The estimation and presentation of standard errors in a survey report

Renè Swanepoel
David J Stoker

Statistics South Africa

July 2000

Acknowledgements

This study was undertaken by the first author with the view to submit it in partial fulfillment of the requirements for the Masters degree in Mathematical Statistics at the University of Pretoria. The study was done under the supervision of the second author, who suggested and initiated the study in his capacity as a consultant to Statistics South Africa.

The first author would like to express her sincere appreciation to Statistics South Africa for making available to her the data sets for the study with the view to develop and test modeling techniques used for the presentation of standard errors in publications. These data sets were: The October Household Surveys (OHS) of 1995, 1996 and 1997, and the Victims of Crime Survey (VOC) of 1998.

It must, however, be emphasized that the three OHS data sets differ from the final released OHS data sets in that the weighting of the data records was based on the adjusted (for growth) 1991 population census data and not on the 1996 population census data. Consequently, the estimates (i.e. estimated values) of population characteristics (such as unemployment rate) appearing in tables in this study, may or will differ from the final released data. For this reason, **all estimates appearing in this study must be considered as privileged and unofficial and may not be quoted in any way whatsoever**.

Note that the purpose of the study was not to estimate the population characteristics as such, but to model standard errors of the estimated population characteristics with the view to enable readers of survey reports to evaluate the precision of such estimated values.

Contents

Abstract

The vast number of different study variables or population characteristics and the different domains of interest in a survey make it impractical and almost impossible to calculate and publish standard errors for each statistic (estimated value of a population variable or characteristic) and for each domain individually. However, it is advisable not to publish standard errors for only a small number of statistics for a few selected domains or to omit them altogether in a survey report. Since the estimated values are subject to statistical variation (resulting from the probability sampling), they can be evaluated only if their precision is known.

The purpose of this research project was to study the feasibility of using mathematical models to estimate the standard errors of estimated values of population parameters or characteristics in survey data sets regularly gathered by Statistics South Africa, and to investigate effective and user-friendly presentation methods of these models in reports.

The following data sets were used in the investigation:

- October Household Survey (OHS) 1995
- OHS 1996
- OHS 1997
- Victims of Crime Survey (VOC) 1998

Note that the OHS data sets consist of various sections, of which the persons section (which contains information on person demographics), workers section (containing information on economic activity, employment, etc.) and the household section (containing information on household characteristics) were used.

The basic methodology followed was to calculate estimates of the standard errors of the statistics considered in the survey for a variety of domains (such as the whole country, provinces, urban/rural areas, population groups, gender and age groups as well as combinations of these). This was done using a computer program that takes into consideration the complexity of the different sample designs. A set of domains covered a large variety of sample sizes, ranging from a very small number of sample records up to the whole data set. The standard errors obtained in this way are referred to as *direct calculated standard errors.*

A regression model was then fitted to such a set of estimated domain values of a statistic and the associated direct calculated domain standard errors, where a function of the standard error value is considered as the dependent variable and a function of the size of the statistic is considered as the independent variable.

A linear model, equating the natural logarithm of the coefficient of relative variation of a statistic to a linear function of the natural logarithm of the size of

the statistic, gave an adequate fit in most cases considered in this study. Well-known tests for the occurrence of outliers were applied in the fitting of the model. When an observation (sample record) was indicated as such, it was established whether the observation could be deleted legitimately (e.g. when the domain sample size was too small, or the estimate biased). After the deletion of such observations, the fitting process was repeated.

The above model is the same model used by the Australian Bureau of Statistics in similar surveys. They derived this model especially for variables that count people belonging to a specific category. It was found that this model performs equally well when the variable of interest counts households instead of persons, or counts incidents as in the case of the VOC.

It is interesting to note that the set of domains considered in the fitting process includes segregated classes, mixed classes and cross-classes. Thus, the model can be used irrespective of the type of subclass domain. This result makes it possible to use the same model to predict the standard error of an estimated value of a study variable for any type of domain.

Although the fitted model can be used directly to approximate the standard error associated with an estimated value of a population characteristic; the model as a mathematical formula, is not a user-friendly method of presenting the precision of estimates. Consequently, user-friendly and effective presentation methods of standard errors are summarized in this report. The suitability of a specific presentation method, however, depends on the extent of the survey and the number of study variables involved.

1. Introduction

Addressing the presentation of standard errors is a common problem every survey statistician has to deal with during the compilation of a survey report. The problem is two-fold. Firstly, standard errors of the published survey statistics needs to be calculated, and then presented in a simple, comprehensive and cost effective way in the publication.

All estimates of population parameters or characteristics derived from sample survey data are subject to errors. These errors are divided into two categories, viz. sampling- and non-sampling errors. Sampling errors refer to the probabilistic nature of a sample and can be explained as the error made when the sample used for the specific survey is only one of a large number of possible samples of the same size and sample design that could have been selected. Non-sampling errors refer to response differences, definitional difficulties, respondent inability to recall information, etc.

It is impractical to include in a survey report standard errors for each and every statistic, for each and every domain of interest and, taking into account the time absorbency of these complex calculation procedures; it would be an impossible task.

The easiest approach would be to omit standard errors totally from the publication, but there are certain criteria to which published results, subject to the above mentioned errors, have to conform (Gonzalez, Ogus, Shapiro and Tepping; March1974):

a) The user must be informed of the different errors which play a role and the limiting effects of these errors on the results. An explanation of how to interpret standard errors and confidence intervals should be included.

b) The implications of the sample design on the various sources of error must be clearly indicated, e.g. what the effect of an old or incomplete sampling frame could be on the data.
c) If missing data was imputed, it should be mentioned as well as the imputation method that was used and the implications this could have on the results.
d) Standard errors should be displayed in an organized manner and be thoroughly explained.
e) If the results in a survey report are subject to large survey errors, users should be adequately warned against lack of reliability of such data.

Alternatively indirect methods can be used, i.e. to model standard errors of the survey estimates instead of calculating standard errors for each statistic individually.

The purpose of this research project is to investigate and introduce alternative methods to generate and present standard errors in an efficient way in a survey report. Different aspects which play a role in choosing an acceptable model to approximate standard errors are investigated. Also included, among other factors, is the influence of the size of the subclass to which the estimate belongs in the model, the effect of the population parameter being estimated in the model and the possible influence that cross-class, segregated class or mixed class domains could have on the model.

2. A Different Approach

2.1 Indirect methods of estimating standard errors

The use of indirect methods to estimate standard errors have been practiced with satisfactory results by several countries, some of which include the USA, Australia and Sweden. Different models are used according to suitability and preferences. Part of this project is to test and examine some of these models for suitability on the data sets made available by Statistics South Africa.

The data sets used in the research project are the October Household Surveys (OHS) of Statistics South Africa of 1995, 1996 and 1997 and the Victims of Crime survey (VOC) of 1998. The OHS consists of more than one section, including the persons section, the workers section and the household section. The sample sizes for the OHS of 1995 and 1997 were 30 000 households and for the OHS of 1996 they were 16000 households. The VOC reports on the crimes committed against members of the households, including the violent and non-violent crimes, in South Africa. The sample for the VOC consisted of 4000 households from which one person, aged 16 years or older, was selected to be interviewed.

Usually South African data sets, e.g. the workers subset of the OHS with a target population of all economically active people between the ages of 15 and 65 years, have a unique composition. This is due to the inclusion of four different race groups in the data sets, the different provinces being covered as well as the substantial differences between urban and rural areas in South Africa. All these different classes lead to a large variety of domains of interest in SA data sets, in addition to the usual gender by age type of domains. This adds to the complexity of calculating standard errors.

2.2 Levels of domains of interest

Table 1: Levels of domains used in this research project

Subclass	Number of categories	Type of class
RSA	1	Segregated class
Province	9	Segregated class
Urban / Rural (U/R)	2	Segregated class
E A type [1]	5	Segregated or cross-class
Race	4	Mixed classes
Gender	2	Cross-class

[1] Includes 5 different types: Type 1 – Urban formal, Type 2 – Urban informal, Type 3 – Tribal, Type 4 – Commercial farms and Type 5 – Other non-urban.

Age group [2]	3	Cross-class
Province by U/R	18	
Province by gender	18	
Province by age group	27	
Province by race	36	
U/R by Race	8	
U/R by gender	4	
U/R by age group	6	
Race by gender	8	
Race by age group	12	
Gender by age group	6	

A large number of categories for a subclass may have the consequence that the sample sizes of some of the subclass categories become too small to be included in the modeling procedure.

2.3 Models proposed by other countries

2.3.1 The United States

In the USA, Generalized Variance Functions were used to estimate standard errors for the SESTAT survey which combines information from three National Science Foundation-sponsored surveys (*Sampling Errors For SESTAT and Its Component Surveys, 1993*):

- The National Survey of College Graduates
- The Survey of Doctorate Recipients, and
- The National Survey of Recent College Graduates

Two other surveys in the United States that also make use of Generalized Variance Functions are the Current Population Survey (CPS) and the National Health Interview Survey (HIS) (*Generalized Variance Functions in Stratified Two-Stage Sampling, Richard Valliant, 1987*).

Generalized Variance Functions (GVFs) are mathematical functions that describe the relationship between a population parameter (such as a population total) and its corresponding variance. GVFs provide users with a quick and simple way to model standard errors. The user inserts the estimated value of the statistic of interest into the fitted GVF model to generate a model-based approximation of the variance.

[2] Includes 3 different age groups: Age group 1 – between 15 and 30 years, Age group 2 – between 31 and 45 years and Age group 3 – between 46 and 65 years.

A GVF depends on the assumption that the relative variance of an estimated population parameter, $\hat{Y}$, is a decreasing function of the magnitude of the estimate:

$$\text{Relvar}(\hat{Y}) = a + bY^{-1} \qquad [1]$$

where a and b are known as the GVF parameters.

The relationship [1] can be derived as follows. Consider a sample of n units from a population of size N, where $\hat{P}$ denotes the estimate of the proportion $P = \frac{Y}{N}$ of a population characteristic, and Y is some counting variable measuring the occurrence of the characteristic. Let D be the design effect accounting for departures from simple random sampling. The probability sampling relative variance[3] of $\hat{P}$ is then:

$$\begin{aligned} \text{Rel}Var_p &= \frac{DP(1-P)}{nP^2} \\ &= \frac{D(1-P)}{nP} \\ &= \frac{D-DP}{n\frac{Y}{N}} \\ &= \frac{-D}{n} + \frac{ND}{nY} \end{aligned} \qquad [2]$$

which is of the form: $\text{Relvar}(\hat{Y}) = a + bY^{-1}$

(*Generalized Variance Functions in Stratified Two-Stage Sampling, Richard Valliant, 1987*)

To derive the estimated standard error from this model is very simple:

$$\begin{aligned} &Var\,(\hat{Y}) \doteq \hat{a}\hat{Y}^2 + \hat{b}\hat{Y} \\ &\therefore SE(\hat{Y}) \doteq \sqrt{\hat{a}\hat{Y}^2 + \hat{b}\hat{Y}} \end{aligned} \qquad [3]$$

(Richard Valliant, 1987)

This formula can be adapted to estimate the standard error (as a %) of an estimated percentage:

$$SE(\hat{P}) \doteq \sqrt{\frac{\hat{b}}{\hat{Y}}\hat{P}(100-\hat{P})}$$

[3] Definition of relative variance: Rel$Var = (CV)^2$ where CV is the coefficient of relative variation

[4]

where $\hat{P}$ is the estimated percentage, $\hat{Y}$ is the estimated value of Y and $\hat{b} = \frac{Nd}{n}$ with d an estimate of D (Richard Valliant, 1987).

Obtaining the GVF parameters requires the calculation of a number of variances of the survey statistics through direct calculation methods, e.g. in the SESTAT survey the successive difference replication method was used. The a and b parameters are then estimated by fitting the model to these survey estimates and their variances.

2.3.2 Models proposed by Lepkovski

Lepkovski introduced the use of Generalized Variance Functions to estimate standard errors in a survey report (*Presentation of Sampling Errors; Lepkovski, 1998*).

Models proposed by Lepkovski are basically the same as model [1] and other mathematical derivations from this model, e.g. continuing with the same proof as in [2]:

$$\text{Rel}Var_p = \frac{D(1-P)}{nP}$$

$$= \frac{DN(1-P)}{nY}$$

and when P is small

$$\doteq \frac{DN}{nY}$$

This approximation gives a model of the form:

$$\text{Rel}Var_p = cY^{-1}$$

[5]

Formula [5] can be converted into the coefficient of variance by taking the square root:

$$\text{Rel}Var_p^{\frac{1}{2}} = (cY^{-1})^{\frac{1}{2}}$$

$$\frac{1}{2}\log(\text{Rel}Var_p) = \frac{1}{2}\log(c) - \frac{1}{2}\log(Y)$$

$$\log(\text{Rel}Var_p) = c' + k\log(Y)$$

[6]

where $c' = \log(c)$ and $k = -1$

(Ghangurde, 1981; Kalton, 1977)

Model [6] is used by the Australian Bureau of Statistics and by Statistics Canada.

(Richard Valliant, 1987)

2.3.3 The Australian Bureau of Statistics (ABS)

The ABS has derived mathematical models by applying smoothing regression techniques on the standard errors calculated through split-half techniques (*Household Collection Support – Standard Error Manual, ABS; 1997*).

The following assumptions in deriving the models were made: Simple random sampling without replacement (SRSWOR) is used to draw the sample of size n from the population of size N . Y_c denotes the number of people in category c and is estimated by:

$$\hat{Y}_c = N\hat{P}_c$$

[7]

where p_c is the proportion of the sample in category c .

$$E(\hat{Y}_c) = NP_c$$

[8]

$$Var(\hat{P}_c) = \frac{1}{n}\left(\frac{N-n}{N-1}\right)P_cQ_c$$

(Cochran, 1977)

Thus:

$$Var(\hat{Y}_c) = \frac{N^2}{n}\left(\frac{N-n}{N-1}\right)P_cQ_c$$

[9]

where $P_c = \frac{Y_c}{N}$ and $Q_c = 1 - P_c$

The relative standard error % $(RSE\%)$ of $\hat{Y}_c$ is:

$$\begin{aligned} RSE\%(\hat{Y}_c) &= \frac{\sqrt{Var(\hat{Y}_c)}}{\hat{Y}_c} \times 100 \\ &= \frac{\sqrt{\frac{N^2}{n}\left(\frac{N-n}{N-1}\right)P_cQ_c}}{NP_c} \times 100 \\ &= \sqrt{\frac{\frac{N^2}{n}\left(\frac{N-n}{N-1}\right)P_cQ_c}{N^2P_c^{\,2}}} \times 100 \\ &= \sqrt{\frac{(N-n)Q_c}{n(N-1)P_c}} \times 100 \end{aligned}$$

$$\cong \sqrt{\frac{(N-n)Q_c}{nNP_c}} \times 100$$
$$= \sqrt{\frac{1-f}{f}\frac{Q_c}{Y_c}} \times 100$$

[10]

where $f = \frac{n}{N}$ and denotes the sampling fraction.
(*Household Collection Support – Standard Error Manual, ABS; 1997*)

However, usually the survey sample is not drawn with SRSWOR and to compensate for the design effect, formula [10] should be adapted to take the design effect into account:

$$RSE\%(\hat{Y}_c) \cong \sqrt{d}\sqrt{\frac{(1-f)}{f}\frac{Q_c}{Y_c}} \times 100$$

[11]

where d is an estimate of D, the design effect.

In exactly the same manner one can derive the relative standard error % for the ratio estimator $\hat{R}$ ($R = \frac{Y}{X}$) by making use of the variance-formula of $\hat{R}$:

$$V(\hat{R}) \doteq \frac{1-f}{n\bar{X}^2}\frac{XR(1-R)}{N-1}$$

[12]

(Cochran, 1977)

Thus:

$$RSE\%(\hat{R}) \doteq \sqrt{d}\sqrt{\frac{1-f}{f}\frac{(1-R)}{Y}} \times 100$$

[13]

If the natural logarithm is taken, we get:

from [11] $$\ln RSE\%(\hat{Y}_c) = a_c - \frac{1}{2}\ln(Y_c) + \frac{1}{2}\ln(1-P_c)$$

[14]

or from [13] $$\ln RSE\%(\hat{R}) = a_c - \frac{1}{2}\ln(Y) + \frac{1}{2}\ln(1-R)$$

[15]

where the factor a_c depends on the category considered through the design effect d and on the population size through f. (*Household Collection Support – Standard Error Manual, ABS; 1997*)

If a_c is correlated with Y_c, or with P_c (or R), the coefficients of $ln(Y_c)$, $\ln(1-P_c)$, or $\ln(1-R)$ would be different from 0.5.

When the model is fitted to the data, population parameters are replaced by their estimated values. Changing from percentage to proportion:

<u>Model 1</u>

$$ln(cv(\hat{Y}_c)) = a + b\,ln(\hat{Y}_c) + c\,ln(1-\hat{P}_c)$$

or

$$\ln(cv(\hat{R})) = a + b\ln(\hat{Y}) + c\ln(1-\hat{R})$$

where cv denotes the estimated coefficient of relative variation.

$\hat{P}_c$ or $\hat{R}$, in addition to $\hat{Y}_c$ *or* $\hat{Y}$, are independent variables in the above models and that adds to the degree of difficulty when the models are used in practice. The above models can be simplified to:

<u>Model 2</u>

$$ln\,cv(\hat{Y}_c) = a + b\,ln(\hat{Y}_c)$$

or

$$\ln cv(\hat{R}) = a + b\ln(\hat{Y})$$

In the cases where the value $\hat{P}_c$ (or $\hat{R}$) gets nearer to 1, Model 2 tends to result into a larger value for $cv(\hat{Y}_c)$ or $cv(\hat{R})$ than really exists and this consequently gives rise to outliers (*Household Collection Support – Standard Error Manual, ABS; 1997*). One possible solution to compensate in Model 2 for the additional term in Model 1, is to include a quadratic term into Model 2 leading to Model 3:

<u>Model 3</u>

$$ln\,cv(\hat{Y}_c) = a + b\,ln(\hat{Y}_c) + c\left(ln(\hat{Y}_c)\right)^2$$

or

$$\ln cv(\hat{R}) = a + b\ln(\hat{Y}) + c\left(\ln(\hat{Y})\right)^2$$

using $\left(ln(\hat{Y}_c)\right)^2$ as a rough substitute for $\ln(1-\hat{P}_c)$ or $\left(\ln(\hat{Y})\right)^2$ as a rough substitute for $\ln(1-\hat{R})$ (*Household Collection Support – Standard Error Manual, ABS; 1997*).

3. The Modeling Procedure

3.1 Estimation of the model parameters

The estimation of the parameters in the standard error models requires the calculation of the variances of a number of typical survey estimates through direct methods. Although it is not necessary to calculate the variance of each survey estimate directly, a larger number of related survey estimates and their variances that cover a wide range of the domains of interest would contribute to a more representative model.

There are several different ways to calculate variances directly. SESTAT made use of successive differences techniques and resampling methods such as random groups, balanced repeated replication and jack-knife replication. The ABS used split-half techniques where the sample is split into two similar sections to calculate standard errors directly.

In this research project SAS programs were used to calculate the relative variances and standard errors of complex sample estimates for different domains of interest. Prof. D.J. Stoker, a consultant to Statistics SA, developed the programs. These programs make it very easy to calculate standard errors and coefficients of variance for every desired set of domains of interest of a specific estimate by simply changing the categorical variable criteria in the program macro.

For a variety of population parameters or characteristics the standard error model is fitted to the estimated coefficients of variance obtained for the set of domains of interest, by making use of Least Squares regression or Maximum Likelihood regression. Survey estimates of both large values and small values should be included in the model-fitting procedure. This will contribute to a good fit of the model at large, small and in-between values of the estimates.

3.2 Procedure of fitting data to the model

There are many software packages that make regression modeling very easy, e.g. Statistica, Statsgraphics, SPSS, SAS, Microsoft Excel and many more. SAS INSIGHT was used to do model fitting for this project.

Choosing the best model mainly depends on finding the model with the highest coefficient of determination R^2. The R^2-value gives the proportion of the variability in the dependent variable that can be explained by the fitted regression line. If the fitting results are not satisfactory, it can either be due to the existence of outliers or a model that is not suitable for the data.

After the fitting procedure, the outliers must be identified, if there are any. If it is justified to exclude the outliers from the calculations, it is recommended to repeat the fitting procedure without the outlier-observations. Consequently, it is very important to first try to establish the reason for the outlier's occurrence. It was found that most outliers occur because of one of the following reasons:

a) The sample size of the domain on which the estimate is based is too small. The size of the domain, whether it is too small, depends on the sizes of the other domains of interest in the survey. Thus, it seems that there does not exist a definite cut-off point in the size of the domain that can be identified as too small.

However, it was found to be usually the case that if the sample size of a specific domain of interest was as small as 10, it had to be discarded from the data set or else it produced outliers in the data. Estimated proportions, $\hat{P}_c$ or $\hat{R}$, that are found to be close to 0 or 1, for modeled coefficients of relative variation from the model $ln(cv(\hat{Y}_c)) = a + b\,ln(\hat{Y}_c)$, generally resulted in values that differ largely from the direct calculated values.

Statistics SA does not publish estimates for too small sample sizes of the domains of interest. Thus, these estimates can be excluded from the modeling procedure of standard errors.

A possible solution for some of these cases would be to use Model 3 (page 13) instead of the above model. The factor $\ln(1-\hat{P}_c)$ becomes important when $\hat{P}_c \approx 1$ or $\hat{R} \approx 1$. To compensate for this, a quadratic term, $\left(ln(\hat{Y}_c)\right)^2$ or $\left(\ln(\hat{Y})\right)^2$, is included in Model 3 which then serves as a rough substitute for $\ln(1-\hat{P}_c)$ or $\ln(1-\hat{R})$ (*Household Collection Support – Standard Error Manual, ABS; 1997*).

b) Survey estimates with direct calculated coefficients of relative variance larger than 0.1 $\left(cv(\hat{X}) \doteq \sqrt{\frac{Var(\hat{X}_c)}{\hat{X}_c^{\,2}}} > 0.1\right)$ may result in outliers, but it depends on the whole data set. However, for ratio estimation the estimate can be biased to the extent that the estimate becomes misleading when $cv(\hat{X}) > 0.15$, with X denoting the variable in the denominator of the ratio. Such cases should thus be excluded in the modeling procedure.

c) Outliers were also observed for subclasses of the domain under consideration where $\hat{P}_c$ or $\hat{R}$ ≈ 1 for some subclasses and $\hat{P}_c$ or $\hat{R}$ ≈ 0 for other subclasses. Such cases are for example " water on site " and " toilet on site " which are applicable to almost all households in the formal urban area, but at the same time, are applicable to almost none of the households in the informal urban area.

d) The set of domains of interest used in the modeling procedure should relate to the study variable of which the standard error is required. Otherwise, it can result in outliers. An example is where the study variable is the total number of households for each different dwelling-type according to province, race and urban or rural area. If estimates of the total number of households in each province, but not according to dwelling-type, are included in the modeling procedure, outliers will surely appear.

3.3 Goodness of the fit and identification of outliers

Apart from the R^2-value as an indication of how well the model fits the data, there are other guidelines and tests that help with deciding if the model is suitable. These tests are easy to perform with the help of a statistical software package such as SAS Insight.

One possibility is to investigate the distribution of the residuals. If the model is suitable for the data, the residuals would follow, or very nearly follow, a normal distribution. A Normal probability plot is very useful in indicating gross departures from normality, which can either be because the data does not fit the model, or because of the presence of outliers.

Another practical guideline to follow is to plot the standardized residuals versus the observed values. Nearly all the residuals should lie between the $-2s$ and $+2s$ confidence bands. In a good fit, the residuals will be scattered randomly around the X-axis with the larger concentration near the X-axis. Residuals lying outside of the $2s$ bands could indicate the presence of outliers.

Alternatively, the absolute values of the standardized residuals are considered. Values larger than 2 could indicate outliers while values larger than 3 should be regarded as outliers; i. e. $|e_i^*| > 3$ where e_i^* denotes the standardized residual.

All these tests will be discussed further in an example.

3.4 Example: Finding the best suitable model for the data

For illustrative purposes, the results of the fitted regression model on the 1997 October Household Survey Workers data set, are summarized and discussed below. The results of all the other investigated surveys are included in Appendix A.

In Appendix B, the results of the Workers data set of the OHS of 1997 are summarized, created after the necessary SAS programs were run to estimate the standard error and coefficient of relative variation for the study variable of interest (number of unemployed, and the unemployment rate in this case). This data set in Appendix B was used in the regression modeling procedure to find a suitable standard error model for the 1997 OHS Worker data set.

The discussion below refers to pages 19 & 20.

Figure 1: Plot of the linear relationship between $\ln(cv(\hat{Y}))$ as the dependent variable, and $\ln(\hat{Y})$ as the predictor, where $\hat{Y}$ denotes the estimated population total of unemployed in South Africa.

The model being fitted to the data is: $\ln(cv(\hat{Y})) = a + b\ln(\hat{Y})$, where a and b are the model parameters that should be estimated by using LS Regression.

Figure 1 shows that a linear relationship between $\ln(cv(\hat{Y}))$ and $\ln(\hat{Y})$ does exist. All the observations were included in this graph without excluding any outliers.

Table 3: The summary of the regression fit results

The R^2 value of 0.9284 shows that the model that was fitted on the data can explain almost 93% of the variation in $\ln(cv(\hat{Y}))$, giving evidence that the model is suitable for this data set.

Table 4: A summary of the estimated parameters

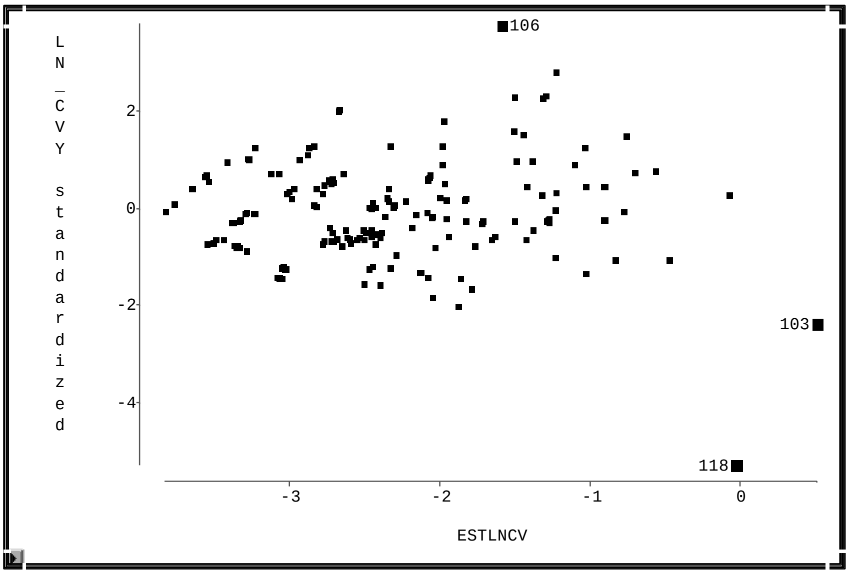
The small exceedance probabilities 0.0001 for both parameters show that both the parameters are significant in the model.

Figure 2: Plot of the residual values of $\ln(cv(\hat{Y}))$ versus the predicted values of $\ln(cv(\hat{Y}))$.

The residual- versus predicted values plot (Figure 2) serves as a test for outliers and to diagnose non-constant error variance. The residual values (take notice: not standardized residuals) seem to be randomly scattered around 0. There is a possibility that observations 103, 106 and 118 could be outliers, because they are lying outside the band containing the majority of residuals. These observations require further testing.

When the standardized residuals are plot against the predicted values of $\ln(cv(\hat{Y}))$, observations 106 and 118 are lying substantially outside the $2s$ bands (refer to Figure 2A).

Figure 2A: Standardized Residual Plot



The test $|e_i^*| > 3$ where e_i^* is the standardized residual, identified values 106 and 118 as outliers. This could be because the subclass sample sizes in both cases were small. In the repeated regression fit model these values are excluded to test whether a significant increase in the results appears when the outliers are excluded.

Figure 3: Residual Normal Quantile Quantile plot of the residual of $\ln(cv(\hat{Y}))$ versus the residual normal quantiles of $\ln(cv(\hat{Y}))$.

The Residual Normal QQ plot displays the extent to which the residuals are normally distributed. The empirical quantiles are plotted against the quantiles of a standard normal distribution. If the residuals follow a normal distribution, which is evident of a good fit, the points tend to fall along a straight line.

From Figure 3 it appears as if the residuals do follow a normal distribution with probable outlier observations at the upper – and the lower end of the plot. This gives further evidence that this model is suitable for this data set.

The next step is to repeat the whole fitting procedure, excluding the identified outliers, to see if there is a significant improvement in the fit.

On page 20 the second set of fitting results is given. The R^2 value increased to 0.9421. The model $\ln(cv(\hat{Y})) = 2.588 - 0.4382\ln(\hat{Y})$ can thus be accepted as a suitable model to approximate the standard errors for $\hat{Y}$, the estimated total number of unemployed people in a subclass for the workers data set of the 1997 OHS.

To derive the standard error from the model, the following conversion needs to be done:

$$\begin{aligned} cv(\hat{Y}) &= e^{2.588} \times e^{-0.4382\ln(\hat{Y})} \\ &= e^{2.588} \times \left(\hat{Y}\right)^{-0.4382} \\ &= 13.303 \times \left(\hat{Y}\right)^{-0.4382} \end{aligned}$$

$$\begin{aligned} \therefore se(\hat{Y}) &= cv(\hat{Y}) \times \hat{Y} \\ &= 13.303 \times \left(\hat{Y}\right)^{1-0.4382} \\ &= 13.303\left(\hat{Y}\right)^{0.5618} \end{aligned}$$

If for example $\hat{Y} = 81091$, then $se(\hat{Y}) = 13.303 \times (81091)^{0.5618}$

$$\begin{aligned} &= 13.303 \times 572.6125 \\ &= 7617 \end{aligned}$$

Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}$, the estimated population total unemployed in South Africa, according to the strict definition of unemployment, as predicted by the natural logarithm of $\hat{Y}$. (Source: OHS 1997 - Workers)

Model: $\ln(cv(\hat{Y})) = 2.5078 - 0.4312\ln(\hat{Y})$

Figure 1:

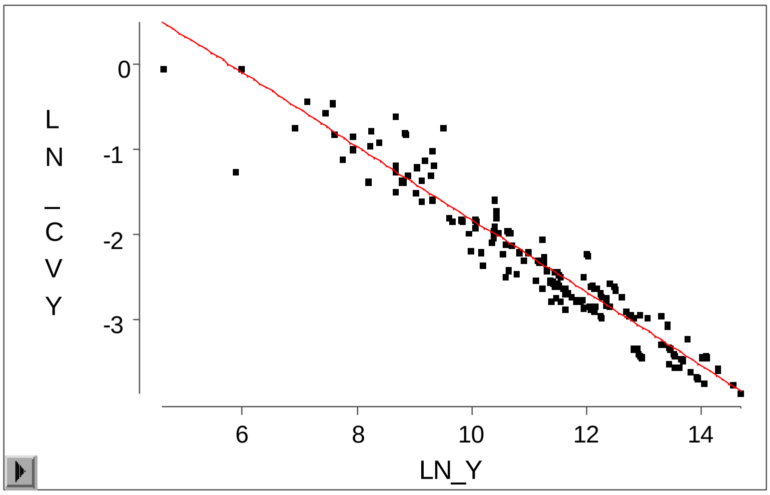


Table 3:

Parametric Regression Fit								
			Model		Error			
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob > F
	1	1	133.8087	195	0.0529	0.9284	2529.1800	0.0001

Table 4:

Parameter Estimates							
Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.5078	0.0974	25.7393	0.0001	.	0
LN_Y	1	-0.4312	0.0086	-50.2910	0.0001	1.0000	1.0000

Figure 2:

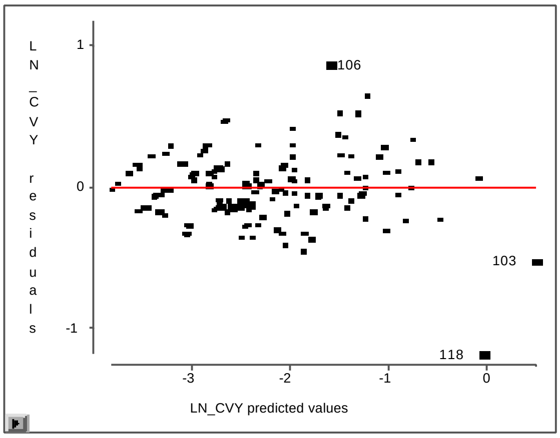
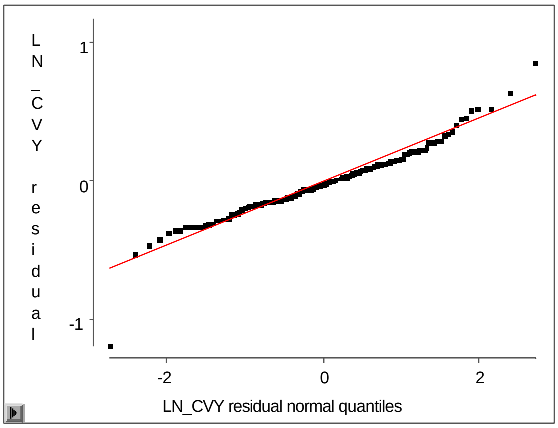


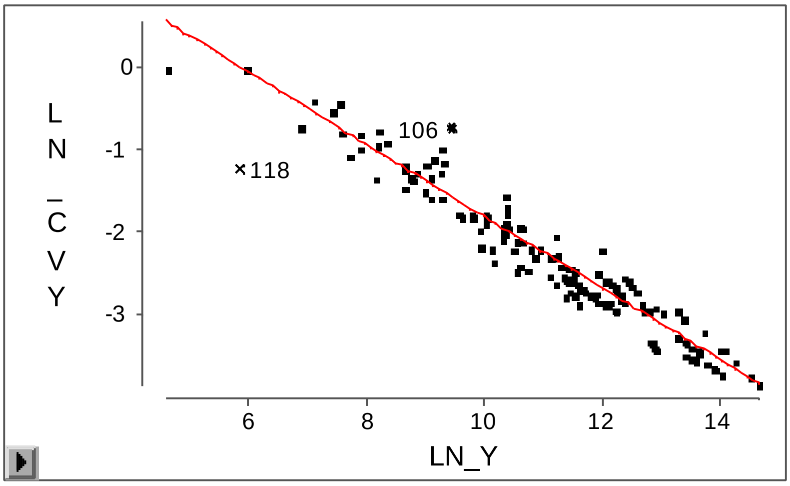
Figure 3:



Model of the natural logarithm of the coefficient of relative variation of $\hat{Y}$, the estimated population total unemployed in South Africa, according to the strict definition of unemployment, as predicted by the natural logarithm of $\hat{Y}$. (Source: OHS 1997 - Workers)

Model: $\ln(cv(\hat{Y})) = 2.588 - 0.4382\ln(\hat{Y})$

Figure 4:



Observations 106 and 118 have been excluded from the calculations.

Table 5:

Parametric Regression Fit								
			Model		Error			
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob > F
	1	1	132.1242	193	0.0421	0.9421	3141.3556	0.0001

Table 6:

Parameter Estimates							
Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	2.5880	0.0891	29.0534	0.0001	.	0
LN_Y	1	-0.4382	0.0078	-56.0478	0.0001	1.0000	1.0000

Figure 5:

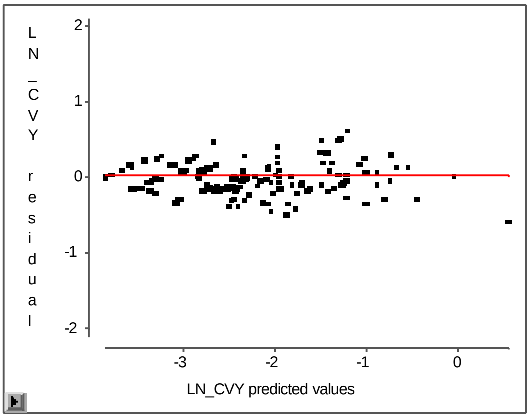
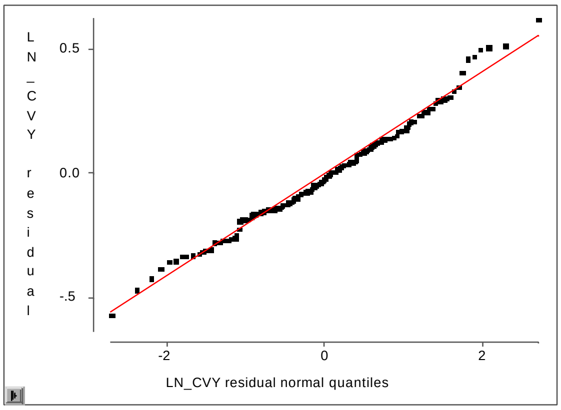


Figure 6:



3.5 The results of fitting the different models

Model proposed by Lepkovski and the United States is:

$$\text{Relvar}(\hat{Y}) = a + bY^{-1}$$

(refer to formula [1] on page 9)

The results obtained when this model was fitted to the previously mentioned data sets were very disappointing. The model was tested on both the worker data set and the household data set of the OHSs considered and in most cases the fitted model resulted in a very unsatisfactory R^2-value of less than 0.5. It led to the conclusion that the GVF used by SESTAT and other US institutes is not suitable for the above data sets.

Model proposed by the Australian Bureau of Statistics:

$$ln(cv(\hat{Y}_c)) = a + b\,ln(\hat{Y}_c)$$

or

$$\ln(cv(\hat{R})) = a + b\ln(\hat{Y})$$

(refer to Model 2 on page 13)

Model 2 has proven to be the best model when fitted to the various data sets. Giving an R^2-value of not less than 0.9 in most cases, it is safe to accept that this model is suitable for the data sets considered in this study.

It was found that, although the ABS had derived Model 2 for estimates of "person counts", the model performed equally well when estimates of counting variables in general are considered, e.g. counted households in the Household data set of the OHS or counted incidents of crimes in the VOC. Note that $\hat{Y}_c$ in the derivation discussed on pages 10 and 11, is a counting variable: $y_{ci} = 1$ if the occurrence appears and $y_{cj} = 0$ if it does not appear. Consequently Model 2 gives satisfactory results in all the cases where a total or a ratio is estimated.

From the fitting results of Model 3, $ln(cv(\hat{Y}_c)) = a + b\,ln(\hat{Y}_c) + c\left(ln(\hat{Y}_c)\right)^2$, as measured by R^2, it seems that the contribution of the additional quadratic term in the model to a better fit is minimal. The quadratic term makes Model 3 also less user friendly compared to Model 2.

The conclusion reached is that Model 2 seems to fit the data equally well for cross-classes, mixed classes and segregated classes. This result makes it possible to find one suitable model for a study variable over all the domains of interest.

3.6 Illustration of results

A number of the survey estimates of the 1997 OHS Workers data set and Household data set were taken and are displayed in the following tables with their directly calculated standard errors. These standard errors are compared with the modeled standard errors to illustrate the functionality of the models.

Table 7: Illustration of results

Study variable: Number of Unemployed –1997	Estimated value $\hat{Y}$	Directly calculated Standard errors	Modeled Standard errors
Domain		$\ln(cv(\hat{Y})) = 2.588 - 0.4382\ln(\hat{Y})$ $\therefore se(\hat{Y}) = e^{(\ln(cv(\hat{Y})))} \times \hat{Y}$	
African	2088753	49835	47257
Coloured	209235	14749	12974
Indian / Asian	41944	6055	5260
White	77277	7824	7414
Western Cape	185061	13824	12110
Eastern Cape	303402	20422	15986
Northern Cape	47209	4161	5621
Free State	156583	9327	11025
KwaZulu / Natal	474734	24780	20558
North West	190619	11402	12313
Gauteng	670552	32244	24960
Mpumalanga	178189	10315	11855
Northern Province	210861	11186	13031

From the 1997 OHS Household file, the estimated values for **dwelling type = "formal house or brick structure on separate yard or stand" according to province** (Table 8) and **main water source = "piped (tap) water in dwelling" according to province** (Table 9), were considered. Model fitting details are given in Appendix A, pages A – 7 and A – 9.

Table 8: Illustration of results

Study variable: **Dwelling Type** = Formal house or brick structure on separate yard or stand	Estimated value $\hat{Y}$	Directly calculated Standard errors	Modeled Standard errors
Domain		$\ln(cv(\hat{Y})) = 2.4389 - 0.3955\ln(\hat{Y})$ $\therefore se(\hat{Y}) = e^{(\ln(cv(\hat{Y})))} \times \hat{Y}$	
Western Cape	633402	25698	36840
Eastern Cape	685917	26561	38657
Northern Cape	155996	4804	15792
Free State	391459	17636	27541
KwaZulu / Natal	912224	35148	45928
North West	553899	15417	33971
Gauteng	1321969	38177	57475
Mpumalanga	429799	13637	29141
Northern Province	733840	17946	40268

Table 9: Illustration of results

Study variable: **Main Water source** = Piped (tap) water, in dwelling	Estimated value $\hat{Y}$	Directly calculated Standard errors	Modeled Standard errors
Domain		$\ln(cv(\hat{Y})) = 2.2365 - 0.3889\ln(\hat{Y})$ $\therefore se(\hat{Y}) = e^{(\ln(cv(\hat{Y})))} \times \hat{Y}$	
Western Cape	776426	23638	41665
Eastern Cape	336955	26641	25155
Northern Cape	95346	7098	11727
Free State	247448	20943	20872
KwaZulu / Natal	640290	38002	37081
North West	188721	18232	17719
Gauteng	1243752	43696	55395
Mpumalanga	221891	19003	19541
Northern Province	114416	19275	13094

The results in Table 8 and Table 9 seem less evident of a good fit (refer to reason c) on page 15). Nevertheless, a R^2-value ≥ 0.85 was obtained.

4. Presentation Methods

There are numerous ways to present standard errors in a survey report. The main requirements for the successful presentation of standard errors in a report are: the method should be cost effective in the sense of taking up as few pages as possible in the publication, easy to apply for the statistician and simple enough for the users to understand. A short introduction to some of the methods adopted by other countries is given along with an example of each. A few of the advantages and disadvantages of each specific presentation method are also discussed.

4.1 A table with estimated parameter values

The U.S. Bureau of the Census used the following method in the 1997 National Survey of College Graduates (*Sampling Errors For SESTAT and Its Component Surveys).*

Having fitted a suitable model to approximate standard errors, the resulting estimated model parameters are displayed in a parameter table in the publication. Each new study variable in the survey, with its own model parameters, becomes an entry in the table.

The following steps describe the procedure to determine the standard errors of an estimated total or percentage:

- Substitute the estimated total or percentage ($\hat{Y}$ or $\hat{R}$) into the standard error model that is provided;

- Find the table entry for the study variable of interest. If different models were fitted according to domains of interest, make sure to use the appropriate model parameters for the subclass on which the estimate is based;

- Substitute the parameter estimates into the model;

- Compute the approximate standard error.

The following example demonstrates the use of the parameter table for calculating the standard error of the estimate of the number of unemployed men in the Western Cape, in the 1997 OHS worker data set.

4.1.1 Example:

Table 10: Parameter Table for the worker data set and the household data set of the October Household Survey of 1997.

Study Variables According to different domains	Model Coefficients for $\hat{Y}$: $\ln(cv(\hat{Y})) = a + bLn(\hat{Y})$		Model Coefficients for $\hat{R}$: $\ln(cv(\hat{R})) = a + bLn(\hat{Y})$	
	Intercept a	Slope b	Intercept a	Slope b
Unemployed Strict definition	2.5880	-0.4382	2.7087	-0.4585
Unemployed Expanded def.	2.8358	-0.4623	2.6269	-0.4601
Dwelling Type	2.4389	-0.3955	2.7167	-0.4297
Water Source	2.2365	-0.3889	2.3443	-0.4067
Light Source	2.5152	-0.4202	2.772	-0.4642

Model: $\ln(cv(\hat{Y})) = a + b\ln(\hat{Y})$ and $\ln(cv(\hat{R})) = a + b\ln(\hat{Y})$

where $\hat{Y}$ denotes the estimated total and $\hat{R}$ denotes the estimated proportion.

The above models were fitted on the data for $\hat{Y}$ and $\hat{R}$ respectively.

To estimate the standard error for the **estimated total of unemployed males in the Western Cape,** proceed as follows:

Obtain the estimate of the total number of unemployed males in the Western Cape for 1997 according to the strict definition of unemployment:

$\hat{Y}$ = 81091[4]

From the above table the parameter values are:

$\hat{a}$ = 2.588 and $\hat{b}$ = -0.4382

Now we have the model:

$$\ln(cv(\hat{Y})) = 2.588 - 0.4382\ln(81091)$$
$$\ln(cv(\hat{Y})) = -2.3651$$

The standard error can be calculated with the following conversion:

[4] Preliminary results were used and may differ from the final published results

$$se(\hat{Y}) = e^{\left(\ln(cv(\hat{Y}))\right)} \times \hat{Y}$$
$$se(\hat{Y}) = 7618$$

The direct calculated standard error for this estimate which is based on the subclass: gender = male and province = Western Cape is 7394.

The standard error for $\hat{R}$ can be calculated in the same way.

Advantages:

- The method is fairly easy for the statistician to apply and is easy to understand.
- The possibility of including a separate pair of parameters for each new domain of interest may contribute to a higher level of accuracy in the modeling of standard errors.

Disadvantages:

- The more study variables and the more domain possibilities there are, the more parameter sets must be included in the table. This takes up space in the publication and can be time consuming. It also complicates the readability of the table.
- The method requires that the user is familiar with the substitution of the correct pair of parameters into the model and the calculation of the standard error with the formulas provided.

4.2 A table with the standard errors according to the size of the estimate

A table which consists of the standard errors according to size of the survey estimates and the confidence intervals of a specific level of significance, is published. These standard errors can be estimated with the suitable model or calculated directly if the size of the data set in terms of number of study variables allows it.

In the example the fitted model was used to estimate the published standard errors and the associated confidence intervals were calculated on a 95% level of significance.

From the table, the user is expected to find the estimate that is nearest in size to the estimate from the survey whose standard error is desired. Note that in the table it is the estimate which is just larger in size that should be chosen rather than the one just smaller than the estimate which the user is interested in. The conservative approach should be followed whenever standard errors are concerned. However, the chosen estimate must compare realistically with the survey estimate.

4.2.1 Example:

A table with standard errors for the worker data set of the 1997 OHS is constructed according to the strict definition of unemployment. The standard errors are calculated using the fitted model:

$$\ln\left(cv(\hat{Y})\right) = 2.588 - 0.4382\ln(\hat{Y})$$

and the conversion formula:

$$se(\hat{Y}) = \exp\left(\ln(cv(\hat{Y}))\right) \times \hat{Y}$$

The confidence intervals are then calculated with the following formula at a level of 95% significance:

$$CI = \hat{Y} \pm 1.96 se(\hat{Y})$$

A chosen range of typical survey estimates with their associated standard errors and confidence intervals are presented in the table.

If, for example, we want to obtain the standard error for the estimate of the number of **unemployed men in the Western Cape for 1997**, we first find the estimate nearest in value to $\hat{Y} = 81091$ and use that value in the table (Table 11). Following the conservative approach, the standard error of 100 000 is used, which is 8569. This value is larger than the direct calculated standard error, 7394, but is still acceptable.

Table 11: Table with the standard errors and confidence intervals for the worker data set of the OHS of 1997, according to the official strict definition of unemployment.

Size of Estimate	Standard Error	Lower Confidence Interval	Upper Confidence Interval
1500	645	236	2764
3000	1195	658	5342
5000	1592	1879	8121
10000	2350	5393	14607
30000	4357	21460	38540
50000	5805	38621	61379
70000	7013	56254	83746
100000	8569	83204	116796
300000	15885	268864	331136

500000	21166	458515	541485
700000	25570	649883	750117
1000000	31243	938764	1061236
1300000	36205	1229038	1370962
1500000	39236	1423098	1576902
1700000	42094	1617496	1782504
2000000	46118	1909608	2090392
2300000	49885	2202225	2397775

Advantages:

- This method of presentation makes it very easy for the user to find the standard error, because no calculation is needed.
- The confidence intervals are immediately available to the user.

Disadvantages:

- Often, when the estimate of interest does not match closely with an estimate from the table, it is necessary for the user to use interpolation to find an acceptable estimate of the standard error.
- This presentation method requires that for each new study variable in the survey, a new table must be set up. This can be very time consuming and also take up much space in the publication. It is therefore recommended that this method be used where the survey consists of a limited number of study variables. A good example is the VOC survey where there are only two main study variables, viz. household crimes that are committed against people living together and individual crimes that affect only a single person.
- The table provides estimates of only the same order in size. This may lead to a loss in accuracy in the prediction of the standard error.

4.3 A table with coefficients of relative variation and factor-lines

This presentation method is used by the ABS and is discussed in their Technical Note on Sampling Variability, Appendix D, HES Summary of Results, 1993-1994.

The table consists of the coefficients of relative variation of each study variable at the highest domain level in the survey, e.g. RSA-level, and the necessary factor-lines to be used at lower domain levels, e.g. province, race or gender level (refer to Table 1). The factor-lines are graphically displayed and are used to obtain the necessary adjustment factor with which the given relative standard error should be multiplied to adjust for the smaller sample size of the subclass on which the estimate is based. The coefficients of relative variation are estimated using the fitted model or are calculated directly.

The adjustment factors are calculated by dividing the estimate at a lower domain of interest level through the same estimate at RSA-level and then raised to a power found in the standard error model, e.g.

$$f_a = \left(\frac{\hat{Y}_{\mathrm{Prov}}}{\hat{Y}_{RSA}} \right)^{Power} \qquad [16]$$

where f_a denotes the adjustment factor, $\hat{Y}_{RSA}$ the estimate at RSA-level and $\hat{Y}_{\mathrm{Prov}}$ the estimate at a lower level, e.g. province-level.

This procedure can be justified mathematically as follows: To estimate the natural logarithm of the coefficient of relative variation at RSA-level, the formula is:

$$\ln(cv(\hat{Y}_{RSA})) = a + b\ln(\hat{Y}_{RSA})$$

and at a lower level, e.g. at province-level:

$$\begin{aligned}
\ln(cv(\hat{Y}_{\mathrm{Prov}})) &= a + b\ln\Big(\frac{\hat{Y}_{\mathrm{Prov}}}{\hat{Y}_{RSA}} \times \hat{Y}_{RSA}\Big) \\
\therefore cv(\hat{Y}_{\mathrm{Prov}}) &= e^{a + b\ln(\hat{Y}_{RSA}) + b\ln(\frac{\hat{Y}_{\mathrm{Prov}}}{\hat{Y}_{RSA}})} \\
&= e^{a} \times e^{b\ln(\hat{Y}_{RSA})} \times e^{b\ln(\frac{\hat{Y}_{\mathrm{Prov}}}{\hat{Y}_{RSA}})} \\
&= e^{a} \times (\hat{Y}_{RSA})^{b} \times \left(\frac{\hat{Y}_{\mathrm{Prov}}}{\hat{Y}_{RSA}} \right)^{b} \\
&= cv(\hat{Y}_{RSA}) \times f_a
\end{aligned} \qquad [17]$$

The following steps must be followed to find the estimated coefficient of relative variation of interest:

- Obtain the estimated value of the study variable from the published table.
- Obtain the estimated coefficient of relative variation for this study variable at RSA-level and its factor-line from the table (Table 12 in the case of the OHS of 1997).

- Read off the adjustment factor for the estimate of interest and the specific factor-line from the factor-line graph which is provided (Figure 7 in the case of the OHS of 1997).

- The estimated coefficient of relative variation for the estimate at a lower level is calculated as: $cv_{lowerlevel} = f_a \times cv_{RSA}$

4.3.1 Example:

To compare the standard error given by this presentation method with the standard errors of the previous methods, again the example of the estimated number of **unemployed males in the Western Cape** from the worker data set of the OHS of 1997 is used. The estimate of interest is: $\hat{Y}_{WC \times M} = 81091$. From Table 12 we obtain the coefficient of relative variation for the study variable at RSA-level: the number of unemployed people in the RSA:

$$cv(\hat{Y}_{RSA}) = 0.0212$$

The factor-line to use is: **I** and from Figure 7 on page 32 we find the factor for $\hat{Y}_{WC \times M} = 81091$, is: $f_a = 4.4$

$$\therefore cv(\hat{Y}_{WC \times M}) = 4.4 \times 0.0212$$

$$cv(\hat{Y}_{WC \times M}) = 0.0933$$

To calculate the standard error, the coefficient of relative variation must be multiplied with the estimate:

$$se(\hat{Y}_{WC \times M}) = 0.0933 \times 81091$$

$$se(\hat{Y}_{WC \times M}) = 7565$$

This value compares well with the direct calculated standard error, 7394, for the same estimate. In the same way the standard error for $\hat{R}$ can be obtained.

Table 12: Table with coefficients of relative variation at RSA level for the worker data set and the household data set of the October Household Survey of 1997, and factor-lines to derive the relative standard errors at lower levels of the domains of interest.

Study Variable from survey	**Coefficient of Relative Variation of $\hat{Y}$ = estimated number**	**Coefficient of Relative Variation of $\hat{R}$ = estimated ratio**	**Factor-lines** At lower levels, e.g. province-, race-, gender-level, etc.
OHS 1997 – Worker data set			
Unemployed in RSA	0.0212	0.0178	I
OHS 1997 – Household data set			
Dwelling Types			

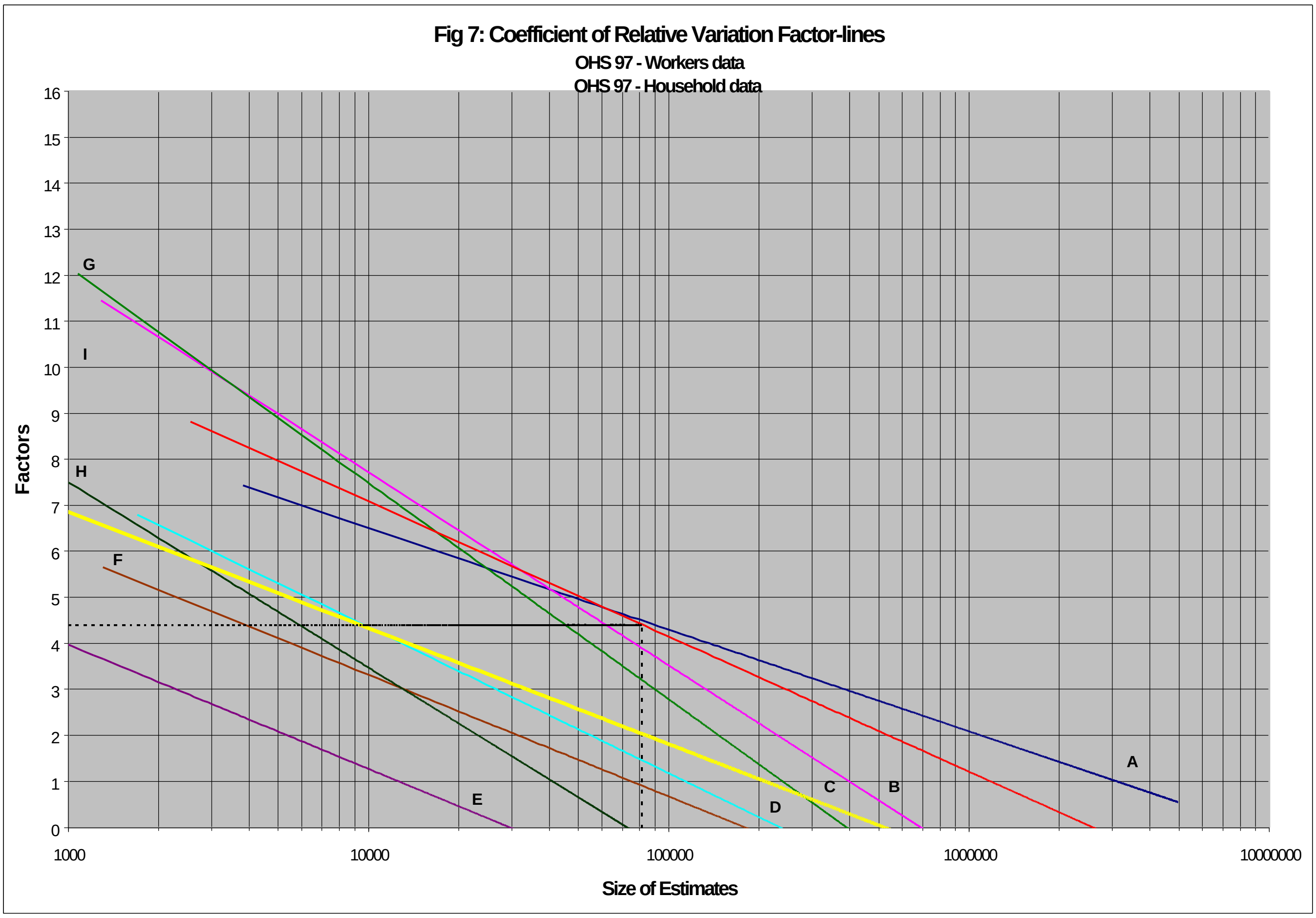
Households with a formal house or brick structure on a separate stand or yard in RSA	0.0242	0.0188	A
Households with traditional dwelling, hut, structure, made of traditional materials	0.0444	0.0363	B
Households living in flats, apartment in block of flats	0.0676	0.0573	C
Town-, cluster-, semi-detached house (simplex, duplex, or triplex)	0.0808	0.0695	D
Households with an informal dwelling, shack, in the back yard	0.2137	0.2	E
Households with an informal dwelling, shack, NOT in the back yard, e.g. in an informal squatter settlement	0.0988	0.0865	F
Room in hostel, compound for workers provided by employer or municipality	0.0923	0.0803	G
Main source of Water	$cv(\hat{Y})$	$cv(\hat{R})$	Factor-line
Piped (tap) water, in dwelling	0.0257	0.0218	A
Piped (tap) water, on site or in yard	0.0327	0.0281	B
Public tap	0.0357	0.0308	B
Water-Carrier, tanker	0.111	0.101	E
Borehole on site	0.1079	0.098	H
Borehole: off site, communal	0.0697	0.0623	F
Rain-water tank on site	0.1845	0.1717	E
Flowing water, stream	0.0516	0.0453	B
Dam, pool, stagnant water	0.0907	0.0817	H
Well	0.1126	0.1024	D
Spring	0.0846	0.076	H

Advantage:

- This presentation method is fairly easy for the user to apply.

Disadvantages:

- Many study variables from the survey require many table entries.
- To set up the table requires much work and time.
- The method requires the user to be familiar with the use of graphs and to do some simple calculations to obtain the estimated value of the standard error.



4.4 Formulas and Graphs

The model for the coefficient of relative variation for each study variable from the survey is published and can also be graphically presented. The user only needs to insert the value of the estimate of interest into the model or has the option to read off the coefficient of relative variation from the published graph. The necessary conversion formula to calculate the standard error from the coefficient of relative variation must also be given with an example that explains to the user how the formulas and the graphs should be used.

The formulas to calculate confidence intervals can also be included and explained to the user as indicated below. This comment is also applicable to the previous presentation methods.

4.4.1 Example

Returning to the example that has already been used, the estimated number of **unemployed men in the Western Cape** from the 1997 OHS worker data set is 81091.

The model to use is:

$$\begin{aligned}\ln\left(cv(\hat{Y})\right) &= 2.588 - 0.4382\ln(\hat{Y}) \\ &= 2.588 - 0.4382\ln(81091) \\ &= -2.3651\end{aligned}$$

To convert this value into the standard error, the following formula is used

$$\begin{aligned}se(\hat{Y}) &= \exp\left(\ln(cv(\hat{Y}))\right)\times\hat{Y} \\ &= 0.0939\times 81091 \\ &= 7614\end{aligned}$$

Alternatively, the formula $se(\hat{Y}) = 13.303\times\hat{Y}^{0.5618}$ as derived on page 18 can be used.

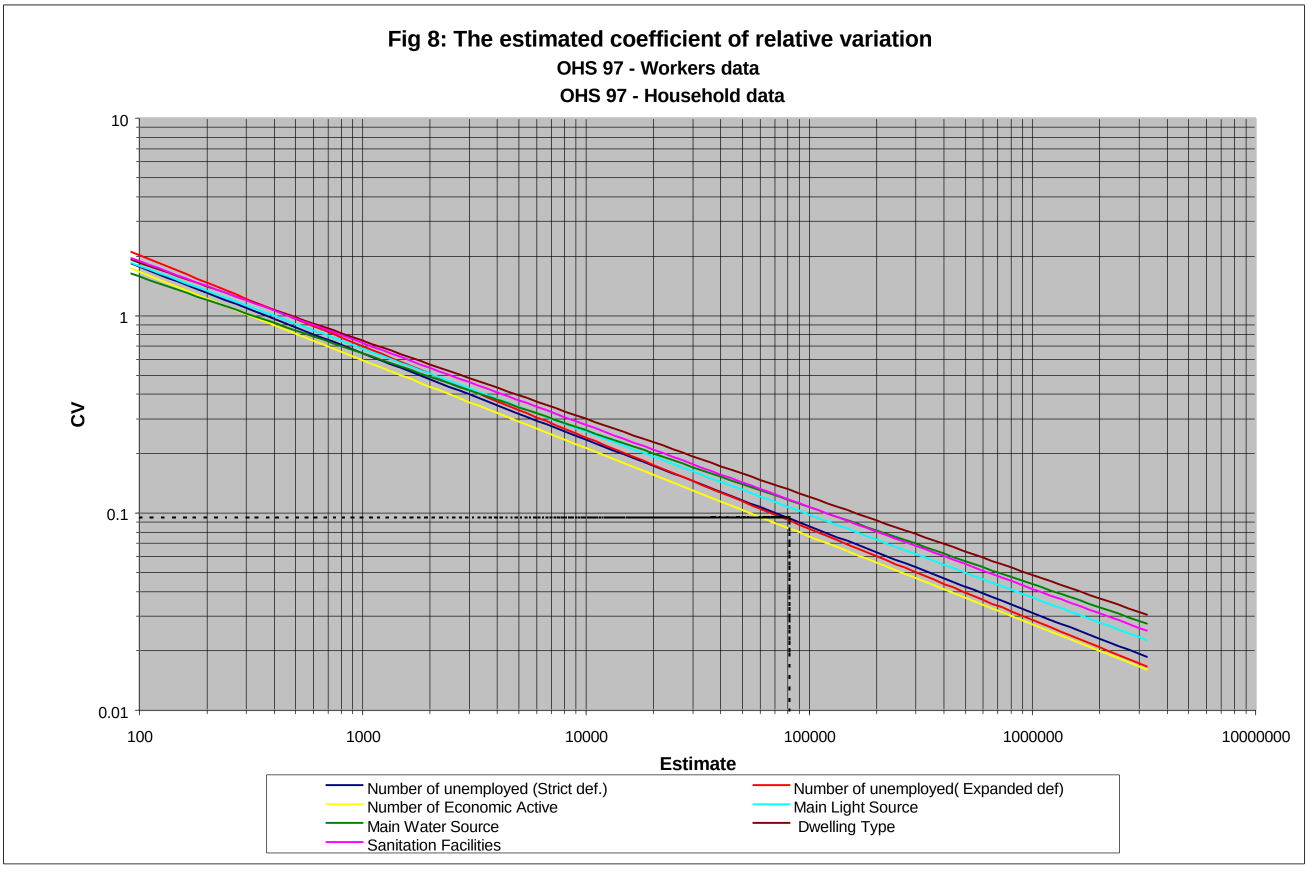
If the graph is used, we find the coefficient of relative variation for $\hat{Y} = 81091$ is: $cv(\hat{Y}) = 0.095$ (see the dotted line on Figure 8,page 35)

To calculate the standard error: $se(\hat{Y}) = cv(\hat{Y})\times\hat{Y}$

$$\begin{aligned}&= 0.095\times 81091 \\ &= 7704\end{aligned}$$

To calculate the 95% confidence interval for this estimate:

$$\begin{aligned}CI &= \hat{Y}\pm 1.96se(\hat{Y}) \\ &= 81091\pm 1.96\times 7704 \\ &= [65991;96191]\end{aligned}$$



If the standard error and confidence interval for $\hat{R}$ are required, $\hat{Y}$ must be replaced with $\hat{R}$ in the above formulas where applicable.

Advantages:

- The formulas presented are very easy to use for the statistician and will result in getting a better estimated value for the standard error.
- The method in graphical form is very easy to understand and to be used by the general user.
- This method is also very space efficient

Disadvantage:

- The method requires users to be familiar with the use of graphs and / or formulas.

4.5 Nomogram

A nomogram is a graphical presentation for mathematical functions consisting of more than one independent variable. The model, $\ln(cv(\hat{Y})) = a + b\ln(\hat{Y})$, is a simple straight line with only one independent variable. It will serve no purpose to construct a nomogram for this model.

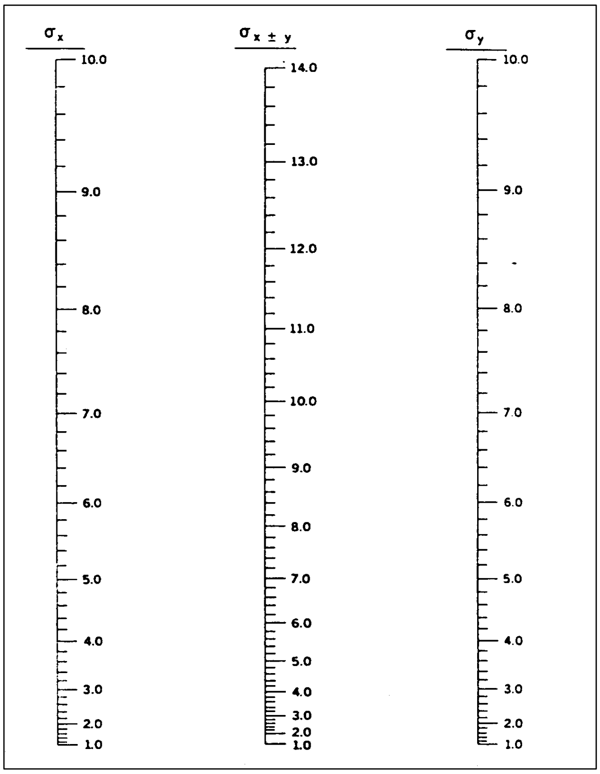
However, a nomogram can be an extremely valuable tool to facilitate calculations. For example, it may be necessary to test whether estimates of the unemployment rate obtained in independent cross-sectional surveys such as the OHS of 1995 and the OHS of 1996, differ significantly. This will require an estimate of the standard error of the difference between the estimated values.

Instructions to use the nomogram: Let $\hat{X}$ and $\hat{Y}$ be two independent estimates of X and Y respectively. $\hat{X} + \hat{Y}$ is an estimate of the sum and $\hat{X} - \hat{Y}$ is an estimate of the difference of $\hat{X}$ and $\hat{Y}$. The nomogram can be used to approximate the standard errors of $\hat{X} + \hat{Y}$ and $\hat{X} - \hat{Y}$ by following the steps:

- Find the point on the s_x- scale that corresponds to the estimated standard error of $\hat{X}$ and the point on the s_y- scale that corresponds to the estimated standard error of $\hat{Y}$
- The scales may be read in any unit (tenths, thousands, millions) as long as the same unit is used on all the scales
- Connect these points on the s_x- scale and the s_y- scale by a straight line. The value where the line crosses the $s_{x \pm y}$- scale is the estimated standard error of $\hat{X} + \hat{Y}$ and $\hat{X} - \hat{Y}$.

If for example $se(\hat{X}) = 6.75$ and $se(\hat{Y}) = 4.7$, a straight line connecting these points, crosses the $s_{x \pm y}$- scale at about 8.25 while an exact computation gives 8.225 (Gonzalez, Ogus, Shapiro and Tepping; March1974).

Figure 9: Nomogram - Standard error of sum or difference



To test whether an observed difference between the unemployment rates, $\hat{R}_1$ and $\hat{R}_2$, obtained in two different OHSs, is statistically significant, the 95% confidence interval for the difference must be calculated, viz.:

$$\left((\hat{R}_1 - \hat{R}_2) - 1.96se(\hat{R}_1 - \hat{R}_2) \; ; \; (\hat{R}_1 - \hat{R}_2) + 1.96se(\hat{R}_1 - \hat{R}_2)\right)$$

[18]

If this interval does not include the value 0, then the estimated unemployment rates $\hat{R}_1$ and $\hat{R}_2$, differ significantly at the 5% level of significance (using two-sided testing).

Nomograms require more effort to set up, but the user of the report will find them very easy to use.

5. Concluding remarks

This research project addressed a very common problem experienced by survey statisticians: How to estimate and present standard errors in a survey report without taking up too much time and too much space in the publication.

The results of the project suggest that it is feasible to estimate standard errors indirectly with the use of mathematical models. Also, there are many statistical packages available in the market, which are very effective for modeling purposes. The combined result is a large reduction in time spent on the calculation of standard errors.

Several practical and effective methods of the presentation of standard errors in the publication are available. These methods can contribute even more to the saving of time and costs in the publication of standard errors. Most importantly, these methods are easy to understand and to apply by the user of the publication.

The research positively suggests a solution to the above problem.

6. References

Bieler, G. S., and Williams, R. L., 1990. '**Generalized Standard Error Models for Proportions in Complex Design Surveys**' in *Proceedings of Section on Survey Research Methods of the American Statistical Association*, 272-277

Cohcran, 1977. '**Sampling Techniques**', John Wiley, New York.

Cox, B. G., Jang, D., Edson, D., 1997. '**Sampling Errors for SESTAT and its Component Surveys: 1993**', Mathematica Policy Research, Inc.

Finamore, J. M., 1999. '**Generalized Variance Parameters for the 1997 National Survey of College Graduates**', U. S. Bureau of the Census, Demographic Statistical Methods Division, Health Surveys and Supplements Branch

Gonzalez, M. E., Ogus, J. L., Shapiro, O. G., and Tepping, B. J., 1975. '**Standards for Discussion and Presentation of Errors in Survey and Census Data**' in *Journal of the American Statistical Association,* **70**(351), Part II, 5-23

Johnson, E. G., and King, B. F., 1987. '**Generalized Variance Functions for a Complex Sample Survey**' in *Journal of Official Statistics*, **3**, 235-250

Lepkovski, 1998. '**Presentation of Sampling Errors**', Methods *of Survey Sampling / Applied Sampling* – Lecture at Statistics South Africa

The Australian Bureau of Statistics, 1993-1994. '**Technical Note on Sampling Variability**' in *ABS – HES Summary of Results*, Appendix D, 43-49

The Australian Bureau of Statistics, 1997. '**Household Collection Support Standard Error Manual**', Section 3, 29-41

Valliant, R., 1987. '**Generalized Variance Functions in Stratified Two-Stage Sampling**' in *Journal of the American Statistical Association*, **82**, 499-508

7. Appendix A

8. Appendix B

OHS: October Househols Survey
N: Population number in subclass
MSWX: Estimated number of economic active
CV-R: Estimated coefficient of relative variation of R

STR: Official strict definition of unemployment
n: Sample size of subclass
SE-R: estimated standard error of R
CV-WY: Estimated coefficient of relative variation of MSWY

U / R: Urban / Rural
R: Estimated ratio
SE-WY: Estimated standard error of MSWY

Type: Urban formal, Urban informal, Tribal, Commercial farms, Other non-urban
MSWY: Estmated number of unemploued
SE-WX: Estimated standard error of MSWX
CV-WY: Estimated coefficient of relative variation of MSWX

Table: OHS 97 Workers data set (Official strict definition of unemployment). Results obtained from the SAS programs.

OHS	STR	PROV	U/R	TYPE	RACE	GENDER	N	n	R	MSWY	MSWX	SE-R	SE-WY	SE-WX	CV-R	CV-WY	CV-WX
97	1	0	0	0	0	0	33105	7504	0.217176587	2417209	11130153	0.004373042	52227.601	98341.3506	0.020135882	0.021606576	0.008835579
97	1	0	1	0	0	0	22261	4698	0.2004055	1624948	8108299	0.005363801	46333.719	81963.68127	0.026764741	0.028513974	0.010108616
97	1	0	2	0	0	0	10844	2806	0.262177109	792261	3021854	0.006883268	23442.181	44591.74787	0.026254267	0.029588968	0.014756422
97	1	0	0	0	0	1	17721	3282	0.18142816	1147325	6323853	0.004490088	29598.395	61034.42713	0.02474857	0.025797741	0.009651462
97	1	0	0	0	0	2	15384	4222	0.264212313	1269884	4806300	0.005800762	31048.098	52712.19143	0.021954926	0.024449562	0.010967312
97	1	0	1	0	0	1	11873	2037	0.166883809	765376	4586282	0.005405693	26010.266	51596.42707	0.032391956	0.033983632	0.011250163
97	1	0	1	0	0	2	10388	2661	0.244056609	859572	3522017	0.007135486	27490.653	44810.80432	0.029237013	0.03198181	0.012723052
97	1	0	2	0	0	1	5848	1245	0.219817677	381949	1737571	0.007596377	13956.204	28252.21556	0.034557626	0.036539466	0.016259606
97	1	0	2	0	0	2	4996	1561	0.319487308	410312	1284283	0.009076794	13958.714	22857.49295	0.028410501	0.034019749	0.017797864
97	1	0	0	0	1	0	22606	6408	0.281014557	2088753	7432900	0.005055938	49835.238	93550.19262	0.017991728	0.023858844	0.012585961
97	1	0	0	0	2	0	5755	831	0.152525126	209235	1371804	0.008269506	14749.158	42250.20162	0.054217338	0.070491023	0.030799011
97	1	0	0	0	3	0	1161	115	0.098898299	41944	424112	0.012920395	6055.3929	24256.1123	0.130643248	0.144368707	0.05719272
97	1	0	0	0	4	0	3583	150	0.04064346	77277	1901337	0.003999025	7823.8743	53841.7158	0.098392837	0.101244663	0.028317822
97	1	0	1	0	1	0	13307	3700	0.277989056	1317282	4738611	0.006559193	43723.178	80832.52602	0.023595148	0.033191965	0.017058274
97	1	0	2	0	1	0	9299	2708	0.28633569	771471	2694289	0.007444732	23067.985	42741.52595	0.026000014	0.02990129	0.015863747
97	1	0	1	0	2	0	4475	748	0.169691452	194602	1146798	0.009221961	14393.218	40963.1595	0.054345465	0.073962409	0.035719598
97	1	0	2	0	2	0	1280	83	0.065032741	14633	225006	0.011140544	2540.4531	8880.726555	0.1713067	0.173614156	0.039468854
97	1	0	1	0	3	0	1142	113	0.098817545	41486	419826	0.013053026	6057.263	24162.34899	0.132092195	0.146006786	0.057553245
97	1	0	2	0	3	0	19	2	0.106808567	458	4286	0	0	0	0	0	0
97	1	0	1	0	4	0	3337	137	0.03969786	71578	1803064	0.003922593	7194.5554	50736.54984	0.098811186	0.100513784	0.028139067
97	1	0	2	0	4	0	246	13	0.057992967	5699	98272	0.028783916	3225.7497	10306.59308	0.496334601	0.56600967	0.104877818
97	1	0	0	0	1	1	11777	2784	0.23763836	988873	4161252	0.00559734	27543.811	53578.43986	0.023554024	0.02785374	0.012875559
97	1	0	0	0	1	2	10829	3624	0.33618529	1099880	3271649	0.006734204	29266.346	47468.47969	0.020031227	0.026608666	0.014509038
97	1	0	0	0	2	1	3140	372	0.128479398	100382	781305	0.009379747	8665.357	24412.52397	0.073005848	0.08632419	0.031245842
97	1	0	0	0	2	2	2615	459	0.184340655	108853	590499	0.010969143	8109.7938	18693.07568	0.059504739	0.07450226	0.0316564
97	1	0	0	0	3	1	739	64	0.086660971	23627	272639	0.013749685	3959.6181	15675.29754	0.158660636	0.167587329	0.057494622
97	1	0	0	0	3	2	422	51	0.120924588	18317	151473	0.01828524	3069.2498	9200.851404	0.151211925	0.16756518	0.060742712
97	1	0	0	0	4	1	2065	62	0.031067566	34443	1108657	0.004391151	4971.9564	31566.53086	0.141341981	0.14435196	0.028472753
97	1	0	0	0	4	2	1518	88	0.054036501	42834	792679	0.006564561	5252.0462	23935.8009	0.121483818	0.122615056	0.030196072
97	1	0	1	0	1	1	6873	1578	0.233818675	615730	2633364	0.007195452	23703.661	45216.43301	0.030773642	0.038496859	0.017170595
97	1	0	1	0	1	2	6434	2122	0.333239922	701552	2105247	0.008927784	25557.369	40567.70966	0.02679086	0.036429744	0.019269813
97	1	0	2	0	1	1	4904	1206	0.244221713	373143	1527887	0.008447066	13784.123	26701.04725	0.034587693	0.036940568	0.017475797
97	1	0	2	0	1	2	4395	1502	0.341501402	398328	1166402	0.009520214	13648.631	21581.98574	0.027877524	0.034264805	0.01850304
97	1	0	1	0	2	1	2385	339	0.145759442	93998	644884	0.010668491	8404.1436	23464.19755	0.073192454	0.089407795	0.036385163
97	1	0	1	0	2	2	2090	409	0.200440456	100604	501914	0.011996717	7808.6047	18032.36066	0.059851775	0.077617306	0.035927177
97	1	0	2	0	2	1	755	33	0.046793924	6384	136421	0.012315674	1698.5148	5516.018039	0.263189593	0.266071564	0.040433773
97	1	0	2	0	2	2	525	50	0.093120587	8249	88585	0.020970961	1891.4117	4693.868597	0.225202196	0.229287695	0.052987241
97	1	0	1	0	3	1	725	62	0.086016069	23169	269362	0.013915128	3953.6447	15585.9085	0.161773583	0.170640619	0.057862406
97	1	0	1	0	3	2	417	51	0.121734758	18317	150464	0.018351645	3065.8921	9195.46308	0.150751066	0.16738187	0.061113865
97	1	0	2	0	3	1	14	2	0.139657331	458	3278	0	0	0	0	0	0
97	1	0	2	0	3	2	5	0	0	0	1008	0	0	0	.	.	0
97	1	0	1	0	4	1	1890	58	0.031269917	32479	1038673	0.004559308	4813.4594	29938.13098	0.14580494	0.148201225	0.028823446
97	1	0	1	0	4	2	1447	79	0.051149944	39099	764392	0.006335423	4858.3345	22942.59409	0.123859822	0.124258579	0.030014192
97	1	0	2	0	4	1	175	4	0.028064383	1964	69985	0.016473633	1298.6687	7302.633134	0.586994295	0.661211248	0.104346252
97	1	0	2	0	4	2	71	9	0.132037034	3735	28288	0.05538339	1773.8595	1848.366028	0.419453458	0.474924863	0.06534155
97	1	1	0	0	0	0	5335	606	0.118159744	185061	1566189	0.008358207	13823.97	30958.75891	0.070736503	0.074699715	0.019766931
97	1	2	0	0	0	0	2819	875	0.290738303	303402	1043555	0.018226476	20422.41	35733.27628	0.062690315	0.067311495	0.03424186
97	1	3	0	0	0	0	1724	336	0.1854019	47209	254629	0.017273045	4160.9167	8820.988505	0.093165413	0.088138597	0.034642446
97	1	4	0	0	0	0	2821	624	0.203465737	156583	769578	0.012917579	9326.9869	23569.73768	0.063487735	0.059565863	0.030626834
97	1	5	0	0	0	0	5462	1361	0.228219463	474734	2080165	0.01113769	24779.863	48127.83597	0.048802543	0.052197342	0.023136545

OHS	STR	PROV	U/R	TYPE	RACE	GENDER	N	n	R	MSWY	MSWX	SE-R	SE-WY	SE-WX	CV-R	CV-WY	CV-WX
97	1	6	0	0	0	0	2798	686	0.240685374	190619	791984	0.01281856	11402.443	21981.18526	0.053258573	0.059818013	0.027754594
97	1	7	0	0	0	0	6736	1553	0.216847661	670552	3092273	0.009544372	32243.916	50205.74779	0.044014182	0.048085622	0.016235873
97	1	8	0	0	0	0	2961	801	0.244494269	178189	728807	0.013375113	10314.641	23515.06747	0.05470522	0.057885884	0.032265133
97	1	9	0	0	0	0	2449	662	0.262600315	210861	802972	0.013835001	11185.715	30484.92199	0.052684632	0.053047914	0.037965129
97	1	1	1	0	0	0	4354	570	0.12934128	178564	1380563	0.009288908	13598.019	30021.37324	0.071817044	0.076152148	0.021745742
97	1	1	2	0	0	0	981	36	0.034998948	6497	185626	0.009279342	1737.0754	5902.536693	0.265132035	0.267377296	0.031797981
97	1	2	1	0	0	0	1646	430	0.244196718	162474	665339	0.025261837	18198.402	26799.90877	0.103448717	0.112008331	0.040280069
97	1	2	2	0	0	0	1173	445	0.37261196	140928	378216	0.022951586	9131.9492	18960.59107	0.061596483	0.064798747	0.050131625
97	1	3	1	0	0	0	1255	293	0.233375875	41394	177372	0.021361072	3771.8218	6552.512963	0.091530763	0.091119417	0.036942274
97	1	3	2	0	0	0	469	43	0.075261279	5815	77258	0.017893163	1365.8859	4892.189831	0.237747258	0.234909393	0.063322907
97	1	4	1	0	0	0	2122	503	0.216331409	126035	582600	0.015299008	8498.9463	20544.51589	0.070720233	0.067433353	0.035263479
97	1	4	2	0	0	0	699	121	0.163377812	30548	186978	0.021572033	3863.9239	10277.34183	0.132037717	0.126486982	0.054965622
97	1	5	1	0	0	0	3203	655	0.190487101	260792	1369081	0.014165656	19903.562	38119.09345	0.074365432	0.076319575	0.027842823
97	1	5	2	0	0	0	2259	706	0.300867277	213942	711084	0.01670471	14331.115	23201.27639	0.055521856	0.066986038	0.032628051
97	1	6	1	0	0	0	1156	261	0.220529746	76852	348489	0.019665104	7735.2814	14694.00261	0.089172117	0.10065151	0.042164938
97	1	6	2	0	0	0	1642	425	0.256523225	113767	443495	0.016461676	8396.54	15964.78471	0.064172265	0.073804855	0.035997665
97	1	7	1	0	0	0	6526	1535	0.221653691	660986	2982066	0.009754929	31930.869	48502.78844	0.044009773	0.048307943	0.016264828
97	1	7	2	0	0	0	210	18	0.086802317	9566	110207	0.028029929	3231.759	11678.15742	0.322916826	0.337830398	0.105965723
97	1	8	1	0	0	0	1421	359	0.219114968	86357	394115	0.017909888	7094.0395	15357.16493	0.081737402	0.082148225	0.038966166
97	1	8	2	0	0	0	1540	442	0.274379575	91833	334692	0.019390341	7260.621	11988.43406	0.07066977	0.07906361	0.035819301
97	1	9	1	0	0	0	578	92	0.150926385	31494	208673	0.020437052	4286.5672	19395.16764	0.135410729	0.136105969	0.092945045
97	1	9	2	0	0	0	1871	570	0.301811929	179366	594298	0.015556624	10357.931	19779.97117	0.051544101	0.057747375	0.033282909
97	1	1	0	0	0	1	2977	246	0.089061403	81091	910511	0.007924057	7394.0026	19997.90854	0.088972961	0.091181111	0.021963391
97	1	1	0	0	0	2	2358	360	0.158567292	103969	655678	0.012588807	8897.9956	18223.38414	0.079390947	0.085583031	0.027793172
97	1	2	0	0	0	1	1446	418	0.266409485	149448	560972	0.017837825	9801.5539	20053.09801	0.066956418	0.065584892	0.035747036
97	1	2	0	0	0	2	1373	457	0.31901902	153953	482583	0.023393363	13058.131	19305.49071	0.073329054	0.08481885	0.040004495
97	1	3	0	0	0	1	965	146	0.140432216	21070	150037	0.017242567	2444.0981	6684.319648	0.122782131	0.115999103	0.044551256
97	1	3	0	0	0	2	759	190	0.249910106	26139	104593	0.0240455	2545.5043	4525.019331	0.096216596	0.097384062	0.043263167
97	1	4	0	0	0	1	1469	256	0.158674308	68829	433776	0.013646896	5608.9718	15257.17397	0.086005709	0.081491288	0.035172937
97	1	4	0	0	0	2	1352	368	0.261325552	87754	335802	0.016545078	5601.485	11603.16989	0.063312134	0.063831927	0.034553601
97	1	5	0	0	0	1	2893	629	0.199970453	231908	1159713	0.011783989	14123.465	28539.61654	0.058928648	0.060901076	0.02460921
97	1	5	0	0	0	2	2569	732	0.263811461	242826	920452	0.013999545	14598.151	26221.2353	0.053066478	0.06011777	0.028487335
97	1	6	0	0	0	1	1512	309	0.202661245	93449	461110	0.014368867	7057.2923	14676.07914	0.070900911	0.075520107	0.031827699
97	1	6	0	0	0	2	1286	377	0.293676388	97170	330873	0.016114338	6497.2906	10758.62615	0.054871072	0.066865395	0.03251584
97	1	7	0	0	0	1	3648	685	0.183214053	327414	1787055	0.009663869	18717.86	32073.12838	0.052746332	0.057168842	0.017947473
97	1	7	0	0	0	2	3088	868	0.262897553	343138	1305217	0.012777804	18312.043	27834.80614	0.048603739	0.05336634	0.021325801
97	1	8	0	0	0	1	1612	306	0.17222659	74944	435150	0.013006747	5572.1675	15649.23004	0.075521132	0.074350703	0.03596285
97	1	8	0	0	0	2	1349	495	0.351582538	103245	293657	0.018658533	6646.6451	11000.33368	0.053070134	0.064377499	0.037459741
97	1	9	0	0	0	1	1199	287	0.233052807	99171	425529	0.01777784	7563.3379	18499.0742	0.076282452	0.07626585	0.043473133
97	1	9	0	0	0	2	1250	375	0.29591217	111690	377443	0.016348578	6450.5398	14818.47098	0.055248076	0.057754009	0.039260181
97	1	1	0	0	1	0	959	220	0.225676101	75771	335752	0.024651818	9930.4828	20369.08247	0.109235397	0.131058641	0.060666974
97	1	1	0	0	2	0	3621	363	0.111867795	98266	878413	0.007860167	8820.1702	31230.10763	0.070263	0.089758028	0.035552889
97	1	1	0	0	3	0	65	5	0.092238429	2010	21790	0.046911711	927.78908	6328.640994	0.50859183	0.461622287	0.290442301
97	1	1	0	0	4	0	690	18	0.027293668	9013	330235	0.007227353	2440.9357	23234.55839	0.264799611	0.270814383	0.070357714
97	1	2	0	0	1	0	2118	757	0.351420935	268559	764208	0.019869647	19582.556	31566.07491	0.056540875	0.072917237	0.041305611
97	1	2	0	0	2	0	498	115	0.227178477	33196	146125	0.034697966	6206.1926	13409.86287	0.152734388	0.186953533	0.091769835
97	1	2	0	0	3	0	18	1	0.05848942	399	6817	0.056588887	398.72165	2733.632947	0.967506386	1	0.401003065
97	1	2	0	0	4	0	185	2	0.009870534	1248	126406	0.006567453	849.83838	14919.60244	0.665359434	0.681129476	0.118029685
97	1	3	0	0	1	0	474	114	0.238387907	20500	85993	0.031421092	2938.6164	7051.66363	0.131806569	0.14334872	0.082002432
97	1	3	0	0	2	0	1041	215	0.202030418	25196	124716	0.019038248	2841.8869	7332.822359	0.094234563	0.112789605	0.058796328
97	1	3	0	0	3	0	2	0	0	0	204	0	0	0	.	.	0
97	1	3	0	0	4	0	207	7	0.034601889	1513	43716	0.015114224	672.05549	5643.023523	0.436803441	0.444286234	0.129083101

OHS	STR	PROV	U/R	TYPE	RACE	GENDER	N	n	R	MSWY	MSWX	SE-R	SE-WY	SE-WX	CV-R	CV-WY	CV-WX
97	1	4	0	0	1	0	2489	598	0.237179056	146514	617734	0.013842554	9360.2396	20275.32264	0.058363307	0.063886526	0.032822104
97	1	4	0	0	2	0	80	9	0.107825879	2752	25518	0.02797965	1063.4469	4836.919515	0.259489187	0.38649059	0.189546154
97	1	4	0	0	3	0	11	1	0.092669769	101	1088	0.044356136	100.81381	568.4560162	0.478647314	1	0.522534438
97	1	4	0	0	4	0	241	16	0.057625341	7217	125238	0.01643173	2068.4229	11570.12187	0.285147633	0.286609157	0.092385148
97	1	5	0	0	1	0	4156	1229	0.286625042	417429	1456358	0.011984963	22993.491	40197.19857	0.04181408	0.055083627	0.027601173
97	1	5	0	0	2	0	94	23	0.271666849	13184	48529	0.093473624	6491.888	11164.46034	0.344074459	0.49241937	0.230058611
97	1	5	0	0	3	0	852	92	0.104109935	32999	316960	0.016375782	5678.5623	20061.42228	0.157293174	0.172084243	0.063293134
97	1	5	0	0	4	0	360	17	0.043059509	11123	258318	0.013002948	3573.4271	29280.16722	0.301976216	0.321263795	0.113349466
97	1	6	0	0	1	0	2589	664	0.257375583	182740	710012	0.012944523	10876.15	20529.34608	0.050294295	0.059517124	0.028914068
97	1	6	0	0	2	0	38	9	0.229834164	2300	10008	0.074964207	796.48474	3058.740415	0.326166508	0.346261963	0.305621943
97	1	6	0	0	3	0	6	0	0	0	1971	0	0	0	.	.	0
97	1	6	0	0	4	0	165	13	0.079706386	5579	69992	0.033489734	2467.5802	9008.717665	0.420163745	0.442313866	0.128710914
97	1	7	0	0	1	0	4775	1392	0.286276725	599997	2095862	0.010173082	32087.458	61993.54302	0.035535835	0.053479404	0.029579018
97	1	7	0	0	2	0	343	89	0.250641759	32643	130239	0.031725945	6943.9856	20316.29048	0.126578845	0.212722359	0.155991945
97	1	7	0	0	3	0	164	12	0.083099626	5768	69411	0.019173921	1838.6128	11530.19669	0.230734144	0.318758983	0.166114649
97	1	7	0	0	4	0	1454	60	0.040343522	32144	796760	0.006252208	5091.3229	29714.65211	0.154974272	0.158390526	0.037294348
97	1	8	0	0	1	0	2650	777	0.279913571	170330	608511	0.012952867	10170.384	20165.05586	0.046274524	0.059709745	0.03313838
97	1	8	0	0	2	0	38	8	0.220178859	1697	7706	0.073127165	1010.0748	2449.106311	0.33212619	0.595281097	0.317798469
97	1	8	0	0	3	0	41	3	0.068716525	356	5174	0.047339957	106.21445	2259.490173	0.68891664	0.298716781	0.436664264
97	1	8	0	0	4	0	232	13	0.054056045	5806	107416	0.016918461	1702.7316	10345.04921	0.312979997	0.293246972	0.096308399
97	1	9	0	0	1	0	2396	657	0.272805203	206914	758470	0.013637743	11188.275	24339.46564	0.049990772	0.054071989	0.032090234
97	1	9	0	0	2	0	2	0	0	0	549	0	0	136.2709791	.	.	0.248081968
97	1	9	0	0	3	0	2	1	0.448669426	312	696	0	0	0	0	0	0
97	1	9	0	0	4	0	49	4	0.084008354	3634	43257	0.025052973	1457.1849	6504.913924	0.298220025	0.400992977	0.150378637
97	1	0	1	0	0	0	21858	4594	0.199417113	1592953	7988046	0.005400697	45983.207	79682.69808	0.027082414	0.028866642	0.009975243
97	1	0	2	0	0	0	11247	2910	0.26232575	824256	3142107	0.006869313	24235.69	46650.14562	0.026186193	0.029403127	0.014846773
97	1	0	1	0	0	1	11659	2000	0.166295423	751430	4518646	0.005454325	25822.82	50307.16093	0.032799008	0.034364896	0.011133238
97	1	0	1	0	0	2	10199	2594	0.242555754	841523	3469400	0.007180894	27306.381	43907.73166	0.029605128	0.032448767	0.012655713
97	1	0	2	0	0	1	6062	1282	0.219307186	395895	1805207	0.007492242	14366.619	29594.36924	0.034163232	0.036288975	0.016393892
97	1	0	2	0	0	2	5185	1628	0.32041343	428361	1336900	0.009073686	14300.697	23506.0159	0.028318683	0.033384711	0.017582481
97	1	0	1	0	1	0	12918	3602	0.278362521	1287902	4626708	0.006600139	43324.464	79521.1732	0.023710587	0.033639565	0.017187421
97	1	0	2	0	1	0	9688	2806	0.285387098	800851	2806193	0.007329576	23796.284	44025.40337	0.025682928	0.02971374	0.01568866
97	1	0	1	0	2	0	4461	745	0.169270474	193723	1144458	0.00925209	14422.789	41054.11821	0.05465862	0.074450572	0.035872093
97	1	0	2	0	2	0	1294	86	0.068228935	15512	227345	0.011289582	2584.7796	8915.991404	0.165466187	0.166635936	0.039217815
97	1	0	1	0	3	0	1134	111	0.098150527	41073	418471	0.013096322	6056.424	24155.68044	0.133430987	0.14745463	0.057723685
97	1	0	2	0	3	0	27	4	0.154370686	871	5641	0	0	0	0	0	0
97	1	0	1	0	4	0	3345	136	0.039065032	70255	1798409	0.00394347	7209.4572	50629.54053	0.100946289	0.102618554	0.028152405
97	1	0	2	0	4	0	238	14	0.068222628	7022	102928	0.026569362	3205.8042	9895.440383	0.389450875	0.456537397	0.096139737
97	1	0	1	0	1	1	6655	1540	0.234847262	602539	2565663	0.007289428	23488.416	44267.65326	0.031039016	0.038982404	0.017253885
97	1	0	1	0	1	2	6263	2062	0.332531885	685363	2061045	0.00898327	25352.132	39919.14046	0.027014763	0.036990803	0.0193684
97	1	0	2	0	1	1	5122	1244	0.242126367	386334	1595589	0.008262656	14190.819	28073.68712	0.034125387	0.03673199	0.017594565
97	1	0	2	0	1	2	4566	1562	0.342405185	414517	1210604	0.009406564	13932.485	21984.90576	0.027472025	0.033611362	0.018160277
97	1	0	1	0	2	1	2378	340	0.146021004	94027	643925	0.01068443	8415.1411	23525.70669	0.073170499	0.089497506	0.036534867
97	1	0	1	0	2	2	2083	405	0.199180375	99696	500534	0.012039366	7825.3748	18086.48772	0.060444541	0.078491994	0.036134413
97	1	0	2	0	2	1	762	32	0.046258699	6355	137380	0.012428826	1725.0566	5624.876033	0.268680842	0.271447969	0.040943938
97	1	0	2	0	2	2	532	54	0.10177813	9157	89965	0.020836874	1910.5908	4700.284304	0.204728407	0.208659053	0.052245411
97	1	0	1	0	3	1	721	62	0.086248416	23169	268636	0.013952779	3953.6447	15583.90169	0.161774323	0.170640619	0.058011234
97	1	0	1	0	3	2	413	49	0.119489572	17904	149835	0.018440891	3065.8921	9195.46308	0.154330546	0.171243374	0.061370621
97	1	0	2	0	3	1	18	2	0.114343905	458	4003	0	0	0	0	0	0
97	1	0	2	0	3	2	9	2	0.252225419	413	1638	0	0	0	0	0	0
97	1	0	1	0	4	1	1905	58	0.030463861	31695	1040422	0.004453482	4691.6678	29915.0171	0.146189023	0.148024151	0.028752759
97	1	0	1	0	4	2	1440	78	0.050871114	38560	757987	0.006367223	4848.5914	23040.46549	0.125163811	0.125742709	0.03039693

OHS	STR	PROV	U/R	TYPE	RACE	GENDER	N	n	R	MSWY	MSWX	SE-R	SE-WY	SE-WX	CV-R	CV-WY	CV-WX
97	1	0	2	0	4	1	160	4	0.04027265	2748	68235	0.01515593	1248.3028	7014.547475	0.37633306	0.454258038	0.102799864
97	1	0	2	0	4	2	78	10	0.123195766	4274	34693	0.045537848	1773.8595	1829.156363	0.36963809	0.41503563	0.052724538
97	1	1	1	0	0	0	4342	569	0.129389461	178405	1378825	0.009298994	13596.595	30016.89232	0.071868248	0.076211761	0.021769902
97	1	1	2	0	0	0	993	37	0.035519485	6655	187364	0.009226983	1741.1453	5880.630415	0.259772421	0.261626644	0.03138611
97	1	2	1	0	0	0	1650	440	0.247678643	165976	670128	0.025150057	18265.486	26706.59404	0.1015431	0.110048748	0.039852994
97	1	2	2	0	0	0	1169	435	0.368010212	137425	373428	0.023101763	8835.6273	18136.69089	0.062774787	0.064294083	0.048568147
97	1	3	1	0	0	0	1255	285	0.220712305	38836	175956	0.019377607	3326.0206	6510.64268	0.08779577	0.08564371	0.037001638
97	1	3	2	0	0	0	469	51	0.106429645	8373	78674	0.034491096	2639.3253	4916.409125	0.32407414	0.315209459	0.062490924
97	1	4	1	0	0	0	2000	477	0.217375785	119520	549831	0.015880227	8367.8173	19558.39785	0.073054258	0.070011837	0.035571627
97	1	4	2	0	0	0	821	147	0.168661182	37063	219747	0.01967461	4091.0057	11609.56038	0.116651679	0.11038062	0.052831587
97	1	5	1	0	0	0	3017	596	0.185191305	242888	1311549	0.01426492	19248.38	36809.59814	0.077028023	0.079248126	0.028065739
97	1	5	2	0	0	0	2445	765	0.301641755	231847	768616	0.016696581	15324.516	24307.63789	0.055352354	0.066097632	0.031625204
97	1	6	1	0	0	0	1126	263	0.227735532	77921	342154	0.020345	7903.2795	13829.60359	0.08933608	0.101427342	0.040419248
97	1	6	2	0	0	0	1672	423	0.250535409	112698	449830	0.015499274	7983.9868	16619.84125	0.061864605	0.070843908	0.036946958
97	1	7	1	0	0	0	6556	1535	0.220498248	659668	2991717	0.009767501	31956.82	48766.88622	0.044297406	0.048443761	0.016300633
97	1	7	2	0	0	0	180	18	0.108235645	10884	100555	0.035248883	4148.1226	8871.927472	0.325667974	0.381132382	0.088229244
97	1	8	1	0	0	0	1407	360	0.221522149	86679	391290	0.01808303	6944.1523	14268.93618	0.081630799	0.080113075	0.036466402
97	1	8	2	0	0	0	1554	441	0.271126254	91510	337517	0.019351204	7319.0956	12284.52591	0.071373405	0.079981513	0.036396713
97	1	9	1	0	0	0	505	69	0.130579552	23060	176596	0.01856503	3529.2502	13495.95519	0.142174096	0.153047774	0.076422875
97	1	9	2	0	0	0	1944	593	0.299821265	187801	626376	0.015805663	10639.806	23602.13312	0.052716953	0.056654738	0.037680462