MINISTRY OF BUSINESS, INNOVATION & EMPLOYMENT
HĪKINA WHAKATUTUKI

Technical paper - empirical equations for IVS relative margin of error

September 18, 2015

Contents

Introduction

The relative margin of errors (RMEs) for major survey estimates are the benchmark of the survey design quality. For example, the international visitor survey (IVS) was designed to achieve a 5 per cent relative margin of error (at the 95 per cent confidence level) for total visitor expenditure, and less than 10 per cent relative margin of error for expenditure from the top six tourism market countries (Australia, United Kingdom, United States, China, Japan and Germany).

In designing of the survey, it is necessary to know what sample sizes are required to achieve the target RMEs under certain confidence level, or when the resources (sample sizes) are limited, what RMEs can be achieved. Therefore, a equation which relates the RME, the sample size and confidence level, will become a handy tool.

In this document, we establish a simple empirical equation, based on historic RME estimates and corresponding sample sizes to serve this purpose. Although the method is illustrated in the context of IVS, it is general enough to be able to apply to other surveys.

Theoretical background

The margin of error (ME) is defined as the half width of confidence interval. When the estimates ($\hat{\theta}$) are assumed to be normally distributed, the margin of error is often expressed as a function of the standard error (SE). E.g., for 95% confident level (significance level $\alpha = 5\%$),

$$\mathsf{ME}(\hat{\theta}) = z_{1-\alpha/2}\mathsf{SE}(\hat{\theta}) = 1.96\mathsf{SE}(\hat{\theta}) = 1.96\sqrt{\mathsf{Var}(\hat{\theta})}.$$

For positive measures of a survey sample such as mean and total of the expenditures, the relative margin of error (RME)

$$\mathsf{RME}(\hat{\theta}) = 1.96\frac{\mathsf{SE}(\hat{\theta})}{\hat{\theta}} \tag{1}$$

are often used.

Empirical equation

By central limit theorem, when sample size (n) increases, an estimate ($\hat{\theta}$) approaches to its true value (θ) in order of $1/\sqrt{n}$. Therefore, $\hat{\theta}$ can be represented as

$$\hat{\theta} = \theta + \frac{\epsilon}{\sqrt{n}}$$

with $\epsilon \sim N(0, \sigma_\epsilon^2)$ being a normally distributed error and terms "smaller" than order of $1/\sqrt{n}$ being ignored.

Consequently,

$$\mathsf{RME}(\hat{\theta}) = \frac{1.96\sigma_\epsilon/\sqrt{n}}{\theta + \epsilon/\sqrt{n}}. \tag{2}$$

Rearrange formula (2), we can have

$$\frac{1}{\text{RME}(\hat{\theta})} = \frac{\theta}{1.96\sigma_\epsilon}\sqrt{n} + \frac{\epsilon}{1.96\sigma_\epsilon} = b\sqrt{n} + e, \tag{3}$$

with $b = \frac{1.96\sigma_\epsilon}{\theta}$ being a constant and $e = \frac{\epsilon}{1.96\sigma_\epsilon}$ being a constant variance error term.

Therefore, the relationship between a set of RMEs ($\{\text{RME}_i\}$) and the corresponding sample sizes ($\{n_i\}$) can be described by a linear regression without intercept model

$$\frac{1}{\text{RME}_i} = b\sqrt{n_i} + e_i, \quad e_i \sim N(0, \sigma_e),$$

or a linear regression with intercept model

$$\frac{1}{\text{RME}_i} = a + b\sqrt{n_i} + e_i, \quad e_i \sim N(0, \sigma_e),$$

if $\hat{\theta}$ is a biased estimate of θ.

In practice, the true value of RME are not available. Rather, the estimated RMEs ($\widehat{\text{RME}}$) can be usd as a proxy, which deviates from the true RME in order of $1/\sqrt{n}$. In this case,

$$\frac{1}{\widehat{\text{RME}}_i} = a + b\sqrt{n_i} + e_i, \quad e_i \sim N(0, \sigma_e) \tag{4}$$

is still an appropriate model relating the sample sizes to RMEs, with the coefficient a accounting for the difference between $\widehat{\text{RME}}$ and RME and the possible biasedness of $\hat{\theta}$ and $\widehat{\text{RME}}$ from θ and RME.

Establishing the empirical equation

In this section, we show how to establish the empirical equation of RMEs vs sample size for expenditure statistics (total and mean) in International Visitor Survey (IVS) based on historic RME estimates.

Estimation of RMEs

As a complexly designed sample survey, the standard errors for estimates (and therefore RMEs calculated by formula 1) of IVS will usually neither be same as the standard errors simply calculated by the observations, nor that in a simple random sampling design (see e.g. p6 in Lumley, 2011)).

Generally, the standard error for an estimate increases as the complexity of the sampling design increases. For example, levels of dimensions that have been included in the weighting (e.g. Australia) will have much better estimates than those that aren't (e.g. Tonga, which is categorised as Other), basically because the estimate is in effect only for average spend for Australia as we know exactly how many came, whereas with Tonga we have to estimate both the number who came (from the sampling process) as well as their spend.

Unlike in simpler sampling designs where analytic formulas can be derived to compute the standard error, the complex sampling designs (such as IVS) do not have analytic formulas available. Therefore, the re-sampling techniques such as bootstrapping and jackknifing become generally applicable and robust approaches for standard error estimation in complex sample survey.

Compute the empirical equation

The RMEs for annual (year ended June) expenditure estimates for the whole population, for the sub-populations grouped by countries and by purpose of visit (POV) from 1998 to 2014 have been calculated by bootstrapping approach. An excerpt of the results are presented below.

```
Emp.Equation.df[c(1:3,222:224,290:292),]
```

```
##          Country_POV_Total     YEJun RMEs SampleSizes     groups
## 1   Africa and Middle East YEJun1998   29          88 by_country
## 2                Australia YEJun1998   13        1279 by_country
## 3                   Canada YEJun1998   24         168 by_country
## 222               Business YEJun1998   12         733     by_POV
## 223                  Other YEJun1998   19         280     by_POV
## 224     Holiday / vacation YEJun1998    7        3380     by_POV
## 290                  Total YEJun1998    6        5377      whole
## 291                  Total YEJun1999    7        5377      whole
## 292                  Total YEJun2000    7        5419      whole
```

In conjunction with the corresponding sample sizes, the linear regression model (formula 4) can be fitted,

```
fit <- lm(1/RMEs~sqrt(SampleSizes), data = Emp.Equation.df)
(sum_fit <- summary(fit))
```

```
##
## Call:
## lm(formula = 1/RMEs ~ sqrt(SampleSizes), data = Emp.Equation.df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.079687 -0.005341 -0.000056  0.006114  0.035896
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.053e-02  1.209e-03    8.71   <2e-16 ***
## sqrt(SampleSizes) 2.168e-03  4.051e-05   53.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01157 on 304 degrees of freedom
## Multiple R-squared:  0.9041,Adjusted R-squared:  0.9038
## F-statistic:  2866 on 1 and 304 DF,  p-value: < 2.2e-16
```

with the fitted coefficients $a = 0.011 \pm 0.002$ and $b = 0.0022 \pm 0.0001$.

The RMEs vs n plots (figure 1) shows that the empirical equation well described the relationship between RMEs and sample sizes (n), regardless of the various magnitudes of sample size for populations (group "whole") and that for sub-populations (groups "by_country" and "by_POV")

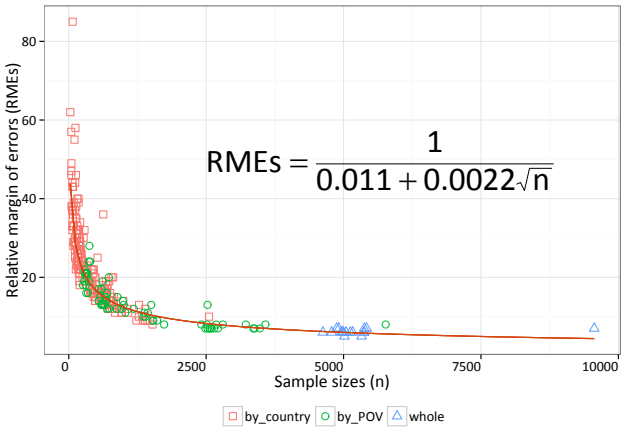

Figure 1: Fitted empirical equation of RMEs vs n for IVS expenditures

To diagnose the validity of linear regression model assumption, figure 2 shows the existence of linear relationship between $\frac{1}{\widehat{\text{RME}}}$ and sample size n.

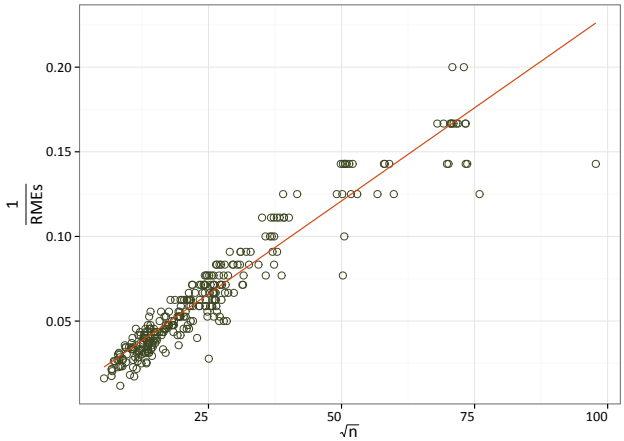

Figure 2: Regression diagnostic - linearity of the linear regression model. The linearity assumption is approximately valid.

Figure 3 shows that the constant variances assumption for residuals is not well-satisfied. The trumpet-shaped residuals-vs-fitted plot suggests that larger size populations/sub-

populations (corresponding to larger sample sizes as well as the large fitted values) are more heteroscedastic than smaller population/sub-population. However, since the RME values decreases as sample sizes increases,and the linearity of the empirical equation is well maintained, the impact of this heteroscedasticity of residuals on the validity of the empirical equation can be ignored.

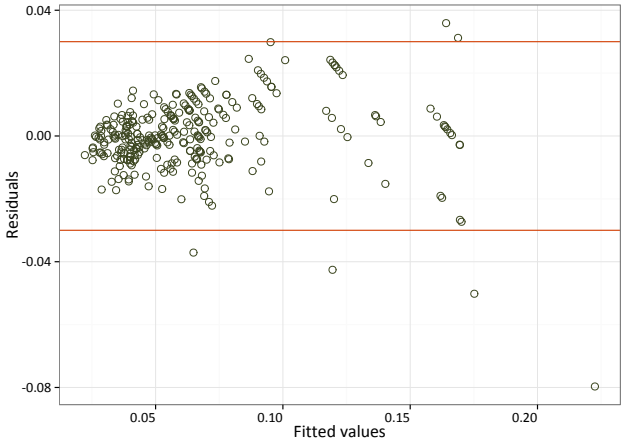

Figure 3: Regression diagnostic - verification of constant variance for residuals. A trumpet-shaped feature is observed.

Figure 4 show the normality assumption of residuals is approximately valid for individual groups ("by_country","by_POV" and "whole"). For the collection of residuals for all groups, the linearity of the qq-plot is not perfect but still acceptable if excluding some outliers.

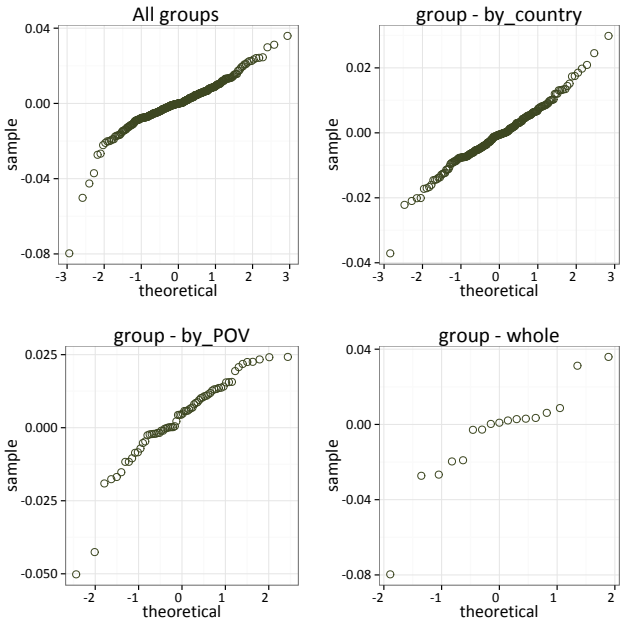

Figure 4: Regression diagnostic - normality of residuals. The qq plots shows the normality is approximately satisfied.

In summary, the proposed empirical equation can (at least approximately) describe the relationship between RMEs and sample size.

Usage of the empirical equation

Once the empirical equation is established, the predicted RME can be computed based on the sample size and vice versa. For example, for a (total or mean) expenditure estimate with sample size 1000, the relative margin of error is expected to be 12.4% at 95% confidence level.

```
n <- 1000
(RME <- round(1/(a+b*sqrt(n)),1))

## [1] 12.4
```

Although current empirical equation was established for 95% confidence level, it can be easily converted to equations for other confidence levels by the formula (5),

$$\text{RME}_{1-\alpha} = \frac{z_{1-\alpha/2}}{1.96} \times \frac{1}{a + b\sqrt{n}} \tag{5}$$

where $1 - \alpha$ is the target confidence level and $z_{1-\alpha/2}$ is the normal score (z-score). Table (1) lists the predicted RMEs for some levels of the sample size.

Table 1: Predicted RMEs for specified sample sizes under confidence level 95%, 90% and 80% for expenditure estiamtes.

Sample Size	Confidence Level		
	95%	90%	80%
20	48.0	40.2	31.3
50	37.7	31.5	24.6
100	30.3	25.4	19.8
200	23.7	19.9	15.5
500	16.6	13.9	10.8
1000	12.4	10.4	8.1
2000	9.1	7.6	6.0
5000	6.0	5.0	3.9
10000	4.3	3.6	2.8

Based on the empirical equation, the minimum required sample sizes to achieve certain level of relative margin of error can be computed by the formula (6), as shown in table (2)

$$n = \frac{1}{b^2}\left(\frac{z_{1-\alpha/2}}{1.96\text{RME}_{1-\alpha}} - a\right)^2. \tag{6}$$

Table 2: Minimum samples size for expected relative margin of error under confidence level 95%, 90% and 80% for expenditure estiamtes.

Expected RMEs	Confidence Level		
	95%	90%	80%
5	7380	5050	2956
10	1637	1091	609
15	640	414	219
20	314	196	97

The relationship can also be described by Figure 5.

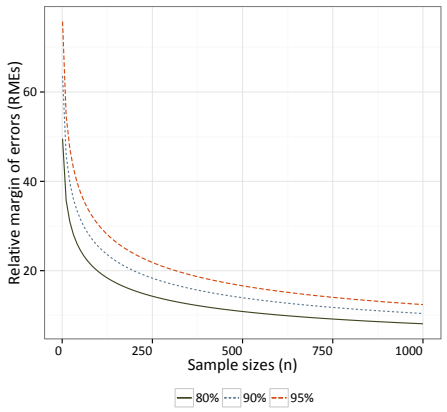

Figure 5: RMEs vs n for IVS expenditures with confidence level 80%, 90% and 95%.

Empirical equation for counts (number of visitors) estimates

To further verify the applicability of the proposed method, we may apply the above approach to the estimates of number of visitors (counts) rather than the expenditure. Since the number of visitors is a different measure and therefore has different inherent variability from the expenditure, it is expected that the coefficients for the empirical equation will be different. But what we are interested in is if the general methodology is still applicable.

The result shows that RMEs for the counts are generally smaller than RMEs for the expenditures. This can be explained by the fact that extra variations due to various expenditures of each visitor. Therefore expenditure estimates are expected less precise than counts estimates.

```
  Emp.Equation.df[c(1:3,222:224,290:292),]

##           Country_POV_Total     YEJun RMEs SampleSizes     groups
## 1    Africa and Middle East YEJun1998   23          88 by_country
## 2                 Australia YEJun1998    7        1279 by_country
## 3                    Canada YEJun1998   17         168 by_country
## 222               Business YEJun1998    9         733     by_POV
## 223                  Other YEJun1998   14         280     by_POV
## 224     Holiday / vacation YEJun1998    6        3380     by_POV
## 290                  Total YEJun1998    5        5377      whole
## 291                  Total YEJun1999    5        5377      whole
## 292                  Total YEJun2000    5        5419      whole
```

For estimated number of visitors, the fitted linear regression model (formula 7) is,

$$\frac{1}{\widehat{\text{RME}}_i} = a + b\sqrt{n_i} + e_i, \quad e_i \sim N(0, \sigma_e) \tag{7}$$

with coefficients $a = 0.017 \pm 0.003$ and $b = 0.0026 \pm 0.0001$.

The pattern of resulting RMEs-vs-n plot (figure 6) is similar as that for expenditure estimates with only different coefficients. This confirmed the applicability of the proposed method.

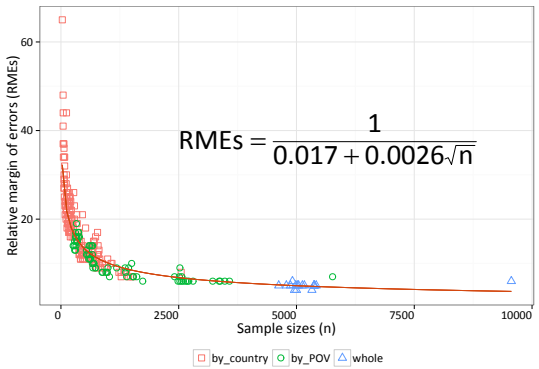

Figure 6: Empirical equation of RMEs vs n for IVS number of visitors.

The features of diagnostic plots (figure 7) for linear regression model assumption are also similar. The linearity maintains well, while the trumpet-shaped residuals is an indication of heteroscedasticity with respect to the population size.

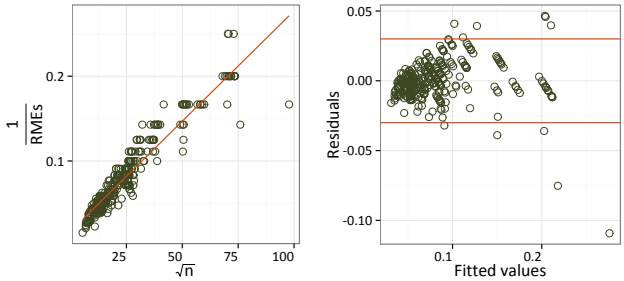

Figure 7: Regression diagnostics - linearity and constant variance of residuals for fitting empirical equation of RMEs vs n for IVS number of visitors. Linearity is satisfied, while the variance of residuals are not constant.

Figure 8 show the normality assumption of residuals. The results are similar to that for expenditure estimates.

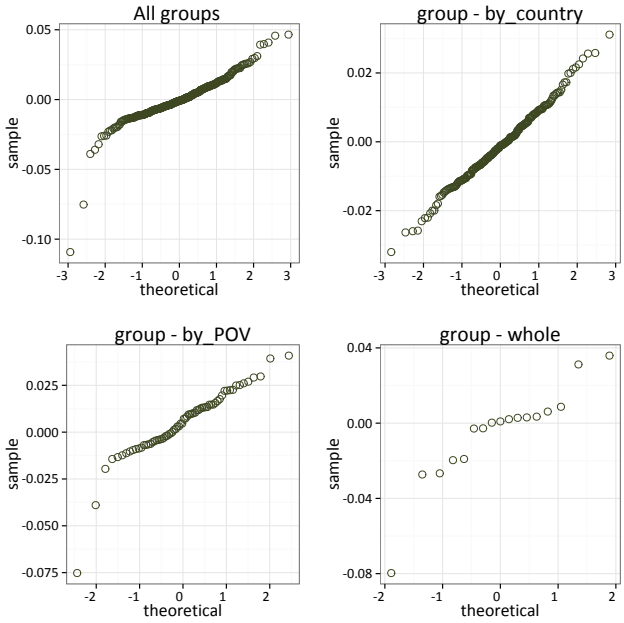

Figure 8: Regression diagnostic - normality of residuals for fitting empirical equation of RMEs vs n for IVS number of visitors. Although not perfect, the linearity of QQ plots are approximately maintained.

The tables of predicted RMEs for specified sample sizes (table 3) and the minimum samples sizes for particular RMEs show that smaller RMEs are achieved for number of visitors (counts) than for expenditures for same sample sizes.

Table 3: Predicted RMEs for specified sample sizes for counts estimates.

Sample Size	Confidence Level		
	95%	90%	80%
20	34.9	29.2	22.8
50	28.3	23.6	18.5
100	23.3	19.5	15.2
200	18.6	15.6	12.1
500	13.3	11.1	8.7
1000	10.1	8.4	6.6
2000	7.5	6.3	4.9
5000	5.0	4.2	3.3
10000	3.6	3.0	2.4

Table 4: Minimum samples size for expected relative margin of error for counts estimates.

Expected RMEs	Confidence Level		
	95%	90%	80%
5	4954	3344	1909
10	1019	658	345
15	365	222	104
20	161	91	36

The relationship can also be described by Figure 9.

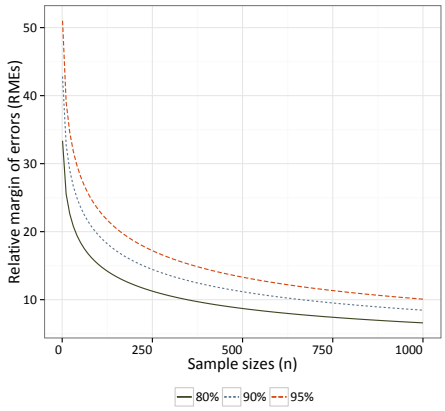

Figure 9: RMEs vs n for number of visitors with confidence level 80%, 90% and 95%

Some further discussions

The applicability of proposed method

- Although the examples are based only on the historic data of IVS, the proposed method was derived on a general basis. Therefore, it is expected that the methodology can apply to other surveys.
- The empirical equation (indexed by the fitted coefficients) depends on the inherent heterogeneity of the target survey measures. Therefore, the equation fitted for one survey measure (e.g. expenditure) cannot be directly used to another measure (e.g. number of visitors).
- Furthermore, the equation established for one particular survey (say IVS) cannot be directly used for other survey.
- An implicit assumption for the validity of the empirical equation is that the variability of the historic populations (e.g. variability of expenditures of visitors from 1998 to 2014) can reflect that of the future population (e.g. variability of expenditures of visitors in 2015). However, this is a reasonable assumption.

The sensitivity of the empirical equation (or fitted curve) to the coefficients

For some audiences, it might be interested in knowing how sensitive the equation (or fitted curve) is to the coefficients. The Figure 10 shows the curve corresponding to the fitted empirical equation for the IVS expenditure estimates (with coefficients $a = 0.011$ and $b = 0.0022$) and curves corresponding to the equations with coefficients taking the critical values ($a = 0.011 \pm 0.002$ and $b = 0.0022 \pm 0.0001$).

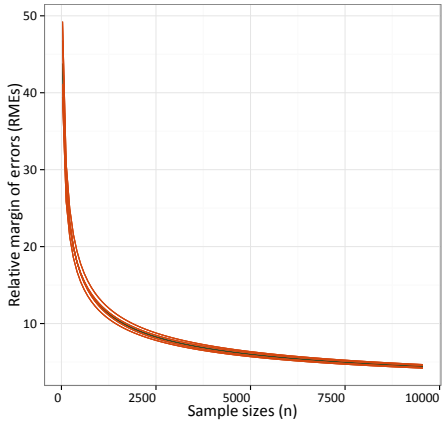

Figure 10: Sensitivity of empirical equation to the coefficients. The black line is the curve of the fitted empirical equation, while the red lines correspond to the equations with coefficients taking the critical values.

Overall, the curves corresponding to the equations with coefficients taking the critical values are close to the curve of the fitted empirical equation.

Summary

- For a particular survey measure, based on the historic RME estimates and corresponding sample size, the relationship between RMEs and sample sizes can be well (though approximately) described by a single empirical equation.
- Once established, the empirical equation becomes a convenient tool to predict survey RME based on the sample size, as well as to decide sample size for target RME.

References

Thomas Lumley. *Complex surveys: A guide to analysis using R*. John Wiley & Sons, 2011.