Testing for the significance of changes in television ratings

Denny H. Meyer
Department of Statistics, University of the Witwatersrand, Johannesburg, Wits 2050, Republic of South Africa

Received 28 November 1989; accepted 8 March 1990

South African television ratings are obtained from the AMPS meter panel. This panel must be viewed as a complex non random sample. For such samples the effective sample size differs from the actual sample size. It has been found that, when equal weights are assigned to strata, the most reliable estimate for effective sample size can be obtained by considering every household as a sample cluster. This estimate of effective sample size can be incorporated directly into a test for significant rating change. For convenience this test is implemented graphically.

Waardebepalings van Suid-Afrikaanse televisieprogramme word vanaf die AMPS-meterpaneel verkry. Die paneel moet as 'n komplekse steekproef beskou word. Vir sulke steekproewe verskil die effektiewe steekproefgrootte van die werklike steekproefgrootte. Deur elke huishouding as 'n steekproeftros te benader, is 'n meer betroubare beraming van effektiewe steekproefgrootte gevind wanneer gelyke gewigstoekenning aan die verskillende strata gegee word. Hierdie beraming van effektiewe steekproefgrootte kan direk gebruik word in 'n toets om betekenisvolle veranderings te bepaal in die waardasies toegeken deur televisiekykers. Vir gerieflikheid is hierdie toets grafies geïmplementeer.

Introduction

Soong (1988) claims that 'in an age with a plethora of viewing choices' repeat viewing of successive episodes of the same programme is not very common. This claim is more true of America than South Africa. In America on average 23,5% of the people who view a programme will view the next episode of the same programme. In South Africa, this same average percentage is 43,4% for TV1 viewers. As indicated in Figure 1 a linear relationship between rating and repeat viewership is evident, particularly at ratings in excess of 10%. This feature of South African viewership must be taken into account when television panel data is used to determine whether a significant change in programme popularity has taken place.

The other factor which must be taken into account is the complex nature of the sampled data. As indicated below, the data sample was far from random.

Data

The data used in this study is typical of that collected from the AMPS meter panel in an average week. We will be considering adult ratings for TV1 for periods of 15 minutes.

The AMPS meter panel of roughly 518 households and 1241 adults was chosen systematically from the television license list after those households not connected to a telephone exchange had been removed. The telephone system is used to transmit all rating data so such households could not be included in the panel. The television list was sorted by postal code, so the systematic choice guaranteed a geographically representative sample.

Sample representativeness was further guaranteed by means of sample stratification. Initial proportionate stratification of the households in the panel produced a sample which was balanced in terms of access to M-Net, households with or without children and metropolitan or non-metropolitan location. Post-stratification weights (Holt & Smith, 1979) were used to ensure representativeness with respect to age, sex, home language, number of housewives and household income. These weights, W_h, were incorporated into the rating calculation, x, as indicated in (1). This formula is used to derive the adult rating for any 15 minute interval. In this formula y_{hi} denotes the viewing time, expressed in minutes, for the ith adult within the hth stratum:

$$x = (\Sigma_h W_h \Sigma_i y_{hi})/(15\Sigma_h W_h) \quad (1)$$

Another feature of the data was clustering of the data within households. The average household contains 2,4 adults. Television viewing is a social activity so the viewing behaviour of the members of a household are somewhat homogeneous and certainly not independent. In Figure 2 social viewing is defined loosely as the percentage of viewing households where more than one adult member of the household views at the same time. It is estimated by taking the proportion of viewing adults for which the next adult in the sample is also viewing. In Figure 2 it is evident, particularly at ratings of above 5%, that social viewing increases roughly linearly with rating.

Method for a random sample

If the members of the AMPS meter panel were randomly chosen the obvious test for significant rating change would be a paired t-test. However, such tests do assume normality of the underlying distribution. The y_{hi} in (1) certainly do not follow a normal distribution. The distribution is bimodal with very strong peaks at zero and 15 minutes. An alternative procedure to the paired t-test would be to set all nonzero y_{hi} values equal to 15

and then apply a McNemar test to the transformed data as indicated by Meyer (1988). The critical rating changes for these tests are given in (2) and (3) respectively. In these formulae x denotes the initial rating proportion for the programme, n denotes the sample size and r denotes the repeat viewing proportion. X^2 and t denote critical values for the chi-square and t distributions with one and (n-1) degrees of freedom respectively.

$$[X^2 +/- \sqrt{\{X^2(X^2 + 8nx(1-r))\}}]/[2n] \quad (2)$$

$$[t^2 +/- \sqrt{\{t^2(t^2 + 8nx(1-r) + 8t^2x(1-r))\}}]/[2n(1+t^2/n)] \quad (3)$$

For sample sizes as large as ours the above two tests are equivalent. In Figure 3 both these tests have been applied to random samples of size 518, the number of households in the panel, and to random samples of size 1241, the number of people in the panel. The least squares regression line indicated in Figure 1 has been used to predict repeat viewing proportion, r, from the initial rating, x. In Figure 3 the shaded region indicates when a significant change in rating has occurred when we consider the sample to be composed of 518 random or independent observations.

Method for a cluster sample

To assume that our panel of 1241 adults is actually equivalent to a random sample of 518 adults is, of course, being too conservative. To assume it is equivalent to a random sample of 1241 adults is too rash. The truth lies somewhere in between as suggested in recent research (Kish and Frankel, 1974, and Kish, 1987). That is, the effective sample size (ESS) is more than 518 but less than 1241. For reasonably large similarly sized clusters the effective sample size of our cluster sample could be approximated by equation (4), (Kish, 1987:41–45).

$$ESS = n/[1 + (b-1)\tau] \quad (4)$$

In this equation b denotes the average cluster size and τ, a measure of cluster homogeneity, is the pairwise correlation of sampled elements. For our data b=2,4 and, as indicated in Figure 4, τ is independent of rating with an average value of about 0,32 for ratings above 5%.

Replacing n by ESS in equations (2) or (3) one obtains a graphical test for determining when a significant change in rating has occurred for a cluster sample.

Method for a systematic stratified sample

The effect of stratification on effective sample size has been well documented (Kish 1987: 194–196). Although proportionate stratification will generally increase the effective sample size slightly, post-stratification with unequal weights serves to decrease effective sample size.

The effect of systematic sampling increases effective sample size, but this effect is more difficult to quantify. In one commonly used approach recommended by Wolter (1985: 250–251) it is assumed that a systematic sample may be regarded as a stratified random sample with all the strata containing only two units. If this assumption is reasonable and the rating is x, then the effective sample size for a systematic stratified sample with weights W_h is given by:

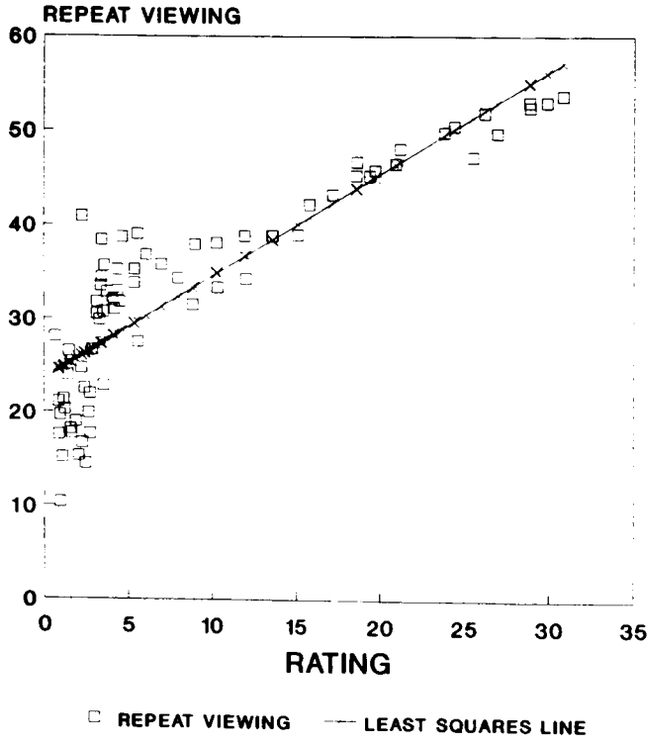

Figure 1 Repeat viewing as a function of rating

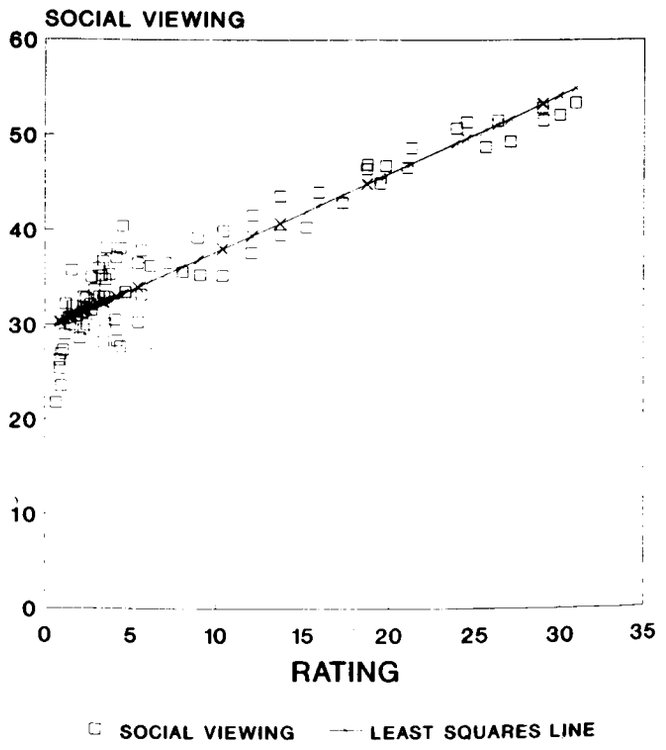

Figure 2 Social viewing as a function of rating

$$ESS = x(1-x)/\sigma^2 \quad (5)$$

where σ^2 is estimated by equation (6), with n_h equal to the number of adults within the hth stratum.

$$\sigma^2 = [\Sigma_h W_h^2 [\Sigma_i (y_{hi} - y_{h,i-1})^2]/\{2n_h(n_h-1)\}] \; / (\Sigma_h 15W_h)^2 \quad (6)$$

Employing this equation for the AMPS data it appears from Figure 5 that for ratings in excess of about 12%, the net effect of having systematic sampling and stratification is to increase the effective sample size. For convenience a generalised variance function has been fitted to the variance estimator. This function takes the form

$$LOG(\sigma^2) = f\{\sqrt{[x(1-x)/n]}\} \quad (7)$$

where f( ) is a fourth order polynomial.

Equations (5) and (7) can be used to obtain the effective sample size for the sample. Replacing n by this value in equation (2) or (3) one obtains another graphical test for the significance of rating change, this time for a systematic, stratified sample.

Method for a clustered, systematic and stratified sample

The effect of a cluster sample on top of a systematic stratified sample can be calculated using various methods. For instance, Wolter (1985) suggests that the variances of such samples be estimated using random group methods, balanced half-sample methods,

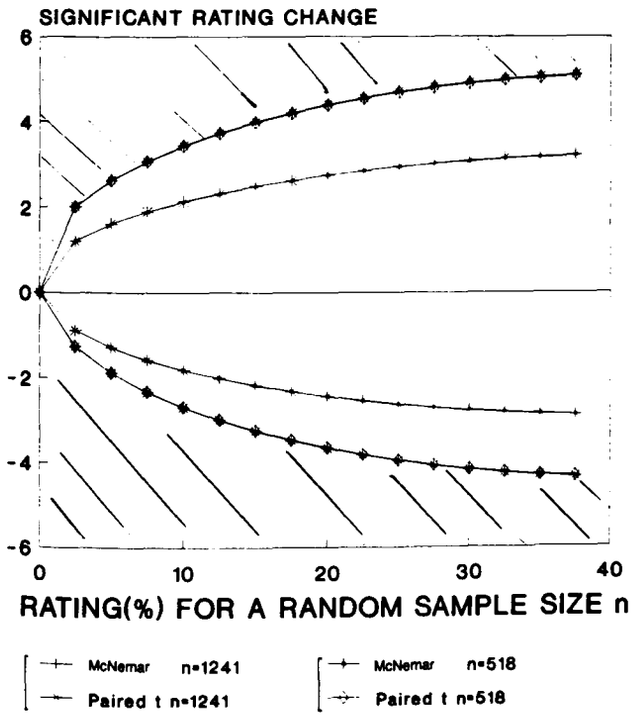

Figure 3 Significant rating changes for random samples at a 5% significance level

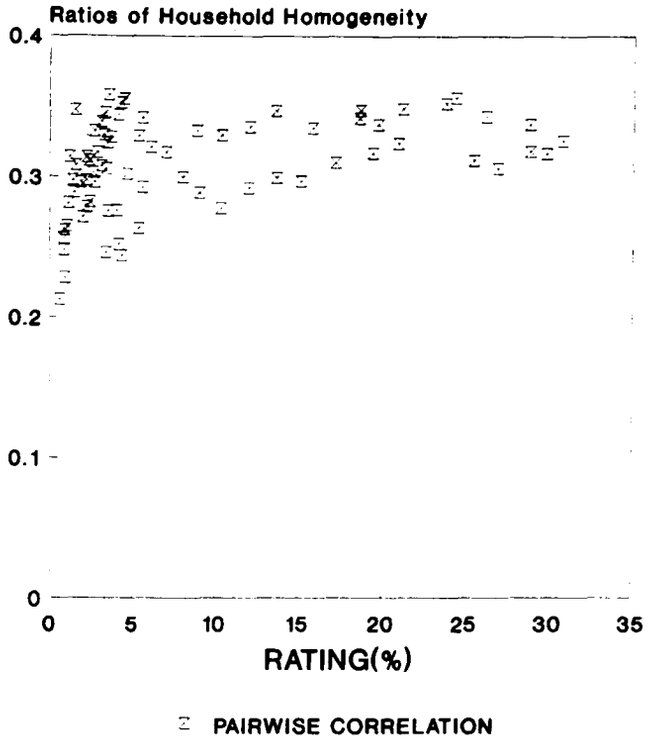

Figure 4 Ratios of household homogeneity (τ = pairwise correlation)

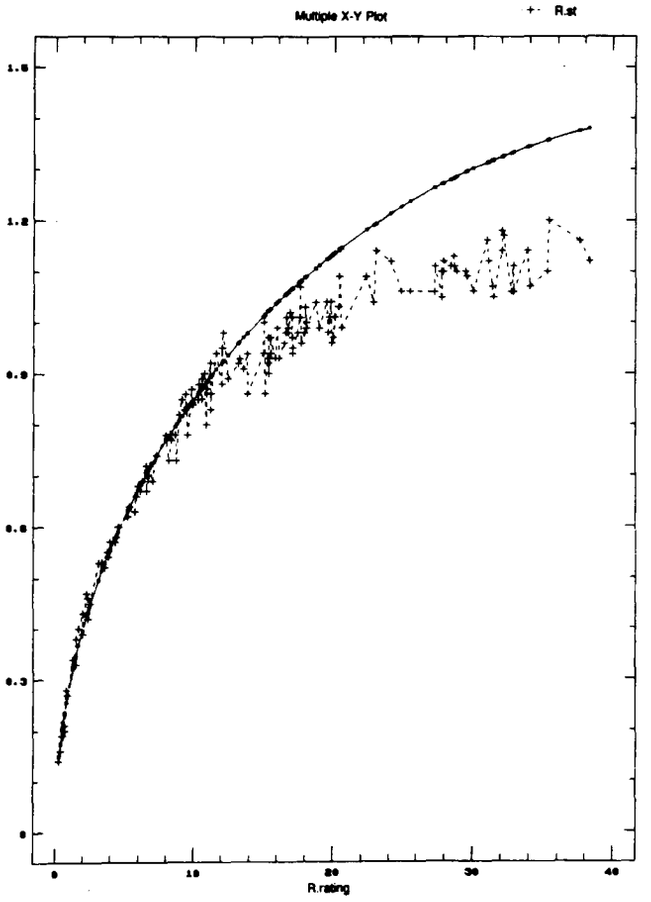

Figure 5 Comparison of rating standard errors calculated by the systematic/stratified method(+) and the random sample method with n=1241(.)

jackknife methods or Taylor series methods. Soong (1988) has used a jackknife estimator for estimating television rating variances so this was the method chosen in our study. This method was expected to have a conservative positive bias (Efron and Stein, 1981) but for our data it was convenient.

The households in our panel were systematically divided into ten groups. Ten ratings were calculated excluding one of the ten groups in turn. In each instance stratification weights were recalculated and the rating obtained by excluding the kth group was denoted by $x_{(k)}$. The jackknife estimator of the rating variance was then obtained from equation (8) using $x_{(.)}$ to denote the mean of the $x_{(k)}$.

$$\sigma^2 = 9(\Sigma_k(x_{(k)} - x_{(.)})^2/10) \quad (8)$$

Figure 6 indicates that this estimate of variance tends to be higher than that for a random sample of the same size. The same form of generalised variance function (7) was found to be appropriate for this data, although there was, of course, a difference in the coefficients. As before the effective sample size was calculated using equation (5) and substituted into equation (2) or (3) to obtain critical rating changes.

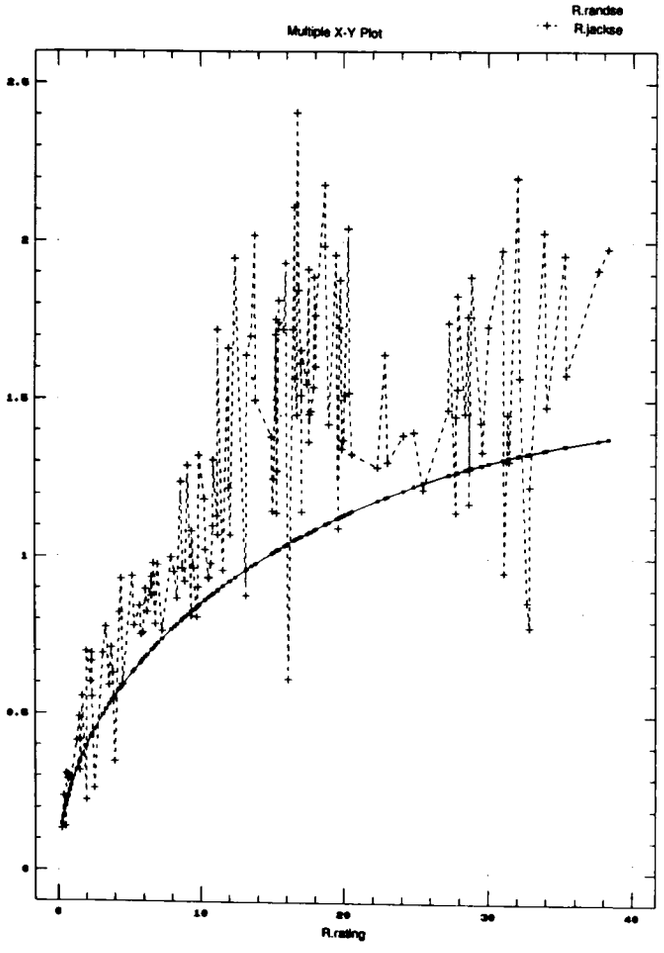

Figure 6 Comparison of rating standard errors calculated by the jackknife method(+) and the random sample method with n=1241(.)

Table 1 Effective sample sizes

	Method		
Rating	Cluster	Syst/Strat	Jackknife
5%	857	1184	829
10%	857	1316	718
15%	857	1421	638
20%	857	1506	633
25%	857	1599	708
30%	857	1710	864
35%	857	1831	1091

Results and discussion

Effective sample sizes for the three methods appear in Table 1. Somewhat surprisingly the jackknife estimator of effective sample size does not lie between the values for the cluster method and the systematic stratified method except at the highest rating levels. This suggests that the jackknife estimator of variance is indeed positively biased or that unequal stratum weightings more than negate the advantages of systematic, stratified sampling.

The accuracy of the three methods was compared using a Monte Carlo simulation. The linear relationships suggested in Figures 1 and 2 were used to generate results for complex samples of size 1241. Equal weights were assumed for the various strata. For 1000 iterations we counted for each method the number of instances in which ratings were found to be significantly different when this was actually false. At a 5% (and 1%) significance level the cluster method was found to be the most reliable in the sense that for no rating did the proportion of incorrect test decisions differ markedly from the chosen significance level.

As indicated in Table 2 the systematic, stratified

Table 2 Percentage of iterations for which rating equality was erroneously rejected at a nominal 5% significance level

	Method		
Rating	Cluster	Syst/Strat	Jackknife
5%	4,5	9,1	4,3
10%	4,1	8,9	2,4
15%	2,8	11,5	1,1
20%	3,8	11,9	1,4
25%	5,5	15,4	3,8
30%	4,0	17,5	4,0
35%	5,5	18,9	8,2

method was unreliable at all rating levels. This supports a claim by Kish (1987, 194) that any reduction in variance due to stratification tends to disappear when means are compared. It also suggests that this approach confuses clustering homogeneity with systematic trends in ratings over postal code. The jackknife method was too conservative in that significantly less than 5% of the test decisions were incorrect for several ratings. At this stage it is impossible to ascertain whether this was due to the fact that strata weights were assumed equal in the simulation or whether it was the result of positive bias in the jackknife variance estimates. In practice, the strata weights used are very similar. This suggests that the conservative results were probably due to positive bias in the jackknife variance estimates.

The cluster method is therefore recommended as the most appropriate test for the significance of rating change. This method is the simplest to apply in practise. The shaded region in Figure 7 indicates significant rating changes for the recommended method at a 5% level of significance. The recommended test is clearly more conservative than a random sample test with n=1241, the number of adults in the panel, but less conservative than a random sample test with n=518, the number of households in the panel.

The quantities average cluster size, b, and cluster homogeneity, τ, can be expected to remain static over time. In addition, τ is 'portable', (Kish 1987, 203) in the sense that its value is unchanged when ratings are required for a subclass of the population, such as males or females. This means that the effective sample size for subclasses of the population can also be obtained from equation (4) after multiplying n and b by the proportion of subclass members in the population. This suggests that the recommended method is flexible as well as being simple to apply.

Acknowledgement

My thanks to Leslie Pels and Igbal Hoosen from IBIS for all the computer runs, to Piet Smit from SAARF for permission to publish, and to Prof. D.J. Stoker, Prof. G.V. Kass and Dr. C.C. Frangos for their advice.

References

Efron B. & Stein C. 1981. The jackknife estimate of variance. *Ann. Stat.:* 9, 3, 586–596.

Holt D. & Smith T.M. 1979. Post stratification. *J.R. Statist. Soc. A:* 142, 1, 33–46.

Kish L. & Frankel M.R. 1974. Inference from complex Samples. *J.R. Statist. Soc. B:* 36, 1–22.

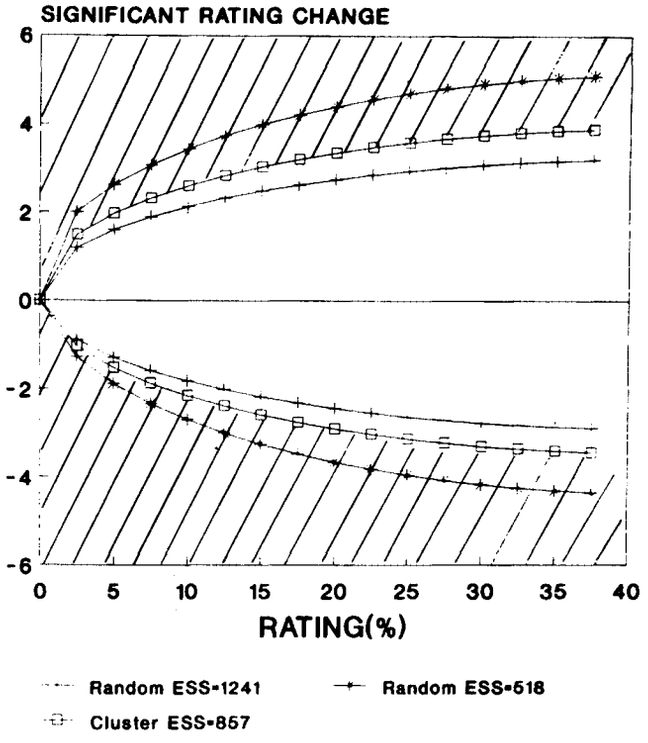

Figure 7 Significant rating changes at a 5% significance level using the recommended cluster method

Kish L. 1987. *Statistical design for research.* John Wiley & Sons, Inc.

Meyer D.H. 1988. Testing for significant changes in popularity. *S. Afr. J. Bus. Mgmt:* 19, 3, 96–98.

Soong R. 1988. *J. Advertising Research,* 50–56.

Wolter K.M. 1985. *Introduction to Variance Estimation.* Springer-Verlag, New York Inc.

Appendix A

Glossary of Terms

Sample size (n). The number of sampled units.

Effective Sample Size (ESS). The effective number of independent information units contained in the sample.

Panel Sample. A sample from which information is sought on several different occasions.

Random Sample. A sample of independently chosen units.

Cluster Sample. A sample with sampling units grouped into somewhat homogeneous clusters.

Systematic Sample. A sample containing every jth unit in the population.

Proportionate Stratification. Division of a population into non-overlapping subpopulations called strata and then sampling of the strata in proportion to their size.

Post-stratification. Weighting of strata after selection.