
III Data Use Guidelines

Sample Weighting Guidelines for Tabulation

Sample Weighting Guidelines for Tabulation, the sample design used for the Provincial Heart Health Surveys, were not self-weighting. When producing simple estimates, including the production of ordinary statistical tables, users must apply the proper sampling weight.

If proper weights are not used, the estimates derived from the micro data tapes cannot be considered to be representative of the survey population.

Guidelines for Statistical Analysis

The Provincial Heart Health Surveys are based upon a complex sample design, with stratification, multiple stages of selection, and unequal probabilities of selection of respondents. Using data from such complex surveys presents problems to analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used. In order for survey estimates and analyses to be free from bias, survey weights must be used.

Approximate co-efficient of variances (CV) for simple estimates such as totals, proportions and ratios (for qualitative variables) are provided in the accompanying approximate CV Tables. Statistical analyses can also be done through the Canadian Heart Health Data Base Centre using the JACKVAR package to calculate variance/co-variance values (see Introduction for User Support information).

Co-efficient Variance Release Guidelines

Before releasing and/or publishing any estimate, users should first determine the quality level of the estimate. The quality levels are , and . Data quality is affected by both sampling and non-sampling errors. However, for this purpose, the quality level of an estimate will be determined only on the basis of sampling error as reflected by the co-efficient of variance as shown in Table 3 below.

First, the number of respondents who contribute to the calculation of the estimate should be determined. If this number is less than 30, the weighted estimate should be considered to be of quality.

For weighted estimates based on sample sizes of 30 or more, users should determine the coefficient of variance of the estimate and follow the guidelines below. These quality level guidelines should be applied to weighted, rounded estimates.

All estimates can be considered releasable. However, those of or quality level must be accompanied by a warning to caution subsequent users.

Table 3: Quality Level Guidelines

Quality Level of Estimate	Guidelines
1. Acceptable	<p>Estimates have:</p> <ul style="list-style-type: none"> \$ a sample size of 30 or more, and \$ low co-efficients of variance in the range 0.0% - 16.5% <p>No warning is required.</p>
2. Marginal	<p>Estimates have:</p> <ul style="list-style-type: none"> \$ a sample size of 30 or more, and \$ high co-efficients of variance in the range 16.6% - 33.3%. <p>Estimates should be flagged with the letter M (or some similar identifier). They should be accompanied by a warning to caution subsequent users about the high levels of error, associated with the estimates.</p>
3. Unacceptable	<p>Estimates have:</p> <ul style="list-style-type: none"> \$ a sample size of less than 30, or \$ very high co-efficients of variance (in excess of 33.3%). <p>It is recommended that estimates of unacceptable quality not be released. However, if the user chooses to do so, then estimates should be flagged with the letter U (or some similar identifier) and the following warning should accompany the estimates:</p> <p>"The user is advised that . . . (specify the data) . . . do not meet quality standards for this statistical program. Conclusions based on these data will be unreliable, and most likely invalid. These data and any consequent findings should not be published. If the user chooses to publish these data or findings, then this disclaimer must be published with the data."</p>

Approximate Co-efficient of Variance (CV) Tables

In order to supply co-efficients of variance, which would be applicable to a wide variety of categorical estimates produced from this micro data file and which could be readily accessed by the user, a set of Approximate CV Tables is provided in this Data Use Guidelines section. These "look-up" tables allow the user to obtain an approximate co-efficient of variance based on the size of the estimate.

The co-efficients of variance are derived using the variance formula for simple random sampling and incorporating a factor which reflects the multi-stage, complex nature of the sample design. This factor, known as the design effect, was determined by first calculating design effects for a wide range of characteristics and then choosing, from among these, a conservative value to be used in the look-up tables.

Table 4 below shows the design effects factors (DEF) for major variables at Canada and single provincial level.

Table 4: Design Effect Factors (DEF) Values

Classificatory Variables	Design Effect Factor (DEF)
Any Age Group	1.00
Sex	1.00
Total cholesterol (High, Medium, Low)	1.25
BMI (— 27, § 27)	1.50
Smoking (Yes or No)	1.25
Years of Education	1.50
Sedentary Lifestyle (Active or Inactive)	1.50
Diabetes (Yes or No)	2.00
Blood Pressure (High, Low)	1.25
Other Home Interview Variables	1.50
Other Clinic Variables	2.00
Cross Tabulation of Other Home Interview and Clinic Variables	3.00

All co-efficients of variance in the Approximate CV Tables are approximate and, therefore, should be used with care.

~~It is not~~ If the number of observations on which an estimate is based is less than 30, the weighted estimate should not be released, regardless of the value of the co-efficient of variance for this estimate.

How to Use the CV Tables for Categorical Estimates

The following rules should enable the user to determine the approximate co-efficients of variance from the CV Tables for estimates of the number, proportion or percentage of the surveyed population possessing a certain characteristic and for ratios and differences between such estimates.

Rule 1: Estimates of Numbers Possessing a Characteristic (Aggregates)

The co-efficient of variance depends only upon the size of the estimate itself. On the CV Table for the appropriate geographic area, locate the estimated number in the left-most column of the table (headed "Numerator of Percentage") and follow the asterisks (if any) across to the first figure encountered. This figure is the approximate co-efficient of variance.

Rule 2: Estimates of Proportions or Percentages Possessing a Characteristic

The co-efficient of variance of an estimated proportion or percentage depends on both the size of the proportion or percentage and the size of the total upon which the proportion or percentage is based. Estimated proportions or percentages are relatively more reliable than the corresponding estimates of the numerator of the proportion or percentage, when the proportion or percentage is based upon a sub-group of the population. For example, the proportion of persons with hypertension is more reliable than the estimated number of persons with hypertension. (Note that in the tables the CVs decline in value, reading from left to right).

When the proportion or percentage is based on the total population of the geographic area covered by the table, the CV of the proportion or percentage is the same as the CV of the numerator of the proportion or percentage. In this case, Rule 1 can be used.

When the proportion or percentage is based on a subset of the total population (e.g. those in a particular sex or age group), reference should be made to the proportion or percentage (across the top of the table) and to the numerator of the proportion or percentage (down the left side of the table). The intersection of the appropriate row and column gives the co-efficient of variance.

Rule 3: Estimates of Differences Between Aggregates or Percentages

The standard error of a difference between two estimates is approximately equal to the square root of the sum of squares of each standard error considered separately. That is, the standard error of a difference ($\hat{d} = \hat{X}_1 - \hat{X}_2$) is:

$$\sigma_{\hat{d}} = \sqrt{(\hat{X}_1 \alpha_1)^2 + (\hat{X}_2 \alpha_2)^2}$$

where \hat{X}_1 is estimate 1, \hat{X}_2 is estimate 2, for example \hat{X}_1 is an estimate of prevalence of hypertensive in a province and \hat{X}_2 is an estimate of prevalence in other provinces and α_1 and α_2 are the co-efficients of variance of \hat{X}_1 and \hat{X}_2 respectively. The co-efficient of variance of \hat{d} is given by $\sigma_{\hat{d}}/\hat{d}$. This formula is accurate for the difference between separate and uncorrelated characteristics, but is only approximate otherwise.

Rule 4: Estimates of Ratios

In the case where the numerator is a subset of the denominator, the ratio should be converted to a percentage and Rule 2 applied. This would apply, for example, to the case where the denominator is the number of persons with hypertension and the numerator is the number of persons who have hypertension and diabetes.

In the case where the numerator is not a subset of the denominator, the standard deviation of the ratio of the estimates is approximately equal to the square root of the sum of squares of each co-efficient of variance considered separately multiplied by R. That is, the standard error of a ratio ($\hat{R} = \hat{X}_1/\hat{X}_2$) is:

$$\sigma_{\hat{R}} = \hat{R} \sqrt{\alpha_1^2 + \alpha_2^2}$$

where α_1 and α_2 are the co-efficients of variance of \hat{X}_1 and \hat{X}_2 respectively. The co-efficient of variance of \hat{R} is given by $\sigma_{\hat{R}}/\hat{R}$. The formula will tend to overstate the error, if \hat{X}_1 and \hat{X}_2 are positively correlated and understate the error if \hat{X}_1 and \hat{X}_2 are negatively correlated.

Rule 5: Estimates of Differences of Ratios

In this case, Rules 3 and 4 are combined. The CVs for the two ratios are first determined using Rule 4, and then the CV of their difference is found using Rule 3.

How to Use the CV Tables to Obtain Confidence Limits

Confidence intervals for an estimate can be calculated directly from the Approximate CV Tables by first determining, from the appropriate table, the co-efficient of variance of the estimate \hat{X} , and then using the following formula to convert to a confidence interval CI:

$$CI_X = [\hat{X} - t \hat{X} \alpha_{\hat{X}}, \hat{X} + t \hat{X} \alpha_{\hat{X}}]$$

where $\alpha_{\hat{X}}$ is the determined co-efficient of variance of \hat{X} , and

t = 1 if a 68% confidence interval is desired

t = 1.6 if a 90% confidence interval is desired

t = 2 if a 95% confidence interval is desired

t = 3 if a 99% confidence interval is desired.

Note: Release guidelines which apply to the estimate also apply to the confidence interval. For example, if the estimate is not releasable, then the confidence interval is not releasable.

How to Use the CV Tables to do the T-Test

Standard errors may also be used to perform hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The sample estimates can be numbers, averages, percentages, ratios, etc. Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the characteristics are different when, in fact, they are identical.

Let \mathbf{X}_1 and \mathbf{X}_2 be sample estimates of characteristics of interest for two groups. Let the standard error on the difference $\hat{d} = \hat{X}_1 - \hat{X}_2$ be $\sigma_{\hat{d}}$.

If

$$t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}}$$

is between -2 and 2, then no conclusion about the difference between the characteristics is justified at the 5% level of significance. If however, this ratio is smaller than -2 or larger than +2, the observed difference is significant at the 0.05 level. That is to say that the characteristics are significant.