MSc - 23

Sample Survey

- ② → Cluster Sampling
- ③ → Two-stage Sampling
- ④ → Double Sampling
- ⑥ → PPS Sampling
- ⑤ → Sampling Design & Inference
- ① → Randomised Response

Randomised Response Technique :

In some sample survey studies, some of the study variables are sensitive in the nature but they are legal to be asked. Examples of such study variables are

(i) addition to alcohol or drugs.

(ii) having a history of abortions at some area.

(iii) habit of gambling.

(iv) habit of being without ticket passenger etc.

Clearly, considering complete and correct information on such sensitive issues is very difficult for the interviewer. It is not a delicad affair to bluntly ask a question (a stranger ~~question~~ question) about such sensitive and highly personal matter. Even if the interviewer doesn't refuse to answer such sensitive queries, we might wonder whether the replies provide by an interviewee are honest. This means that the collected data with the questionaires are very incomplete. In order to avoide excessive refusals or misleading responses to sensitive questions, we need new statistical methodology capable of (i) allowing the interviwee a satisfactory level of privacy and ~~thw~~ thus providing a basis which encourages greater coordination and

(ii) providing valid estimates of the population parameters.

under study

Varions statistical methods have been ~~desin~~ devised by the Statisticians dealing with such situations. All such techniques are known as "Randomized Response Techniques" the first study of these kinds were undertaken by Warner (1965) who developed a clever interviewing process design to reduce or eliminate non-sampling biases. in sample surveys at human population.

□ Randomized Response Techinque for ~~the~~ one dichotomous qualitative sensitive characteristic :

>> <u>What is the problem?</u>

Let. Y be a qualitative characteristic with two versions A (sensitive) and A^c (non-sensitive). For convenience, let

$$Y_i = \begin{cases} 0 & \text{, if the } i\text{th population unit have the non-sensitive version } A^c \\ 1 & \text{, otherwise.} \end{cases}$$

With respect to the characteristic, the population $\mathcal{U}$ can be divided into two subpopulations $\mathcal{U}_0$ & $\mathcal{U}_1$ having the size N_0 and N_1 respectively. All individual in $\mathcal{U}_1$ have the sensitive version of the characteristic and the goal of survey is to make a statistical study of the unknown population parameter $\phi = \frac{N_1}{N}$, which is the proportion of units in the population having the sensitive version of the characteristic Y.

Suppose we pose the ~~p~~ following question to the ith population unit :

Do you belong to $\mathcal{U}_1$?

Let, Y_i be the corresponding response variable, Then it is not ~~unrea~~ unreasonable to ~~post~~ postulate the following ~~mode~~ modes of ~~a~~ response

$$P(Y_i = 1 \mid i \in U_0) = 0 \qquad P(Y_i = 1 \mid i \in U_1) = q_{i1}$$

and

$$P(Y_i = 0 \mid i \in U_0) = 1 \qquad P(Y_i = 0 \mid i \in U_1) = q_{i2}$$

$$P(\text{No response from individual } i \mid i \in U_1) = q_{i3}$$

In other words, if the ith unit belongs to U_0 (ie. the non-sensitive group), then he must tell the truth ($Y_i = 0$) & never claims that he belongs to U_1.

But if he/she belongs to U_1 (ie. the sensitive group), then he/she may tell the truth (ie. $Y_i = 1$) or may tell a lie (ie $Y_i = 0$) or may simply refuse to answer. Each unit in the population has his own way of responding to the question. This model allows different likelihoods for telling the truth, telling a lie or no response for different individuals or units in U ($i = 1, 0$)

Unfortunately, this reasonable assumption on the response complicat the structure of the data collected on the sensitive issue and deprives the statistician from making any meaningful inference on the unknown parameter. To overcome this difficulty the sample units are allowed to keep their privacy by randomizing their responses but the freedom of the individual randomization is replaced by one or two common randomization. Since a randomized response is less revealing than a direct response, it is reasonable to expect greater co-operation.

There are two popular methods, (i) related question method and (ii) unrelated question method.

Related Question Method

To protect the privacy of the interviwee and to attract his collaboration, we don't measure the sensitive characteristic Y, rather we measure a randomized version of it, say Z. To do this each interviwee is presented with two complementary questions:

Q_0 : Do you belong to $\mathcal{U}_0$?

Q_1 : Do you belong to $\mathcal{U}_1$?

Answers to both the questions is either Yes $(Z=1)$ or No $(Z=0)$ and if the reply to Q_0 is Yes, then it would be No to Q_1, and vice-versa. In deciding which question to answer, each interviewee is provided with an identical randomization device, e.g. a spinner with its face divided into two mutually exclusive and exhaustive parts labelled as $\mathcal{U}_0$ & $\mathcal{U}_1$. Unseen by the interviewer, each interviwee spins the spinner. He is told to answer Q_0 and Q_1 depending on whether the spinner stops at $\mathcal{U}_0$ and $\mathcal{U}_1$. Therefore the outcome of each interviwee is either 1 or 0 and of course without the interviewer knowing which question has been answered. To be able to utilize the corrected data, it is necessary to assume that the answers are truthful. This means that if the random selection resulted in Q_0, then

$$P(Z_i = 1 \mid i \in \mathcal{U}_0) = 1 \qquad P(z_i = 1 \mid i \in \mathcal{U}_1) = 0$$

$$P(z_i = 0 \mid i \in \mathcal{U}_0) = 0 \qquad P(z_i = 0 \mid i \in \mathcal{U}_1) = 1$$

On the other hand if the random selection resulted in Q_1,

$$P(z_i = 1 \mid i \in \mathcal{U}_1) = 1 \qquad P(z_i = 1 \mid i \in \mathcal{U}_0) = 0$$

$$P(z_i = 0 \mid i \in \mathcal{U}_1) = 0 \qquad P(z_i = 0 \mid i \in \mathcal{U}_0) = 1$$

The above are assumed to be satisfied for every population unit i interviewed with corresponding z_i.

Estimation of ϕ

Let, π be the probability that the spinner stops at $\mathcal{U}_1$, ie. Q_1 is selected for answering. The choice of π is under the control of the interviewer and its value is assumed to be known to the data analyst. Under strict randomization $\pi \neq 1$ or $\pi \neq 0$. Further it is chosen to be different from 0·5. Then the Bernoulli random variable Z_i associated with the population unit i is characterised by the following randomized response distribution denoted by P_R.

$$P_R(Z_i=1) = P(Q_0 \text{ is selected})\, P(Z_i=1 \mid Q_0 \text{ is selected}) + P(Q_1 \text{ is selected})\, P(Z_i=1 \mid Q_1 \text{ is selected})$$

$$= (1-\pi)(1-\phi_i) + \pi\phi_i \quad , \text{where } \phi_i = \begin{cases} 1 & ,\text{if } i \in \mathcal{U}_1 \\ 0 & ,\text{if } i \in \mathcal{U}_0 \end{cases}$$

$$= (2\pi-1)\phi_i + (1-\pi)$$

$$= \theta_i \quad (\text{say}).$$

$$\therefore \quad P_R(Z_i=0) = 1-\theta_i$$

Hence,
$$\phi_i = \frac{\theta_i - (1-\pi)}{(2\pi-1)}$$

Therefore, if our data allow a response based unbiased estimation of θ_i by $\hat{\theta}_i$, then a responsed based unbiased estimator of ϕ_i based on θ_i will be

$$\hat{\phi}_i = \frac{\hat{\theta}_i - (1-\pi)}{(2\pi-1)} \quad , i=1(1)N$$

Note that, $\phi = \frac{1}{N}\sum_{i=1}^{N}\phi_i$

$\therefore$ ϕ is estimated by $\hat{\phi} = \frac{1}{n}\sum_{i=1}^{n}\hat{\phi}_i$

$\therefore \hat{\phi} = \dfrac{\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i - (1-\pi)}{(2\pi-1)}$, where, $\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i$ is the sample proportion of 'YES' response.

Note that, under SRSWR, the randomized response variables $z_1, z_2, \ldots, z_n$ are all i.i.d Bernoulli random variables with

$$P_R(z_i=1) = (2\pi-1)\phi + (1-\pi) = \theta \text{ (say)} \quad , i = 1(1)n$$

Then, $n_1 = \sum_{i=1}^{n} z_i \sim \text{Bin}(n, \theta)$

giving $\hat{\theta} = \dfrac{n_1}{n}$

Accordingly, $\hat{\phi} = \dfrac{\hat{\theta} - (1-\pi)}{(2\pi-1)} = \dfrac{\frac{n_1}{n} - (1-\pi)}{(2\pi-1)}$

with $E(\hat{\phi}) = \dfrac{\theta - (1-\pi)}{(2\pi-1)} = \phi$

& $V(\hat{\phi}) = \dfrac{\theta(1-\theta)}{n(2\pi-1)^2}$

The variance is estimated by $\widehat{V(\hat{\phi})} = \dfrac{\hat{\theta}(1-\hat{\theta})}{n(2\pi-1)^2}$

$$= \frac{\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)}{n(2\pi-1)^2}$$

Remark :

Note that, $\text{Var}(\hat{\phi}) = \dfrac{\theta(1-\theta)}{n(2\pi-1)^2}$

$$= \frac{\{(1-\pi) + (2\pi-1)\phi\}\{\pi - (2\pi-1)\phi\}}{n(2\pi-1)^2}$$

$$= \frac{\pi(1-\pi) + (2\pi-1)\phi\{\pi - 1 + \pi\} - (2\pi-1)^2\phi^2}{n(2\pi-1)^2}$$

$$= \frac{\phi(1-\phi)}{n} + \frac{\pi(1-\pi)}{n(2\pi-1)^2}$$

Therefore, Var$(\hat{\phi})$ is symmetric about $\pi = 0{\cdot}5$ and this is minimized by taking π as far from 0·5 as possible. To minimize the suspicion of the respondents the value of π should not be close to 0 or 1, otherwise the respondent might get a strong feeling of being identified as the members belonging to the sensitive group in case the response is 'yes' for large π & 'no' for small π. From this point of view, choice of π close to 0·5 would be ~~most~~ most acceptable but this produces a very large variance of the estimate $\hat{\phi}$. This suggests that the choice of π in the interval [0·6, 0·7] or [0·2, 0·3] would be most reasonable.

Unrelated Question Method

In the related question method, the proportion under study is divided into two mutually exclusive group $\mathcal{U}_1$ and $\mathcal{U}_0$. One with and the other without the stigmatising version of the ~~sensible~~ sensitive characteristic and based on this division of the population each respondent is asked to respond 'yes' or 'no' depending on whether or not he belongs to the group indicated by one of the two complementary questions Q_0 and Q_1 which is selected for him to answer. To increase the respondance likelihood of cooperation the related question procedure has been revised. by replacing Q_0 ~~to~~ with a harmless dichotomous question as described.

The required form is called the Unrelated Question Procedure in which the population is ~~de~~ divided into two overlapping groups $\mathcal{U}_1$ and $\mathcal{U}_2$. The group $\mathcal{U}_1$ contains all units with the stigmatising version of the sensitive characteristic and the group $\mathcal{U}_2$ contains all units having a version of some other dichotomous characteristic. which is completely innocuous and unrelated to the sensitive characteristic under study. Therefore in this procedure $\mathcal{U}_1 \cup \mathcal{U}_2 \neq \mathcal{U}$ and for an obvious reason. $\mathcal{U}_2 \subset \mathcal{U}_1$ is not allowed.

For example, Q_1 and Q_2 can be the following

Q_1: Do you use illegal drugs?

Q_2: Do you prefer football?

It seems reasonable to expect more cooperation from the sample units under this procedure in comparison with the related question procedure.

Under truthful reporting for both procedures the unrelated question procedure can be made to provide a more efficient estimator of the proportion of units ϕ associated with the stigmatising version of the sensitive characteristic.

Let, N_1 and N_2 be the ~~no~~ number of units in $\mathcal{U}_1$ and $\mathcal{U}_2$ respectively. As before we are interested in estimating $\phi = \frac{N_1}{N}$. We consider the cases separately δ (known) and δ (unknown) where $\delta = \frac{N_2}{N}$ (Proportion of individual having some version of the innocuous character).

Case-1 : δ known

In this case we select n units by a sampling design of our choice, Each interviewee in provided with the following two questions :

Q_1 : Do you belong to $\mathcal{U}_1$?

Q_2 : Do you belong to $\mathcal{U}_2$?

Unobserved by the interviewer each respondent selects randomly Q_1 ~~and~~or Q_2 with known probabilities q and $(1-q)$ respectively. The interviewer provides the randomization device. Each respondents answers in either 'YES' or 'NO' depending on which question is randomly selected,

Suppose n_1 be the ~~no~~ number of 'YES' responses reported, then we have the following ~~resut~~ results :

Under SRSWR of size n an unbiased estimator of ϕ is given by, ~~$\hat{\phi} = \frac{n_1}{n} q + (1-q)\delta$~~ $\widehat{P(\text{YES})} = \frac{n_1}{n} = q\hat{\phi} + (1-q)\delta$

$$\Rightarrow \hat{\phi} = \left\{\frac{n_1}{n} - (1-q)\delta\right\} / q$$

with $\quad Var(\hat{\phi}) = \dfrac{\alpha(1-\alpha)}{nq^2}$, where $\alpha = q\phi + (1-q)\delta$ —— (*)

The variance is unbiasedly estimated by.

$$\widehat{Var(\hat{\phi})} = \frac{\hat{\alpha}(1-\hat{\alpha})}{nq^2} \quad ; \quad \hat{\alpha} = \frac{n_1}{n} \quad [\text{proportion of 'YES' response}]$$

Let, $\quad x_i = \begin{cases} 1 & \text{, if } i^{th} \text{ individual replies 'YES'} \\ 0 & \text{, o.w.} \end{cases}$

So, $x_i \sim Ber\,(p = P(x_i = 1) = \alpha)$

$$\therefore \quad n_1 = \sum_{i=1}^{n} x_i \sim Bin(n, \alpha)$$

$$E(n_1) = n\alpha \Rightarrow E\left(\frac{n_1}{n}\right) = \alpha$$

$$Var(n_1) = n\alpha(1-\alpha) \Rightarrow Var\left(\frac{n_1}{n}\right) = \frac{\alpha(1-\alpha)}{n}$$

$$\therefore \quad E(\hat{\phi}) = \frac{E\left(\frac{n_1}{n}\right) - (1-q)\delta}{q} = \frac{\alpha - (1-q)\delta}{q} = \frac{q\phi + (1-q)\delta - (1-q)\delta}{q}$$

$$E(\hat{\phi}) = \phi$$

Scanned with CamScanner

$$\text{Var}(\hat{\phi}) = \frac{\text{Var}\left(\frac{n_1}{n}\right)}{q^2} = \frac{\alpha(1-\alpha)}{nq^2}$$

$$= \frac{\{q\phi + (1-q)\delta\}\{1 - q\phi - (1-q)\delta\}}{nq^2}$$

$$= \frac{1}{nq^2} \{q\phi + (1-q)\delta\}\{1 + q(1-\phi) - (1-q)\delta - q\}$$

$$= \frac{1}{nq^2} \{q\phi + (1-q)\delta\}\{q(1-\phi) + (1-q)(1-\delta)\}$$

$$= \frac{q^2\phi(1-\phi)}{nq^2} + \frac{(1-q)^2\delta(1-\delta)}{nq^2} + \frac{q\phi(1-q)(1-\delta) + q(1-\phi)\delta(1-q)}{nq^2}$$

$$= \frac{\phi(1-\phi)}{n} + \frac{(1-q)^2}{nq^2}\delta(1-\delta) + \frac{(1-q)}{nq}\{\phi(1-\delta) + \delta(1-\phi)\}$$

Case : 2: δ unknown but $\hat{\delta}$ is available with known variance estimator $\widehat{\text{Var}}(\hat{\delta})$

The data collection procedure is same, here we have the following results,

$$\hat{\phi} = \frac{\frac{n_1}{n} - (1-q)\hat{\delta}}{q}$$

$$\text{Var}(\hat{\phi}) = \frac{\frac{\alpha(1-\alpha)}{n} + (1-q)^2\,\text{Var}(\hat{\delta})}{q^2}$$ [ignoring the covariance between $\hat{\delta}$ and $\frac{n_1}{n}$]

$$= \frac{\alpha(1-\alpha)}{nq^2} + \frac{(1-q)^2}{q^2}\text{Var}(\hat{\delta})$$

and $$\widehat{\text{Var}(\hat{\phi})} = \frac{\hat{\alpha}(1-\hat{\alpha})}{nq^2} + \frac{(1-q)^2}{q^2}\widehat{\text{Var}(\hat{\delta})}$$

Case 3: δ is unknown and no ~~pro~~ prior estimate of it is available

Here if the same randomization device is followed in connecting data from n sampled units, the expression $(*)$ will continue

$$\hat{\phi} = \frac{\frac{n_1}{n} - (1-q)\delta}{q}$$

with $Var(\hat{\phi}) = \frac{\alpha(1-\alpha)}{nq^2}$ where $\alpha = q\phi + (1-q)\delta$ —$(*)$

Having an unbiased estimator of α in the usual manner will not permit us to extract exactly one unbiased estimator of ϕ from the RHS of $(*)$, Since we have one equations and two unknowns. So the idea is to generate two consistent and independent linear equations of the type. This can be achieved as follows:

We split the total sample size n into two parts n_1 and n_2, next we perform the randomised response survey on a group n_1 sample units selected according to SRSWR using probabilities q_1 and $(1-q_1)$ of selecting the questions Q_1 and Q_2 respectively. We carry out as similar survey on another group of n_2 sample units, selected independently from the entire population again using SRSWR. For this group however the randomization device is made to select the questions Q_1 and Q_2 with probabilities q_2 and $(1-q_2)$ respectively with $q_1 \neq q_2$.

This randomization procedure results in two equations of the form

$$\alpha_1 = q_1\phi + (1-q_1)\delta$$

$$\alpha_2 = q_2\phi + (1-q_2)\delta$$

let us denote by n_{11} and n_{21}, the no. of 'yes' responses for the two groups respectively. Then it is clear that $\frac{n_{11}}{n_1}$ is an u.e. of α_1 with variance $\frac{\alpha_1(1-\alpha_1)}{n_1}$

Similarly, $\frac{n_{21}}{n_2}$ is an u.e. of α_2, with variance $\frac{\alpha_2(1-\alpha_2)}{n_2}$

$\therefore$ To estimate ϕ, we solve the equations,

$$\frac{n_{11}}{n_1} = q_1\hat{\phi} + (1-q_1)\hat{\delta}$$

$$\frac{n_{21}}{n_2} = q_2\hat{\phi} + (1-q_2)\hat{\delta}$$

which gives, $q_2 q_1 \hat{\phi} + q_2(1-q_1)\hat{\delta} = q_2 \frac{n_{11}}{n_1}$ — ①

$q_1 q_2 \hat{\phi} + q_1(1-q_2)\hat{\delta} = q_1 \frac{n_{21}}{n_2}$ — ②

① - ② $\hat{\delta}\left\{q_2(1-q_1) - q_1(1-q_2)\right\} = q_2\frac{n_{11}}{n_1} - q_1\frac{n_{21}}{n_2}$

ie. $\hat{\delta} = \dfrac{q_2\frac{n_{11}}{n_1} - q_1\frac{n_{21}}{n_2}}{q_2 - q_1}$

Again, $\hat{\phi} = \dfrac{\frac{n_{21}}{n_2} - (1-q_2)\hat{\delta}}{q_2} = \dfrac{\frac{n_{21}}{n_2} - (1-q_2)\left\{\dfrac{q_2\frac{n_{11}}{n_1} - q_1\frac{n_{21}}{n_2}}{(q_2-q_1)}\right\}}{q_2}$

~~$= \hat{\phi}\left\{\frac{1}{q_2(q_2-q_1)}\left[\frac{n_{21}}{n_2}q_2 - \frac{n_{11}}{n_1}q_1 - q_2\frac{n_{11}}{n_1} + q_1\frac{n_{21}}{n_2} + q_2\left(q_2\frac{n_{11}}{n_1} - q_1\frac{n_{21}}{n_2}\right)\right]\right.$~~

$$= \frac{1}{q_2(q_2-q_1)}\left[\frac{n_{21}}{n_2}q_2 - \cancel{\frac{n_{21}}{n_2}q_1} - \frac{n_{11}}{n_1}q_2 + \cancel{\frac{n_{21}}{n_2}q_1} + q_2\left(q_2\frac{n_{11}}{n_1} - q_1\frac{n_{21}}{n_2}\right)\right]$$

$$\hat{\phi} = \frac{(1-q_1)\frac{n_{21}}{n_2} - (1-q_2)\frac{n_{11}}{n_1}}{(q_2-q_1)}$$

$$\hat{\phi} = \frac{(1-q_2)\frac{n_{11}}{n_1} - (1-q_1)\frac{n_{21}}{n_2}}{q_1-q_2}$$

$$Var(\hat{\phi}) = \frac{1}{(q_1-q_2)^2}\left[(1-q_2)^2\frac{\alpha_1(1-\alpha_1)}{n_1} + (1-q_1)^2\frac{\alpha_2(1-\alpha_2)}{n_2}\right]$$

For a fix total number of obeervations, $n = n_1+n_2$ and for fixed valneds of q_1, q_2 and δ find the optimum n_1 and n_2 that minimized the variance, we have $n = n_1+n_2$, so $n_2 = n-n_1$

Thus, $\operatorname{Var}(\hat{\phi}) = \dfrac{(1-q_2)^2 \, \dfrac{\alpha_1(1-\alpha_1)}{n_1} + (1-q_1)^2 \, \dfrac{q_2(1-\alpha_2)}{n-n_1}}{(q_1-q_2)^2}$

Randomized Response Techrique for one polychotomous qualitative characteristic

We assume that the sensitive characteristic has t versions (some of them being stigmatizing in nature) and accordingly the population is partitioned into t mutually exclusive and exhaustive parts $U_1, U_2, \ldots, U_t$. We denote by ϕ_i the true proportion (unknown) of individuals in the subpopulation U_i. Clearly, $\sum_{i=1}^{t} \phi_i = 1$.

Related Question Method

We prepare $(t-1)$ randomization devices denoted by $D_1, D_2, \ldots, D_{t-1}$. Under any device, the respondent is supposed to respond any of the questions $Q_1, Q_2, \ldots Q_t$ where,

Q_j :- Do you belong to U_j ? $(j = 1(1)t)$

However, under the device i, the chance of selecting Q_j is π_{ij}, where π_{ij}'s are so chosen that

(i) $0 \leq \pi_{ij} \leq 1$

(ii) $\sum_{j=1}^{t} \pi_{ij} = 1$ for each $i = 1, 2, \ldots, t-1$

(iii) the matrix $\underset{(t-1)\times(t-1)}{A} = \left(\left(\pi_{ij} - \pi_{it}\right)\right)$, $i = 1, 2, \ldots, \overline{t-1}$, $j = 1, 2, \ldots, \overline{t-1}$

is ~~orthogonal~~. non-singular.

We select $(t-1)$ independent samples of size $n_1, n_2, \ldots, n_{t-1}$ respectively, each drawn according to SRSWR. Each member of the first set of n_1 selected respondents is given the randomization device D_1 and requested to report the response of "yes" or "no" depending on whether or not he/she belongs to the category yes or no" for the randomly chosen question. Similarly the other randomization device are used to collect data from the other groups of respondents.

Estimation of $\phi_1, \phi_2, \dots, \phi_t$'s

Clearly, it is enough to estimate $\phi_1, \phi_2, \dots, \phi_{t-1}$, since

$$\phi_t = 1 - \phi_1 - \phi_2 - \dots - \phi_{t-1}$$

Notation: n_{i1} = The no. of 'yes' responses from the individuals using the randomization device D_i

$$\alpha_i = \sum_{j=1}^{t} \pi_{ij}\, \phi_j$$

$$y_i = \alpha_i - \pi_{it} \quad , i = 1(1)\overline{t-1}$$

$$\underset{\overline{t-1}\times 1}{\underset{\sim}{y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{t-1} \end{pmatrix} \quad , \quad \underset{\overline{t-1}\times 1}{\underset{\sim}{\Phi}} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{t-1} \end{pmatrix}$$

Result

An unbiased estimator of $\underset{\sim}{\Phi}$ is given by

$$\hat{\underset{\sim}{\Phi}} = A^{-1} \hat{\underset{\sim}{y}}$$

where, $\hat{\underset{\sim}{y}} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_{t-1} \end{pmatrix}$, $\hat{y}_i = \dfrac{n_{i1}}{n_i} - \pi_{it}$ $\left[\because \hat{\alpha}_i = \dfrac{n_{i1}}{n_i} \right.$, $i = 1(1)\overline{t-1}$

Proof:

$$P(\text{Yes Response from device } i) = \sum_{j=1}^{t} P(j\text{th question selected from device } i)\; P(\text{Yes} \mid j\text{th question selected from device } i)$$

$$= \sum_{j=1}^{t} \pi_{ij}\, \phi_j$$

$$= \alpha_i$$

α_i is unbiasedly estimated by $\alpha_i = \frac{n_{i1}}{n_i}$, $i = 1(1)\overline{t-1}$

We have $(t-1)$ equations

$$\frac{n_{i1}}{n_i} = \sum_{j=1}^{t} \pi_{ij}\hat{\phi}_j \quad (i = 1(1)\overline{t-1})$$

or, $$\frac{n_{i1}}{n_i} = \sum_{j=1}^{t-1} \pi_{ij}\hat{\phi}_j + \pi_{it}\hat{\phi}_t$$

$$= \sum_{j=1}^{t-1} \pi_{ij}\hat{\phi}_j + \pi_{it}\left(1 - \sum_{j=1}^{t-1}\hat{\phi}_j\right)$$

$$= \sum_{j=1}^{t-1} \left(\pi_{ij} - \pi_{it}\right)\hat{\phi}_j + \pi_{it}$$

or, $$\sum_{j=1}^{t-1} \underbrace{\left(\pi_{ij} - \pi_{it}\right)}_{a_{ij}}\hat{\phi}_j = \frac{n_{i1}}{n_i} - \pi_{it} = \hat{\ell}_{y_i}$$

or, $$\sum_{j=1}^{t-1} a_{ij}\hat{\phi}_j = \hat{\ell}_{y_i} \quad , i = 1(1)\overline{t-1}$$

The system of equations can be written in term of matrix notation as

$$A\hat{\underset{\sim}{\Phi}} = \hat{\underset{\sim}{\ell_y}}$$

or. $$\hat{\underset{\sim}{\Phi}} = A^{-1}\hat{\underset{\sim}{\ell_y}}$$ [Since, π_{ij} are so chosen that A is non-singular]

Variance of $\hat{\underset{\sim}{\Phi}}$

$$\text{Disp}\left(\hat{\underset{\sim}{\Phi}}\right) = A^{-1}\,\text{Disp}\left(\hat{\underset{\sim}{\ell_y}}\right)\left(A^{-1}\right)'$$

$$\underset{\overline{t-1}\times 1}{\hat{\underset{\sim}{\ell_y}}} = \begin{pmatrix} \hat{\ell}_{y_1} \\ \vdots \\ \hat{\ell}_{y_{t-1}} \end{pmatrix} , \quad \hat{\ell}_{y_i} = \frac{n_{i1}}{n_i} - \pi_{it}$$

$$\text{Var}\left(\hat{\ell}_{y_i}\right) = \text{Var}\left(\frac{n_{i1}}{n_i}\right)$$

$n_{i1} \sim \text{Bin}\left(n_i, \alpha_i\right)$ independently

So, $Var\left(\hat{\ell}_{y_i}\right) = \dfrac{\alpha_i(1-\alpha_i)}{n_i}$ $, i = 1(1)\,\overline{t-1}$

& $Cov\left(\hat{\ell}_{y_i}, \hat{\ell}_{y_j}\right) = 0 \quad \forall\, i \neq j = 1(1)\,\overline{t-1}$

$$\therefore\ Disp\left(\hat{\underset{\sim}{\ell_y}}\right) = Diag\left(\frac{\alpha_1(1-\alpha_1)}{n_1}, \frac{\alpha_2(1-\alpha_2)}{n_2}, \dots, \frac{\alpha_{t-1}(1-\alpha_{t-1})}{n_{t-1}}\right)$$

Hence,

$$Disp\left(\hat{\underset{\sim}{\Phi}}\right) = A^{-1}\, Diag\left(\frac{\alpha_1(1-\alpha_1)}{n_1}, \frac{\alpha_2(1-\alpha_2)}{n_2}, \dots, \frac{a_{t-1}(1-\alpha_{t-1})}{n_{t-1}}\right)(A^{-1})'$$

Remark :

1. Here we explicitly assume that our randomization devices $D_1, \dots, D_{t-1}$ result in non-singularity of the square matrix A. Indeed there are plenty of π_{ij}'s satisfying the above requirements, For example, we may choose $\pi_{it} = 0$ for $i = 1(1)t$ and choose the other π_{ij}'s such that A is a circulant matrix of the form

$$A = \begin{pmatrix} a_1 & a_2 & a_3 & \cdots & a_{t-1} \\ a_{t-1} & a_1 & a_2 & \cdots & a_{t-2} \\ \vdots & \vdots & \vdots & & \vdots \\ a_2 & a_3 & \cdots & a_{t-1} & a_1 \end{pmatrix}$$

Here, $0 \leq a_j \leq 1$ and $\sum_{j=1}^{t-1} a_j = 1$ with atleast of. the a_j's being strictly positive,

2. To differentiate among various possible choices of the randomization devices $D_1, \dots, D_{t-1}$, it is necessary to formulate a criterion for comparing the relative performances of the estimators $\hat{\underset{\sim}{\Phi}}$ based on the various choices, Since, the precision of the estimators depends on the choice of A, a rational way of choosing A would have been to minimize the determinant of the covariance matrix of the estimators.

Unrelated Question Method with 'yes' or 'no' response

Here we assume that there are t stigmatizing versions of the sensitive characteristic thereby partitioning the population $\mathcal{U}$ into t subpopulations $\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_t$. However, there is ~~no~~ one unrelated dichotomous characteristic with unknown incidence rate δ for one of its versions, say V.

Procedure of estimating the proportions by "Greenbarg's procedure"

Using SRSWR scheme of sampling, we draw t independent samples of sizes $n_1, n_2, \dots, n_t$ respectively from t randomization devices $D_1, D_2, \dots, D_t$.

⊡ Each member of the ith group containing n_i individuals asked to select one of the t alternative questions $Q_1, Q_2, \dots, Q_t$ with the respective selection probabilities $\pi_{i1}, \pi_{i2}, \dots, \pi_{it}$ $(i=1(1)t)$, where

Q_j :- Do you belong to $\mathcal{U}_j$? $(j = 1(1)\overline{t-1})$

and Q_t :- Do you belong to V?

Define

ϕ_j : population proportion of individuals belonging to $\mathcal{U}_j$ $(j=1(1)t)$

δ : population proportion of individuals belonging to V

$$\underset{\sim}{\Phi} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{t-1} \end{pmatrix}$$

$\therefore$ P (Yes response from ith randomization device)

$$= \sum_{j=1}^{t} P(Q_j \text{ is selected from ith device})\; P(\text{Yes} \mid Q_j \text{ is selected})$$

~~$= \sum_{j=1}^{t-1} q_{ij}\phi_j + q_{it}\delta$~~ $= \sum_{j=1}^{t-1} \pi_{ij}\phi_j + \pi_{it}\delta$

$= \alpha_i^*$ (say)

α_i^* is unbiasedly estimated by $\hat{\alpha}_i^* = \dfrac{n_{i1}}{n_i}$

where n_{i1} is the number of individuals giving 'yes' response from device D_i

So we have t equations with t unknowns

$$\frac{n_{i1}}{n_i} = \sum_{j=1}^{t-1} \pi_{ij}\hat{\phi}_i + \pi_{it}\hat{\delta}$$

In matrix notation, $A^* \begin{pmatrix} \hat{\underset{\sim}{\Phi}} \\ \hat{\delta} \end{pmatrix} = \hat{\underset{\sim}{\alpha}}^*$

where, $\underset{t\times t}{A^*} = \left((\pi_{ij})\right)_{t\times t}$ and $\hat{\underset{\sim}{\alpha}}^* = \begin{pmatrix} \hat{\alpha}_1^* \\ \hat{\alpha}_2^* \\ \vdots \\ \hat{\alpha}_t^* \end{pmatrix}$,

π_{ij}'s are so choosen that A^* is non-singular

$$\therefore \begin{pmatrix} \hat{\underset{\sim}{\Phi}} \\ \hat{\underset{\sim}{\delta}} \end{pmatrix} = A^{*-1}\hat{\underset{\sim}{\alpha}}^*$$

Unrelated Question Method with multiple response

This is a procedure due to Bourke applicable in situations where the respondents are willing to participate in a multiple choice survey provided an unrelated question with the same number of multiple choices can be blended with the sensitive characteristic. Suppose that there are t stigmatizing versions of the sensitive characteristic as well as the unrelated characteristic. Let U_i & U_i^* respectively denote the subpopulations corresponding to the ith component of the sensitive characteristic ~~as well as~~ and the innocuous characteristic $(i=1(1)t)$. It is assumed that the proportions $\delta_1, \delta_2, \dots, \delta_t$ of $U_1^*, U_2^*, \dots, U_t^*$ are known. It is described to estimate the proportions $\phi_1, \phi_2, \dots, \phi_t$ based on the data collected from an SRSWR sample of size n.

We prepare a randomization device using, for example, a deck of cards. A known proportion π of the cards are of type 1, while the rest are of type 2. Each respondent is supposed to select one card at random. If the selected card is of type 1, then the respondent has to pick up the sensitive issue and is assumed to provide a truthful response to which category he/she belongs. If the selected card is of type-II, then the respondent has to pick up the unrelated question and provide a similar true response. Thus, for example, a respondent belonging to either of the categories U_i and U_i^*, would simply respond with the serial number i for data recording

$\therefore$ P (Getting Response i) = P (Sensitive issue selected) P (Response is i | Sensitive issue selected) + P (unrelated issue selected) P (Response is i | Unrelated issue selected)

$$= \pi \phi_i + (1-\pi)\delta_i$$

$$= \alpha_i \text{ (say)}$$

α_i is unbiasedly estimated by $\hat{\alpha}_i = \frac{n_i}{n}$, which is the sample proportion of individuals giving response i $(i = 1(1)t)$

Thus we obtain t equations in t unknowns,

$$\pi \hat{\phi}_i + (1-\pi)\hat{\delta}_i = \frac{n_i}{n}$$

$$\Rightarrow \hat{\phi}_i = \frac{\frac{n_i}{n} - (1-\pi)\delta_i}{\pi}$$

$$\Rightarrow Var(\hat{\phi}_i) = \frac{1}{\pi^2} Var(\hat{\alpha}_i) = \frac{\alpha_i(1-\alpha_i)}{n\pi^2}$$

Since, $(n_1, n_2, \ldots, n_t) \sim$ Multinomial $(n, \alpha_1, \alpha_2, \ldots, \alpha_t)$