Basic concepts of point estimation

Contents

1 Statistic, estimator and estimation

Definition 1.1. Let $X = (X_1, \ldots, X_n)$ be a sample. Any (measurable) function $T(X)$ of X, i.e. any quantity that can be calculated solely from the observed data, is called a **statistic**.

An estimator is any statistic used to estimate a given parameter. Typically, we use the notation $\hat{\theta}_n(X) \equiv \hat{\theta}_n \equiv \hat{\theta}$ to denote an **estimator** of θ.

Any realization $\hat{\theta}_n(x)$ of $\hat{\theta}_n(X)$ is an **estimation** (a guess) of θ.

Example 1.1 (Example of statistics).

X_1, $(X_1, \ldots, X_n)$, $\sum_i X_i$, $\overline{X}_n = n^{-1} \sum_i X_i$, $n^{-1} \sum_i X_i^2$, $n^{-1} \sum_i I(X_i \geq 0)$,

$X_{(1)} = \min_i X_i$, $X_{(n)} = \max_i X_i$, $X_{(k)}$ the kth order statistic, i.e. kth-smallest value,

$\tilde{\sigma}_n^2 = n^{-1} \sum_i (X_i - \mu)^2$ (assuming that μ is known), $S_n^2 = (n-1)^{-1} \sum_i (X_i - \overline{X}_n)^2$, $(\overline{X}_n, S_n^2)$,

$\arg\min_a \sum_i |X_i - a|$. $\square$

To evaluate the performance of an estimation procedure, one examines the theoretical properties of the estimator on which it is based. Since an estimator is a random variable, it possesses a probability distribution, referred to as the *sampling distribution*. The analytical characteristics of this distribution are fundamental for determining the adequacy of an estimator in a given inferential context and for identifying those that should be avoided due to undesirable statistical properties.

Example 1.2 (Uniform model).

Consider the uniform density $f(x; \theta) = \frac{1}{\theta} I(0 \leq x \leq \theta); \ \theta > 0$.

As an estimator of θ, based on a sample $X_1, \ldots, X_n$, one can, for example, consider one of the following :

$$\hat{\theta}_1 = X_{(n)}, \quad \hat{\theta}_2 = \frac{n+1}{n} X_{(n)}$$

$$\hat{\theta}_3 = X_{(1)} + X_{(n)}, \quad \hat{\theta}_4 = 2\overline{X}_n$$

$$\hat{\theta}_5 = 2\hat{q}_{0.5},$$

where $\hat{q}_{0.5}$ is the sample median, i.e.

$$\hat{q}_{0.5} = \begin{cases} X_{(k+1)} & \text{if } n = 2k+1 \text{ is odd,} \\ \frac{X_{(k)} + X_{(k+1)}}{2} & \text{if } n = 2k \text{ is even.} \ \square \end{cases}$$

3

This example suggests questions like:

- If many estimators are available—as is always the case—how can we compare them ?
- Are there any general and reliable methods for constructing estimators ?
- Given a statistical model, does a "best estimator" exist, and if so, how can it be identified ?

These questions (and many others of the same nature) will be the subject of our next readings.

2 Risk and loss function

It seems reasonable that we want an estimate $\hat{\theta}$ which generally comes quite close to the true value of θ, and dislike an estimate $\hat{\theta}$ which generally misses the true value of θ by a large amount. The question is how to make this precise and quantifiable?

We quantify the idea of $\hat{\theta}$ being close to θ, by measuring the **risk**, that is the *average distance*, between these two quantities. The distance is measured using what is called a **loss function**.

Examples of loss functions include:

- squared error loss: $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$
- absolute error loss: $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$
- absolute relative loss: $L(\hat{\theta}, \theta) = |\hat{\theta}/\theta - 1|$

Once the loss function is chosen, we calculate the risk as follows

$$E_\theta(L(\hat{\theta}, \theta)) = \int L(\hat{\theta}(x), \theta) f_n(x, \theta) dx,$$

where E_θ means that *the expectation is taken under the assumption that θ is the true parameter*; that is, the pd of X is $f_n(x, \theta)$.

3 MSE, bias, variance, and relative efficiency

When the squared-error loss is used, the risk associated with an estimator is precisely its *mean squared error*:

$$MSE_\theta(\hat{\theta}) := E_\theta[(\hat{\theta} - \theta)^2]$$
$$= E_\theta\left[(\hat{\theta} - E_\theta(\hat{\theta})) + (E_\theta(\hat{\theta}) - \theta)\right]^2$$
$$= Bias_\theta^2(\hat{\theta}) + Var_\theta(\hat{\theta}),$$

where $Bias_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$ is the bias of the estimator $\hat{\theta}$.

A large bias indicates low accuracy ($\hat{\theta}$ lies far from θ, i.e. some systematic error), while a large variance indicates low precision (too much fluctuation). If the bias of $\hat{\theta}$ is always zero, i.e. $Bias_\theta(\hat{\theta}) = 0 \ \forall \theta \in \Theta$, then $\hat{\theta}$ is called **unbiased**. This means that **on average** the estimator will yield the true value of the unknown parameter (whatever the true value is). *In this case, the MSE reduces to variance.*

Attention. In the following, in order to ease the notation, if no confusion is possible, we drop the index θ and write E, Var and MSE instead of E_θ, Var_θ and MSE_θ, respectively.

The choice of an estimator is very often restricted to the class of unbiased estimators. But there are cases where a small bias is accepted, in particular if the bias converges to zero when the sample size tends to infinity. Moreover, there are cases where no unbiased estimator exists. Example: $X_1, \ldots, X_n \sim Bernoulli(p)$, and $\theta = \dfrac{1}{p}$.

Example 3.1 (Unbiased does not necessarily mean a good estimator).

Let X_i, $i = 1, \ldots, n$, be iid rv with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Then X_1, $\overline{X}_n$ and $\frac{X_1 + \overline{X}_n}{2}$ are unbiased estimators of μ. Which one should we use?

It is clear that all three are unbiased. So to compare these estimators, we have to compare their variances (i.e. their MSE). It easy to see that

$$Var(X_1) = \sigma^2, \ Var(\overline{X}_n) = \frac{\sigma^2}{n}, \ \text{and} \ Var\left(\frac{X_1 + \overline{X}_n}{2}\right) = \frac{1}{4}\left(1 + \frac{3}{n}\right)\sigma^2.$$

To check the last equality, observe that

$$Var(X_1 + \overline{X}_n) = Var\left(\frac{n+1}{n}X_1 + \frac{1}{n}\sum_{i=2}^{n}X_i\right) = \frac{(n+1)^2}{n^2}\sigma^2 + \frac{n-1}{n^2}\sigma^2 = \left(1 + \frac{3}{n}\right)\sigma^2.$$

$\rightarrow \overline{X}_n$ is better than the other two. $\square$

Bias and transformation

If $\hat{\theta}$ is an unbiased estimator of θ, then for any constants a and b, the transformed estimator $a + b\hat{\theta}$ remains unbiased for $a + b\theta$; in other words, *unbiasedness is preserved under linear transformations*. This property allows us to correct certain biased estimators: if $E(\hat{\theta}) = a\theta + b$, then the adjusted estimator $(\hat{\theta} - b)/a$ is unbiased for θ. In contrast, nonlinear transformations do not generally preserve unbiasedness: even when $\hat{\theta}$ is unbiased for θ, a nonlinear function $g(\hat{\theta})$ will typically fail to be unbiased for $g(\theta)$. For instance, although the sample mean

$\overline{X}_n$ is an unbiased estimator of μ, Jensen's inequality implies that, provided $\overline{X}_n$ is not almost surely constant, $\left(E\,\overline{X}_n\right)^2 = \mu^2 < E\left(\overline{X}_n^2\right)$; consequently, $\overline{X}_n^2$ is a biased estimator of μ^2. $\square$

Example 3.2. Let X_i, $i = 1, \dots, n$, be iid rv with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Let's find an unbiased estimator for μ^2. To do so, observe that

$$E\left(\overline{X}_n^2\right) = \left(E\left(\overline{X}_n\right)\right)^2 + Var\left(\overline{X}_n\right) = \mu^2 + \frac{\sigma^2}{n}.$$

This means that $Bias\left(\overline{X}_n^2\right) = \sigma^2/n$. This also implies that an unbiased estimator of μ^2 is given by

$$\overline{X}_n^2 - \frac{S_n^2}{n},$$

provided that S_n^2 is an unbiased estimator of σ^2 (see the next example). $\square$

Example 3.3 (The sample variance). Let X_i, $i = 1, \dots, n$, be iid rv with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. A natural estimator of this latter is given by $\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. We have that

$$E(\hat{\sigma}_n^2) = Var\left(X_1 - \overline{X}_n\right) = Var\left(\frac{n-1}{n}X_1 - \frac{1}{n}\sum_{i=2}^n X_i\right) = \frac{n-1}{n}\sigma^2$$

With the correction factor $\frac{n}{n-1}$ we obtain an unbiased estimator, namely the well-known empirical variance (or sample variance):

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2. \; \square$$

The MSE is a global measure of estimation quality that combines variance and squared bias, two components of different nature. *Comparing the MSE of two estimators is meaningful only when their biases are comparable* (i.e., both are unbiased or have roughly the same bias); otherwise, differences in MSE become difficult to interpret. Moreover, comparing MSEs is valid only when the two estimators address the same estimation problem—estimating the same parameter from the same data, under the same model and with the same sample size.

A standard tool for such comparison is the **relative efficiency** :

$$RE(\hat{\theta}_1, \hat{\theta}_2) = \frac{MSE(\hat{\theta}_2)}{MSE(\hat{\theta}_1)}.$$

This quantity expresses the performance of $\hat{\theta}_2$ relative to $\hat{\theta}_1$. When both estimators are unbiased, the relative efficiency reduces to

$$RE(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}.$$

If this quantity is less than one $\forall \theta \in \Theta$, then $\hat{\theta}_1$ has uniformly a larger variance than $\hat{\theta}_2$, and the latter is said to be *more **efficient*** than the former.

Example 3.4.

Let $X_1, \ldots, X_n$, be an iid sample form $Unif[0, \theta]$. Let $\hat{\theta}_1 = 2\overline{X}_n$ and $\hat{\theta}_2 = \dfrac{n+1}{n} X_{(n)}$, two estimators of θ.

Recall that an uniform distribution in $[a, b]$ is characterized by its cdf

$$F(x) = \begin{cases} 0, & \text{if } x < a \\ \frac{x-a}{b-a}, & \text{if } a \leq x \leq b \\ 1, & \text{if } x > b. \end{cases}$$

Its mean and variance are given by

$$E(X) = \frac{a+b}{2} \text{ and } Var(X) = \frac{(b-a)^2}{12}.$$

From this, it follows that $\hat{\theta}_1$ is unbiased and $Var(\hat{\theta}_1) = \frac{\theta^2}{3n}$.

To find the expectation and variance of $\hat{\theta}_2$ first observe that

$$P(X_{(n)} \leq x) = (P(X_1 \leq x))^n.$$

So the cdf of $X_{(n)}$ is given by

$$F_{X_{(n)}}(x) = \begin{cases} 0, & \text{if } x < 0 \\ (x/\theta)^n, & \text{if } 0 \leq x \leq \theta \\ 1, & \text{if } x > \theta. \end{cases}$$

Hence, the pd of $X_{(n)}$ is given by $f_{X_{(n)}}(x) = n\dfrac{x^{n-1}}{\theta^n}I(0 \leq x \leq \theta)$. It follows that $E(X_{(n)}) = \dfrac{n}{n+1}\theta$ and $E\left(X_{(n)}^2\right) = \dfrac{n}{n+2}\theta^2$. Therefore, $\hat{\theta}_2$ is unbiased and its variance is $Var(\hat{\theta}_2) = \dfrac{\theta^2}{n(n+2)}$. Finally, the relative efficiency is

$$\frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)} = \frac{3}{n+2}.$$

Indicating that for $n > 1$, $\hat{\theta}_2$ is more efficient than $\hat{\theta}_1$. $\square$

Note that, in the example above, the relative efficiency does not depend on the parameter θ. This feature is specific to that setting and does not hold in general; a point to which we will return in the subsequent section after the next.

Remark. To compare estimators that are not defined on the same scale, the Mean Absolute Relative Error (MARE)

$$MARE(\hat{\theta}) = E\left(\left|\frac{\hat{\theta} - \theta}{\theta}\right|\right)$$

is particularly useful because it expresses the average absolute relative deviations making the metric **unit-free** and comparable across heterogeneous settings. Its interpretation is straightforward: for example, a MARE of 0.10 indicates that, *on average*, the estimator deviates from the target by 10%, with lower values reflecting better performance and higher values indicating proportionally larger errors. □

4 Simulation-Based Evaluation of Estimator Performance

A practical way to assess the performance of an estimator is to approximate its bias, variance, MSE, etc., through simulation. The idea is to repeatedly generate data from a model with a known parameter value θ, compute the estimator $\hat{\theta}$ for each simulated dataset, and then summarize the resulting collection of estimates. If $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(N)}$ denote the estimates obtained from N independent simulations, the empirical bias, empirical variance, empirical

MSE, and empirical MARE are given, respectively, by

$$\widehat{\text{Bias}} = \bar{\hat{\theta}} - \theta$$

$$\widehat{\text{Var}} = \frac{1}{N} \sum_{s=1}^{N} \left(\hat{\theta}^{(s)} - \bar{\hat{\theta}} \right)^2$$

$$\widehat{\text{MSE}} := \frac{1}{N} \sum_{s=1}^{N} \left(\hat{\theta}^{(s)} - \theta \right)^2 = \widehat{\text{Bias}}^2 + \widehat{\text{Var}}$$

$$\widehat{\text{MARE}} := \frac{1}{N} \sum_{s=1}^{N} \left| \frac{\hat{\theta}^{(s)}}{\theta} - 1 \right|$$

where $\bar{\hat{\theta}} = \frac{1}{N} \sum_{s=1}^{N} \hat{\theta}^{(s)}$ is the average of the simulated estimates.

The number of simulations N must be chosen sufficiently large to ensure that these empirical estimates are stable and accurately reflect their theoretical counterparts.

To obtain a reliable picture of the estimator's performance, and since these measures typically depend on the value of θ, one must repeat the entire simulation procedure across a broad range of parameter values. This can become computationally demanding, especially when the estimator itself is costly to compute.

The R code below demonstrates the simulation procedure outlined above, applied to the setting of Example 3.4.

```r
mse <- function(estimates, theta) {
  mean <- mean(estimates)
  bias <- mean - theta                      # empirical bias
  var <- mean((estimates - mean)^2)         # empirical variance
  mse <- bias^2 + var                       # empirical mean squared error
  mare <- mean(abs(estimates / theta - 1))  # empirical mean absolute relative error
  c(bias = bias, var = var, mse = mse, mare = mare)
}


estSim <- function(dataFun, estFun, N = 10000) {
  estimates <- replicate(N, dataFun() |> estFun())
  return(estimates)
}
```

- Performances of $\hat{\theta}_1$:

```r
n <- 10; theta <- 3
estSim(dataFun = \() runif(n, min = 0, max = theta),
       estFun = \(data) 2 * mean(data)) |> mse(theta)
```

14

```
    bias      var       mse      mare
0.000695 0.301129 0.301130 0.146706
```

- Performances of $\hat{\theta}_2$:

```r
n <- 10; theta <- 3
estSim(dataFun = \() runif(n, min = 0, max = theta),
       estFun = \(data) ((n + 1) / n) * max(data)) |> mse(theta)
```

```
   bias     var      mse     mare
-0.0049  0.0772   0.0773   0.0706
```

These results closely match the theory which tells us that, for $n = 10$, the MSE of $\hat{\theta}_1$ is $3^2/(3 \times 10) = 0.3$ and that the MSE of $\hat{\theta}_2$ is $3^2/(10 \times (10+2)) = 0.075$.

5 The best unbiased estimator (MVUE)

It is natural to prefer estimators that have a small MSE.

If an estimator $\hat{\delta}$ has a larger MSE than another estimator $\hat{\theta}$ for every possible value of the parameter – that is, if

$$MSE_\theta(\hat{\theta}) \leq MSE_\theta(\hat{\delta}), \ \forall \theta \in \Theta,$$

then $\hat{\delta}$ is called ***inadmissible***.

One might hope to find an estimator that has the smallest possible MSE for every value of θ. However, this turns out to be impossible. For an estimator $\hat{\theta}$ to be the "best" for every θ, it would need to satisfy

$$E_\theta\big[(\hat{\theta} - \theta)^2\big] = 0 \quad \forall \theta \in \Theta.$$

But the only way for this expectation to be zero is if $\hat{\theta} = \theta$ with probability 1. This would mean the estimator always returns the true parameter exactly, which is impossible in any non-trivial statistical model.

Typically, two estimators $\hat{\theta}$ and $\hat{\delta}$ are not uniformly comparable. In general, one can find parameter values $\theta_1, \theta_2 \in \Theta$ such that

$$\text{MSE}_{\theta_1}(\hat{\theta}) < \text{MSE}_{\theta_1}(\hat{\delta}) \quad \text{and} \quad \text{MSE}_{\theta_2}(\hat{\theta}) > \text{MSE}_{\theta_2}(\hat{\delta}).$$
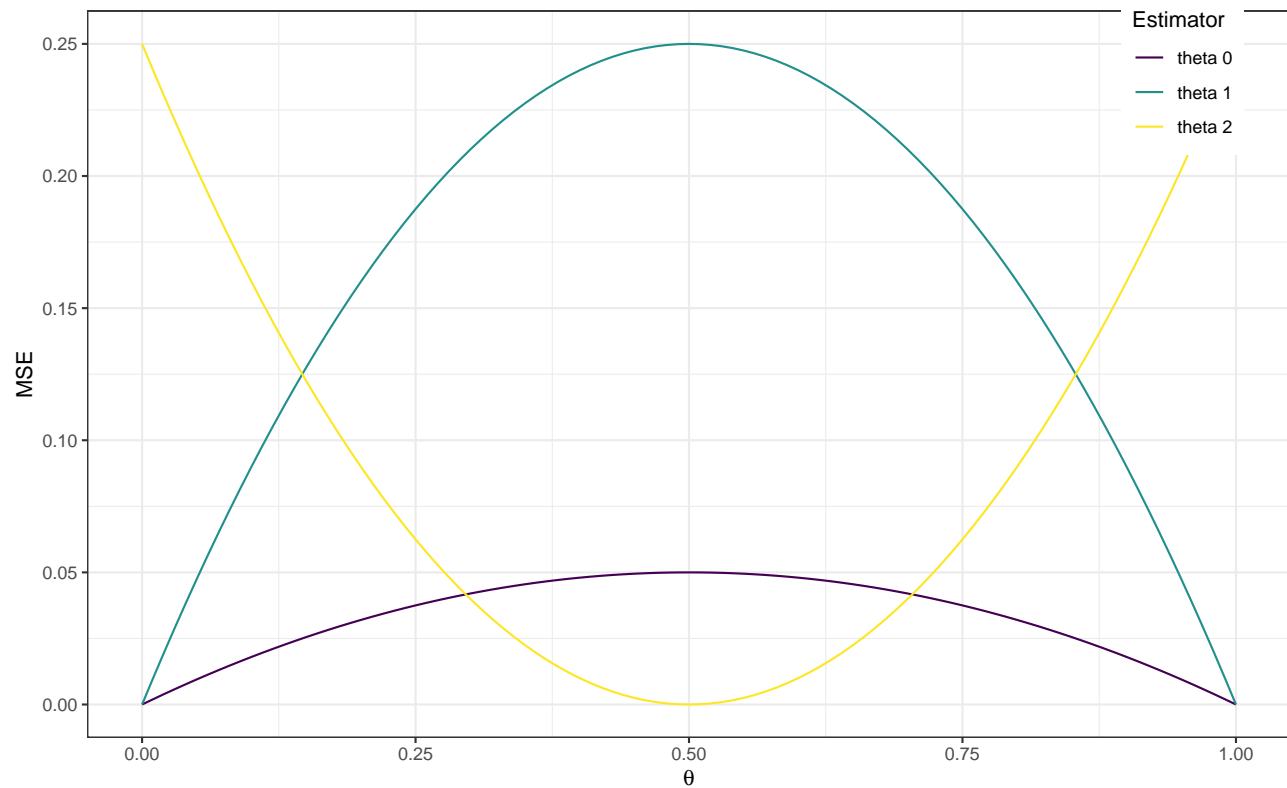
This means that each estimator performs better for some values of θ and worse for others, so neither dominates the other uniformly.

Example 5.1.

Suppose an iid simple $X_1, \ldots, X_n$ from a Bernoulli distribution with an unknown parameter θ, $0 < \theta < 1$. Let $\hat{\theta}_0 = \overline{X}_n$, $\hat{\theta}_1 = X_1$ and $\hat{\theta}_2 = 1/2$. It is easy to see that

$$MSE(\hat{\theta}_0) = \frac{\theta(1-\theta)}{n}, \quad MSE(\hat{\theta}_1) = \theta(1-\theta), \quad MSE(\hat{\theta}_2) = (\theta - 1/2)^2$$

These three MSE (quadratic risk) functions, of θ, are plotted below (for $n = 5$).

$\hat{\theta}_1$ is inadmissible as it is uniformly less efficient than $\hat{\theta}_0$. The estimators $\hat{\theta}_0$ and $\hat{\theta}_2$ are not uniformly comparable. Near $\theta = 1/2$, $\hat{\theta}_2$ is the best, and away from $\theta = 1/2$, $\hat{\theta}_0$ is the best. □

In view of the fact that there is no estimator that has the smallest possible MSE for every value of the parameter,

statisticians adopt other strategies to choose good estimators. One such strategy is to restrict attention to the *class of unbiased estimators* and then search for the best estimator within this restricted group. By focusing on unbiased estimators, it becomes somewhat easier to compare their performance and identify the one that performs best.

Definition 5.1 (MVUE). An <u>unbiased</u> estimator $\hat{\theta}$ of $\theta \in \Theta$ is the uniformly ***Minimum Variance Unbiased Estimator*** (MVUE) if for any other <u>unbiased</u> estimator $\hat{\delta}$

$$Var_\theta(\hat{\theta}) \leq Var_\theta(\hat{\delta}), \ \forall \theta \in \Theta.$$

In other words, the MVUE is the best (***most efficient***) *unbiased* estimator that can be found.

Facts to know

- The MVUE may not exist (in some problems, even an unbiased estimator does not exist). However, *when an MVUE does exist, it is unique.*
- In terms of MSE, the MVUE is not necessarily the best estimator. There may be *biased estimators* whose MSE is smaller than that of the MVUE. In fact, a small increase in bias can sometimes lead to a large reduction in variance, resulting in a lower overall MSE. □

The question now is how to find the MVUE (when it exists). To address this, several techniques have been developed in the statistical literature. One important approach relies on a variance inequality known as the *Cramér–Rao bound*. Before introducing this method, we first need to define the concept of *Fisher information*, which plays a central role in statistical inference.