Berkeley Engineering | BerkeleyHaas

PROFESSIONAL CERTIFICATE IN MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Module 2

Fundamentals of Statistics and Distribution Functions

Office Hours with Viviana Márquez
April 10, 2025

Let’s give everyone a couple of minutes to join…

AGENDA

- Meet your Learning Facilitator!
- Office Hour Expectations
- Walkthrough Canvas
- FAQ Python
- Content review Module 2: Fundamentals of Statistics and Distribution Functions
- Questions

Berkeley Engineering | BerkeleyHaas

Meet your Learning Facilitator!

**Hi everyone,
I'm Viviana Márquez,
your Learning Facilitator!
👋**

- MSc. in Data Science from the University of San Francisco, CA, USA. | BSc. in Mathematics and AA. in Media Production
- Founder, AI Consultant & Educator @ Miss Factorial Academy
- Six years of industry experience in the United States across different sectors, including marketing, streaming services, cybersecurity, hospitality, and startups
- I've taught AI & Data Science to over 3,000 students in North America, Europe, Africa, and South America

Office hours: Every other Thursday
Always check Canvas for the most up-to-date information!
Tool to convert to your timezone:
https://www.worldtimebuddy.com/

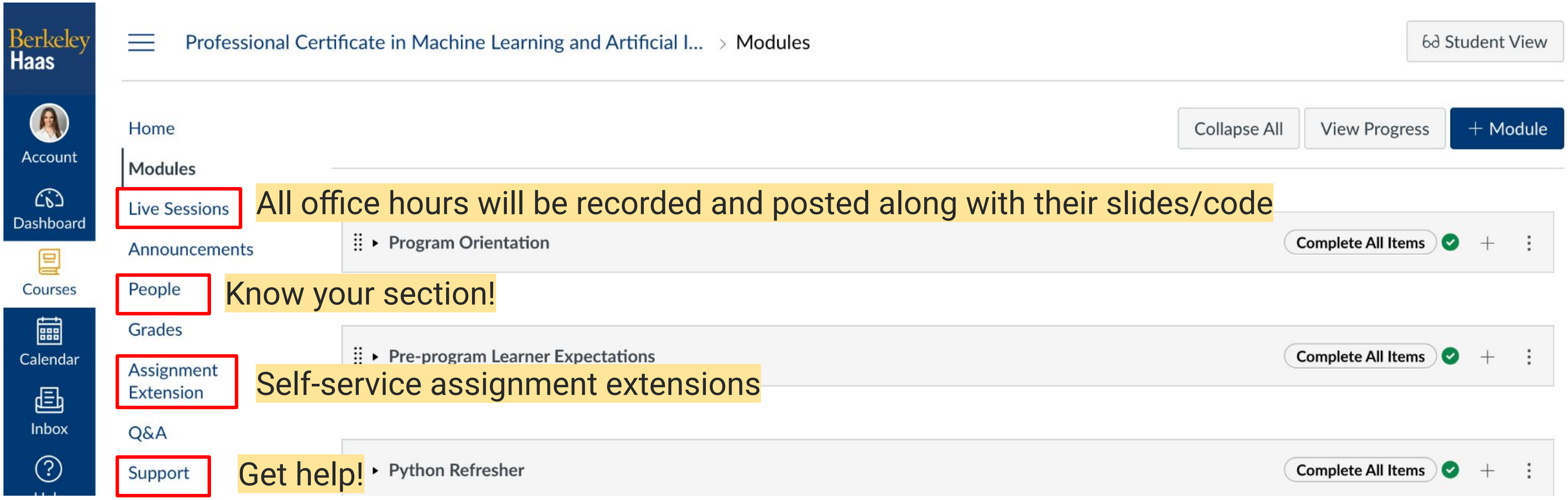
Office Hours Expectations

- Content is released every Wednesday
    - Each Learning Facilitator has their own style, so Office Hours won't be repetitive but will be based on that week's module content
    - I will provide you with a content review, share industry insights/code, and open it up for questions
    - My commitment to you is to make complex technical topics not just understandable but also relatable to the industry and genuinely fun! 👾

- For questions that are unique to you, please submit a ticket

- Everyone can come to any of the Office Hours (they're not mandatory but highly encouraged!) Your section only influences who grades your Practical Applications/Capstone project and who you'll meet for your 1:1 sessions later in the course. I'm responsible for **Section C**.

CANVAS



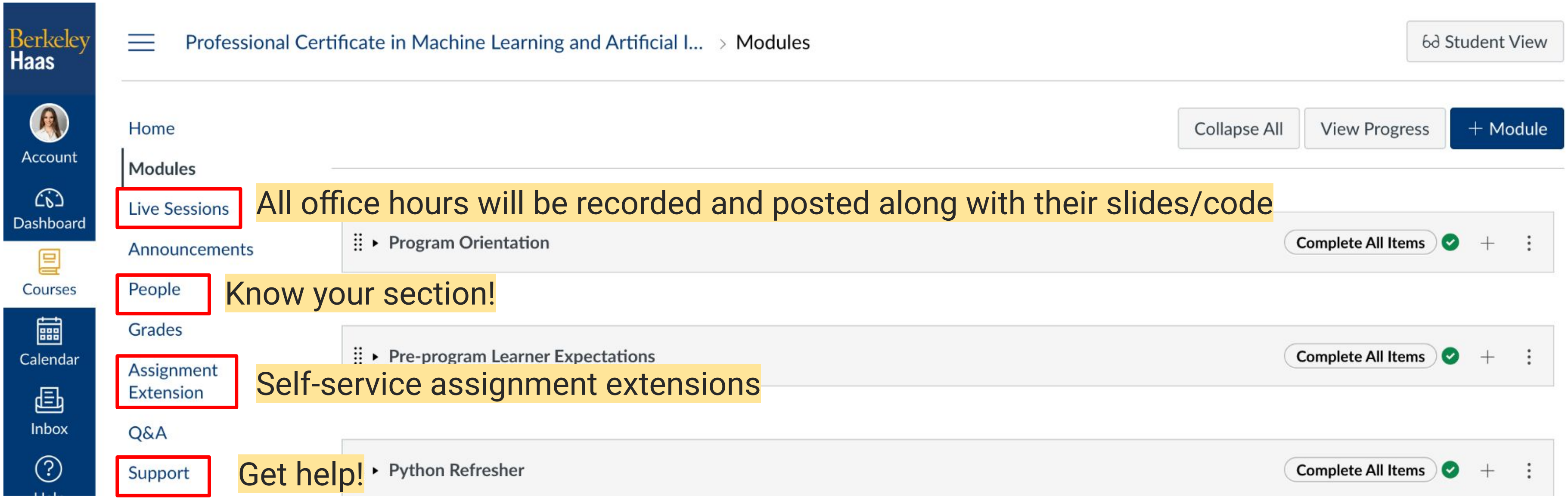
All office hours will be recorded and posted along with their slides/code

Know your section!

Self-service assignment extensions

Get help!

CANVAS



Professional Certificate in Machine Learning and Artificial I... › Modules

All office hours will be recorded and posted along with their slides/code

Know your section!

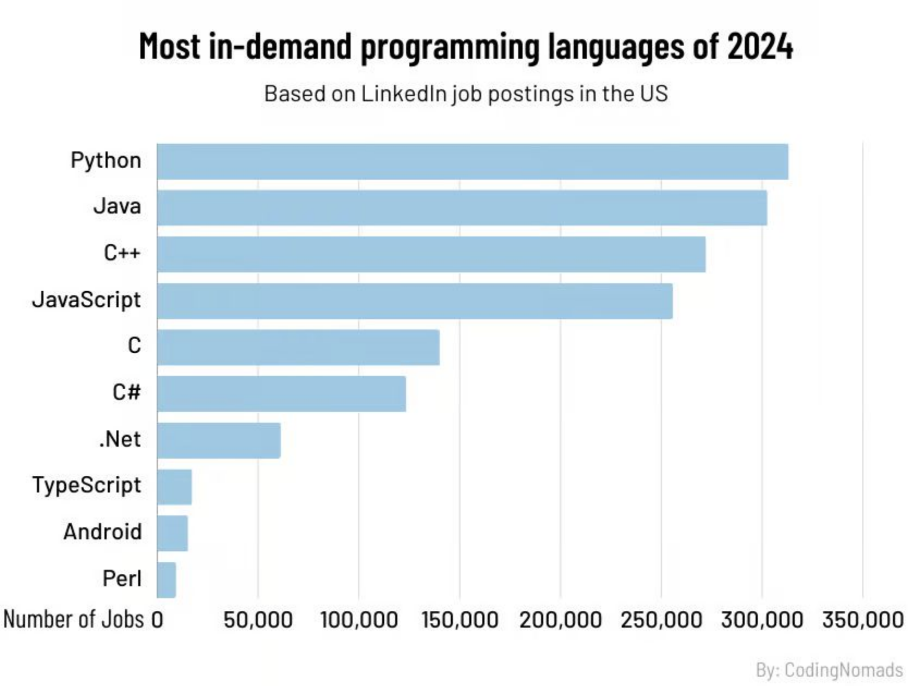
Self-service assignment extensions

Get help!

Only use the support tab to get help.
Canvas inboxes are **not** monitored! Your message might go unanswered.

Python

- Python is a programming language known for its readability and versatility
- Widely used for machine learning algorithms



Most in-demand programming languages of 2024

Based on LinkedIn job postings in the US

By: CodingNomads

Python: Where can you write and execute it?

- **IDE (Integrated Development Environment)**
  IDEs are all-in-one solutions, providing comprehensive facilities for software development. They include a code editor but also integrations for debugging, project management, version control, testing, and building executables, among other tools.
  Examples: PyCharm, Spyder, Cursor

- **Code editors**
  Code editors are typically lightweight and fast. They are primarily used for editing code, with a focus on syntax highlighting and auto-completion.
  Examples: Visual Studio Code, Notepad++, Text Wrangler

- **Web-based interactive platforms**
  Online environments that facilitate the writing, running, and sharing of code directly within a web browser
  Examples: Jupyter Notebook (local/hosted) , Jupyter Lite, Google Colab

- **Terminal/Command Line**
  Allows for running Python scripts and using Python's interactive mode
    - Additionally, in Unix-like systems, the terminal is often used for installing Python packages, managing virtual environments, and performing various development-related task

Python files can have two extensions: .py and .ipynb

- `.py` **Files**
  - Standard Python script files
  - These are text files containing Python code
  - When a `.py` file is run, the Python interpreter executes all the code in the file from top to bottom. There's no concept of "cells" like in a notebook; it's a continuous script.
  - Files are interacted with using text editors, IDEs, or the command line. They don't support rich text or interactive visualization out-of-the-box.

- `.ipynb` **Files**
  - JSON documents used by Jupyter Notebook, an interactive computational environment that allows users to mix executable code, rich text, visualizations, and other media
  - Each code cell can be run individually, and the output will be displayed directly below the corresponding cell
  - These files are typically viewed and edited using a web browser through Jupyter's web application

When to use .py and .ipynb?

- `py` **Files**
  - Production Code (they can be integrated into larger applications or workflows)
  - Large-Scale Projects (better modularity and code reusability)
  - Automation
  - Software development

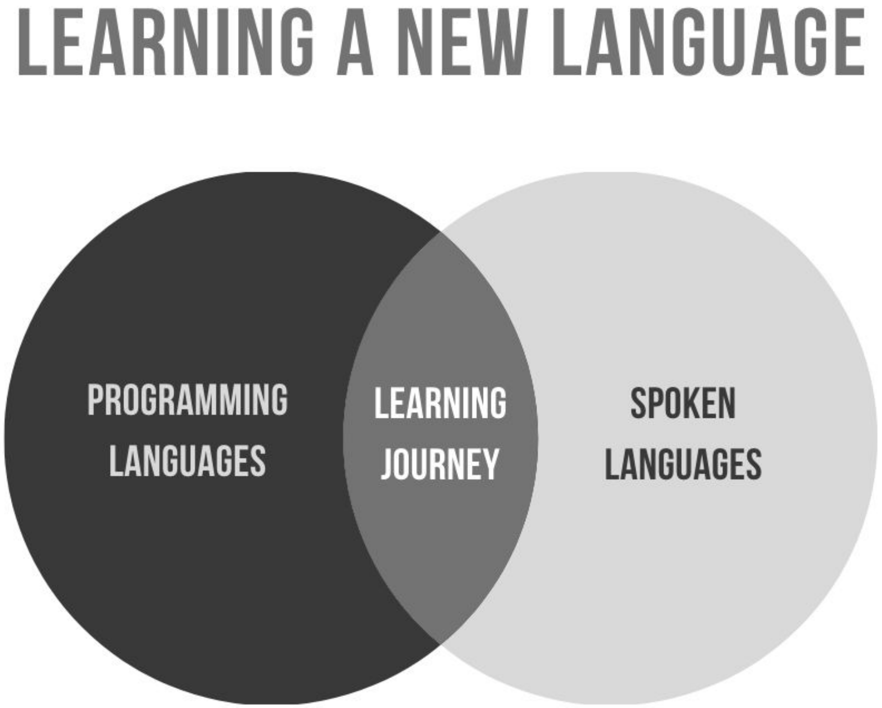
- `.ipynb` **Files**
  - Exploratory Data Analysis (EDA)
  - Documentation and Training
  - Presentations and Reporting
  - Collaborative Work

Getting started

- Installing Anaconda
  https://www.anaconda.com/products/distribution
  Video: https://www.youtube.com/watch?v=IMrxB8Mq5KU

- Alternatively, you could use Google Colab (the Google Docs of code)
  https://colab.research.google.com/

Learning a programming language is just like learning a language!



- Programming languages allow you to communicate with the computer and give instructions to it
- Just like learning a natural language, don't be afraid to make mistakes!
- Practice the language every time you have an opportunity. Try to think in that language.
- Be patient but persistent

Resources:

- https://www.hackerrank.com/domains/python
- https://projecteuler.net/archives
- https://adventofcode.com/
- https://www.kaggle.com/
- Create your own project!

AGENDA

- Meet your Learning Facilitator! ✅
- Office Hour Expectations ✅
- Walkthrough Canvas ✅
- FAQ Python ✅
- Content review Module 2: Fundamentals of Statistics and Distribution Functions
- Questions

Content review Module 2: Fundamentals of Statistics

- What’s data science and why study statistics/probability?
- Probability vs Statistics
- Variables
- Descriptive statistics
- Correlation
- Probability distributions
- Central limit theorem

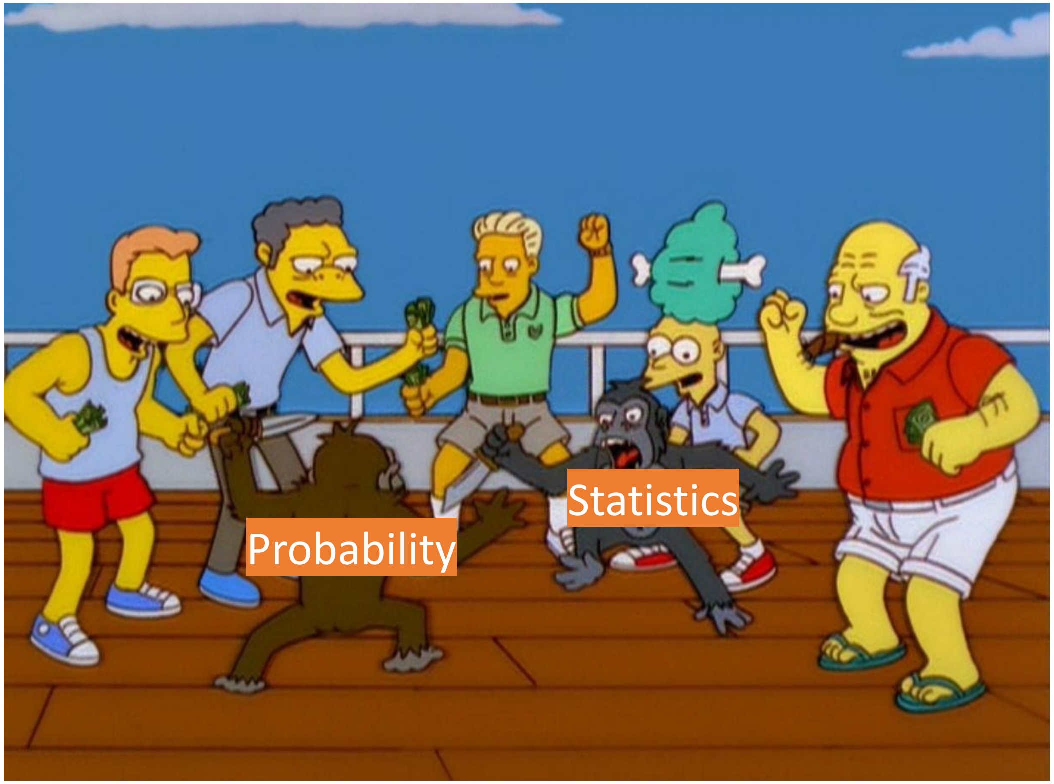
What is a data scientist?
What is a data professional?

A data scientist is someone who works with **data** to unlock its immense potential!

Data is the new oil, a valuable resource that, when refined, can power insightful decisions!

How do extract value from data?

- Statistics, mathematics, probability, linear algebra, machine learning, deep learning, coding, databases, data visualization, software engineering, and more!
- Ultimately, any skill that allows us to understand, model, and draw actionable insights from data is part of a data scientist's arsenal!


Statistics
Probability

Probability vs Statistics

- **Probability:**
  - Predicts likelihood of future events
  - **Example:** Likelihood of getting either a head or a tail using a fair coin is 0.5
  - Provides a theoretical model of what **should** happen

- **Statistics:**
  - Analysis of the frequency past events
  - **Example:** Collect data from a series of actual fair coin flips and analyze it
  - Examines what **does** happen in the real world

Over a small number of flips, it’s possible to get a result that doesn’t exactly match the theoretical probability. But if you do it a “huge” number of times you would expect the result to be very close to the theoretical probability.

Statistics and probability essentials Specifically tailored for data science and machine learning	https://docs.google.com/spreadsheets/d/1iV2F0RJ7AgtNxPw8CI69nE82r5MKBtCMnEI5LM_wKmg/edit?usp=sharing

Variable

A variable is a symbol or quantity that can take different values.
It may represent a measurable quantity, a category, or an unknown.

Variables
Dependent variable

A variable that you observe or measure in an experiment, and its value depends on one or more other variables (typically called independent variables)

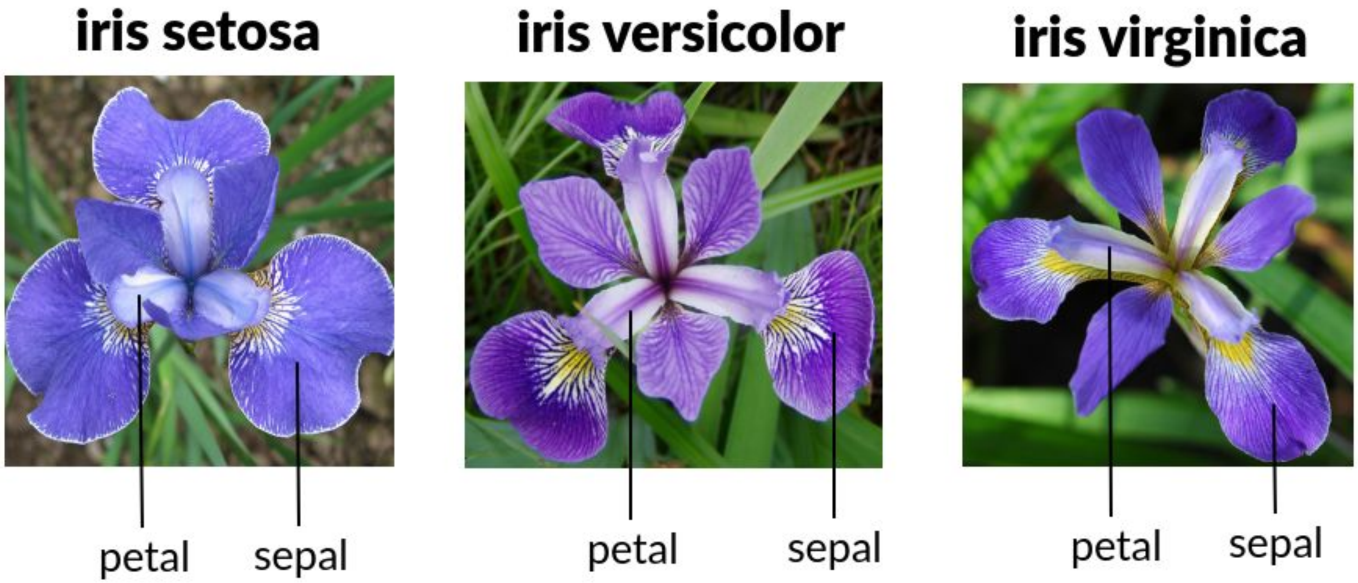
Example: In a study looking at how exercise impacts weight loss, the weight loss is the dependent variable. It’s called “dependent” because its value depends on the amount of exercise the subject does (which would be the independent variable in this case).

Variables
Independent variable

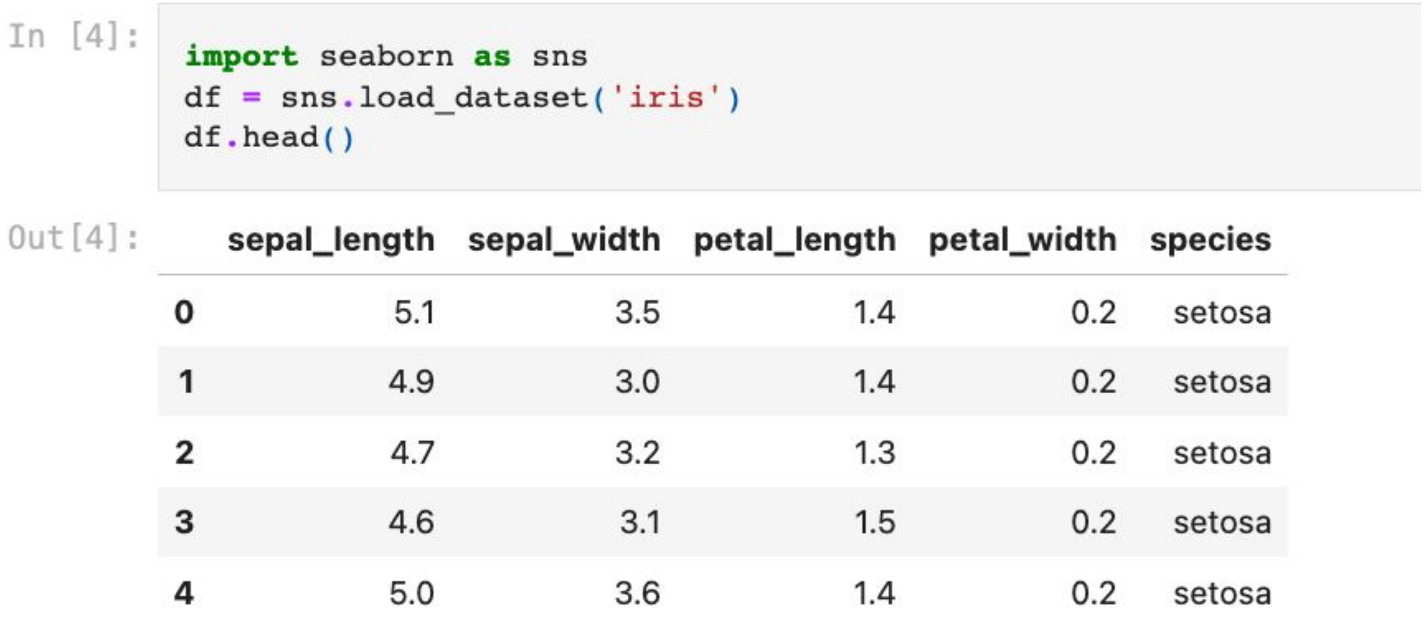
A variable that you control or manipulate in an experiment to see its impact on the dependent variable. It’s called “independent” because its variation does not depend on other variables in your experiment.

Example: In a study looking at the effect of temperature on the growth rate of bacteria, the temperature would be the independent variable because you can set it to different levels to see how the bacteria’s growth rate changes.

Parts of a Machine Learning model



Parts of a Machine Learning model



In [4]:

```
import seaborn as sns
df = sns.load_dataset('iris')
df.head()
```

Out[4]:

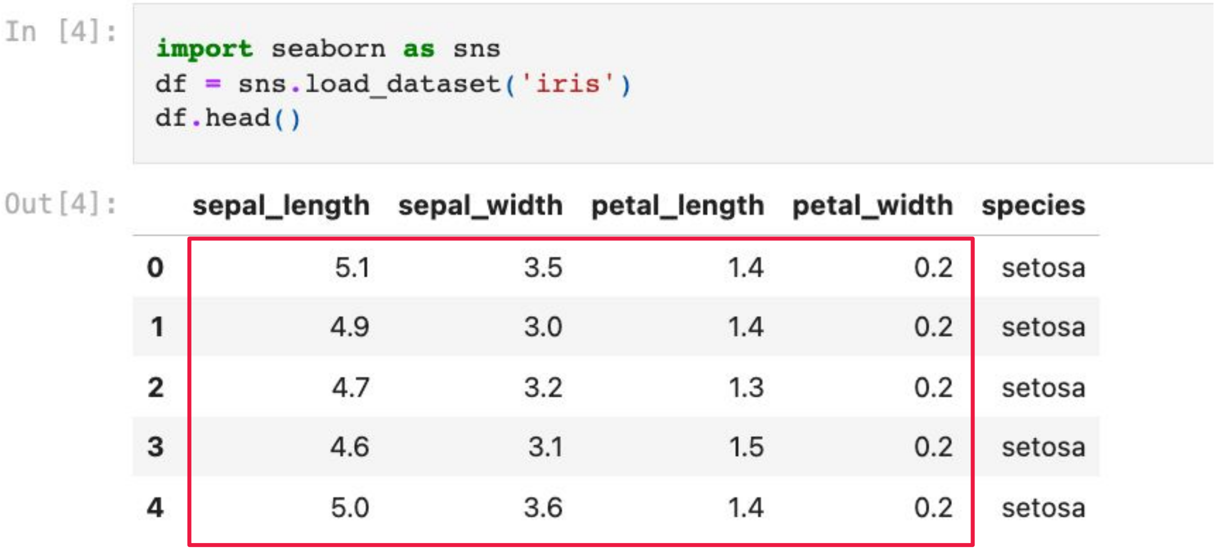
	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Parts of a Machine Learning model

Model Inputs

Also known as:

- Features
- Attributes
- Predictors
- Inputs
- **Independent Variables**
- Dimensions
- X
- Probably more…



In [4]:

```python
import seaborn as sns
df = sns.load_dataset('iris')
df.head()
```

Out[4]:

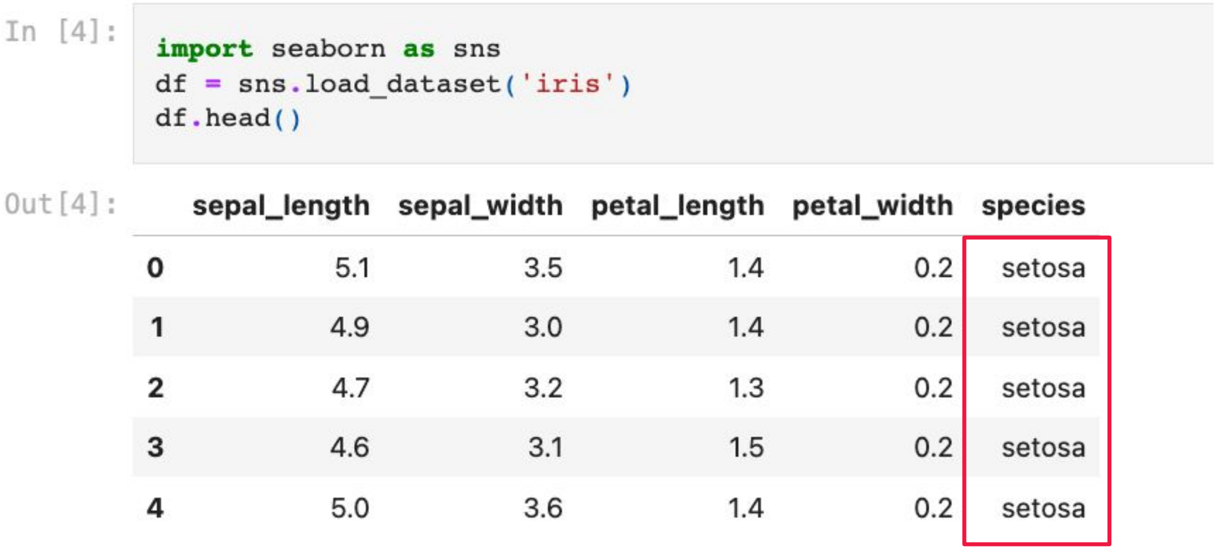
	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Parts of a Machine Learning model

Model Outputs (What you're trying to predict)

Also known as:

- Target
- Response
- Output
- **Dependent Variable**
- Labels
- Y
- Probably more…



In [4]:

```
import seaborn as sns
df = sns.load_dataset('iris')
df.head()
```

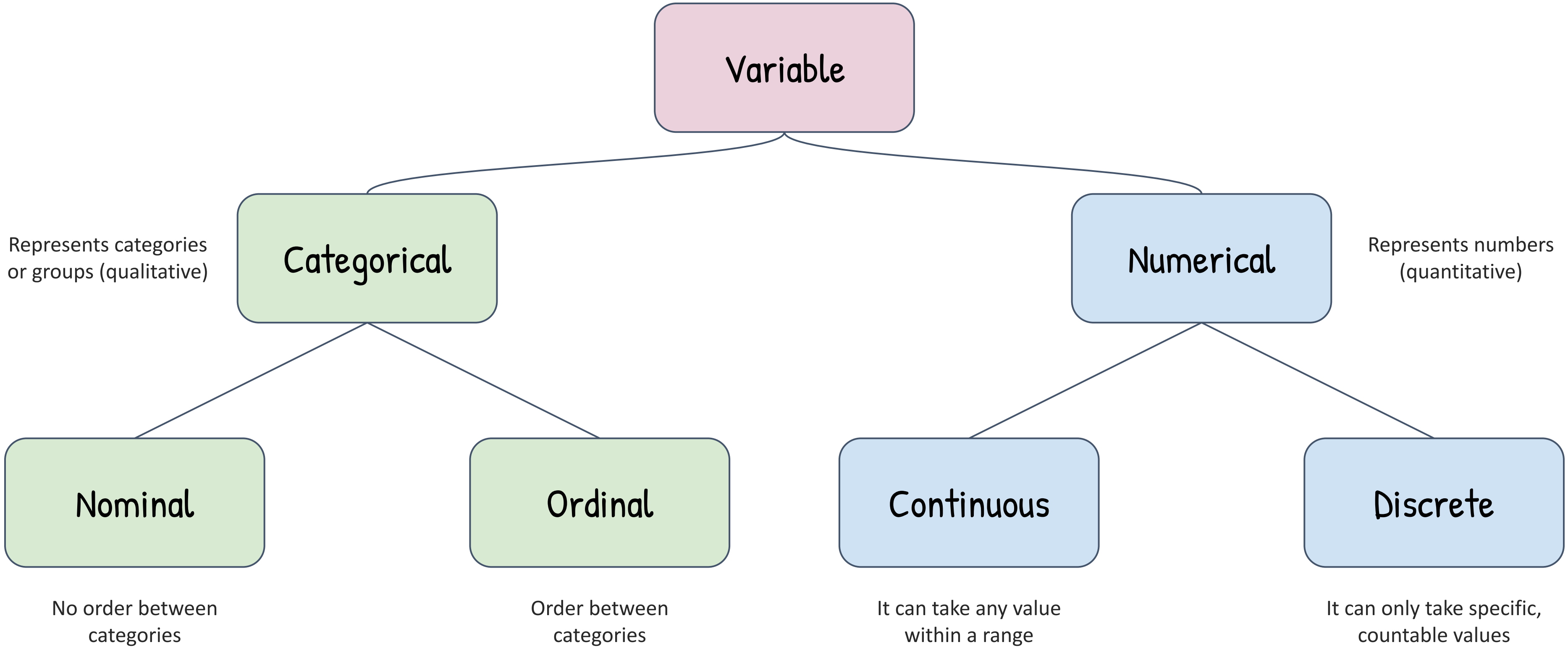
Out[4]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

A variable based on its type
can be categorical or numerical

A variable based on its nature
can be deterministic or random

A variable based on its type can be categorical or numerical



Variable

Categorical

Represents categories or groups (qualitative)

Numerical

Represents numbers (quantitative)

Nominal

No order between categories

Ordinal

Order between categories

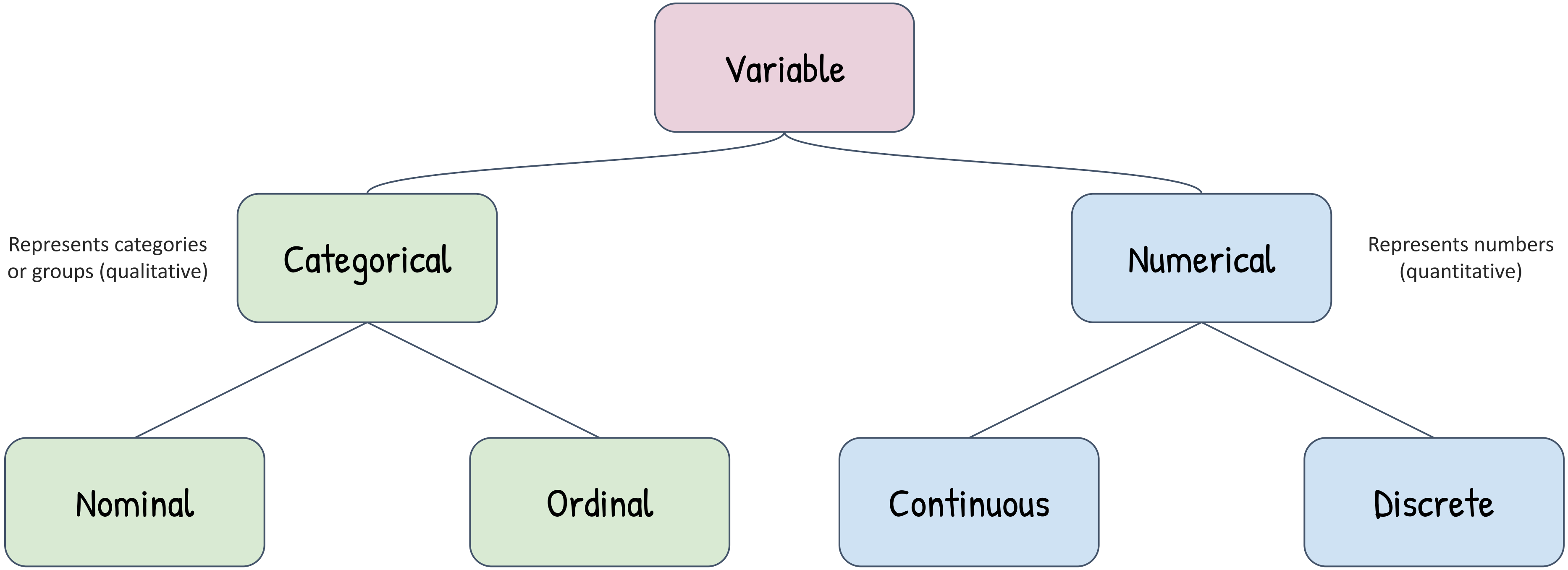
Continuous

It can take any value within a range

Discrete

It can only take specific, countable values

A variable based on its type can be categorical or numerical



Variable

Categorical

Represents categories or groups (qualitative)

Numerical

Represents numbers (quantitative)

Nominal	Ordinal	Continuous	Discrete
No order between categories	Order between categories	It can take any value within a range	It can only take specific, countable values
Example • Eye color (Blue/Green/Brown)	**Example** • Education level (Bachelor’s/Master’s/PhD)	**Example** • Weight in pounds	**Example** • Number of cows a farmer owns
As a target variable, in ML It’s a **Classification** model	As a target variable, in ML It’s a **Regression or Classification** model	As a target variable, in ML It’s a **Regression** model	As a target variable, in ML It’s a **Regression or Classification** model

A variable based on its nature can be deterministic or random

Variables
Deterministic variable

Takes a specific, fixed value

- **Example:**

  In `y = 2x + 3`, if `x=5`, then `y` is deterministically `13`

A variable based on its nature can be deterministic or random

Variables
Random variable

A random variable is a variable whose possible values are numerical outcomes of a random phenomenon. It’s associated with a **probability distribution**

- **Discrete random variable**
  - One that has countable number of possible outcomes
  - **Example:** Rolling a die- Any numbers 1 through 6, but nothing in between

- **Continuous random variable**
  - One that can take on any value in a given range
  - **Example:** Amount of time it takes for a computer to boot up

Descriptive statistics

Method to summarize and describe the main features of a dataset

Measures of central tendency

- **Mean**
  Arithmetic average of the data.
  Calculated by adding all the values and dividing them by the number of values
- **Median**
  Middle value when the data points are arranged in order
- **Mode**
  The most frequently occurring value in the dataset

Measures of variability (dispersion)

- **Range**
  The difference between the maximum and minimum values
- **Variance**
  A measure of how much the data points differ from the mean on average
- **Standard deviation**
  Square root of the variance, giving a sense of how much data points typically deviate from the mean in the same units as the data

Correlation

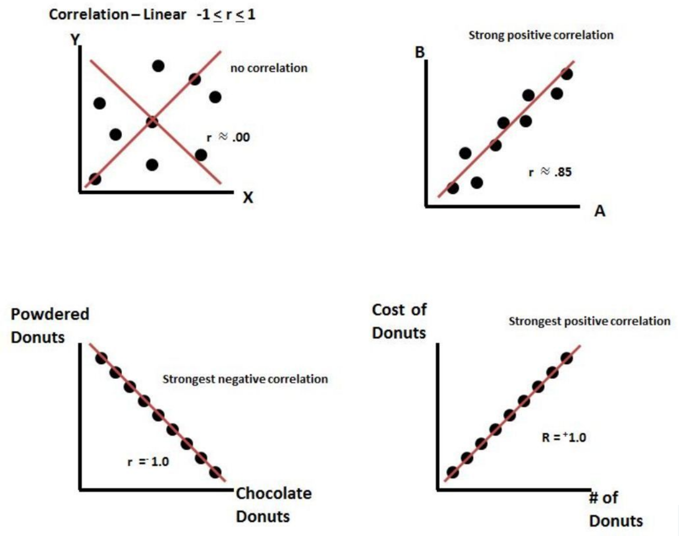
Correlation is a statistical measure that quantifies the **strength** and **direction** of the linear relationship between two variables. It ranges from -1 to 1, where:

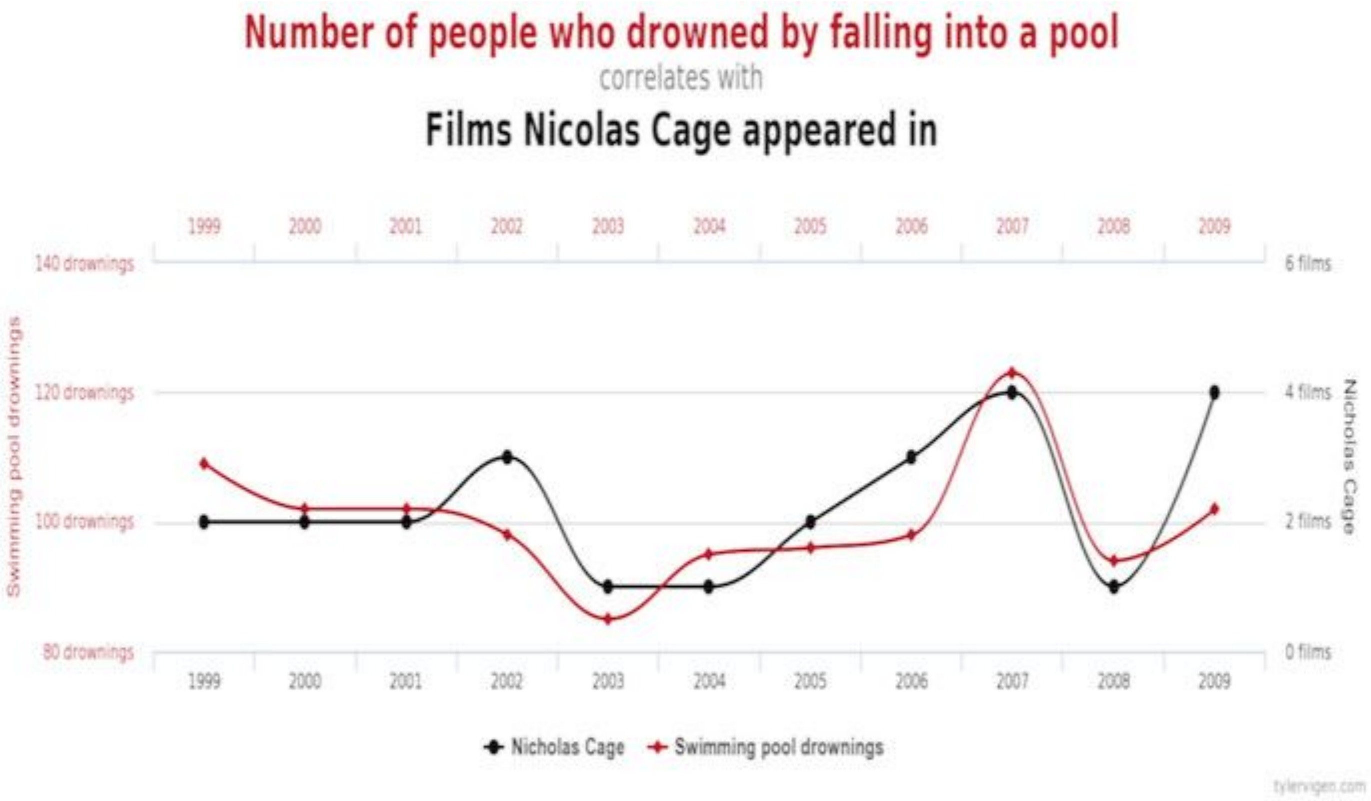
- A value of 1 indicates a perfect **positive correlation**, meaning that as one variable increases, the other variable increases proportionally.

- A value of -1 indicates a perfect **negative correlation**, implying that as one variable increases, the other variable decreases proportionally.

- A value of 0 indicates **no linear correlation** between the variables.

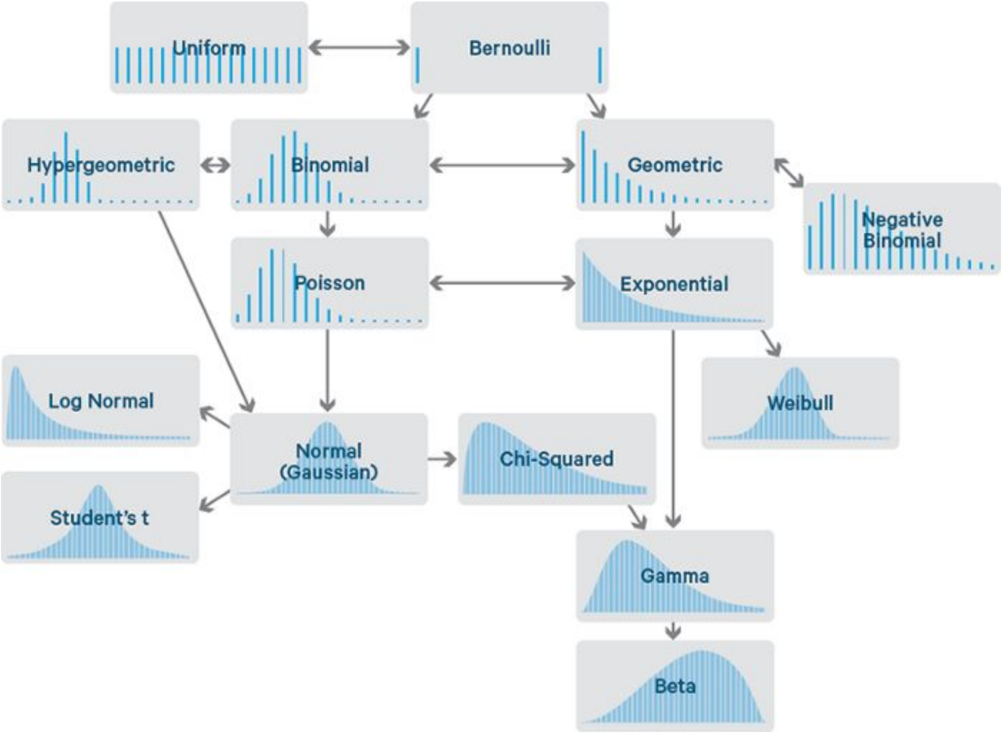
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$



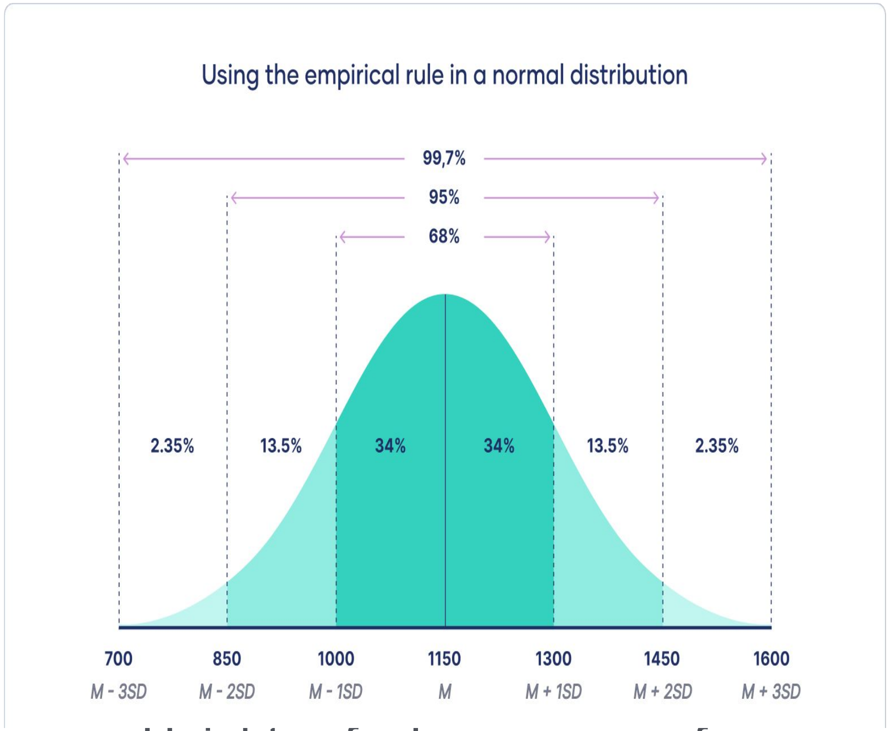

Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
140 drownings
120 drownings
100 drownings
80 drownings
Swimming pool drownings
6 films
4 films
2 films
0 films
Nicholas Cage
Nicholas Cage
Swimming pool drownings
tylervigen.com

Probability distributions

- Mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment



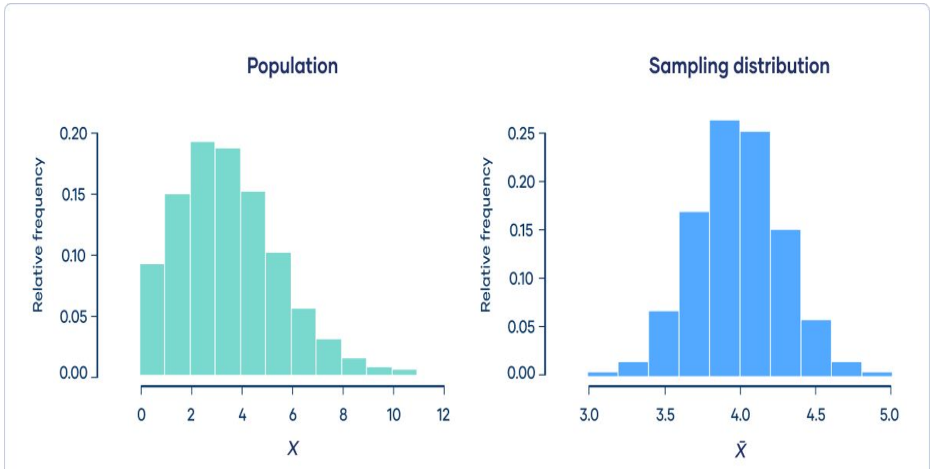
Normal (Gaussian) distribution



- Heights of a large group of people
- Measurement of errors in scientific experiments

- Mean, Median, and Mode are all equal.
- Approximately 68% of values fall within 1 standard deviation of the mean, about 95% within 2 standard deviations, and about 99.7% within 3 standard deviations. This is known as the Empirical Rule.
- It's determined by two parameters: the mean (μ), which determines the center of the distribution, and the standard deviation (σ), which determines the spread.

Central Limit Theorem



- It states that no matter what the distribution of the sample is if you sample batches of data from that distribution and take the mean of each batch then the mean values that we got from all those batches will be normally distributed.

- Example: imagine you are measuring the height of people in a town. The distribution may not be perfectly normal (perhaps it's slightly skewed because there are more adults than children living in the town), but if you take multiple samples of people and calculate the mean height of each sample, the distribution of those means would look approximately normal due to CLT

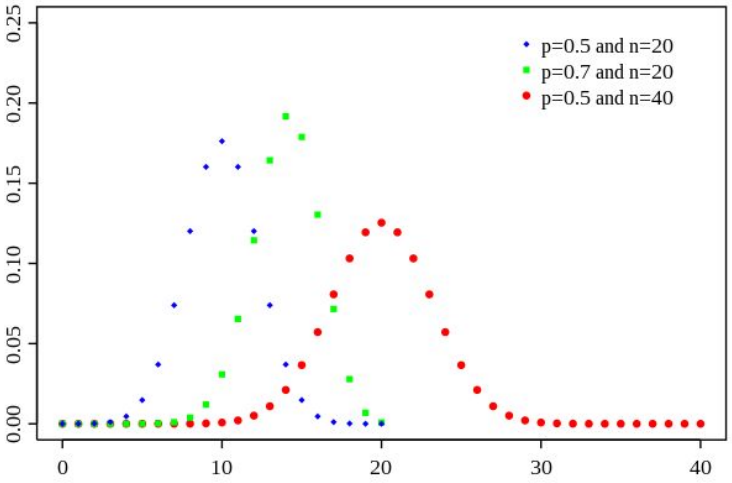
- Why is it important? In practice, many data sets are not normally distributed. However, the CLT allows us to assume normality for large enough sample sizes. This makes statistical techniques that rely on the assumption of normality more applicable.

Code

https://colab.research.google.com/drive/1SCrBt5-c8ShJbfuvhLYQeyCQ59HjWj1n?usp=sharing

QUESTIONS?

Binomial distribution



Examples:

- Coin flip

- Discrete probability distribution
- There are a fixed number of trials (n).
- Each trial is independent of the others.
- Each trial has only two possible outcomes: success or failure.
- The probability of success (p) is the same for each trial.