Bias of an estimator

From Wikipedia, the free encyclopedia

In statistics, the **bias** (or **bias function**) of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated. An estimator or decision rule with zero bias is called **unbiased**. Otherwise the estimator is said to be **biased**. In statistics, "bias" is an objective statement about a function, and while not a desired property, it is not pejorative, unlike the ordinary English use of the term "bias".

Bias can also be measured with respect to the median, rather than the mean (expected value), in which case one distinguishes *median*-unbiased from the usual *mean*-unbiasedness property. Bias is related to consistency in that consistent estimators are convergent and *asymptotically* unbiased (hence converge to the correct value), though individual estimators in a consistent sequence may be biased (so long as the bias converges to zero); see bias versus consistency.

All else equal, an unbiased estimator is preferable to a biased estimator, but in practice all else is not equal, and biased estimators are frequently used, generally with small bias. When a biased estimator is used, the bias is also estimated. A biased estimator may be used for various reasons: because an unbiased estimator does not exist without further assumptions about a population or is difficult to compute (as in unbiased estimation of standard deviation); because an estimator is median-unbiased but not mean-unbiased (or the reverse); because a biased estimator reduces some loss function (particularly mean squared error) compared with unbiased estimators (notably in shrinkage estimators); or because in some cases being unbiased is too strong a condition, and the only unbiased estimators are not useful. Further, mean-unbiasedness is not preserved under non-linear transformations, though median-unbiasedness is (see effect of transformations); for example, the sample variance is an unbiased estimator for the population variance, but its square root, the sample standard deviation, is a biased estimator for the population standard deviation. These are all illustrated below.

Contents

Definition

Suppose we have a statistical model, parameterized by a real number θ, giving rise to a probability distribution for observed data, $P_\theta(x) = P(x \mid \theta)$, and a statistic $\hat{\theta}$ which serves as an estimator of θ based on any observed data x. That is, we assume that our data follow some unknown distribution $P_\theta(x) = P(x \mid \theta)$ (where θ is a fixed constant that is part of this distribution, but is unknown), and then we construct some estimator $\hat{\theta}$ that maps observed data to values that we hope are close to θ. Then the **bias** of this estimator (relative to the parameter θ) is defined to be

$$\operatorname{Bias}_\theta[\hat{\theta}] = \mathrm{E}_\theta[\hat{\theta}] - \theta = \mathrm{E}_\theta[\hat{\theta} - \theta],$$

where E_θ denotes expected value over the distribution $P_\theta(x) = P(x \mid \theta)$, i.e. averaging over all possible observations x. The second equation follows since θ is measurable with respect to the conditional distribution $P(x \mid \theta)$.

An estimator is said to be **unbiased** if its bias is equal to zero for all values of parameter θ.

In a simulation experiment concerning the properties of an estimator, the bias of the estimator may be assessed using the mean signed difference.

Examples

Sample variance

The sample variance of a random variable demonstrates two aspects of estimator bias: firstly, the naive estimator is biased, which can be corrected by a scale factor; second, the unbiased estimator is not optimal in terms of mean squared error (MSE), which can be minimized by using a different scale factor, resulting in a biased estimator with lower MSE than the unbiased estimator. Concretely, the naive estimator sums the squared deviations and divides by n, which is biased. Dividing instead by $n - 1$ yields an unbiased estimator. Conversely, MSE can be minimized by dividing by a different number (depending on distribution), but this results in a biased estimator. This number is always larger than $n - 1$, so this is known as a shrinkage estimator, as it "shrinks" the unbiased estimator towards zero; for the normal distribution the optimal value is $n + 1$.

Suppose $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables with expectation μ and variance σ^2. If the sample mean and uncorrected sample variance are defined as

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2,$$

then S^2 is a biased estimator of σ^2, because

$$\begin{aligned}\mathrm{E}[S^2] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left((X_i-\mu) - (\overline{X}-\mu)\right)^2\right] \\ &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - 2(\overline{X}-\mu)(X_i-\mu) + (\overline{X}-\mu)^2\right] \\ &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\overline{X}-\mu)^2\right] = \sigma^2 - \mathrm{E}\left[(\overline{X}-\mu)^2\right] < \sigma^2.\end{aligned}$$

In other words, the expected value of the uncorrected sample variance does not equal the population variance σ^2, unless multiplied by a normalization factor. The sample mean, on the other hand, is an unbiased[1] estimator of the population mean μ.

The reason that S^2 is biased stems from the fact that the sample mean is an ordinary least squares (OLS) estimator for μ: $\overline{X}$ is the number that makes the sum $\sum_{i=1}^{n}(X_i - \overline{X})^2$ as small as possible. That is, when any other number is plugged into this sum, the sum can only increase. In particular, the choice $\mu \neq \overline{X}$ gives,

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

and then

$$\mathrm{E}[S^2] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right] < \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] = \sigma^2.$$

Note that the usual definition of sample variance is

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

and this is an unbiased estimator of the population variance. This can be seen by noting the following formula, which follows from the Bienaymé formula, for the term in the inequality for the expectation of the uncorrected sample variance above:

$$\mathrm{E}\left[(\overline{X}-\mu)^2\right] = \frac{1}{n}\sigma^2.$$

The ratio between the biased (uncorrected) and unbiased estimates of the variance is known as Bessel's correction.

Estimating a Poisson probability

A far more extreme case of a biased estimator being better than any unbiased estimator arises from the Poisson distribution.[2][3] Suppose that X has a Poisson distribution with expectation λ. Suppose it is desired to estimate

$$\mathrm{P}(X=0)^2 = e^{-2\lambda}$$

with a sample of size 1. (For example, when incoming calls at a telephone switchboard are modeled as a Poisson process, and λ is the average number of calls per minute, then $e^{-2\lambda}$ is the probability that no calls arrive in the next two minutes.)

Since the expectation of an unbiased estimator δ(X) is equal to the estimand, i.e.

$$E(\delta(X)) = \sum_{x=0}^{\infty} \delta(x)\frac{\lambda^x e^{-\lambda}}{x!} = e^{-2\lambda},$$

the only function of the data constituting an unbiased estimator is

$$\delta(x) = (-1)^x.$$

To see this, note that when decomposing $e^{-\lambda}$ from the above expression for expectation, the sum that is left is a Taylor series expansion of $e^{-\lambda}$ as well, yielding $e^{-\lambda}e^{-\lambda} = e^{-2\lambda}$ (see Characterizations of the exponential function).

If the observed value of X is 100, then the estimate is 1, although the true value of the quantity being estimated is very likely to be near 0, which is the opposite extreme. And, if X is observed to be 101, then the estimate is even more absurd: It is −1, although the quantity being estimated must be positive.

The (biased) maximum likelihood estimator

$$e^{-2X}$$

is far better than this unbiased estimator. Not only is its value always positive but it is also more accurate in the sense that its mean squared error

$$e^{-4\lambda} - 2e^{\lambda(1/e^2-3)} + e^{\lambda(1/e^4-1)}$$

is smaller; compare the unbiased estimator's MSE of

$$1 - e^{-4\lambda}.$$

The MSEs are functions of the true value λ. The bias of the maximum-likelihood estimator is:

$$e^{-2\lambda} - e^{\lambda(1/e^2-1)}.$$

Maximum of a discrete uniform distribution

The bias of maximum-likelihood estimators can be substantial. Consider a case where *n* tickets numbered from 1 through to *n* are placed in a box and one is selected at random, giving a value *X*. If *n* is unknown, then the maximum-likelihood estimator of *n* is *X*, even though the expectation of *X* is only (*n* + 1)/2; we can be certain only that *n* is at least *X* and is probably more. In this case, the natural unbiased estimator is 2*X* − 1.

Median-unbiased estimators

The theory of median-unbiased estimators was revived by George W. Brown (http://www.universityofcalifornia.edu/senate/inmemoriam/georgewbrown.htm) in 1947:[4]

> An estimate of a one-dimensional parameter θ will be said to be median-unbiased, if, for fixed θ, the median of the distribution of the estimate is at the value θ; i.e., the estimate underestimates just as often as it overestimates. This requirement seems for most purposes to accomplish as much as the mean-unbiased requirement and has the additional property that it is invariant under one-to-one transformation.

Further properties of median-unbiased estimators have been noted by Lehmann, Birnbaum, van der Vaart and Pfanzagl. In particular, median-unbiased estimators exist in cases where mean-unbiased and maximum-likelihood estimators do not exist. They are invariant under one-to-one transformations.

There are methods of construction median-unbiased methods for probability distributions that have monotone likelihood-functions, such as one-parameter exponential families, to ensure that they are optimal (in a sense analogous to minimum-variance property considered for mean-unbiased estimators). Such constructions exist for probability distributions having monotone likelihoods.[5][6] One such procedure is an analogue of the Rao--Blackwell procedure for mean-unbiased estimators: The procedure holds for a smaller class of probability distributions than does the Rao--Blackwell procedure for mean-unbiased estimation but for a larger class of loss-functions.[7]

Bias with respect to other loss functions

Any minimum-variance *mean*-unbiased estimator minimizes the risk (expected loss) with respect to the squared-error loss function (among mean-unbiased estimators), as observed by Gauss.[8] A minimum-average absolute deviation *median*-unbiased estimator minimizes the risk with respect to the absolute loss function (among median-unbiased estimators), as observed by Laplace.[9][10] Other loss functions are used in statistics, particularly in robust statistics.[11][12]

Effect of transformations

As stated above, for univariate parameters, median-unbiased estimators remain median-unbiased under transformations that preserve order (or reverse order).

Note that, when a transformation is applied to a mean-unbiased estimator, the result need not be a mean-unbiased estimator of its corresponding population statistic. By Jensen's inequality, a convex function as transformation will introduce positive bias, while a concave function will introduce negative bias, and a function of mixed convexity may introduce bias in either direction, depending on the specific function and distribution. That is, for a non-linear function f and a mean-unbiased estimator U of a parameter p, the composite estimator $f(U)$ need not be a mean-unbiased estimator of $f(p)$. For example, the square root of the unbiased estimator of the population variance is *not* a mean-unbiased estimator of the population standard deviation: the square root of the unbiased sample variance, the corrected sample standard deviation, is biased. The bias depends both on the sampling distribution of the estimator and on the transform, and can be quite involved to calculate – see unbiased estimation of standard deviation for a discussion in this case.

Bias, variance and mean squared error

While bias quantifies the *average* difference to be expected between an estimator and an underlying parameter, an estimator based on a finite sample can additionally be expected to differ from the parameter due to the randomness in the sample.

One measure which is used to try to reflect both types of difference is the mean square error,

$$\operatorname{MSE}(\hat{\theta}) = \operatorname{E}\left[(\hat{\theta} - \theta)^2\right].$$

This can be shown to be equal to the square of the bias, plus the variance:

$$\begin{aligned}\operatorname{MSE}(\hat{\theta}) =& (\operatorname{E}[\hat{\theta}] - \theta)^2 + \operatorname{E}[\,(\hat{\theta} - \operatorname{E}[\,\hat{\theta}\,])^2\,] \\ =& (\operatorname{Bias}(\hat{\theta}, \theta))^2 + \operatorname{Var}(\hat{\theta})\end{aligned}$$

When the parameter is a vector, an analogous decomposition applies:[13]

$$\operatorname{MSE}(\hat{\theta}) = \operatorname{trace}(\operatorname{Var}(\hat{\theta})) + \left\|\operatorname{Bias}(\hat{\theta}, \theta)\right\|^2$$

where

$$\operatorname{trace}(\operatorname{Var}(\hat{\theta}))$$

is the trace of the covariance matrix of the estimator.

An estimator that minimises the bias will not necessarily minimise the mean square error.

Example: Estimation of population variance

For example,[14] suppose an estimator of the form

$$T^2 = c\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = cnS^2$$

is sought for the population variance as above, but this time to minimise the MSE:

$$\begin{aligned}\operatorname{MSE} &= \operatorname{E}\left[(T^2 - \sigma^2)^2\right] \\ &= \left(\operatorname{E}\left[T^2 - \sigma^2\right]\right)^2 + \operatorname{Var}(T^2)\end{aligned}$$

If the variables X_1 ... X_n follow a normal distribution, then nS^2/σ^2 has a chi-squared distribution with $n - 1$ degrees of freedom, giving:

$$\operatorname{E}[nS^2] = (n-1)\sigma^2 \text{ and } \operatorname{Var}(nS^2) = 2(n-1)\sigma^4.$$

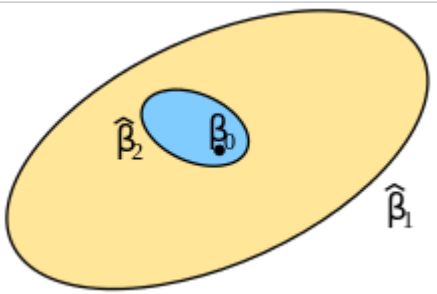
and so

$$\operatorname{MSE} = (c(n-1) - 1)^2\sigma^4 + 2c^2(n-1)\sigma^4$$

With a little algebra it can be confirmed that it is $c = 1/(n + 1)$ which minimises this combined loss function, rather than $c = 1/(n - 1)$ which minimises just the bias term.

More generally it is only in restricted classes of problems that there will be an estimator that minimises the MSE independently of the parameter values.

However it is very common that there may be perceived to be a *bias–variance tradeoff*, such that a small increase in bias can be traded for a larger decrease in variance, resulting in a more desirable estimator overall.



Sampling distributions of two alternative estimators for a parameter β_0. Although $\hat{\beta}_1$ is unbiased, it is clearly inferior to the biased $\hat{\beta}_2$.

Ridge regression is one example of a technique where allowing a little bias may lead to a considerable reduction in variance, and more reliable estimates overall.

Bayesian view

Most bayesians are rather unconcerned about unbiasedness (at least in the formal sampling-theory sense above) of their estimates. For example, Gelman *et al* (1995) write: "From a Bayesian perspective, the principle of unbiasedness is reasonable in the limit of large samples, but otherwise it is potentially misleading."[15]

Fundamentally, the difference between the Bayesian approach and the sampling-theory approach above is that in the sampling-theory approach the parameter is taken as fixed, and then probability distributions of a statistic are considered, based on the predicted sampling distribution of the data. For a Bayesian, however, it is the *data* which is known, and fixed, and it is the unknown parameter for which an attempt is made to construct a probability distribution, using Bayes' theorem:

$$p(\theta \mid D, I) \propto p(\theta \mid I) p(D \mid \theta, I)$$

Here the second term, the likelihood of the data given the unknown parameter value θ, depends just on the data obtained and the modelling of the data generation process. However a Bayesian calculation also includes the first term, the prior probability for θ, which takes account of everything the analyst may know or suspect about θ *before* the data comes in. This information plays no part in the sampling-theory approach; indeed any attempt to include it would be considered "bias" away from what was pointed to purely by the data. To the extent that Bayesian calculations include prior information, it is therefore essentially inevitable that their results will not be "unbiased" in sampling theory terms.

But the results of a Bayesian approach can differ from the sampling theory approach even if the Bayesian tries to adopt an "uninformative" prior.

For example, consider again the estimation of an unknown population variance σ^2 of a Normal distribution with unknown mean, where it is desired to optimise *c* in the expected loss function

$$\mathbf{ExpectedLoss} = \mathbf{E}\left[\left(cnS^2 - \sigma^2\right)^2\right] = \mathbf{E}\left[\sigma^4\left(cn\frac{S^2}{\sigma^2} - 1\right)^2\right]$$

A standard choice of uninformative prior for this problem is the Jeffreys prior, $p(\sigma^2) \propto 1/\sigma^2$, which is equivalent to adopting a rescaling-invariant flat prior for $\ln(\sigma^2)$.

One consequence of adopting this prior is that S^2/σ^2 remains a pivotal quantity, i.e. the probability distribution of S^2/σ^2 depends only on S^2/σ^2, independent of the value of S^2 or σ^2:

$$p\left(\frac{S^2}{\sigma^2} \mid S^2\right) = p\left(\frac{S^2}{\sigma^2} \mid \sigma^2\right) = g\left(\frac{S^2}{\sigma^2}\right)$$

However, whilst

$$\mathbf{E}_{p(S^2|\sigma^2)}\left[\sigma^4\left(cn\frac{S^2}{\sigma^2} - 1\right)^2\right] = \sigma^4\,\mathbf{E}_{p(S^2|\sigma^2)}\left[\left(cn\frac{S^2}{\sigma^2} - 1\right)^2\right]$$

in contrast

$$\mathrm{E}_{p(\sigma^2|S^2)}\left[\sigma^4\left(cn\frac{S^2}{\sigma^2}-1\right)^2\right] \neq \sigma^4\,\mathrm{E}_{p(\sigma^2|S^2)}\left[\left(cn\frac{S^2}{\sigma^2}-1\right)^2\right]$$

— when the expectation is taken over the probability distribution of σ^2 given S^2, as it is in the Bayesian case, rather than S^2 given σ^2, one can no longer take σ^4 as a constant and factor it out. The consequence of this is that, compared to the sampling-theory calculation, the Bayesian calculation puts more weight on larger values of σ^2, properly taking into account (as the sampling-theory calculation cannot) that under this squared-loss function the consequence of underestimating large values of σ^2 is more costly in squared-loss terms than that of overestimating small values of σ^2.

The worked-out Bayesian calculation gives a scaled inverse chi-squared distribution with $n - 1$ degrees of freedom for the posterior probability distribution of σ^2. The expected loss is minimised when $cnS^2 = \langle\sigma^2\rangle$; this occurs when $c = 1/(n - 3)$.

Even with an uninformative prior, therefore, a Bayesian calculation may not give the same expected-loss minimising result as the corresponding sampling-theory calculation.

See also

- Omitted-variable bias
- Consistent estimator
- Estimation theory
- Expected loss
- Expected value
- Loss function
- Median
- Statistical decision theory
- Optimism bias

Notes

1. Richard Arnold Johnson; Dean W. Wichern (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall. ISBN 978-0-13-187715-3. Retrieved 10 August 2012.
2. J. P. Romano and A. F. Siegel (1986) *Counterexamples in Probability and Statistics*, Wadsworth & Brooks / Cole, Monterey, California, USA, p. 168
3. Hardy, M. (1 March 2003). "An Illuminating Counterexample". *American Mathematical Monthly*. **110** (3): 234–238. doi:10.2307/3647938. ISSN 0002-9890. JSTOR 3647938.
4. Brown (1947), page 583
5. Pfanzagl, Johann. "On optimal median unbiased estimators in the presence of nuisance parameters." The Annals of Statistics (1979): 187-193.
6. Brown, L. D.; Cohen, Arthur; Strawderman, W. E. A Complete Class Theorem for Strict Monotone Likelihood Ratio With Applications. Ann. Statist. 4 (1976), no. 4, 712--722. doi:10.1214/aos/1176343543. http://projecteuclid.org/euclid.aos/1176343543.
7. Page 713: Brown, L. D.; Cohen, Arthur; Strawderman, W. E. A Complete Class Theorem for Strict Monotone Likelihood Ratio With Applications. Ann. Statist. 4 (1976), no. 4, 712--722. doi:10.1214/aos/1176343543. http://projecteuclid.org/euclid.aos/1176343543.

8. Dodge, Yadolah, ed. (1987). *Statistical data analysis based on the L1-norm and related methods: Papers from the First International Conference held at Neuchâtel, August 31–September 4, 1987*. Amsterdam: North-Holland Publishing Co.
9. Dodge, Yadolah, ed. (1987). *Statistical data analysis based on the L1-norm and related methods: Papers from the First International Conference held at Neuchâtel, August 31–September 4, 1987*. Amsterdam: North-Holland Publishing Co.
10. Jaynes, E.T. (2007). *Probability theory : the logic of science* (5. print. ed.). Cambridge [u.a.]: Cambridge Univ. Press. p. 172. ISBN 978-0-521-59271-0.
11. Dodge, Yadolah, ed. (1987). *Statistical data analysis based on the L1-norm and related methods: Papers from the First International Conference held at Neuchâtel, August 31–September 4, 1987*. Amsterdam: North-Holland Publishing Co.
12. Chapter 3: Robust and Non-Robust Models in Statistics by Lev B. Klebanov, Svetlozar T. Rachev and Frank J. Fabozzi, Nova Scientific Publishers, Inc. New York, 2009.
13. Taboga, Marco (2010). "Lectures on probability theory and mathematical statistics".
14. Morris H. DeGroot (1986), *Probability and Statistics* (2nd edition), Addison-Wesley. ISBN 0-201-11366-X. Pp. 414–5.
    But compare it with, for example, the discussion in Casella and Berger (2001), *Statistical Inference* (2nd edition), Duxbury. ISBN 0534243126. P. 332.
15. A. Gelman *et al* (1995), *Bayesian Data Analysis*, Chapman and Hall. ISBN 0-412-03991-5. p. 108.

References

- Brown, George W. (http://www.universityofcalifornia.edu/senate/inmemoriam/georgewbrown.htm) "On Small-Sample Estimation." *The Annals of Mathematical Statistics*, vol. 18, no. 4 (Dec., 1947), pp. 582–585. JSTOR 2236236 (https://www.jstor.org/stable/2236236).
- Lehmann, E. L. "A General Concept of Unbiasedness" *The Annals of Mathematical Statistics*, vol. 22, no. 4 (Dec., 1951), pp. 587–592. JSTOR 2236928 (https://www.jstor.org/stable/2236928).
- Allan Birnbaum, 1961. "A Unified Theory of Estimation, I", *The Annals of Mathematical Statistics*, vol. 32, no. 1 (Mar., 1961), pp. 112–135.
- Van der Vaart, H. R., 1961. "Some Extensions of the Idea of Bias" *The Annals of Mathematical Statistics*, vol. 32, no. 2 (June 1961), pp. 436–447.
- Pfanzagl, Johann. 1994. *Parametric Statistical Theory*. Walter de Gruyter.
- Stuart, Alan; Ord, Keith; Arnold, Steven [F.] (2010). *Classical Inference and the Linear Model*. Kendall's Advanced Theory of Statistics. **2A**. Wiley. ISBN 0-4706-8924-2..
- Voinov, Vassily [G.]; Nikulin, Mikhail [S.] (1993). *Unbiased estimators and their applications*. 1: Univariate case. Dordrect: Kluwer Academic Publishers. ISBN 0-7923-2382-3.
- Voinov, Vassily [G.]; Nikulin, Mikhail [S.] (1996). *Unbiased estimators and their applications*. 2: Multivariate case. Dordrect: Kluwer Academic Publishers. ISBN 0-7923-3939-8.
- Klebanov, Lev [B.]; Rachev, Svetlozar [T.]; Fabozzi, Frank [J.] (2009). *Robust and Non-Robust Models in Statistics*. New York: Nova Scientific Publishers. ISBN 978-1-60741-768-2.

External links

- Hazewinkel, Michiel, ed. (2001), "Unbiased estimator", *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4

Categories: Statistical theory | Point estimation performance | Bias