Estimator

Deterministic function of the data $x_1, x_2, \cdots, x_n$,

$$y := T(x_1, x_2, \cdots, x_n)$$

We refer to $T(x_1, x_2, \cdots, x_n)$ as a point estimator.

**Aim**: Use $y$ to estimate a quantity $\theta$ related to the underlying distribution (population)

- $\theta$ is the population mean and $y$ is the sample mean
- $\theta$ is the population variance and $y$ is the sample variance
- $\theta$ is the population median and $y$ is the sample median
If data are samples from a probabilistic model, then $y$ is a realization of the random variable. 

$$\hat{\theta}_n := T(X_1, \cdots, X_n)$$

Remember that

- $\hat{\theta}_n$ depends on data and is a random variable
- $\theta$ is a fixed but unknown quantity. We use $\hat{\theta}_n$ to estimate $\theta$. 
Evaluate the quality of estimator under probabilistic assumption

How to evaluate the estimator $\hat{\theta}_n$? How well does it approximate $\theta$?

We discuss a few criteria to evaluate the quality of estimator

- Bias
- Mean squares error
- Consistency
- Asymptotically normal
Note that
\[
\text{bias}(\hat{\theta}_n) = \mathbb{E}_\theta(\hat{\theta}_n) - \theta.
\]
Here the expectation is taken w.r.t. the joint pdf under the true parameter \( \theta \):
\[
f(x_1, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).
\]
We say \( \hat{\theta}_n \) is unbiased if
\[
\text{bias}(\hat{\theta}_n) = 0 \iff \mathbb{E}_\theta(\hat{\theta}_n) = \theta.
\]
The estimator \( \hat{\theta}_n \) is unbiased if the expected value of the estimator is the same as the true underlying quantitative parameter being estimated.
Example of unbiased estimators

- Sample mean:

\[ \mathbb{E} \overline{X}_n = \frac{1}{n} \mathbb{E} \left( \sum_{i=1}^{n} X_i \right) = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu. \]

- Sample variance:

\[ S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2, \quad \mathbb{E}(S_n^2) = \sigma^2. \]

Why?

\[
(n - 1) \mathbb{E} S_n^2 = \mathbb{E}(X_i - \overline{X}_n)^2 = \sum_{i=1}^{n} \mathbb{E} X_i^2 - n \mathbb{E} \overline{X}_n^2
\]

\[
= n \mathbb{E}(X^2) - n \mathbb{E}(\overline{X}_n - \mu + \mu)^2
\]

\[
= n \mathbb{E}(X^2) - n(\text{Var}(\overline{X}_n) + \mu^2)
\]

\[
= n(\sigma^2 + \mu^2) - n(\sigma^2 / n + \mu^2) = (n - 1)\sigma^2
\]
The mean square error (MSE) of an estimator $\hat{\theta}_n$ that approximates $\theta$ is

$$\text{MSE}(\hat{\theta}_n) := \mathbb{E}_\theta (\hat{\theta}_n - \theta)^2 = \int_{\mathcal{X}} (\hat{\theta}_n - \theta)^2 f(x_1, \ldots, x_n, \theta) \, dx_1 \cdots dx_n$$

where $\hat{\theta}_n$ is a function of $\{x_i\}_{i=1}^n$.

- MSE is a measure of “risk”.
- The larger MSE is, the worse the quality of $\hat{\theta}_n$ is.
- Aim: To find $\hat{\theta}_n$ with small MSE, i.e., risk minimization.
- Important in statistical decision theory. $\hat{\theta}_n$ is also referred to as the policy/decision.
- MSE is one choice of risk. Other risk functions include absolute loss, hinge loss (used in support vector machine), etc.
Bias-variance decomposition

Remember $\hat{\theta}_n$ is a random variable. Is the $\text{MSE}(\hat{\theta}_n)$ related to the variance of $\hat{\theta}_n$ and bias?

\[
\text{MSE}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n - \theta)^2
\]
\[
= \mathbb{E}(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n) + \mathbb{E}(\hat{\theta}_n) - \theta)^2
\]
\[
= \mathbb{E}(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))^2 + (\mathbb{E}(\hat{\theta}_n) - \theta)^2
\]
\[
+ 2 \mathbb{E}(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))(\mathbb{E}(\hat{\theta}_n) - \theta)
\]
\[
= \mathbb{E}(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))^2 + (\mathbb{E}(\hat{\theta}_n) - \theta)^2
\]
\[
+ 2(\mathbb{E}(\hat{\theta}_n) - \theta) \mathbb{E}(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))
\]
\[
= \mathbb{E}(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))^2 + (\mathbb{E}(\hat{\theta}_n) - \theta)^2
\]
\[
= \text{Var}(\hat{\theta}_n) + \text{bias}^2(\hat{\theta}_n).
\]

In other words, the MSE consists of two parts: bias and variance.
We are interested in whether $\hat{\theta}_n$ will “converge” to $\theta$ as the sample size goes to infinity.

**Consistency**

An estimator $\hat{\theta}_n$ that approximates $\theta$ is consistent if $\hat{\theta}_n$ converges to $\theta$ as $n \to \infty$ in probability, i.e., for any $\varepsilon > 0$

$$P(|\hat{\theta}_n - \theta| \geq \varepsilon) \to 0$$

as $n \to \infty$.

In particular, if $\hat{\theta}_n$ converges to $\theta$ in MSE, then $\hat{\theta}_n$ is a consistent estimator of $\theta$. Why?
Consistency

- The empirical mean of an i.i.d. sequence $X$ with mean $\mu$
  \[ \overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \]
  is consistent by the law of large numbers if the variance is bounded.
- The sample variance is a consistent estimator of $\sigma^2$ given the fourth moment exists $\mathbb{E}(X_i^4) < \infty$. (See homework).
  \[ S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 = \frac{1}{n-1} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} \overline{X}_n^2 \rightarrow \mathbb{E}(X^2) - \mu^2 = \sigma^2 \]
  in probability.
- Sample covariance matrix is a consistent estimator of the underlying covariance matrix.
Empirical median is consistent

The empirical median of an i.i.d. sequence $X$ is consistent even if the mean is not well defined or the variance is unbounded.

In general, the $\alpha$-empirical quantile is a consistent estimator of the actual $\alpha$-quantile.

Suppose $X_i \sim$ Cauchy distribution, i.e.,

$$f_X(x) = \frac{1}{\pi} \cdot \frac{1}{1 + x^2}, \quad x \in \mathbb{R}.$$ 

Would the sample average converge to 0?

Note that neither $\mathbb{E} X$ nor $\mathbb{E} X^2$ exists. The median is 0, i.e.,

$$F_X(0) = \mathbb{P}(X \leq 0) = \frac{1}{2}.$$
Empirical median and mean for Cauchy distribution
Empirical median and mean for Cauchy distribution
Empirical median and mean for Cauchy distribution

Sample median and mean of Cauchy distribution

- Sample mean
- Sample median
The empirical distribution function

Why empirical quantile is consistent? Quantile function is the inverse function of cdf.

Empirical cumulative distribution function

The empirical cdf is defined as

\[ F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\} \]

where \(\{X_i\}_{i=1}^{n}\) are samples from distribution \(F_X(x)\) and

\[ 1\{X_i \leq x\} = \begin{cases} 
1, & X_i \leq x, \\
0, & X_i > x.
\end{cases} \]
Convergence of empirical cdf to cdf

- $F_n(x)$ is an unbiased estimator of $F_X(x)$:

$$
\mathbb{E} F_N(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} 1\{X_i \leq x\} = \mathbb{P}(X_i \leq x) = F_X(x).
$$

- Consistency:

$$
\mathbb{E}(F_n(x) - F_X(x))^2 = \frac{F_X(x)(1 - F_X(x))}{n} \rightarrow 0, \quad n \rightarrow \infty.
$$

Therefore, $F_n(x)$ converges to $F_X(x)$ in probability.

- If empirical cdf is a consistent estimator of the actual cdf, the empirical quantile should be a consistent estimator of the quantile function.