International Labour Organization

ILO-IPEC Interactive Sampling Tools No. 1

Sample size and margin of error

Version 1

August 2014

International Programme on the Elimination of Child Labour (IPEC)

Fundamental Principles and Rights at Work (FPRW) Branch
Governance and Tripartism Department

ILO-IPEC

ILO-IPEC Interactive Sampling Tools No. 1 – Sample size and margin error / International Labour Office, International Programme on the Elimination of Child Labour (IPEC) - Geneva: ILO, 2014

ACKNOWLEDGEMENTS

This publication was elaborated by Mr. Farhad Mehran, consultant, for ILO-IPEC and coordinated by Mr. Federico Blanco Allais from IPEC Geneva Office.

Funding for this ILO publication was provided by the United States Department of Labor (Projects GLO/13/21/USA & GLO/10/55/USA).

This publication does not necessarily reflect the views or policies of the United States Department of Labor, nor does mention of trade names, commercial products, or organizations imply endorsement by the United States Government.

Visit our website: www.ilo.org/ipec

Available in electronic PDF format only.

Photocomposed by ILO-IPEC Geneva.

1. Introduction

This document describes the use of the template "Sample Size" of the SIMPOC Interactive Sampling Tools. The template assists the user to calculate the required sample size for a reporting domain of a child labour survey based on alternative sets of parameters.

If the child labour survey is conducted as a module attached to a broader household-based survey such as a labour force survey, the sample size of the broader survey determines the sample size of the child labour module. In this situation, the template assists to calculate the margin of errors of child labour estimates for the given sample size of the broader survey.

The problem of sample size determination for a given margin of errors is described in Section 2. The reverse problem of margin of error determination for a given sample size is described in Section 3. Section 4 gives step-by-step instructions on the use of the template with numerical illustrations.

2. Sample size determination

Sample size determination in most household-based surveys with multi-stage stratified design is based on the principle of first calculating the required sample size for a single «domain» assuming a simple random sample design and no non-response. A domain is a well-defined population group for which estimates with pre-determined accuracy are sought. The results are then extended to allow for non-response and deviation from simple random sampling. Finally, the total sample size is obtained from the addition of the required sample size for single domains over all reporting domains of the survey.

The required sample size for a reporting domain is determined by the following formulae:

$$n = \frac{4 \times \sigma^2 \times deff}{ME^2 \times AveHH \times RR}$$

where σ^2 is the assumed value of the standard deviation of the underlying variable defining the main indicator of the survey, *deff* is the design effect, *ME* is the specified margin of error at 95% confidence level, *AveHH* is the estimated average number of persons in the target or base population per household, and *RR* is the expected response rate of the survey. The above formula for sample size calculation ignores the finite population correction, assumed to be negligible in most national child labour surveys. The various parameters of the formulae are now described in turn below.

The multiplier 4 in the expression is the rounded squared value of the tail of the standard normal distribution α = 1.96 corresponding to a 5% two-tailed significance level (4=1.96^2).

The above formulation allows the determination of the sample size for estimation of both ratios and amounts. For example, suppose the main indicator of the child labour survey is considered to be the percentage of children 5-17 years old who are engaged in activities defined as child labour. Let r be an assumed value of this percentage then

$$\sigma^2 = r \times (1 - r)$$

More generally, suppose the main indicator of the survey is the total income from employment generated by working children. Then

$$\sigma^2 = Var(y)$$

where *Var* is the variance of the underlying variable y, the income from employment of a working child.

The design effect is the ratio of the variance of the estimator of the main indicator obtained under the sample design to the variance of the estimate that would have been obtained under simple random sampling with the same size.

$$deff = \frac{Var_d(estimator)}{Var_o(estimator)}$$

where $Var_d(estimator)$ is the variance of the estimator under the proposed sample design for the survey and $Var_o(estimator)$ is the variance of the estimator that would have been obtained under a simple random sample design. The design effect expresses how well the proposed sample design compares to the reference design (simple random sample). The design effect of national child labour surveys with conventional two-stage stratified sample design is often set in the range from 3 to 4.

The design effect may be expressed in terms of the intra-class correlation of the units within the same cluster or primary sampling unit. The intra-class correlation measures the degree in which the work statuses of children in the same area are similar. In terms of the intra-class correlation, the design effect may be expressed as

$$deff = 1 + (\frac{b}{h} - 1) \times roh$$

where b is the average number of sample households per cluster (the sample take), h is the expected number of households required to find one base population unit, and *roh* is the intra-cluster correlation. For example, if the sample design

envisages to sample 15 households in each primary sampling unit (that is b=15) and for every three households there are two children in the age group 5 to 17 years old, then an intra-class correlation, *roh* = 0.22, implies a design effect, *deff* = 3, and a intra-class correlation, *roh* = 0.33, implies a design effect, *deff* = 4.

In the extreme situation of one sample household per primary sampling unit (b=1) and one sample child per household (h=1), the ratio b/h is equal to 1, the intra-class correlation does not intervene and the design effect reduces to one, *deff* = 1.

The margin of errors is a value chosen to reflect the required precision of the survey estimate. It can be specified in absolute or relative terms. In absolute terms, the margin of errors (ME) is about half of the length of the confidence interval at 95% level of significance of the survey estimate around the true value of the indicator. For example, if the margin of error is set at 1% (i.e., 1 percentage point),

$$ME = 1\%$$

This means that the confidence interval of the estimate of the child labour rate (percentage of children engaged child labour) would be in the interval ($r \pm 1\%$), where r is the predicted value of the child labour rate.

The alternative specification of the margin of errors in relative terms is more common when the main indicator to be estimated from the survey is an absolute figure, such as the total number of children engaged in child labour or the total income from employment generated by working children. The relative margin of errors (RME) may be expressed as

$$RME = \frac{ME}{y}$$

where y is the predicted value of the main indicator. For example, let the main indicator be the total number of working children in the country, predicted to be around 20'000. Suppose that it should be estimated with a margin of error of 100, i.e., *ME* = 100. The corresponding relative margin of error for estimation of the total number of working children is

$$RME = \frac{100}{20'000} = 0.5\%.$$

The next term in the sample size formulae is *AveHH*, the estimated average number of persons in the target or base population per household. This parameter accounts for the difference between the sampling unit and the analytic unit of the survey. The sampling unit of the survey is a household while the analytical unit is a child 5 to 17 years old.

Thus, *AveHH* is the average number of children 5 to 17 years old that can be found in a given household. Its value is generally estimated from previous surveys or the most recent population census. For example, suppose the last population and

housing census of the country recorded 80'000 households and 60'000 children in the age group 5-17 years old, then *AveHH* may be estimated as

$$AveHH = \frac{60'000}{80'000} = 0.75$$

The term *AveHH* may also be expressed as the product of two other parameters,

$$AveHH = pb \times AveSize$$

where *pb* is the proportion of the base population in total population and *AveSize* is the average household size. In the context of child labour surveys, the base population would generally be the children in the age group 5-17 years old, and *pb* would be the proportion of children 5-17 years old in the total population. *AveSize* is simply the average number of household members per household. The values of *pb* and *AveSize* can similarly be estimated from previous surveys or from the most recent population census.

Thus, if the country has 80'000 households and 400'000 inhabitants of whom 60'000 children 5-17 years old, then

$$pb = \frac{60'000}{400'000} = 15\%$$

$$AveSize = \frac{400'000}{80'000} = 5$$

It can then be verified that *AveHH*, the average number of children 5-17 years old per household, can be obtained as the product of *pb* and *AveSize*,

$$AveHH = 0.15 \times 5 = 0.75$$

It can also be verified that the parameter *h* in the expression of design effect in terms of the intra-class correlation is equal to the inverse of *AveHH*,

$$h = \frac{1}{AveHH}$$

Finally, the last term in the sample size formulae is the response rate *RR.* It accounts for possible non-response of sample households due to absence after repeated visits of the interviewer or due to refusal to participate in the survey. The response rate can be estimated from earlier experience during the fieldwork of similar household-based surveys, for example, previous child labour surveys or labour force surveys. It can also be estimated as part of a pilot survey prior to the child labour force. Typically, the response rate of SIMPOC child labour surveys ranged from about 80% to about 95%.

$$RR = 80\% - 95\%$$

Generally, response rates in rural areas are higher than response rates in urban areas. Similarly, response rates in countries at higher levels of development are generally higher than in countries at lower levels of development.

A numerical illustration of the use of the sample size formulae is given below. Consider the design of a national child labour survey in a country where it is predicted that the child labour rate is about 16%, that is 16% of the children 5-17 years old are engaged in activities considered as child labour. It is required to determine the sample size of a household-based survey that would estimate the child labour rate within a margin of error of 2 percentage points. It is assumed that the design effect of the survey is 4, and the response rate is estimated at about 90%. Furthermore, based on the latest population and housing census, it is known that there were on average 1 child 5 to 17 years old per household at the time of the census.

Using these parameters (σ^2=0.16*(1-0.16), *deff*=4, *ME*=0.02, *AveHH*=1, and *RR*=0.9) in the sample size formulae gives the following result,

$$n = \frac{4 \times 0.16 \times (1-0.16) \times 4}{(0.02)^2 \times 1 \times 0.9} = 5'973 \qquad households$$

Thus 5’973 sample households are required for the child labour survey to estimate the child labour rate with a 2 percentage-points margin of error. If the specified precision of the estimate were higher (for example, a margin of error of 1 percentage-point), the required sample size would have been four times larger n = 23’893 households.

In general, the higher the specified precision of the estimate, or the higher the variability of the indicator to be estimated, the larger is the required sample size. Similarly, the higher the design effect, the lower the response rate, and the smaller the average size of the base population per household, the larger is the required sample size to achieve the same degree of precision of the estimate.

In practice, the choice of the sample size is also determined on the basis of the available resources for the survey and the statistical infrastructure of the country. The sample must of course be large enough to yield information with sufficient sampling precision to be useful to the various types of analysis of the results. However, the choice of inappropriately large sample sizes can adversely affect the overall quality of data. The desire to produce too many breakdowns in too much detail or over-emphasis on sampling precision and neglect of the needs to control non-sampling errors often lead to inappropriately large sample sizes. A good final advice is moderation in the choice of sample size.

3. Margin of error determination

Margin of error determination is the reverse problem of sample size determination. It is relevant, for example, when the child labour survey is conducted as a module attached to a broader household-based survey such as a labour force survey. In such situations, the sample size of the broader survey determines the sample size of the child labour module. The problem then is to calculate the margin of error of child labour estimates for the given sample size of the broader survey.

The reverse formulae for the calculation of margin of errors for a given sample size is given by

$$ME = \sqrt{\frac{4 \times \sigma^2 \times deff}{n \times AveHH \times RR}}$$

The corresponding formulae for the calculation of the relative margin of errors is given by

$$RME = \sqrt{\frac{4 \times (\sigma / y)^2 \times deff}{n \times AveHH \times RR}}$$

Consider the following numerical example. A child labour survey is conducted as part of a module attached to a labour force survey. The sample size of the labour force survey is 5'000 households. What would be the margin of error of the estimate of the child labour rate? The child labour rate is the percentage of children 5 to 17 years old engaged in activities considered as child labour.

Assuming that the child labour module has the same design effect (4) and the same response rate (90%) as the labour force survey, the formulae for the calculation of the margin error shown above gives,

$$ME = \sqrt{\frac{4 \times 0.16 \times (1-0.16) \times 4}{5000 \times 0.75 \times 0.90}} = 2.52 \qquad ppts$$

where in the numerator the variance of the main indicator is replaced by σ^2 = r*(1-r) with r, the percentage of children engaged in child labour, predicted to be around 16%, and in the denominator the parameter *AveHH*, the average number of children 5-17 years old per household, set at 0.75, the ratio of the number of children 5-17 years old (60'000) to the total number of households in the country (80'000).

The relative margin of error is similarly calculated,

$$RME = \sqrt{\frac{4 \times (1-0.16)/0.16 \times 4}{5000 \times 0.75 \times 0.90}} = 15.78\%$$

Using the relationship between the relative and absolute margin of error RME=ME/y, it can be verified that 15.78% = 2.52/0.16 in rounded figures,

The margin of error of an estimate is half the standard error of the estimate. The standard error of an estimate measures the difference between the estimate and the average value that would have been obtained by repeated sampling of the population under otherwise identical conditions. The standard error of an estimate is the square root of the variance of the estimate.

An important use of the standard error is for the calculation of confidence intervals. Under certain broad assumptions, it can be stated that the true value of a variable of interest lies in between the survey estimate and a multiple of the standard error, with certain degree of probability. In general, if y represents the survey estimate of a variable of interest, the true value of the variable represented say by θ lies with (1-α)% confidence in the following interval,

$$y - 2 \times se \leq \theta \leq y + 2 \times se$$

where *se* is the standard error of the estimate and 2 is the approximate value of the standard normal distribution corresponding to the (1-α)% confidence probability. For a 95% confidence probability, α=5% and the corresponding standard normal distribution value is about 1.96, approximated here at 2.

With respect to the numerical illustration given earlier, it can be stated that the estimate of the child labour rate is within the following confidence interval at 95% level,

$$0.16 - 2 \times (0.0252/2) \leq \theta \leq 0.16 + 2 \times (0.0252/2)$$

$$0.1348 \leq \theta \leq 0.1852$$

$$13.48\% \leq \theta \leq 18.52\%$$

Another use of the standard error is for determining the statistical significance of differences in survey estimates. An approximate significance test of the difference between two estimates may be done by simply comparing the confidence intervals of the estimates and checking their overlap. If they do not overlap, the two estimates may be said to be significantly different. For example, the child labour rate among boys may be compared with the corresponding rate among girls, to conclude whether the two rates are significantly different or not.

4. Template user instructions

The Excel template for calculating the sample size is divided into four parts: input values, output values, additional input values, and additional output values. Each part is described below with numerical examples.

• Input and output values

The standard table of input and output values is presented in Diagram 1. There are five standard input values shown in the left panel of the diagram:

- Predicted value of the main indicator (say the child labour rate), assumed here to be 10% (r=0.1)
- Design effect set here at 4 (*deff*=4)
- Margin of error of the survey estimate of the main indicator (child labour rate), specified here at 3% (*ME*=0.03)
- Average number of persons in the base population (children 5-17 years old) per household derived from external sources and set here to be 3 children 5=17 years old for every 4 households (*AveHH*=0.75)
- Response rate from past experience, set here at 90% (*RR*=0.9).

Diagram 1. Standard table of input and output values in the calculation of the sample size for one domain

INPUT VALUES		
Parameter		**Value**
Predicted value of main indicator	*r*	**0.1**
Standard deviation of underlying variable	*σ*	
Design effect	*deff*	**4**
Intraclass correlation	*rho*	
Number of households per cluster	*b*	
Margin of error at 95% confidence	*ME*	**0.03**
	RME	
Average No. of persons of base population per HH	*AveHH*	**0.75**
Average household size	*AveSize*	
Proportion of base population in total population	*pb*	
Response rate	*RR*	**0.9**
Sample size (number of households)	*n*	

OUTPUT VALUES		
Estimate		**Value**
Sample size (number of households)	*n*	**2370**
Standard deviation of underlying variable	*σ*	**0.3**
Design effect	*deff*	**4**
Intraclass correlation	*rho*	**0.27**
Standard error of estimate	*se*	**0.015**
Margin of error at 95% confidence	*ME*	**0.03**
	RME	**0.30**
	AveHH	**0.75**
Confidence limits (at 95% confidence)	*Lower*	**0.07**
	Upper	**0.13**
	RR	**0.9**

The output values are shown in the right panel of the diagram. They are calculated on the basis of the input values and recorded in bold red color. The Excel contents of cells marked in red should be changed.

- Sample size. Calculated based on the parameters specified as input values (here n = 2370 sample households).
- Standard deviation of the underlying variable, here a dichotomous indicator variable (y=1 if child in child labour and y=0 otherwise). If the value of the standard deviation is not given as part of the input values, it is calculated by σ=sqrt(r*(1-r)) with r=0.1 (here σ=0.3).
- Design effect. Value transferred directly from the input values (here *deff*=4). If not specified, it is calculated on the basis of the intra-class correlation (*rho*) and other parameters of the input values.
- Intra-class correlation. Value either transferred directly from the input values if specified, or calculated on the basis of the design effect and other parameters of the input values (here, *rho*=(*deff*-1)/(*b*AveHH*-1)= 0.27).
- Standard error of estimate. Half of the margin of error.
- Margin of error at 95% confidence. Value either transferred directly from the input values if specified (here *ME*=0.03), or calculated on the basis of *RME*, the relative margin of error (*ME=r*RME*).
- Average number of persons in base population per household. . Value either transferred directly from the input values if specified (here *AveHH*= 0.75), or calculated on the basis of the average household size and the proportion of base population in total population (*AveHH=pb*AveSize*).
- Confidence limits (at 95% confidence). Calculated from the predicted value of the main indicator and the relative margin of error of the survey estimate (here Lower=r*(1-RME)=7% and Upper=r*(1+RME)=13%).
- Finally, response rate. Value transferred directly from the input values (here *RR*= 0.9). If not specified, it is set at 1, that is 100% response.

Alternative sets of parameters to be specified as input values in the template. For example, the standard deviation of the underlying indicator may be specified directly as part of the input values, instead of being derived as part of the output values. This feature is particularly relevant when the main indicator is an amount rather than a percentage, e.g. total income from employment. A numerical example is shown in Diagram 1b below.

Diagram 1b. Input and output values for sample size determination when main indicator is an amount rather than a percentage

INPUT VALUES		
Parameter		**Value**
Predicted value of main indicator	*r*	**400**
Standard deviation of underlying variable	*σ*	**250**
Design effect	*deff*	**4**
Intraclass correlation	*rho*	
Number of households per cluster	*b*	
Margin of error at 95% confidence	*ME*	**20**
	RME	
Average No. of persons in base population per HH	*AveHH*	**0.75**
Average household size	*AveSize*	
Proportion of base population in total population	*pb*	
Response rate	*RR*	**0.9**
Sample size (number of households)	*n*	

OUTPUT VALUES		
Estimate		**Value**
Sample size (number of households)	*n*	**3704**
Standard deviation of underlying variable	*σ*	**250**
Design effect	*deff*	**4**
Intraclass correlation	*rho*	**27.3%**
Standard error of estimate	*se*	**10.000**
Margin of error at 95% confidence	*ME*	**20.00**
	RME	**5.0%**
	AveHH	**0.75**
Confidence limits (at 95% confidence)	*Lower*	**380.00**
	Upper	**420.00**
	RR	**90.0%**

In Diagram 1b the underlying indicator is the income from employment of a child with standard deviation set at σ=250. The specified margin of error of the survey estimate is set at 20 and the predicted total income from employment at 400. Other parameters have been left unchanged. The new output values are given in the left panel of the diagram.

Similarly, the following three diagrams (Diagram 1c, 1d and 1e) show the numerical examples of input and output values under other alternative sets of parameter specification. The table details are self-explanatory.

Diagram 1c. Intra-class correlation (rho) rather than design effect (deff)

INPUT VALUES		
Parameter		**Value**
Predicted value of main indicator	*r*	**0.1**
Standard deviation of underlying variable	σ	
Design effect	*deff*	
Intraclass correlation	*rho*	**0.25**
Number of households per cluster	*b*	**16**
Margin of error at 95% confidence	*ME*	**0.03**
	RME	
Average No. of persons in base population per HH	*AveHH*	**0.75**
Average household size	*AveSize*	
Proportion of base population in total population	*pb*	
Response rate	*RR*	**0.9**
Sample size (number of households)	*n*	

OUTPUT VALUES		
Estimate		**Value**
Sample size (number of households)	*n*	**2222**
Standard deviation of underlying variable	σ	**0.3**
Design effect	*deff*	**3.75**
Intraclass correlation	*rho*	**25.0%**
Standard error of estimate	*se*	**0.015**
Margin of error at 95% confidence	*ME*	**0.03**
	RME	**0.30**
	AveHH	**0.75**
Confidence limits (at 95% confidence)	*Lower*	**0.07**
	Upper	**0.13**
	RR	**90.0%**

Diagram 1d. Relative (RME) rather than absolute margin of error (ME)

INPUT VALUES		
Parameter		**Value**
Predicted value of main indicator	*r*	**0.1**
Standard deviation of underlying variable	σ	
Design effect	*deff*	**4**
Intraclass correlation	*rho*	
Number of households per cluster	*b*	
Margin of error at 95% confidence	*ME*	
	RME	**0.2**
Average No. of persons in base population per HH	*AveHH*	**0.75**
Average household size	*AveSize*	
Proportion of base population in total population	*pb*	
Response rate	*RR*	**0.9**
Sample size (number of households)	*n*	

OUTPUT VALUES		
Estimate		**Value**
Sample size (number of households)	*n*	**5333**
Standard deviation of underlying variable	σ	**0.3**
Design effect	*deff*	**4**
Intraclass correlation	*rho*	**27.3%**
Standard error of estimate	*se*	**0.000**
Margin of error at 95% confidence	*ME*	**0.02**
	RME	**0.20**
	AveHH	**0.75**
Confidence limits (at 95% confidence)	*Lower*	**0.08**
	Upper	**0.12**
	RR	**90.0%**

Diagram 1e. AveSize and pb rather than AveHH

INPUT VALUES		
Parameter		**Value**
Predicted value of main indicator	*r*	**0.1**
Standard deviation of underlying variable	σ	
Design effect	*deff*	**4**
Intraclass correlation	*rho*	
Number of households per cluster	*b*	
Margin of error at 95% confidence	*ME*	**0.03**
	RME	
Average No. of persons in base population per HH	*AveHH*	
Average household size	*AveSize*	**5**
Proportion of base population in total population	*pb*	**0.15**
Response rate	*RR*	**0.9**
Sample size (number of households)	*n*	

OUTPUT VALUES		
Estimate		**Value**
Sample size (number of households)	*n*	**2370**
Standard deviation of underlying variable	σ	**0.3**
Design effect	*deff*	**4**
Intraclass correlation	*rho*	**27.3%**
Standard error of estimate	*se*	**0.015**
Margin of error at 95% confidence	*ME*	**0.03**
	RME	**0.30**
	AveHH	**0.75**
Confidence limits (at 95% confidence)	*Lower*	**0.07**
	Upper	**0.13**
	RR	**90.0%**