KUK's MODEL ADJUSTED FOR EFFICIENCY AND PROTECTION USING TWO NON-SENSITIVE CHARACTERISTICS UNRELATED TO THE MAIN CHARACTERISTIC OF INTEREST

Shu-Ching Su, Stephen A. Sedory and Sarjinder Singh
Department of Mathematics
Texas A&M University-Kingsville
Kingsville, TX 78363
E-mail: sarjinder@yahoo.com

ABSTRACT

In this paper, we adjust the Kuk (1990) model for both protection and efficiency by making use of proportions of two non-sensitive characteristics which are unrelated to the main sensitive characteristic of interest. The situations where the proportions of the two non-sensitive characteristics in the population of interest are known and unknown are investigated. To avoid any confusion, we first briefly explain the Kuk's model. Then we discuss an adjustment in this model that makes use of two non-sensitive characteristics. We compare the adjusted model and Kuk's model through a simulation study from both the protection and efficiency points of views.

Key words: Randomized response, estimation of proportion, protection and efficiency.

1. INTRODUCTION

Kuk (1990) proposed a randomized response model by making use of two randomization devices. The first randomization device R_1 (say) consists of two possible outcomes, say a deck of cards each, bearing one of two possible questions: "(i) Are you a member of group A?" and "(ii) Are you a member of group A^c?" with known probabilities θ_1 and $(1-\theta_1)$ respectively. The second randomization device R_2 (say) also consists of two possible outcomes, say a deck of cards each bearing one of two possible questions: "(i) Are you a member of group A^c?", and "(ii) Are you a member of group A?" with known probabilities θ_2 and $(1-\theta_2)$ respectively. Assume a simple random and with replacement (SRSWR) sample of n respondents is selected from the population of interest. Each respondent selected in the sample is provided with both randomization devices R_1 and R_2 along with instructions on how to use these devices. Each respondent is given the instruction that if he/she belongs to the sensitive group A then he/she should make use of the first randomization device R_1, otherwise he/she should make use of the second randomization device R_2, without disclosing to the interviewer, which device he/she is using. The choice between the two devices is made by the interviewee in the absence of the interviewer; hence the privacy of the respondent is maintained. The true probability of a 'Yes' answer θ_K is given by:

$$\theta_K = \pi\theta_1 + (1-\pi)\theta_2 \tag{1.1}$$

and the maximum likelihood and unbiased estimator of π is given by:

$$\hat{\pi}_K = \frac{\hat{\theta}_K - \theta_2}{\theta_1 - \theta_2}, \qquad \theta_1 \neq \theta_2 \tag{1.2}$$

The variance of the estimator $\hat{\pi}_K$ is given by:

$$V(\hat{\pi}_K) = \frac{\theta_K(1-\theta_K)}{n(\theta_1 - \theta_2)^2} \tag{1.3}$$

Note that if $\theta_2 = (1-\theta_1)$ then the Kuk's randomized response model reduces to the Warner (1965) model.

In this paper, we suggest an adjustment to the original Kuk's model that makes use of two non-sensitive characteristics, say Y_1 and Y_2. Then we further investigate different situations: (a) The population proportions π_{Y_1} and π_{Y_2} of the characteristics Y_1 and Y_2 are known (b) The population proportions π_{Y_1} and π_{Y_2} of the characteristics Y_1 and Y_2 are unknown (c) The non-sensitive characteristics Y_1 and Y_2 are related to each other, but are unrelated to the characteristic of interest, A and (d) The non-sensitive characteristics Y_1 and Y_2 are unrelated to each other as well as to the characteristic of interest, A. Due to limited space only the situations (a) is discussed here, other situations (b) – (d) are available in Su (2013).

2 KUK'S MODEL ADJUSTED WITH NON-SENSITIVE CHARACTERISTICS

In the adjusted Kuk's model, each respondent selected in a simple random and with replacement sample are provided with a pair of randomization devices, say $R_1^{(a)}$ and $R_2^{(a)}$. The first device, $R_1^{(a)}$, may consist of a deck of cards with two types of cards each bearing one of two types of statements: (a) I belong to the sensitive group A and (b) I belong to the first non-sensitive group Y_1, with probabilities P and $(1-P)$ respectively. Similarly, the second device, $R_2^{(a)}$, may consist of another deck of cards with two types of cards each bearing one of two types of statements: (a) I belong to the sensitive group A and (b) I belong to the second non-sensitive group Y_2, with probabilities T and $(1-T)$ respectively. Each respondent selected in the sample is asked to follow the following instructions. If the respondent belongs to the sensitive group A, then he/she is instructed to utilize the device $R_1^{(a)}$, and if he/she belongs to non-sensitive group A^c then he/she is instructed to use the second device $R_2^{(a)}$. The respondents are also instructed not to disclose to the interviewer which randomization device that they are using in providing a response. A pictorial representation of such a setup is shown in Fig. 1.

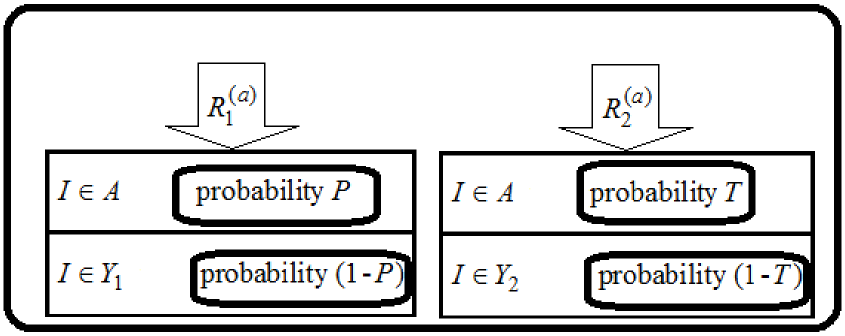

Fig. 1. Adjusted Kuk's randomized response model

2.1. POPULATION PROPORTIONS OF THE NON-SENSITIVE CHARACTERISTICS ARE KNOWN

In a situation where the proportions π_{y_1} and π_{y_2} of the both non-sensitive characteristics are known, we take simple random and with replacement sample of size n. The probability of a "Yes" answer from a given respondent is given by:

$$P(\text{Yes}) = \theta = \pi[P + (1-P)\pi_{y_1}] + (1-\pi)[T + (1-T)\pi_{y_2}] \qquad (2.1)$$

Or equivalently,

$$\theta = \pi[(P-T) + (1-P)\pi_{y_1} - (1-T)\pi_{y_2}] + T + (1-T)\pi_{y_2} \qquad (2.2)$$

Let X be the number of "Yes" answers observed out of n responses taken from the selected n respondents. Obviously $X \sim B(n,\theta)$, and the probability mass function of X is given by:

$$P(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad (2.3)$$

On setting $\dfrac{\partial \log(P(x))}{\partial \pi} = 0$, and by the method of moments, we have the following theorems:

Theorem 2.1. An unbiased estimator of the population proportion of the sensitive characteristic π is given by:

$$\hat{\pi}_c = \frac{\hat{\theta} - T - (1-T)\pi_{y_2}}{(P-T) + (1-P)\pi_{y_1} - (1-T)\pi_{y_2}} \qquad (2.4)$$

Proof. Obvious by taking expectation on both sides of (2.4).

Theorem 2.2. The variance of the estimator $\hat{\pi}_c$ is given by

$$\mathrm{V}(\hat{\pi}_c) = \frac{\theta(1-\theta)}{n[(P-T)+(1-P)\pi_{y_1} - (1-T)\pi_{y_2}]^2} \tag{2.5}$$

Proof. It follows from $X \sim B(n, \theta)$.

Theorem 2.3. An unbiased estimator of variance of the estimator $\hat{\pi}_c$ is given by

$$\hat{\mathrm{V}}(\hat{\pi}_c) = \frac{\hat{\theta}(1-\hat{\theta})}{(n-1)[(P-T)+(1-P)\pi_{y_1} - (1-T)\pi_{y_2}]^2} \tag{2.6}$$

Proof. It follows from $E[\hat{V}(\hat{\pi}_c)] = V(\hat{\pi}_c)$.

2.1.1. PROTECTION OF RESPONDENTS

We adopt Lanke (1975, 1976), also cited in Singh (2003), to define a protection criterion as follows. In Kuk's pioneer model, if a respondent reports "Yes", then the conditional probability that this particular respondent belongs to group "A" is given by:

$$P_k(A\,|\,Yes) = \frac{P(A \cap Yes)}{P(Yes)} = \frac{\pi\theta_1}{\theta_k} = \frac{\pi\theta_1}{\pi\theta_1 + (1-\pi)\theta_2} \tag{2.7}$$

Again, in Kuk's pioneer model if a respondent reports "No", then the conditional probability that this particular respondent belongs to sensitive group A is given by:

$$P_k(A\,|\,No) = \frac{P(A \cap No)}{P(No)} = \frac{\pi(1-\theta_1)}{1-\pi\theta_1 - (1-\pi)\theta_2} \tag{2.8}$$

Then the least protection (or greatest jeopardy) of a respondent in the pioneer Kuk's model is given by:

$$\mathrm{PROTK} = Max[P_k(A\,|\,Yes), P_k(A\,|\,No)] \tag{2.9}$$

In the adjusted-Kuk's model, if a respondent reports "Yes", then the conditional probability that this particular respondent belongs to the sensitive group A is given by:

$$P_c(A\,|\,Yes) = \frac{P(A \cap Yes)}{P(Yes)} = \frac{\pi[P+(1-P)\pi_{y_1}]}{\pi[P+(1-P)\pi_{y_1}] + (1-\pi)[T+(1-T)\pi_{y_2}]} \tag{2.10}$$

In the same way, in the adjusted-Kuk's model, if a respondent reports "No", then the conditional probability that this particular respondent belongs to sensitive group A is given by:

$$P_c(A \mid No) = \frac{P(A \cap No)}{P(No)} = \frac{\pi(1-P)(1-\pi_{y_1})}{1-\pi[P+(1-P)\pi_{y_1}]-(1-\pi)[T+(1-T)\pi_{y_2}]} \qquad (2.11)$$

Then the least protection of a respondent in the adjusted-Kuk's model is given by:

$$\text{PROTC} = Max[P_c(A \mid Yes), P_c(A \mid No)] \qquad (2.12)$$

Note, small values of PROT indicate a model with greater protection.

In the next section, we compare the adjusted-Kuk's model with the pioneer Kuk's model through an extensive simulation study. The motivation of the simulation study is to investigate situations where the adjusted-Kuk's model can perform better than the pioneer Kuk's model for different choice of parameters involved.

2.1.2. COMPARISON OF THE MODELS

We define the percent relative protection (RP) of the adjusted-Kuk's model with respect to the pioneer Kuk's model as:

$$RP = \frac{PROTK}{PROTC} \times 100\% \qquad (2.13)$$

Also we define the percent relative efficiency (RE) of the adjusted-Kuk's model with respect to the pioneer Kuk's model as:

$$RE = \frac{V(\hat{\pi}_K)}{V(\hat{\pi}_C)} \times 100\% \qquad (2.14)$$

We wrote SAS code to compare the adjusted-Kuk's model and the pioneer Kuk's model for at least equal protection of the respondents. In the program, we changed the values of π, π_{y_1}, π_{y_2}, P and T between 0.1 to 0.9 with a step of 0.1, for fixed values of $\theta_1 = 0.7$ and $\theta_2 = 0.2$ in Kuk's model. There are values for which the Kuk model performs better than the Warner model. We retained all results with percent relative protection (RP) and percent relative efficiency (RE) values more that 101%. There were 2604 cases where both the RP and RE values are observed to be more than 101%. It is not very useful to display a table of these individual outcomes, thus we provide only basic summary statistics and graphical representations of all the raw data.

In Table 2.1 we see that for $\pi = 0.1$ there were 105 different choices of parameters P, T, π_{y_1} and π_{y_2} that met our criterion. Among these, relative protection range from a minimum of 101.50% to a maximum of 126.15% with a median of 108.18%, mean of 116.46% and standard deviation of 8.99%. For these same 105 cases, Table 2.2 shows that percent relative efficiency (RE) values ranged from 101.14% to 152.85% with median 117.49%, mean 122.07% and standard deviation 16.09% . If π is equal to 0.2, then there were 146 different choices of parameters P, T, π_{y_1} and π_{y_2} that met out criterion for comparing to the Kuk's model. Among these, relative protection range from a minimum of 101.11% to a maximum of 129.73% with a median of 109.65%, mean of 111.29% and standard deviation of 7.73%. For these same 105 cases, Table 2.2 shows that percent relative efficiency (RE) values ranged from 101.78% to 159.88% with median 118.14%, mean 122.39% and standard deviation 17.49%

Table 2.1. Summary of percent relative protection (RP) for different levels of proportions of sensitive attribute (π) in a population.

π	f	Mean	StDev	Minimum	Median	Maximum
0.1	105	111.46	8.99	101.50	108.18	126.15
0.2	146	111.29	7.73	101.11	109.65	129.73
0.3	170	111.37	7.17	101.26	112.86	125.71
0.4	211	110.54	6.25	101.11	110.89	125.15
0.5	252	109.82	5.40	101.19	109.06	121.43
0.6	325	108.55	4.46	101.04	108.09	118.29
0.7	391	107.23	3.45	101.24	107.20	115.32
0.8	462	105.11	2.33	101.02	105.00	111.01
0.9	541	102.87	1.16	101.00	102.88	105.63

Table 2.2. Summary of percent relative efficiency (RE) for different levels of proportions of sensitive attribute (π) in a population.

π	f	Mean	StDev	Minimum	Median	Maximum
0.1	105	122.07	16.09	101.14	117.49	152.85
0.2	146	122.39	17.49	101.78	118.14	159.88
0.3	170	124.57	17.31	102.19	119.41	163.64
0.4	211	127.03	18.91	102.12	123.92	168.58
0.5	252	130.23	20.95	101.82	126.36	179.11
0.6	325	134.90	25.68	101.01	129.94	205.27
0.7	391	141.74	32.36	101.68	132.64	246.51
0.8	462	159.08	47.03	101.21	144.57	324.26
0.9	541	188.12	72.67	102.50	167.94	452.94

Continuing reading of these Tables 2.1 and 2.2 shows that if π is equal to 0.9, then there are 541 choices of different parameters P, T, π_{y_1} and π_{y_2} such that the adjusted Kuk's model can have minimum protection level 101.0% to maximum of 105.63% with a median protection of 102.88% , mean value equal to 102.87% with a standard deviation of 1.16% in comparison to the Kuk's pioneer model with $\theta_1 = 0.7$ and $\theta_2 = 0.2$. At the same time for $\pi = 0.9$, Table 2.2 shows that the percent relative efficiency (RE) value ranges between 102.50% and 452.94% with a median value of 167.94%, average value of 188.12% with a standard deviation of 72.67%.

A close look of these Tables 2.1 and 2.2 indicates that as the value of π changes from 0.1 to 0.9, the maximum value of RP decreases from 126.15% to 105.63% whereas the corresponding maximum value of RE increases from 152.85% to 452.94%.

Figure 2.1 shows that if the value of π is close to 0.1 then it is possible to adjust Kuk's model for both more protection and more relative efficiency by utilizing appropriate choices of values of P, T, π_{y_1} and π_{y_2}. As the value of π becomes close to one, the relative protection of the proposed mode remains close to 100%, but the percent relative efficiency takes on very large values (more than 400%). It is not a surprise thatv there is little gain in protection for large values of π; such a characteristic with such a large probability is most likely not sensitive. Greater gains are possible with characteristics is close to zero then a study characteristic can be considered as a sensitive one, whereas if the value of π is close to one then a study characteristic remains no longer sensitive characteristic.

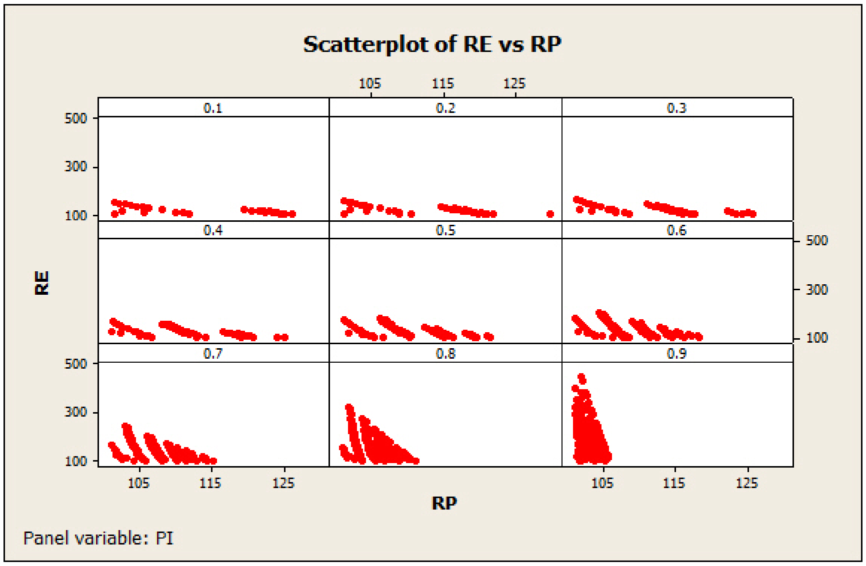

Fig. 2.1. Relationship between RP and RE for different values of π.

Figure 2.2 conveys further evidence that, for values of π close to zero, the value of RE remains higher, and the value of RE remains lower in comparison to those values when π is close to one. Note that the values of RE are not symmetric around the value of $\pi = 0.5$ because the proposed model's parameters are set for at least equal protection of the respondents in comparison to Kuk's pioneer model with $\theta_1 = 0.7$ and $\theta_2 = 0.2$.

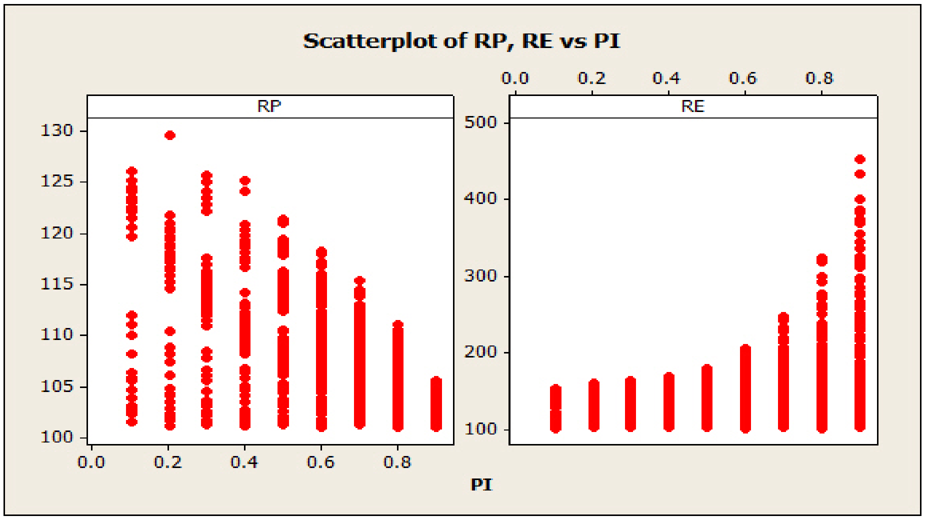

Fig. 2.2. Another look at RE and RP values for different choice of values of π.

An additional picture on the data may be gained by considering box plots. Sometimes it is more convenient to look at the true picture of a dataset by using a box plot. Box plots for the values of percent RP are presented in Figure 2.3. The distribution of RP for each value of π considered is observed to be skewed to the right as might be expected but in some cases skewed to the left. Figure 2.3 shows that there are more outliers of RP if the value of π is close to zero, which indicates that there are a few combinations of the choices of the parameters π_{y_1}, π_{y_2}, P and T where the RP is very high. A thorough search of the raw data set indicates that for $\pi = 0.1$ a choice of $P = 0.5$, $T = 0.3$, $\pi_{y_1} = 0.9$ and $\pi_{y_2} = 0.1$ gives maximum RP value of 126.15% with percent RE value of 103.06%. In the same way, using SAS codes, a combination of these parameters can be sought for that are likely to provide more protection and better percent efficiency depending on whether a good guess of the value of π is available.

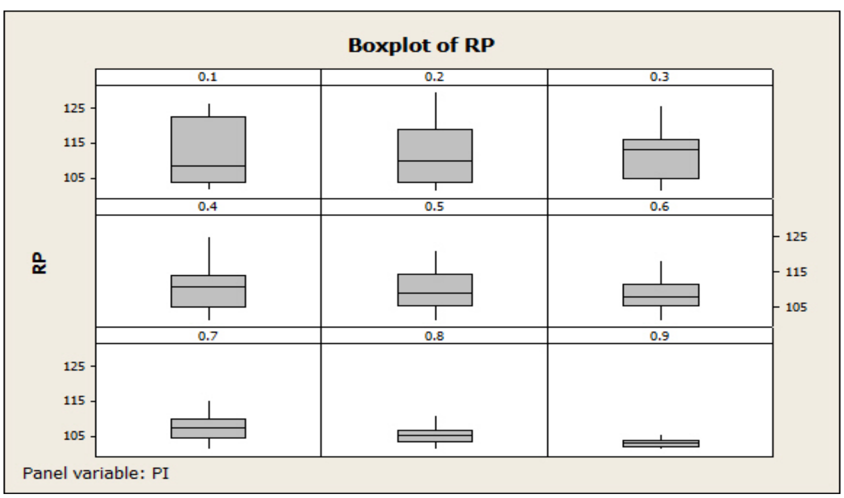

Fig. 2.3. Box plots for detecting outliers in RP values.

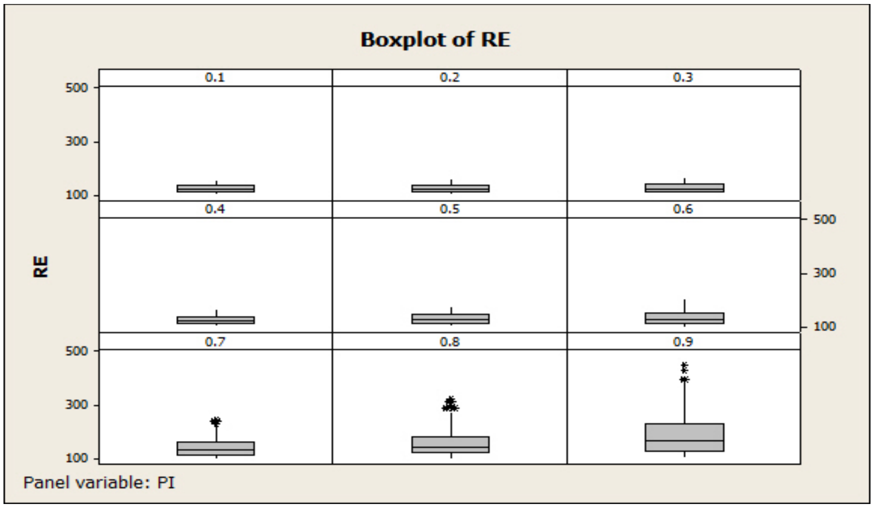

Fig. 2.4. Box plots for detecting outliers in RE values.

Figure 2.4 shows that more outliers can be found if the value of π is close to one. Also for each level of value of π considered, the distribution of these values are observed to be skewed to the right indicating that majority of the RE values are higher than the median RE value.

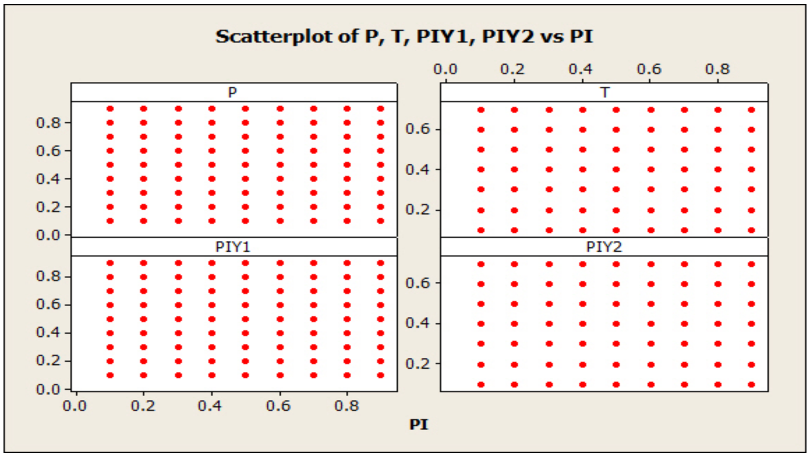

Fig. 2.5. The P, T, π_{y_1} and π_{y_2} values for different levels of π leading to the RP and RE values.

Figure 2.5 shows values of P, T, π_{y_1} and π_{y_2} for different levels of the sensitive proportion π considered in the simulation study such that the values of the percent relative protection (RP) and the percent relative efficiency (RE) remain higher than 101%. It seems that the parameters P and T can take any values while showing more efficient and more protective results.

Figure 2.6 shows the behavior of the values of percent relative protection while the values of π_{y_1} and π_{y_2} changes regardless change in the values of other three parameters P, T and π. It seems that when π_{y_1} is close to 0.77 and π_{y_2} is close to 0.2 then there is a combination of P, T and π which could lead to more protection from respondents.

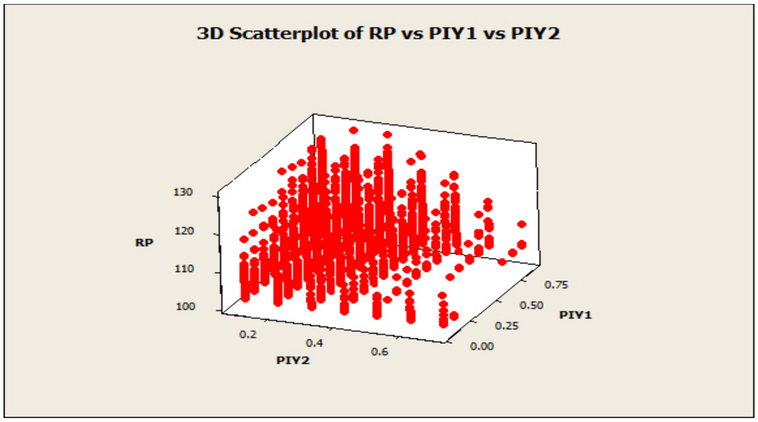

Fig. 2.6. Relative protection (RP) as a function of π_{y_1} and π_{y_2}

Figure 2.7 shows the behavior of the values of percent relative efficiency while the values of π_{y_1} and π_{y_2} changes regardless change in the values of other three parameters P, T and π. Again it seems that when π_{y_1} is close to 0.2 and π_{y_2} is close to 0.75 then there is a combination of P, T and π which could lead to more relative efficiency of the adjusted Kuk's model.

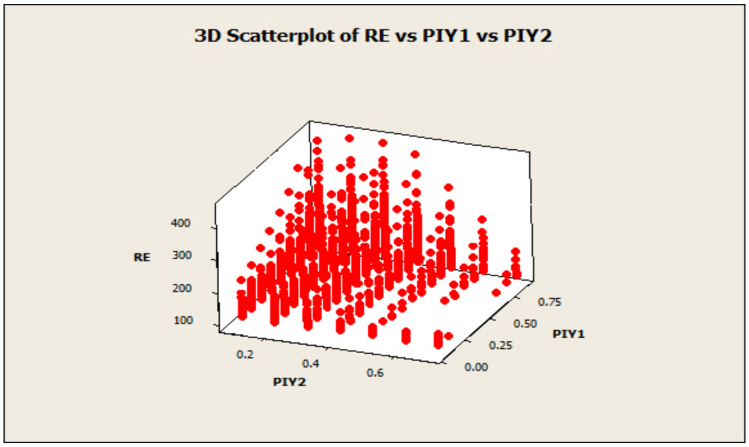

Fig. 2.7. Relative efficiency (RE) as a function of π_{y_1} and π_{y_2}

Thus based on our discussion of the simulation results we conclude that for the given values of $\theta_1 = 0.7$ and $\theta_2 = 0.2$ in the Kuk's pioneer model, and there can be found a choice of the other four parameters P, T, π_{y_1} and π_{y_2} such that the adjusted Kuk's model can be made more efficient and more protective for any level of the proportion of a sensitive characteristic π in a population.

Next we narrow our search of the parameters by setting $\theta_1 = P = 0.7$ and $\theta_2 = T = 0.2$ to investigate the additional gain in relative protection and relative efficiency that comes about solely from the additional flexibility of utilizing unrelated characteristics. Then we end up with 108 situations where the adjusted Kuk's model is better than the Kuk's pioneer model for various choice of values of π_{y_1} and π_{y_2}.

We observed that if we keep $\theta_1 = P = 0.7$ and $\theta_2 = T = 0.2$, but are allowed to change the other two known parameters π_{y_1} and π_{y_2} between 0.1 and 0.9 with a step of 0.1, then Table 2.3 shows that for $\pi = 0.1$, there are six choices of π_{y_1} and π_{y_2} that meet out criterion. Among these, the value of RP ranges between 103.06% to 124.51% with a median value of 109.60%, average value of 112.31% with standard deviation of 8.76%.

Table 2.3. Summary of percent relative protection (RP) for different levels of proportions of sensitive attribute (π) in a population.

π	f	Mean	StDev	Minimum	Median	Maximum
0.1	6	112.31	8.76	103.06	109.60	124.51
0.2	8	110.70	6.78	102.27	109.28	120.51
0.3	9	109.75	7.02	101.70	107.80	123.51
0.4	11	108.90	6.05	101.28	108.97	119.15
0.5	12	108.03	5.32	101.71	107.10	119.47
0.6	14	106.30	3.82	101.23	105.80	114.02
0.7	16	105.65	3.34	101.24	104.99	112.71
0.8	17	103.81	1.99	101.02	103.58	107.77
0.9	15	102.36	1.03	101.05	102.13	104.47

Table 2.4. Summary of percent relative efficiency (RE) for different levels of proportions of sensitive attribute (π) in a population.

π	f	Mean	StDev	Minimum	Median	Maximum
0.1	6	120.67	15.19	103.71	117.40	144.38
0.2	8	121.65	17.00	102.01	119.75	151.04
0.3	9	126.56	18.03	106.42	123.02	158.87
0.4	11	129.81	20.79	104.17	124.79	168.58
0.5	12	130.83	20.97	103.51	130.04	165.36
0.6	14	138.22	25.75	106.11	134.06	187.06
0.7	16	143.51	32.84	101.68	139.74	219.76
0.8	17	158.80	48.50	102.60	141.20	277.10
0.9	15	188.50	68.80	110.10	172.80	335.90

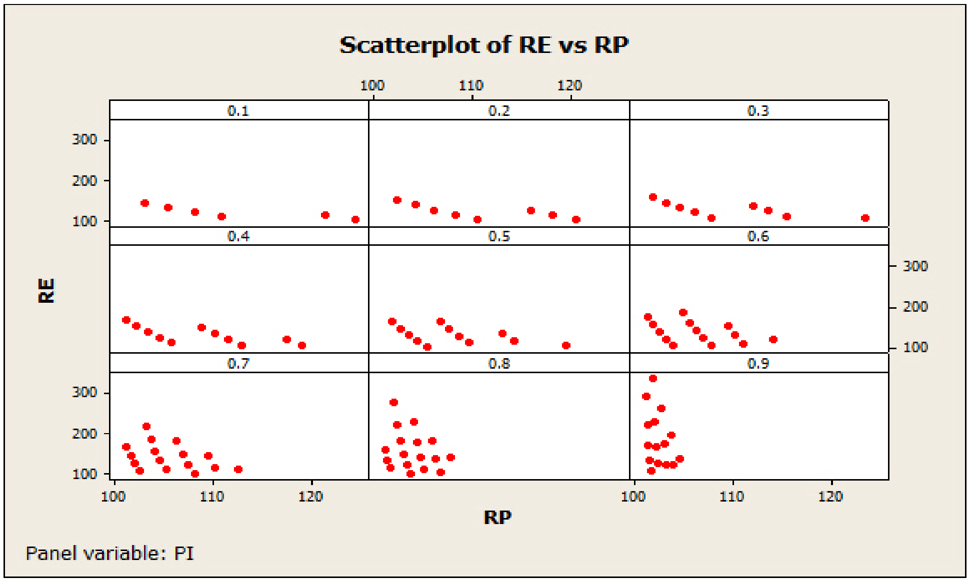

Fig. 2.8. RP versus RE for different values of π with $\theta_1 = P = 0.7$ and $\theta_2 = T = 0.2$.

At the same time, the value of percent RE ranges from 103.71% to 144.38% with a median of 117.40% and average value of 120.67% with a standard deviation of 15.19%. For $\pi = 0.2$, there are 8 cases where the value of RP varies from 102.27% to 120.51% with a median value of 109.28%, and the RE value changes from 102.01% to 151.04% with a median value of 119.75%.

Again the results show that if the value of π is close to 0.0 then the adjusted Kuk's model shows better protection as well as better relative efficiency for certain choice of model parameters which are assumed to be known. If the value of π is close to 1.00 then the adjusted Kuk's model shows greater efficiency, but protection level remains close to 100%. For example, another direct look at raw results shows that if $\pi = 0.1$, then a choice of π_{y_1} greater than 0.5, and a choice π_{y_2} of less than or equal to 0.2 leads to more protection and more efficient adjusted Kuk's model. The RP value changes from 103.06% to 124.51% and the RE value changes from 103.71% to 144.38%. Now if $\pi = 0.9$, then a choice of π_{y_1} greater than or equal to 0.4 and a choice of π_{y_2} less than or equal to 0.6 may lead to more protective and efficient results. Thus we recommend choosing π_{y_1} close to one and π_{y_2} close to zero for the adjusted Kuk's model to be more protective and more efficient than the pioneer Kuk's model. In nutshell, based on our discussion of the simulation results we conclude that for a given values of $\theta_1 = P = 0.7$ and $\theta_2 = T = 0.2$, there can be found a choice of the other two known parameters π_{y_1} and π_{y_2} such that the adjusted Kuk's model can be made more efficient and more protective for any value of π in a population.

We also consider a situation when the population proportions π_{y_1} and π_{y_2} are unknown, but can be estimated from another independent sample by following the concept of Greenberg et al. (1969) and Moors (1971). The detail can be found in Su (2013).

REFERENCES

Greenberg, B. G., Abul-Ela, A. L. A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model-Theoretical framework. *J. Amer. Stat. Assoc*., *64*,520-539.

Kuk, A. Y. C. (1990). Asking sensitive questions indirectly. *Biomerika*, *77*(2): 436-438.

Lanke, J. (1975). On the choice of the unrelated question in Simmons version of randomized response. *J. Amer. Statist. Assoc.*, *70*, 80-83.

Lanke, J. (1976). On the degree of protection in randomized interviews. *Int. Stat.. Rev*. 197-203.

Moors, J. J. A. (1971). Optimization of the unrelated question randomized response model. *J. Amer. Statist. Assoc.*, 66, 627-629.

Singh, S. (2003). *Advanced Sampling Theory with Applications: How Michael Selected Amy.* Kluwer Academic Press, The Netherlands.

Su, Shu-Ching (2013).*On protection and efficiency of randomized response strategies.* Unpublished MS thesis submitted to the Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX.

Warner, S. L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*: 63-69.