1. What are the most important topics in statistics?

* **Ans:** Descriptive Statistics (mean, median, mode, variance, standard deviation)
* Inferential Statistics (hypothesis testing, confidence intervals)
* Probability Distributions (normal, binomial, Poisson)
* Regression Analysis (linear, logistic)
* ANOVA (Analysis of Variance)
* Bayesian Statistics

1. What is exploratory data analysis?  
   **Ans**: Exploratory Data Analysis (EDA) is the initial phase of data analysis where researchers summarize key characteristics, detect patterns, identify outliers, and test assumptions using visual methods (box plots, histograms, scatter plots) and statistical summaries. EDA helps in understanding data structure, guiding further analysis, and selecting appropriate modeling techniques.
2. What are quantitative data and qualitative data?  
   **Ans**: Quantitative data consists of numerical values that can be measured (e.g., height, temperature, sales figures) and is categorized as discrete (countable) or continuous (infinite range). Qualitative data, on the other hand, is non-numerical and descriptive (e.g., colors, customer feedback, satisfaction levels), classified as nominal (no order) or ordinal (ordered categories). While quantitative data allows statistical testing, qualitative data provides contextual insights.
3. What is the meaning of KPI in statistics?  
   **Ans**: KPI (Key Performance Indicator) in statistics refers to measurable metrics used to evaluate performance, track progress toward objectives, and compare results over time. Examples include conversion rates, customer retention, error rates, and revenue growth. KPIs help organizations make data-driven decisions by quantifying success and identifying improvement areas.
4. What Is the Difference Between Univariate, Bivariate, and Multivariate Analysis?  
   **Ans:** Univariate analysis examines a single variable to understand its distribution (e.g., mean, histogram). Bivariate analysis explores the relationship between two variables (e.g., correlation, scatter plot). Multivariate analysis studies interactions among three or more variables (e.g., multiple regression, factor analysis).
5. How Would You Approach a Dataset That’s Missing More Than 30 Percent of Its Values?

**Ans:** First, analyze the missingness pattern (random or systematic). For high missingness:

* *Drop variables* if missingness is too high and irrelevant
* Use *imputation* (mean/median for numerical, mode for categorical) if data is missing randomly
* Apply *advanced methods* (MICE, KNN imputation) for structured missingness
* Consider *domain knowledge* to fill gaps or flag missingness as a separate category

1. Give an example where the median is a better measure than the mean  
   **Ans:** In income data where most people earn $30k–$60k but a few CEOs earn $10M+, the *median* (~$50k) better represents "typical" earnings than the *mean* (skewed upward by outliers). The median resists distortion from extreme values in skewed distributions.

**Key Insight:**  
Median is robust for:  
Asymmetric distributions (e.g., wealth, house prices)  
Ordinal data (e.g., survey ratings)  
Outlier-prone metrics (e.g., hospital wait times)

1. What is the difference between Descriptive and Inferential Statistics?

**Ans.** Descriptive statistics describe some sample or population.  
 Inferential statistics attempts to infer from some sample to the larger population

1. What are descriptive statistics?  
   **Ans:** Distribution – refers to the frequencies of responses.  
   Central Tendency – gives a measure or the average of each response.  
   Variability – shows the dispersion of a data set.
2. Can you state the method of dispersion of the data in statistics?  
   **Ans:** Dispersion measures how spread out data points are. Common methods include:

**Range** (Max – Min)

**Variance** (Average squared deviation from mean)

**Standard Deviation** (Square root of variance)

**Interquartile Range (IQR)** (Q3 – Q1, resistant to outliers)

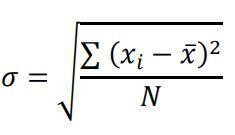
**Mean Absolute Deviation (MAD)** (Average absolute differences from mean)

1. How can we calculate the range of the data?  
   **Ans:** Range=Maximum Value−Minimum Value
2. Is the range sensitive to outliers?  
   **Ans: Yes**, the range is **highly sensitive to outliers** because it depends only on the extreme values.  
   **Example**:

Original data 10,12,15,18,2010,12,15,18,20 → Range = **10**

With an outlier 10,12,15,18,100,10,12,15,18,100 → Range = **90**

1. What is the meaning of standard deviation?  
   **Ans.** Standard deviation is a statistic that measures the dispersion of a dataset relative to its mean. It is the average amount of variability in your dataset. It tells you, on average, how far each value lies from the mean.  
   A high standard deviation means that values are generally far from the mean, while a low standard deviation indicates that values are clustered close to the mean.  
   The standard deviation is calculated as the square root of variance by determining each data point’s deviation relative to the mean.



1. What are the scenarios where outliers are kept in the data?  
   **Ans:**1. Genuine Extreme Values (Not Errors)  
   Example: In income data, billionaires are true outliers but reflect real economic disparity.  
   Action: Keep them to avoid underestimating variability.

2. Domain-Specific Importance  
Example: In fraud detection, unusual transactions (outliers) signal potential fraud.  
Action: Preserve outliers for anomaly detection.

3. Scientific or Medical Relevance  
Example: A drug trial where a few patients show extreme reactions (may indicate side effects).  
Action: Investigate rather than discard.

4. Skewed but Natural Distributions  
Example: House prices in a city (a few luxury homes skew data).  
Action: Use robust statistics (median, IQR) instead of removing.

5. Predictive Modeling (If Outliers Affect Outcomes)  
Example: Stock market crashes (rare but critical for risk models).  
Action: Include them to improve model realism.

6. Small Sample Sizes  
Example: A dataset with 20 observations where 1 is an outlier.  
Action: Removing it may distort the remaining data.

1. What is Bessel's correction?  
   **Ans:** In statistics, Bessel’s correction is the use of n-1 instead of n in several formulas, including the sample variance and standard deviation, where n is the number of observations in a sample. This method corrects the bias in the estimation of the population variance. It also partially corrects the bias in the estimation of the population standard deviation, thereby, providing more accurate results.
2. What do you understand about a spread out and concentrated curve?  
   **Ans:**   
   **Spread Out Curve**: Data points are widely dispersed (high variance/standard deviation). Example: Test scores ranging from 20 to 100.

**Concentrated Curve**: Data clusters tightly around the mean (low variance). Example: Heights of adults in a small region (e.g., 165–175 cm).

**Visual Cue**:

Spread out → Flatter, wider distribution.

Concentrated → Taller, narrower peak.

1. Can you calculate the coefficient of variation?  
   Ans: **Formula**:

CV=(Standard DeviationMean)×100%*CV*=(MeanStandard Deviation​)×100%

**Example**:

Mean = 50, SD = 5 → CV=(5/50)×100=10%*CV*=(5/50)×100=10%.  
**Use**: Compares variability across datasets with different units/scales (e.g., risk analysis in finance).

1. State the case where the median is a better measure when compared to the mean.  
   Ans: **Use median when**:

Data is **skewed** (e.g., income, house prices).

**Outliers exist** (e.g., CEO salaries in a company).

**Ordinal data** (e.g., survey ratings like "1–5 stars").

**Example**:

Dataset: [30K, 35K, 40K, 45K, **10M**]

Mean = ~2M (distorted by outlier).

Median = 40K (better represents "typical" value).

1. How is missing data handled in statistics?  
   Ans: **Methods**:  
   **Deletion**:  
   Listwise (remove entire rows with missing values).  
   Pairwise (use available data points).

**Imputation**:  
Mean/Median/Mode (for numerical/categorical data).  
Predictive models (e.g., KNN, MICE).

**Flagging**: Mark missingness as a separate category.  
**Domain-Based**: Use expert knowledge to fill gaps.

1. What is meant by mean imputation for missing data? Why is it bad?  
   **Ans:** Replacing missing values with the mean of observed data.  
   **Why Bad?**:  
   **Reduces variance** (underestimates spread).  
   **Creates artificial spikes** at the mean.  
   **Ignores relationships** between variables.  
   **Better Alternatives**: Median imputation, regression imputation, or multiple imputation.
2. What is the benefit of using box plots?  
   **Ans:**  
   **Visualizes 5-number summary** (min, Q1, median, Q3, max).  
   **Identifies outliers** (points beyond whiskers).  
   **Compares distributions** across groups.  
   **Robust to skewness** (uses median/IQR).  
   **Example**: Comparing exam scores across classes.
3. What is the meaning of the five-number summary in Statistics?  
   Ans: **Minimum** (smallest value).  
   **Q1 (25th percentile)**: Median of lower half.  
   **Median (50th percentile)**: Middle value.  
   **Q3 (75th percentile)**: Median of upper half.  
   **Maximum** (largest value).  
   **Used in box plots** to show distribution shape, spread, and outliers.
4. What is the difference between the First quartile, the IInd quartile, and the IIIrd quartile?  
   **Ans:**   
   **First Quartile (Q1)**: 25th percentile → 25% of data lies below this value.  
   **Second Quartile (Q2)**: 50th percentile (Median) → 50% of data lies below.  
   **Third Quartile (Q3)**: 75th percentile → 75% of data lies below.
5. What is the difference between percent and percentile?
6. What is an Outlier?  
   **Ans:**  
   A data point **far outside the typical range** of a dataset.  
   **Example**: In [10, 12, 13, 15, 100], **100** is an outlier.
7. What is the impact of outliers in a dataset?  
   **Ans:**  
   **Skew Averages**: Mean shifts toward outliers (median is robust).  
   **Affect Models**: Regression/ML models may perform poorly.  
   **Misleading Conclusions**: Distort correlations and variability.
8. Mention methods to screen for outliers in a dataset.  
   **Ans:**

**Visual:** Box plots, scatter plots.  
**Z-Score:** Values beyond ±3 standard deviations.  
**IQR Method:**  
Lower Bound = Q1 − 1.5×IQR  
Upper Bound = Q3 + 1.5×IQR  
**Domain Knowledge:** Unusual but valid (e.g., billionaires in income data).

1. How you can handle outliers in the datasets.  
   Ans:  
   **Remove**: If errors (e.g., data entry mistakes).

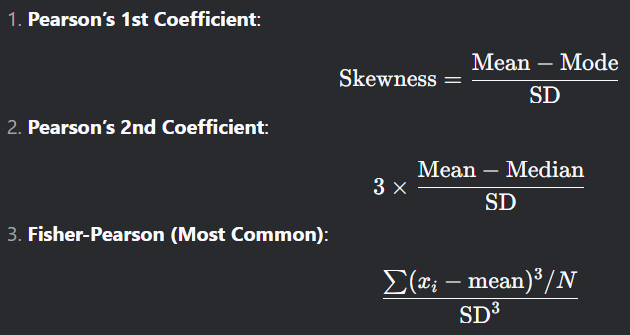
**Transform**: Log, square root (reduce skewness).

**Cap/Winsorize**: Replace outliers with nearest non-outlier values.

**Keep**: If meaningful (e.g., fraud detection).

1. What is the empirical rule?  
   **Ans:**

68% data within ±1σ (standard deviation) of mean.  
95% within ±2σ.  
99.7% within ±3σ.  
Example: Mean = 100, σ = 15 → 95% of values between 70 and 130.

1. How to calculate range and interquartile range?  
   Ans:   
   **Range**: Max − Min (e.g., [3, 5, 7, 9] → 9 − 3 = **6**)  
   **IQR**: Q3 − Q1 (e.g., Q1=5, Q3=8 → IQR = **3**).
2. What is skewness?  
   Ans:  
   Measures **asymmetry** in data distribution:  
   **Positive Skew**: Right tail longer (mean > median).  
   **Negative Skew**: Left tail longer (mean < median)  
   **Zero Skew**: Symmetric (e.g., normal distribution).
3. What are the different measures of Skewness?  
   Ans:  
   ​
4. What is kurtosis?  
   Ans: Measures if the distribution is peaked or flat   
   There is 3 types of kurtosis  
   Leptokurtic  
   Mesokurtic  
   platykurtic
5. Where are long-tailed distributions used?  
   Ans:  
   **Long-tailed distributions** (where extreme values/outliers are more frequent than in a normal distribution) are common in:  
   **Finance & Economics**  
   Stock market returns (extreme crashes/booms).  
   Insurance claims (rare but costly events).

**Natural Phenomena**  
Earthquake magnitudes (most are small, few are catastrophic).  
Rainfall/flood data (most days have light rain, rare heavy storms).

**Internet & Social Networks**  
Website traffic (few viral posts, many with low views).  
Wealth distribution (Pareto principle: 80% of wealth held by 20%).

**Risk Management**  
Modeling black swan events (e.g., pandemics, market collapses).  
**Example**: Power-law distributions (e.g., Pareto, Zipf’s law) describe these scenarios.

1. What is the central limit theorem?  
   Ans:  
   **Definition**:Regardless of the population’s distribution, the sampling distribution of the mean will approximate a **normal distribution** as the sample size grows (typically n≥30*n*≥30).

**Key Points**:  
**Requirements**:  
Random sampling.  
Independent observations.  
Sample size n*n* ≥ 30 (or population is normally distributed).

**Implications**:  
Enables use of parametric tests (e.g., t-tests) even for non-normal data.  
Justifies confidence intervals and hypothesis testing.

1. Can you give an example to denote the working of the central limit theorem?

**Scenario: Rolling Dice (Non-Normal Population)**

1. **Population Distribution**:
   * Single die roll = **Uniform distribution** (1 to 6, equal probability).
   * Clearly **non-normal** (flat, not bell-shaped).
2. **Sampling Process**:
   * Take **small samples** (e.g., roll 2 dice → n=2*n*=2), calculate the **mean of each sample**.
   * Repeat thousands of times to see the distribution of sample means.
3. **Results**:
   * ***n*=2**: Distribution of means starts triangular (not normal).
   * ***n*=30**: Distribution becomes **approximately normal**, centered at 3.5 (true population mean).

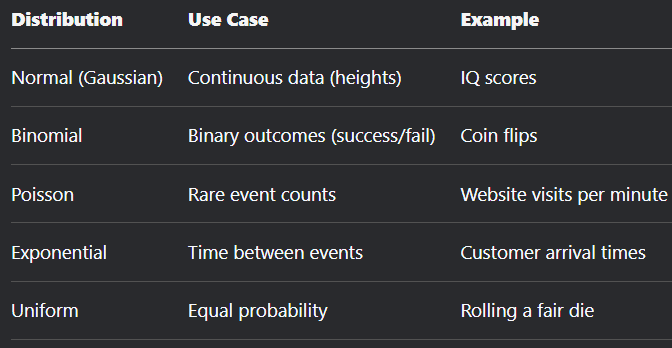
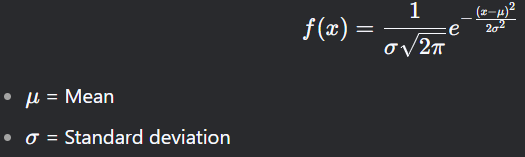
**Step-by-Step Simulation**

**1. Population Parameters (Single Die)**

* **Mean (μ)** = (1+2+3+4+5+6)/6=3.5(1+2+3+4+5+6)/6=3.5.
* **Standard Deviation (σ)** = (1−3.5)2+⋯+(6−3.5)26≈1.716(1−3.5)2+⋯+(6−3.5)2​​≈1.71.

**2. Sample Means for Different n**

| **Sample Size (n*n*)** | **Distribution Shape** | **Mean of Means** | **Standard Error (σ/√n)** |
| --- | --- | --- | --- |
| n=1(raw die) | Uniform (flat) | 3.5 | 1.71 |
| n=5 | Approaching normal | 3.5 | 1.71/sqrt(5)≈0.7 |
| n=30 | Nearly normal (bell curve) | 3.5 | 1.71/sqrt(30)≈0.31 |

1. What general conditions must be satisfied for the central limit theorem to hold?  
   **Ans:** The data must be sampled randomly  
   The sample values must be independent of each other  
   The sample size must be sufficiently large, generally it should be greater or equal than 30
2. What is the meaning of selection bias?  
   Ans: **Definition:**  
   Selection bias occurs when the sample collected is not representative of the population due to systematic errors in sampling. This leads to distorted or inaccurate results.  
   **Example:**  
   Surveying only smartphone users to study "internet usage habits" excludes non-smartphone users, introducing bias.
3. What are the types of selection bias in statistics?  
   Ans:  
   **Sampling Bias**: Non-random sampling (e.g., surveying only daytime shoppers).  
   **Self-Selection Bias**: Participants volunteer (e.g., online polls favoring enthusiastic respondents).  
   **Survivorship Bias**: Focusing only on "survivors" (e.g., studying successful companies while ignoring failed ones).  
   **Time Interval Bias**: Data collected during atypical periods (e.g., retail sales during holidays).
4. What is the probability of throwing two fair dice when the sum is 8?  
   Ans:   
   Total outcomes: 6×6=366×6=36.  
   Favorable combinations: (2,6), (3,5), (4,4), (5,3), (6,2) → **5 outcomes**.  
   Probability:  
   P(Sum=8)=536≈13.89%*P*(Sum=8)=365​≈13.89%
5. What are the different types of Probability Distribution used in Data Science?  
   
6. What do you understand by the term Normal Distribution or What is a bell-curve distribution??  
   Ans: A symmetric, bell-shaped distribution where: Mean = Median = Mode.  
   68% of data within ±1σ, 95% within ±2σ.  
   **Example:**  
   Human heights cluster around the mean (e.g., 170 cm) with fewer extremes.
7. Can you state the formula for normal distribution?  
   
8. What type of data does not have a normal distribution or a Gaussian distribution?  
   Ans:   
   **Skewed**: Income, house prices (right-skewed).  
   **Multimodal**: Data with multiple peaks (e.g., exam scores from two classes).  
   **Discrete**: Count data (e.g., number of children).  
   **Heavy-Tailed**: Stock market returns.
9. What is the relationship between mean and median in a normal distribution?  
   Ans: Mean=Median=Mode
10. What are some of the properties of a normal distribution?  
    Ans:  
    Symmetric around the mean.  
    Total area under curve = 1.  
    Defined by mean (μ*μ*) and SD (σ*σ*).  
    Follows Empirical Rule (68-95-99.7).
11. What is the assumption of normality?  
    **Ans:**  
    Statistical tests (e.g., t-tests, ANOVA) assume data is normally distributed.  
    **Violations**: Use non-parametric tests (Mann-Whitney U, Kruskal-Wallis).
12. How to convert normal distribution to standard normal distribution?  
    Ans: *Z*=*X*−*μ*​/ σ

**Example**: If μ=100*μ*=100, σ=15*σ*=15, then X=130*X*=130 → Z=2.0*Z*=2.0.

1. Can you tell me the range of the values in standard normal distribution?  
   Ans: Theoretical range: −∞−∞ to +∞+∞, but practically:

99.7% of values lie between **-3 and +3**.

1. What is the Pareto principle?  
   Ans:   
   **Definition**: 80% of effects come from 20% of causes.  
   **Applications**: Business: 80% of sales from 20% of customers.  
   Software: 80% of bugs from 20% of code.  
   **Statistical Basis**: Power-law distribution (long-tailed).
2. What are left-skewed and right-skewed distributions?  
   Ans: Skewness is a way to describe the symmetry of a distribution.  
   A left-skewed (Negative Skew) distribution is one in which the left tail is longer than that of the right tail. For this distribution, ***mean < median < mode*.**  
   Similarly, right-skewed (Positively Skew) distribution is one in which the right tail is longer than the left one. For this distribution, ***mean > median > mode.***
3. If a distribution is skewed to the right and has a median of 20, will the mean be greater  
   than or less than 20?  
   Ans: **Mean > Median** (20).

**Why?** Right skew means the tail pulls the mean upward.

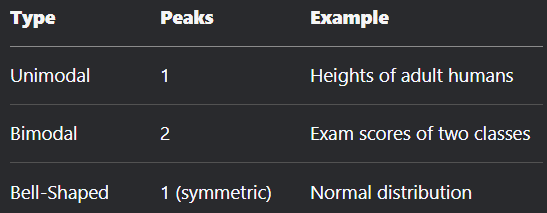
**Example**: Data: [10, 15, 20, 25, **100**], Median = 20, Mean = 34 → Higher due to outlier (100).

1. Given a left-skewed distribution that has a median of 60, what conclusions can we draw about the mean and the mode of the data?

**Ans**: Mean<Median (60)<Mode

1. Imagine that Jeremy took part in an examination. The test has a mean score of 160, and it has a standard deviation of 15. If Jeremy’s z-score is 1.20, what would be his score on the test?  
   **Ans:** *X*=*μ*+(*Z*×*σ*) and *X*=160+(1.20×15)=178
2. The standard normal curve has a total area to be under one, and it is symmetric around zero. True or False? **True**
3. Briefly explain the procedure to measure the length of all sharks in the world.  
   Ans:   
   **Sampling**: Collect data from diverse locations (oceans, species).

Use stratified sampling for species/subgroups.  
**Tools**:  
Satellite tags, underwater cameras, or direct measurements.  
**Statistical Analysis**:  
Calculate mean, SD, and confidence intervals.  
Account for measurement errors.  
**Challenge**: Impossible to measure every shark → Reliance on representative samples.

1. Can you tell me the difference between unimodal bimodal and bell-shaped curves?  
   
2. Does symmetric distribution need to be unimodal?  
   **Answer**: No.  
   **Explanation**:  
   **Unimodal Symmetric**: Normal distribution.  
   **Multimodal Symmetric**: Multiple peaks (e.g., mixture of two normal distributions with equal means/SDs).  
   **Example**:  
   Data combining male and female heights (if mean heights are equal) → Bimodal but symmetric.
3. What are some examples of data sets with non-Gaussian distributions?

**Ans.** When data follows a non-normal distribution, it is frequently non-Gaussian. A non-Gaussian distribution is often seen in many statistics processes. This occurs when data is naturally clustered on one side or the other on a graph. For instance, bacterial growth follows an exponential or non-Gaussian distribution, which is non-normal.

1. What is the Binomial Distribution Formula?  
   **Ans**: *P*(*X*=*k*)=(*kn*​)*pk*(1−*p*)*n*−*k*
2. What are the criteria that Binomial distributions must meet?  
   **Ans:** A Binomial Distribution must satisfy:

**Fixed Trials (n*n*)**: Number of trials is predetermined (e.g., 10 coin flips).

**Binary Outcomes**: Only two outcomes per trial (success/failure).

**Constant Probability (p*p*)**: Probability of success remains the same across trials.

**Independence**: Trials do not influence each other.

**Violation Example**:

**Non-binary**: Rolling a die (6 outcomes) → Not Binomial.

**Dependent trials**: Drawing cards without replacement → Not Binomial.

1. What are the examples of symmetric distribution?  
   Ans:   
   **Normal (Gaussian) Distribution**:  
   Bell-shaped, symmetric around the mean.  
   Example: Heights of adult humans.  
   **Uniform Distribution**:  
   All outcomes equally likely.  
   Example: Rolling a fair die.

**Student’s t-Distribution**:  
Symmetric, heavier tails than Normal.  
Used for small sample sizes.

**Cauchy Distribution**:  
Symmetric but with heavy tails (no defined mean/variance).

1. How to find the mean length of all fishes in the sea?  
   Ans:  
   Define the confidence level (most common is 95%)  
   Take a sample of fishes from the sea (to get better results the number of fishes > 30)  
   Calculate the mean length and standard deviation of the lengths  
   Calculate t-statistics  
   Get the confidence interval in which the mean length of all the fishes should be.
2. What are the types of sampling in Statistics?  
   Ans:  
   **Random Sampling**:  
   Simple Random: Every member has an equal chance (e.g., lottery draw).  
   Systematic: Select every k*k*-th item (e.g., every 10th student in a list).

**Stratified Sampling**:  
Divide population into subgroups (strata) and sample from each (e.g., age groups).

**Cluster Sampling**:  
Divide population into clusters, randomly select clusters, and sample all within them (e.g., schools in a district).

**Convenience Sampling**:  
Use readily available data (e.g., surveying friends). *Prone to bias!*

**Snowball Sampling**:  
Participants recruit others (e.g., rare disease studies).

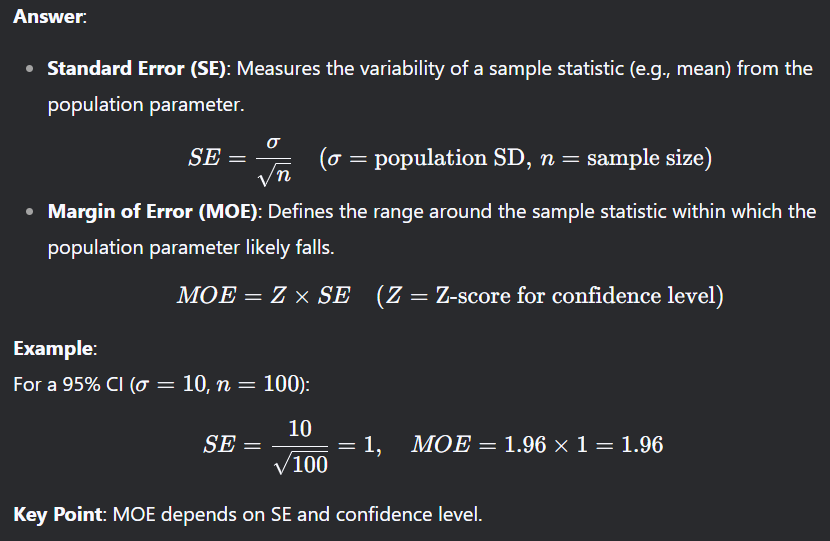
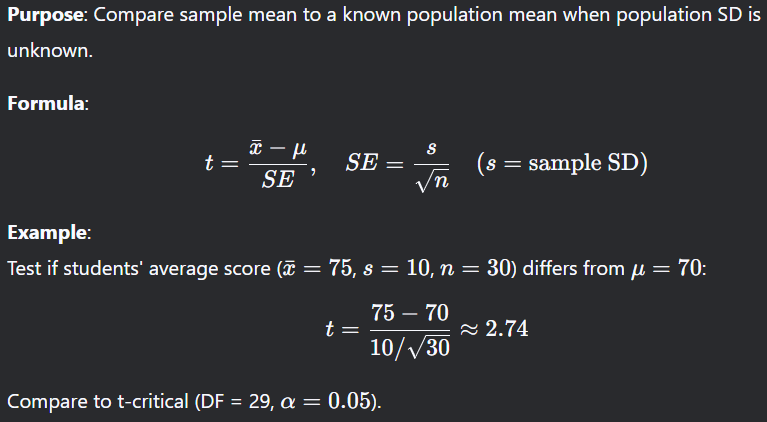
1. Why is sampling required?  
   Ans:  
   **Cost & Time**: Cheaper/faster than surveying entire populations.  
   **Feasibility**: Impossible to study all sharks, all voters, etc.  
   **Accuracy**: Proper sampling reduces errors vs. haphazard data collection.  
   **Example**:  
   Testing all smartphones for defects destroys inventory → Sample testing preserves resources.
2. How do you calculate the needed sample size?  
   Ans: *n*=(*E*⋅*σ*​)^2/E  
   *Z* = Z-score (e.g., 1.96 for 95% confidence).  
   *σ* = Population SD (estimate).  
   *E* = Margin of error.
3. Where is inferential statistics used?  
   Ans:   
   **Hypothesis Testing**: Compare groups (e.g., Drug A vs. Placebo).

**Confidence Intervals**: Estimate population parameters (e.g., average income).

**Regression Analysis**: Predict outcomes (e.g., sales based on ad spend).

**Quality Control**: Monitor manufacturing processes.

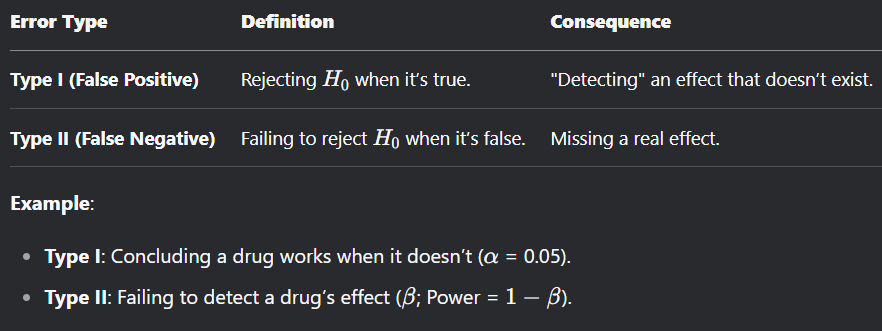
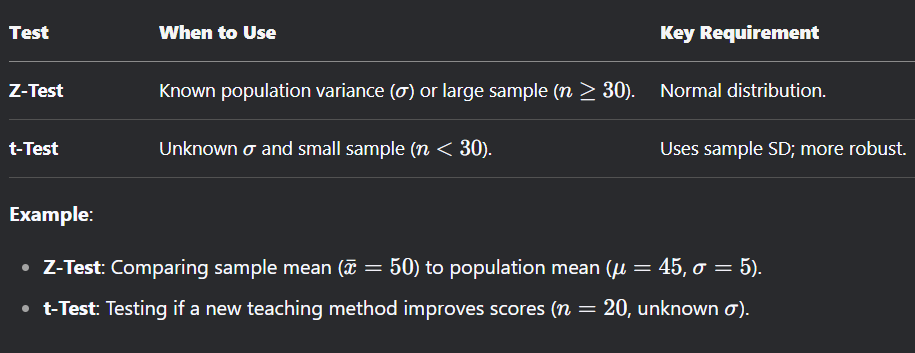
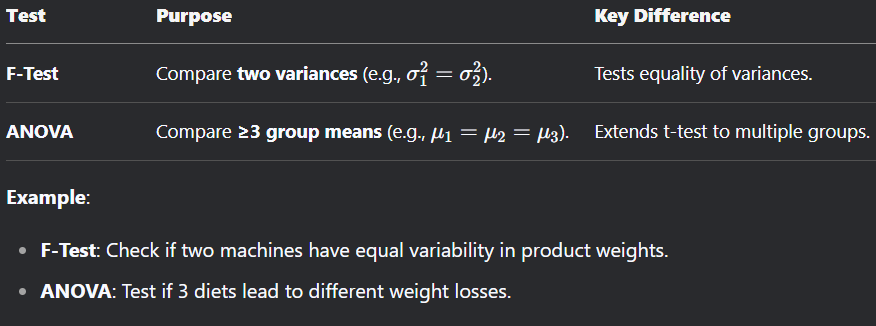
**Real-World Application**:  
Clinical trials use inferential stats to generalize drug effects from samples to populations.

1. What is the difference between Point Estimate and Confidence Interval Estimate?  
   Ans: Point = Single value estimate (e.g, sample mean)  
    CIE : Range of plausible values (e.g., 45–55 kg).
2. What do you understand about biased and unbiased terms?  
   Ans:   
   **Unbiased Estimator**: Expected value = Population parameter (e.g., sample mean xˉ*x*ˉ for μ*μ*).  
   **Biased Estimator**: Systematic deviation (e.g., using n*n* instead of n−1*n*−1 for sample variance).
3. How does the width of the confidence interval change with length?  
   Ans: Width∝1/sqrt(n)​
4. What is a Sampling Error and how can it be reduced?  
   Ans: Difference between sample statistic and population parameter.  
   **Reduction Methods**:  
   **Increase Sample Size**: Lowers SE.  
   **Random Sampling**: Minimizes bias.  
   **Stratification**: Ensures subgroup representation.  
   **Example**:  
   Polling 1,000 people vs. 100 → Smaller sampling error.
5. How do the standard error and the margin of error relate?  
    
6. What is hypothesis testing?  
   Ans: A statistical method to test claims about population parameters using sample data.
7. What is an alternative hypothesis?  
   Ans: The hypothesis that contradicts *H*0​, representing the effect or difference you seek evidence for.
8. **What is the difference between one-tailed and two-tail hypothesis testing?**
9. What is one sample t-test?  
   
10. What is the meaning of degrees of freedom (DF) in statistics?  
    **Definition**:  
    The number of independent values in a calculation that can vary.  
    **For t-Test**:  
    DF=n−1(n=sample size) *DF*=*n*−1(*n*=sample size)  
    **Why**: One parameter (mean) is estimated from the sample, reducing variability.  
    **Example**:  
    Sample size *n*=20 → DF = 19 for a one-sample t-test.
11. What is the p-value in hypothesis testing?  
    **Ans**: p-value is a number that describes the probability of finding the observed or more extreme results when the null hypothesis (H0) is True.  
    P-values are used in hypothesis testing to help decide whether to reject the null hypothesis or not. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.
12. How can you calculate the p-value?  
    Ans: **Steps**:  
    **Choose Test**: e.g., t-test, Z-test.  
    **Compute Test Statistic**: e.g., t=2.0*t*=2.0.  
    **Use Distribution Table**: Find area beyond the test statistice  
    **Example**:  
    For t=2.0*t*=2.0 (DF = 20, two-tailed):  
    p-value = 2×P(T>2.0)≈0.0592×*P*(*T*>2.0)≈0.059.

**Software**:  
Python: scipy.stats.t.sf(abs(t\_score), df) \* 2 (two-tailed).  
Excel: =T.DIST.2T(2.0, 20).

1. If there is a 30 percent probability that you will see a supercar in any 20-minute time interval, what is the proba bility that you see at least one supercar in the period of an hour (60 minutes)?  
   **Ans**:   
   Probability in **20-minute interval** = 30% (p=0.3*p*=0.3).  
   **Total time** = 60 minutes → **3 intervals**.  
   **Approach**:  
   Calculate probability of **not** seeing a supercar in one interval:  
   P(None)=1−0.3=0.7*P*(None)=1−0.3=0.7For 3 independent intervals:  
   P(None in 60 mins)=0.73=0.343*P*(None in 60 mins)=0.73=0.343  
   Probability of **at least one** supercar:  
   P(≥1)=1−0.343=0.657(65.7%)*P*(≥1)=1−0.343=0.657(65.7%)  
   **Answer**: **65.7% chance** to see at least one supercar in 60 minutes.
2. How would you describe a ‘p-value’?  
   Ans: The **p-value** is the chance of seeing your data (or something more extreme) if the null hypothesis (H0​) is true.

**Analogy**:  
**Null Hypothesis (H0​)**: "This coin is fair."  
**p-value**: "If the coin is fair, the probability of getting 8 heads in 10 flips is 5%."  
Low p-value (e.g., 0.01) → Unlikely H0 is true.  
High p-value (e.g., 0.30) → No evidence against *H*0​.  
**Key Point**:  
**p < 0.05** → Reject H0​ (standard threshold).

1. What is the difference between type I vs type II errors?  
   
2. When should you use a t-test vs a z-test?  
   
3. What is the difference between the f test and anova test?  
   
4. What is Resampling and what are the common methods of resampling?  
   Ans: Techniques to draw repeated samples from data to estimate accuracy (e.g., confidence intervals)  
   **Common Methods**:  
   **Bootstrapping**  
   Randomly sample **with replacement** to estimate statistics (e.g., median).  
   Example: Estimating confidence intervals for a skewed dataset.  
   **k-Fold Cross-Validation**:  
   Split data into k subsets; train on *k*−1, test on 1. Repeat *k* times.  
   Example: Validating a machine learning model’s performance.  
   **Key Benefit**: Avoids assumptions about data distribution
5. What is the proportion of confidence intervals that will not contain the population parameter?  
   Ans: The proportion of confidence intervals (CIs) that **will not** contain the true population parameter is equal to the **significance level (α*α*)**.  
   **For a 95% CI**: α=0.05*α*=0.05 → **5%** of CIs will miss the parameter.  
   **For a 99% CI**: α=0.01*α*=0.01 → **1%** of CIs will miss the parameter.

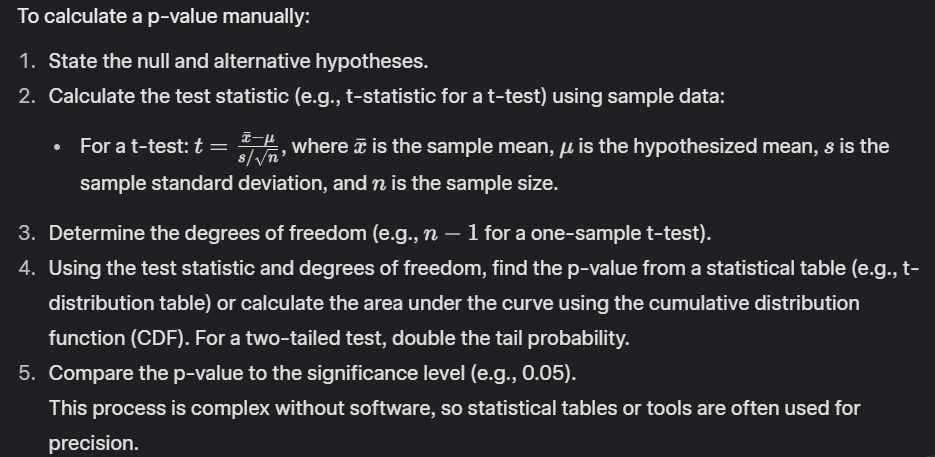
**Why?**  
A 95% CI means that if you repeat the sampling process infinitely, 95% of intervals will capture the true parameter, while 5% will not.

This aligns with the **Type I error rate (α*α*)** in hypothesis testing.

**Example**:  
If you construct 100 independent 95% CIs, ~5 intervals are expected to exclude the population mean.

1. What is a confounding variable?  
   **Ans**: A confounding variable in statistics is an ‘extra’ or ‘third’ variable that is associated with both the dependent variable and the independent variable, and it can give a wrong estimate that provides useless results.  
   For example, if we are studying the effect of weight gain, then lack of workout will be the independent variable, and weight gain will be the dependent variable. In this case, the amount of food consumption can be the confounding variable as it will mask or distort the effect of other variables in the study. The effect of weather can be another confounding variable that may later the experiment design.
2. What are the steps we should take in hypothesis testing?   
   **Ans.**  
   1. State the null hypothesis  
   2. State the alternate hypothesis  
   3. Which test and test statistic to be performed  
   4. Collect Data  
   5. Calculate the test statistic  
   6. Construct Acceptance / Rejection regions  
   7. Based on steps 5 and 6, draw a conclusion about H0
3. How would you describe what a ‘p-value’ is to a non-technical person or in a layman term?  
   **Ans**: The best way to describe the p-value in simple terms is with an example. In practice, if the p-value is less than the alpha, say of 0.05, then we’re saying that there’s a probability of less than 5% that the result could have happened by chance. Similarly, a p-value of 0.05 is the same as saying “5% of the time, we would see this by chance.”
4. What does interpolation and extrapolation mean? Which is generally more accurate?  
   **Ans:**  
   Interpolation is a prediction made using inputs that lie within the set of observed values. Extrapolation is when a prediction is made using an input that’s outside the set of observed values. Generally, interpolations are more accurate.
5. You roll a biassed coin (p(head)=0.8) five times. What’s the probability of getting three or more heads?  
   **Ans:** It's always good to cross-verify. The total probability must be 1, so the probability of getting less than 3 heads (0,1,2) plus P(X≥3) should be 1.  
   P(X=0) = C(5,0) × (0.8)^0 × (0.2)^5 = 1 × 1 × 0.00032 = 0.00032  
   P(X=1) = C(5,1) × (0.8)^1 × (0.2)^4 = 5 × 0.8 × 0.0016 = 0.0064  
   P(X=2) = C(5,2) × (0.8)^2 × (0.2)^3 = 10 × 0.64 × 0.008 = 0.0512  
   Sum of P(X=0,1,2) = 0.00032 + 0.0064 + 0.0512 = 0.05792  
   Now, P(X≥3) = 1 - P(X<3) = 1 - 0.05792 = 0.94208
6. Chi-square test?  
   Ans: A statistical method is used to find the difference or correlation between the observed and expected categorical variables in the dataset.  
   Example: A food delivery company wants to find the relationship between gender, location and food choices of people in India.  
   It is used to determine whether the difference between 2 categorical variables is:

* Due to chance or
* Due to relationship

1. What is the ANOVA test?  
   Ans: ANOVA (Analysis of Variance) is a statistical test used to compare means among three or more groups to determine if there are significant differences between them. It assesses whether the observed differences in group means are due to random chance or an actual effect. One-way ANOVA is used for one independent variable, while two-way ANOVA includes two independent variables.
2. How to calculate p-value using a manual method?  
   
3. What is the goal of A/B testing?  
   Ans:   
   The goal of A/B testing is to compare two versions (A and B) of a variable (e.g., a webpage design, marketing email) to determine which performs better based on a specific metric (e.g., click-through rate, conversion rate). It uses statistical analysis to identify significant differences in performance, helping optimize decisions.
4. What is the difference between a box plot and a histogram  
    **Box Plot**: A graphical summary of data distribution showing the median, quartiles (Q1, Q3), and potential outliers. It highlights the spread and skewness of data but doesn’t show the frequency distribution.
5.  **Histogram**: A plot showing the frequency distribution of continuous data by dividing it into bins. It visualizes the shape of the distribution (e.g., normal, skewed) but doesn’t directly show quartiles or outliers.  
   **Key Difference**: Box plots summarize key statistics; histograms show the full distribution shape.
6. What is a confidence interval and how do you interpret it?   
   A confidence interval (CI) is a range of values, derived from sample data, that is likely to contain the true population parameter (e.g., mean) with a specified confidence level (e.g., 95%).  
   Interpretation: A 95% CI means that if we repeated the sampling process many times, about 95% of the intervals would contain the true parameter value. It does not mean there’s a 95% chance the true value lies within this specific interval.
7. How do you stay up-to-date with the new and upcoming concepts in statistics?  
   Ans:  
   Read academic journals (e.g., *Journal of the American Statistical Association*).  
   Follow statistical blogs and websites (e.g., Statistical Modeling, R-bloggers).  
   Participate in online courses or webinars (e.g., Coursera, edX).  
   Engage with communities on platforms like X or Stack Exchange for discussions.  
   Attend conferences or workshops (e.g., JSM, useR!).  
   Experiment with new tools and methods in R, Python, or other statistical software.
8. What is correlation?  
   Ans:  
   Correlation measures the strength and direction of a linear relationship between two variables. It ranges from -1 (perfect negative linear relationship) to 1 (perfect positive linear relationship), with 0 indicating no linear relationship.
9. What types of variables are used for Pearson’s correlation coefficient?  
   Ans:  
   Pearson’s correlation coefficient is used for two continuous variables that are assumed to have a linear relationship and are approximately normally distributed.
10. In an observation, there is a high correlation between the time a person sleeps and the amount of productive work he does. What can be inferred from this?  
    **ANS:**  
    A high correlation suggests a strong linear relationship between sleep time and productive work, but **correlation does not imply causation**. It could mean:  
    More sleep leads to higher productivity.  
    Higher productivity allows for more sleep.  
    A third factor (e.g., health) affects both.  
    Further study (e.g., controlled experiments) is needed to establish causality.
11. What is the meaning of covariance?  
    Ans:  
    Covariance measures how two variables change together. Positive covariance indicates that as one variable increases, the other tends to increase. Negative covariance indicates that as one increases, the other tends to decrease. Unlike correlation, covariance is not standardized and depends on the units of the variables.
12. What does autocorrelation mean?  
    Ans:  
    Autocorrelation (or serial correlation) is the correlation of a time series with its own past values (lags). It measures how a variable’s current value is related to its previous values, often used in time series analysis to detect patterns or dependencies.
13. What can be the reason for non normality of the data?  
    **Ans:**  
    Reasons for non-normal data include:  
    **Skewed distributions**: Outliers, heavy tails, or asymmetric data.  
    **Small sample size**: May not approximate normality.  
    **Mixture of populations**: Data from multiple subgroups with different distributions.  
    **Non-linear relationships**: Data following exponential, logarithmic, or other patterns.  
    **Measurement errors**: Inconsistent data collection or outliers.  
    **Inherent properties**: Some variables (e.g., time-to-event, counts) are naturally non-normal.
14. why is there no such thing like 3 samples t- test?? why t-test failed with 3 samples  
    Ans:  
    There’s no “3-sample t-test” because the t-test is designed for comparing **two groups** (e.g., two means). For three or more groups:

**Why t-test fails**: Performing multiple t-tests (e.g., pairwise comparisons) increases the risk of Type I errors (false positives) due to multiple testing, inflating the overall error rate.

**Solution**: Use **ANOVA** (Analysis of Variance) to compare means across three or more groups simultaneously, controlling the error rate. If ANOVA shows significance, post-hoc tests (e.g., Tukey’s HSD) can identify which groups differ.

**Non-parametric alternative**: For non-normal data, use the Kruskal-Wallis test for three or more groups.