# Sampling

Sagnik Chakravarty

# Table of contents

# Lecture 1- Applied Sampling

## Definition of Terms

1. **Sampling:** Sampling is not mere substitution of a partial coverage for a total coverage. Sampling is the science and art of controlling and measuring the reliability of useful statistical information through the theory of probability
2. **Statistics:** population quantity of interest , e.g., population total, mean, percentiles, regression coefficient, etc.
3. **Sampling Distribution:** the distribution of different values of the statistics $\hat{t}$ obtained by the process of taking all possible samples from the population, i.e., repeated sampling
4. (Design Based) Randomization theory: the population data are fixed, and the sampling inclusion indicators are random variables. The probabilities of selection for units in the samples give the sampling distribution of the statistics $\hat{t}$
5. Model-based sampling inference: the population is a set of random variables following some probability distribution, and the actual sample values are realizations of these random variables. The sample data are fixed, and the population distribution is unknown.

## Probability Formulas

1. A random variable takes on different values in different samples
2. The probability distribution of $Y$ depicts the set of possible values $y$ and the probability if each value occurring P(Y=y) where the quantity y is called a realization of Y.
3. The **expectation** of Y is defined as $E(y) = \sum_Y y \times P(Y = y)$
4. The **variance** of Y is: $V(y) = E[\{Y - E(Y)\}^2] = E(Y^2) - [E(Y)]^2$
5. The **covariance** is: $E[\{X - E(X)\}\{Y - E(Y)\}] = E(XY) - E(X)E(Y)$
6. The **Correlation** is: $Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$
7. The **coefficient of variation** is defined as: $CV(Y) = \frac{SD(Y)}{E(Y)}$
8. The **joint probability** is: $P(X = x, Y = y)$
9. The **Conditional Probability** is: $P(Y = y|X = x) = \frac{P(X=x,Y=y)}{P(X=x)}$
10. The **Conditional Expectation** is: $E(Y|X = x) = \sum_y y \times P(Y = y|X = x)$
11. The **Conditional Variance** is: $V(Y|X = x) = \sum_y [y = E(Y|X = x)]^2 \times P(Y = y|X = x)$
12. Successive Conditioning: $E(Y) = E[E(Y|X)]$
13. Total Variability: $V(Y) = V[E(Y|X)] + E[V(Y|X)]$

## Population Formulas

1. A population of N elements $Y_1, \cdots, Y_N$
2. Population total: $t = Y = Y_1 + Y_2 + \cdots + Y_N$
3. Population mean: $\bar{Y} = t/N$
4. Population element variance:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N}(\sum_{i=1}^{N} Y_i^2 - t^2/N)$$

5. For binary deviation $Y_i$
    1. Population proportion: $P = t/N$
    2. Population element variance:

$$S^2 = \frac{N}{N-1}P(P-1)$$

## Sample Formulas

1. One sample of n elements: $y_1, y_2, \cdots, y_n$
2. Sample total: $t = y_1 + y_2 + \cdots + y_n$
3. Sample mean: $\bar{y} = t/n$
4. Sample element variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y - \bar{y})^2 = \widehat{S^2}$
5. Sample fraction: $f = n/N$
6. Finite population correction factor: $f_{PC} = 1 - f$
7. These are the calculated/observed summary statistics based on the one sample dataset

## Sampling Inference

1. Denote the selection probability for one sample as ( ) and obtain the sample estimate $\hat{t}_s$
2. **Expectation:** The mean of the sampling distribution of $\hat{t}$:

$$E(\hat{t}) = \sum_{\text{all possible samples s}} \hat{t}_s P(s)$$

3. **Variance:** $V(\hat{t}) = E[\{\hat{t} - E(\hat{t})\}^2] = \sum_s (\hat{t}_s - E(\hat{t}))^2 P(s)$
4. **Standard Error:** $SE(\hat{t}) = \sqrt{V(\hat{t})}$, i.e the standard deviation of sampling distribution of $\hat{t}$
5. **Confidence interval:** $CI(s) = [low_s, up_s]$, if we repeatedly take samples from the population, construct a 95% CI for each possible sample, we expect 95% of the resulting intervals to include the true value, i.e., a 95% chance that the sample containing the true value

## Quality Measure

1. **Bias:** the difference between the true and the expected value: $Bias(\hat{t}) = E(\hat{t}) - t$
2. **Mean Squared Error:** $MSE(\hat{t}) = E[(\hat{t} - t)^2] = V(\hat{t}) + Bias^2(\hat{t})$
3. An estimator $\hat{t}$ of $t$ is **unbiased** if $E(\hat{t}) = t$, **precise** if $V(\hat{t})$ is small, and **accurate** if $MSE(\hat{t})$ is small and the CI coverage probability is close to the nominal level.
4. Statistical inference validity assessment calculates bias, variance, MSE and CI

## Lecture 2- Simple Random Sampling

### Implementation

1. SRS is the most basic form of probability sampling and provides the basis for the more complicated forms.
2. Select a random sample of size n from a population of N
    1. SRS with replacement (SRSWR): the same unit can be included more than once in the sample, with equal selection probability $\pi_i = n/N$
    2. SRS without replacement (SRS): all units in the sample are distinct with the equal selection probability $\pi_i = n/N$

### Inference: Population mean

1. Support we collect the systolic blood pressure (SBP) measurements of the five individuals selected via SRSWOR from a population of 20 and would like to estimate the population average SBP value: 110, 125, 145, 90, 135
2. The finite population correction factor $f_{pc} = 1 - f = 1 - n/N = 1 - 5/20 = 15/20$
3. The sample total $t = 110 + 125 + 145 + 90 + 135 = 605$
4. The sample mean $\bar{y} = t/n = 605/5 = 121$
5. The element variance estimate $s^2 = \frac{1}{n-1}\sum(y_i - \bar{y}) = 467.5$
6. The standard deviation $s = \sqrt{s^2} = 21.622$
7. Population total estimate: $\hat{t} = N\bar{y} = 20 \times 121$
8. The sampling variance of the population total estimate:

$$var(\hat{t}) = N^2 var(\bar{y}) = N^2(1-f)\frac{s^2}{n} = 20^2 \times 70.125$$

9. The standard error: $se(\hat{t}) = \sqrt{var(\hat{t})} = N\sqrt{var(\bar{y})} = 20 \times 8.374$
10. the $1-\alpha$ confidence interval: $\hat{t} \pm t_{1-\alpha/2,n-1} se(\hat{t})$

### Inference: Population Proportion

1. Now we are interested in estimating the proportion of individuals with hypertension, and hypertension indicator values are: 0, 0, 0, 1, 0
2. The proportion estimate: $p = \bar{y} = 1/5$
3. The theoretical sampling variance (UNKNOWN):

$$Var(p) = (1-f)\frac{S^2}{n} = \frac{1-f}{n}\frac{N}{N-1}P(P-1)$$

4. The sampling variance estimate:

$$var(p) = \frac{1-f}{n}\frac{n}{n-1}p(p-1)$$

5. The sampling error: $se(p) = \sqrt{var(p)}$
6. The $1-\alpha$ confidence interval: $p \pm t_{1-\alpha/2.n-1} se(p)$
7. The desired precision is often expressed as $P(|\bar{y} - \bar{y}_u| < e) = 1 - \alpha$, where  is called the *margin of error*, as one-half of the width of a 95% CI.

### Design Effects

1. **Design Effect:** Ratio of variance under a new design to SRS variance with the same sample size
    1. Often used to compare SRS when evaluating complex sample survey design
    2. Inflate the projected SRS sample size by the design effect for complex sample survey sample size projection

## Sampling Frame

1. **Frame:** Set of materials used to designate a sample of units
2. Rule links frame elements to population elements
3. Accurate and up-to-date frames located in one location preferred
4. Numbered, computerized lists are best
5. The population and the list/frame may not match up

## Frame Problems

### Non coverage

- Some population elements are not on the frame
  - Have zero chance of selection
- Potential solutions
  - Use supplemental frames that cover noncovered elements
  - Use noncoverage weighting adjustments

### Blanks

- Frame elements do not have corresponding population elements
- Know frame element is blank after selection
  - Screening to find eligible list elements
- Potential solutions
  - Reject blanks by adjusting the sampling rate and size
  - Substitute with the next element on the listing

### Duplicates

- Occur when a single population element is linked to two or more frame elements
  - Potential solutions
    - If only a few readily identified, remove from the list before selection
      - Eliminating duplicates from the sample still leaves unequal probabilities of selection
    - Choose unique listing
      - First, last, largest, or randomly chosen frame listing
    - Determine how many duplicates for a given selected element and weight

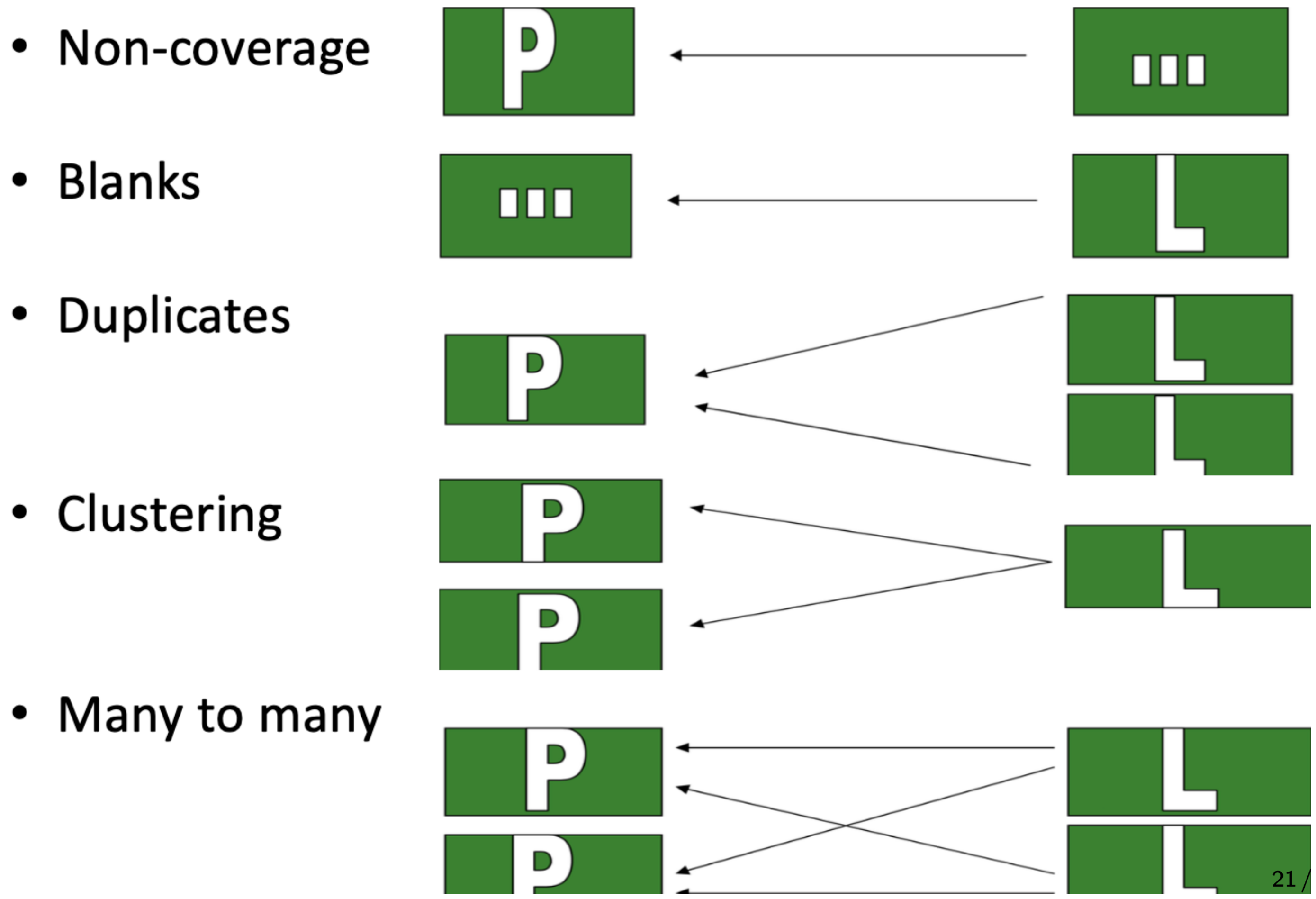
### Clustering

- Occurs when more than one population element can be selected by a sample frame element
- Potential solutions
  - Take all elements within selected clusters
    - The sample size varies with unequal-size clusters
    - Can adjust the sample size in advance
  - Use cluster sampling
  - Weighting adjustment

### Many to many matching

- Occurs when more than one population element can be selected by more than one frame element
- Potential solutions
  - Combinations of weighting and subsampling

### Summary



- Non-coverage
- Blanks
- Duplicates
- Clustering
- Many to many

## Withing household selection: The Kish method

- Interviewers list eligible household members by gender and age
- Use selection table
- Selection tables “rotated” across households
- Maximum of four eligibles per household can be handled
    - Can be expanded to handle households with five/six eligibles

# Lecture 3- Stratified Sampling

## Implementation

1. Dividing our population of elements into subgroups (strata) using auxiliary information that is available prior to drawing the sample
2. Simple random sampling of elements WITHIN each of the strata (or population subgroups): **Independent** across strata
3. Need the auxiliary information on the frame to create mutually exclusive and exhaustive subgroups (strata)

- Avoid selecting a really bad SRS sample
- Desire precision for subgroups
- More convenient to administer and may results in a lower survey cost
- Often gives more precise estimates for population means and totals

## Inference

- We can apply everything that we've learned about for SRS within each of the strata
- Stratum index: $h = 1, \cdots, H$
- Denote the variable of interest for $i^{th}$ element in stratum $h$ as $Y_{hi}$
- For each population stratum, population mean $\bar{Y_h} = \sum_{i=1}^{N_h} Y_{hi}/N_h$ and element variance $S_h^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(Y_{hi} - \bar{Y}_h)^2$
- For each stratum in the sample, we can compute $\bar{y_h}, s_h^2, t_h, N_h, n_h$ etc., in addition to sampling variances, etc.; **all specific to** h

## Inference: Population mean

- We can rewrite the population mean as a weighted sum of the population means for each stratum, where the weight $W_h$ is the relative proportion of the population within each stratum:

$$\bar{Y} = \sum_h \frac{N_h}{N}\bar{Y}_h = \sum_h W_h \bar{Y}_h$$

- We can write the sample mean in the same way, assuming that we have good (unbiased) estimates of the means in each stratum:

$$\bar{y}_w = \sum_h W_h \bar{y}_h$$

**Within Stratum** h: SRS

- Sampling fraction; $f_h = n_h/N_h$
- Mean Estimate; $\bar{y_h}$
- Element variance estimate: $s_h^2 = \frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2$ [or $s_h^2 = \frac{n_h}{n_h-1}p_h(1-p_h)$ if proportion]
- Sampling variance estimate: $var(\bar{y}_h) = (1 - f_h)\frac{s_h^2}{n_h}$
- Standard Error: $se(\bar{y}_h) = \sqrt{var(\bar{y}_h)}$

**Combine across strata**

The sampling variance of the overall estimated mean is entirely a function of the within-stratum sampling variances only

$$\mathrm{var}(\bar{y}_w) = \mathrm{var}\left(\sum_h W_h \bar{Y}_h\right) = \sum_h W_h^2 \mathrm{var}(\bar{y}_h) = \sum_h W_h^2 (1-f_h)\frac{s_h^2}{n_h}$$

## Inference: Sampling Weight

- **Element-level weighting:**

$$\bar{y}_w = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi} y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}}$$

- Here we have introduced the survey weight $w_{hi}$, the sampling weight for unit $i$ in stratum $h$.
- **The sampling weight is often the reciprocal of the inclusion probability:** $w_{hi} = \frac{1}{\pi_{hi}}$, where $\pi_{hi}$ is the inclusion probability of unit $i$ in stratum $h$.
- For stratified sampling, $\pi_{hi} = \frac{n_h}{N_h}$, so we have: $w_{hi} = \frac{N_h}{n_h}$ and $\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi} = N$.
- A stratified sample is self-weighting if the sampling fraction $\frac{n_h}{N_h}$ is the same across strata, where the sampling weight for each observation is $\frac{N}{n}$, exactly the same as in simple random sampling (SRS).

## Analysis of Variance (ANOVA)

The sum of squares

$$\begin{aligned}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(Y_{hi} - \bar{Y}) =& (N-1)S^2 \\ =& \sum_{h=1}^{H}(N_h - 1)S_h^2 + \sum_{h=1}^{H} N_h(\bar{Y_h} - \bar{Y}) \\ =& SSW + SSB\end{aligned}$$

hence $SSTO = SSW + SSB$

**We have the simplification as**

$$\begin{aligned}S^2 &= \sum_{h=1}^{H}\frac{N_h - 1}{N-1}S_h^2 + \sum_{h=1}^{H}\frac{N_h-1}{N-1}(\bar{Y}_h - \bar{Y})^2 \\ &\approx \sum_{h=1}^{H} W_h S_h^2 + \sum_{h=1}^{H} W_h(\bar{Y}_h - \bar{Y})^2 \\ &= \text{Within-stratum variance} + \text{Between-stratum variance}\end{aligned}$$

- The overall $S^2$ is fixed; if we define strata such that the between-stratum variance component becomes large, the within-stratum variance will necessarily become smaller.
- Hence, the sampling variance will go down based on Equation (1), which only depends on the within-stratum variance.
- Decrease the sampling variance of the mean by making strata heterogeneous between and homogeneous within.