#Why Probability & Statistics Matter in Data Science?

- Understanding data distributions, uncertainty, and inference is key to building ML models that generalize well.
- Statistical methods help in hypothesis testing, evaluating model performance, and making data-driven decisions.

#Probability Distributions:

- **Normal Distribution (Gaussian):** Most natural phenomena follow this bell-shaped curve. Ex: Heights of people
  - Formula: $f(x) = (1 / (\sigma\sqrt{(2\pi)})) * \exp(- (x - \mu)^2 / (2\sigma^2))$
- **Uniform Distribution:** All outcomes equally likely. Ex: Rolling a fair die → P(1) = P(2) = ... = P(6) = 1/6
- **Binomial Distribution**: Probability of success in a series of trials. Ex: Flipping a coin 10 times → P(5 heads)
- **Poisson Distribution**: Probability of a number of events in a fixed interval (time/space).
  Ex: Number of customer arrivals per hour.

#Z-Score (Standardization):

- Formula: $Z = (X - \mu) / \sigma$

Where: X = Data point, μ = Mean of dataset, σ = Standard deviation

- Intuition: Measures how many standard deviations a data point is from the mean.
- Example: Dataset: [60, 70, 80, 90, 100] → μ = 80, σ ≈ 15.81 & Z-score of 90 → (90 - 80) / 15.81 ≈ 0.63

#Conditional Probability:

- P(A | B): Probability of event A given B happened.
- Formula: $P(A \mid B) = P(A \cap B) / P(B)$
- Example: From a deck of 52 cards, P(Queen | Face Card) = 4/12 = 1/3

Example 1.2.1 (Flights and rain). JFK airport hires you to estimate how the punctuality of flight arrivals is affected by the weather. You begin by defining a probability space for which the sample space is

$$\Omega = \{\text{late and rain}, \text{late and no rain}, \text{on time and rain}, \text{on time and no rain}\} \tag{1.21}$$

and the σ-algebra is the power set of Ω. From data of past flights you determine that a reasonable estimate for the probability measure of the probability space is

$$\mathrm{P}(\text{late, no rain}) = \frac{2}{20}, \qquad \mathrm{P}(\text{on time, no rain}) = \frac{14}{20}, \tag{1.22}$$

$$\mathrm{P}(\text{late, rain}) = \frac{3}{20}, \qquad \mathrm{P}(\text{on time, rain}) = \frac{1}{20}. \tag{1.23}$$

The airport is interested in the probability of a flight being late if it rains, so you define a new probability space conditioning on the event *rain*. The sample space is the set of all outcomes such that *rain* occurred, the σ-algebra is the power set of {on time, late} and the probability measure is $\mathrm{P}(\cdot|\text{rain})$. In particular,

$$\mathrm{P}(\text{late}|\text{rain}) = \frac{\mathrm{P}(\text{late, rain})}{\mathrm{P}(\text{rain})} = \frac{3/20}{3/20 + 1/20} = \frac{3}{4} \tag{1.24}$$

and similarly $\mathrm{P}(\text{late}|\text{no rain}) = 1/8$.

#Bayes' Theorem

- Formula: P(A | B) = [P(B | A) * P(A)] / P(B)
- **Ex:** A person has undertaken a job. The probability of completing the job on time if it rains is 0.44, and the probability of completing the job on time if it does not rain is 0.95. If the probability that it will rain is 0.45, then determine the probability that the job will be completed on time.

Let: R: event that it rains, Rc: event that it does not rain, C: event that the job is completed on time

We are given:

P(R)=0.45, P(Rc)=1−0.45=0.55

P(C|R)=0.44, P(C|Rc)=0.95

By the law of total probability:

P(C)=P(R)P(C|R) + P(Rc)P(C|Rc)

Substitute values:

P(C)= (0.45) (0.44) + (0.55) (0.95)

P(C)=0.198+0.5225=0.7205

Example 2: There are three urns containing 3 white and 2 black balls, 2 white and 3 black balls, and 1 black and 4 white balls, respectively. There is an equal probability of each urn being chosen. One ball is equal probability chosen at random. What is the probability that a white ball will be drawn?

Solution:

Let E_1, E_2, and E_3 be the events of choosing the first, second, and third urn respectively. Then,
$P(E_1) = P(E_2) = P(E_3) = 1/3$

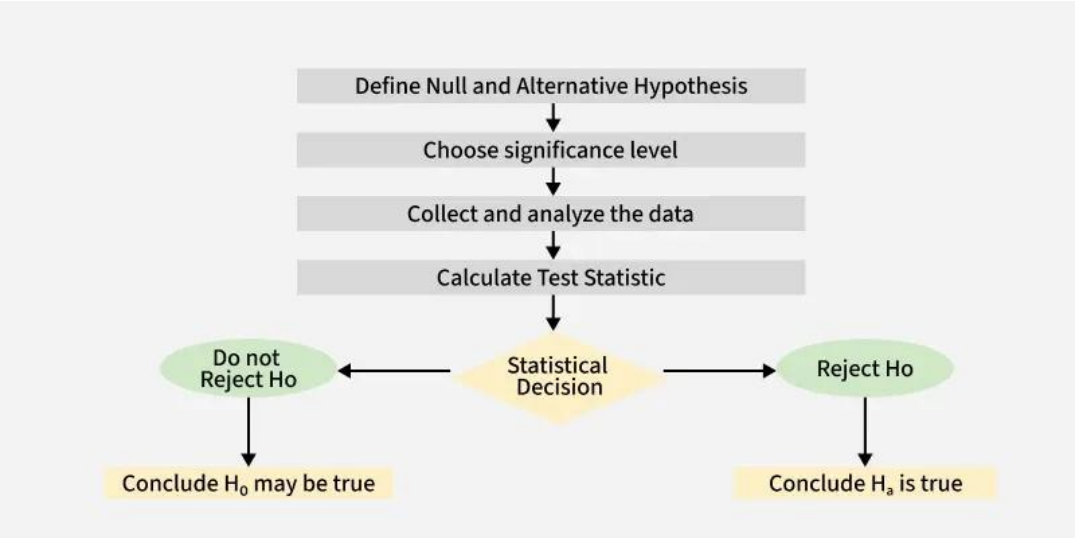
Let E be the event that a white ball is drawn. Then,
$P(E/E_1) = 3/5$, $P(E/E_2) = 2/5$, $P(E/E_3) = 4/5$

By theorem of total probability, we have
$P(E) = P(E/E_1) \cdot P(E_1) + P(E/E_2) \cdot P(E_2) + P(E/E_3) \cdot P(E_3)$
$\Rightarrow P(E) = (3/5 \times 1/3) + (2/5 \times 1/3) + (4/5 \times 1/3)$
$\Rightarrow P(E) = 9/15 = 3/5$

#Hypothesis Testing:

- Null Hypothesis (H_o): No effect or difference.
- Alternative Hypothesis (H_1): There is an effect or difference.
- P-value interpretation: $P < 0.05 \rightarrow$ Reject H_o & $P \geq 0.05 \rightarrow$ Fail to reject H_o

- Example: A/B Test on website CTR → Test if design B performs better than A.



The test statistic measures how much the sample data deviates from what we did expect if the null hypothesis were true. Different tests use different statistics:

- **Z-test:** Used when population variance is known and sample size is large.
- **T-test:** Used when sample size is small or population variance unknown.
- **Chi-square test:** Used for categorical data to compare observed vs. expected counts.

#Maximum Likelihood Estimation (MLE):

- Concept: Choose parameters θ that maximize P(Data | θ).
- Example: Estimating mean μ of normal distribution given sample data → μ = sample mean.

MLE is widely applied in various machine learning tasks and models:

- **Logistic Regression**: For classification tasks, MLE finds parameters that max the likelihood of observed outcomes.
- **Naive Bayes Classification**: MLE estimates the means and variances of conditional probabilities for classification.
- **Linear Regression**: When combined with the assumption of normally distributed errors, the least squares solution for linear regression is equivalent to the MLE.