1. Events & Probability

1.1 Axioms of Probability

- Sample Space (Ω): The sample space Ω represents all the possible outcomes of a random experiment. For example, if we roll a die, $\Omega=\{1,2,3,4,5,6\}$
- Event (E): An event is any subset of the sample space. It could consist of one or more outcomes. For example, in the same die-rolling experiment, if we are interested in rolling an even number, the event is $E=\{2,4,6\}$.
- Probability Function (Pr): The probability function assigns a probability to each event. It must satisfy these conditions (axioms):
    - Axiom 1: $0\leq Pr(E)\leq 1$ for any event E. This means probabilities are always between 0 and 1.
    - Axiom 2: $Pr(\Omega)=1$. The probability that something in the sample space will occur is always 1.
    - Axiom 3: For mutually exclusive events $E_1, E_2, \ldots$, where no two events can occur simultaneously (i.e., $E_i \cap E_j = \emptyset$ for $i = j$), the probability of their union is the sum of their individual probabilities:

$$Pr\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} Pr(E_i)$$

This is called the countable additivity property.

1.2 Union and Intersection of Events

- Union of Events (Addition Rule): The union of two events A and B is the event that either A or B occurs (or both). The probability of the union is given by:

$$Pr(A\cup B)=Pr(A)+Pr(B)-Pr(A\cap B)$$

Here, $Pr(A \cup B)$ represents the probability that either event A, event B, or both occur. $Pr(A \cap B)$ is the probability that both events occur simultaneously. The reason we subtract $Pr(A \cap B)$ is that it gets counted twice when we add $Pr(A)$ and $Pr(B)$, so we need to correct for this double-counting.

1.3 Union Bound

- The union bound provides an upper bound for the probability of the union of multiple events:

$$Pr\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} Pr(E_i)$$

This inequality tells us that the probability of the union of several events cannot be greater than the sum of their individual probabilities. It's a useful approximation when we don't have exact probabilities for intersections of events.

1.4 Inclusion-Exclusion Principle

- The inclusion-exclusion principle is used to calculate the exact probability of the union of multiple events by accounting for over-counted intersections:

$$Pr(A \cup B \cup C) = Pr(A) + Pr(B) + Pr(C) - Pr(A \cap B) - Pr(B \cap C) - Pr(A \cap C) + Pr(A \cap B \cap C)$$

This principle becomes more complicated as more events are involved, but it's a precise way to compute probabilities for unions. It balances adding the probabilities of individual events and subtracting the intersections (which were counted too many times) before adding back the intersection of all events.

1.5 Conditional Probability

- Definition: Conditional probability gives the probability of one event happening given that another event has already occurred. The conditional probability of event A given that event B has occurred is:

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}, \ provided\ Pr(B) > 0$$

Here, $Pr(A|B)$ means "the probability of A, given that B has occurred." The numerator $Pr(A \cap B)$ is the probability that both A and B happen, and the denominator $Pr(B)$ ensures that we are only considering cases where B happens.

1.6 Multiplication Rule

- The multiplication rule is a way to find the probability of the intersection of two events:

$$Pr(A \cap B) = Pr(A) \cdot Pr(B|A)$$

This formula shows that the probability of both A and B occurring is the probability that A happens, multiplied by the probability that B happens given that A already occurred. This rule can be extended to more events, multiplying conditional probabilities in a chain.

1.7 Independence of Events

- Definition of Independence: Two events A and B are said to be independent if the occurrence of one does not affect the probability of the other. Mathematically:

$$Pr(A \cap B) = Pr(A) \cdot Pr(B)$$

Independence implies that knowing whether one event happened does not change the likelihood of the other event happening. For instance, flipping a coin and rolling a die are independent events because the outcome of one does not influence the outcome of the other.

1.8 Law of Total Probability

- The law of total probability helps break down a complex probability into simpler components when we have a partition of the sample space. If $E_1, E_2, \ldots, E_n$ are mutually exclusive events that form a partition of the sample space, then for any event B:

$$Pr(B) = \sum_{i=1}^{n} Pr(B|E_i) * Pr(E_i)$$

- This theorem allows us to compute the probability of B by considering how B behaves under each scenario (partition event E_i) and weighing these scenarios by the likelihood of each scenario occurring (i.e., $Pr(E_i)$).

1.9 Bayes' Theorem

- Formula: Bayes' theorem allows us to reverse conditional probabilities:

$$Pr(E_j|B) = \frac{Pr(B|E_j) * Pr(E_j)}{\sum_{i=1}^{k} Pr(B|E_j) * Pr(E_j)}$$

- This theorem is particularly useful when we want to update our beliefs about an event E_j based on new evidence B. For example, in medical testing, Bayes' theorem helps calculate the probability that a patient has a disease (event E_j) given a positive test result (event B).

---

2. Discrete Random Variables

2.1 Random Variables

- A random variable (often denoted as X) is a function that assigns a real number to each outcome of a random experiment. Formally, for a sample space Ω, a random variable is a function X:$\Omega \rightarrow$R.
  - Example: Suppose we roll two dice. Let X be the sum of the dice. Then $X(2,5) = 7$ because the outcome of rolling 2 and 5 gives a sum of 7.

- Discrete Random Variable: A random variable is discrete if it takes on a countable number of possible values, such as the integers. For instance, the number of heads in 10 coin flips is a discrete random variable because the possible values are finite: 0, 1, 2, ..., 10.

2.2 Probability Mass Function (PMF)

- The probability mass function (PMF) $p_X(x)$ for a discrete random variable X is the function that gives the probability of each possible value of X. Formally:

$$(x)=Pr(X=x)$$

- This means that $p_X(x)$ is the probability that the random variable X takes the value x. The PMF must satisfy:
  - $p_X(x) \geq 0$ for all x
  - $\sum_{x} px(x) = 1$

2.3 Cumulative Distribution Function (CDF)

- The cumulative distribution function (CDF) $F_X(x)$ of a discrete random variable X gives the probability that X will take a value less than or equal to x:

$$F_X(x) = Pr(X \leq x)$$

The CDF is essentially a running total of the probabilities up to a certain value. It is defined as:

$$Fx(x) = \sum_{y \leq x} px(y)$$

The CDF is a non-decreasing function, and as $x \longrightarrow \infty$, $F_X(x) \longrightarrow 1$ because the total probability of all possible outcomes sums to 1.

2.4 Joint Probability Function

- For two discrete random variables X and Y, the joint probability mass function (PMF) $p_{X,Y}(x,y)$ gives the probability that $X=x$ and $Y=y$ simultaneously:

$$p_{X,Y}(x,y) = Pr(X = x \text{ and } Y = y)$$

Marginal PMF: The marginal PMF for X is obtained by summing over the possible values of Y:

$$px(x) = \sum_{y} px, y(x, y)$$

This follows from the law of total probability, as it considers all possible ways the value $X=x$ can occur by summing over all possible Y values.

2.5 Independence of Random Variables

- Two random variables X and Y are independent if the joint probability function factorizes into the product of the individual (marginal) probabilities:

$$p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$$

This means that knowing the value of X gives no information about Y and vice versa. Independence in random variables is analogous to independence in events.

2.6 Expectation of a Discrete Random Variable

- The expectation (or expected value) of a discrete random variable X is the average or mean value it would take if the experiment were repeated infinitely many times. It is calculated by weighting each possible value of X by its probability:

$$E(X) = \sum_{x} x * px(x)$$

- Example: For a fair die, the expectation of the outcome is: $E(X) = \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6) = 3.5$
- The expectation represents the "long-run average" outcome of rolling a die repeatedly.

2.7 Linearity of Expectation

- Linearity of Expectation states that the expectation of a sum of random variables is equal to the sum of their expectations, regardless of whether the variables are independent:

$$E(X+Y)=E(X)+E(Y)$$

More generally, for any constants a and b:

$$E(aX+b)=aE(X)+b$$

This is a crucial property because it allows us to break down complex expressions into simpler parts when calculating expectations.

2.8 Jensen's Inequality

- Jensen's Inequality applies to convex functions. A function $f(x)$ is convex if for any x_1,x_2 and $\lambda \in [0,1]$, the following holds:

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

- Theorem: For a convex function f and any random variable X,

$$E[f(X)] \geq f(E[X])$$

This inequality tells us that the expectation of a convex transformation of a random variable is greater than or equal to the transformation of its expectation. It's particularly useful in optimization and risk management.

2.9 Binomial Random Variable

- A binomial random variable counts the number of successes in a fixed number of independent Bernoulli trials (where each trial has two possible outcomes: success or failure). The probability mass function for a binomial random variable X, with parameters n (number of trials) and p (probability of success), is:

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the number of ways to choose k successes out of n trials.

- Expectation and Variance:

$$E(X) = np, Var(X) = np(1-p)$$

These formulas give the average number of successes and the spread (variance) around that average.

2.10 Conditional Expectation

- Conditional Expectation is the expected value of a random variable given that some condition holds. If Y is a random variable and X is another random variable or event, the conditional expectation of Y given $X=x$ is:

$$E(Y|X=x) = \sum_{y} y \cdot p_{Y|X}(y|x)$$

  - Law of Total Expectation: This law states that the total expectation of a random variable Y can be computed by taking the conditional expectation given X and then averaging over the possible values of X:

$$E(Y) = E[E(Y \mid X)]$$

2.11 Geometric Distribution

- A geometric random variable counts the number of trials until the first success in a sequence of independent Bernoulli trials, each with success probability p. The probability mass function for a geometric random variable X is:

$$p_X(k) = (1-p)^{k-1}p, \quad k = 1, 2, 3, \ldots$$

  - Expectation: The expected number of trials until the first success is:

$$E(x) \;\; = \frac{1}{p}$$

- The geometric distribution is memoryless, meaning that the probability of success on the k-th trial is independent of what happened before.

2.12 Poisson Distribution

- A Poisson random variable models the number of events that occur in a fixed interval of time or space, where events happen at a constant average rate and independently of each other. The PMF of a Poisson random variable X with rate parameter λ (the average number of events) is:

$$p_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

  - Expectation and Variance:

$$E(X) = \lambda, Var(X) = \lambda$$

Poisson distributions are often used to model rare events, such as the number of emails you receive per hour or the number of accidents at an intersection in a day.

---

3. Moments & Deviations

3.1 Markov's Inequality

- Markov's Inequality gives an upper bound for the probability that a non-negative random variable X takes a value greater than or equal to a constant a:

$$Pr(X \geq a) \leq \frac{E(X)}{a}$$

This inequality is useful when we only know the expectation of X and need a rough estimate of how likely X is to exceed a certain threshold. For instance, if $E(X)=10$, then $Pr(X \geq 20) \leq 10/20 = 0.5$

3.2 Variance and Moments

- Variance measures the spread or variability of a random variable around its mean. It is defined as:

$$Var(X) = E[(X - E(X))^2] = E(X^2) - (E(X))^2$$

The square root of the variance is the standard deviation, which gives a measure of dispersion in the same units as the original variable.

- Moments are expectations of powers of the random variable. The k-th moment of a random variable X is: $E(X^k)$
- Moments provide detailed information about the shape of the distribution of X.

3.3 Covariance

- Covariance measures the degree to which two random variables change together. For two random variables X and Y, the covariance is:

$$Cov(X,Y)=E[(X-E(X))(Y-E(Y))]$$

  - If $Cov(X,Y) > 0$, then X and Y tend to increase together.
  - If $Cov(X,Y) < 0$, then when X increases, Y tends to decrease.
  - If $Cov(X,Y) = 0$, then X and Y are uncorrelated (although they may not be independent).

3.4 Chebyshev's Inequality

- Chebyshev's Inequality provides a bound on how much a random variable deviates from its mean. Specifically, it states that for any random variable X with finite variance:

$$Pr(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}$$

- where σ is the standard deviation of X and k is a positive constant. This inequality is useful because it applies to any distribution, not just normal distributions.

3.5 Median and Mean

- The median of a random variable X, denoted $Md(X)$, is a value m such that:

$$Pr(X \leq m) \geq \frac{1}{2} \quad \text{and} \quad Pr(X \geq m) \geq \frac{1}{2}$$

  The median represents the "middle" value of the distribution, where half the probability mass is below and half is above.

- The mean $E(X)$ is the value that minimizes the expected squared distance:

$$E(X) = \arg\min_c E[(X - c)^2]$$

  Similarly, the median minimizes the expected absolute deviation:

$$Md(X) = \arg\min_c E[|X - c|]$$

3.6 Relationship Between Median and Mean

- For a random variable X with finite mean $\mu = E(X)$ and median $m = Md(X)$, we have the following relationship:

  $$|E(X) - Md(X)| \leq \sigma$$

  This inequality tells us that the difference between the mean and median is at most the standard deviation, highlighting how the two measures of central tendency relate to the spread of the data.