A TWO-PHASE SAMPLING PLAN TO IDENTIFY PSYCHIATRIC DISORDERS IN COMMUNITY

Yuk-Miu Lam, Jessie Wong, and Nancy Lee
The Chinese University of Hong Kong

1. BACKGROUND

The Psychiatric Epidemiology Research Unit (PERU), funded by the Keswick Charity Foundation was established under the Department of Psychiatry of the Chinese University of Hong Kong in 1981. A large scale community psychiatric epidemiological survey was conducted by the Unit in Shatin, a new town in Hong Kong. One of the aims of such a large scale screening of the community is to estimate the prevalences of various disorders in the community followed by case identification. This requires to draw a sufficiently large and representative random sample from the community and have comprehensive instruments applied to the sample for screening and case identification.

The diagnosing of mental disorders is a very complex process, the instrument for screening must be comprehensive enough to cover a wide range of disorders with acceptable validity and reliability. However a community survey usually requires a large sample which imposes a cost impossible if a detailed instrument is applied. To solve the dilemma, a two-stage screening method (Blum 1962; Deming 1977) was sought to be the alternative in which a simple instrument such as the Self-Reporting Questionnaire (SRQ) (Harding 1980) is used to screen as many as the Unit can afford and a comprehensive instrument such as the Diagnostic Interview Schedule (DIS) (Robins 1981) is used to obtain definitive diagnosis of the potential cases as screened in the first stage. In order to estimate the proportion missed by the SRQ a sample of the non-cases (with low SRQ scores) are also included in the second stage and followed-up with the DIS.

Shortcomings were found when the usual two-stage screenig method was applied to a pilot survey. Especially the problem arises on how to draw a random subsample from the negatives (with low SRQ scores) after the first stage screening. The effect of time lapse between the two instruments on their reliability could not allow the drawing of a truly random sample from the negatives and following-up on them after the completion of the first stage. Therefore a modified two-stage screening method is designed to overcome this problem.

In the modified method a proportion of the original sample is flagged to be included into the second stage regardless of their results in the first stage screening. As for the unflagged group, only those subjects with high SRQ scores (i.e. the positives) in the first screening will enter the second stage. The advantages of the modified method over the classical method are:

(1) two instruments can be applied to a respondent in one interviewing session, thus reducing the non-response rate of the second stage;

(2) the total coverage of the flagged group by both instruments serves as a built-in gauge for the concordance between the instruments.

The purpose of this paper is to present some theory based on the modified method. This involves in developing estimates of prevalences and finding the properties of these estimates. Optimal allocation of the flagged subsample size and the determination of a cut-off point for defining the low and high SRQ scores are also investigated in this study.

2. THE MODIFIED TWO-PHASE SAMPLING METHOD

Suppose a random sample of n residents in a large population of size N is drawn. Within the sample a random subsample of size n_1 is selected and flagged where $r=n_1/n$, the proportion of the subsample size to the total sample size is fixed beforehand. The optimal choice of r will be discussed in the later section. The case-identification instrument (DIS) is applied to the flagged subsample immediately after the first stage screening (SRQ). The remaining unflagged sampled individuals will all respond to the SRQ and only those with SRQ-positives will proceed to the spot for the DIS interview.

We now proceed with the following notations:

	Population	Sample
Size	N	n
Prevalence rate (proportion of cases)	P	p
Proportion of SRQ-positives	P^*	p^*
Prevalence among SRQ-positives	P_1	p_1
Prevalence among SRQ-negatives	P_2	p_2

P_2 can be considered as the expected proportion of false negatives and $1-P_1$ as the expected proportion of false positives. The SRQ screening is meaningful if $P_2<P<P_1$.

With regard to the sampling outcome we further denote:

- n_1 = subsample size (flagged)
- n_2 = subsample size (unflagged)
- n_{11} = number of SRQ-positives in the flagged subsample
- n_{12} = number of SRQ-negatives in the flagged subsample
- n_{21} = number of SRQ-positives in the unflagged subsample
- n_{22} = number of SRQ-negatives in the unflagged subsample
- m_{11} = number of (DIS) cases among n_{11} SRQ-positives
- m_{12} = number of (DIS) cases among n_{12} SRQ-negatives
- m_{21} = number of (DIS) cases among n_{21} SRQ-positives
- c_1 = cost of using SRQ to screen one person
- c_2 = cost of applying DIS to interview one person

Based on the notations, we have

$$n=n_1+n_2$$

$$n_1=n_{11}+n_{12}$$

$$n_2=n_{21}+n_{22};$$

and the prevalence rate P can be expressed as

$$P=P^*P_1+(1-P^*)P_2. \quad (1)$$

We can then estimate P, P^*, P_1 and P_2 by the corresponding sample proportions p, p^*, p_1 and p_2, where

$$p^*=(n_{11}+n_{21})/n$$

$$p_1=(m_{11}+m_{21})/(n_{11}+n_{21})$$

$$p_2=m_{12}/n_{12}$$

and

$$p=p^*p_1+(1-p^*)p_2$$

$$=(m_{11}+m_{21}+m_{12}+n_{22}m_{12}/n_{12})/n \quad (2)$$

where the term $n_{22}m_{12}/n_{12}$ is the expected number of cases among n_{22} SRQ-negatives in the unflagged subsample.

In the following, we assume that the population is large in relation to the sample so that variances of the above estimates can be derived from the binomial probability theory.

Theorem 1

The sample proportions p^*, p_1 and p_2 are unbiased estimates of their corresponding population proportions; their variances are:

$$Var(p^*)=P^*(1-P^*)/n$$

$$Var(p_1)\doteq P_1(1-P_1)/(nP^*)$$

$$Var(p_2)\doteq P_2(1-P_2)/(n_1(1-P^*))$$

Proof: (see Appendix)

Theorem 2

The sample proportion p is an unbiased estimate of the population proportion (prevalence) P with variance,

$$Var(p)\doteq P(1-P)/n+P_2(1-P_2)(1-P^*+P^*/n)*(1-r)/(nr) \quad (3)$$

where $r=n_1/n$ is the fraction of the sample being flagged.

Proof: (see Appendix)

The population (or expected) variance P(1-P) can be partitioned into three components, i.e.

$$P(1-P)=P^*(1-P^*)(P_1-P_2)^2+P^*P_1(1-P_1)+(1-P^*)P_2(1-P_2) \quad (4)$$

where $P^*(1-P^*)(P_1-P_2)^2$ is the variance between SRQ-positives and SRQ-negatives; $P^*P_1(1-P_1)$ is the variance within the SRQ-positives; and $(1-P^*)P_2(1-P_2)$ is the variance within the SRQ-negatives.

The first component, P(1-P)/n of Var(p) can be considered as the variance of the sample prevalence with which all subjects in the total sample are undergone the two stages. In this case the first stage screening will be unnecessary since cases will be identified by using the DIS alone. With regard to the modified method, only a fraction of the SRQ-negatives will be included in the second stage, thus the second term of Var(p) can be considered as the additional variance due to the lack of information from all the SRQ-negatives in the unflagged subsample.

If only the flagged subsample is used to estimate the prevalence P, then the estimate is simply

$$p_s=(m_{11}+m_{12})/n_1$$

which is also an unbiased estimate of P with variance

$$Var(p_s)=P(1-P)/n_1.$$

We can compare p with p_s in terms of the relative efficiency defined as the ratio of their variances. From equation (4),

$$P(1-P) > (1-P^*)P_2(1-P_2);$$

with this inequality, it can be shown that the relative efficiency $Var(p)/Var(p_s) < 1$ for sufficiently large sample size n. Thus the proposed estimate in equation (2) is more efficient.

3. OPTIMAL ALLOCATIONS

An effective screening procedure should have a high probability of correctly discriminating the cases and the non-cases leaving both the false positives and false negatives low. Since the proportion P_2 of false negatives is critical thus we should plan the screening so that P_2 can be kept very low while the proportion of false positives, $1-P_1$ is held to a moderately low level. Since the SRQ scores define the positives and the negatives on screening according to a specific cut-off point. Changing the cut-off point will change the proportions P^*, P_1, and P_2. Ideally, for specific values of P_1 and P_2, P^* can be obtained from equation (1) so that

$$P^*=(P-P_2)/(P_1-P_2).$$

This can only be served as a reference point for the SRQ-scores, since in practive, we would not be able to manipulate the values of P_1 and P_2. Once P^* is fixed, P_1 and P_2 will also be fixed, this is because of the inherent validity of the screening procedure. The values of P_1 and P_2 can only be improved by modifying the screening instrument.

Example 1

A preliminary study of psychiatric morbidity has been carried out in a community which gives a

prevalence of 0.1. Suppose we would like to set the proportion of false negatives not greater than 0.05 and keep the proportion of false positives as small as 0.1, then P^* is found to be 0.06. This would allow us to determine an appropriate cut-off point of the SRQ scores for the full scale study.

In terms of costs, the total cost required for sampling is

$$C=nc_1+(n_1+n_{21})c_2.$$

Since n_{21} SRQ-positives in the unflagged sample is a random variable with its expectation $E(n_{21})=n_2P^*$, thus the expected total cost is

$$E(C)=nc_1+(n_1+n_2P^*)c_2$$
$$=nc_1+n[r(1-P^*)+P^*]c_2. \quad (5)$$

The optimal allocation of the flagged subsample involves finding a value r which minimizes both the variance of p and the total expected cost E(C).

Since the amount of information in the estimate p is defined by the Fisher's information number I(P) which is simply $I(P)=1/Var(p)$, thus the amount of information per unit cost can be determined by

$$I(P)/E(C)=1/(E(C)Var(p)) \quad (6)$$

This defines the efficiency of the procedure which can be used as a criterion to compare the procedure with different allocations r of the flagged proportions or with other sampling procedures as well. The optimal allocation of the flagged proportion r of the total sample can be obtained by maximizing the efficiency defined in equation (6)

Theorem 3

In the modified two-phase sampling, the amount of information per unit cost is maximized if the proportion r of the flagged sample to the total sample is set to

$$r=\sqrt{\frac{(K+P^*)P_2(1-P_2)(1-P^*+P^*/n)}{(1-P^*)[P(1-P)-P_2(1-P_2)(1-P^*+P^*/n)]}} \quad (7)$$

where $K=c_1/c_2$.

Proof: Since the amount of information per unit cost, I(P)/E(C) can be expressed as a function of r, finding the value of r to maximize I(P)/E(C) can be achieved by setting the derivative to 0, i.e.

$$\frac{d}{dr}\left(\frac{I(P)}{E(C)}\right)=0.$$

The solution is given in equation (7) which can be obtained through some simple calculations.

Example 2

From example 1, we have P=0.1, $P^*=0.06$, $P_1=0.9$ and $P_2=0.05$. Suppose the cost for screening is \$5.00 per person and for each DIS interviewing is \$45.00 which give the ratio of the costs K=0.111. With the total sample of 1000 we find the proportion of the flagged sample from equation (7) as r=0.423 which indicates a subsample of size 423 should be selected and flagged.

Since the amount of information per unit cost is a function of r, in this example we have

$$I(P)/E(C)=r/(1.918r^2+2.238r+0.344)$$

Figure 1 shows this functional relationship (the line with "X" symbol on it) which indicates the optimal allocation of r= 0.423. However the graph also shows that there is a range of choosing the proportion r with only a slight reduction of the amount of information per unit cost. In particular if a smaller flagged subsample is preferred then r may be chosen as small as 0.25 so that only 1/4 of the total sample needs to be flagged. As a result, the efficiency reduces from 0.259 to 0.244.

4. COMPARISON OF OTHER SAMPLING METHODS

If the usual two-stage sampling procedure is performed then all subjects in the whole sample are screened by SRQ, while all SRQ-positives and only a fraction, r_2 of SRQ-negatives (i.e. a subsample) will be interviewed by DIS. Let m_1 be the DIS-cases among all SRQ-positives and m_2 be the DIS-cases among the selected subsample of the SRQ-negatives then the prevalence estimate is simply

$$p_0=(m_1+m_2/r_2)/n$$

and its variance is

$$Var(p_0)=P(1-P)/n+P_2(1-P_2)(1-P^*)(1-r_2)/(nr_2).$$

Thus we may compare the two two-stage methods in terms of their variances so that

$$Var(p)-Var(p_0)$$
$$=P_2(1-P_2)(1-P^*)[(1-r)/(nr)-(1-r_2)/(nr_2)]+$$
$$P_2(1-P_2)P^*(1-r)/(n^2r).$$

In particular if $r=r_2$ then

$$Var(p)-Var(p_0)=P_2(1-P_2)P^*(1-r)/(n^2r),$$

thus Var(p) is larger than $Var(p_0)$ by a small amount. An example will show that this amount is negligible.

Taking into consideration the cost, the total expected cost for the usual two-stage sampling is

$$E_0(C)=nc_1+n[r_2(1-P^*)+P^*]c_2.$$

If $r=r_2$, then $E(C)=E_0(C)$. The optimal fraction, r_2 of the subsample taken from the SRQ-negatives can also be found by minimizing the efficiency of the procedure defined in (6). This gives

$$r_2 = \sqrt{\frac{(K+P^*)P_2(1-P_2)(1-P^*)}{(1-P^*)[P(1-P)-P_2(1-P_2)(1-P^*)]}}$$

Example 3

From example 2, if the usual two-stage method is applied then r_2=0.423 is the optimal subsample fraction of the SRQ-negatives which is also the optimal proportion of the flagged subsample in the modified method. In this case, their total expected costs are the same and the variances of their prevalence estimates are approximately the same, since

$$\text{Var}(p)-\text{Var}(p_0)=0.003887/n^2$$

and the amount of information per unit cost (i.e. efficiency) is a function of r_2, that is

$$\text{Efficiency} = r_2/(1.918r_2^2+2.229r_2+0.335).$$

With r_2=r, the efficiencies of the two-stage methods are very much the same. This is seen from Figure 1. Thus the modified procedure lose very little efficiency and yet it has advantages over the usual method.

If the whole sample is investigated without screening we can also calculate its efficiency and compare with the two-stage methods (the usual and the modified methods). The calculations are shown in Table 1. Variances and total expected costs are also included in the table. In this particular example, we see that costs can be cut down by using the two-stage methods while the efficiency remains.

5. DISCUSSIONS AND CONCLUSION

The usual two-stage sampling procedure has been proposed for community psychiatric epidemiological survey (Deming 1977, Duncan-Jones 1978). However its practicality has not been put through very easily and effectively in our local preliminary survey. Problems arise which relate to the time delay between the two stages of interviews, the non-response rate and a random subsampling from the SRQ-negatives for the second stage interview. The modified two-stage sampling is designed to overcome these problems.

Under this modified two-phase sampling plan, an estimate of the prevalence is proposed and shown to be unbiased. Its variance is derived which can be used as a guideline to determine an appropriate sample size with acceptable sampling error.

A preliminary study can be used to choose an appropriate cut-off point for the screening scores to define the SRQ-positives and the SRQ-negatives. With a proper choice, the false negatives can be controlled and reduced to a very low proportion while the false positives are kept to a reasonably low level.

Taking into account the cost effectiveness, an optimal allocation of the flagged subsample is found which maximizes the number of information per unit cost. However the choice of this flagged proportion could be flexible taken into consideration of other factors.

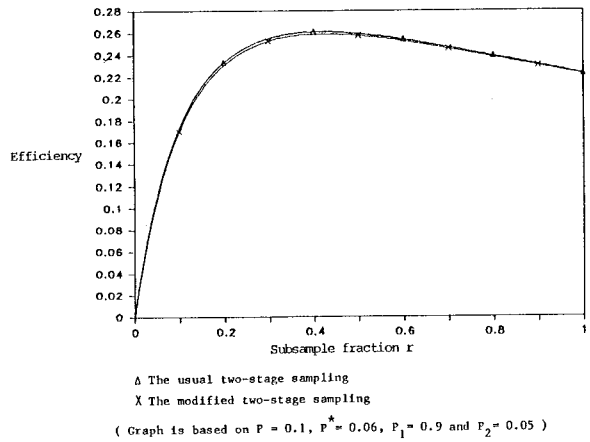
In comparing this modified method with the usual two-stage sampling procedure, we find that, with the same subsample proportion, they require the same costs and the variances of the prevalence estimates are about the same. Thus both methods have almost the same efficiency and yet the modified method is more advantageous in this study. With respect to the fixed sample procedure the two-stage methods with good screening instrument for the first stage can always cut down the cost while attaining the same efficiencies.

TABLE 1
Variance, Total Expected Cost and Efficiency

	Two-stage Methods (subsample proportion)		No screening
	r=0.25	r=0.42	
Var(p)	0.224/n	0.151/n	0.09/n
E(C)	18.275n	25.593n	45.00n
I(P)/E(C)	0.244	0.259	0.247

* Results are based on P=0.1, P^*=0.06, P_1=0.9 and P_2=0.05.

FIGURE 1
Efficiency of Two-Stage Sampling Methods



Δ The usual two-stage sampling
X The modified two-stage sampling
(Graph is based on P = 0.1, P^*= 0.06, P_1= 0.9 and P_2= 0.05)

APPENDIX

1. Proof of Theorem 1:

The unbiasedness of the estimates can easily be proved and $\text{Var}(p^*)=P^*(1-P^*)/n$ can also be derived from the definitions of variance and binomial distribution. To find $\text{Var}(p_1)$, we apply the theory of conditional probability such that

$$\text{Var}(p_1)=\text{Var}(E(p_1|\text{SRQ}))+E(\text{Var}(p_1|\text{SRQ}))$$

where $E(p_1|\text{SRQ})$ and $\text{Var}(p_1|\text{SRQ})$ denote correspondingly the conditional expectation and the conditional variance of p_1 given that the SRQ screening is done and n_{11}, n_{12}, n_{21} and n_{22} are

all known and fixed.

Since $p_1=(m_{11}+m_{21})/(n_{11}+n_{21})$, m_{11} and m_{21} are independent,

$$\mathrm{Var}(E(p_1|SRQ))=\mathrm{Var}(p_1)=0$$

and

$$E(\mathrm{Var}(p_1|SRQ))$$

$$=E[(\mathrm{Var}(m_{11}|SRQ)+\mathrm{Var}(m_{21}|SRQ))/(n_{11}+n_{21})^2]$$

$$=E[(n_{11}P_1(1-P_1)+n_{21}P_1(1-P_1))/(n_{11}+n_{21})^2]$$

$$\doteq P_1(1-P_1)/[nP^*-(1-P^*)]$$

$$\doteq P_1(1-P_1)/(nP^*)$$

for sufficiently large n. Hence

$$\mathrm{Var}(p_1)\doteq P_1(1-P_1)/(nP^*).$$

Similarly we can show that

$$\mathrm{Var}(p_2)=P_2(1-P_2)/[n_1(1-P^*)-P^*]$$

$$\doteq P_2(1-P_2)/[n_1(1-P^*)]$$

for sufficiently large n_1.

2. Proof of Theorem 2:

p can be shown to be an unbiased estimate of P. To find its variance, suppose all subjects in the sample were included in the two-stage interviewing, let m_{22} be the number of cases identified by the DIS among n_{22} SRQ-negatives in the unflagged subsample. Then the estimate of the prevalence rate is

$$p_A=(m_{11}+m_{12}+m_{21}+m_{22})/n$$

with variance, $\mathrm{Var}(p_A)=P(1-P)/n$. The estimate p can then be written as

$$p=p_A+(p-p_A)$$

and its variance is

$$\mathrm{Var}(p)=\mathrm{Var}(p_A)+\mathrm{Var}(p-p_A)+2\mathrm{Cov}(p_A,p-p_A)$$

Since $p-p_A=(n_{22}m_{12}/n_{12}-m_{22})/n$, applying the conditional probability theory as given in the proof of Theorem 1 and following the similar steps, we can show that

$$\mathrm{Var}(p-p_A)=(1-P^*+P^*/n)(P_2(1-P_2)(1-r)/(nr)$$

where $r=n_1/n$. The covariance can be expressed as

$$\mathrm{Cov}(p_A,p-p_A)=E(\mathrm{Cov}(p_A,p-p_A|SRQ))+$$

$$\mathrm{Cov}(E(p_A|SRQ),E(p-p_A|SRQ)).$$

Given the result of the SRQ screening, m_{11}, m_{12}, m_{21} and m_{22} are independent, the conditional covariance can be written as

$$\mathrm{Cov}(p_A,p-p_A|SRQ)$$

$$=\mathrm{Cov}(m_{12}+m_{22},n_{22}m_{12}/n_{12}-m_{22}|SRQ)/n^2$$

$$=(n_{22}\mathrm{Var}(m_{12}|SRQ)/n_{12}-\mathrm{Var}(m_{22}|SRQ))/n^2$$

$$=(n_{22}n_{12}P_2(1-P_2)/n_{12}-n_{22}P_2(1-P_2))/n^2$$

$$=0$$

and since $E(p-p_A|SRQ)=0$, this implies that

$$\mathrm{Cov}(E(p_A|SRQ),E(p-p_A|SRQ))=0$$

thus $\mathrm{Cov}(p_A,p-p_A)=0$ and hence

$$\mathrm{Var}(p)\doteq P(1-P)/n+P_2(1-P_2)(1-P^*+P^*/n)(1-r)/(nr)$$

REFERENCES

Blum, R. H. (1962) Case identification in psychiatric epidemiology: methods and problems. Milbank Mem Fund A. 40:253-288.

Cochran, W. G. (1977) Sampling Techniques, third edition. John Wiley & Sons.

Deming, W. E. (1977) An essay on screening, or on two-phase sampling, applied to surveys of a community. Inter. Stat. Review 45:29-39.

Duncan-Jones, P. and Henderson S (1978) The use of a two-phase design in a prevalence survey. Social Psyciatry 13:231-237.

Harding, T. W., et al (1980) Mental disorders in primary health care : a study of their frequency and diagnosis in four developing countries. Psychol. Med. 10:231-241.

Robins, L. N., Helzer, J. E., Croughan, J. and Ratcliff, K. S. (1981) National Institute of Mental Health Diagnostic Interview Schedule : its history, characteristics and validity. Arch. Gen. Psychiatry 38:381-389.

Acknowledgment

This research under the Psychiatric Epidemiology Research Unit (PERU) was supported by the Keswick Charity Foundation. The authors would like to thank Professor C N Chen for his advice and support, Mr. Frederick Ho for his initiative idea and Dr. H. K. Lam for reviewing the paper.