VIDYASAGAR UNIVERSITY


# A Comparison between two Methods of Estimating Population Proportion

Dissertation Paper (DSE 4)

REGISTRATION NO.:1160531 OF 2020-2021

ROLL:1126116 NO.:200144

DEPT. OF STATISTICS, HALDIA GOVT. COLLEGE

# Acknowledgement: -

The success and final outcome this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along with the completion of my project. All that I have done is only due to their supervision and assistance and I would not forget to thank them. I respect and thank for providing me an opportunity to do the project work under the department of statistics, Haldia Government College and giving you all the support and guidance that I required, which made me complete the project duly on time. I owe my deep gratitude to our project guides Dr. Shyamsundar Sahoo, Mr. Sibsankar Karan, Mr. Tanmay Maity, Mr. Bijitesh Halder, who took keen interest on my project work and guided me all along, till the completion of our project work by providing all the necessary information for developing a good system. I would not forget to remember my parents, and also my friends for their encouragement and more over for their timely support and guidance till the completion of our project work.

Contents: -

# OBJECTIVE:-

The objective of the project is to compare the effectiveness of two methods for estimating the population proportion viz. Estimating population proportion by indirect sampling and estimating population proportion by Randomize response technique. We have taken the Werner's model as our Randomized Response Technique.

Now, although the latter is used when the character of study is sensitive to answer directly or the survey may relate to stigmatizing issues, here we have taken the character less sensitive and collected the data in such a way that we can use both the methods.

Thus, we have considered two methods on same and compared them with purpose of answering the question "**Have you cheated in your last semester exam**?"

# INTRODUCTION:-

As the objective says we are to do a comparative study of two method we have learnt so far for estimating population proportion. First being the indirect sampling and the second, the Warner's model which is a model to estimate population proportion by randomize response technique.

Now with this objective we collect data from two different groups of respondents where we asked each groups different groups accordingly. For the indirect sampling part unlike case of Warner's model we have to ask the respondent a less sensitive question which could be answer by the respondents without any hesitation, ensuring the availability of the actual population proportion value. And here in this project, we have asked them if they done any type of cheating in their last university examination.

# Data source: -

The raw data for the two different procedures was collected in two different modes. The first being online mode i.e., through goggle form and later by offline mode.

Offline mode data collection procedure:

1. Firstly, we would arrange a well shuffled deck of 35 cards with 20 red cards and 15 black cards.
2. Each respondent will pick a card and answer yes or no according to the rule given below

Rule:

Red card: Answer the question, "**Have you cheated in your last university exam?**"

**Black card**: Answer the question, "**Have you not cheated in your last university exam?**"

Such 100 response was collected: -

| Sl No. | Responses | | Sl No. | Response |
|---|---|---|---|---|
| **1** | yes | | 51 | no |
| **2** | no | | 52 | yes |
| **3** | yes | | 53 | yes |
| **4** | no | | 54 | no |
| **5** | no | | 55 | no |
| **6** | yes | | 56 | no |
| **7** | yes | | 57 | no |
| **8** | yes | | 58 | yes |
| **9** | no | | 59 | yes |
| **10** | no | | 60 | no |
| **11** | yes | | 61 | yes |
| **12** | yes | | 62 | no |
| **13** | yes | | 63 | no |

| | | | | |
|---|---|---|---|---|
| **14** | no | | 64 | yes |
| **15** | yes | | 65 | no |
| **16** | no | | 66 | yes |
| **17** | no | | 67 | no |
| **18** | no | | 68 | no |
| **19** | yes | | 69 | no |
| **20** | yes | | 70 | yes |
| **21** | no | | 71 | no |
| **22** | yes | | 72 | yes |
| **23** | yes | | 73 | yes |
| **24** | no | | 74 | yes |
| **25** | no | | 75 | no |
| **26** | no | | 76 | yes |
| **27** | yes | | 77 | no |
| **28** | yes | | 78 | yes |
| **29** | yes | | 79 | yes |
| **30** | no | | 80 | yes |
| **31** | no | | 81 | no |
| **32** | no | | 82 | yes |
| **33** | no | | 83 | yes |
| **34** | yes | | 84 | yes |
| **35** | no | | 85 | no |
| **36** | no | | 86 | no |
| **37** | no | | 87 | yes |
| **38** | yes | | 88 | yes |

| | | | | |
|---|---|---|---|---|
| **39** | yes | | 89 | no |
| **40** | no | | 90 | no |
| **41** | yes | | 91 | yes |
| **42** | no | | 92 | yes |
| **43** | no | | 93 | yes |
| **44** | no | | 94 | yes |
| **45** | yes | | 95 | no |
| **46** | yes | | 96 | no |
| **47** | yes | | 97 | no |
| **48** | yes | | 98 | yes |
| **49** | yes | | 99 | yes |
| **50** | yes | | 100 | yes |

## Online data collection (Indirect sampling):

This data was collected via goggle form. We have asked the respondents five questions, related to their overall behavior in their exam hall including question to cross validate their previous response

**Goggle form link**: https://forms.gle/vQBMTSP2JraCpN659

77 Responses:

| Did you find your last university exam difficult? | Did you complete the exam entirely on your own? | Did you refer to any external materials during the exam? | Did you discuss the exam questions repeatedly with any classmates during the exam? | Was your exam better than your expectation? |
|---|---|---|---|---|
| **No** | No | No | Yes | Yes |
| **No** | Yes | No | Yes | No |

| | | | | |
|---|---|---|---|---|
| **No** | No | No | Yes | No |
| **No** | Yes | No | Yes | Yes |
| **No** | No | No | Yes | Yes |
| **No** | Yes | No | Yes | No |
| **No** | Yes | No | No | No |
| **Yes** | No | No | Yes | Yes |
| **Yes** | Yes | No | No | Yes |
| **No** | Yes | No | No | No |
| **No** | Yes | No | No | Yes |
| **No** | Yes | No | No | Yes |
| **No** | Yes | No | Yes | No |
| **Yes** | Yes | No | Yes | No |
| **No** | Yes | No | No | No |
| **Yes** | Yes | No | No | Yes |
| **No** | Yes | No | No | Yes |
| **No** | Yes | No | No | No |
| **No** | Yes | No | Yes | Yes |
| **Yes** | Yes | No | No | No |
| **Yes** | Yes | No | No | No |
| **No** | No | No | Yes | Yes |
| **No** | Yes | Yes | No | Yes |
| **No** | Yes | No | No | Yes |
| **No** | No | Yes | Yes | No |
| **Yes** | Yes | No | Yes | Yes |
| **Yes** | Yes | No | Yes | No |

| | | | | |
|---|---|---|---|---|
| **Yes** | No | No | No | Yes |
| **No** | No | No | No | No |
| **Yes** | Yes | No | No | No |
| **No** | Yes | No | No | Yes |
| **No** | Yes | No | No | No |
| **No** | No | Yes | No | No |
| **No** | Yes | No | No | No |
| **Yes** | Yes | No | Yes | No |
| **Yes** | Yes | Yes | Yes | Yes |
| **No** | Yes | No | No | Yes |
| **No** | Yes | No | Yes | Yes |
| **Yes** | Yes | No | No | No |
| **Yes** | Yes | No | Yes | No |
| **Yes** | Yes | No | No | No |
| **No** | Yes | No | No | Yes |
| **Yes** | Yes | No | No | No |
| **Yes** | No | No | Yes | No |
| **No** | Yes | No | Yes | Yes |
| **No** | Yes | No | No | No |
| **No** | Yes | No | Yes | Yes |
| **Yes** | Yes | No | No | No |
| **Yes** | Yes | No | Yes | Yes |
| **No** | No | No | Yes | Yes |
| **No** | No | No | Yes | Yes |
| **Yes** | No | No | Yes | No |

| | | | | |
|---|---|---|---|---|
| **Yes** | Yes | Yes | Yes | No |
| **No** | Yes | No | No | Yes |
| **Yes** | Yes | No | No | Yes |
| **No** | Yes | No | Yes | Yes |
| **No** | Yes | No | Yes | Yes |
| **Yes** | Yes | No | No | No |
| **Yes** | Yes | No | No | No |
| **Yes** | Yes | No | No | No |
| **No** | No | No | Yes | No |
| **No** | No | No | Yes | Yes |
| **Yes** | No | Yes | Yes | No |
| **Yes** | No | No | Yes | No |
| **No** | Yes | No | Yes | Yes |
| **No** | No | No | Yes | Yes |
| **No** | No | No | Yes | No |
| **No** | No | No | Yes | No |
| **No** | Yes | No | No | Yes |
| **No** | Yes | No | No | No |
| **No** | Yes | No | Yes | Yes |
| **No** | No | Yes | Yes | Yes |
| **Yes** | Yes | No | No | No |
| **Yes** | Yes | No | Yes | Yes |
| **Yes** | No | No | Yes | Yes |
| **No** | Yes | No | No | Yes |
| **Yes** | Yes | No | No | No |

# Estimating P by the method of direct sampling:-

# Theory:-

Here we have a finite population size N (known), of which some units, say $N_1$ (unknown)``, cheated in their last exam and the rest N- $N_1$ did not cheated in their last exam

Then, $p = \frac{N_1}{N}$

Is the population is the population proportion or individual possessing that character.

Therefore, $N_1$=N.P

And N-$N_1$=N- N. P=N (1-P) =NQ (say)

Purpose:

To estimate P (or equivalently, N1 when N is known) using a sample size n (say drawn by employing the sampling scheme SRSWR.

Let us define a marker variable y on the unit of the population as follows:

$$Y_i = \begin{cases} 1, \text{if 'i'th student cheated in his/her exam} \\ 0, \ otherwise, for\ i = 1(1)N \end{cases}$$

[Note that then the population would consist of 0's and 1's]

Then,$\sum_1^N Y_i$ = $N_1$ [ $N_1$ population values are 1 and rests are 0]

$\Rightarrow \bar{Y} = \frac{N_1}{N} = P$

And the population variance $Y_\alpha^2$

$\sigma^2 = \frac{\sum Y_i^2 - \bar{Y}^2}{N}$

$= \frac{N_1}{N} - P^2$ [$As\ Y_\alpha$ = 1, if its unit cheated in their exam]

$= P\ (1 - P) = PQ$

To estimate the P let us draw a sample size n (say) by SRSWR.

If $y_i$ denotes the value of y for the ith unit,

Then,

$y_i$= 1, if ith selected unit cheated in their exam

0, otherwise for i=1(1)n

Then, $\sum_1^n yi$=$n_1$ (say) is the number of persons who have cheated in their exam and once the sample is taken $n_1$ is known.

It may be noted that p= $\frac{n1}{n}$

Is the sample proportion of the unit possessing character A (and can be computed from the sample observations).

So the sample mean is $\bar{y}$= $\frac{\Sigma yi}{n}$= $\frac{n1}{n}$= $p$

And also the sample variance is $s^2$= $\frac{\sum y_i^2 - \bar{y}^2}{n}$

$= \frac{n_1}{n} - p^2$

$= p\ (1 - p)$

$= pq$ [with q = (1-p)]

And $s'^2 = n\ (pq)\ /\ (n-1)$

In case of SRSWR we know that $E\ (\bar{y}) = \bar{Y}$

Hence, $E\ (p) = P$

Thus $\hat{P} = p$, i.e. an unbiased estimator of P is p.

Also, for SRS we have proved the results $var\ (\bar{y})$ = $\sigma^2/n$

So $\widehat{varwr}(p) = PQ/$ n

Since, in case of SRSWR $(s'\ 2) = \sigma^2$

We have $PQ = npq/\ (n-1)$

And hence, an unbiased estimator of $varwr\ (p)$ will be

$\widehat{varwr}(p) = \frac{pq}{n-1}$

**Computation:**

Now from the observed sample

$n_1$=46

Estimate of the proportion of student took sick leave without being sick using direct response is

p=$\frac{46}{77}$=0.5974

And estimate of $varw(p)$ will be

$\widehat{\text{varwr}}(p) = \frac{pq}{n-1}$=0.000512

# Estimating P by Warner's Model:-

Suppose it is desired to estimate the proportion (P) of a population belonging to a class $C$. Let, the class of the population not having character $C$ be donated by $\bar{C}$.

The respondent is given a spinner with a mark, so that the spinner points to the letter $C$ with probability $p0$ (known) and to $\bar{C}$with probability $q0= (1- p0)$.

The respondents is required to spin the spinner, unobserved by the interviewer and report only whether or not the spinner points to the letter representing the group he belongs to.

Suppose an SRSWR of n respondents is selected to estimate P.

Let, $yi$ be the response for the i-th selected individual according to the rule given below –

$yi$ = 1 if spinner points $C$ and the individual belongs to the class $C$ or, if the spinner points $\bar{C}$and the individual belongs to $\bar{C}$.

= 0 for any other cases [∀ i=1(1) n]

$yi$ = 1 with probability $p0. P+ (1-p0). (1-P) = \pi$

= 0 with probability (1-π) [∀ i=1(1) n]

Now, if $n_c$ be number of individuals responding 1

Then, $n_c=\sum_1^n yI$ and $nc$ ~ Binomial (n, π)

$\therefore (nc) = n.\pi$

$\Rightarrow \hat{\pi} = nc/n= pc$ (say)

Now, we know, $\pi = p0.P+ (1-p0).(1-P)=(p0- q0)P+ q0$

$\Rightarrow (p0- q0) \hat{P}= \hat{\pi}- q0= pc- q0$

$\Rightarrow \hat{P}= (\boldsymbol{pc- q0})/(\boldsymbol{p0- q0})$

Now, $(\hat{P})=V ((pc- q0)/ (p0- q0))$

$= V (pc)/ (p0- q0)2$

$= 1 (po-qo)^2 V (nc/n)$

$= (nc) / n^2 (p0-q0)^2$

$= n. \pi (1-\pi)/n^2(p0-q0)^2$ [As $nc$ ~ Binomial (n, π)]

$= \hat{\pi}.(1-\hat{\pi})/n.(p0-q0)2$

Now, $(nc)=n.\hat{\pi}$

and, $V(nc)=n.\hat{\pi}.(1-\hat{\pi})$

$\Rightarrow E(n_c^2)-E^2(n_c)$

$=n.\hat{\pi}-n.\hat{\pi}^2$

$\Rightarrow E(n_c^2)-n^2\hat{\pi}^2$

$=E(nc)-n.\hat{\pi}^2$

$\Rightarrow E(n_c^2)--E(n_c)$

$=n^2\hat{\pi}^2-n.\hat{\pi}2$

$=\hat{\pi}^2(n^2-n)$

$\Rightarrow E\left(\frac{n_c^2-n_c}{n^2-n}\right)$

$=\hat{\pi}^2$

So, from $E(n_c/n)=\hat{\pi}$

and $E\left(\frac{n_c^2-n_c}{n^2-n}\right)=\hat{\pi}^2$

$E\left(\frac{n_c}{n}\right)-E\left(\frac{n_c^2-n_c}{n^2-n}\right)=\hat{\pi}-\hat{\pi}^2$

$\Rightarrow E\left(\frac{n_c}{n}-\frac{n_c^2-n_c}{n^2-n}\right)=\hat{\pi}.(1-\hat{\pi})$

$\Rightarrow E\left(\frac{n.nc-nc^2}{n(n-1)}\right)=\hat{\pi}.(1-\hat{\pi})$

$\Rightarrow E\left(\frac{n^2\left(\frac{n_c}{n}-\frac{nc^2}{n^2}\right)}{n(n-1)}\right)=\hat{\pi}.(1-\hat{\pi})$

$\Rightarrow E\left(\frac{n(p_c-p_c^2)}{n-1}\right)=\hat{\pi}.(1-\hat{\pi})$

$\therefore \pi.(1-\pi)=\frac{n(p_c-p_c^2)}{n-1}$

So, $\boldsymbol{V(P)}=\frac{\pi.(1-\pi)}{n(p0-q0)^2}$

$=\frac{1}{n(p_0-q_0)^2}\times\frac{n(p_c-p_c^2)}{n-1}$

$$= \frac{\mathbf{pc\,(1-pc)}}{\mathbf{(n-1)\,(p0-q0)\,2}}$$

Here we tackle our 'spinning a spinner' situation by providing each respondent with a deck containing 35 cards with 20red cards and 15 black cards and the respondents picking a red cards pointing to letter 'C 'and black cards pointing '$\bar{C}$' and C denotes the class of students who have cheated in their exam.

So $p_0$=20/35=0.5714

$q_0$=15/35=0.4285

$n_c$=$\sum_1^n y$i=53

$p_c$ =$nc/n$

= 53/100 = 0.53

$\therefore \frac{p_c - q_0}{p_0 - q_0} = 0.71028$

So, the estimate of population proportion possessing the character under study by the Warner's Model is –

$P=\frac{p_c - q_0}{p_0 - q_0} =$ 0.71028

The estimate of the variance of the population proportion possessing the character of study

$V(P) = p_c(1-p_c)/(n-1)(p0-q0)2$ =0.12321

# Conclusion:

We can clearly see that the results differ in the two cases.

• While comparing the estimates of the variance of the estimator we see that the estimate is 0.000512in the first case while it is 0.12321 in the case of Warner's Model.

So, we can say that variance in the case of the method of direct sampling is significantly less than the second method.

Hence, by using the sense of variability, the first method is better. So, comparing the two methods, we can easily conclude that the method of estimation of population proportion by **direct sampling method is more efficient than the method of estimation by Warner's Model.**

# Discussion:

′Theoretical Support:

We can represent the variance of the estimator in the case of Warner's Model in a different way –

$$V(\widehat{P}) = V\left(\frac{p_c - q0}{p0 - q0}\right)$$

$$= V(pc)/(p0 - q0)\,2$$

$$= \pi(1-\pi)/n(p0-q0)2$$

$$= p0q0 + (p0-q0)2.\frac{P(1-P)}{n(p0-q0)2}$$

$$= p0.P + (1-p0).(1-P)]$$

$$= P(1-P)/n + p0q0\,/n(p0-q0)\,2$$

Here, the first term of the above expression is the variance of the estimator of the population proportion when the question can be asked directly i.e., as in the first case.

And the second term represents the increase in the variance due to the fact that the question has been posed indirectly.

So, we can clearly see that the variance of the estimator increases by a non-negative quantity when we shift our method from direct sampling to Warner's Model concluding that the effectivity of the former will be greater than the latter. And here this project also we have seen the same conclusion, supporting the above-mentioned theoretical conclusion by practical methods.

# Variation in $p_0$:

We have seen the expression-

$$V(\widehat{p}) = \frac{p(1-p)}{n} + \frac{p_0 q_0}{n(p_0 - q_0)^2}$$

Note that if $p_0 = 1 \; or \; 0$, the second term in the above expression vanishes and in that case there will be no confidentiality involved in the responses. The respondents might feel that their privacy is not sufficiently protected and they will hesitate to answer truthfully in case $p_0$ is close to 0 or 1.But in this case $\hat{P}$ will have higher efficiency.

On the other hand if $p_0$ is taken close to 0.5 (note that it can't be equal to 0.5 as in that case the denominator in the second term becomes 0), the respondents will feel secured while answering but $\hat{P}$will have lower efficiency.

So, we can see that we can't simultaneously make the respondents feel secure and increase our efficiency of the Warner's Model. This can be an area of development where we can show the change in efficiency of the estimator in Warner's Model with varying $p_0$.