

UNIT 6

Sample Survey Methods

(10 Hours) — Probability and Statistics (CACS202)

1. Types of Sampling

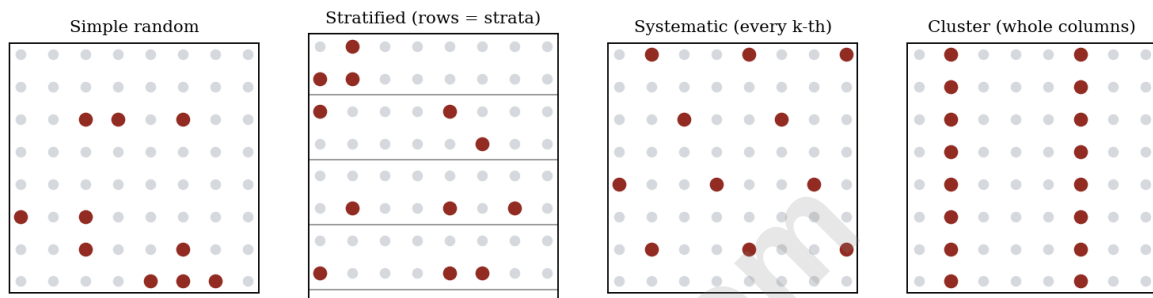


Figure 6.1 — Four common probability sampling methods (red points = selected units)

Probability (random) sampling	Non-probability sampling
Every unit has a known, non-zero chance of selection	Selection is based on judgement or convenience; probabilities unknown
Simple random, stratified, systematic, cluster, multistage, PPS	Convenience, judgement (purposive), quota, snowball sampling
Sampling error can be measured; valid statistical inference	Cannot measure error; results may be biased; used for exploratory studies

Notation: N = population size, n = sample size, $f = n/N$ (sampling fraction), $1 - f$ = finite population correction (fpc), $\bar{Y}$ = population mean, $\bar{y}$ = sample mean, S^2 = population variance (divisor $N - 1$), s^2 = sample variance (divisor $n - 1$), $Y = N\bar{Y}$ = population total.

2. Simple Random Sampling (SRS)

In **simple random sampling** every possible sample of size n has the same chance of being selected (and every unit has equal chance n/N). Units are drawn by lottery or random-number tables. **SRSWOR** (without replacement): a selected unit is not returned; there are ${}^N C_n$ possible samples. **SRSWR** (with replacement): a unit may be selected more than once; there are N^n possible samples.

	SRS without replacement (WOR)	SRS with replacement (WR)
Estimate of mean $\bar{Y}$	$\bar{y} = \Sigma y / n$ (unbiased)	$\bar{y} = \Sigma y / n$ (unbiased)

	SRS without replacement (WOR)	SRS with replacement (WR)
Variance of $\bar{y}$	$V(\bar{y}) = (1 - f) S^2/n = (N - n)S^2/(Nn)$	$V(\bar{y}) = \sigma^2/n, \sigma^2 = (N - 1)S^2/N$
Estimated variance	$v(\bar{y}) = (1 - f) s^2/n$	$v(\bar{y}) = s^2/n$
Estimate of total Y	$\hat{Y} = N\bar{y}$	$\hat{Y} = N\bar{y}$
Variance of total	$V(\hat{Y}) = N^2(1 - f)S^2/n$	$V(\hat{Y}) = N^2\sigma^2/n$
Standard error	$SE = \sqrt{v}$	$SE = \sqrt{v}$

Note: SRSWOR is more precise than SRSWR because the fpc $(1 - f)$ is less than 1. If the sampling fraction is below 5%, the fpc can be ignored.

Example 6.1 — Estimation under SRSWOR

A sample of $n = 25$ households is drawn from $N = 500$. The sample mean monthly income is Rs. 60 (thousand) and $s^2 = 100$. Estimate the population mean, total and their standard errors, and give a 95% confidence interval for the mean.

Solution

$f = 25/500 = 0.05$, $fpc = 0.95$. Estimated mean = **60**

$v(\bar{y}) = (1 - f)s^2/n = 0.95 \times 100/25 = 3.80$; **$SE(\bar{y}) = 1.949$**

Estimated total $\hat{Y} = N\bar{y} = 500 \times 60 = \mathbf{30,000}$; **$SE(\hat{Y}) = N \times SE(\bar{y}) = 974.7$**

95% CI for mean $\approx \bar{y} \pm 1.96 SE = 60 \pm 3.82 = \mathbf{(56.18, 63.82)}$

3. Stratified Random Sampling

When the population is heterogeneous, it is divided into non-overlapping, internally homogeneous groups called **strata** (e.g., urban/rural, class levels), and an independent simple random sample is drawn from each. N_h = size of stratum h , n_h = sample size in stratum h , $W_h = N_h/N$, $n = \sum n_h$. **Advantages:** greater precision, estimates for each stratum, administrative convenience. It works best when strata are homogeneous within and different between.

Quantity	Formula
Estimate of mean	$\bar{y}_{st} = \sum W_h \bar{y}_h = \sum N_h \bar{y}_h / N$
Variance of $\bar{y}_{st}$ (WOR)	$V(\bar{y}_{st}) = \sum W_h^2 (1 - f_h) S_h^2 / n_h, f_h = n_h/N_h$
Estimate of total	$\hat{Y}_{st} = N \bar{y}_{st} = \sum N_h \bar{y}_h; V(\hat{Y}_{st}) = N^2 V(\bar{y}_{st})$
Equal allocation	$n_h = n / k$ (k = number of strata)
Proportional allocation	$n_h = n N_h / N = n W_h$ (self-weighting)

Quantity	Formula
Neyman (optimum) allocation	$n_h = n N_h S_h / \sum N_k S_k$ — larger samples for larger and more variable strata

Precision order (for the same n): **V(Neyman) $\leq$ V(proportional) $\leq$ V(SRS)**. Under proportional allocation $V(\bar{y}_{st}) = [(1 - f)/n] \sum W_h S_h^2$; under Neyman allocation $V(\bar{y}_{st}) = (\sum W_h S_h)^2/n - \sum W_h S_h^2/N$.

Example 6.2 — Stratified sampling and allocation

A population of $N = 1000$ is divided into three strata with $N_h = 200, 300, 500$, standard deviations $S_h = 10, 20, 30$ and sample means $50, 60, 70$. A total sample $n = 100$ is to be taken. Find the estimate of the mean, the allocation under proportional and Neyman allocation, and the variances.

Solution

$W_h = 0.2, 0.3, 0.5$. $\bar{y}_{st} = 0.2(50) + 0.3(60) + 0.5(70) = 10 + 18 + 35 = \mathbf{63}$

Proportional: $n_h = 100 \times W_h = \mathbf{20, 30, 50}$. $\sum W_h S_h^2 = 0.2(100) + 0.3(400) + 0.5(900) = 590$; $V = (1 - 0.1)/100 \times 590 = \mathbf{5.31}$

Neyman: $N_h S_h = 2000, 6000, 15000$, total 23000 $\rightarrow n_h = 100 \times N_h S_h / 23000 = 8.7, 26.1, 65.2 \approx \mathbf{9, 26, 65}$

$\sum W_h S_h = 2 + 6 + 15 = 23$; $V = 23^2/100 - 590/1000 = 5.29 - 0.59 = \mathbf{4.70}$ (smaller than 5.31). SE (Neyman) = 2.168

4. Ratio and Regression Methods of Estimation

If an auxiliary variable X (e.g., last year's yield, area) is correlated with the study variable Y (e.g., this year's yield) and the population mean $\bar{X}$ is known, using X can improve the precision of the estimate.

4.1 Under Simple Random Sampling

	Ratio estimator	Regression (difference) estimator
Best when	Y is roughly proportional to X (line through origin)	Y is linearly related to X (line need not pass through origin)
Estimator of $\bar{Y}$	$\bar{y}_R = R \bar{X}$, $R = \bar{y} / \bar{x}$	$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x})$, $b = \sum(x - \bar{x})(y - \bar{y}) / \sum(x - \bar{x})^2$
Estimate of total	$\hat{Y}_R = N \bar{y}_R = R X$	$\hat{Y}_{lr} = N \bar{y}_{lr}$
Variance (approx.)	$V(\bar{y}_R) \approx [(1 - f)/n] \cdot \sum(y_i - R x_i)^2 / (n - 1)$	$V(\bar{y}_{lr}) \approx [(1 - f)/n] \cdot S_y^2 (1 - \rho^2)$

	Ratio estimator	Regression (difference) estimator
Nature	Biased (bias negligible for large n) — better than $\bar{y}$ if $\rho > Cx/(2Cy)$	Approximately unbiased; never less precise than $\bar{y}$; at least as good as the ratio estimator

Example 6.3 — Ratio and regression estimation

From a population of $N = 50$ farms, a sample of $n = 5$ gives x (area) = 10, 12, 8, 15, 10 and y (yield) = 20, 26, 15, 31, 18. The known population mean area is $X = 11$. Estimate the mean yield using (a) the sample mean, (b) the ratio estimator, (c) the regression estimator.

Solution

$\bar{x} = 11.0$, $\bar{y} = 22.0$. (a) Simple mean estimate: **22.0**

(b) $R = \bar{y}/\bar{x} = 22.0/11.0 = 2.0000$. $\bar{y}R = R X = 2.0000 \times 11 = \mathbf{22.00}$; estimated variance = 0.450

(c) $b = 2.3929$, $r = 0.9827$. $\bar{y}lr = \bar{y} + b(X - \bar{x}) = 22.0 + 2.3929(11 - 11.0) = \mathbf{22.00}$; estimated variance ≈ 0.256

Since the sample mean area (11.0) equals X , the ratio and regression estimators equal or nearly equal $\bar{y}$ here; when $\bar{x}$ differs from X they correct for it.

4.2 Under Stratified Random Sampling

- **Separate ratio estimator:** $\bar{y}Rs = \sum W_h (\bar{y}_h/\bar{x}_h) X_h$ — uses a ratio in each stratum; needs the stratum means X_h ; best when the ratios differ between strata and each stratum sample is large.
- **Combined ratio estimator:** $\bar{y}Rc = (\bar{y}_{st} / \bar{x}_{st}) X$, with $\bar{y}_{st} = \sum W_h \bar{y}_h$ and $\bar{x}_{st} = \sum W_h \bar{x}_h$ — needs only the overall X ; safer when stratum samples are small.
- **Separate regression estimator:** $\bar{y}lrs = \sum W_h [\bar{y}_h + b_h(X_h - \bar{x}_h)]$; **combined regression estimator:** $\bar{y}lrc = \bar{y}_{st} + b(X - \bar{x}_{st})$ with a common b .
- Variances are obtained by applying the SRS variance formulas within strata and combining with weights W_h^2 .

5. Systematic, Cluster and Multistage Sampling

5.1 Systematic Sampling

Select a random start r between 1 and k , where $k = N/n$ is the **sampling interval**, then take every k -th unit: $r, r + k, r + 2k, \dots$. It is simple, spreads the sample evenly over the frame and needs only one random number. Estimator: $\bar{y}_{\text{sys}} = \Sigma y/n$. It is equivalent to SRS if the list is in random order, more precise if the list is ordered by a related characteristic, but may be biased if the list has a **periodic pattern** matching k .

Example 6.4 — Systematic sample

Select a systematic sample of 10 from $N = 100$ students with random start 4.

Solution

$k = 100/10 = 10$. Selected serial numbers: **4, 14, 24, 34, 44, 54, 64, 74, 84, 94**.

5.2 Cluster Sampling

The population is divided into natural groups (**clusters**) such as villages, wards, schools or households; a random sample of whole clusters is chosen and **all units in the chosen clusters** are surveyed. It is cheaper and needs a frame of clusters only, but precision is lower than SRS when units within a cluster are similar (high intra-cluster correlation). Clusters should therefore be **heterogeneous within and similar between** (opposite to strata). For n clusters of equal size M chosen from K clusters: $\bar{y}_{\text{cl}} = \Sigma \bar{y}_i / n$, $V(\bar{y}_{\text{cl}}) = (1 - n/K) s^2_b / n$, where s^2_b is the variance between cluster means.

Example 6.5 — Cluster sampling

Four clusters (wards) are chosen at random, each of 5 households, with mean monthly expenses of 12, 15, 11 and 14 (thousand rupees). Estimate the mean expense.

Solution

$\bar{y}_{\text{cl}} = (12 + 15 + 11 + 14)/4 = \mathbf{13.0}$. $s^2_b = [1 + 4 + 4 + 1]/3 = 3.33$; ignoring fpc, $SE = \sqrt{(3.33/4)} = \mathbf{0.91}$.

5.3 Multistage Sampling

Sampling is done in **stages** from larger to smaller units: e.g., district (first-stage unit) → municipality/VDC (second stage) → ward (third stage) → household (final stage), with random selection at each stage. It needs frames only for the selected units, is economical for large geographic areas and is used in national surveys; however, the estimation is more complex and errors from every stage add up. For a two-stage design with equal-sized units, $\bar{y} = (1/n) \sum \bar{y}_i$ and $V(\bar{y}) = (1 - f_1) S_1^2/n + (1 - f_2) S_2^2/(nm)$, where n first-stage and m second-stage units are chosen.

6. PPS Sampling and Estimation of Population Total

In **probability proportional to size (PPS)** sampling, the chance of selecting a unit is proportional to its size (measure M_i such as population, area or number of employees): $p_i = M_i / \Sigma M$. Large units (villages, firms) have larger chances, which improves precision when Y is related to size. Units are selected by the **cumulative total method**: form cumulative totals of M_i , draw a random number from 1 to ΣM and pick the unit whose range contains it (with replacement).

Quantity (PPS with replacement, n draws)	Formula
Estimate of population total	$\hat{Y}_{pps} = (1/n) \sum (y_i / p_i)$ (Hansen–Hurwitz)
Variance of $\hat{Y}_{pps}$	$V(\hat{Y}_{pps}) = (1/n) \sum p_i (y_i/p_i - Y)^2$
Unbiased estimate of variance	$v(\hat{Y}_{pps}) = [1 / (n(n - 1))] \sum (y_i/p_i - \hat{Y}_{pps})^2$
Estimate of mean	$\bar{Y} = \hat{Y}_{pps} / N$

Example 6.6 — PPS estimate of a total

Four villages have populations 100, 200, 300, 400 (total 1,000), so $p = 0.1, 0.2, 0.3, 0.4$. Two draws select villages with $p = 0.2$ and 0.4 , where the number of school children (y) observed is 60 and 90. Estimate the total number of school children and its standard error.

Solution

$$y_i/p_i = 60/0.2 = 300 \text{ and } 90/0.4 = 225.$$

$$\hat{Y}_{pps} = (300 + 225)/2 = 262.5$$

$$v(\hat{Y}_{pps}) = [(300 - 262.5)^2 + (225 - 262.5)^2] / (2 \times 1) = 1406.25; \text{SE} = 37.5$$

Design	Estimate of population total Y	Key variance factor
SRSWOR	$\hat{Y} = N \bar{y}$	$N^2(1 - f)s^2/n$

Design	Estimate of population total Y	Key variance factor
Stratified	$\hat{Y}_{st} = \sum N_h \bar{y}_h$	$\sum N_h^2(1 - f_h)s_h^2/n_h$
Ratio	$\hat{Y}_R = R X$	uses $\sum (y - Rx)^2$
PPS (WR)	$\hat{Y}_{pps} = (1/n)\sum y_i/p_i$	$\sum (y_i/p_i - \hat{Y})^2 / [n(n-1)]$

7. Sampling Distributions (z, t, χ^2)

A **sampling distribution** is the probability distribution of a statistic (such as $\bar{x}$ or s^2) over all possible samples of the same size. Its standard deviation is called the **standard error (SE)**. The **Central Limit Theorem** states that for a large sample ($n \geq 30$) the distribution of $\bar{x}$ is approximately normal with mean μ and $SE = \sigma/\sqrt{n}$, whatever the shape of the population.

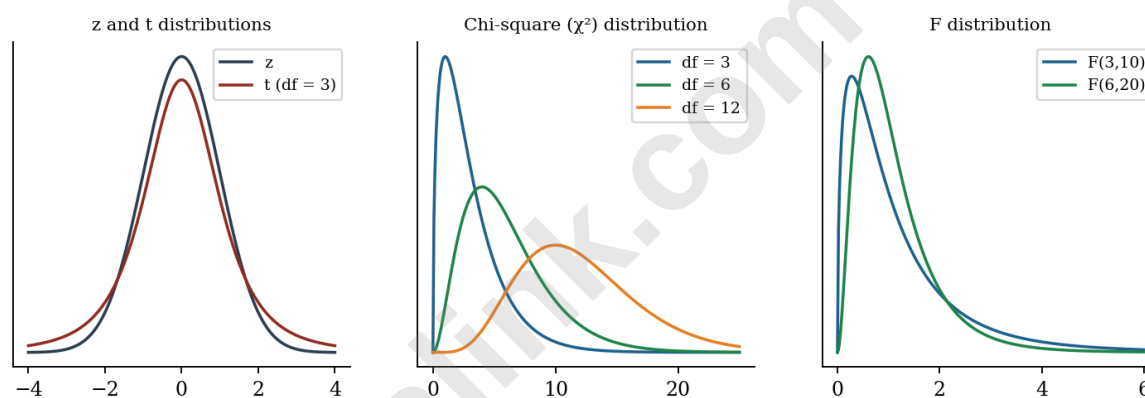


Figure 6.2 — z, t, chi-square and F distributions

Distribution	Statistic	Conditions and properties
Normal (z)	$z = (\bar{x} - \mu) / (\sigma/\sqrt{n})$	Large n (≥ 30) or σ known and population normal; $z \sim N(0, 1)$
Student's t	$t = (\bar{x} - \mu) / (s/\sqrt{n})$	Small n (< 30), σ unknown, normal population; degrees of freedom $v = n - 1$; symmetric, bell-shaped with heavier tails than z; tends to z as $v \rightarrow \infty$; mean 0
Chi-square (χ^2)	$\chi^2 = (n - 1)s^2 / \sigma^2$	Follows χ^2 with $n - 1$ d.f.; non-negative and right-skewed; mean = v , variance = $2v$; used for variance intervals and tests of goodness of fit / independence
F	$F = (s_1^2/\sigma_1^2) / (s_2^2/\sigma_2^2)$	Ratio of two independent χ^2 variables divided by their d.f.; d.f. ($n_1 - 1, n_2 - 1$); studied in Unit 7

Degrees of freedom = number of independent pieces of information in the statistic (n minus the number of estimated parameters).

Example 6.7 — z (sampling distribution of the mean)

A population has $\mu = 50$ and $\sigma = 10$. For samples of size $n = 25$, find the probability that the sample mean exceeds 53.

Solution

$$SE = 10/\sqrt{25} = 2. \quad z = (53 - 50)/2 = 1.50. \quad P(\bar{x} > 53) = P(Z > 1.50) = 1 - 0.9332 = \mathbf{0.0668}$$

Example 6.8 — t distribution

A sample of $n = 10$ has mean 52 and standard deviation 6. Test whether the population mean could be 50, and give a 95% confidence interval for μ .

Solution

$$SE = 6/\sqrt{10} = 1.897. \quad t = (52 - 50)/1.897 = \mathbf{1.054} \text{ with } v = 9.$$

Critical t (two-tailed, 5%, 9 d.f.) = 2.262. Since $|1.054| < 2.262$, the mean of 50 is **not rejected**.

$$95\% \text{ CI} = 52 \pm 2.262 \times 1.897 = \mathbf{(47.71, 56.29)}$$

Example 6.9 — Chi-square (variance)

A sample of $n = 15$ gives $s^2 = 25$. Construct a 95% confidence interval for the population variance σ^2 .

Solution

$$v = 14. \quad \chi^2(0.025, \text{upper}) = 26.119, \quad \chi^2(0.975, \text{lower}) = 5.629.$$

$$CI = [(n - 1)s^2/\chi^2_{\text{upper}}, (n - 1)s^2/\chi^2_{\text{lower}}] = [350/26.119, 350/5.629] = \mathbf{(13.40, 62.18)}$$

Key Takeaways

- ✓ SRSWOR: $V(\bar{y}) = (1 - f)S^2/n$; SRSWR: $V(\bar{y}) = \sigma^2/n$; total $\hat{Y} = N\bar{y}$.
- ✓ Stratified sampling: $\bar{y}_{st} = \sum W_h \bar{y}_h$; allocation may be equal, proportional ($n_h \propto N_h$) or Neyman ($n_h \propto N_h S_h$); Neyman is the most precise.
- ✓ Ratio ($\bar{y}_R = R\bar{X}$) and regression ($\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x})$) estimators use an auxiliary variable to gain precision.
- ✓ Systematic uses a random start and interval $k = N/n$; cluster samples whole groups; multistage samples in stages; PPS gives larger units larger chances, $\hat{Y} = (1/n)\sum y_i/p_i$.
- ✓ z is used for large n , t ($n - 1$ d.f.) for small n with unknown σ , χ^2 for variances, F for ratios of variances.

— End of Unit 6 —

bcalink.com