bias

Canonical name	Bias
Date of creation	2013-03-22 15:00:21
Last modified on	2013-03-22 15:00:21
Owner	CWoo (3771)
Last modified by	CWoo (3771)
Numerical id	8
Author	CWoo (3771)
Entry type	Definition
Classification	msc 62A01
Synonym	systematic error
Defines	unbiased estimator
Defines	biased estimator
Defines	asymptotically unbiased estimator

Background. In estimating a parameter from a statistical model, one is interested in how the estimates deviate from the true value of the parameter. The deviations generally come from two sources. One source is known as the *noise*, or *random error*, which has to do with the random nature of observations or measurements in general. For example, when a fair coin is tossed 100 times and the number of heads is counted. One might get 51 even though the true parameter is 50. The difference of 1 is due to the random nature of coin tossing.

The other source of deviation is known as the *bias*, or *systematic error*, which has to do with how the observations are made, how the instruments are set up to make the measurements, and most of all, how these observations or measurements are tallied and summarized to come up with an estimate of the true parameter. For example, a rating scheme is proposed for an online collaborative encyclopedia on entries contributed by individuals who are members of the online website hosting the encyclopedia. The purpose of this rating scheme is to give the readers, members or non-members inclusive, a better idea on the quality of the entries by their corresponding numerical values. Suppose that members are asked to rate an entry from a scale of 1 to 10. For simplicity, members who are intimately familiar with the concept in the entry rate it with a perfect 10. Next, members who are not that familiar with the entry give it a 5. Finally, the remaining members choose to not participate and the rating scale from them default to a 0. A simple arithmetic average is computed and a rating of 2.5 is produced. Would this 2.5 be a good indicator of the overall quality of the entry? Maybe not. Here, biases are introduced. First, the participants of the rating scheme do not include non-members, who, collectively, may very well represent a different level of understanding of the rated entry than members. Secondly, even among the members, there is a considerable amount of differences in terms of levels of understanding of the entry, etc... To some, the entry may be accurately and perfectly written, not everyone will rate it the same way in the end. Finally, there are the non-raters. We have no idea as to how they would rate the entries. Their votes should certainly count if they decide to rate in the last minute. The final rate, however, would most likely be different.

The difference between the bias and the noise is that one can be reduced while the other can not. Mathematically, we have the following:

Definition. If θ is a parameter in a statistical model, the bias of an estimator $\hat{\theta}$ of θ, is the difference between expectation of $\hat{\theta}$ and the value of

θ, which, by abuse of notation, is also denoted θ:

$$\mathrm{Bias}(\hat{\theta}) := \mathrm{E}[\hat{\theta}] - \theta.$$

An estimator is called an *unbiased estimator* if its bias is zero at *all* values of θ. Otherwise, it is a *biased estimator*.

Note that the random error does not appear in the above definition because its expectation is zero.

Examples.

1. If observations $X_1, \ldots, X_n$ are iid from a normal distribution with mean μ and variance σ^2, then the sample mean estimator $\overline{X}$ is an unbiased estimator for μ. To see this, recall the definition of a sample mean

   $$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

   so that

   $$\mathrm{E}[\overline{X}] = \frac{1}{n}(\mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n]).$$

   But $\mu = \mathrm{E}[X_1] = \cdots = \mathrm{E}[X_n]$, the above expression reduces to

   $$\mathrm{E}[\overline{X}] = \frac{1}{n}(n\mu) = \mu,$$

   showing that the bias of $\overline{X}$ is zero. Note that even though $\overline{X}$ depends on the size of the sample n, its expectation, however, does not, and is identically μ, for all values of μ.

2. Here is another example of an unbiased estimator. Again, assume observations X_i, $i = 1, \ldots, n$ are iid as normal distribution $N(\mu, \sigma^2)$. The sample variance estimator s^2 is defined by

   $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

   Expressing s^2 explicitly in terms of the random variables X_i, we have

   $$\begin{aligned}(n-1)s^2 &= \sum_{i=1}^{n} X_i^2 - \frac{1}{n}(\sum_{i=1}^{n} X_i)^2 & (1)\\ &= \frac{1}{n}[(n-1)\sum_{i=1}^{n} X_i^2 - 2\sum_{i<j} X_i X_j] & (2)\end{aligned}$$

Now, for $i \neq j$, X_i and X_j are independent so that

$$\mathrm{E}[X_i X_j] = \mathrm{E}[X_i]\,\mathrm{E}[X_j] = \mu^2 = \mathrm{E}[X_i]^2,$$

for all $i = 1, \ldots, n$. Hence

$$\begin{aligned}
(n-1)\,\mathrm{E}[s^2] &= \frac{1}{n}\{(n-1)\sum_{i=1}^{n}\mathrm{E}[X_i^2] - 2\sum_{i<j}\mathrm{E}[X_i X_j]\} && (3)\\
&= \frac{1}{n}\{(n-1)\sum_{i=1}^{n}\mathrm{E}[X_i^2] - 2\sum_{i<j}\mu^2\} && (4)\\
&= \frac{1}{n}\{(n-1)\sum_{i=1}^{n}\mathrm{E}[X_i^2] - 2\cdot\frac{n(n-1)}{2}\mu^2\} && (5)\\
&= \frac{n-1}{n}\sum_{i=1}^{n}\{\mathrm{E}[X_i^2] - \mu^2\} && (6)\\
&= \frac{n-1}{n}\sum_{i=1}^{n}\{\mathrm{E}[X_i^2] - \mathrm{E}[X_i]^2\} && (7)\\
&= \frac{n-1}{n}\sum_{i=1}^{n}\mathrm{Var}[X_i] = (n-1)\sigma^2. && (8)
\end{aligned}$$

This shows that s^2 is an unbiased estimator for σ^2.

3. However, s^2 would be biased if we were to define it by

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

since

$$\mathrm{E}[s^2] = \frac{n-1}{n}\sigma^2$$

would depend on the sample size n and would not equal to σ^2 at any n.

Remark. In practice, unbiased estimators are rare. There is another, larger class of estimators that are biased with smaller samples, but the bias gets smaller and tends to 0 as the sample size gets larger. Such an estimator is called an *asymptotically unbiased estimator*. For example, if we were to define s^2 as in Example 3 above, s^2 would be an asymptotically unbiased estimator for σ^2.