Bias

One of the foremost expectations of an estimator is that it gives accurate estimates.

Definition

An estimator $\hat{\theta}$ is said to be *accurate* or *unbiased* if and only if

$$\mu_{\hat{\theta}} = E(\hat{\theta}) = \theta_0 \tag{19}$$

In plain language, the average of estimates across the records should yield the true value.

The difference $\triangle\hat{\theta} = E(\hat{\theta}) - \theta_0$ is said to be the **bias** of that estimator.

Example: Unbiased estimator

Example

The sample mean estimator $\bar{y} = \dfrac{1}{N}\sum_{k=0}^{N-1} y[k]$ is unbiased since

$$E(\bar{y}) = E\left(\frac{1}{N}\sum_{k=0}^{N-1} y[k]\right) = \frac{1}{N}\sum_{k=0}^{N-1} E(y[k]) = \mu \qquad (20)$$

assuming $y[k]$ to be **stationary**.

Remarks

- The averaging in (19) is across all possible records of data and *not* along time. From this viewpoint, the definition has limited practical value since it is extremely rare to obtain multiple records of data.
- Unbiasedness is nevertheless a useful requirement for comparing performance of two estimators.

An unbiased estimator is desirable. However, what is generally more important is the **spread** of estimates when different realizations are presented. This is measured by the **variance** of the estimator.

Variance of estimators

Definition

The variance of an estimator (estimate) is defined as

$$\sigma^2_{\hat{\theta}} = E((\hat{\theta} - \mu_{\hat{\theta}})^2) = E((\hat{\theta} - E(\hat{\theta}))^2) \tag{21}$$

- Observe that the definition is w.r.t the average of the estimator, $\mu_{\hat{\theta}}$ and *not* with respect to its true value, θ_0. When the estimator is unbiased, $E(\hat{\theta}) = \theta_0$.
- The square root of the variance in (21) is the **standard error** in an estimate.
- It is obviously desirable to have the variance of estimate much lower than that in the data itself.

Remarks

The variance expression is useful in many different ways:

i. Computation of error in estimates.
ii. Constructing confidence regions for the true parameters.
iii. Design of experiments, i.e., knowing how experimental factors can be adjusted to achieve more reliable (precise) estimates.

Example: Variance

Variance of sample mean

Using Definition 3,

$$\sigma_{\bar{y}}^2 = E((\bar{y} - E(\bar{y}))^2) = E\left(\left(\frac{1}{N}\sum_{k=0}^{N-1} y[k] - \mu_y\right)^2\right) = E\left(\left(\frac{1}{N}\sum_{k=1}^{N}(y[k] - \mu_y)\right)^2\right)$$

$$= \frac{1}{N^2}E\left(\sum_{k=1}^{N}(y[k] - \mu_y)^2\right) + \frac{1}{N^2}E\left(\sum_{n=1}^{N}\sum_{m=1,m\neq n}^{N}(y[n] - \mu_y)(y[m] - \mu_y)\right)$$

$$= \frac{1}{N^2}\left(\sum_{k=1}^{N}E(y[k] - \mu_y)^2\right) + \frac{1}{N^2}\left(\sum_{n=1}^{N}\sum_{m=1,m\neq n}^{N}E(y[n] - \mu_{y,n})(y[m] - \mu_{y,n})\right)$$

Example ... contd.

The summand in the second term can be easily recognized as the ACVF $y[k]$.

When the signal is WN, *i.e.*,

$$y[k] = c + e[k], \quad e[k] \sim \text{GWN}(0, \sigma_e^2)$$

the variability of sample mean is

$$\boxed{\sigma_{\bar{y}}^2 = \frac{\sigma_y^2}{N} = \frac{\sigma_e^2}{N}} \tag{22}$$

Remarks

- var$(\bar{y}) \propto \sigma_y^2$ (for a fixed sample size). Intuitively this is a meaningful result. However, we have no control over this factor.
- var$(\bar{y}) \propto 1/N, \Longrightarrow \sigma_{\bar{y}}^2 \to 0$ as $N \to \infty$. This is an interesting result and also a good feature of the estimator. Thus, we are able to shrink the variability (and the error) in the estimate by collecting more samples.
- As we shall shortly learn, (unbiased) estimators that possess this feature are known to be **consistent**.
- The true mean has no bearing on the variability (of the sample mean), which is again a sensible result.

Variance of vector of parameters

When $\hat{\boldsymbol{\theta}}$ is a $p \times 1$ vector, we have a variance-covariance **matrix,**

$$\boxed{\begin{aligned}\text{Var}(\hat{\boldsymbol{\theta}}) = \Sigma_{\hat{\boldsymbol{\theta}}} &= E((\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}}))(\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}}))^T) \\ &= \begin{bmatrix} \sigma^2_{\hat{\theta}_1} & \sigma_{\hat{\theta}_1\hat{\theta}_2} & \cdots & \sigma_{\hat{\theta}_1\hat{\theta}_p} \\ \sigma_{\hat{\theta}_2\hat{\theta}_1} & \sigma^2_{\hat{\theta}_2} & \cdots & \sigma_{\hat{\theta}_2\hat{\theta}_p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{\hat{\theta}_p\hat{\theta}_1} & \sigma_{\hat{\theta}_p\hat{\theta}_2} & \vdots & \sigma^2_{\hat{\theta}_p} \end{bmatrix}\end{aligned}} \tag{23}$$

Remarks

- It is a symmetric matrix with the off-diagonal elements reflecting the error incurred in estimating a pair of parameters jointly.
- **A diagonal $\Sigma_{\hat{\theta}}$ connotes that the parameters can be estimated on an individual basis. In practice, the trace$(\Sigma_{\hat{\theta}})$ and the diagonal elements of $\Sigma_{\hat{\theta}}$ find wider utility.**

Minimum Variance Unbiased Estimator (MVUE)

Definition

An estimator $\hat{\theta}(\mathbf{Z})$ is said to be *minimum variance unbiased estimator* (MVUE) if and only if

C1. $E(\hat{\theta}) = \theta_0$ (unbiased)

C2. $\text{var}(\hat{\theta}) \leq \text{var}(\hat{\theta}_i) \;\; \forall i$ satisfying C1 (least variance)

- The class is restricted to unbiased since biased estimators can always be tuned to have lower variance by sacrificing the bias. Then the comparison becomes difficult.
- The efficiency of an estimator is used to measure how well it performs relative to an *unbiased* estimator that has the least variance.

Comparing two estimators: Efficiency

Formally, the efficiency of an estimator $\hat{\theta}$ is defined as

$$\boxed{\text{Efficiency}(\hat{\theta}) = \eta_{\hat{\theta}} = \frac{\text{var}(\hat{\theta}^\star)}{\text{var}(\hat{\theta})}} \tag{24}$$

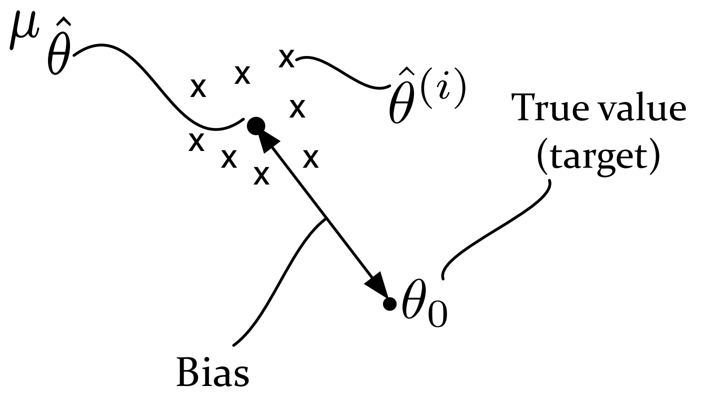
where $\hat{\theta}^\star(\mathbf{y})$ has theoretically the lowest variance among all estimators.

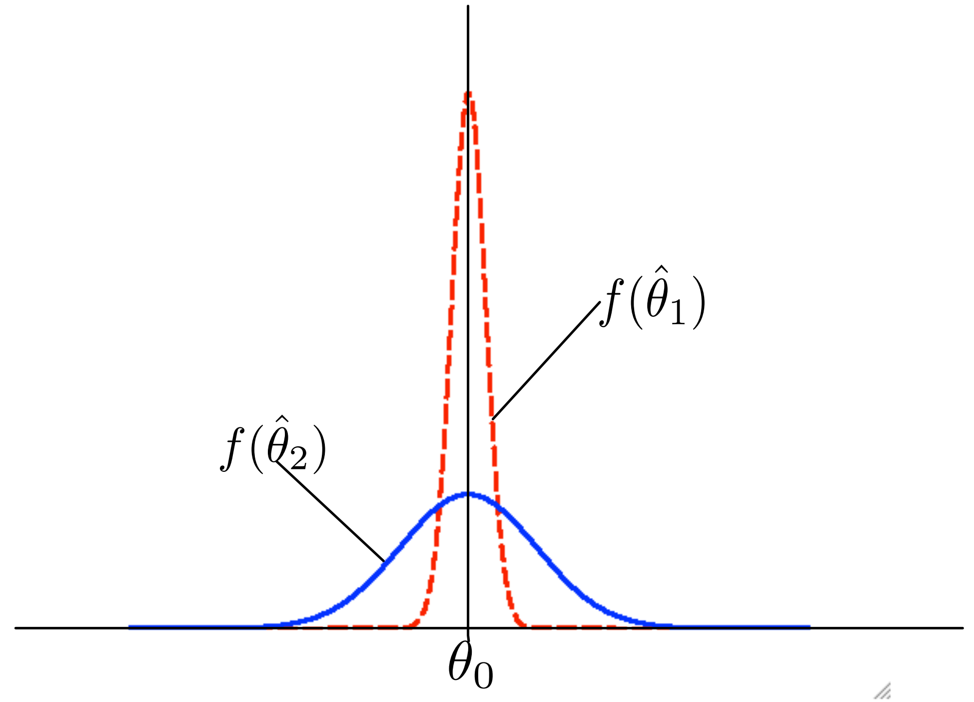
Remarks

- The Cramer-Rao's inequality dictates the bound (on achievable variance) and also stipulates the condition under which such an estimator exists.
- An estimator that achieves this lower bound is said to be the *most efficient* or *fully efficient*.
- When it is not possible to find an efficient estimator, *relative efficiency* is used.

$$\text{Relative efficiency (\%)} = 100 \times \frac{\sigma^2_{\hat{\theta}_1}}{\sigma^2_{\hat{\theta}_2}} \tag{25}$$

Bias, Variance and Efficiency





Bias is the distance between the center of estimates and the true value, while the variance is a measure of spread around its own center.

The estimator $\hat{\theta}_1$ has lesser spread than $\hat{\theta}_2$, and is therefore relatively more efficient. It produces estimates that have a higher probability (than those of $\hat{\theta}_2$) of being closer to θ_0.

Most efficient estimator

In seeking the most efficient estimator, it is important to answer the question: what is the minimum variance achievable by any unbiased estimator? The celebrated C-R inequality answers this question.

Cramer-Rao inequality

Theorem

Suppose $\hat{\theta}(\mathbf{y})$ is an unbiased estimator of a single parameter θ. Then, if the p.d.f. $f(\mathbf{y};\theta)$ is regular, the variance of any unbiased estimator is bounded below by $I(\theta))^{-1}$

$$\boxed{\mathit{var}(\hat{\theta}(\mathbf{y})) \geq (I(\theta))^{-1}} \tag{26}$$

where $I(\theta)$ is the information measure in (6) (or (9)). Further, an estimator $\hat{\theta}^\star(\mathbf{y})$ that can achieve this lower bound exists if and only if

$$S(Y_N,\theta) = I(\theta)(\hat{\theta}^\star(\mathbf{y}) - \theta) \tag{27}$$

Then, $\hat{\theta}^\star(\mathbf{y})$ is the **most efficient** *estimator of θ.*

Cramer-Rao lower bound

The C-R inequality gives us:

i. Lowest variance achievable by any unbiased estimator
ii. Means of deriving that most efficient estimator, if it exists.

The role of Fisher information introduced earlier is clear now.

Larger the information on θ in a given data, lower is the variability and hence the error in $\hat{\theta}$.

Existence of efficient estimator

An alternative form of the condition of existence of an efficient estimator can be given. From (27), the MVUE that achieves the C-R bound exists if and only if

$$\frac{S(Y_N, \theta)}{I(\theta)} + \theta \tag{28}$$

is **independent of θ (sufficiency) and only dependent on the observations $\mathbf{y}$.**

Example: C-R bound

Efficient estimator of mean

Consider the standard problem of estimating the mean of a GWN process $y[k] \sim \mathcal{N}(\mu, \sigma^2)$ from N observations. Find the most efficient estimator of μ.

Solution: Recall from (1)

$$I(\mu) = \frac{N}{\sigma^2} \implies \text{var}(\hat{\mu}) \geq (I(\mu))^{-1} = \frac{\sigma^2}{N} \tag{29}$$

Example: C-R bound ... contd.

To determine the existence of an estimator that achieves this minimum, construct (28)

$$\frac{S(\mathbf{y},\theta)}{I(\theta)} + \theta = \frac{\sum_{k=0}^{N-1}(y[k]-\mu)}{N} + \mu = \frac{1}{N}\sum_{k=0}^{N-1} y[k] \tag{30}$$

which is only dependent on the observations $\mathbf{y}$.

This is none other than the sample mean! In a previous example, we showed that the variance of this estimator is indeed σ^2/N. Thus, we conclude that the sample mean is the most efficient estimator of the mean of a GWN.

Existence of an efficient estimator

Whether it is possible to arrive at an efficient estimator depends on two factors:

1. *The parameter θ, or in general, its function $g(\theta)$*. For *e.g.*, in the case of exponentially distributed WN, it turns out that there exists an efficient estimator if $1/\lambda$ is estimated instead of λ.

   In parametric modelling, this means that the form of parametrization, i.e., how the parameters enter the model, has an important say in estimation and the estimate.

2. *The probabilistic characteristics of the observed data*. In reality, it is difficult to know the p.d.f. a priori. Then, the existence of an efficient estimator depends on the assumed density function.

Mean Square Error

The mean square error (MSE) of an estimator is its variance with reference to its true value θ_0.

Definition

The MSE of an estimator is defined as

$$\text{MSE}(\hat{\theta}) = E(||\hat{\theta} - \theta_0||_2^2) = E(\sum_{i=1}^{p}(\hat{\theta}_i - \theta_{i0})^2) \tag{31}$$

A classical result in estimation relates the bias, variance and MSE.

MSE ... contd.

Theorem

For any estimator $\hat{\theta}$, the following identity holds

$$MSE(\hat{\theta}) = trace(\Sigma_{\hat{\theta}}) + ||\triangle\hat{\theta}||_2^2 \tag{32}$$

Proof:

$$\begin{aligned}
E(||\hat{\theta} - \theta_0||_2^2) &= E(tr((\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^T)) \\
&= tr(E((\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^T)) \\
&= tr(E((\hat{\theta} - E(\hat{\theta}))(\hat{\theta} - E(\hat{\theta})^T)) + tr(E((E(\hat{\theta}) - \theta_0)(E(\hat{\theta}) - \theta_0)^T)) \\
&+ 2tr(E((\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta_0)^T)) \\
&= trace(\Sigma_{\hat{\theta}}) + ||\triangle\hat{\theta}||_2^2
\end{aligned}$$

MSE ...contd.

The last identity comes about by recognizing the first term as the trace of Var$(\hat{\theta})$ and that $E(\hat{\theta} - \theta_0)$ is a deterministic quantity. Consequently the expectation on the second term disappears

$$\text{tr}(E((E(\hat{\theta}) - \theta_0)(E(\hat{\theta}) - \theta_0)^T)) = \text{tr}(\triangle\hat{\theta}\triangle\hat{\theta}^T) = \text{tr}(\triangle\hat{\theta}^T\triangle\hat{\theta}) = ||\triangle\hat{\theta}||_2^2$$

and the third term vanishes to zero.

MSE ... contd.

1. For unbiased estimators, $\triangle\hat{\theta} = 0$, therefore MSE and $\Sigma_{\hat{\theta}}$ are identical.
2. Since both terms on the RHS of (32) are positive-valued, *estimators that have small MSE naturally require good accuracy and precision*.
3. When MSE$(\hat{\theta}) \to 0$ as $N \to \infty$, the estimator is said to be consistent.
4. It is ideally desirable to build an estimator $\hat{\theta}$ with minimum mean square error. The MMSE problem can be set up by assuming the parameter θ to be a random variable. Therefore, this is useful in a Bayesian estimation framework. The resulting estimator, as it turns out is the conditional expectation $E(\theta|\mathbf{y})$.

Minimum Mean Square Error estimator

Theorem

The MMSE estimator of θ given $\mathbf{y}$ is the conditional expectation

$$\boxed{\hat{\theta}_{MMSE}(Y) = E(\theta|Y)} \tag{33}$$

As in the case of MVUE, the form of MMSE could be non-linear or linear. For practical reasons, linear MMSE estimators are more popular. In fact, when $\boldsymbol{\theta}$ and $\mathbf{y}$ follow a joint Gaussian distribution, the linear MMSE is also the optimal MMSE.

Asymptotic bias

Statistical unbiasedness is a desirable property; however, *it is not necessarily the most desirable property*. A *biased* estimator is also considered acceptable provided the bias vanishes for very large samples. For this purpose, asymptotic unbiasedness is defined.

Definition

An estimator is said to be asymptotically unbiased if

$$\lim_{N\to\infty} \triangle\hat{\theta} = 0 \qquad \textit{i.e.,} \qquad \lim_{N\to\infty} E(\hat{\theta}) = \theta_0 \tag{34}$$

- Asymptotic bias is a large sample property. Therefore it is of little interest in situations concerning small samples.

Asymptotic bias ... contd.

A standard estimator of variance

$$\hat{\sigma}_y^2 = \frac{1}{N}\sum_{k=0}^{N-1}(y[k] - \bar{y})^2 \tag{35}$$

where $\bar{y}$ is the sample mean, is a *biased* estimator of σ_y^2 but is asymptotically unbiased.

- A *statistically* biased estimator can achieve a variance lower than that of a MVU estimator. However, the variance is no longer a measure for comparing the performance of such estimators since in principle one can shrink the variance to an arbitrarily low (non-zero) value by increasing the bias to a very large value.
- Thus, a better universal metric is the MSE.

Consistency

An important and desirable large sample property is **consistency**, which examines the convergence of $\hat{\theta}$ to θ_0 as $N \to \infty$.

An estimator is said to be *consistent* if $\hat{\theta}$ (a RV) converges to θ_0 (a fixed value).

Different forms of consistency arise depending on the notion of convergence one uses:

1. **In probability:** $\hat{\theta}_N \xrightarrow{p} \theta_0$ iff $\lim\limits_{N\to\infty} \Pr(|\hat{\theta}_N - \theta_0| \geq \varepsilon) = 0, \quad \forall \varepsilon > 0.$
2. **In mean square sense:** $\hat{\theta}_N \xrightarrow{m.s.} \theta_0$ iff $\lim\limits_{N\to\infty} E((\hat{\theta}_N - \theta_0)^2) = 0.$
3. **Almost sure convergence:** $\hat{\theta}_N \xrightarrow{\text{a.s.}} \theta_0$ iff $\hat{\theta}_N \longrightarrow \theta_0$ w.p.1

Order of implication: Almost sure $\Longrightarrow$ Mean square $\Longrightarrow$ Probabilistic

Convergence of sequences of random variables

Definition

A sequence of real numbers $\{x_n\}$ is a realization of the sequence of random variables $\{X_n\}$ if x_n is a realization of the RV X_n.

Sequences of RVs on a sample space Ω

$\{X_n\}$ is a sequence of random variables on a sample space Ω if all the RVs belonging to the sequence are mappings from Ω to $\mathbb{R}$.

One can then have i.i.d or stationary or weakly stationary sequences, etc.

Pointwise convergence

The requirement is that there exist a random variable to which all possible sequences converge on Ω.

Pointwise convergence

Let $\{X_n\}$ be a sequence of random variables defined on a sample space Ω. Then it is pointwise convergent to a random variable X if and only if $\{X_n(\omega)\}$ converges to $X(\omega)$ for all $\omega \in \Omega$. X is called the pointwise limit of the sequence and denoted as

$$\boxed{X_n \xrightarrow{\text{pointwise}} X} \tag{36}$$

Example: PC

Let $\Omega = \{\text{blue}, \text{red}\}$ be the sample space with two sample points. Suppose $\{X_n\}$ is a sequence of RVs such that

$$X_n(\omega) = \begin{cases} \dfrac{2}{n}, & \omega = \text{blue} \\ 2 + \dfrac{1}{n}, & \omega = \text{red} \end{cases}$$

Then the sequence converges to a random variable

$$X(\omega) = \begin{cases} 0, & \omega = \text{blue} \\ 2, & \omega = \text{red} \end{cases}$$

Convergence in probability

Idea: The sequence gets very close to a RV X with "high probability".

Convergence in probability

Let $\{X_n\}$ be a sequence of random variables defined on a sample space Ω and ϵ be a strictly positive number. Then, $\{X_n\}$ is said to be convergent in probability if and only if

$$\lim_{n\to\infty} \text{Pr}(|X_n - X| > \epsilon) = 0 \tag{37}$$

and denoted by

$$\boxed{X_n \xrightarrow{p} X \qquad \text{or} \qquad \underset{n\longrightarrow\infty}{\text{plim}}\, X_n = X} \tag{38}$$

Example

Consider a sequence of RVs $X_n = \left(1 + \dfrac{1}{n}\right) X$, where X on $\Omega = \{0, 1\}$ is a discrete RV with p.m.f.

$$p_X(X) = \begin{cases} \frac{1}{5}, & x = 1, \\ \frac{4}{5}, & x = 0 \end{cases}$$

Then $|X_n - X| = 0$ when $X = 0$ (with probability 4/5) and $|X_n - X| = \frac{1}{n}$ when $X = 1$ (with prob. 1/5). Therefore,

$$\Pr(|X_n - X| \leq \epsilon) = \begin{cases} \dfrac{4}{5}, & n < \frac{1}{\epsilon} \\ 1, & n \geq \frac{1}{\epsilon} \end{cases}$$

Mean square convergence

Idea: The sequence gets very close to a RV X in a "squared distance" sense.

Convergence in mean-square

Let $\{X_n\}$ be a sequence of random variables defined on a sample space Ω. Then, $\{X_n\}$ is said to be convergent in mean-square sense if and only if there exists a RV (with finite variance), such that

$$\underset{n\to\infty}{n} E\left[(X_n - X)^2\right] = 0 \tag{39}$$

i.e., in the sense of a distance metric. The convergence is denoted by

$$\boxed{X_n \xrightarrow{m.s.} X} \tag{40}$$

Example

Consider a sequence of RVs $X_n = \dfrac{1}{n}\sum_{k=0}^{N-1} x[k]$, where $x[k]$'s are uncorrelated random variables with mean μ and variance σ^2.

Then, the sequence $\{X_n\}$ converges to a random variable μ in the mean square sense since

$$E((X_n - \mu)^2) = \frac{\sigma^2}{n} \qquad \Longrightarrow \qquad \lim_{n\to\infty} \frac{\sigma^2}{n} = 0 \tag{41}$$

Almost sure convergence

Almost sure convergence is relaxed version of pointwise convergence except that it does accommodate points in Ω where the sequence does not converge.

However, the points $\omega \in \Omega$ on which $\{X_n(\omega)\}$ does not converge pointwise to $X(\omega)$ should be **zero-probability events**.

In other words, **sequences should converge over an interval** (of arbitrarily finite length) in Ω unlike at every point that is required in pointwise convergence.

Then, we write

$$\boxed{X_n \xrightarrow{a.s.} X} \tag{42}$$

Example: a.s. convergence

Consider a sample space $\Omega = [0, 1]$ and a sequence $\{X_n(\omega)\}$ constructed on Ω as

$$X_n(\omega) = \begin{cases} 1, & \omega = 0 \\ \frac{1}{n}, & \omega \neq 0 \end{cases} \tag{43}$$

Examine if the sequence a.s. converges to a (constant) random variable $X(\omega) = 0$.

Deduce that $\lim\limits_{n\to\infty} X_n(\omega) = \begin{cases} 1, & \omega = 0 \\ 0, & \text{otherwise} \end{cases}$. Note that $\Pr(\omega = 0) = 0$.

Clearly X_n converges to the given random variable $X(\omega)$ except for the event $\omega = 0$, **which is a zero-probability event.** Hence, $X_n(\omega)$ converges to $X(\omega)$ almost surely.

Example 1: Consistency

Sample mean

The sample mean estimator for a WN process has the MSE

$$\text{MSE}(\bar{y}) = \text{var}(\bar{y}) = \frac{\sigma_e^2}{N} \tag{44}$$

This is obviously a m.s. consistent estimator since its MSE $\to 0$ as $N \to \infty$.

Example 1: Consistency

Sample variance

The biased estimator of the variance of a random process was shown to be earlier asymptotically unbiased. For a GWN process with variance σ_e^2, this estimator is known to have a variance

$$\text{var}(\hat{\sigma}_N^2) = \frac{2(N-1)\sigma_e^4}{N^2} \tag{45}$$

Therefore, it is mean-square consistent.

Remarks

- Consistency essentially guarantees that increasing the number of observations takes us "closer" to the true value. Therefore, it is practically one of the most important properties of an estimator.
- There are several estimators that are not consistent. A popular one is the *periodogram*, which estimates the power spectral density of a signal.
- For biased estimators, mean square consistency also implies asymptotic unbiasedness because

$$\text{MSE}(\hat{\theta}) = \text{bias}^2 + \text{var}(\hat{\theta})$$

Running summary

To recap the key points until now:

- The goodness (accuracy, precision, etc.) of an estimate depends on two factors: (i) information content in the data and (ii) properties of the estimator
  - Information content is measured by Fisher's information, which is based on the likelihood function. It is a measure of the quality of a given dataset w.r.t. estimating θ and is regardless of the form of $\hat{\theta}$.
- Six properties of an estimator are usually important: *bias*, *variance, efficiency, mean square error, asymptotic bias and consistency*.
  - Efficiency and consistency are the two most important criteria

Running summary ...contd.

To recap the key points until now:

- C-R inequality gives us the lowest variability (error) that can be achieved by an unbiased estimator. The bound is the inverse of FI.
- Whenever it becomes difficult to estimate or design a 100% efficient estimator or even a MVUE, **a best linear unbiased estimator** is sought.
- Consistency guarantees convergence of the estimate to the true value

Motivation

Up to this point we studied metrics for quantifying the goodness of data and estimators. Now we raise an important question:

Given an observation vector $\mathbf{y}$ and a *point* estimate $\hat{\theta}$ what can be said about the true value θ_0?

Interval estimates and hypothesis testing

Two related problems are:

1. **Confidence intervals:** What is the interval in which the true value resides? Only intervals are sought since the true value cannot be estimated precisely.
2. **Hypothesis testing:** Given $\hat{\theta}$ how do we test claims on the true parameters?

To be able to answer the above questions it is necessary to determine the probability distribution of an estimate.

Introductory remarks

The distribution of estimate generally depends on three factors:

1. *Randomness in observations:* It is a crucial factor since it is the "feed" to the estimator. It is the source of uncertainty in estimate.
2. *Form of estimator:* When the estimator is linear (e.g., sample mean, BLUE estimator) the transformation of the $f(\mathbf{y};\boldsymbol{\theta})$ can be easily studied. Non-linear estimators naturally pose a challenge, except under very special conditions.
3. *Sample size:* A large body of estimation literature is built on the large sample size assumption. Small sample sizes not only affect the distribution but also the consistency property of an estimator!

Multiplication by $\sqrt{N}$

Distributions are quite often stated for $\sqrt{N}(\hat{\theta} - \theta_0)$ (or at times $\sqrt{N}\hat{\theta}_N$) instead of $\hat{\theta}_N$ itself. This is because asymptotically $(\hat{\theta}_N - \theta_0)$ converges to a constant (mostly zero), whereas $\sqrt{N}(\hat{\theta} - \theta_0)$ converges to a random variable with a meaningful distribution.

Example

From the previous sections, we know that the sample mean is a consistent estimator of the mean. This means $\bar{y}_N - \mu_y$ converges to zero as $N \to \infty$. On the other hand, $\sqrt{N}(\bar{y}_N - \mu_y)$ converges to a random variable with mean zero and finite variance.

$$E(\hat{\theta}_N) = \mu_y; \quad \text{var}(\bar{y}) = \frac{\sigma_y^2}{N} \quad \Longrightarrow \quad (\bar{y} - \mu_y) \xrightarrow{m.s.} 0$$

$$\text{but,} \quad \sqrt{N}(\bar{y} - \mu_y) \xrightarrow{m.s.} \sigma_y^2$$

Convergence in distribution

In order to study the asymptotic distributional properties of an estimator, it is necessary to first understand the notion of **convergence in distribution** of a sequence of RVs.

Definition

A sequence of random variables $\{X_N\}$, each possessing a distribution function $F(x_N)$ **converges in distribution** if the sequence of those distributions $\{F(x_N)\}$ (sometimes written as $F_N(x)$) converges to a distribution function $F(x)$. The random variable X associated with $F(x)$ is said to be the *limit in distribution* of the sequence, indicated as

$$X_n \xrightarrow{d} X \tag{46}$$

Note that the theorem speaks of convergence of distributions, not the RVs themselves.

Central Limit Theorem

The CLT is one of the most celebrated and landmark results in estimation theory. Historically it is nearly seven decades old and has undergone several modifications. The basic version due to Lindeberg and Levy is as follows.

Theorem (Central Limit Theorem)

The uniformly weighted sum of N independent and identically distributed (i.i.d.) random variables $\{X_n,\ n = 1, \cdots, N\}$

$$\bar{X} = \sum_{n=1}^{N} \frac{X_n}{N} \tag{47}$$

converges in distribution as

$$\sqrt{N}\left(\bar{X} - \mu\right) \xrightarrow{d} \mathcal{N}(0, 1) \tag{48}$$

CLT ... contd.

- The conditions of independence and identical distributions are not heavily restrictive. Versions of CLT which place some minor additional requirements on the moments or the autocorrelation functions are also available.
- Generalizations and extensions to other random objects such as matrices and polytopes are available.
- Note that the sum is none other than the sample mean of the N random variables.
- It is also conventional to state that the distribution of $\bar{X}$ is *asymptotically normal* or simply write as

$$\sqrt{N}(\bar{X} - \mu) \sim A\mathcal{N}(0, \sigma^2) \tag{50}$$

- Sums of transformed RVs that have more symmetric distributions and have lighter

Limitations

- The CLT provides us with a tool for deriving distributions of parameter estimates from linear estimators with known distributional properties of the data. Many standard results on distributions of estimators such as sample mean, sample variance, linear least squares estimates can be derived through CLT.
- However, when the estimator is a complicated function of the observations, further simplifying approximations or the use of modern (Monte-Carlo) methods such as bootstrapping or surrogate data analysis have to be employed.

Confidence regions

The term "confidence region" essentially refers to the interval containing the true value with $< 100\%$ confidence. Ideally one would like to have a narrow interval with maximum confidence. However, these are conflicting requirements because a higher degree of confidence is associated with a wider band.

The procedure for constructing a confidence interval is a two-step process.

Step 1: Construct a probabilistic interval for the error $\hat{\theta}_N - \theta_0$ using the knowledge of the distribution (or density) of $\hat{\theta_N}$, the bias and variance properties and the specified degree of confidence $100(1-\alpha)\%$.

Step 2: Convert this *probabilistic* interval into a *confidence* region for θ_0 by an algebraic manipulation.

Confidence interval for mean

Assume that the sample mean $\bar{y}$ is used as an estimator of the mean μ_y from a single record of data.

Goal: To obtain a confidence region for μ_y

Assume that σ_y^2 is known. Invoking CLT, $\sqrt{N}\left(\dfrac{\bar{y}-\mu_y}{\sigma_y}\right) \sim \mathcal{N}(0,1)$

Confidence interval for mean ... contd.

From the properties of a Gaussian distribution,

$$-1.96 \leq \sqrt{N}\frac{\bar{y}-\mu_y}{\sigma_y} \leq 1.96 \qquad \text{(with 95\% probability)}$$

$$\Longrightarrow \boxed{\mu_y \in [\bar{y} - \frac{1.96}{\sqrt{N}}\sigma_y, \bar{y} + \frac{1.96}{\sqrt{N}}\sigma_y]} \qquad \text{(with 95\% confidence)} \qquad (51)$$

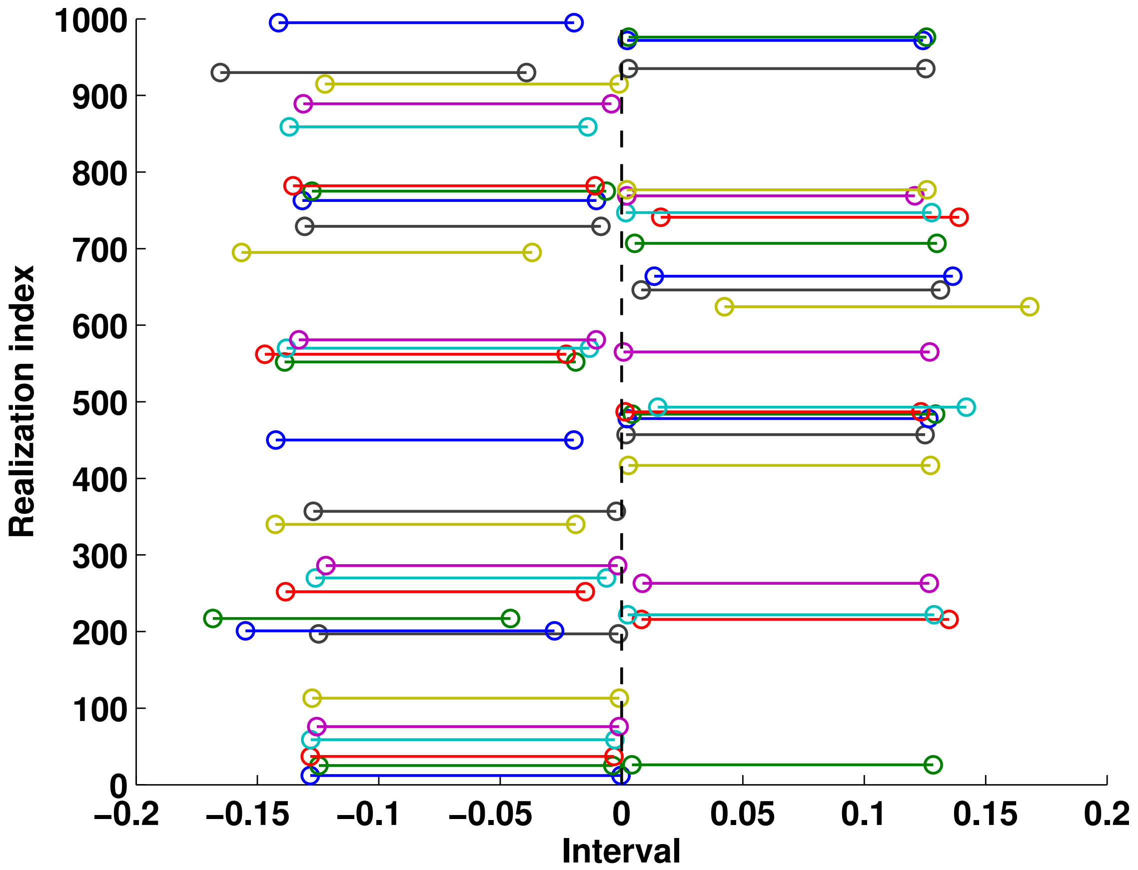
The $100(1-\alpha)\%$ CI for the mean is obtained by replacing 1.96 with ζ_c such that Pr$(\zeta > \zeta_c) = \alpha/2$ (using the standard Gaussian distribution).

Interpretation

The confidence interval (CI) should be interpreted with care. Consider the case of a 95% CI for mean. Suppose that we have 1000 records of data, from each of which we can obtain an estimate $\bar{y}^{(i)}, \quad i = 1, \cdots, 1000$, from each of which a 95% C.I. can be constructed. Then, out of 1000 such CIs, roughly 950 intervals would have correctly captured the true mean.

Simulation example

In one such simulation study, it turns out that 51 intervals do not contain the true value $\mu_0 = 0$ as shown below.



Remarks

- The width of the CI is only dependent on the standard error in the estimate $\sigma_y/\sqrt{N}$. In general, **the width depends on the variability of the process and the sample size** (for a consistent estimator)
- Narrower the width of the interval at a fixed α, better is the estimator. A consistent estimator produces zero-width CI asymptotically.
- For correlated processes, the CI has to be re-derived because var$(\bar{y})$ is influenced by the correlation structure.

Confidence intervals

1. **Mean:** Small sample, variance unknown.

$$\boxed{\mu_y \in [\bar{y} - t_{\alpha/2}(N-1)\hat{\sigma}_y, \bar{y} + t_{\alpha/2}(N-1)\hat{\sigma}_y]} \quad \text{(with 95\% confidence)} \tag{52}$$

2. **Variance:** Gaussian population, random samples

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \tag{53}$$

One-sided: Lower and upper confidence bounds

$$\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}} \leq \sigma^2, \qquad \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}} \tag{54}$$

Remarks ... contd.

Several standard texts on statistics present the theory of hypothesis testing and confidence interval construction (see Johnson, 2011; Ogunnaike, 2010).

For non-linear estimators, modern empirical methods are used to obtain the distributions of estimates via the generation of **surrogate data** or *pseudo-population*. Monte-Carlo simulations and boostrapping methods are increasingly being used for this purpose.

Hypothesis testing

Once the distributional properties of an estimator are known, it is possible to answer the two questions raised earlier, i.e., pertaining to hypothesis testing and confidence interval construction. Both are in fact related problems.

Hypothesis testing involves a statistical test for a claim made by the analyst with regards to the properties of the process of interest or model parameters using the observations as an evidence.

Examples:

- Average temperature of a reactor is at a specified value.
- Model parameters are truly zero.
- The given series is white (unpredictable)

Procedure for hypothesis testing

A hypothesis test typically consists of the following steps.

1. *Formulate the null hypothesis* H_0 based on the postulate or the claim. Choose an appropriate alternate hypothesis H_a.
2. *Choose an appropriate statistic* ζ for the test. The statistic is generally a linear or non-linear function of the parameter(s) involved in the hypothesis.
3. *Compute the test statistic* from the given observations. Denote this by ζ_o.
4. *Make a decision.* Retain or discard the null hypothesis by applying a certain criterion to the observed statistic (three different approaches).

No hypothesis test can result in a perfect decision!

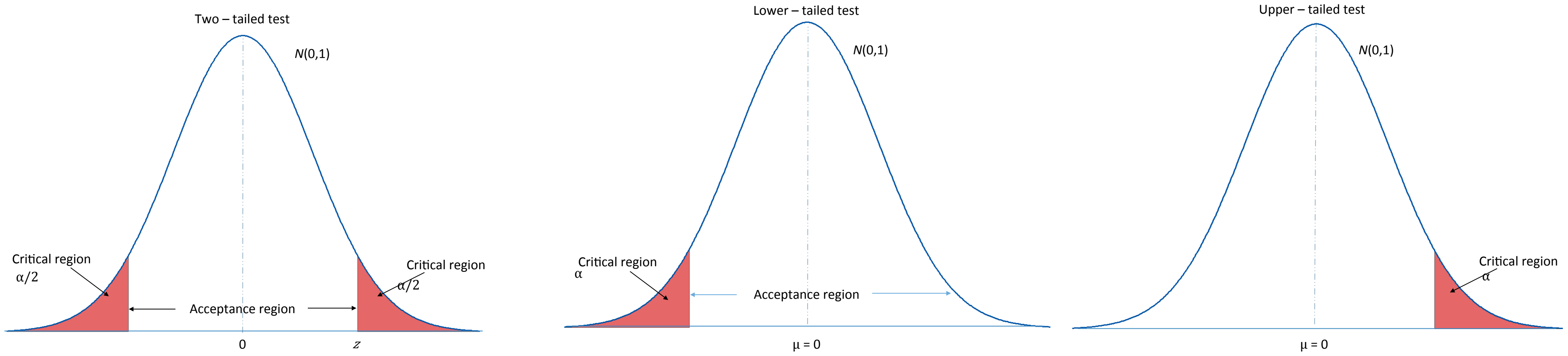
Errors in hypothesis testing

Any hypothesis test is marred by two errors - Type I and II errors. Typically, the first type, known as the α risk or the **significance level** is specified.

Decision $\longrightarrow$ Truth $\downarrow$	Fail to Reject H_0	Reject H_0
H_0 True	Correct Decision Probability: $1-\alpha$	**Type I Error** Probability (Risk): α
H_a True	**Type II Error** Risk: β	Correct Decision Probability: $1-\beta$

One of the two errors has to be specified for making a decision in hypothesis testing. **It is a common practice to specify the α risk.**

Graphical understanding: One sample test for mean



(a) Two tailed test (b) Lower-tailed test (c) Upper-tailed test

The α risk (probability of making Type I error) depends on the type of **alternate hypothesis** and the **sampling distribution** of the statistic.

Decision making in hypothesis testing

There are **three** different approaches to making decisions in hypothesis testing, *all of which lead to the same result.*

1. **Critical value approach:** Determine a critical value (for a given risk) and compare the observed statistic against it.
2. p-**value method:** Determine the probability of obtaining a value more extreme than the observed and compare this probability against a user-specified value (risk).
3. **Confidence interval approach:** Construct the confidence region (for a given risk) and determine if the postulated value falls within the region.

p-value

The p-value is the probability of observing a more extreme value of the statistic than the observed value.

- It is computed under the null hypothesis being held to be true.
- The sign or the direction of the extreme value depends on the alternate hypothesis just as the way the critical value does.
- The p-value computation is fairly straightforward. For example, in an upper tail test, the p-value is the $\Pr(\zeta \geq \zeta_o)$. If the p-value $\leq \alpha$, then H_0 is rejected.

Example: Hypothesis testing

An engineer measures the (controlled) temperature of a reactor over a period of 3 hours at a sampling interval of $T_s = 15$ sec. The sample average of the $N = 720$ readings is calculated to be $\bar{y} = 90.1826$°C. Based on this observation, the engineer claims that the temperature is at its set-point $T_0 = 90$°C on the average.

To test this claim, the formal hypothesis test is set up as follows.

$$H_0 : \mu_y = 90 \quad H_a : \mu_y \neq 90 \qquad \text{two-tailed test}$$

Example: Hypothesis testing ...contd.

Assume that the temperature series has white-noise characteristics. Then we know that for the large sample case,

$$\sqrt{N}\left(\frac{\bar{Y}-\mu}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0,1)$$

An appropriate test statistic suited for the purpose is therefore,

$$Z = \frac{\bar{y}-\mu_0}{\sigma/\sqrt{N}} \tag{55}$$

where μ_0 is the true value assumed in H_0. For the example, $\mu_0 = T_0 = 90^\circ$C. Assume that σ is known to be 2°C. Then the observed statistic is $z_0 = 2.45$.

Example ...contd.

Decision making

i. **Critical value approach:** The critical value at $\alpha = 0.05$ is $z_c = 1.96$. Since $z_o > z_c$, the null hypothesis is rejected.

ii. p-**value approach:** The $\Pr(|Z| > z_0 = 2.45) = 0.0143 < \alpha = 0.05$. Hence H_0 stands rejected in favour of H_a.

iii. **C.I. approach:** The $100(1-\alpha)$ C.I. for the average temp. is $(90.0365, 90.3287)$, which does not include the postulated value. Hence H_0 stands rejected in favour of H_a.

On the average the temperature is not at its set-point, i.e., the engineer's claim that $H_0 : \mu = 90^\circ$ (set-point), stands rejected in favour of the alternate hypothesis.

Remarks

- How would you adjust the significance level just enough so that H_0 is not rejected?
- **Note:** The assumption of known σ can be relaxed. It can be estimated σ from data, in which case ζ has a t-distribution with $\nu = N - 1$ degrees of freedom.

Confidence intervals in Hypothesis testing

In a general situation, the C.I. approach for testing hypothesis (all three forms) is as follows :

Procedure

1. Specify the significance level.
2. Depending on the hypothesis, construct the appropriate C.I. and apply the test
   - 2.1 **Two-sided:** $100(1-\alpha/2)\%$ C.I. If postulated value is not within the C.I., reject H_9.
   - 2.2 **One-sided:** Reject H_0 if the postulated value is greater or lesser than the bound, for the upper- and lower-tailed test, respectively.

Summary

- The end goal of an estimation exercise is to arrive at an interval estimate or a confidence region for the true value
- Distributions of estimates provide the necessary information to move from a point to an interval estimate
- The distributional properties also facilitate hypothesis testing, which involves a systematic and statistical way of testing the claims related to a process and/or a model
- Arriving at distributions is non-trivial, but the CLT comes to the rescue for linear estimators
- The case of non-linear estimation is more complicated and calls for the use of modern tools such as Monte-Carlo simulations and Bootstrapping methods.

Bibliography I

- Fisher, R. (1922). "On the mathematical foundations of theoretical statistics". In: *Philosophical Transactions Royal Society London A* 222, pp. 309–368.
- – (1950). *Contributions to mathematical statistics*. New York, USA: John Wiley & Sons, Inc.
- Johnson, R. A. (2011). *Miller and Freund's: Probability and Statistics for Engineers*. Upper Saddle River, NJ, USA: Prentice Hall.
- Ogunnaike, B. A. (2010). *Random Phenomena: Fundamentals of Probability and Statistics for Engineers*. Boca Raton, FL, USA: CRC Press, Taylor & Francis Group.
- Tangirala, A. K. (2014). *Principles of System Identification: Theory and Practice*. Boca Raton, FL, USA: CRC Press, Taylor & Francis Group.