

2025

4Geeks Academy: data science cohort 12

# DAY 10: PROBABILITY

# TODO

**PROBABILITY** Intro to frequentist and Bayesian probability in data science

**PROBABILITY  
PROJECT** Work on Probability Exercises Project in Python (Probability module) plan to finish before class Friday

# TOPICS

01 PROBABILITY IN DATA SCIENCE

02 FREQUENTIST PROBABILITY

03 BAYESIAN PROBABILITY

# PROBABILITY IN DATA SCIENCE

**WHAT** Part of statistics dealing with how likely events are to occur

- WHY**
- **Building models**
    - Does applying data transformation X increase the performance of my model?
    - Is model type A better than model type B for some problem?
  - **Interpreting models**
    - An email is given a 89.3% chance of being a scam
    - An investment is predicted yield a return of 4.2 +/- 0.3%
  - **Evaluating models**
    - The model scores 6.5 RMSE on test data

- HOW**
- **Statsmodels:** module with statistical models & tests
  - **Scipy.stats:** module with probability distributions, summary statistics, statistical functions and tests
  - **Scikit-learn:** library for statistical modeling and machine learning

# FREQUENTIST PROBABILITY

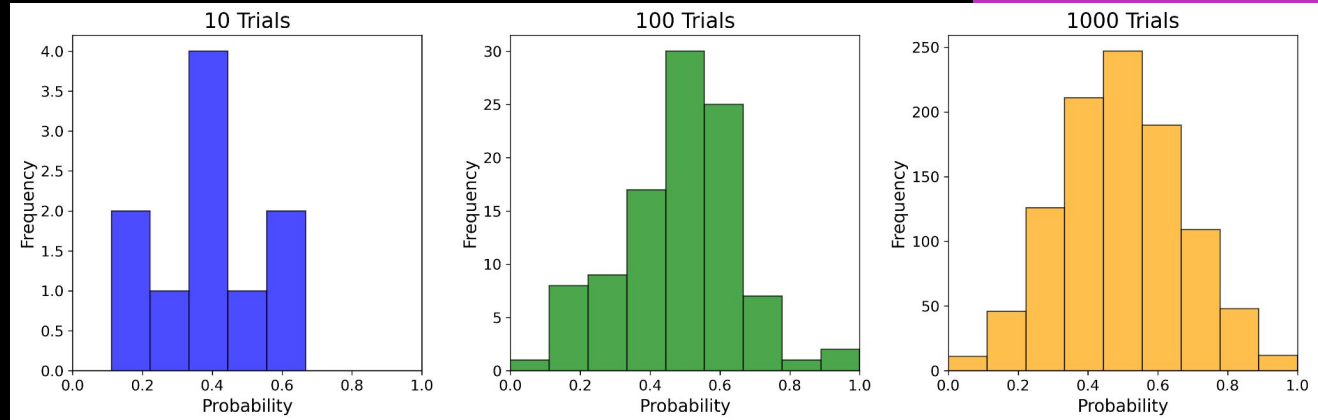
**PROBABILITY** Frequency of events in an infinite number of trials

**EMPIRICAL  
PROBABILITY** Observed frequency of event. Ex - flip a coin 10 times:

$$P_{heads} = \frac{heads}{flips} = \frac{6}{10} = 0.6$$

Wait, that can't be right...?!

**SAMPLING**



# BAYESIAN PROBABILITY

**WHAT** Treats probability as an expectation given a state of knowledge

**WHY** Allows for the inclusion of prior knowledge and updating of probability based on new knowledge (**Bayesian inference**)

**HOW** Bayes' theorem:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

## EXAMPLE

A model is trained to detect spam on a dataset of 100,000 emails, 1000 of which are spam. Of flagged emails, 95% are actually spam. It also incorrectly identifies 1% of non-spam emails as spam. If the model flags a new email as spam, what is the probability that it really is spam?

There is only a 2.4% chance the new email is spam!

$$P(flagged|spam) = \frac{P(spam|flagged) \times P(spam)}{P(flagged)}$$

$$P(flagged|spam) = \frac{0.95 \times 0.050}{0.019} = 0.024$$