Theoretical Questions: Statistics (Part 1)

---

1. What is a random variable in probability theory?

A random variable is a variable whose value is a numerical outcome of a random phenomenon. It assigns a numerical value to each possible outcome in a sample space.

Example: In a coin toss experiment, a random variable could be defined to be 1 if the outcome is "heads" and 0 if the outcome is "tails."

---

2. What are the types of random variables?

Random variables can be broadly classified into two main types:

- Discrete Random Variable: A variable that can take on a finite or countably infinite number of distinct values. Examples include the number of heads in a series of coin flips or the number of defective items in a batch.

- Continuous Random Variable: A variable that can take on any value within a given range or interval. Examples include the height of a person, the temperature of a room, or the time it takes to complete a task.

---

3. What is the difference between discrete and continuous distributions?

The primary difference lies in the nature of the random variable they describe:

- Discrete Distribution: Describes the probability of occurrence for each possible value of a discrete random variable. The probabilities are represented by a probability mass function (PMF), and the sum of all probabilities is equal to 1.

- Continuous Distribution: Describes the probabilities for the possible values of a continuous random variable. The probability is represented by a probability density function (PDF), where the probability of the variable falling within a specific range is the area under the curve of the PDF over that range. The total area under the curve is 1.

---

4. What are probability distribution functions (PDF)?

A Probability Density Function (PDF) is a function associated with a continuous random variable.

- Purpose: To describe the relative likelihood that the random variable will take on a particular value.
- Calculation: The probability of the variable falling within a specific interval is given by the integral (the area under the curve) of the PDF over that interval.
- Note: The value of the PDF at any single point is not the probability of that point, but rather its density; the probability of any single point for a continuous variable is zero.

---

5. How do cumulative distribution functions (CDF) differ from probability distribution functions (PDF)?

A CDF and a PDF provide different types of information about a probability distribution:

- PDF (Probability Density Function): Indicates the relative likelihood of a random variable taking a specific value. For a discrete variable, the equivalent is the PMF, which gives the probability of a specific outcome.

- CDF (Cumulative Distribution Function): Gives the probability that a random variable X will take a value less than or equal to a specific value x. It is defined as:

$F(x)=P(X\leq x)$

The CDF accumulates the probabilities up to the value x. For a continuous variable, the CDF is the integral of the PDF from negative infinity to x.

---

6. What is a discrete uniform distribution?

A discrete uniform distribution is a probability distribution where a finite number of values are equally likely to be observed. If a random variable can take any of n distinct values, then the probability of observing each value is $1/n$.

Example: The roll of a fair six-sided die, where each outcome (1, 2, 3, 4, 5, or 6) has a probability of 1/6.

---

7. What are the key properties of a Bernoulli distribution?

A Bernoulli distribution is a discrete probability distribution for a random variable which takes the value 1 with probability p and the value 0 with probability $1-p$.

Key Properties:

- Two Outcomes: Often termed "success" (1) and "failure" (0).
- Single Trial: It models the outcome of a single experiment.

- Probability Parameter: Defined by a single parameter, p.
- Expected Value: $E[X]=p$.
- Variance: $Var(X)=p(1-p)$.

---

8. What is the binomial distribution, and how is it used in probability?

The binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent Bernoulli trials. It is defined by two parameters:

- n: The total number of trials.
- p: The probability of success in a single trial.

The probability of getting exactly k successes in n trials is:

$P(X=k)=(kn)pk(1-p)n-k$

---

9. What is the Poisson distribution and where is it applied?

The Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate (λ) and independently of the time since the last event.

Applications:

- Number of phone calls received by a call center per hour.
- Number of emails received per day.
- Number of defects in a manufactured product.
- Number of accidents at an intersection in a week.

---

10. What is a continuous uniform distribution?

A distribution where all outcomes within a certain range are equally likely. It is defined by parameters a (minimum) and b (maximum). The probability density function is constant over the interval $[a,b]$ and zero everywhere else.

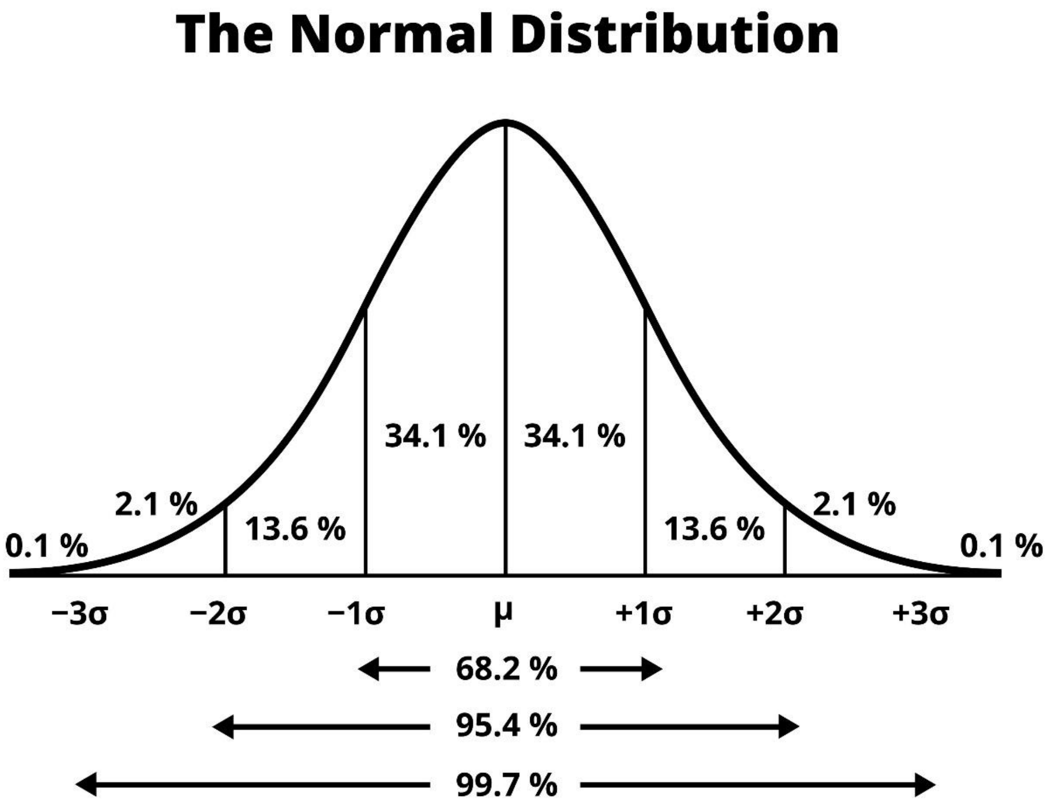
---

11. What are the characteristics of a normal distribution?

The normal (Gaussian) distribution is characterized by:

- Bell-Shaped Curve: Symmetric about its mean.
- Central Tendency: Mean, median, and mode are all equal and located at the center.

- Parameters: Defined by mean (μ) and standard deviation (σ).
- Empirical Rule: * ~68% of data within $\pm 1\sigma$
    - ~95% of data within $\pm 2\sigma$
    - ~99.7% of data within $\pm 3\sigma$
- Asymptotic: Tails extend indefinitely without touching the horizontal axis.



Shutterstock

---

12. What is the standard normal distribution, and why is it important?

The standard normal distribution is a special case where μ=0 and σ=1. It is important for:

- Standardization: Converting any normal distribution using the Z-score: $Z=(X-\mu)/\sigma$.
- Simplification: Allows comparison of scores from different distributions.
- Probability Calculation: Simplifies finding probabilities using a Z-table.

---

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

The CLT states that the distribution of sample means from a large number of samples will be approximately normally distributed, regardless of the population's distribution shape, provided the sample size is sufficiently large (usually $n \geq 30$).

Why it is critical:

- Foundation for Inference: Allows estimation of population means without knowing the population's shape.
- Hypothesis Testing: Basis for confidence intervals and many statistical tests.

---

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

The CLT establishes that the sampling distribution of the sample mean will tend toward a normal distribution as the sample size increases, even if the original population is not normal. This allows statisticians to use normal distribution properties for real-world data analysis.

---

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

Z-statistics are used to determine if there is a statistically significant difference between a sample mean and a hypothesized population mean. This is used when the population standard deviation is known and the sample size is large.

---

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

Formula:

$Z = \sigma(X - \mu)$

Representation:

- Positive Z: Value is above the mean.
- Negative Z: Value is below the mean.
- Zero Z: Value equals the mean.

---

17. What are point estimates and interval estimates?

- Point Estimate: A single value (e.g., sample mean $\bar{x}$) used to estimate a population parameter.
- Interval Estimate: A range of values (Confidence Interval) within which the parameter is expected to lie with a certain level of confidence.

18. What is the significance of confidence intervals?

- Quantify Uncertainty: A wider interval implies more uncertainty.
- Range of Plausible Values: Provides more context than a single number.
- Decision Making: If a hypothesized value falls outside the interval, it provides evidence against that hypothesis.

---

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

Z-scores act as the "critical value" defining the boundaries of the interval. Formula:

Confidence Interval=$\bar{x} \pm Z \times n \quad \sigma$

For a 95% confidence interval, the Z-score is approximately 1.96.

---

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

By converting values from different datasets into Z-scores, they are placed on a common scale (Standard Normal Distribution). This allows for relative comparison, such as comparing test scores from two different classes with different averages and spreads.

---

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

- Random Sampling: Samples must be drawn randomly.
- Independence: Observations must be independent (sample size < 10% of population).
- Sample Size: Sufficiently large (usually $n \geq 30$).

---

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

The expected value is the long-run average value of repetitions of an experiment. For a discrete random variable X:

$E[X]=i\sum xipi$

It represents the "mean" of the probability distribution.

---

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

The distribution provides the full picture of all possible outcomes and their likelihoods. The expected outcome condenses this information into a single number representing the central tendency or long-term average.