Chap 4- Summary

- **Sample Space:** defining all possible outcomes / can be discrete → number of values is finite, or continuous → infinitely many values in an interval
- **Event:** subset of sample space / specific situation. If outcome is in a subset we say **event has occurred**
  - certain event → always occurs (Ω)
  - impossible event → never occurs ($\emptyset$)
- **Combining events:**
  - Union ($A \cup B$): event that either A **or** B **or** both occur
  - Intersection ($A \cap B$): event that both A **and** B occur
  - Complement (A^c): event that A does not occur
  - Difference ($A - B$): event that A occurs but B does not occur
- **De Morgan's Laws:**
  - $(A \cup B)^c = A^c \cap B^c$
  - $(A \cap B)^c = A^c \cup B^c$
- **Conditional Probability** of A given B:
  - $P(A|B) = \frac{P(A \cap B)}{P(B)}$, for $P(B) > 0$
- **Independence of events:** meaning knowing one gives us no information about the other
  - $P(A|B) = P(A)$
  - $P(B|A) = P(B)$
  - $P(A \cap B) = P(A)P(B)$
- **Bayes' Theorem:** B is true once A is observed
  - $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$ for $P(A) > 0$
  - with:
    - $P(B)$ prior probability (initial belief in event B before seeing any data)
    - $P(A|B)$ likelihood (probability of observed data A, given B being true)
    - $P(B|A)$ posterior probability (updated belief in event B after seeing data A)
- **Random Variable**: a function that assigns a real number to each outcome in the sample space (denoted by capital letters, e.g., X, Y)

- **Discrete** RV: finite or countably infinite values (e.g., number of heads in coin tosses)
- **Continuous** RV: uncountably infinite values (e.g., height, weight), described using **PDF**: $P(a \leq Z \leq b) = \int_a^b f_Z(z)\,dz$, where $f_Z(z)$ is the probability density function

- **Mean** (average): denoted by μ or $E(X)$, tells where the distribution of X tends to balance out
  - for discrete RV: $E(X) = \sum x_i P(X = x_i)$
  - for continuous RV: $E(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx$
- **Variance**: denoted by σ^2 or $Var(X)$, measures spread of distribution around the mean
  - $Var(X) = E[(X - \mu)^2]$ or $Var(X) = E(X^2) - (E(X))^2$
  - discrete RV: $Var(X) = \sum (x_i - \mu)^2 P(X = x_i)$
  - continuous RV: $Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x)\,dx$
  - high variance = unstable prediction
- **Standard Deviation**: $\sigma = \sqrt{Var(X)}$
  - Law of Large Numbers (LLN): states that more trials = less randomness = truer picture of reality
- **Probability Distributions**: how probabilities are distributed across the possible values of a random variable
  - **Discrete Distributions:** e.g., Bernoulli, Binomial, Poisson
  - **Continuous Distributions:** e.g., Uniform, Normal
- **Bernoulli Distribution:** denoted $X \sim \text{Bernoulli}(p)$, has 2 outcomes

  X random variable, $X = \begin{cases} 1 & \text{if success} \\ 0 & \text{if failure} \end{cases}$

  - **PMF:** $P(X = k|p) = p^k (1 - p)^{1-k}$ for $k = 0, 1$
  - **Mean**: $E(X) = p$
  - **Variance**: $Var(X) = p(1 - p)$
- **Binomial Distribution**: denoted $X \sim \text{Binomial}(n, p)$, represents the number of successes in n repeated Bernoulli trials with probability of success p
  - **PMF**: $P(X = k|n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$ for $k = 0, 1, 2, ..., n$
  - **Mean**: $E(X) = np$
  - **Variance**: $Var(X) = np(1 - p)$
- **Poisson Distribution:** denoted $X \sim \text{Poisson}(\lambda)$, models the number of events occurring in a fixed interval of time/space with known average rate λ

- **PMF**: $P(X = k|\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, 2, \ldots$
- **Mean**: $E(X) = \lambda$
- **Variance**: $Var(X) = \lambda$

- **Normal (Gaussian) Distribution:** denoted $X \sim \mathrm{N}(\mu, \sigma^2)$, models continuous data with a symmetric bell-shaped curve
  - **PDF**: $f_X(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for $-\infty < x < \infty$
  - **Mean**: $E(X) = \mu$
  - **Variance**: $Var(X) = \sigma^2$
- **Standard Normal Distribution:** special case with $\mu = 0$ and $\sigma^2 = 1$, denoted $Z \sim \mathrm{N}(0, 1)$
  - **PDF**: $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ for $-\infty < z < \infty$
  - Any normal random variable can be standardized to a standard normal variable using $Z = \frac{X-\mu}{\sigma}$ so $P(X \leq x) = P(Z \leq \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma})$
- In **scipy.stats**, you can use:
  - `scipy.stats.binom` for Binomial distribution
  - `scipy.stats.poisson` for Poisson distribution
  - `scipy.stats.norm` for Normal distribution
  - `scipy.stats.bernoulli` for Bernoulli distribution
  - Methods:
    - `.pmf(k)` for PMF of discrete distributions
    - `.pdf(x)` for PDF of continuous distributions
    - `.cdf(x)` for CDF of both discrete and continuous distributions
    - `.rvs(size=n)` to generate `n` random samples from the distribution