NON-RESPONSE IN SAMPLE SURVEYS

By

Prof. G. N. Singh

Professor & HOD

Department of Mathematics & Computing

Indian Institute of Technology (Indian School of Mines), Dhanbad.

❖ Missing Data Problems in Sample Surveys

- Missing data is a common problem and challenge for analysts.
- Non-response error arises from failure to obtain the responses from the respondents (sample units) during the survey.

❖ Causes of Missing Data Problems in Sample Surveys

- There are many reasons why data could be missing, including:

- Unavailable (No Contact)
- Unwillingness (Refusal)
- Unable (Not able)

- A sensor failed.
- Someone purposefully turned off recording equipment.
- There was a power cut.
- The method of data capture was changed.

- An internet connection was lost.
- A network went down.
- A hard drive became corrupt.
- A data transfer was cut short.

❖ Methods to Handel the Non-Response Problems

- To deal with the problem of non-response following techniques are frequently being used by survey practitioners

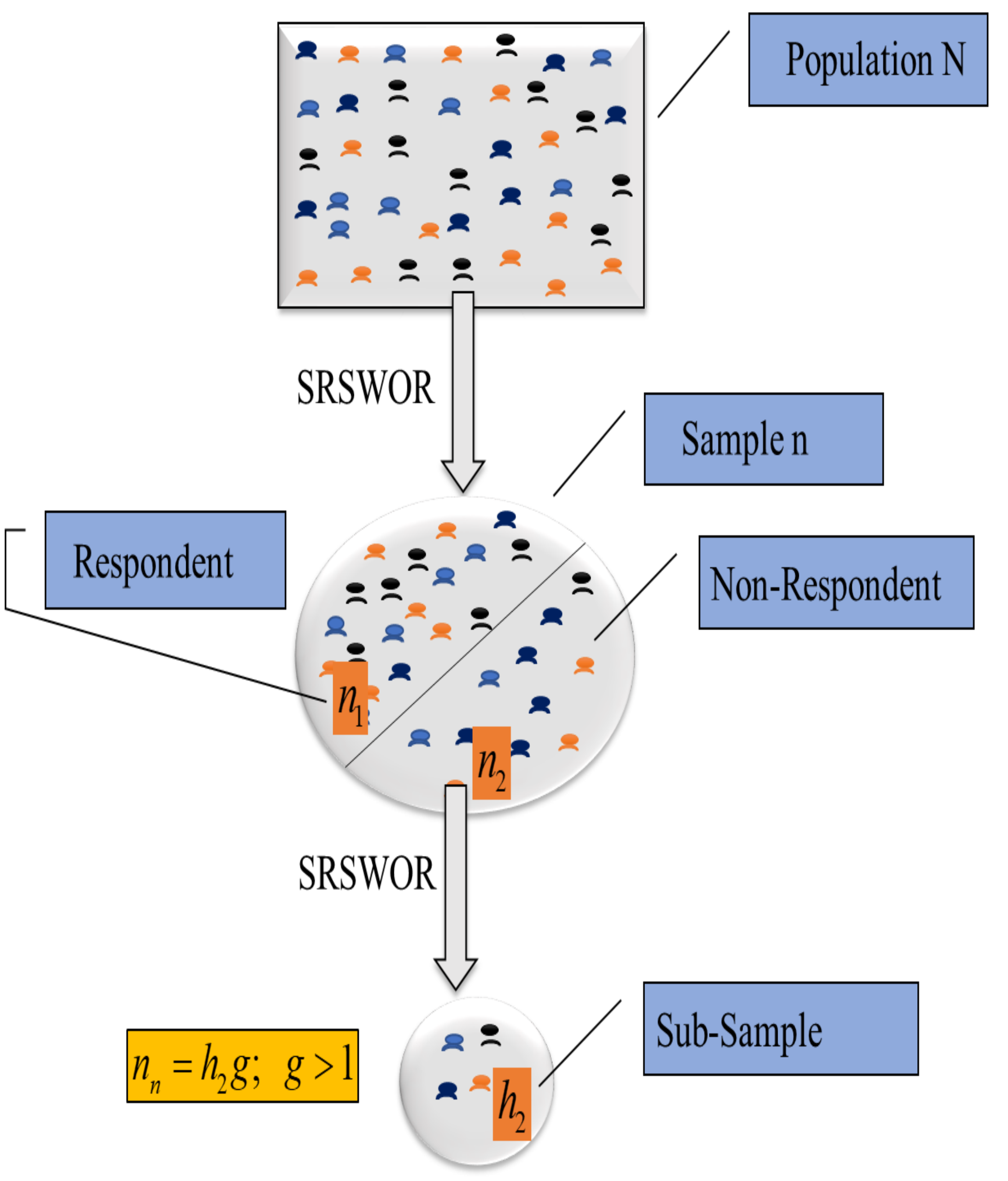
❖ Hansen and Hurwitz (1946) Technique

- Hansen and Hurwitz (1946) suggested a sub-sampling of non-respondents technique to deal with the problems of non-response which often occur in mail surveys. This technique is applicable for human surveys.

The Hansen and Hurwitz (1946) technique may be summarized in the following steps:

STEP 1: Select a sample of respondents from the population and seek their responses through postal mail, email or online survey etc. by a fixed deadline.

STEP 2: Once the deadline is over, identify the non-respondents.



STEP 3: Select a sub-sample from the non-respondents and seek their responses through more engaging mode of surveying such as personal interview by trained interviewers.

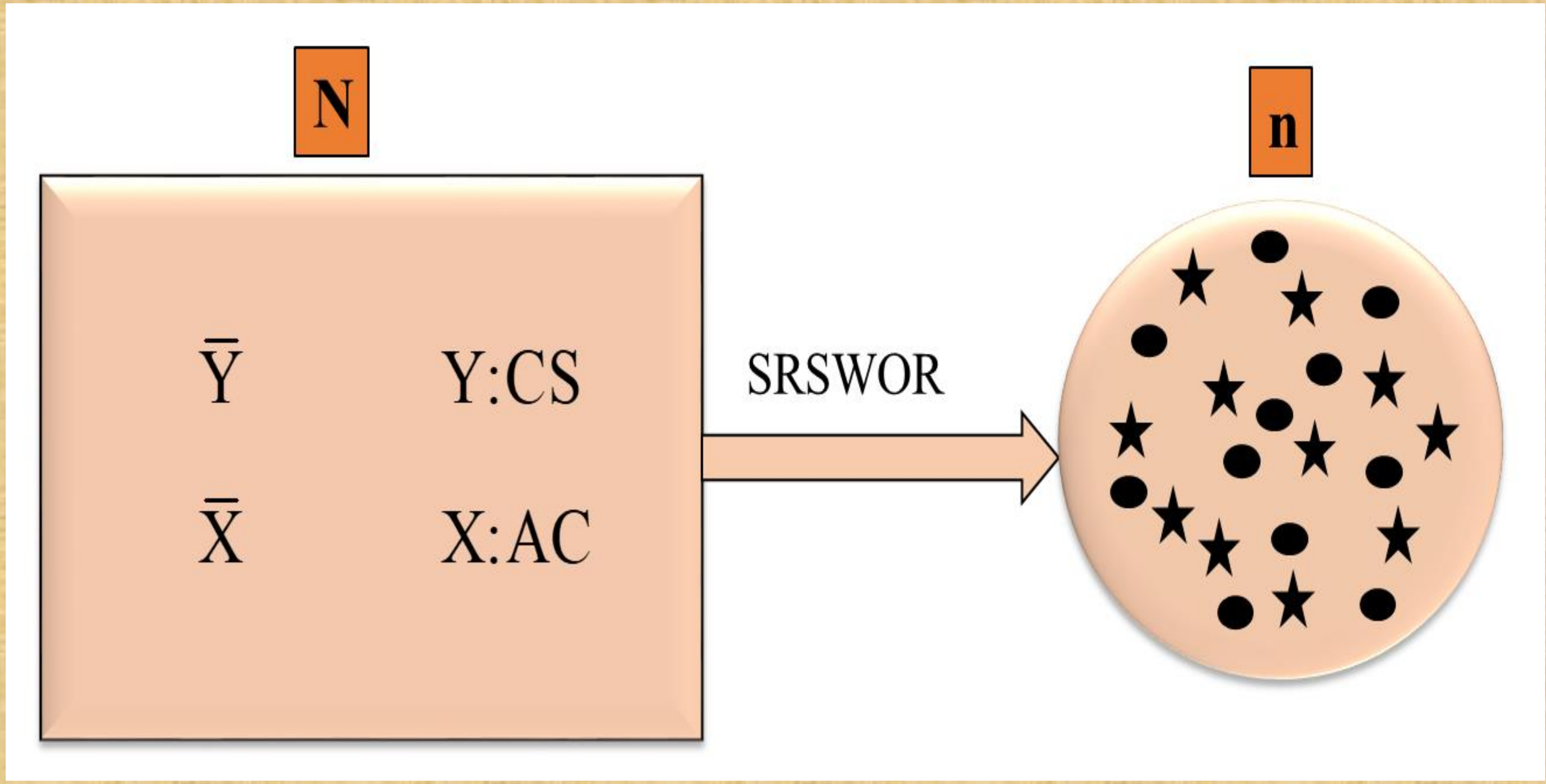
STEP 4: Combine the data of both the rounds of survey to develop an estimate for the population parameter.

The unbiased estimator of population mean $\bar{Y}$ is given as

$$\bar{y}_{HH} = \frac{n_1\bar{y}_1 + n_2\bar{y}_{h_2}}{n}$$

where $\bar{y}_{h_2}$ denotes the mean of h_2 observations from the sub-sample of non-responding units.

❖ Imputation Technique



A : Set of responding units

A= {★★★★★}

$\bar{A}$: Set of non-responding units

$\bar{A}$ = {● ● ● ●●}

● ● ● ●● (n-r) units

★★★★★ r units

n Sample Size

❖ General Imputation Method

$$y_{.i} = \begin{cases} y_i & if \quad i \in A \\ \hat{y}_i & \text{if} \quad i \in \bar{A} \end{cases}$$

$\hat{y}_i$ denotes the imputed value for the i^{th} non-responding unit

- The general point estimator of population mean takes the form

$$\bar{y}_s = \frac{1}{n}\left[\sum_{i=1}^{r} y_i + \sum_{i=1}^{n-r} \hat{y}_i\right]$$

where the value $\hat{y}_i$ is *the imputed value*

❖ Mean Method of Imputation

- In this method, no auxiliary information is used.
- The missing values are replaced with the mean of responding units of the study variable.
- Under this method the data after imputation becomes:

$$y_{.i} = \begin{cases} y_i & if \quad i \in A \\ \bar{y}_r & \text{if} \quad i \in \bar{A} \end{cases}$$

$\bar{y}_r$ is the response mean

- The point estimator of the population mean is :

$$\bar{y}_m = \frac{1}{n}\left[\sum_{i=1}^{r} y_i + \sum_{i=1}^{n-r} \bar{y}_r\right] = \bar{y}_r$$

❖ Ratio Method of Imputation

- In the ratio method of imputation, we assume that imputation is carried out with the aid of an auxiliary variable x.
- The data after imputation becomes:

$$y_{.i} = \begin{cases} y_i & if \quad i \in A \\ \hat{b}\, x_i & \text{if} \quad i \in \bar{A} \end{cases}$$

$$\text{where } \hat{b} = \frac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} x_i} = \frac{\bar{y}_r}{\bar{x}_r}$$

- The point estimator of the population mean becomes

$$\bar{y}_{RAT} = \frac{1}{n}\left[\sum_{i=1}^{r} y_i + \sum_{i=1}^{n-r} \hat{b}\, x_i\right] = \frac{\bar{y}_r}{\bar{x}_r}\bar{x}_n$$

$$\Rightarrow \bar{y}_{RAT} = \frac{\bar{y}_r}{\bar{x}_r}\bar{x}_n$$

❖ Hot Deck (HD) Method of Imputation

$$y_{.i} = \begin{cases} y_i & \text{if } i \in A \\ \bar{y}_{g(i)} & \text{if } i \in \bar{A} \end{cases}$$

where $y_{g(i)}$ is the y value given by the donor unit $g(i) \in A$, drawn at random (with replacement) from the r responding units.

- Under the HD method of imputation the point estimator of population mean

$$\bar{y}_{HD} = \frac{1}{n}\left[\sum_{i=1}^{r} y_i + \sum_{i=1}^{n-r} y_{g(i)}\right]$$

❖ Nearest Neighbour (NN) Method of Imputation

- Under the NN method the data after imputation becomes

$$y_{.i} = \begin{cases} y_i & if \quad i \in A \\ \bar{y}_{g(i)} & \text{if} \quad i \in \bar{A} \end{cases}$$

where $y_{g(i)}$ is the y value given by the donor unit $g(i)$ such that $\underset{g \in R_P}{Min}\left|x_g - x_i\right|$ occors for $g = g(i)$.

If it results in more than one unit a donor is randomly selected from them.

- Under the NN method of imputation the point estimator of the population mean becomes

$$\bar{y}_{NN} = \frac{1}{n}\left[\sum_{i=1}^{r} y_i + \sum_{i=1}^{n-r} y_{g(i)}\right]$$

❖ Regression Method of Imputation

$$y_{.i} = \begin{cases} y_i & if \quad i \in A \\ \hat{y}_i & \text{if} \quad i \in \bar{A} \end{cases}$$

$$where\ \hat{y}_i = \hat{a} + \hat{b}x_i$$

$$\hat{b} = \frac{s_{yx}(r)}{s_x^2(r)}; \quad \hat{a} = \bar{y}_r - \hat{b}\bar{x}_r$$

- Under the Regression method of Imputation the point estimator of the population mean

$$\bar{y}_{Reg} = \frac{1}{n}\left[\sum_{i=1}^{r} y_i + \sum_{i=1}^{n-r}\left(\hat{a} + \hat{b}x_i\right)\right]$$

$$\Rightarrow \bar{y}_{reg} = \bar{y}_r + \hat{b}\left(\bar{x}_n - \bar{x}_r\right)$$

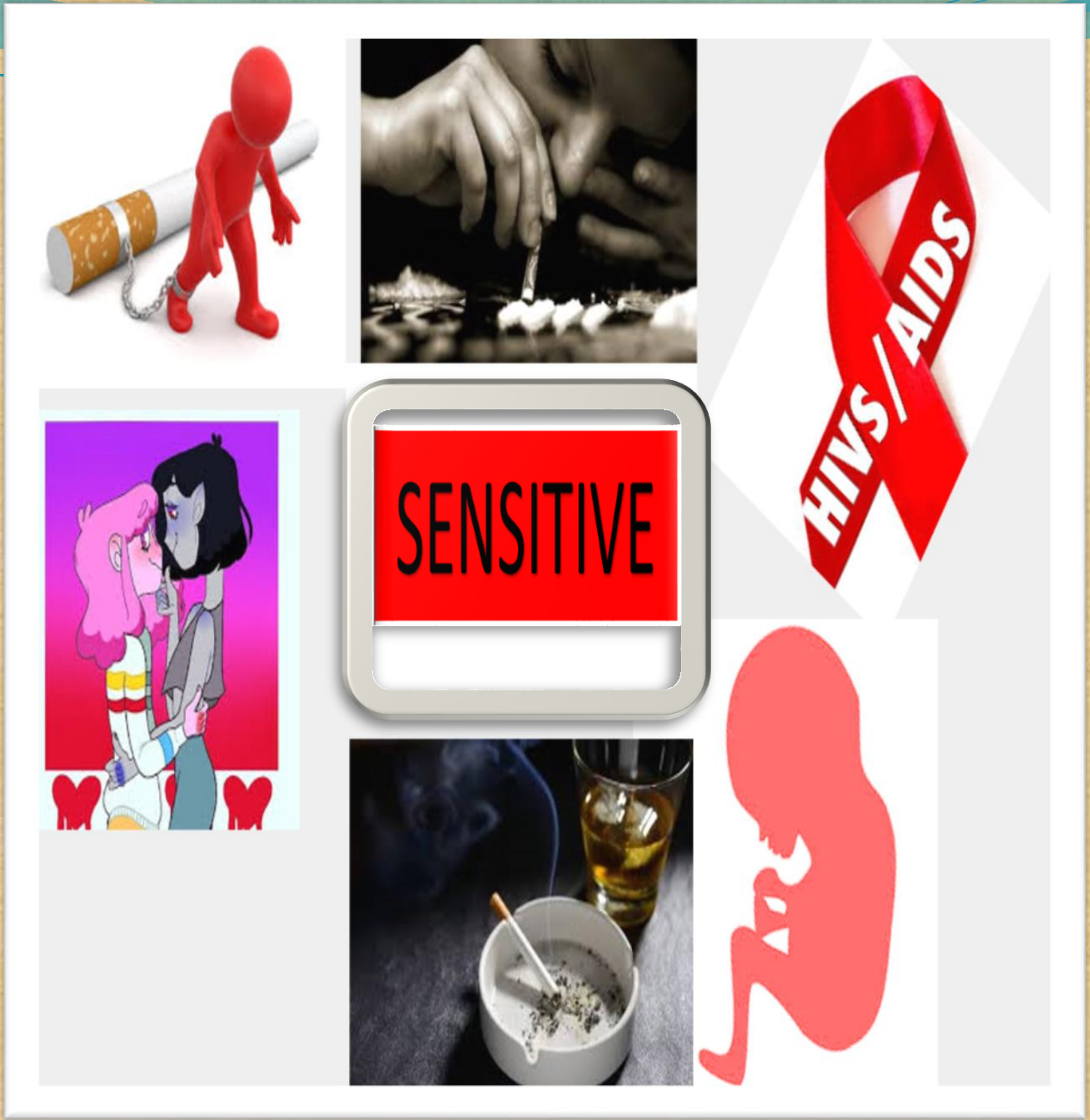
❖ Dealing with non-response arise due to sensitive issues

- Economists, psychologists, sociologists, managers, and policy makers have many reasons for asking personal questions.
- Biometric sample surveys need get ready-made information for future planning and policy implementations related to the subject matters of highly sensitive issues.

- Highly sensitive issues such as sexual behavior, domestic violence, tax offender, HIV infection status, drug addiction, extra marital affair etc.

- Actual answers of these questions are hidden or misguided by people.

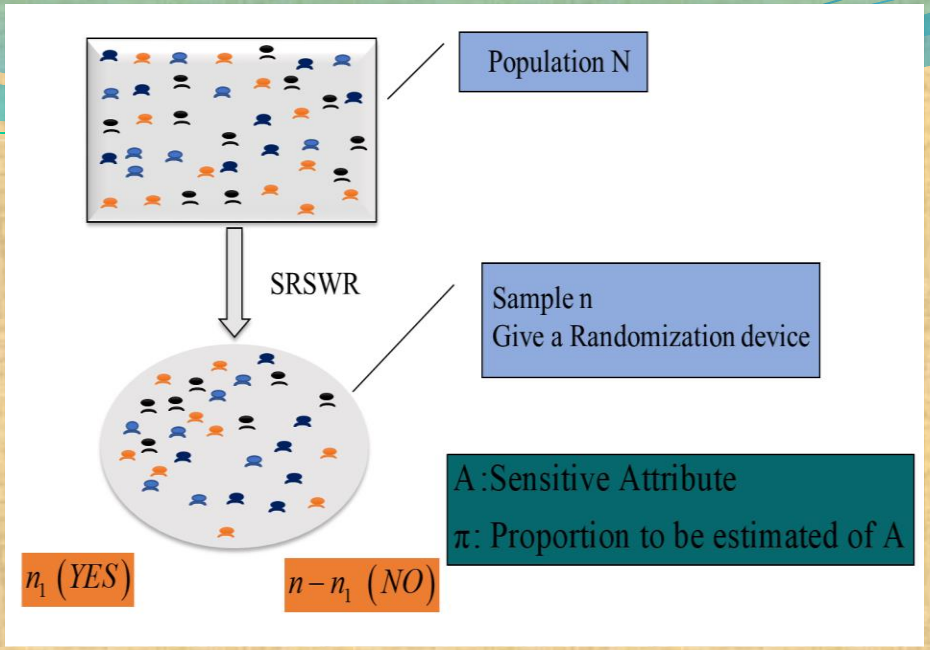
- Data obtained are definitely open to error if surveys are conducted through classical methods.



❖ Randomized Response Technique

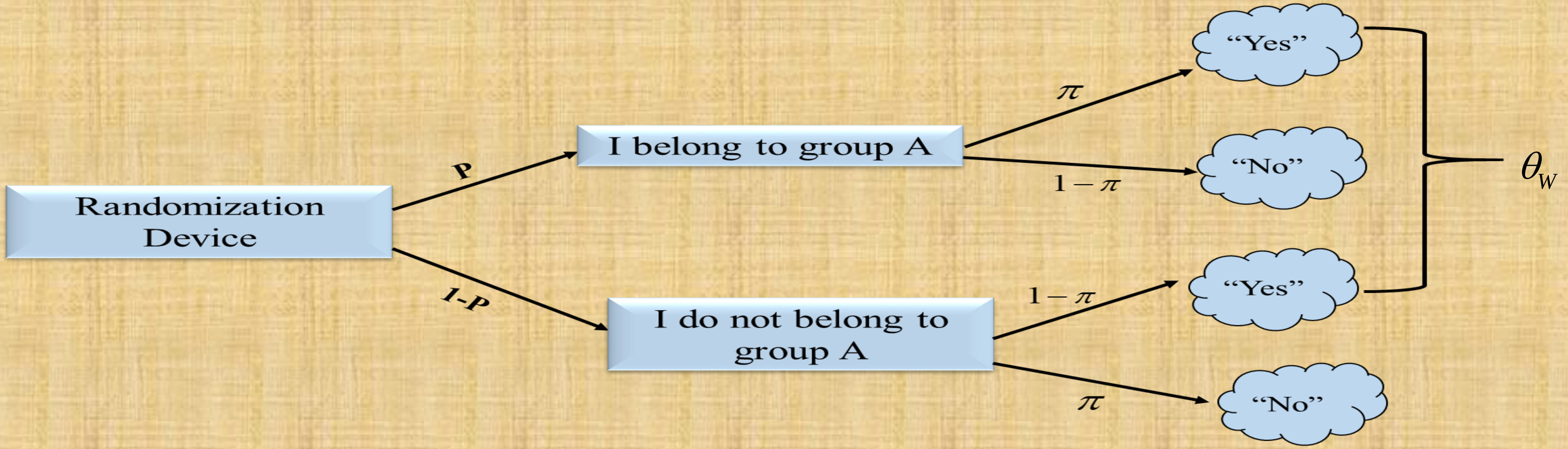
- Introduced by Warner (1965)
- When study under characteristics is sensitive in nature.
- Use randomization device to acquire the truthful response from respondents.
- To estimate the Proportion of the population possessing sensitive characteristic.

❖ Gradual Development: RRT



❖ Warner (1965) Technique

- Each respondent is provided an identical randomization device as:



- With the help of a randomized device, the respondent replies only “Yes” or “No” answers in a random sample of n respondents. The probability of “Yes” answer:

$$\theta_W = P\pi + (1-P)(1-\pi)$$

- The unbiased estimator of population proportion

$$\hat{\pi}_W = \frac{\hat{\theta}_W - (1-P)}{2P-1} \qquad P \neq 0.5$$

$$\hat{\theta}_W = \frac{n_1}{n}$$

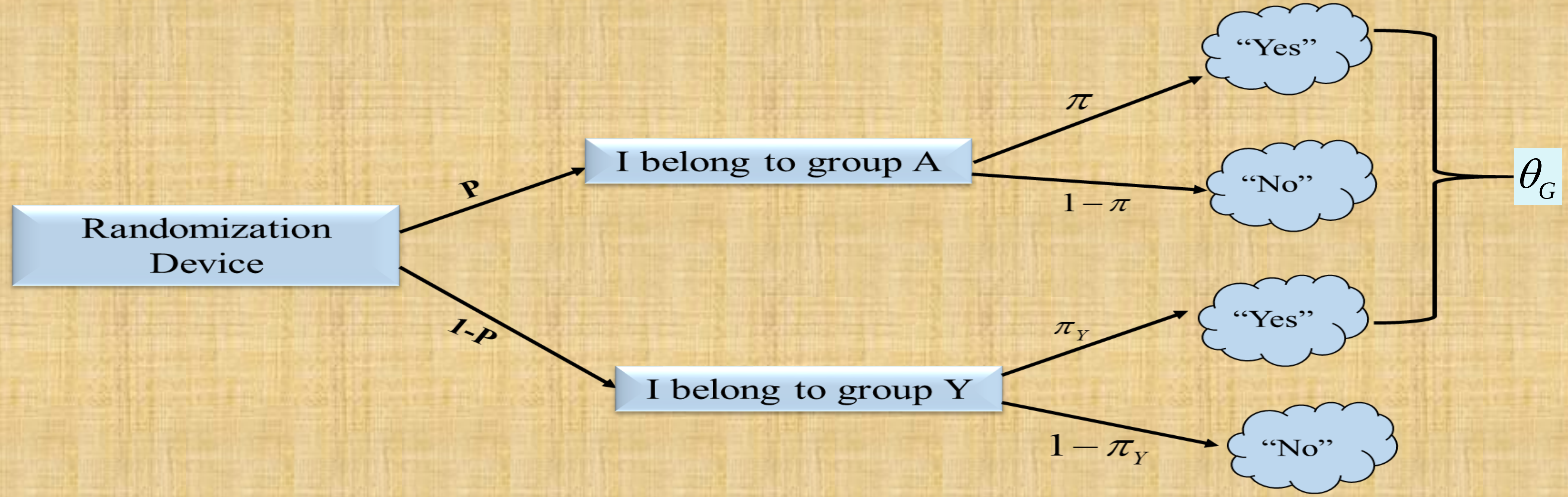
where $\hat{\theta}_W$ is the observed proportion of "Yes" answer in the sample of n units drawn by the SRSWR sampling.

- The Variance is :

$$V(\hat{\pi}_W) = \frac{\pi(1-\pi)}{n} + \frac{P(1-P)}{n(2P-1)^2}$$

❖ Greenberg et al. (1969) Technique

- Each respondent is provided an identical randomized device as:



- For example, in estimating the proportion of persons having the extra marital relations in a certain community the two questions may be:

i. Are you having extra marital relations?

ii. Did you born in the month of March?

❖ When the Proportion of unrelated character is known

- With the help of a randomized device, the respondent replies only "Yes" or "No" answers in a random sample of n respondents. The probability of "Yes" answer in the population:

$$\theta_G = P\pi + (1-P)\pi_Y$$

Let n_1 be the number of observed "Yes" answer in the sample of n units. So that $\hat{\theta}_G = \frac{n_1}{n}$

- The unbiased estimator of π is:

$$\hat{\pi}_G = \frac{\hat{\theta}_G - (1-P)\pi_Y}{P}$$

- The variance of the estimator $\hat{\pi}_G$ is:

$$V(\hat{\pi}_G) = \frac{\theta_G - (1-\theta_G)}{nP^2}$$

Population: $(N < \infty)$; A: Sensitive Attribute

Y: Non-Sensitive Unrelated Attribute

π: Proportion to be estimated of A

π_Y : Proportion of Non-Sensitive unrelated attribute Y

❖ When the Proportion of unrelated character is unknown

Here π_Y the proportion of unrelated character Y in the population is unknown

- In this case the probability of "Yes" answer:

Probability of "yes" answer using device 1

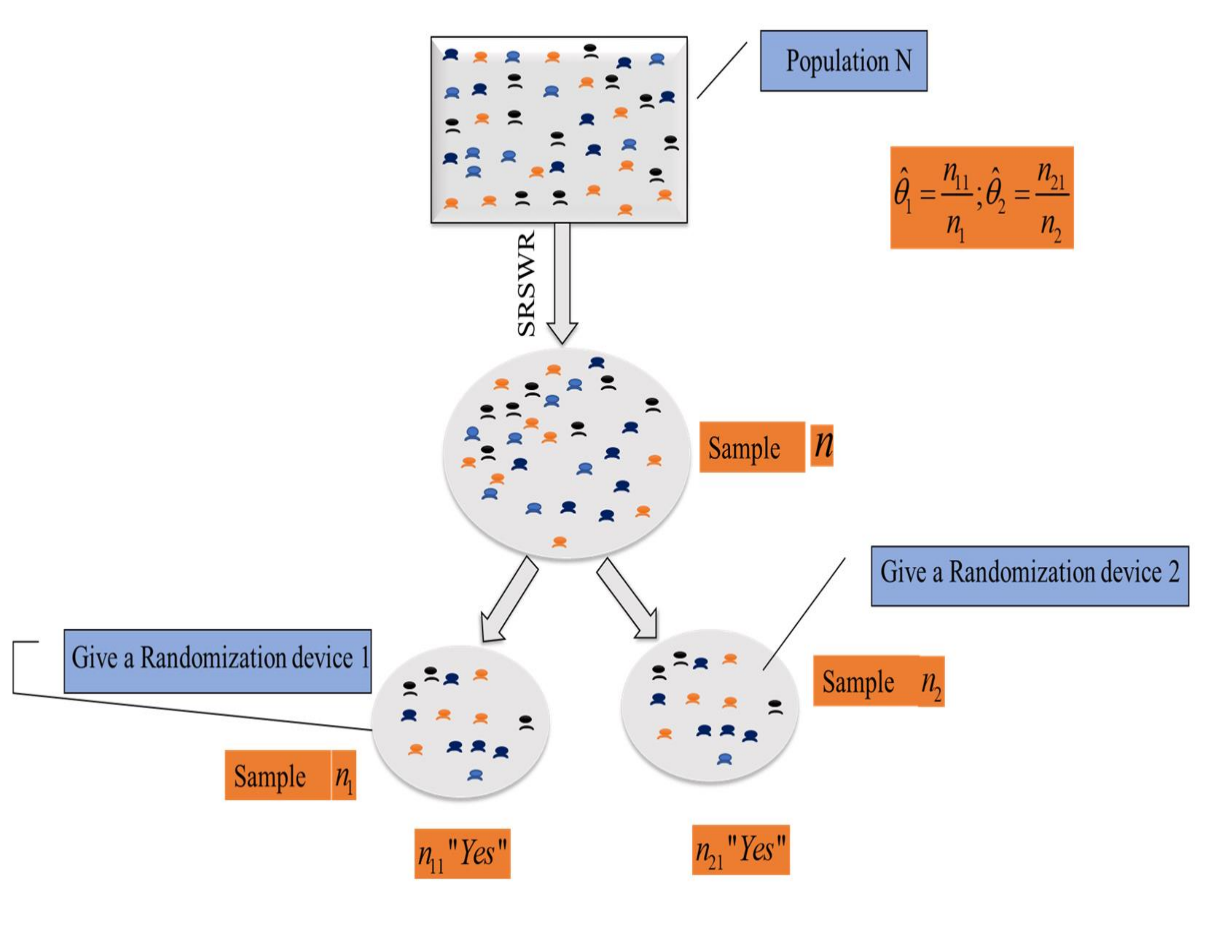
$$\theta_1 = P_1\pi + (1-P_1)\pi_Y$$

Probability of "yes" answer using device 2

$$\theta_2 = P_2\pi + (1-P_2)\pi_Y$$

- The unbiased estimator of π is:

$$\hat{\pi}_G = \frac{(1-P_2)\hat{\theta}_1 - (1-P_1)\hat{\theta}_2}{P_1 - P_2}$$



where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the observed proportion of "Yes" answer in the first and second sample respectively.

- The variance of the estimator $\hat{\pi}_G$ is:

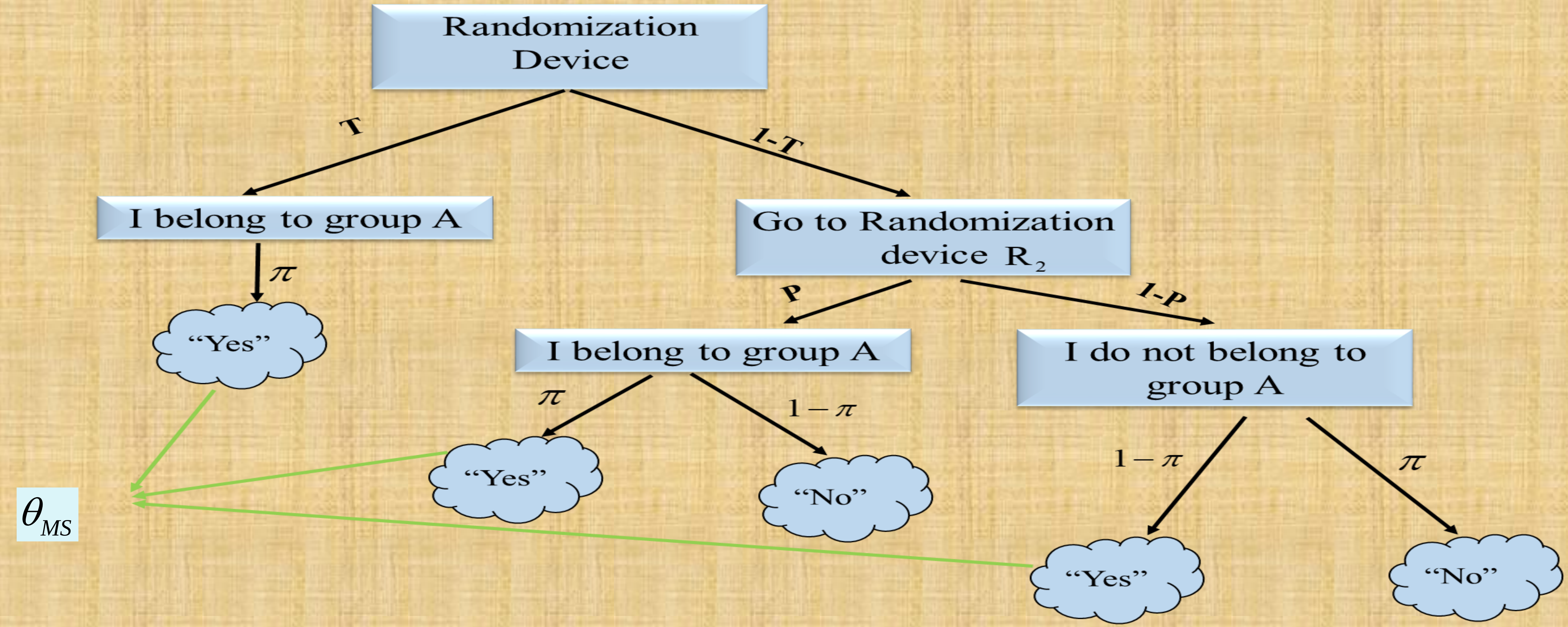
$$V\left(\hat{\pi}_G\right)=\frac{1}{\left(P_1-P_2\right)^2}\left[\frac{\left(1-P_2\right)^2\theta_1\left(1-\theta_1\right)}{n_1}+\frac{\left(1-P_1\right)^2\theta_2\left(1-\theta_2\right)}{n_2}\right]$$

For the best choice of n_1 and n_2, it should follow the relation

$$\frac{n_1}{n_2}=\sqrt{\frac{\theta_1\left(1-\theta_1\right)}{\theta_2\left(1-\theta_2\right)}}\frac{\left(1-P_2\right)}{\left(1-P_1\right)}$$

❖ Mangat and Singh (1990) Technique

- Each respondent is provided an identical randomized device



- With the help of a randomized device, the respondent replies only "Yes" or "No" answers in a random sample of n respondents. The probability of "Yes" answer:

$$\theta_{MS} = T\pi + (1-T)\left\{P\pi + (1-P)(1-\pi)\right\}$$

$$\hat{\theta}_{MS} = \frac{n_1}{n}$$

- The unbiased estimator of π is:

$$\hat{\pi}_{MS} = \frac{\hat{\theta}_{MS} - (1-T)(1-P)}{2P - 1 + 2T(1-P)}$$

- The Variance is :

$$V\left(\hat{\pi}_{MS}\right)=\frac{\pi\left(1-\pi\right)}{n}+\frac{\left(1-T\right)\left(1-P\right)\left\{1-\left(1-T\right)\left(1-P\right)\right\}}{n\left\{2P-1+2T\left(1-P\right)\right\}^2}$$

Population: $\left(N<\infty\right)$

A:Sensitive Attribute

π: Proportion to be estimated of A

THANK YOU