.7$$

- Otherwise consider the test is invalid

$$P(\text{random value} > 0.7) = 0.3$$

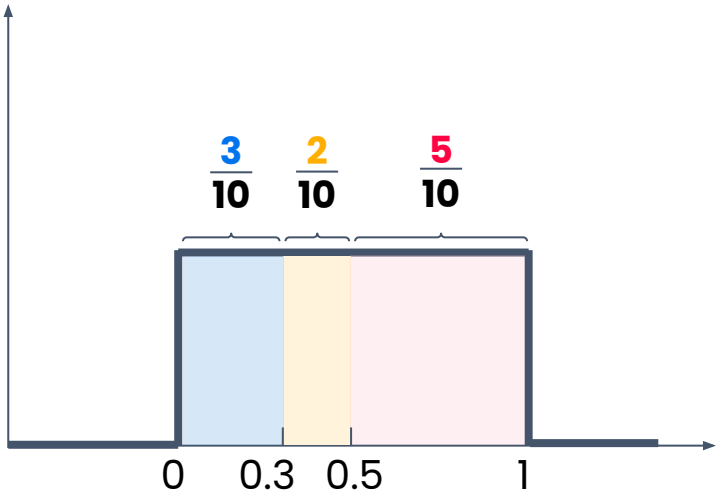


Scenario



Goal: Simulate action taken by one customer

- Divide range from 0 to 1 into three segments:
  - **Basic**
  - **Premium**
  - **Canceled**
- Length of each segment → proportional to probability of the outcome



Power of simulation

Repeat experiment as many times as you want

Perform descriptive analytics

Generate data needed to inform a decision

Simulated data is just another version of sample data

Change parameters to analyze different scenarios

Probability and simulation

Demo: spreadsheet simulation – discrete

DeepLearning.AI

Probability and simulation

Demo: LLM simulation - discrete

DeepLearning.AI

LLMs for simulation

Only LLMs that can write and run code are useful for simulation

LLMs struggle with math

⚠️ Generating random samples is a mathematical task

⚠️ LLMs aren't suited for it unless they can also write and run code

✅ If they can run their own code, then they can calculate probabilities

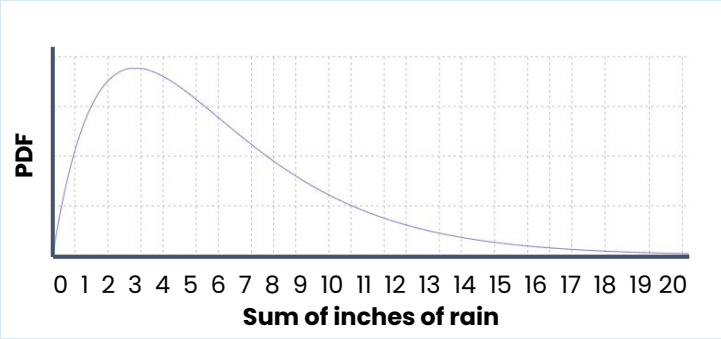
Probability and simulation

Continuous probability distributions

DeepLearning.AI

Continuous vs. discrete probability distributions

Continuous probability distribution



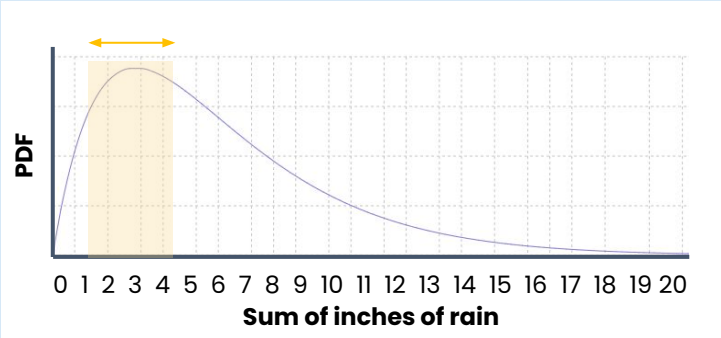
- Visualized with a smooth curve
- Outcomes between each outcome
- Probability density function (**PDF**)

Discrete probability distribution

- Visualized with a column chart
- Distinct, countable values
- Probability mass function (**PMF**)

Continuous vs. discrete probability distributions

Continuous probability distribution



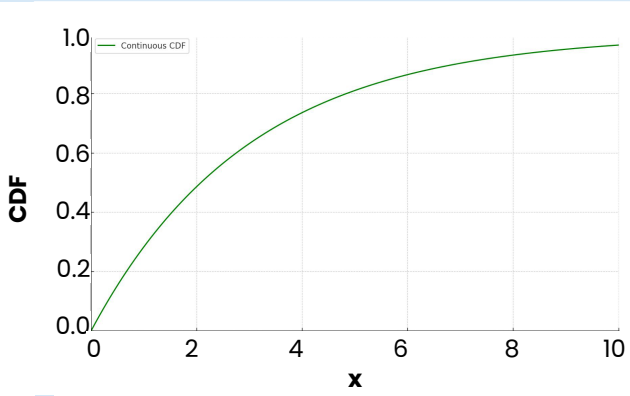
- Calculate probability of value falling within a range
- Probability of any exact value is zero

Discrete probability distribution

- Calculating probability of specific values

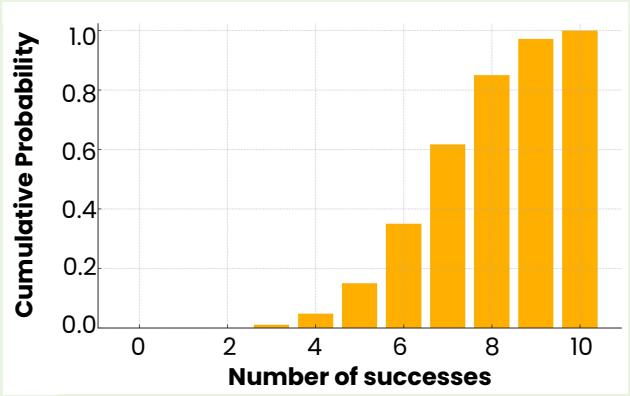
Continuous vs. discrete probability distributions

Continuous probability distribution



- **CDF** is a smooth curve
- Strictly increasing function as x increases
- Range from 0 to 1

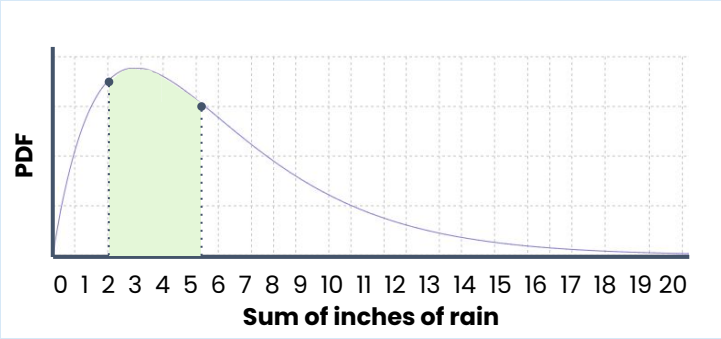
Discrete probability distribution



- **CDF** is a column chart
- Strictly increasing function as x increases
- Range from 0 to 1

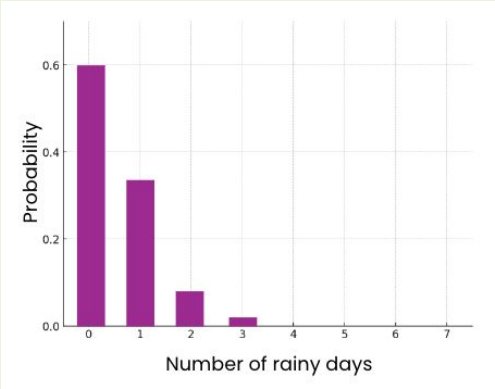
Continuous vs. discrete probability distributions

Continuous probability distribution



- Area under the curve represents probability of value within that range
- Requires calculus to define the area under the curve

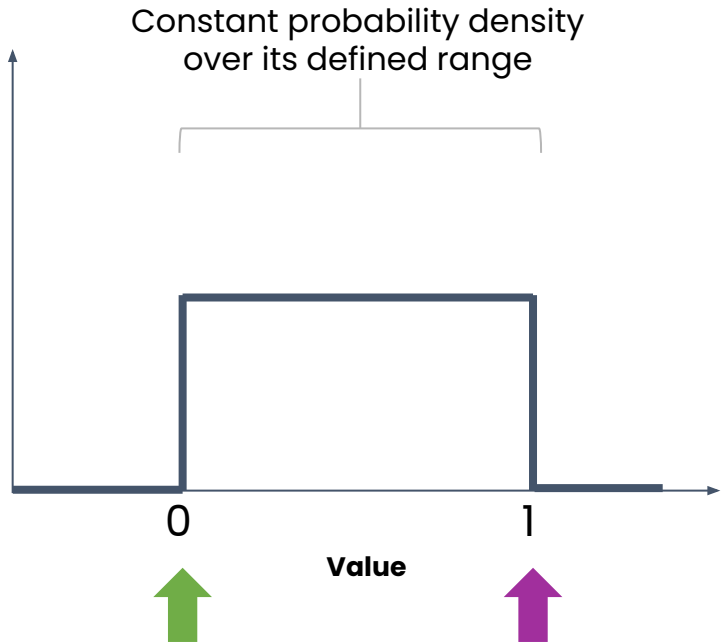
Discrete probability distribution



- Calculating probability of values in a range is straightforward
- Sum the different probabilities

Uniform distribution

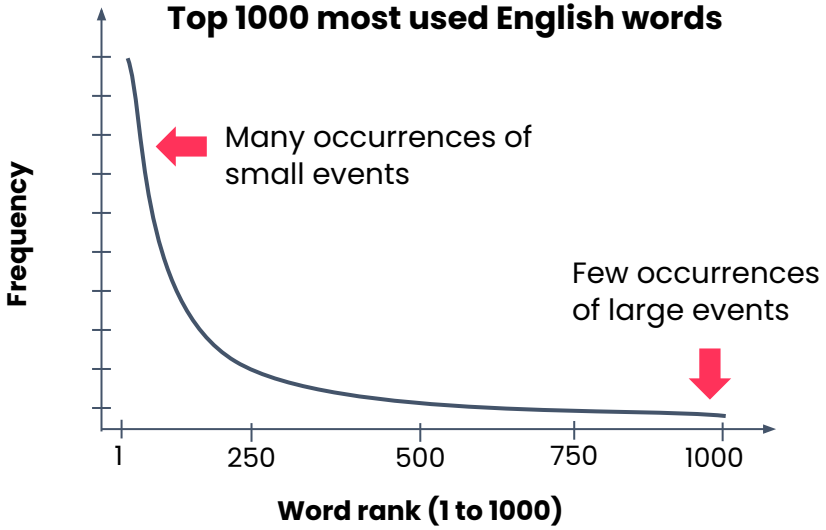
- Use to model a distribution where all outcomes are **equally likely**
- Used when you have little information about behavior of the random variable



Power law distribution

- Also called a **skewed distribution**
- Used to model data where probability is inversely proportional to size
- Characterized by “long tail” where rare events still have a meaningful probability of occurring
- Often associated with 80-20 rule

80% of effects ← **20%** of causes



City population sizes

Earthquake magnitude

Income distribution

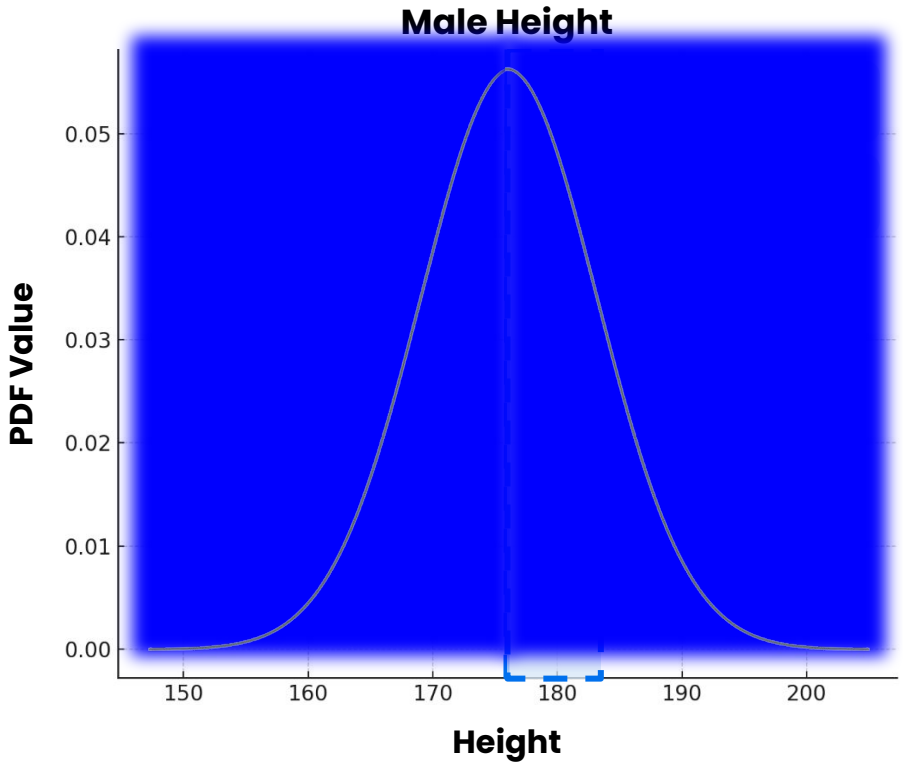
Probability and simulation

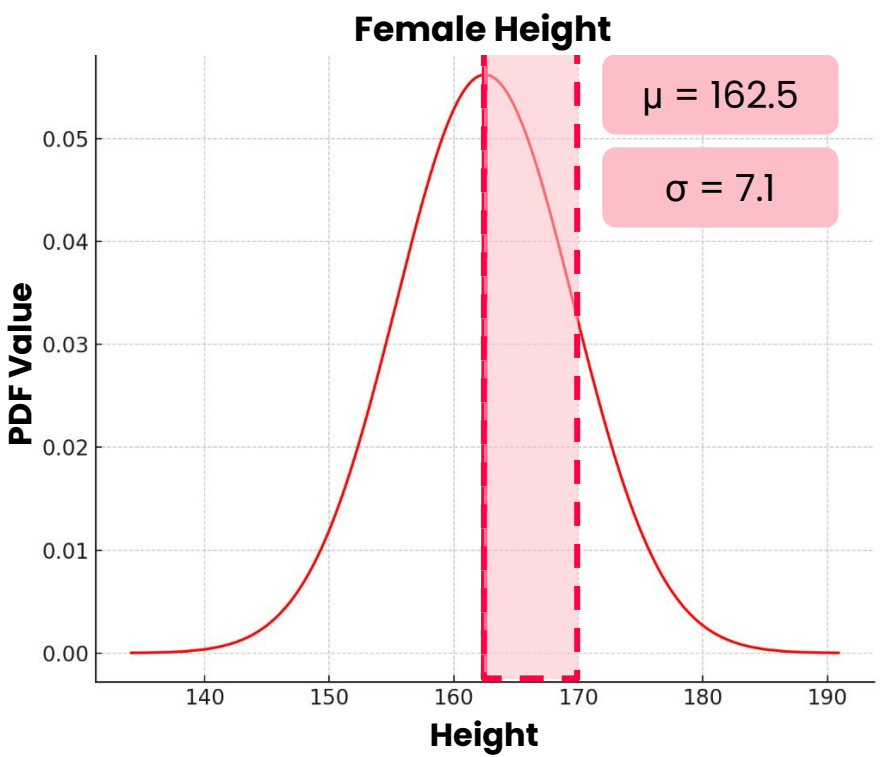
The normal distribution

DeepLearning.AI

Normal distribution: height

$$H \sim \text{Normal}(162.5, 7.1)$$





Using the normal distribution

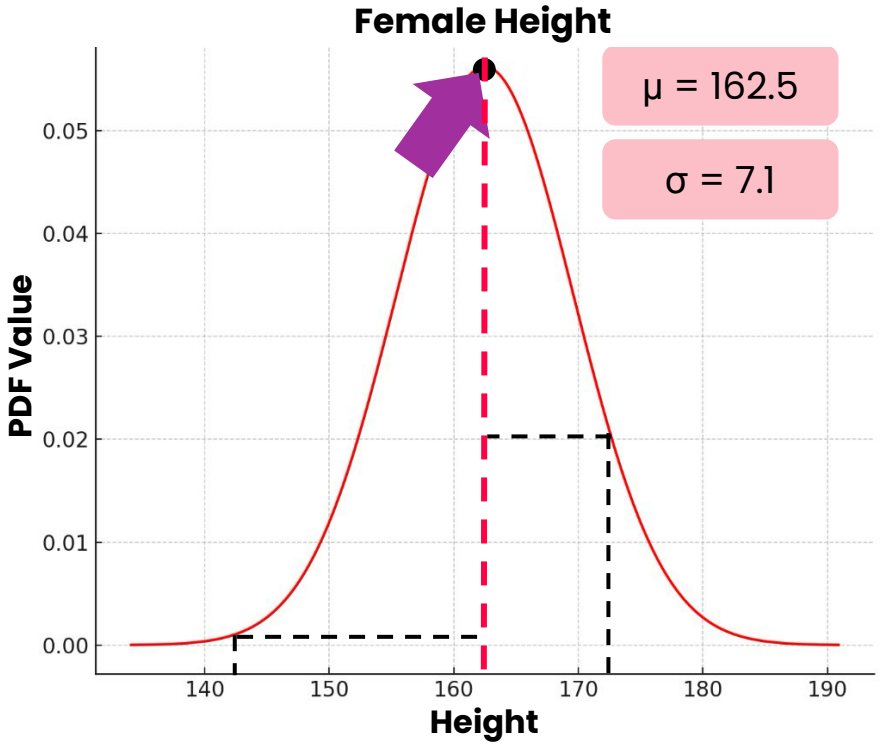
$$H \sim \text{Normal}(162.5, 7.1)$$

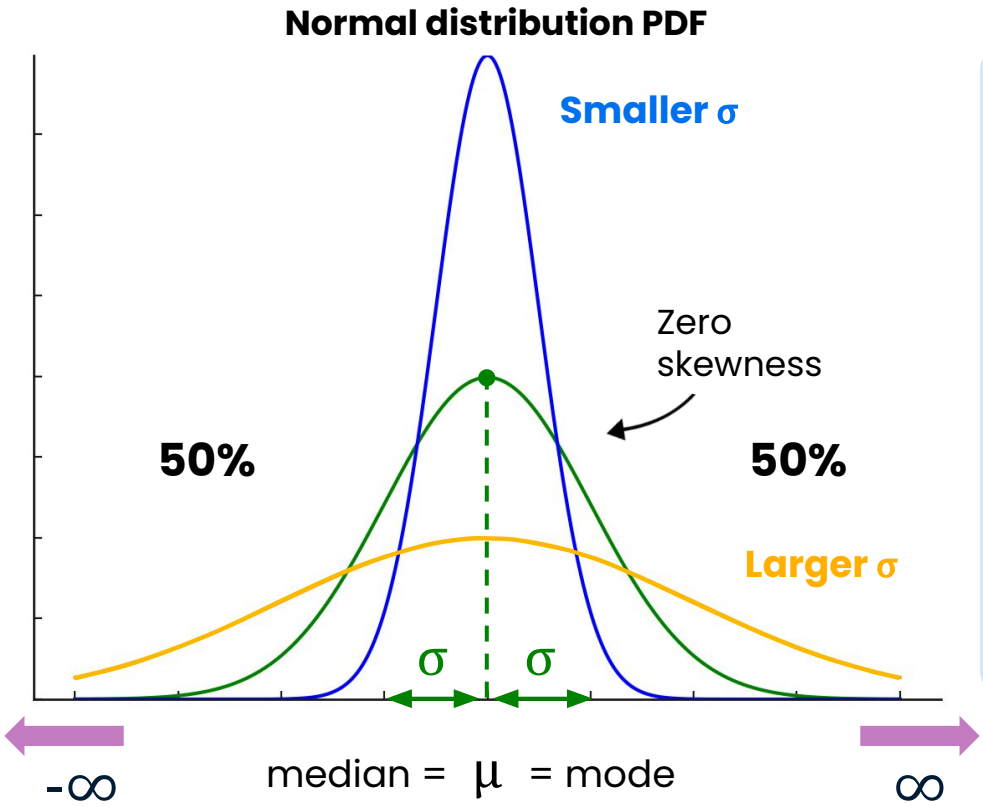
Q: Which female height is most common?

162.5 cm

Q: Is it more likely that any given female is 142 or 173 cm?

173 cm

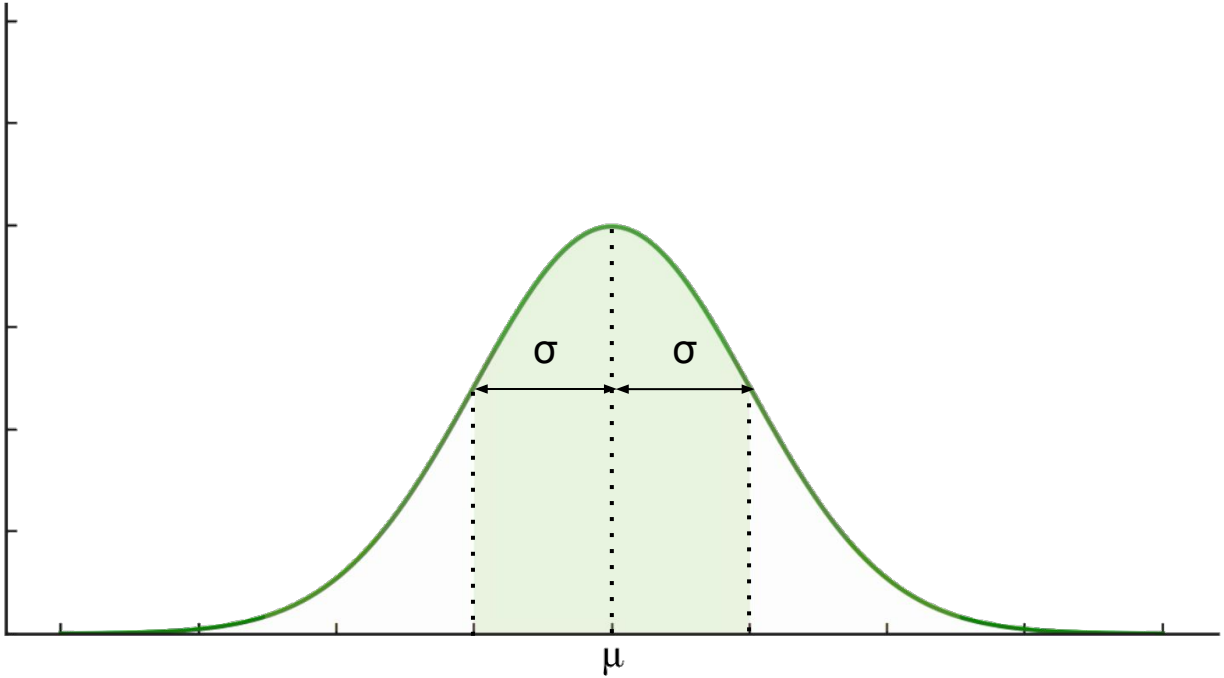




Properties of normal distribution

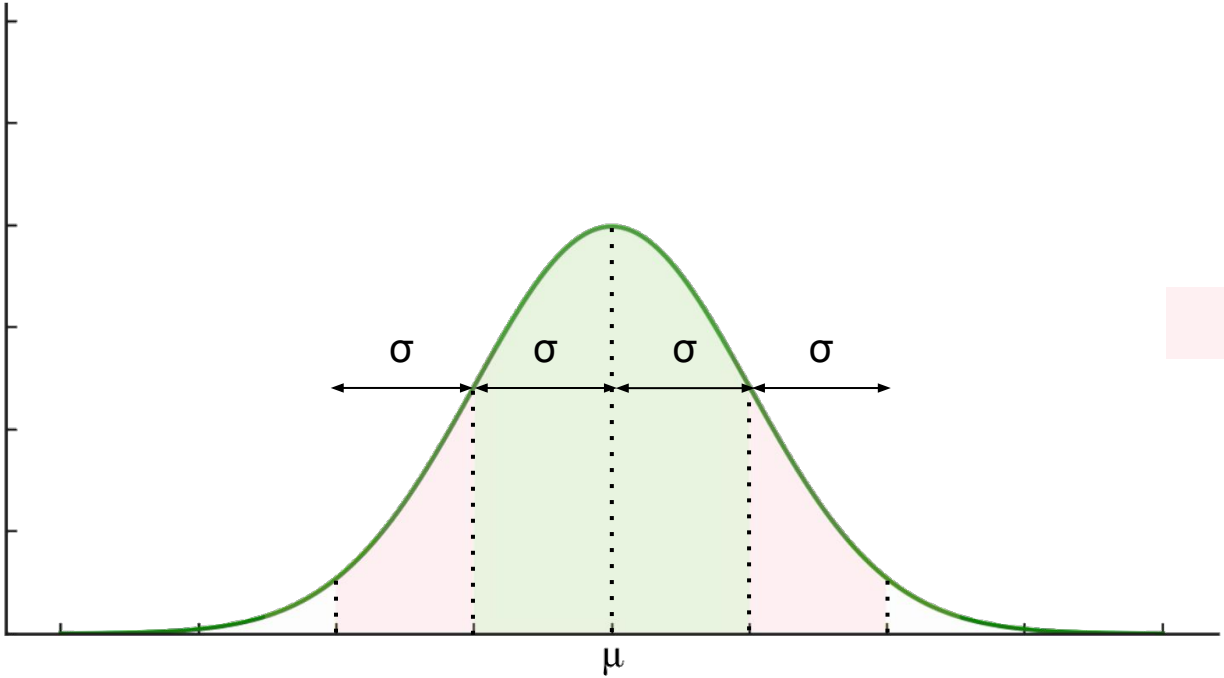
1. Symmetrical around the mean μ
   - 50% of the data is to the left of the mean and 50% is to the right
2. Mean, median, and mode are all equal
   - Zero skewness
3. Spread of the distribution is defined by the standard deviation σ

Sigma rules



68% of the data

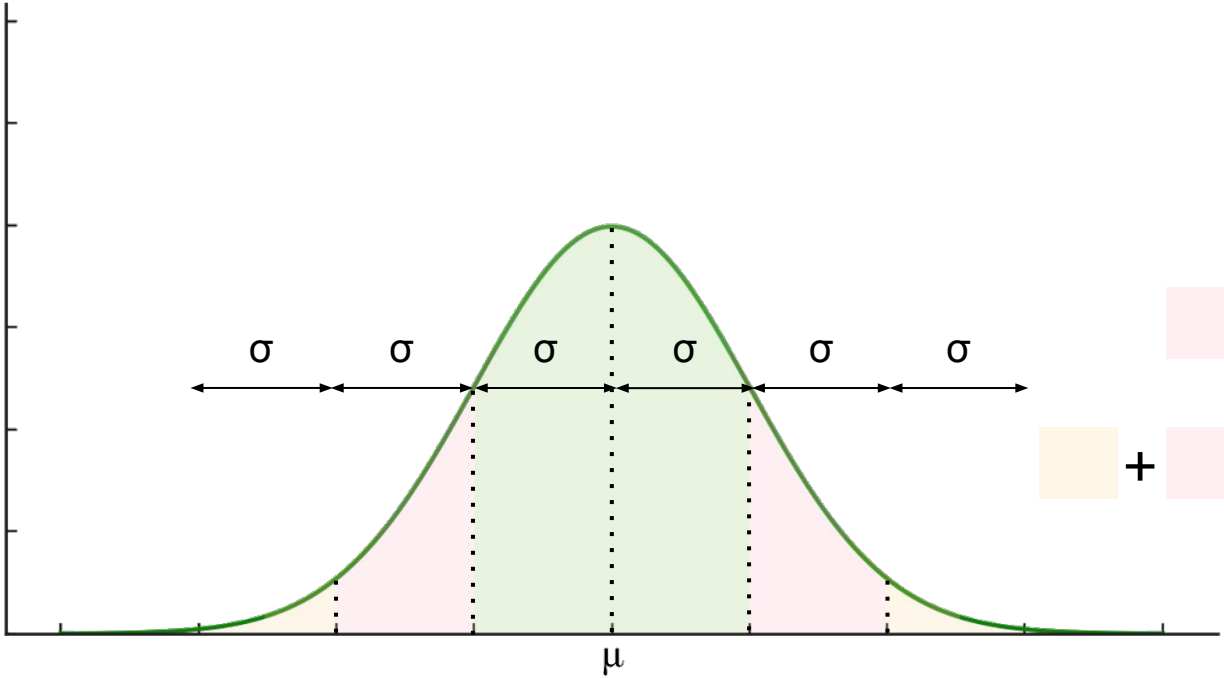
Sigma rules



68% of the data

\+ 95% of the data

Sigma rules

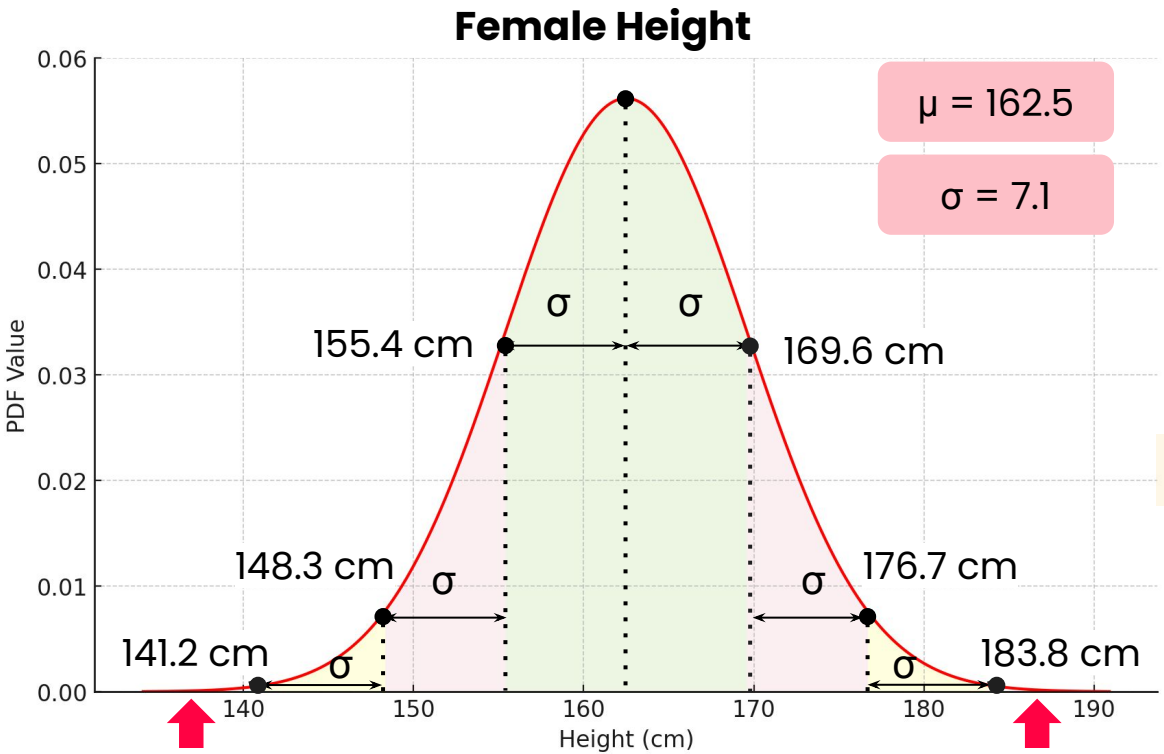


68% of the data

\+ 95% of the data

\+ \+ 99.7% of the data

Sigma rules



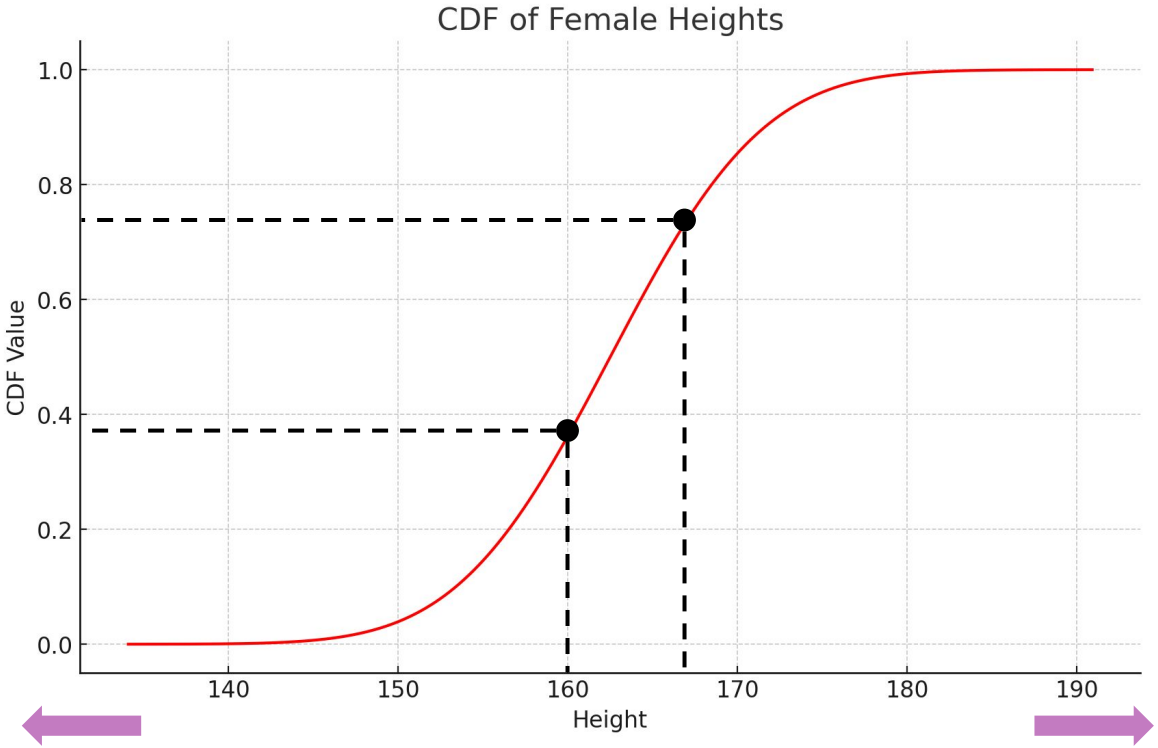
68% of females

\+ 95% of females

\+ \+ 99.7% of females

0.3% of females is still more than 11 million people globally

Normal distribution CDF



Q: What is the probability that a given female is 160cm or less?

A: about 39.2%

Q: What is the height of the 75th percentile of females?

A: about 167.29cm

Probability and simulation

The standard normal distribution

DeepLearning.AI

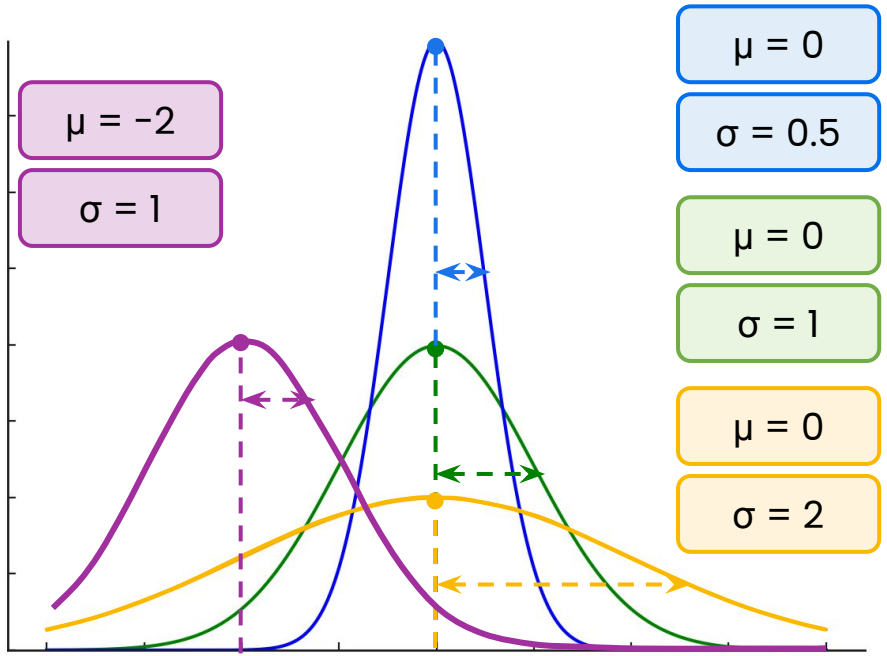
Normal distributions

- Infinitely many normal distributions
- Center is determined by the mean
- Shape is determined by the standard deviation

Properties of normal distribution

✅ mean = median = mode

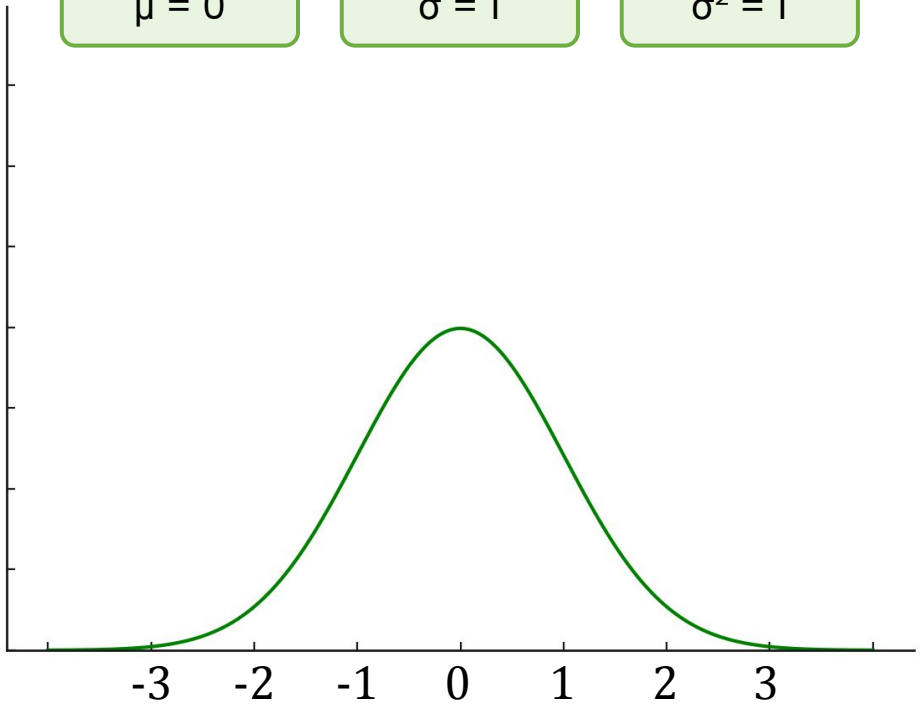
✅ Follow the sigma rules



The standard normal distribution

$$Z \sim \text{Normal}(0, 1)$$

$\mu = 0$ $\sigma = 1$ $\sigma^2 = 1$

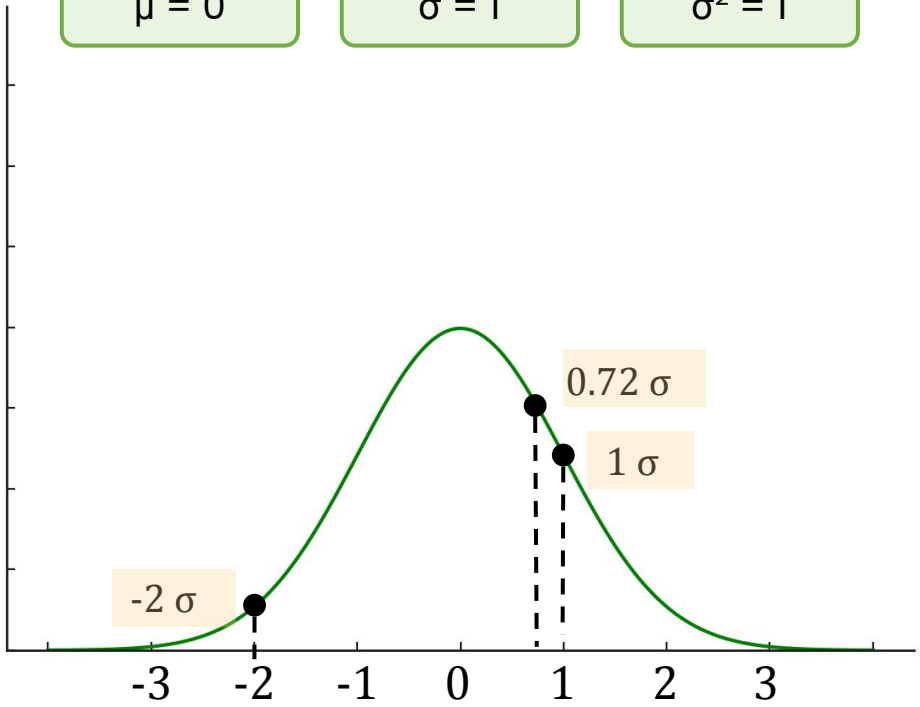


σ

The standard normal distribution

$$Z \sim \text{Normal}(0, 1)$$

μ = 0 | σ = 1 | $\sigma^2 = 1$

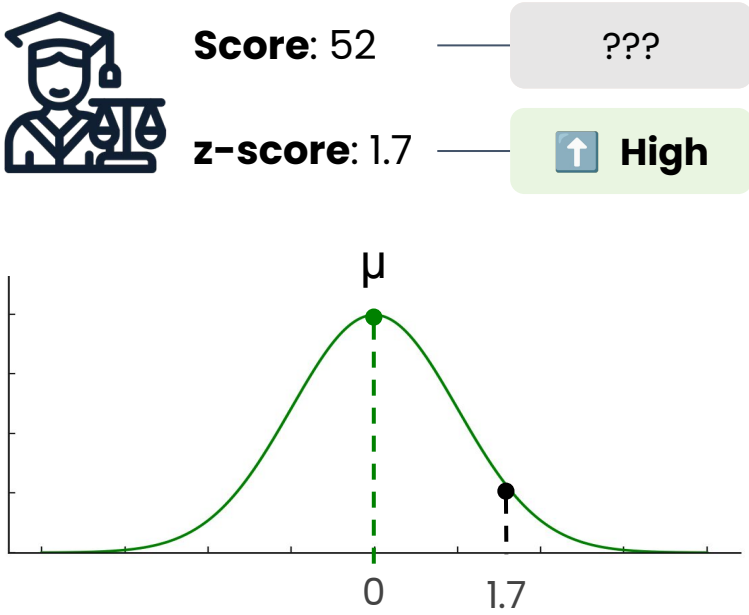


- ✅ Important mathematical properties
- ✅ Value of outcome corresponds with standard deviations from mean

z-scores

z-score

- Provide a common reference for normally distributed data
- **Similar to percentiles:** intuition about position of data point in distribution
- **Z scores:** provide information about how far away value is from mean
- **Percentiles:** only provide information about rank of that value



Standardization

Score: 52

μ = 43.5

σ = 5

- To **transform** data into z-scores
- To interpret distribution of data according to a consistent scale
- After this transformation, all normal distributions will have:
  - Mean of 0
  - Standard deviation of 1

$$z = \frac{x - \mu}{\sigma}$$

Centers distribution at 0

Expresses values in units of σ

$$z = \frac{52 - 43.5}{5} = 1.7 \text{ z-score}$$

Reversing z-scores

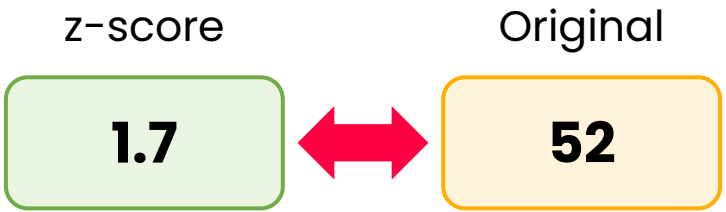
- If you have:
  - z (z score)
  - μ
  - σ

- It's the opposite of the calculation

$$z = \frac{x - \mu}{\sigma}$$

Inverse transformation

$$x = (z * \sigma) + \mu$$



Probability and simulation

Random sampling – normal

DeepLearning.AI

Simulation scenario

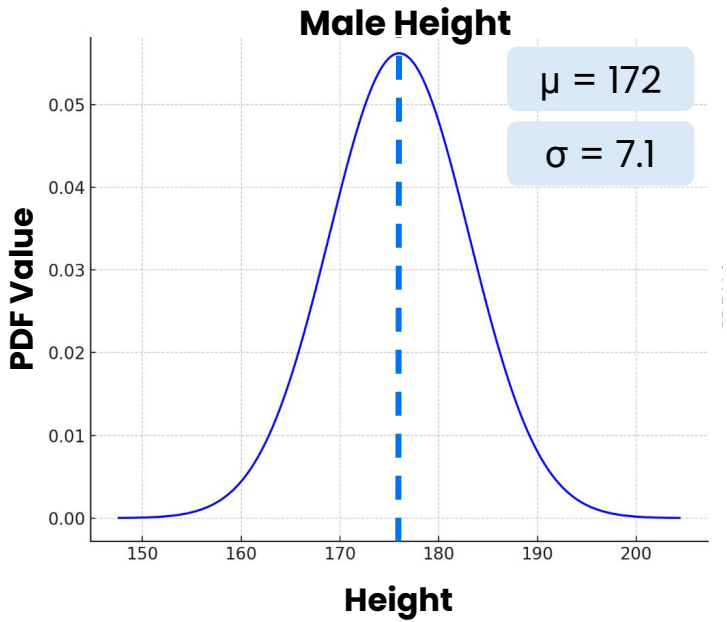
Diving company that supplies diving suits

Goal: Simulate heights of diving group of 10 for suit sizing

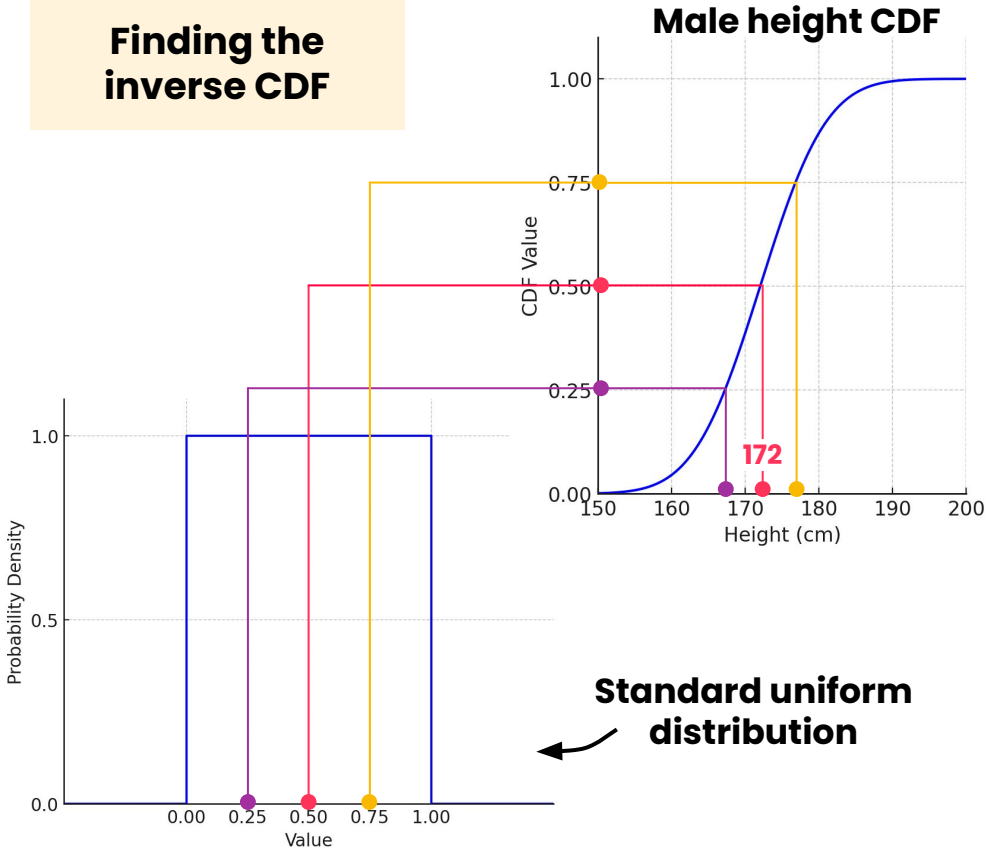
Step 1: Use a random number generator to produce a sample

0.22189401

Step 2: Use inverse transformation to convert value to normal distribution



Finding the inverse CDF



NORM.INV

Identify corresponding **x** for a given CDF value and parameters

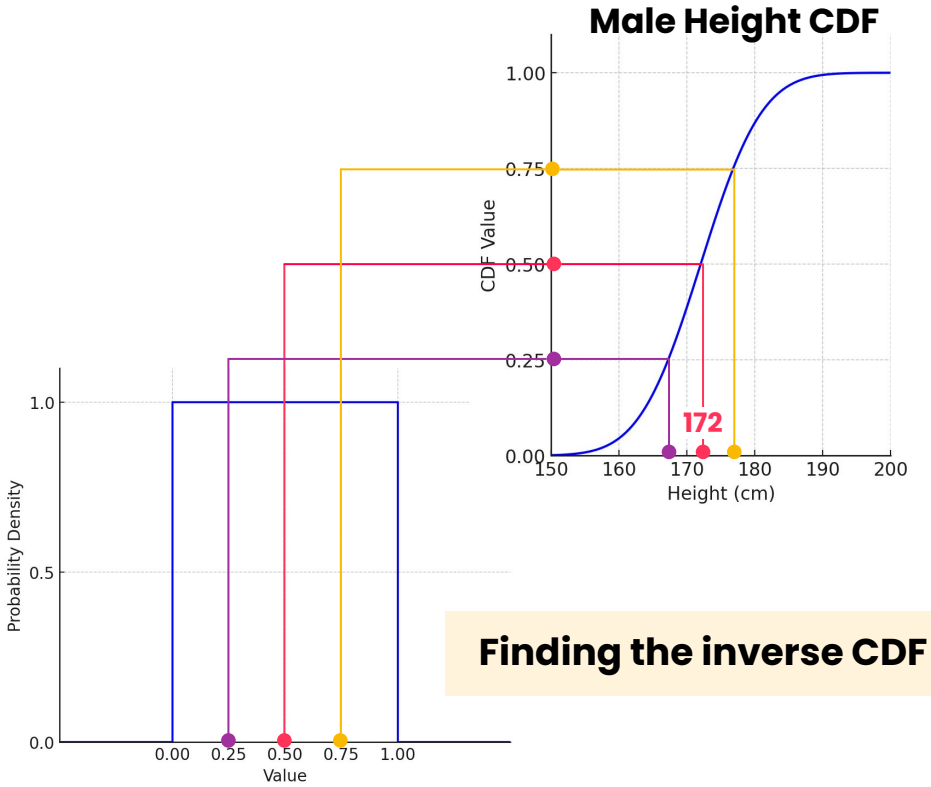
- Repeat 9 more times to simulate 10 customers
- By sampling many times, you can estimate how many of each size you need

Simulation scenario

NORM.INV

Identify corresponding **x** for a given CDF value and parameters

- Repeat 9 more times to simulate 10 customers
- By sampling many times, you can estimate how many of each size you need



Finding the inverse CDF

Probability and simulation

Demo: Spreadsheet simulation - normal

DeepLearning.AI

Probability and simulation

Demo: LLM simulation - normal

DeepLearning.AI

Probability and simulation

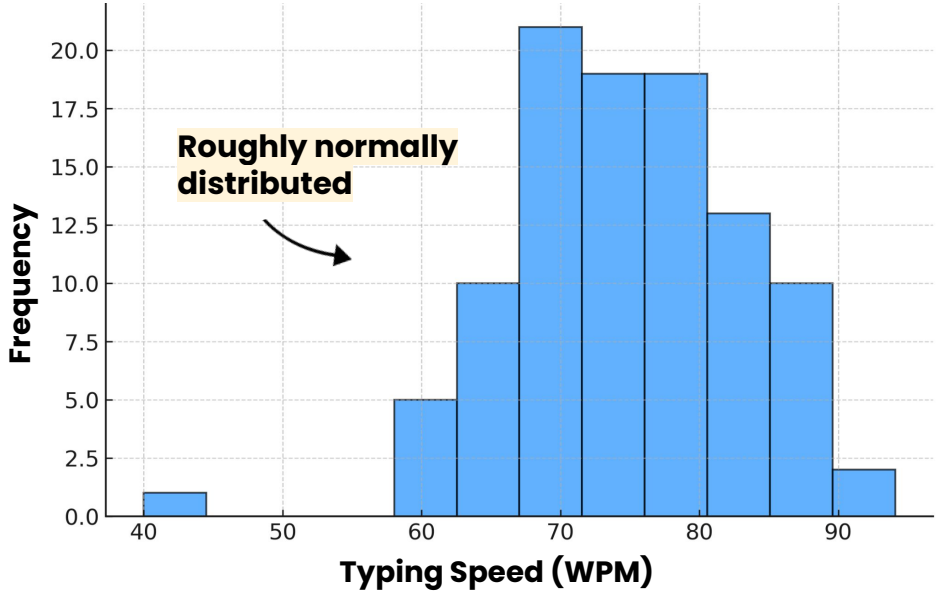
Making decisions with distributions

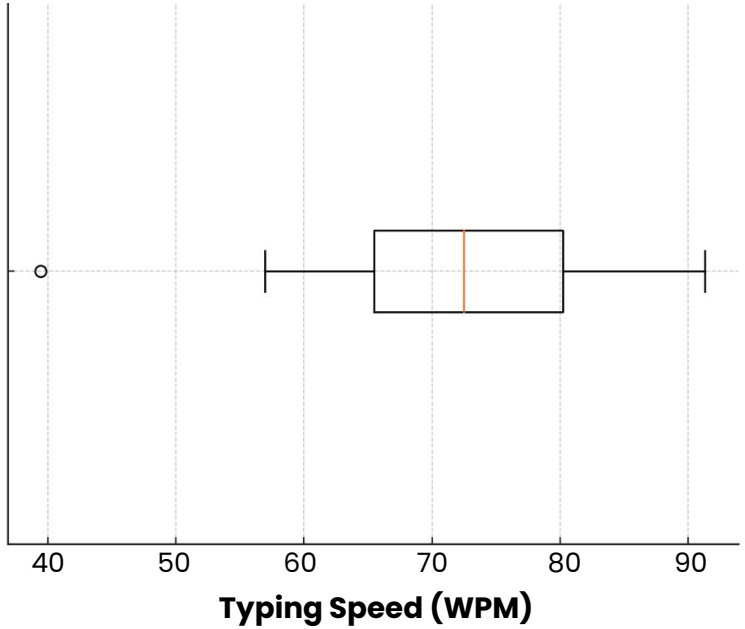
DeepLearning.AI

Concern: User's typing speed might affect their ability to score well

Random sample of **100** users

Measure typing speed





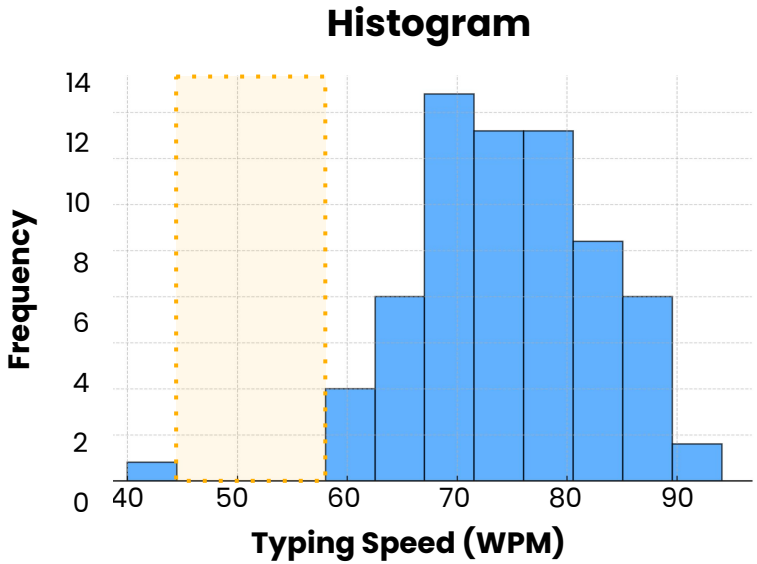
1 Descriptive statistics:

- If population does follow the normal distribution:

$\bar{x} = 72.47$ $s = 9.59$

μ σ

Population parameters



2 Address some of your business questions:

Q. "What's the likelihood that any given user types below 40 words per minute?"

✅ Estimate probability of observing that user given the parameters

➡ About **0.035%**, or **1 in 2800 users** ➡

Scenario

1 Descriptive statistics:

- If population does follow the normal distribution:

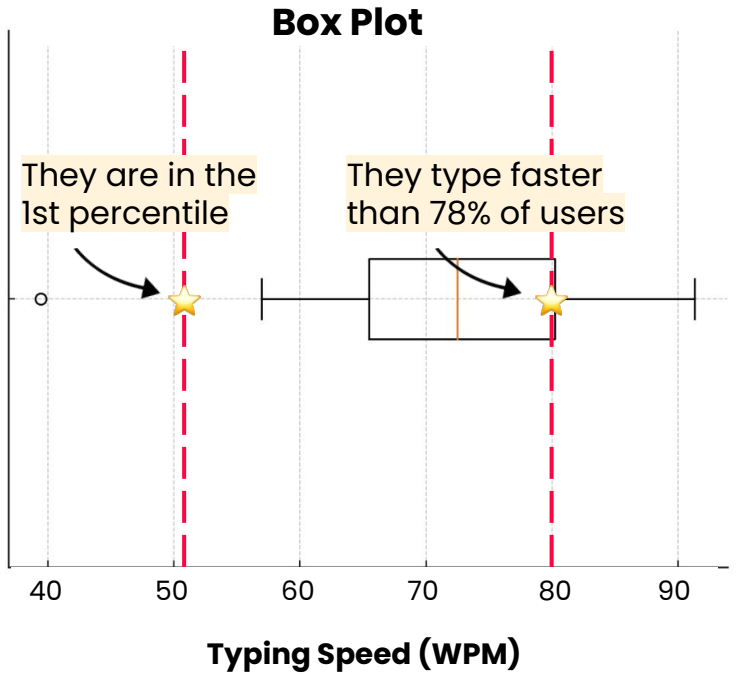
$\bar{x} = 72.47$ $s = 9.59$

μ σ

Population parameters

2 Address some of your business questions:

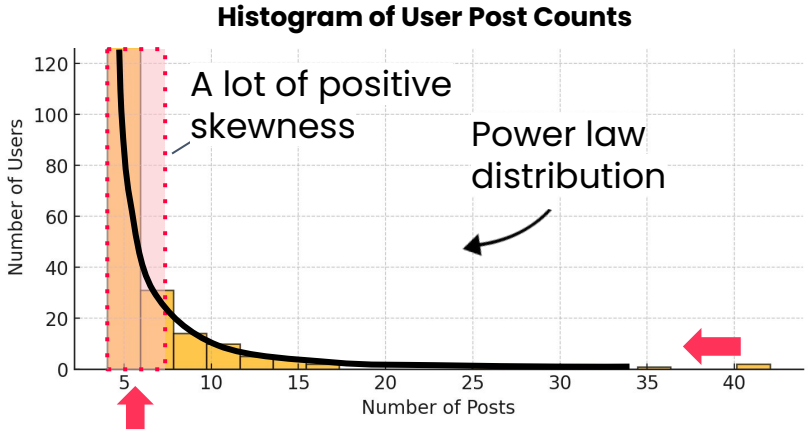
Complaint: "The test timing was unfair due to my typing speed."

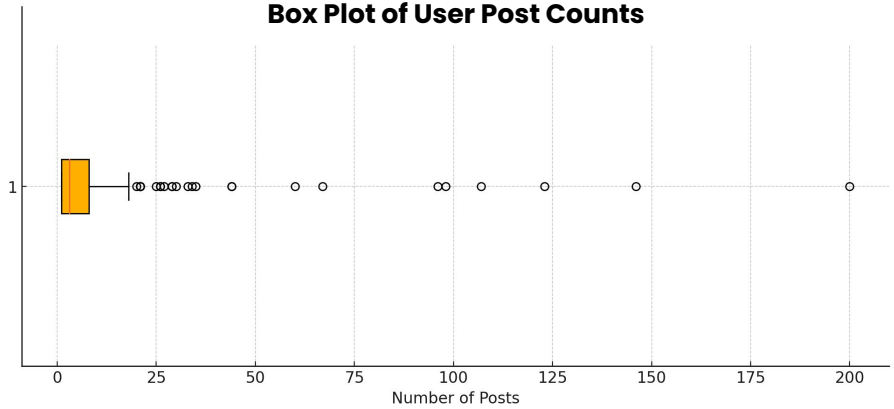


Scenario

Task: Figure out how to incentivize people to post more

- Random sample of **200** users
- Follow each user for a week
- Collect data about how much they are posting



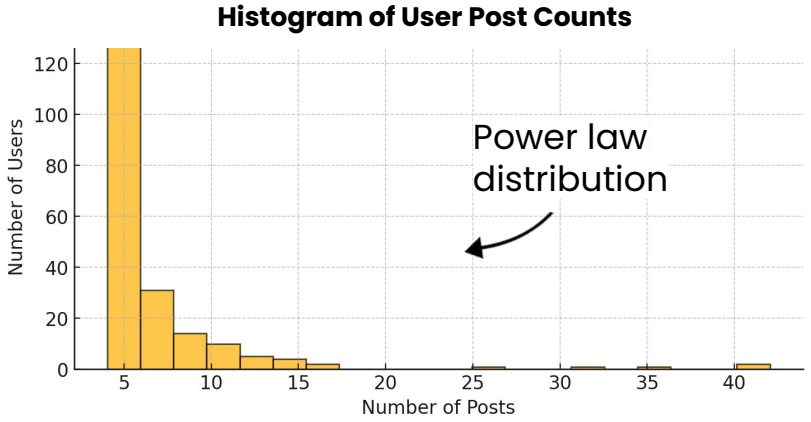


Scenario

Use **power law distribution** as a model for how population behaves on the social media site:

1. Sample statistics: $\bar{x} = 14.465$ $s = 23.8$

2. Address some of your business questions:

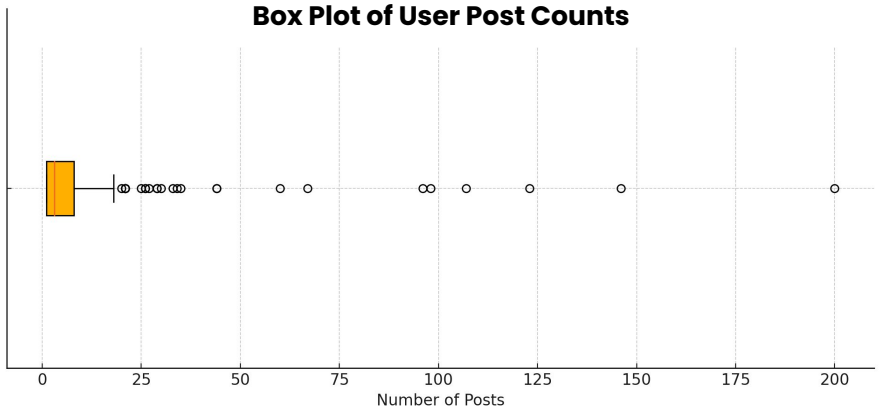


"A particular incentive causes users in the bottom 50% of activity to create one more post per week"

Median posts would increase by **33%**

Mean posts would increase by about **5%**

Total posts would increase by about **5%**



Probability and simulation

Coursera dialogue item introduction

DeepLearning.AI