NACIONES UNIDAS

CEPAL

Quality criteria for the publication of estimates from Household Surveys

Andrés Gutiérrez
Statistics Division
ECLAC

Always using the CVE?

What is the coefficient of variation?

The coefficient of variation is a measure of error relative to an estimator, it is defined as:

$$cve(\hat{\theta}) = \frac{se(\hat{\theta})}{\hat{\theta}}$$

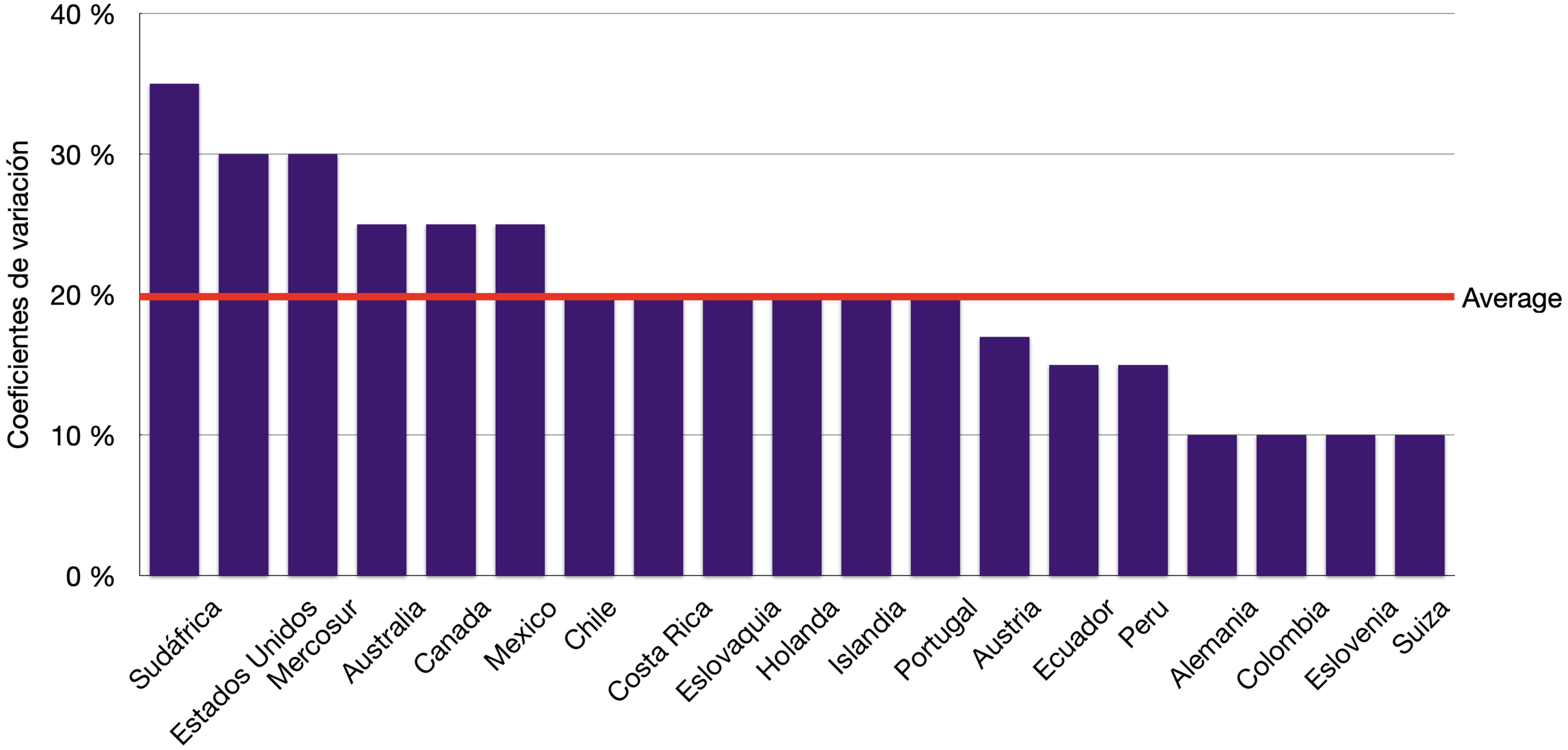
Many times it is expressed as a percentage, although it is not bounded to the right, and for this reason it is convenient when talking about the precision of a statistic that comes from a survey.

Use of the coefficient of variation

Sarndal et. al. (2003) state that *a statistician can express his opinion that a value of the coefficient of variation of 2% is* ***good*** *, considering the restrictions of the survey, while a value of the coefficient of variation of 9% can be considered* ***unacceptable***.

In this way, many national statistical institutes around the world have considered that the precision of the statistics resulting from a survey is subject to the behavior of its coefficient of variation.

Disclaimer thresholds per country (household surveys)



Some disclaimers

When the coefficient of variation threshold is exceeded, some of the following alerts appear:

- **not published**
- Use with caution.
- Coefficients of variation greater than 20% mean that the estimate is subject to high variability, which limits its scope for analytical purposes.
- Estimates require revision, are not precise, and should be used with caution.
- Unreliable, less accurate. Use with care.
- Does not meet publication standards.
- With reserve, referential, questionable.
- Very random values, poor estimate.

Review of other standards for disaggregating estimates

Precision of the estimators

As a survey is a partial observation over a subset in the finite population, it is necessary to know that:

- From a survey, indicators are not **calculated, but are estimated** with the help of the survey data.

- It is necessary to calculate the degree of error that is committed by not being able to carry out an exhaustive observation. This error is known as the **sampling error** .

- The precision of an estimator is highly related to the **confidence interval** . The narrower the interval, the more precision is generated and therefore there is less sampling error.

The confidence interval in subpopulations

If the parameter of interest is θ_d , and a subpopulation of interest U_d has been defined, then a 95% confidence interval on that subpopulation is given by the following expression:

$$(\hat{\theta}_d - t_{0.975,gl} * se(\hat{\theta}_d),\ \ \hat{\theta}_d + t_{0.975,gl} * se(\hat{\theta}_d))$$

The use of the coefficient of variation as an indicator of the reliability of the statistics from household surveys should be complemented with some other measures that allow the creation of reliability and precision rules.

Confidence interval

Note that the length of the confidence intervals induces confidence that an estimator is accurate:

- The incidence of poverty in the department of the country was estimated at 5.2%, with a confidence interval of *(5.15%, 5.25%)* .
- The unemployment rate in the country for men was 7.5%, with a confidence interval of *(7.1%, 7.9%)* ; while for women it was 9.2%, with a confidence interval of *(8.8%, 9.6%)* .
- The net student attendance rate in primary school for the last income quintile was estimated at 85%, with a confidence interval of *(48.2%, 100.0%)*

A. Sample size

- The sample size directly affects the width of the confidence interval, through the standard error, which generally decreases as the sample size gets larger.

- An adequate sample size guarantees the distribution convergence of the estimators to the theoretical distribution from which the percentiles are calculated.

- For example, it is possible to suggest that all estimates based on a sample size less than a predefined threshold should be suppressed or marked as unreliable.

B. The effective sample size

- In household surveys, with complex sampling designs, **there is no** succession of independent and identically distributed variables.

- The sample $(y_1, \ldots, y_n)$ is not a vector in n-dimensional space, where it is assumed that each component of the vector can vary by itself.

- The final dimension of the sample is much smaller than n, due to the hierarchical selection of households within PSU and the intra-class correlation with the variable of interest.

B. The effective sample size

The effective sample size is defined as follows:

$$n_{efectivo} = \frac{n}{Deff}$$

Where *Deff* is the design effect that depends on:

1. The average number of households selected within the PSU.
2. The existing correlation between the variable of interest and the PSU themselves.

It is possible to consider that if the effective sample size is not greater than a predefined threshold, then the estimate should not be considered for publication.

C. Degrees of freedom

They are a measure of how many independent units of information are in the inference. Note that:

- In the extreme case of carrying out a census in each PSU, regardless of the cluster size, the number of independent units will only be the number of PSU selected in the first sampling stage.

- In household surveys, the variability of the estimate is mainly due to the cluster contribution in the first stage while the second sampling stage contribution to the variance is (many times) considered negligible.

C. Degrees of freedom

Regarding subpopulations, the degrees of freedom are not considered fixed but variable.

$$gl_{sub} = \sum_{h=1}^{H} v_h * (n_{Ih} - 1)$$

Note that v_h is an indicator variable that takes the value one if stratum h contains one or more cases of the subpopulation of interest, n_{Ih} is the number of PSU in the stratum. In the most general case, the degrees of freedom are reduced to the following expression:

$$gl = \#Strata - \#PSU$$

C. Degrees of freedom

Consider the 0.975th percentile for which the critical values of the *t -distribution* vary with respect to their degrees of freedom.

- t -student $_{df = 1}$ = 12.7
- t -student $_{df = 2}$ = 4.30
- t -student $_{df = 40}$ = 2.02
- t -student $_{df = oo}$ = Z = 1.96

It is possible to consider that if the degrees of freedom induced by the subpopulation are less than a predefined threshold, the figure should be suppressed.

What happens when the software estimate a CV = 0 with one degree of freedom?

D. Logarithmic coefficient of variation

- For a proportion, the coefficient of variation is not symmetric around P=0.5. This implies a possible contradiction, since for a dichotomous variable it is possible to have at the same time coefficients of variation that lead to different conclusions.

- This measure is defined as a logarithmic transformation on the proportion and prevents estimates close to zero from being penalized with a high coefficient of variation even when their variation is small.

$$\hat{L} = -\log(\hat{P}) \quad \Rightarrow \quad CV(\hat{L}) = \frac{SE(\hat{L})}{\hat{L}} = \frac{CV(\hat{P})}{\hat{L}}$$

E. Unweighted Count of Cases

- It is determined by the number of cases in the sample affected by the phenomenon of interest (without considering the expansion factor or the complex design of the survey).

$$n_y = \sum_s \delta_k^y$$

Standards: ECLAC and NSOs

Thresholds in ECLAC (2020)

- Coefficient of variation > 20%.
- Logarithmic coefficient of variation > 17.5%.
- Unweighted count fo cases < 50.
- Degrees of freedom > 8.
- Sample size < 100.
- Effective sample size < 68.

Example: thresholds in an ONS

- Degrees of freedom > 14: imply at least 15 PSU , which guarantee the beginning of the convergence in distribution (for linearized estimators: means, proportions, ratios).
- Sample size < 150: if a minimum of 15 PSU with a subsample of 10 households is expected.
- Effective sample size < 60: taking into account a design effect of 2.5 (mean of the survey), an effective n of 150/2.5 = 60 would be obtained.
- Logarithmic coefficient of variation > 18.5%, since to obtain a sample size of 150, with a proportion p = 0.5, the minimum threshold of the logarithmic coefficient of variation must be 0.185.
- Unweighted case count < 40.
- Coefficient of Variation > 30%.

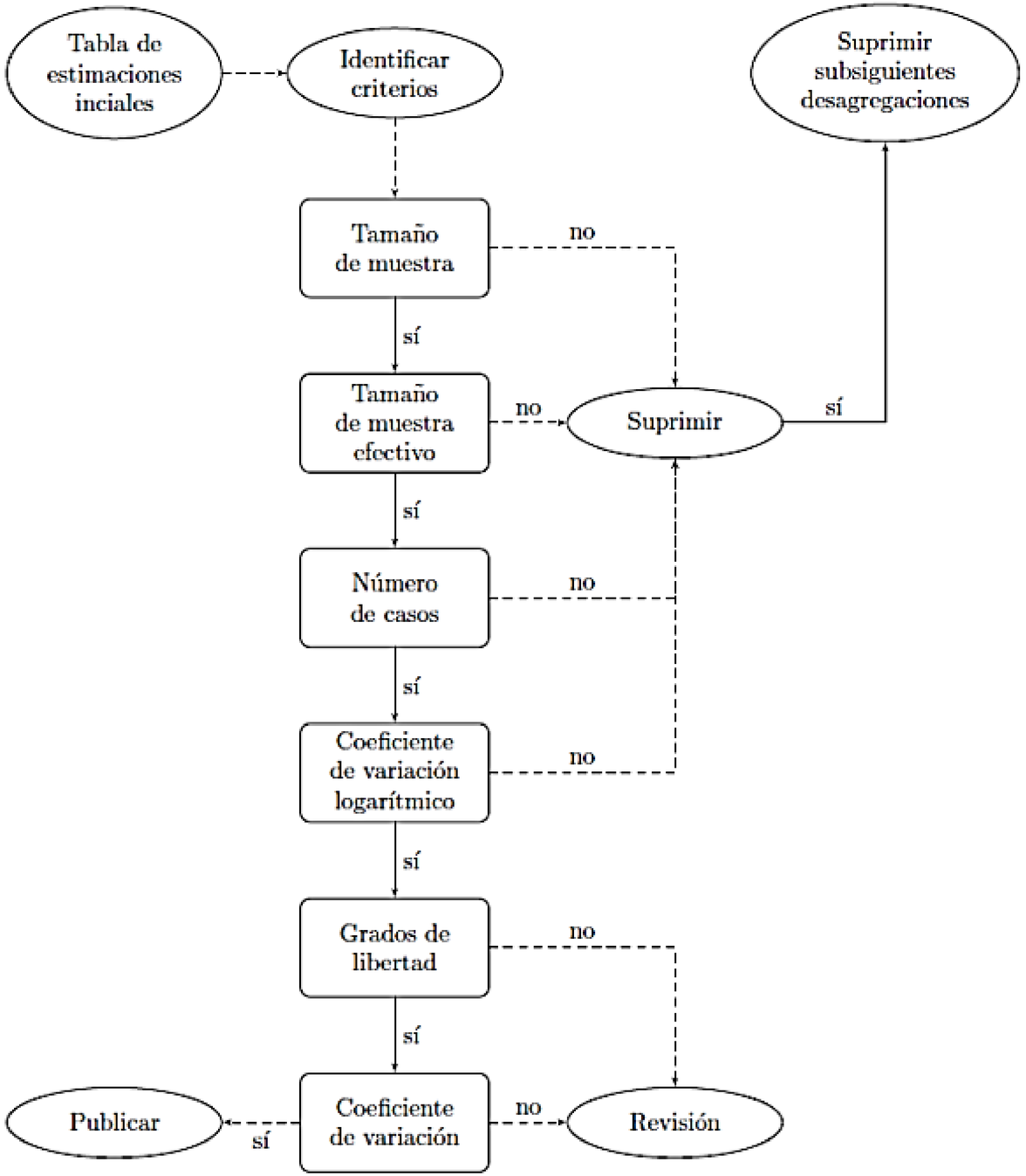


Tabla de estimaciones inciales
Identificar criterios
Suprimir subsiguientes desagregaciones
Tamaño de muestra
no
sí
Tamaño de muestra efectivo
no
Suprimir
sí
sí
Número de casos
no
sí
Coeficiente de variación logarítmico
no
sí
Grados de libertad
no
sí
Publicar
sí
Coeficiente de variación
no
Revisión

NACIONES UNIDAS

CEPAL

Thank you!

andres.gutierrez@un.org