Pacific Training on Sampling Methods for Producing Core Data Items for Agricultural and Rural Statistics

13-17 August, Suva, Fiji

Module 2: Review of Basics of Sampling Methods
Session 2.7: Standard Errors & Bias

By
Chris Ryan
Statistician (ESCAP Pacific Office)

UNITED NATIONS SIAP Statistical Institute for Asia and the Pacific

FAO FIAT PANIS

UNITED NATIONS ESCAP Economic and Social Commission for Asia and the Pacific

Topics Covered

- Things which contribute to the error of an estimate
  - Sampling error
  - Non-sampling error
- Examples of Sampling error formulas
- What is Bias?
  - Impact on estimates

How accurate are the estimates

* There are two main sources of error which impact on the quality of our estimates
  * Sampling error
  * Non-sampling error

* Sampling error – the error attributed to the fact that a sample of units were selected for the survey as opposed to a census
* Non-sampling error – all other errors associated with the results

Non-sampling errors

* Some common forms of non-sampling error include:
  * Field enumeration error
  * Respondent error
  * Questionnaire design problems
  * Data processing errors
  * Sample selection and response bias
  * etc

Sampling error

- As discussed, this is the error in the estimates, generated by taking a sample of units as opposed to complete enumeration
- Unlike the non-sampling error, which is extremely difficult to measure, we can estimate a measure for the magnitude of the sampling error if we know things like:
  - Population size
  - Sample size
  - Sample selection methodology
  - Degree of variation in the response data

Sampling Error (cont)

- In order to determine what the sampling error (or standard error) is for an estimate, we first need to calculate the variance of the estimate
- The standard error is then simply the square root of the variance

$$SE(Y) = \sqrt{Var(Y)}$$

Sample Error for a Simple Random Sample

$$Var(\hat{Y}) = \frac{N^2}{n}(1 - n/N)s_y^2$$

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n - 1}$$

Sampling error

* The formula for the variance of a two-stage survey design involving PPS sampling at the first stage and a fixed cluster size of households at the second stage is:

$$\text{Var}(Y) = \frac{M^2}{m}\overline{N}\sigma_b^2 + \frac{N}{n}\sum_{i=1}^{M} N_i\left(1 - \frac{\overline{n}}{N_i}\right)S_i^2$$

Link between Estimate, Sampling Error and Relative Sampling Error

$$SE(\hat{Y}) = SQRT\ (Var(\hat{Y}))$$

$$RSE(\hat{Y}) = \frac{SE(\hat{Y})}{Est(\hat{Y})}\ x\ 100$$

Interpreting a RSE

* A user will often want to know how to interpret the accuracy of an estimate, with respect to the RSE
* If RSE < 5% It's a reliable estimate
* If 5% < RSE < 10% It's still good
* If 10% < RSE 20% It's usable
* If RSE > 20% Not overly reliable

Bias

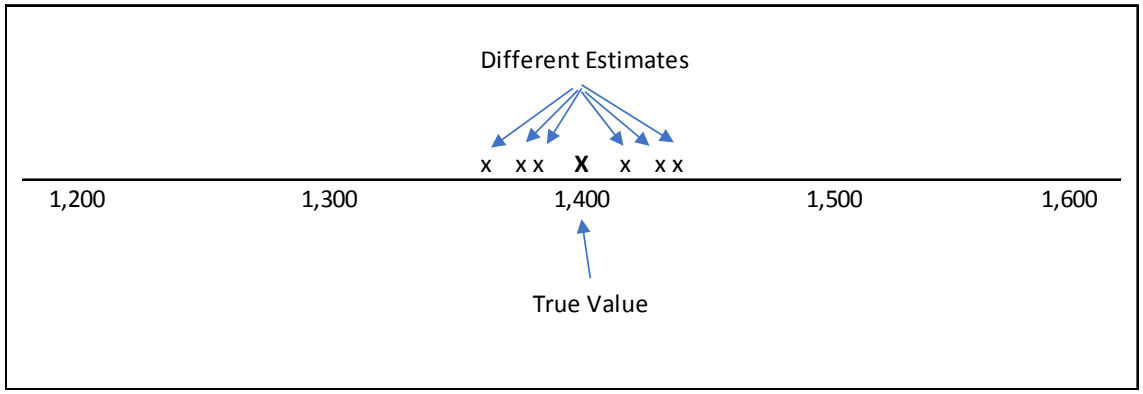
Definition

* Bias is occuring if you produce numerous values of the same estimate, and you repeatedly come up with something higher (or lower) than the true value

* Two main things often contribute to a bias being generated in a survey
  * Poor sample selection
  * Non-response

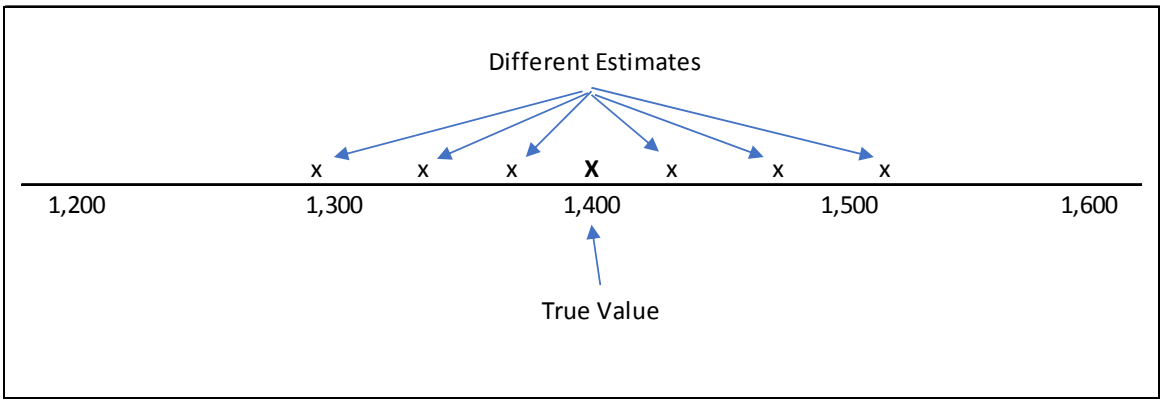
What happens to the estimates?

Scenario 1: Small sample error & small bias



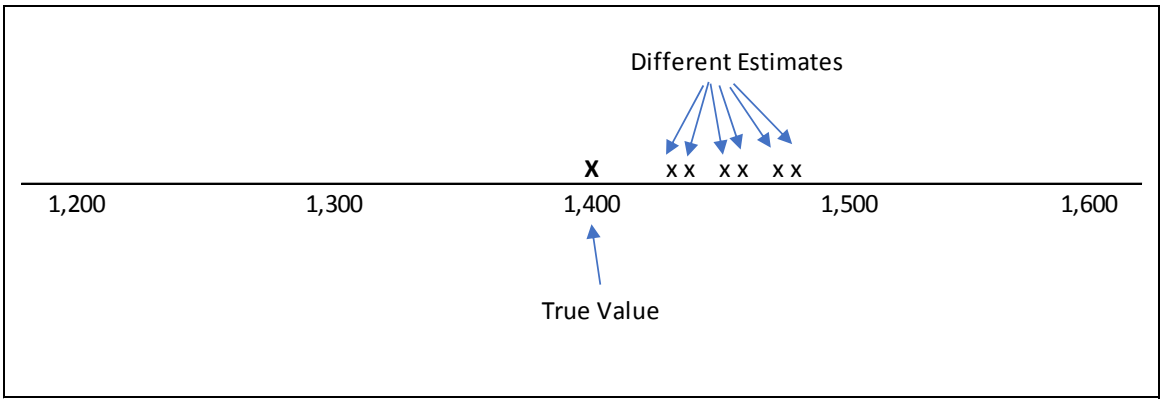
What happens to the estimates?

Scenario 2: Large sample error & small bias



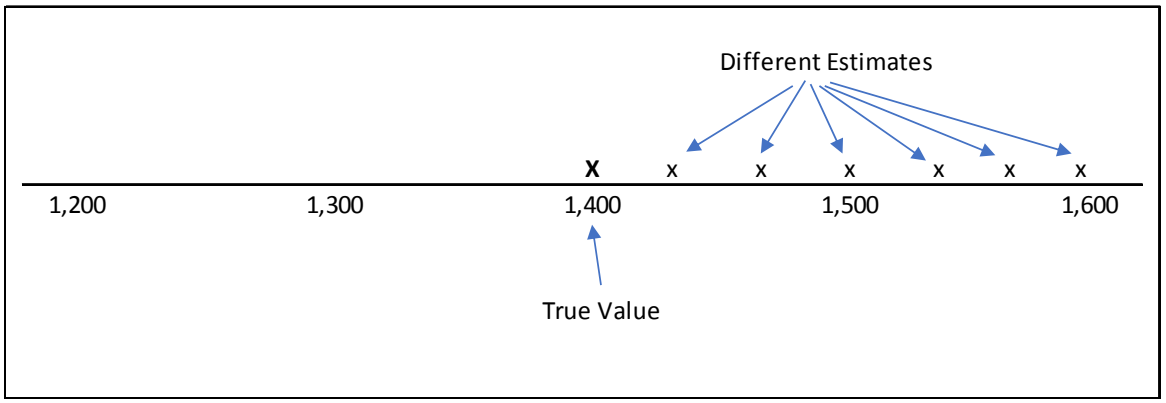
What happens to the estimates?

Scenario 3: Small sample error & large bias



What happens to the estimates?

Scenario 4: Large sample error & Large bias



An example of selection bias

* Suppose the Department of Fisheries want to run a household survey to determine fish catches by non-commercial fisherman in Samoa
  * The question they ask relates to the fish catches over the last week
* Also suppose that Samoa is made up of:
  * 150 coastal villages
  * 150 inland villages

An example of selection bias (cont)

* The Department of Fisheries decide to select only coastal villages because they believe they make-up most of the non-commercial fishing activity

* Let us assume the following:
  * Average fish catches by households living in coastal villages for the last week was: *70kg*
  * Average fish catches by households living in inland villages for the last week was: *30kg*

An example of selection bias (cont)

* The population and sample counts for the two regions are:

Coastal villages
* Population: 10,000 households
* Sample: 1,000 households

Inland villages
* Population: 10,000 households
* Sample: 0 households

An example of selection bias (cont)

The true value of Fish catch

True Value (Fish catch) = (10,000 x 70) + (10,000 x 30)
= 700,000 + 300,000
= 1,000,000

The estimation procedure would be

If the DoF weighted up to the total population

Est(Fish catch) = N/n x fish catch in sample
= 20,000/1,000 * (70 x 1,000)
= 1,400,000