If \( \hat{\theta} = T(X) \) is an estimator of \( \theta \), then the bias of \( \hat{\theta} \) is the difference between its expectation and the 'true' value: i.e.

\[
\text{bias}(\hat{\theta}) = E_{\theta}(\hat{\theta}) - \theta.
\]

An estimator \( T(X) \) is unbiased for \( \theta \) if \( E_{\theta} T(X) = \theta \) for all \( \theta \), otherwise it is biased.

In the above example, \( E_{\mu}(T) = \mu \) so \( T \) is unbiased for \( \mu \).

[Notation note: when a parameter subscript is used with an expectation or variance, it refers to the parameter that is being conditioned on. i.e. the expectation or variance will be a function of the subscript]

2. Estimation and bias

2.1. Estimators

Suppose that \( X_1, \ldots, X_n \) are iid, each with pdf/pmf \( f_X(x \mid \theta) \), \( \theta \) unknown.

- We aim to estimate \( \theta \) by a statistic, i.e. by a function \( T \) of the data.
- If \( X = x = (x_1, \ldots, x_n) \) then our estimate is \( \hat{\theta} = T(x) \) (does not involve \( \theta \)).
- Then \( T(X) \) is our estimator of \( \theta \), and is a rv since it inherits random fluctuations from those of \( X \).
- The distribution of \( T = T(X) \) is called its sampling distribution.

**Example**

Let \( X_1, \ldots, X_n \) be iid \( N(\mu, 1) \).

A possible estimator for \( \mu \) is \( T(X) = \frac{1}{n} \sum x_i \).

For any particular observed sample \( x \), our estimate is \( T(x) = \frac{1}{n} \sum x_i \).

We have \( T(X) \sim N(\mu, 1/n) \).

\[\square\]

2.2. Bias

If \( \hat{\theta} = T(X) \) is an estimator of \( \theta \), then the bias of \( \hat{\theta} \) is the difference between its expectation and the 'true' value: i.e. the expectation or variance will be a function of the subscript.

\[
\text{bias}(\hat{\theta}) = E_{\theta}(\hat{\theta}) - \theta.
\]

An estimator \( T(X) \) is unbiased for \( \theta \) if \( E_{\theta} T(X) = \theta \) for all \( \theta \), otherwise it is biased.

In the above example, \( E_{\mu}(T) = \mu \) so \( T \) is unbiased for \( \mu \).

[Notation note: when a parameter subscript is used with an expectation or variance, it refers to the parameter that is being conditioned on. i.e. the expectation or variance will be a function of the subscript]

2.3. Mean squared error

Recall that an estimator \( T \) is a function of the data, and hence is a random quantity. Roughly, we prefer estimators whose sampling distributions “cluster more closely” around the true value of \( \theta \), whatever that value might be.

**Definition 2.1**

The mean squared error (mse) of an estimator \( \hat{\theta} \) is \( E_{\theta} [(\hat{\theta} - \theta)^2] \).

For an unbiased estimator, the mse is just the variance. In general

\[
E_{\theta}[(\hat{\theta} - \theta)^2] = E_{\theta}[(\hat{\theta} - E_{\theta}\hat{\theta} + E_{\theta}\hat{\theta} - \theta)^2] = E_{\theta}[(\hat{\theta} - E_{\theta}\hat{\theta})^2] + [E_{\theta}(\hat{\theta}) - \theta]^2 + 2[E_{\theta}(\hat{\theta}) - \theta]E_{\theta}[(\hat{\theta} - E_{\theta}\hat{\theta})] = \text{var}_{\theta}(\hat{\theta}) + \text{bias}_{\theta}^2(\hat{\theta}).
\]

where bias(\( \hat{\theta} \)) = E_{\theta}(\hat{\theta}) - \theta.

[NB: sometimes it can be preferable to have a biased estimator with a low variance - this is sometimes known as the 'bias-variance tradeoff'.]

2.4. Estimation and bias

2.4.1. Efficiency

2.4.2. Consistency

2.4.3. Sufficiency

2.4.4. Asymptotic distribution

2.4.5. Delta method

2.4.6. Cramér–Rao bound

2.4.7. Maximum likelihood estimation

2.4.8. Bayesian estimation

2.4.9. Non-parametric estimation

2.4.10. Instrumental variables

2.4.11. Monte Carlo estimation

2.4.12. Other estimation methods
Example: Alternative estimators for Binomial mean

- Suppose \( X \sim \text{Binomial}(n, \theta) \), and we want to estimate \( \theta \).
- The standard estimator is \( T_U = X/n \), which is Unbiassed since 
  \[ \mathbb{E}_\theta(T_U) = n\theta/n = \theta. \]
- \( T_U \) has variance \( \text{var}_\theta(T_U) = \text{var}_\theta(X)/n^2 = \theta(1-\theta)/n \).
- Consider an alternative estimator \( T_B = \frac{X+1}{n+2} = w \frac{X}{n} + (1-w) \frac{1}{2} \), where 
  \( w = n/(n+2) \). \( T_B \) is a weighted average of \( X/n \) and \( \frac{1}{2} \).
- e.g. if \( X \) is 8 successes out of 10 trials, we would estimate the underlying success probability as \( T(8) = 9/12 = 0.75 \), rather than 0.8.
- Then \( \mathbb{E}_\theta(T_B) - \theta = \frac{n+1}{n+2} - \theta = (1-w)(\frac{1}{2} - \theta) \), and so it is biased.
- \( \text{var}_\theta(T_B) = \frac{\text{var}_\theta(X)}{(n+2)^2} = w^2 \theta(1-\theta)/n \).
- \( \text{mse}(T_B) = \text{var}_\theta(T_B) + \text{bias}^2(T_B) = w^2 \theta(1-\theta)/n + (1-w)^2 (\frac{1}{2} - \theta)^2 \)

So the biased estimator has smaller MSE in much of the range of \( \theta \). \( T_B \) may be preferable if we do not think \( \theta \) is near 0 or 1.
So our prior judgement about \( \theta \) might affect our choice of estimator.
Will see more of this when we come to Bayesian methods.

Why unbiasedness is not necessarily so great

Suppose \( X \sim \text{Poisson}(\lambda) \), and for some reason (which escapes me for the moment), you want to estimate \( \theta = \left[ \mathbb{P}(X=0) \right]^2 = e^{-2\lambda} \).
Then any unbiased estimator \( T(X) \) must satisfy 
\[ \mathbb{E}_\lambda(T(X)) = \theta, \]
or equivalently 
\[ \mathbb{E}_\lambda(T(X)) = e^{-\lambda} \sum_{x=0}^{\infty} T(x) \frac{\lambda^x}{x!} = e^{-2\lambda}. \]
The only function \( T \) that can satisfy this equation is \( T(X) = (-1)^X \) [coefficients of polynomial must match].
Thus the only unbiased estimator estimates \( e^{-2\lambda} \) to be 1 if \( X \) is even, -1 if \( X \) is odd.
This is not sensible.