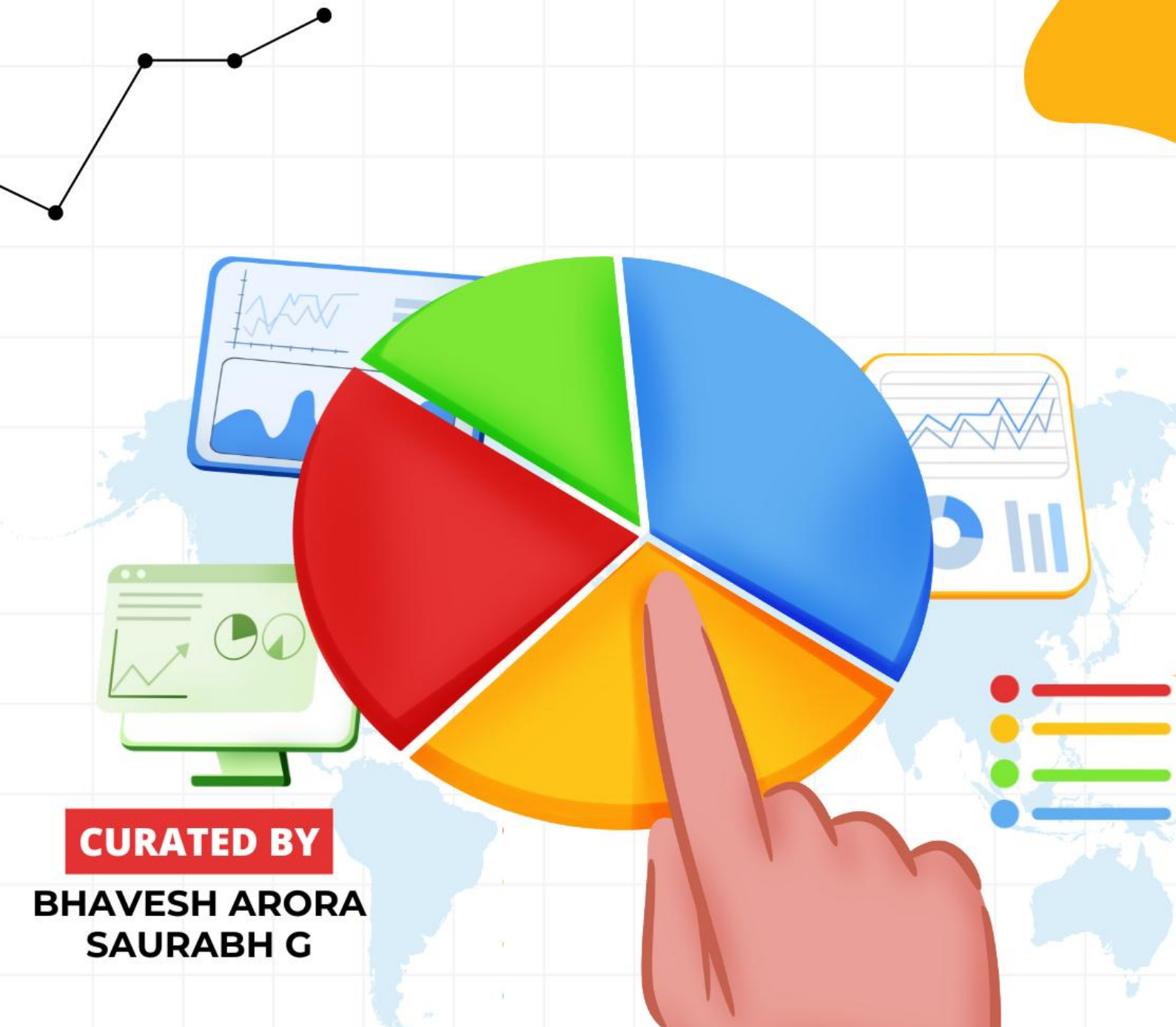
LEARN COMPLETE

STATISTICS

WITH MOST ASKED INTERVIEW QUESTIONS



CURATED BY

BHAVESH ARORA
SAURABH G

Crack Top Analyst Roles with This 15-Day STATISTICS Series!

Learn Complete STATISTICS with Most ASKED INTERVIEW Questions

Are you aiming for roles like Business Analyst, Data Analyst, Power BI Developer, or BI Consultant at top companies?

Curated by

SAURABH G
Founder at DataNiti
6+ Years of Experience | Senior Data Engineer
Linkedin: www.linkedin.com/in/saurabhgghatnekar

BHAVESH ARORA
Senior Data Analyst at Delight Learning Services
M.Tech – IIT Jodhpur | 3+ Years of Experience
Linkedin: www.linkedin.com/in/bhavesh-arora-11b0a319b

Connect with us: https://topmate.io/bhavesh_arora/

Let’s embark on this journey together and make your dreams a reality, starting today.

◇ Day 1: Descriptive Statistics – Mean, Median, Mode, Range

Descriptive statistics summarize and describe the main features of a dataset. These measures help in understanding data distribution.

☑ 1. Mean (Average)

The sum of all values divided by the number of values.

Formula:

$$\text{Mean} = \frac{\sum x_i}{n}$$

Python Code:

```
import numpy as np
data = [10, 20, 30, 40, 50]
print("Mean:", np.mean(data))
```

☑ 2. Median

The middle value when data is sorted.

- If odd: middle value
- If even: average of two middle values

Python Code:

```
print("Median:", np.median(data))
```

☑ 3. Mode

The most frequently occurring value in a dataset.

Python Code:

```
from scipy import stats
print("Mode:", stats.mode(data).mode[0])
```

☑ 4. Range

The difference between the highest and lowest value.
Formula:

$$\text{Range} = \max(x) - \min(x)$$

Python Code:
print("Range:", max(data) - min(data))

💬 INTERVIEW QUESTIONS (Medium to High)

◇ **Q1:** When would you prefer median over mean?
☑ When the data has **outliers or is skewed**, median gives a better central tendency.

◇ **Q2:** Can a dataset have more than one mode?
☑ Yes — it's called **bimodal** or **multimodal** depending on the number of modes.

◇ **Q3:** Why is range not always a reliable measure of dispersion?
☑ Because it's **highly affected by outliers**; it only considers the extremes.

◇ **Q4:** In what scenarios is the mean misleading?
☑ In **skewed distributions** (e.g., income, house prices), mean doesn't reflect the real central value.

◇ Day 2: Measures of Dispersion – Variance, Standard Deviation, IQR

Understanding how **spread out** your data is can be just as important as knowing the average.

☑ 1. Variance (σ^2 or s^2)

Variance tells how far each number in the dataset is from the mean.

Formulas:

- **Population Variance (σ^2):**

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

- Sample Variance (s^2):

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

Python Code:

```
import numpy as np
data = [10, 20, 30, 40, 50]
print("Sample Variance:", np.var(data, ddof=1))  # ddof=1 for sample
```

☑ 2. Standard Deviation (σ or s)

The square root of variance. It's in the **same unit** as the data.
Formula:

$$\sigma = \sqrt{\text{Variance}}$$

Python Code:
print("Standard Deviation:", np.std(data, ddof=1))

☑ 3. Interquartile Range (IQR)

IQR measures the middle 50% spread of the data.

Formula:

$$\text{IQR} = Q3 - Q1$$

Python Code:
Q1 = np.percentile(data, 25)
Q3 = np.percentile(data, 75)
print("IQR:", Q3 - Q1)

💬 INTERVIEW QUESTIONS (Medium to High)

◇ **Q1:** What is the relationship between variance and standard deviation?
☑ Standard deviation is the **square root of variance**, making it more interpretable in real-world terms.

◇ **Q2:** Why do we divide by (n - 1) in sample variance?
☑ To **correct the bias** in estimation — this is called **Bessel's correction**.

◇ **Q3:** Why is IQR preferred over range in some cases?
☑ IQR **ignores outliers**, making it more robust for **skewed distributions**.

◇ **Q4:** When is a high standard deviation a concern?
☑ When **consistency** is desired (e.g., product quality, risk in investments), a high SD means high variability.

◇ Day 3: Probability Basics – Types, Rules & Real-World Use Cases

Probability forms the **foundation of statistics** and is critical in data analysis, risk modeling, and machine learning.

☑ 1. Types of Probability

Type	Description	Example
Theoretical	Based on reasoning or known outcomes	Coin flip: P(Heads) = 0.5
Experimental	Based on actual experiments/data	3 out of 10 students are left-handed → P(LH) = 0.3
Subjective	Based on intuition or experience	"I feel there's a 70% chance of rain"

☑ 2. Basic Probability Formula

$$P(E) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$$

Python Code:

```
favorable = 3
total = 10
print("Probability:", favorable / total)
```

✅ 3. Rules of Probability

📌 Rule 1 – Complement Rule:

$$P(\text{Not A}) = 1 - P(A)$$

📌 Rule 2 – Addition Rule (Mutually Exclusive):

$$P(A \cup B) = P(A) + P(B)$$

📌 Rule 3 – Multiplication Rule (Independent):

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

✅ 4. Independent vs Dependent Events

Term	Explanation	Example
Independent	One event does not affect the other	Tossing two coins
Dependent	One event affects the outcome of another	Drawing cards **without** replacement

☑ 5. Real-World Use Cases

- **Spam filters** use conditional probability (Bayes' Theorem)
- **Risk analysis** in insurance and banking
- **Predictive modeling** in ML algorithms
- **A/B testing** and marketing success prediction

💬 INTERVIEW QUESTIONS (Medium to High)

🔸 **Q1:** What's the difference between independent and mutually exclusive events?

☑ Independent: Events don't influence each other.

☑ Mutually Exclusive: Both events **cannot occur together**.

🔸 **Q2:** What is the probability of either A or B occurring if they are **not mutually exclusive**?

✅

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

🔸 **Q3:** Why is conditional probability important in machine learning?

✅ Used in **Naive Bayes**, **recommendation engines**, **fraud detection**, etc.

🔸 **Q4:** Give an example where probability helps in real-world decision making.

✅ In **A/B testing**, probability determines if a new version is significantly better.

🔸 Day 4: Conditional Probability & Bayes' Theorem – Intuition to Application

Conditional probability and Bayes' Theorem are the backbone of **predictive analytics**, **machine learning**, and **risk assessment**.

✅ 1. Conditional Probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

> "What is the probability of A happening, **given** that B has happened?"

📌 Example:

If 30% of emails are spam (P(B)), and 10% of those contain "Free" (P(A ∩ B)),

$$P(\text{'Free'} \mid \text{Spam}) = \frac{0.10}{0.30} = 0.33$$

✅ 2. Bayes' Theorem

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Used to **update belief** in event A after observing event B.

🔍 **Example:** Medical Test

- P(Disease) = 0.01
- P(Positive Test | Disease) = 0.99
- P(Positive Test | No Disease) = 0.05

What's the probability a person **actually has the disease** given a positive test?

Let's say:

$$P(\text{Disease} \mid \text{Positive}) = \frac{0.99 \cdot 0.01}{0.99 \cdot 0.01 + 0.05 \cdot 0.99} \approx 0.167$$

✅ Even with a positive result, the actual chance is just ~16.7% due to the **base rate** being very low!

☑ 3. Python Code Example

```
# Bayes' Theorem Calculation
P_A = 0.01  # Prior: Disease
P_B_given_A = 0.99  # Test positive if disease
P_B_given_not_A = 0.05  # False positive
P_not_A = 1 - P_A

P_B = (P_B_given_A * P_A) + (P_B_given_not_A * P_not_A)
P_A_given_B = (P_B_given_A * P_A) / P_B
print("P(Disease | Positive):", P_A_given_B)
```

☑ 4. Use Cases of Bayes' Theorem

- **Spam filters** (Gmail, Outlook)
- **Medical diagnostics**
- **Recommendation systems**
- **Credit scoring & fraud detection**

💬 INTERVIEW QUESTIONS (Medium to High)

◇ **Q1:** What's the difference between P(A|B) and P(B|A)?
☑ They are not the same! Bayes' Theorem connects them.

◇ **Q2:** Why is Bayes' Theorem important in ML?
☑ It allows updating predictions based on **new evidence** (e.g., Naive Bayes Classifier).

◇ **Q3:** What's the biggest trap in interpreting conditional probability?
☑ Ignoring the **base rate** (prior probability). It can skew results drastically.

◇ **Q4:** How would you explain conditional probability to a non-technical person?
☑ "If it's raining, the chance of traffic increases. That's conditional probability—knowing one thing changes the likelihood of another."

◇ Day 5: Probability Distributions – Discrete vs Continuous

Probability distributions tell us how likely outcomes are—**the foundation for modeling uncertainty** in real-world data.

☑ 1. Discrete Distributions

These apply when outcomes are **countable** (whole numbers).

📌 Common Examples:

- **Bernoulli Distribution** (Single binary trial: Success/Failure)
- **Binomial Distribution** (Multiple independent binary trials)
- **Poisson Distribution** (Number of events in a fixed interval)

🔸 Bernoulli Example:

$$P(X = 1) = p, \quad P(X = 0) = 1 - p$$

🔸 Binomial Example:

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

🧠 Think: Tossing a coin 10 times (n = 10), how many heads?

✅ 2. Continuous Distributions

Outcomes are **uncountable**, like height, weight, time.

📌 Common Examples:

- **Uniform Distribution** (Equal probability in a range)
- **Normal Distribution** (Bell curve – most natural phenomena)
- **Exponential Distribution** (Time between events)

🔸 **Normal Distribution Equation:**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

🧠 Used in: IQ scores, test results, height measurements

☑ 3. Python Code Example

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import binom, norm


# Binomial (n=10, p=0.5)
x = np.arange(0, 11)
binom_probs = binom.pmf(x, 10, 0.5)
plt.bar(x, binom_probs)
plt.title('Binomial Distribution (n=10, p=0.5)')
plt.show()

# Normal
x = np.linspace(-3, 3, 100)
plt.plot(x, norm.pdf(x, 0, 1))
plt.title('Normal Distribution (μ=0, σ=1)')
plt.show()
```

☑ 4. When to Use Which

Scenario	Use
Count of events (e.g., calls per hour)	Poisson
Repeated binary trials	Binomial
Heights, weights	Normal
Waiting times	Exponential

💬 INTERVIEW QUESTIONS (Medium to High)

◇ **Q1:** What's the difference between discrete and continuous variables?
☑ Discrete = Countable, Continuous = Measurable

◇ **Q2:** How is the binomial distribution related to Bernoulli?
☑ Binomial is a **sum of independent Bernoulli trials**

◇ **Q3:** Why is the normal distribution so widely used?
☑ Central Limit Theorem: Means of samples from any distribution become normal as sample size grows.

◇ **Q4:** Can probabilities be assigned to exact values in continuous distributions?
☑ No. For continuous variables, **P(X = x) = 0**. Only ranges matter (e.g., P(2 < X < 5)).

◇ Day 6: Central Limit Theorem (CLT) – Why the Bell Curve Appears Everywhere

The **Central Limit Theorem (CLT)** is the *bedrock* of statistical inference and hypothesis testing.

☑ 1. What is the Central Limit Theorem?

When you take **many random samples** from *any* population (regardless of its shape), the **distribution of sample means** will:

- Tend toward a **Normal Distribution**
- Have a **mean = population mean (μ)**
- Have a **standard deviation = σ / √n** (called Standard Error)

As n (sample size) increases, normality improves.

☑ 2. Why It's Powerful

Even if your population is:

- Skewed
- Bimodal
- Non-normal

👉 The **mean of large enough samples** will still follow a **normal distribution**.
This allows:

- Hypothesis testing
- Confidence intervals
- Z-tests

☑ 3. Visualize CLT in Python

```
import numpy as np
import matplotlib.pyplot as plt

population = np.random.exponential(scale=2, size=10000)
sample_means = []

for _ in range(1000):
    sample = np.random.choice(population, size=50)
    sample_means.append(np.mean(sample))

plt.hist(sample_means, bins=30, density=True)
plt.title("Sampling Distribution of the Mean")
```

```
plt.xlabel("Sample Mean")
plt.ylabel("Frequency")
plt.grid(True)
plt.show()
```

💡 Even though the **population is exponential** (skewed), the **sample means form a bell curve**!

✅ 4. Formula Recap

- Mean of sampling distribution:

$$\mu_{\bar{x}} = \mu$$

- Standard deviation of sampling distribution (Standard Error):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

✅ 5. Real-World Examples

Scenario	CLT Usage
Estimating average delivery time	Use CLT to construct confidence intervals
Polling 1000 people about voting	CLT tells us the sampling distribution is normal
Quality control in manufacturing	Sample averages help detect shifts in process

💬 INTERVIEW QUESTIONS (Medium to High)

🔸 **Q1:** Why is CLT important even when data isn't normally distributed?

✅ Because it allows us to use normal-based methods on non-normal populations.

🔸 **Q2:** How large should the sample size be?

✅ Generally **n ≥ 30** is considered sufficient, but depends on skewness.

🔸 **Q3:** What is Standard Error, and how is it related to CLT?

✅ It's the standard deviation of sample means:

$$SE = \frac{\sigma}{\sqrt{n}}$$

🔸 **Q4:** Can CLT be applied to medians or variances?

✅ CLT applies *directly* to sample means, not necessarily to medians/variances.

◇ Day 7: Hypothesis Testing – From Assumptions to Conclusions

Hypothesis Testing helps you make decisions based on data — whether you're testing marketing strategies, A/B experiments, or scientific claims.

☑ 1. What is Hypothesis Testing?

It's a statistical method to **test an assumption** (claim) about a population using sample data.

- **Null Hypothesis (H_0)**: Status quo / no effect
- **Alternative Hypothesis (H_1)**: Your claim / there is an effect

Example:

H_0: The average delivery time = 2 days

H_1: The average delivery time ≠ 2 days

☑ 2. Steps in Hypothesis Testing

1. **Define H_0 and H_1**
2. **Select significance level (α)** – typically 0.05
3. **Choose test type (z-test, t-test, etc.)**
4. **Calculate test statistic**
5. **Compare with critical value or p-value**
6. **Make decision: Reject or fail to reject H_0**

☑ 3. Types of Errors

Type	Meaning
Type I Error	Rejecting H_0 when it's actually true (False Positive)
Type II Error	Failing to reject H_0 when it's false (False Negative)

☑ 4. One-Tailed vs Two-Tailed Test

- **One-tailed**: Testing in one direction (e.g., H_1: $\mu > 2$)
- **Two-tailed**: Testing both directions (e.g., H_1: $\mu \neq 2$)

☑ 5. Example in Python: One Sample t-test

```
from scipy import stats
import numpy as np


data = np.random.normal(loc=2.1, scale=0.5, size=30)

t_stat, p_val = stats.ttest_1samp(data, popmean=2)

print("t-statistic:", t_stat)
print("p-value:", p_val)

if p_val < 0.05:
```

```
    print("Reject H₀")
else:
    print("Fail to reject H₀")
```

☑ 6. Common Tests

Test	Use Case
Z-test	Known population std dev, n > 30
T-test	Unknown std dev, small samples
Chi-square test	Categorical variables
ANOVA	Comparing >2 groups
Proportion test	Proportion-based hypotheses

💬 INTERVIEW QUESTIONS (Medium to High)

◇ **Q1:** What does p-value mean?

☑ Probability of observing the test results under H_0. Low p-value ⇒ unlikely under H_0 ⇒ reject it.

◇ **Q2:** How do you interpret a t-test result?

☑ If $p < \alpha$, reject H_0 ⇒ the sample mean significantly differs from the population mean.

◇ **Q3:** When to use z-test vs t-test?

☑ Use **z-test** when population standard deviation is known; **t-test** otherwise.

◇ **Q4:** What's the difference between one-tailed and two-tailed tests?

☑ One-tailed tests check for deviation in one direction; two-tailed in both.

◇ Day 8: T-Test Types – One-Sample, Independent, Paired

The **T-test** is a fundamental statistical tool used to compare **means** and check if differences are statistically significant. It's essential in A/B testing, medical trials, and user behavior analysis.

☑ 1. When to Use T-Test?

- Population standard deviation is **unknown**
- Sample size is **small** (typically n < 30)
- Data is **approximately normally distributed**

☑ 2. One-Sample T-Test

Checks if the **sample mean** is significantly different from a **known value** (usually the population mean).

```
from scipy.stats import ttest_1samp
import numpy as np

data = np.array([22, 21, 23, 20, 24])
t_stat, p_val = ttest_1samp(data, popmean=21)
print("T-Statistic:", t_stat)
print("P-Value:", p_val)
```

☑ 3. Independent Two-Sample T-Test

Checks if **two independent groups** have significantly different means.
Use case: Comparing conversion rates of **Group A vs Group B**

```
from scipy.stats import ttest_ind

group_a = np.random.normal(50, 5, 30)
```

```
group_b = np.random.normal(52, 5, 30)

t_stat, p_val = ttest_ind(group_a, group_b)
print("T-Statistic:", t_stat)
print("P-Value:", p_val)
```

☑ 4. Paired T-Test (Dependent Samples)

Used when **same subjects** are tested **before and after** a treatment.
Use case: Pre-test vs Post-test scores

```
from scipy.stats import ttest_rel

before = np.array([70, 68, 72, 71, 69])
after = np.array([75, 70, 73, 74, 72])

t_stat, p_val = ttest_rel(before, after)
print("T-Statistic:", t_stat)
print("P-Value:", p_val)
```

☑ 5. Assumptions of T-Test

- Data is continuous and normally distributed
- Observations are independent
- Equal variances (in some cases)

💬 INTERVIEW QUESTIONS (Medium to High)

◇ **Q1:** What is the main difference between Independent and Paired T-test?
☑ Independent t-test compares two different groups; paired t-test compares the same group twice.

◇ **Q2:** What are the assumptions of a T-test?
☑ Normality, independence, and equal variances (for independent t-test).

◇ **Q3:** When should you not use a T-test?
☑ When data is not normally distributed or when using ordinal/categorical data.

◇ **Q4:** Why is the T-distribution used instead of the normal distribution?
☑ It adjusts for small sample sizes by having heavier tails, providing more conservative results.

◇ Day 9: Z-Test vs T-Test vs ANOVA – Know When to Use What

Confused between **Z-Test**, **T-Test**, and **ANOVA**? You're not alone. These statistical tests help determine if group differences are real or due to chance—but each has its **own use-case** and **assumptions**.

☑ 1. Z-Test

Used when:

- Population standard deviation is **known**
- Sample size is **large** ($n \geq 30$)
- Data is **normally distributed**

Use case: Testing population mean with large sample

```
from statsmodels.stats.weightstats import ztest
import numpy as np

data = np.random.normal(loc=50, scale=10, size=100)
z_stat, p_val = ztest(data, value=52)
```

```
print("Z-Statistic:", z_stat)
print("P-Value:", p_val)
```

☑ 2. T-Test

Used when:

- Population standard deviation is **unknown**
- Sample size is **small** (n < 30)
- Data is **approximately normal**

We've already covered:

- One-Sample T-Test
- Independent Two-Sample T-Test
- Paired T-Test

☑ 3. ANOVA (Analysis of Variance)

Used to compare **means of 3 or more groups**.

Use case: Comparing test scores across 3 departments

```
from scipy.stats import f_oneway

group1 = [88, 92, 85, 90]
group2 = [78, 79, 74, 77]
group3 = [92, 94, 96, 91]

f_stat, p_val = f_oneway(group1, group2, group3)
print("F-Statistic:", f_stat)
print("P-Value:", p_val)
```

☑ 4. Key Differences

Feature	Z-Test	T-Test	ANOVA
Groups Compared	1 or 2	1 or 2	3 or more
Sample Size	Large (n≥30)	Small (n<30)	Any
Std Dev Known?	Yes	No	No
Output	Z-statistic	T-statistic	F-statistic

💬 INTERVIEW QUESTIONS (Medium to High)

◇ **Q1:** When should you use a Z-Test instead of a T-Test?
☑ When the population standard deviation is known and sample size is large.

◇ **Q2:** Why is ANOVA better than multiple T-Tests?
☑ Reduces Type I error; multiple t-tests increase false positives.

◇ **Q3:** What's the null hypothesis in ANOVA?
☑ All group means are equal.

◇ **Q4:** What do you do after a significant ANOVA result?
☑ Perform **post-hoc tests** like Tukey's HSD to identify which groups differ.

◇ **Q5:** Can ANOVA be used for 2 groups?
☑ Yes, but it's equivalent to a t-test in that case.

◇ Day 10: Correlation vs Covariance – Know the Real Difference

Both **correlation** and **covariance** measure how two variables move together—but they are **not the same**. Understanding the difference is key for data interpretation and choosing the right statistical tools.

☑ 1. Covariance – The Raw Relationship

Covariance tells you **how two variables vary together**, but not how strong the relationship is.

- **Positive covariance** → variables move in the **same** direction
- **Negative covariance** → variables move in **opposite** directions
- **Magnitude** is not standardized (depends on units)

```
import numpy as np

x = [1, 2, 3, 4, 5]
y = [2, 4, 6, 8, 10]

cov_matrix = np.cov(x, y)
print("Covariance Matrix:\n", cov_matrix)
```

☑ 2. Correlation – The Standardized Relationship

Correlation measures the **strength and direction** of the linear relationship.

- Always between **-1 and 1**
- **+1:** Perfect positive linear relation
- **-1:** Perfect negative linear relation
- **0:** No linear relation

```
correlation = np.corrcoef(x, y)
print("Correlation Matrix:\n", correlation)
```

☑ 3. Key Differences Table

Feature	Covariance	Correlation
Range	No fixed range	[-1, 1]
Unit-dependent	Yes	No
Interpretation	Direction only	Direction + Strength
Standardized?	No	Yes
Use Cases	PCA, Portfolio Analysis	Feature selection, EDA

☑ 4. When to Use What?

Use Case	Use This
To measure pure directional movement	Covariance
To understand strength of relation	Correlation
In PCA (Principal Component Analysis)	Covariance
In EDA & Feature Selection	Correlation

💬 INTERVIEW QUESTIONS (Easy to Medium)

◇ **Q1:** Can two variables have high covariance but low correlation?
☑ Yes, if the units or scales are different, correlation may still be low.

◇ **Q2:** Why do we prefer correlation in EDA?
☑ It standardizes values, making it easier to compare relationships.

◇ **Q3:** What happens if correlation is 0?
☑ There's **no linear** relationship, but **non-linear** may still exist.

◇ **Q4:** Is correlation always causation?
☑ Nope! Strong correlation ≠ cause-effect relationship.

◇ Day 11: Chi-Square Test – For Categorical Data Only

The **Chi-Square Test (χ²)** is a powerful statistical method used to determine if there's a **significant association between categorical variables**. It's your go-to test when analyzing survey results, demographics, marketing funnels, and more!

☑ 1. What is the Chi-Square Test?

- Tests **independence** between two categorical variables
- Compares **observed frequencies** to **expected frequencies**

☑ 2. Types of Chi-Square Tests

Test Type	Use Case
Chi-Square of Independence	Are two categorical variables related?
Chi-Square Goodness of Fit	Does a sample match a population?

☑ 3. Formula

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- O = Observed value
- E = Expected value

☑ 4. Using Chi-Square in Python (SciPy)

```
import pandas as pd
from scipy.stats import chi2_contingency
```

```
# Sample contingency table
data = [[30, 10],
        [20, 40]]

chi2, p, dof, expected = chi2_contingency(data)

print("Chi-Square Value:", chi2)
print("P-Value:", p)
print("Degrees of Freedom:", dof)
print("Expected Frequencies:\n", expected)
```

☑ 5. Interpretation

- **p < 0.05** → Reject null hypothesis → Significant association
- **p ≥ 0.05** → Fail to reject → No significant association

📌 Applications

- Is **gender** related to **product preference**?
- Are **age groups** independent of **newsletter signup**?
- Do **regions** and **voting patterns** relate?

💬 INTERVIEW QUESTIONS (Medium Level)

◇ **Q1:** Can Chi-Square be used for numerical data?
☑ No. It's strictly for **categorical** data.

◇ **Q2:** What are the assumptions of Chi-Square?
☑ Observations should be **independent**, and expected frequency should be **≥ 5** in each cell (preferably).

◇ **Q3:** What's the role of degrees of freedom?
☑ It determines the distribution curve and affects p-value calculation.

◇ **Q4:** What if the expected frequency is < 5 in many cells?
☑ Use **Fisher's Exact Test** instead.

◇ Day 12: Central Limit Theorem (CLT) – The Heart of Statistics 🫀

The **Central Limit Theorem (CLT)** is one of the most fundamental ideas in statistics and data science. It explains **why normal distribution appears so often**, even when the data itself isn't normal!

☑ 1. What is the Central Limit Theorem?

No matter the original distribution of the data, the **sampling distribution of the sample mean** will approach a **normal distribution** as the sample size becomes large (typically $n \geq 30$).

☑ 2. Why is CLT so Important?

- It allows **parametric tests** (like Z-tests and t-tests) even on non-normal data
- Powers **confidence intervals** and **hypothesis testing**
- Makes large-sample inference possible!

☑ 3. CLT in Action: Python Example

```
import numpy as np
import matplotlib.pyplot as plt

# Generate skewed data (exponential)
data = np.random.exponential(scale=2, size=10000)
```

```
sample_means = []

# Take 1000 samples of size 50 each
for _ in range(1000):
    sample = np.random.choice(data, size=50)
    sample_means.append(np.mean(sample))

# Plot the distribution of sample means
plt.hist(sample_means, bins=30, color='skyblue', edgecolor='black')
plt.title("Sampling Distribution Approaches Normality")
plt.xlabel("Sample Mean")
plt.ylabel("Frequency")
plt.show()
```

☑ 4. Key Terms

Term	Meaning
Population	Entire group
Sample	Subset of the population
Sampling Distribution	Distribution of sample statistics

☑ 5. CLT Assumptions

- Samples are **random and independent**
- Sample size **n ≥ 30** is usually sufficient
- Population must have **finite variance**

📌 Applications of CLT

- A/B testing
- Quality control in manufacturing
- Estimating population mean from sample
- Predictive analytics in ML workflows

💬 INTERVIEW QUESTIONS (Medium-High)

◇ **Q1:** Why is the Central Limit Theorem important for hypothesis testing?
☑ Because it allows us to use the **normal distribution** even when data isn't normal!

◇ **Q2:** What's the difference between population distribution and sampling distribution?
☑ Population: Original data distribution.
☑ Sampling: Distribution of sample means.

◇ **Q3:** Does CLT apply to the median or standard deviation?
☑ It primarily applies to **sample means**.

◇ **Q4:** What if the population is already normal?
☑ Then the sampling distribution of the mean is **also normal** for **any sample size**.

◇ Day 13: t-Tests – One Sample, Two Sample, Paired t-Test

The **t-Test** is your go-to method when comparing **means**—whether it's comparing a group to a benchmark, or comparing two related/unrelated groups.

☑ 1. What is a t-Test?

A **t-Test** determines whether the **difference in means is statistically significant** when the sample size is small and the population standard deviation is unknown.

☑ 2. Types of t-Tests

Test Type	Use Case
One Sample t-Test	Compare **one group mean** to a known value
Two Sample (Independent) t-Test	Compare **two independent group means** (e.g., A/B testing)
Paired Sample t-Test	Compare **two related groups** (e.g., before vs. after)

☑ 3. Assumptions of t-Test

- Data is **normally distributed** (CLT helps here!)
- Observations are **independent**
- Variances are **equal** (for two-sample t-test)
- Data is measured on **interval/ratio** scale

☑ 4. Python Examples

📌 One-Sample t-Test

```
from scipy import stats
import numpy as np

sample = [23, 25, 21, 22, 24, 26]
t_stat, p_val = stats.ttest_1samp(sample, popmean=20)
print("t-statistic:", t_stat, "p-value:", p_val)
```

📌 Two-Sample (Independent) t-Test

```
group1 = [30, 32, 29, 35, 28]
group2 = [25, 26, 27, 30, 22]

t_stat, p_val = stats.ttest_ind(group1, group2)
print("t-statistic:", t_stat, "p-value:", p_val)
```

📌 Paired t-Test

```
before = [80, 85, 88, 90]
after = [82, 87, 89, 91]

t_stat, p_val = stats.ttest_rel(before, after)
print("t-statistic:", t_stat, "p-value:", p_val)
```

☑ 5. Interpreting p-Value

p-value	Interpretation
$p < 0.05$	Reject the null hypothesis (significant)
$p \geq 0.05$	Fail to reject the null (not significant)

📌 Real-World Applications

- **A/B Testing** (website layout A vs. layout B)
- **Medical Studies** (treatment vs. control)
- **Before/After Improvements** (new teaching method vs. old)

💬 INTERVIEW QUESTIONS (Medium to High)

◇ **Q1:** What is the difference between independent and paired t-tests?

☑ Independent: different groups; Paired: same group at two times.

◇ **Q2:** When can you not use a t-test?

☑ When assumptions (normality, independence, equal variance) are violated.

◇ **Q3:** What's the null hypothesis in a t-test?

☑ That the **means are equal** (no significant difference).

◇ **Q4:** Why use t-distribution instead of normal?
☑ t-distribution accounts for **extra uncertainty in small samples**.

◇ Day 14: ANOVA – Analysis of Variance

ANOVA (Analysis of Variance) is your go-to test when you want to compare the **means of 3 or more groups** at once instead of doing multiple t-tests.

☑ 1. What is ANOVA?

ANOVA helps us determine whether the **mean differences across multiple groups** are statistically significant.
It avoids the risk of **Type I error** that comes from performing multiple t-tests.

☑ 2. Types of ANOVA

Type	Use Case
One-Way ANOVA	1 categorical IV (grouping) → 1 continuous DV (outcome)
Two-Way ANOVA	2 categorical IVs → 1 continuous DV
Repeated Measures	Same subjects measured multiple times (like paired t-test)

☑ 3. Assumptions of ANOVA

- Groups are **independent**
- **Normal distribution** of the outcome variable
- **Equal variances** across groups (homogeneity of variances)
- Outcome variable is **continuous**

☑ 4. Python Example: One-Way ANOVA

```
from scipy.stats import f_oneway

group_A = [25, 30, 28, 35, 33]
group_B = [22, 29, 27, 24, 26]
group_C = [35, 40, 38, 37, 42]

f_stat, p_val = f_oneway(group_A, group_B, group_C)
print("F-statistic:", f_stat, "p-value:", p_val)
```

☑ 5. Interpreting Results

- **F-statistic**: Ratio of between-group variance to within-group variance
- **p-value**:
  - **p < 0.05** → At least **one group mean** is significantly different
  - **p ≥ 0.05** → No significant difference between group means

---

📌 Post-hoc Tests (If ANOVA is Significant)

Use **Tukey's HSD** or **Bonferroni** test to find **which pairs of groups** differ significantly.

📊 Real-World Applications

- Comparing **test scores** across 3 different teaching methods
- Measuring **user engagement** across 3 website layouts
- Comparing **sales** performance in different regions

💬 INTERVIEW QUESTIONS (Medium to High)

◇ **Q1:** Why not just use multiple t-tests instead of ANOVA?
☑ Multiple t-tests increase the **chance of Type I error**.

◇ **Q2:** What does a high F-statistic mean?
☑ Greater variation **between group means** compared to **within groups**.

◇ **Q3:** When to use ANOVA over t-test?
☑ When comparing **3 or more group means**.

◇ **Q4:** What should you do after a significant ANOVA result?
☑ Perform a **post-hoc test** to identify which groups differ.

◇ Day 15: Chi-Square Test – Categorical Data Analysis

The **Chi-Square (χ^2) Test** helps us analyze **relationships between categorical variables**. It's widely used in surveys, marketing, healthcare, and more!

☑ 1. What is Chi-Square Test?

It checks whether two **categorical variables** are **independent** or **associated**.

📌 Example:
Does gender influence product preference?

☑ 2. Types of Chi-Square Tests

Type	Use Case
Goodness of Fit	One categorical variable vs expected distribution
Test of Independence	Two categorical variables (e.g., Gender vs Choice)

☑ 3. Assumptions of Chi-Square

- Data should be **frequencies** (counts), not percentages
- Categories must be **mutually exclusive**
- Expected frequency in each cell should be **≥ 5** for validity
- Observations are **independent**

☑ 4. Python Example: Test of Independence

```
import pandas as pd
from scipy.stats import chi2_contingency

# Contingency table
data = pd.DataFrame({
    'Product A': [30, 10],
    'Product B': [20, 40]
}, index=['Male', 'Female'])

chi2, p, dof, expected = chi2_contingency(data)
print("Chi2 Statistic:", chi2)
print("p-value:", p)
print("Degrees of Freedom:", dof)
print("Expected Frequencies:\n", expected)
```

☑ 5. Interpreting Results

- **Chi² value**: Larger = more difference between observed and expected
- **p-value**:

- $p < 0.05$ → Variables are likely **associated**
- $p \geq 0.05$ → Variables are likely **independent**

🔲 Real-World Applications

- Is **gender** associated with product preference?
- Does **education level** impact voting behavior?
- Do different **ads** get clicked by different age groups?

💬 INTERVIEW QUESTIONS (Medium to High)

◇ **Q1:** Can Chi-Square test be used on numerical data?
☑ No, it's only for **categorical data**.

◇ **Q2:** What's the difference between Chi-Square Goodness-of-Fit and Test of Independence?
☑ Goodness-of-Fit → One variable; Independence → Two variables.

◇ **Q3:** Why should expected frequency be ≥ 5?
☑ To ensure **accuracy** of the Chi-Square approximation.

◇ **Q4:** What does a low p-value mean in Chi-Square?
☑ There's a **statistical association** between the variables.

👏 Congratulations! You've completed the 15-Day Statistics Crash Course!

📚 You now have hands-on knowledge of:

- Descriptive & Inferential Stats
- Central Tendency & Dispersion
- Distributions
- Hypothesis Testing
- ANOVA & Chi-Square

20 Medium-level Statistics Interview Questions

📊 1. What is the Central Limit Theorem and why is it important?
The Central Limit Theorem states that the sampling distribution of the sample mean approaches a normal distribution as the sample size becomes large, regardless of the population's distribution. This is crucial because it allows for the use of inferential statistics, such as confidence intervals and hypothesis tests, even when the population distribution is unknown.

📈 2. Explain the difference between Type I and Type II errors.
- **Type I Error (False Positive):** Rejecting the null hypothesis when it is actually true.
- **Type II Error (False Negative):** Failing to reject the null hypothesis when it is actually false.

Understanding these errors is vital for evaluating the reliability of statistical tests.

📌 3. How do you interpret a p-value?
A p-value represents the probability of obtaining a test statistic at least as extreme as the one observed, assuming the null hypothesis is true. A smaller p-value indicates stronger evidence against the null hypothesis. Typically, a p-value less than 0.05 is considered statistically significant. [Learn R, Python & Data Science Online+1DataLemur+1]

📊 4. What is the difference between a t-test and a z-test?
- **Z-test:** Used when the population variance is known and the sample size is large ($n > 30$).
- **T-test:** Used when the population variance is unknown and/or the sample size is small ($n \leq 30$).

Both tests assess whether there is a significant difference between sample means.

📉 5. Describe the assumptions of linear regression.

1. **Linearity:** The relationship between independent and dependent variables is linear.
2. **Independence:** Observations are independent of each other.
3. **Homoscedasticity:** Constant variance of errors.
4. **Normality:** The residuals (errors) are normally distributed.
5. **No multicollinearity:** Independent variables are not highly correlated with each other.

📈 6. What is multicollinearity and how can it be detected?

Multicollinearity occurs when independent variables in a regression model are highly correlated, leading to unreliable coefficient estimates. It can be detected using:

- **Variance Inflation Factor (VIF):** Values greater than 5 or 10 indicate multicollinearity.
- **Correlation Matrix:** High correlation coefficients between variables suggest multicollinearity.

📊 7. Explain the difference between correlation and causation.

- **Correlation:** Measures the strength and direction of a linear relationship between two variables.
- **Causation:** Implies that one variable directly affects another.

Correlation does not imply causation; two variables may be correlated due to coincidence or a third variable.

📊 8. What is the purpose of an ANOVA test?

ANOVA (Analysis of Variance) tests whether there are statistically significant differences between the means of three or more independent groups. It analyzes the variance within groups and

between groups to determine if at least one group mean is different.

📈 9. How do you interpret R-squared and adjusted R-squared in regression analysis?

- **R-squared:** Represents the proportion of variance in the dependent variable explained by the independent variables.
- **Adjusted R-squared:** Adjusts R-squared for the number of predictors in the model, providing a more accurate measure when comparing models with different numbers of predictors.

🧪 10. What is a p-value and how is it used in hypothesis testing?

A p-value is the probability of obtaining test results at least as extreme as the observed results, assuming the null hypothesis is true. In hypothesis testing, if the p-value is less than the chosen significance level (e.g., 0.05), the null hypothesis is rejected.Learn R, Python & Data Science Online

---

📊 11. Describe the difference between a one-tailed and two-tailed test.

- **One-tailed test:** Tests for the possibility of the relationship in one direction.
- **Two-tailed test:** Tests for the possibility of the relationship in both directions.Data Interview

The choice depends on the research hypothesis.

📈 12. What is the difference between parametric and non-parametric tests?

- **Parametric tests:** Assume underlying statistical distributions (e.g., normal distribution).

- **Non-parametric tests:** Do not assume specific distributions and are used when data doesn't meet parametric test assumptions.

🧮 13. Explain the concept of statistical power.

Statistical power is the probability that a test correctly rejects a false null hypothesis (i.e., detects an effect when there is one). Higher power reduces the risk of Type II errors.

📊 14. What is the difference between a population and a sample?

- **Population:** The entire group being studied.
- **Sample:** A subset of the population used to make inferences about the whole.Data Interview

---

📈 15. Describe the difference between descriptive and inferential statistics.

- **Descriptive statistics:** Summarize and describe the features of a dataset.
- **Inferential statistics:** Make predictions or inferences about a population based on a sample.

🧪 16. What is a confidence interval?

A confidence interval is a range of values, derived from the sample data, that is likely to contain the value of an unknown population parameter. For example, a 95% confidence interval suggests that there is a 95% chance the interval contains the true parameter.

📊 17. What is stratified sampling, and when is it used?

Stratified sampling involves dividing the population into distinct subgroups (strata) based on shared characteristics and then sampling from each stratum proportionally. This method ensures representation from all subgroups, leading to more precise estimates, especially when there are significant differences between strata. DataLemur+1Learn R, Python & Data Science Online+1

📈 18. Explain the difference between a population parameter and a sample statistic.

- **Population Parameter:** A value that describes a characteristic of the entire population (e.g., population mean μ).
- **Sample Statistic:** A value that describes a characteristic of a sample drawn from the population (e.g., sample mean $\bar{x}$).

Sample statistics are used to estimate population parameters.Learn R, Python & Data Science Online+1GitHub+1

🧮 19. What is the law of large numbers?

The law of large numbers states that as the size of a sample increases, the sample mean will get closer to the population mean. This principle underpins many statistical practices, ensuring that larger samples provide more accurate estimates.

📊 20. How do you determine if a dataset is normally distributed?

To assess normality:

- **Visual Methods:**
  - **Histogram:** Should resemble a bell-shaped curve.
  - **Q-Q Plot:** Data points should lie approximately along the reference line.Learn R, Python & Data Science Online

- **Statistical Tests:**
  - **Shapiro-Wilk Test:** Tests the null hypothesis that the data is normally distributed.
  - **Kolmogorov-Smirnov Test:** Compares the sample distribution with a normal distribution.

A combination of visual and statistical methods provides a comprehensive assessment of normality.

20 Hard-level Statistics Interview Questions

1. What is the difference between Maximum Likelihood Estimation (MLE) and Bayesian Estimation?
MLE seeks parameter values that maximize the likelihood function, relying solely on observed data. **Bayesian Estimation** incorporates prior beliefs and updates them with observed data to form a posterior distribution.

---

2. Explain the concept of the Bias-Variance Tradeoff.

The Bias-Variance Tradeoff describes the balance between:Data Interview

- **Bias**: Error from erroneous assumptions in the learning algorithm.
- **Variance**: Error from sensitivity to small fluctuations in the training set.GitHub+3Learn R, Python & Data Science Online+3Data Interview+3

High bias can cause underfitting, while high variance can cause overfitting. The goal is to find a model with optimal complexity to minimize total error.

---

3. What is the purpose of regularization in regression models?
Regularization adds a penalty term to the loss function to prevent overfitting by discouraging complex models:

- **Lasso Regression (L1)**: Encourages sparsity, potentially reducing some coefficients to zero.

- **Ridge Regression (L2)**: Shrinks coefficients towards zero but doesn't eliminate them.

---

4. How do you interpret the Area Under the ROC Curve (AUC-ROC)?
AUC-ROC measures a classifier's ability to distinguish between classes:
- **AUC = 1**: Perfect classification.
- **AUC = 0.5**: No discriminative ability (equivalent to random guessing).

Higher AUC indicates better model performance.

---

5. What is the Central Limit Theorem and why is it important?
The Central Limit Theorem states that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the population's distribution. This is crucial for making inferences about population parameters.

---

6. Explain the concept of multicollinearity and how to detect it.
Multicollinearity occurs when independent variables in a regression model are highly correlated, leading to unreliable coefficient estimates. It can be detected using:
- **Variance Inflation Factor (VIF)**: Values above 5 or 10 indicate high multicollinearity.
- **Correlation Matrix**: High correlation coefficients between variables suggest multicollinearity.

---

7. What is the difference between parametric and non-parametric tests?
- **Parametric Tests**: Assume underlying statistical distributions (e.g., t-test assumes normality).
- **Non-Parametric Tests**: Make fewer assumptions about data distribution (e.g., Mann-Whitney U test).

Non-parametric tests are useful when data doesn't meet parametric test assumptions.

---

8. Describe the concept of overfitting and how to prevent it.
Overfitting occurs when a model learns noise in the training data, performing well on training data but poorly on unseen data. Prevention techniques include:

- **Cross-validation**: Ensures model generalization.
- **Regularization**: Adds penalty terms to discourage complexity.
- **Pruning**: Reduces complexity in decision trees.
- **Early Stopping**: Halts training when performance on validation set deteriorates.

---

9. What is the difference between Type I and Type II errors?

- **Type I Error (False Positive)**: Rejecting a true null hypothesis.
- **Type II Error (False Negative)**: Failing to reject a false null hypothesis.DataLemur+7LinkedIn+7Learn R, Python & Data Science Online+7

Balancing these errors is crucial in hypothesis testing.

---

10. How do you handle missing data in a dataset?
Strategies include:

- **Deletion**: Removing records with missing values (listwise or pairwise).
- **Imputation**: Filling in missing values using methods like mean, median, mode, or predictive models.
- **Model-Based Methods**: Using algorithms that handle missing data inherently.

The choice depends on the nature and extent of missingness.

---

11. Explain the concept of statistical power.
Statistical power is the probability of correctly rejecting a false null hypothesis (i.e., detecting an effect when it exists). Higher power reduces the risk of Type II errors. Factors influencing power include sample size, effect size, significance level, and variability.

---

12. What is the purpose of cross-validation in model evaluation?
Cross-validation assesses a model's ability to generalize to unseen data by partitioning the data into training and validation sets multiple times. Common methods include k-fold and leave-one-out cross-validation. It helps in selecting models and tuning hyperparameters.

---

13. Describe the differences between bagging and boosting.

- **Bagging (Bootstrap Aggregating)**: Builds multiple independent models on random subsets of data and aggregates their predictions (e.g., Random Forest).
- **Boosting**: Builds models sequentially, each trying to correct the errors of the previous one (e.g., AdaBoost, Gradient Boosting).

Bagging reduces variance, while boosting reduces bias.

---

14. What is Principal Component Analysis (PCA) and when is it used?
PCA is a dimensionality reduction technique that transforms correlated variables into a set of uncorrelated components, ordered by the amount of variance they capture. It's used to simplify datasets, reduce noise, and prevent overfitting.