<u>Statistics Advance Part 1</u>

---

1. **What is a random variable in probability theory**
    - A random variable is defined as a set possible values obtained from a random experiment.
    - The experiment is random which means it’s outcomes will also be random.
    - It is denoted by Capital Letter “X”
    - It assigns numerical values to outcomes in a sample space.
    - There are two types: discrete (countable outcomes) and continuous (uncountable outcomes).

---

2. **What are the types of random variables**
    - **Discrete Random Variables**: Take specific, separate values (e.g., number of heads in coin tosses, dice rolled). It takes non decimal values
    - **Continuous Random Variables**: Take any value within a range (e.g., height, weight, temperature). It takes decimal values

---

3. **What is the difference between discrete and continuous distributions**
    - **Discrete distributions** deal with variables that have countable outcomes (e.g., 0, 1, 2).
    - **Continuous distributions** deal with variables that can take any value in a range or interval (e.g., 2.35, 3.67).
    - Discrete distributions use **probability mass functions**, while continuous use **probability density functions**.

---

4. **What are probability distribution functions (PDF)**
    - A PDF shows how probabilities are distributed over the values of a random variable. It also refers to that probability can be plot in graph
    - For **discrete variables**, it's called a **probability mass function (PMF)**.
    - For **continuous variables**, the PDF gives the **density**, and probabilities are found by integrating over intervals.

---

5. **How do cumulative distribution functions (CDF) differ from probability distribution functions (PDF)**
    - A **CDF** gives the probability that a random variable is less than or equal to a certain value.
    - A **PDF** (or PMF) gives the probability for a specific value (discrete) or density at a point (continuous).
    - CDF is the cumulative total of the PDF or PMF up to a given point.

---

6. **What is a discrete uniform distribution**
    - A distribution where all outcomes are equally likely.
    - Example: Rolling a fair die — each side (1 to 6) has a 1/6 chance.
    - It’s simple and used when all outcomes are known and equally probable.

7. **What are the key properties of a Bernoulli distribution**
    - Only two outcomes: success (1) or failure (0).
    - One trial with a fixed probability of success (p).
    - Mean = p; Variance = p(1 - p).
    - It's the simplest discrete distribution.

---

8. **What is the binomial distribution, and how is it used in probability**
    - It is series of n^{th} independent Bernoulli Trials.
    - Describes the number of successes in a fixed number of independent trials.
    - Each trial has two outcomes (like Bernoulli), with a constant probability of success.
    - Example: Flipping a coin 5 times and exactly getting 2 heads
      $P(X=3)= {}^{n}C_{k} . p^{k} .(1-p)^{n-k}$
      $P(X=3)={}^{5}C_{2} . p^{2} .(1-p)^{5-2} = 0.375$
    - It helps in modeling repeated experiments with the same probability.

---

9. **What is the Poisson distribution and where is it applied**
    - Models the number of times an event occurs in a fixed interval of time or space.
    - Used for rare events (e.g., number of calls to a helpdesk per hour).
    - Assume number of calls to a customer care center fluctuate throughout the day, rather than staying constant. The poisson distribution is useful here because it allows to calculate probability of a specific number of calls occurring within an hour. It uses average number of calls per hour(λ) lambda to predict these probabilities.
    - Mean = Variance = λ.

---

10. **What is a continuous uniform distribution**
    - All values within a certain interval are equally likely.
    - This applies to situations where the outcomes can take any value within a continuous range.
    - Within this defined range, every value has an equal change of being observed, this means probability will be uniform across height.
    - Example: Probability of getting an exact time within a given moment.
      Waiting time at a Bus Stop: Bus arrival is continuous and consistent. Bus is coming every 30 mins at a stop.

---

11. **What are the characteristics of a normal distribution**
    - Bell-shaped, symmetric curve centered at the mean.
    - Mean = Median = Mode.
    - Described by its mean (μ) and standard deviation (σ).
    - Many natural phenomena (e.g., heights, test scores) follow this pattern.

---

12. **What is the standard normal distribution, and why is it important**
    - A normal distribution with mean = 0 and standard deviation = 1.
    - Denoted as **Z-distribution**.

- Allows for comparison across different normal distributions.
- Used in Z-tests and calculating probabilities using Z-scores.

---

13. What is the Central Limit Theorem (CLT), and why is it critical in statistics

- CLT states that taking large number of samples form population, calculating sample mean($\bar{x}$) for each sample and plotting all sample mean in distribution forms a normal distribution.
- States that the sampling distribution of the sample mean becomes normal as the sample size increases (typically $n > 30$).
- Important because it allows us to use normal distribution for hypothesis testing and confidence intervals, even if the original data isn't normal.

---

14. How does the Central Limit Theorem relate to the normal distribution

- CLT explains why the normal distribution appears so often in statistics.
- Even if data is not normally distributed, the mean of large samples will follow a normal distribution.
- This is why normal-based methods are widely applicable.

---

15. What is the application of Z statistics in hypothesis testing

- Z-statistics, calculated using a Z-test, are applied in hypothesis testing to determine if there's a significant difference between a sample's mean and a population's mean, or between the means of two independent samples.
- This is particularly useful when the population variance is known, or when dealing with large sample sizes (typically $n \geq 30$).
- Used to calculate how far a sample statistic is from the population parameter, in standard deviation units.
- Helps determine whether to reject the null hypothesis.

---

16. How do you calculate a Z-score, and what does it represent

- **$Z = (X - \mu) / \sigma$**
  Where, x is the individual data point.
  μ(mu) is the population mean.
  σ(sigma) is the population standard deviation.
- It shows how many standard deviations a value (X) is from the mean (μ).
- A Z-score of 0 means the value is exactly the mean; positive or negative Z-scores show how far and in which direction the value is from the mean.

---

17. What are point estimates and interval estimates in statistics

- **Point estimate**: A single value used to estimate a population parameter (e.g., sample mean).
- **Interval estimate**: A range of values (like a confidence interval) used to estimate the population parameter with a certain level of confidence.
- Interval estimates give more information and reliability than point estimates.

---

18. What is the significance of confidence intervals in statistical analysis

- Confidence intervals show the range in which a population parameter likely falls.

- For example, a 95% confidence interval means we are 95% confident the true value lies within that range.
- It reflects both the estimate and the uncertainty.
- Example: if we want to know the salary of all IT employees then in reality it is not possible but using interval estimate it is possible. We take sample from population and from population we calculate the average salary that fall between this range(30k to 40k). This range(30k to 40k) is your interval estimate.

---

19. What is the relationship between a Z-score and a confidence interval

- Z-scores define the range of values in a confidence interval for a normal distribution.
- For example, a Z-score of ±1.96 corresponds to a 95% confidence level.
- Confidence intervals use Z-scores to determine their boundaries when standard deviation is known.

---

20. How are Z-scores used to compare different distributions

- Z-scores standardize values from different distributions for comparison.
- They allow you to compare scores on different scales by converting them into the same standard scale.
- Example: Comparing test scores from different subjects with different averages and variances.

---

21. What are the assumptions for applying the Central Limit Theorem

- Samples must be independent.
- Number of samples should be large.
- The sample size should be large (usually $n > 30$).
- The population should have a finite variance.
- If the population is normal, CLT applies even for small samples.

---

22. What is the concept of expected value in a probability distribution

- The expected value is the long-term average outcome of a random variable over many trials.
- Calculated as the sum of all possible values multiplied by their probabilities.
- It represents the “mean” of a random variable.

---

23. How does a probability distribution relate to the expected outcome of a random variable?

- A probability distribution shows all possible outcomes and their likelihoods.
- The expected outcome (expected value) is calculated using this distribution.
- It gives a summary of what to expect on average over many trials of the random process.
