# Sampling Techniques: Comprehensive Viva Guide

Quick Reference for Oral Examination

## 1 Definitions and Core Concepts

### Random Sampling

**Definition:** A sampling method where each element in the population has an *equal and independent* probability of being selected.
**Key Properties:**

- Probability of selection: $P(x_i) = \frac{1}{N}$ for population size $N$
- Each selection is independent of others
- Simple Random Sampling (SRS) is the foundation of statistical inference

### Stratified Sampling

**Definition:** The population is divided into homogeneous subgroups (strata) based on a characteristic, and random samples are drawn from each stratum.
**Key Properties:**

- Population $= \bigcup_{h=1}^{L} S_h$ where $S_h$ are disjoint strata
- Sample from each stratum: $n_h$ from stratum $h$
- Total sample size: $n = \sum_{h=1}^{L} n_h$
- Ensures representation from all subgroups

### Hybrid Sampling (CRITICAL)

**Definition:** A two-stage approach combining stratified sampling with random sampling within strata. First stratify, then apply different sampling rates or methods per stratum.
**Key Characteristics:**

- Combines advantages of both stratification and random sampling
- Allocation can be proportional or optimal
- Reduces variance compared to simple random sampling
- Flexible: can oversample rare strata

**Stratified Clustering**

**Definition:** A multi-stage technique where the population is first divided into strata, then clusters are formed within strata, and samples are drawn from selected clusters.
**Structure:**

- Stage 1: Divide population into strata
- Stage 2: Identify clusters within each stratum
- Stage 3: Sample clusters, then sample elements within clusters
- Useful for geographically dispersed populations

# 2 Mathematical Foundations

## 2.1 Random Sampling

**Sample Mean Estimator:**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

**Variance of Sample Mean:**

$$\text{Var}(\bar{x}) = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right) = \frac{\sigma^2}{n} \cdot \text{FPC}$$

where FPC = finite population correction = $(1 - \frac{n}{N})$

**Standard Error:**

$$SE(\bar{x}) = \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}$$

## 2.2 Stratified Sampling

**Stratified Mean Estimator:**

$$\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h$$

where $W_h = \frac{N_h}{N}$ is the weight of stratum $h$, and $\bar{x}_h$ is the sample mean in stratum $h$.

**Variance of Stratified Mean:**

$$\text{Var}(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)$$

**Allocation Methods:**

- *Proportional:* $n_h = n \cdot W_h = n \cdot \frac{N_h}{N}$
- *Optimal (Neyman):* $n_h = n \cdot \frac{N_h \sigma_h}{\sum_{k=1}^{L} N_k \sigma_k}$
- Optimal allocation minimizes variance for fixed cost

## 2.3 Hybrid Sampling

Hybrid sampling variance depends on the specific combination used. For stratified random sampling hybrid:

$$\text{Var}(\bar{x}_{hybrid}) \leq \text{Var}(\bar{x}_{SRS})$$

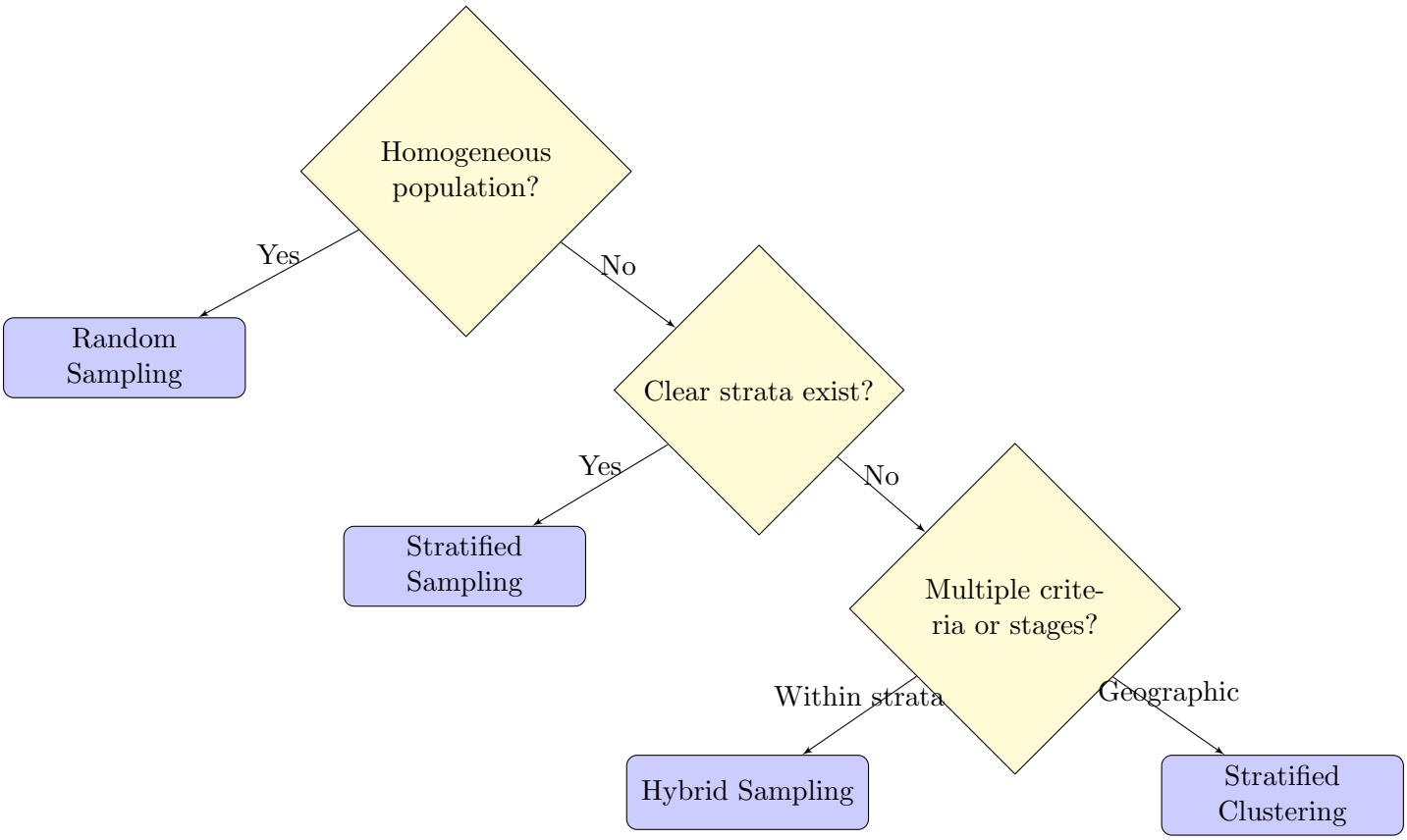
The reduction factor depends on stratum homogeneity and allocation efficiency.

# 3 Practical Understanding

## 3.1 When to Use Each Technique

| Technique | Best Used When | Example |
|---|---|---|
| Random Sampling | Homogeneous population, no prior knowledge | Selecting students from a single class |
| Stratified | Known heterogeneous subgroups, want precision | Income survey across age groups |
| Hybrid | Need flexibility, multiple characteristics | Customer satisfaction with regional and age stratification |
| Stratified Clustering | Large geographic area, cost constraints | National health survey by state then city |

## 3.2 Decision Flowchart




# 4 Common Viva Questions with Detailed Answers

### Q1: What is the main advantage of stratified sampling over random sampling?

**Answer:** Stratified sampling provides *greater precision* (lower variance) than simple random sampling for the same sample size. This occurs because:

- Within-stratum variance is smaller than population variance
- We ensure representation from all subgroups
- Mathematical proof: $\text{Var}(\bar{x}_{st}) \leq \text{Var}(\bar{x}_{SRS})$ when strata are homogeneous
- The gain is maximum when strata are internally homogeneous but differ from each other

### Q2: Explain hybrid sampling and why it's useful.

**Answer:** Hybrid sampling combines multiple sampling techniques, typically stratified sampling followed by different methods within strata.

**Key benefits:**

- *Flexibility:* Can use different sampling rates for different strata
- *Efficiency:* Oversample rare but important subgroups
- *Cost-effective:* Apply expensive methods only where needed
- *Precision:* Maintains stratification benefits while adapting to constraints

**Example:** In a national survey, stratify by region, then use cluster sampling in rural areas (cost-effective) and simple random sampling in urban areas (more precise).

### Q3: How does stratified clustering differ from stratified sampling?

**Answer:**

- **Stratified Sampling:** Sample individuals directly from each stratum
- **Stratified Clustering:** Sample clusters from each stratum, then sample individuals from selected clusters

**Key differences:**

1. *Stages:* Stratified clustering is multi-stage; stratified sampling is single-stage
2. *Sampling unit:* Clusters first, then individuals vs. individuals directly
3. *Cost:* Clustering reduces travel/administrative costs
4. *Variance:* Clustering typically increases variance due to intra-cluster correlation

### Q4: What is the finite population correction (FPC) and when is it important?

**Answer:** FPC = $(1 - \frac{n}{N})$ accounts for sampling without replacement from finite populations.

**When important:**

- When sampling fraction $\frac{n}{N} > 0.05$ (5% rule)
- Reduces estimated variance
- As $n \to N$, variance $\to 0$ (census)
- Often ignored for large populations where $\frac{n}{N} \approx 0$

### Q5: Derive or explain the stratified mean estimator.

**Answer:** The stratified mean is a weighted average of stratum means:

$$\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h$$

**Intuition:**

- Each stratum contributes proportionally to its size in the population
- $W_h = \frac{N_h}{N}$ ensures proper weighting
- If we sampled proportionally, $\bar{x}_{st} = \bar{x}$ (equal to simple mean)
- Unbiased: $E(\bar{x}_{st}) = \mu$ (population mean)

### Q6: What is optimal (Neyman) allocation and when should you use it?

**Answer:** Neyman allocation minimizes variance for fixed total sample size:

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum_{k=1}^{L} N_k \sigma_k}$$

**Interpretation:**

- Allocate more samples to larger strata ($N_h$)
- Allocate more samples to more variable strata ($\sigma_h$)
- *Use when:* You know or can estimate stratum variances and want minimum variance
- *Challenge:* Requires prior knowledge of $\sigma_h$

### Q7: Why might stratified clustering be preferred over stratified sampling despite higher variance?

**Answer: Cost and logistics:**

- Dramatically reduces travel and data collection costs
- More practical for geographically dispersed populations
- Easier administrative implementation
- Can sample more elements for the same budget

**Trade-off:** Accept higher variance per unit in exchange for larger affordable sample size, potentially achieving lower overall variance.

### Q8: How do you determine the number of strata in stratified sampling?

**Answer:** Consider:

1. **Heterogeneity:** More strata if population is highly variable
2. **Sample size:** Need sufficient $n_h \geq 2$ in each stratum for variance estimation
3. **Known characteristics:** Use natural divisions (age groups, regions)
4. **Diminishing returns:** Beyond 6-8 strata, gains are often minimal
5. **Practical constraints:** Cost and complexity of managing many strata

**Rule of thumb:** More homogeneous within strata than between strata.

### Q9: Compare proportional vs. optimal allocation.

**Answer:**

**Proportional ($n_h \propto N_h$):**

- Simpler to implement
- Self-weighting: $\bar{x}_{st} = \bar{x}$
- Good when stratum variances are similar
- No prior variance knowledge needed

**Optimal ($n_h \propto N_h \sigma_h$):**

- Minimizes variance
- Better when stratum variances differ substantially
- Requires variance estimates
- More complex weighting in estimation

### Q10: What are the assumptions for valid stratified sampling?

**Answer:**

1. **Disjoint strata:** Each population element belongs to exactly one stratum
2. **Exhaustive:** Strata cover entire population
3. **Known stratum sizes:** $N_h$ must be known for all $h$
4. **Random sampling within strata:** Each element in a stratum has equal probability
5. **Independent selections:** Selections across strata are independent

## 5 Comparison Points

### 5.1 Variance Hierarchy (Best to Worst)

For well-designed surveys with homogeneous strata:

$$\text{Var(Optimal Stratified)} < \text{Var(Proportional Stratified)} < \text{Var(Random)} < \text{Var(Clustering)}$$

### 5.2 Cost Hierarchy (Cheapest to Most Expensive)

$$\text{Clustering} < \text{Stratified Clustering} < \text{Stratified} < \text{Random}$$

**Key insight:** Trade-off between statistical efficiency and practical cost.

## 6 Real-World Applications

1. **Political Polling:** Stratify by state/region, age, gender; hybrid approach with different methods for urban/rural
2. **Quality Control:** Stratify by production line, shift, or product type
3. **Medical Research:** Stratify by age groups, severity of condition; ensures adequate representation
4. **Market Research:** Stratified clustering by geographic region then retail locations
5. **Educational Assessment:** Stratify by school type, grade level; random sampling within strata
6. **Agricultural Surveys:** Stratified clustering by county then farms; accounts for regional variation

# 7 Common Pitfalls and Misconceptions

**Common Mistakes**

1. **Confusing stratified and cluster sampling:**
   - Stratified: Sample from *all* strata
   - Cluster: Sample *some* clusters, then all or sample within
2. **Forgetting weights in stratified estimation:** Must use $W_h$; simple averaging gives wrong answer if allocation isn't proportional
3. **Ignoring FPC:** Can overestimate variance when sampling fraction is substantial
4. **Using too many strata:** Results in small $n_h$, unstable variance estimates
5. **Poor stratification criteria:** Choosing variables unrelated to outcome reduces efficiency gains
6. **Assuming clustering reduces variance:** Clustering typically *increases* variance due to within-cluster homogeneity
7. **Not checking stratum assumptions:** Ensure strata are truly disjoint and exhaustive
8. **Optimal allocation without variance knowledge:** Can't use Neyman allocation without estimating $\sigma_h$

**Key Takeaways**

- **Stratification reduces variance** when strata are homogeneous internally
- **Hybrid sampling offers flexibility** to combine methods optimally
- **Clustering trades precision for cost** in geographically dispersed surveys
- **Stratified clustering combines both approaches** for practical large-scale surveys
- **Choice depends on:** population structure, budget, precision requirements