Basic concepts of point estimation

Contents

1 Statistic, estimator and estimation

Definition 1.1. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample. Any (measurable) function $T(\boldsymbol{X})$ of $\boldsymbol{X}$, 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 $\hat{\theta}_n(\boldsymbol{X}) \equiv \hat{\theta}_n \equiv \hat{\theta}$ to denote an ***estimator*** of θ.

Any realization $\hat{\theta}_n(\boldsymbol{x})$ of $\hat{\theta}_n(\boldsymbol{X})$ 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), $S_n^2 = (n-1)^{-1} \sum_i (X_i - \overline{X}_n)^2$, $(\overline{X}_n, S_n^2)$,

$\arg\min_a \sum_i |X_i - a|$. □

To assess the usefulness of an estimation procedure, we examine the properties of the estimator used. Being random, any estimator has a distribution that is referred to as the ***sampling distribution***. The properties of the sampling distribution determine which estimator is best for a particular problem and which estimator to avoid.

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 consider one of the following estimators:

$$\begin{aligned}
&\hat{\theta}_1 = X_{(n)}, \quad \hat{\theta}_2 = \frac{n+1}{n} X_{(n)} \\
&\hat{\theta}_3 = X_{(1)} + X_{(n)}, \quad \hat{\theta}_4 = 2\overline{X}_n \\
&\hat{\theta}_5 = 2\hat{q}_{0.5},
\end{aligned}$$

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

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

This example suggests questions like:

- If many estimators are available, how can we compare them?
- Are there general methods for constructing estimators?
- How to find the best possible estimator for a given model?

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

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

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

3 MSE, bias, variance, and relative efficiency

If the squared error loss is used, then the risk is the ***mean squared error***:

$$\begin{aligned} MSE_\theta(\hat{\theta}) &:= E_\theta[(\hat{\theta} - \theta)^2] \\ &= E_\theta\big[\big(\hat{\theta} - E_\theta(\hat{\theta})\big) + \big(E_\theta(\hat{\theta}) - \theta\big)\big]^2 \\ &= Bias_\theta^2(\hat{\theta}) + Var_\theta(\hat{\theta}), \end{aligned}$$

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

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.

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. $Bias(\hat{\theta}) = 0 \, \forall \theta \in \Theta$, 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*.

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 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. □

Bias and transformation

If $\hat{\theta}$ is an unbiased estimator of θ, then $a + b\hat{\theta}$ is unbiased estimator for $a + b\theta$. In general, however, if $\hat{\theta}$ is an unbiased estimator of θ, then $g(\hat{\theta})$ is not necessarily an unbiased estimator of $g(\theta)$.

For example, we know that $\overline{X}_n$ is an unbiased estimator of μ, but, by Jensen's inequality, $E\left(\overline{X}_n^2\right) > \left(E\left(\overline{X}_n\right)\right)^2 = \mu^2$. Thus, $\overline{X}_n^2$ is a biased estimator of μ^2.

It is sometimes easy to adjust a biased estimator and transform it into an unbiased one. For example, if we known that $E(\hat{\theta}) = a\theta + b$ then $(\hat{\theta} - b)/a$ is unbiased for θ.

Example 3.2. Let X_i, $i = 1, \ldots, n$, 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). □

Example 3.3 (The sample variance)**.** Let X_i, $i = 1, \ldots, n$, 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

$$\begin{aligned} E(\hat{\sigma}_n^2) &= \frac{1}{n}\sum_{i=1}^n \left(E(X_i^2) + E\left(\overline{X}_n^2\right) - \frac{2}{n}\sum_{j=1}^n E(X_i X_j) \right) \\ &= (\sigma^2 + \mu^2) + \left(\frac{\sigma^2}{n} + \mu^2\right) - \frac{2}{n}\left(\sigma^2 + \mu^2 + (n-1)\mu^2\right) \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

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

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

We can compare the quality of two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ by looking at the ratio of their MSE and we call this quantity the ***relative efficiency*** of $\hat{\theta}_1$ to $\hat{\theta}_2$:

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

This ratio is particularly meaningful if both $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased or have about same bias. For unbiased estimators,

RE reduces to

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

If this quantity is less than 1 $\forall \theta \in \Theta$, then $\hat{\theta}_1$ has 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 $\hat{\theta}_2 = \frac{n+1}{n} X_{(n)}$, 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 $f_{X_{(n)}}(x) = n\frac{x^{n-1}}{\theta^n} I(0 \leq x \leq \theta)$. It follows that $E(X_{(n)}) = \frac{n}{n+1}\theta$ and $E\left(X_{(n)}^2\right) = \frac{n}{n+2}\theta^2$. Therefore, $\hat{\theta}_2$ is unbiased and its variance is $Var(\hat{\theta}_2) = \frac{\theta^2}{n(n+2)}$. 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 $\hat{\theta}_1$. □

Note that in the above example, the relative efficiency does not depend on θ (the parameter of interest), but this is not the case in general; more on this to follow in the next section.

In practice, when it is difficult or impossible to obtain an explicit/exact formula for MSE, we use asymptotic expansions or simulations to estimate it. The R code below illustrates the second approach in the context of Example 3.4.

```r
mse <- function(estimator, truth) mean((estimator - truth)^2)
bias <- function(estimator, truth) mean(estimator) - truth
var <- function(estimator) mean((estimator - mean(estimator))^2)

sim <- function(n, theta, estFun, N = 5000) {
  est <- replicate(N, {
    x <- runif(n, min = 0, max = theta)
    estFun(x) } )
  c(mse = mse(est, theta), bias = bias(est, theta), var = var(est))
}

restht1 <- replicate(500, sim(n = 10, theta = 3, estFun = \(x) 2 * mean(x)))
```

```
restht2 <- replicate(500,
                     sim(n <- 10, theta = 3, estFun = \(x) ((n + 1) / n) * max(x)))

restht1 |> rowMeans()
```

```
     mse     bias      var
3.00e-01 7.96e-05 3.00e-01
```

```
restht2 |> rowMeans()
```

```
      mse      bias       var
 0.075204 -0.000231  0.075188
```

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 $3^2/(10 \times (10 + 2)) = 0.075$.

4 The best unbiased estimator (MVUE)

It seems very natural to prefer estimators that have a small MSE. An estimator $\hat{\delta}$ whose MSE is uniformly larger than another estimator (i.e. $\exists \hat{\theta} : \ MSE_\theta(\hat{\theta}) \leq MSE_\theta(\hat{\delta})$, $\forall \theta \in \Theta$) is called ***inadmissible***. Actually, it would be best if we could find an estimator $\hat{\theta}$ that has the smallest MSE among all possible estimators for each possible value of θ. *Unfortunately, this is impossible.* Because in such a case $\hat{\theta}$ should satisfies $E_\theta\big(\hat{\theta} - \theta\big)^2 = 0$, $\forall \theta \in \Theta$ which is simply not possible.

Typically, two estimators $\hat{\theta}$ and $\hat{\delta}$ are not (uniformly) comparable. We can always find $\theta_1, \theta_2 \in \Theta$ such that

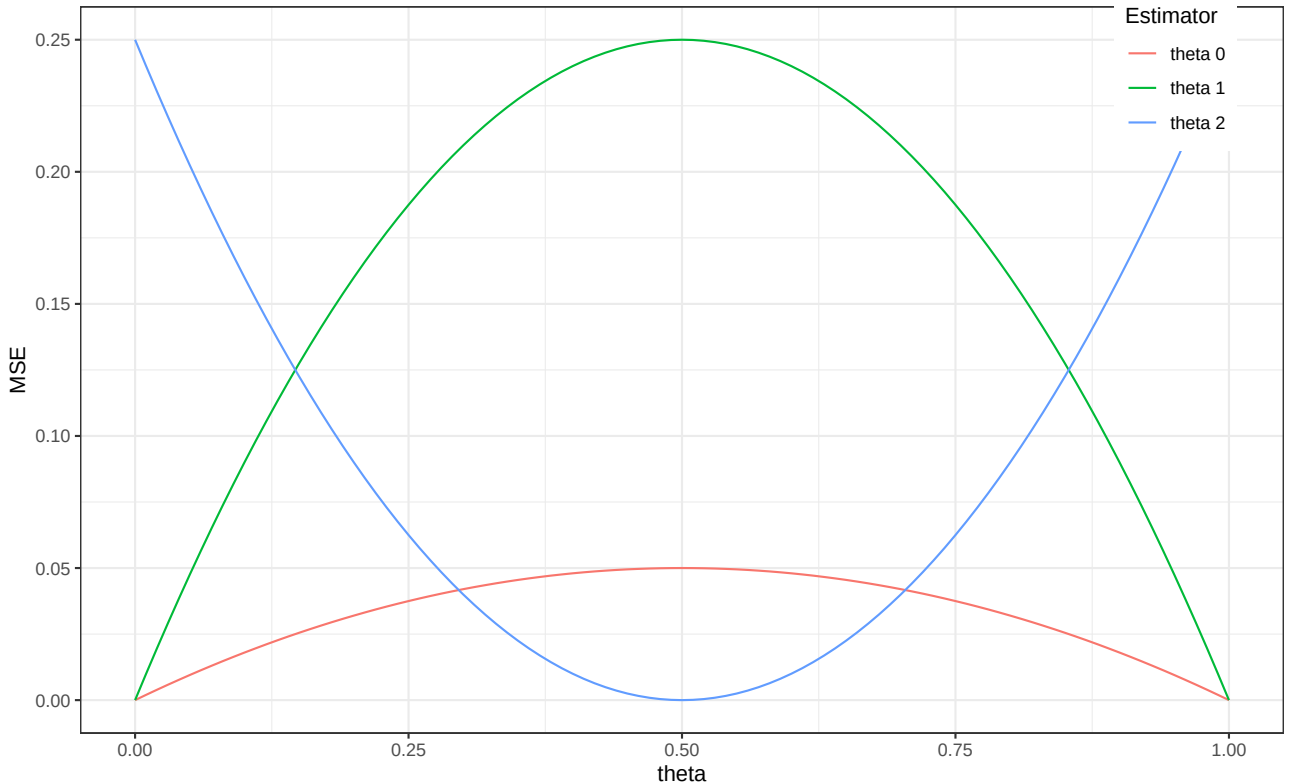
$$MSE_{\theta_1}(\hat{\theta}) < MSE_{\theta_1}(\hat{\delta}) \text{ and } MSE_{\theta_2}(\hat{\theta}) > MSE_{\theta_2}(\hat{\delta}).$$

Example 4.1.

Suppose an iid simple $X_i, \ldots, X_n$ 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 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 uniformly minimum MSE (best) estimator, statisticians adopt other strategies. One of this strategies is to restrict attention to the *class of unbiased estimators*, and then to search for the best estimator in this restricted group.

Definition 4.1 (MVUE)**.** An *unbiased* estimator $\hat{\theta}$ of $\theta \in \Theta$ is the uniformly ***Minimum Variance Unbiased Estimator*** (MVUE) if for any other *unbiased* 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

- MVUE may not exist (even an unbiased estimator may not exist!), but when it does, it's unique.
- In terms of MSE, the MVUE is not necessarily the best estimator as there may be *biased estimators* that achieve lower MSE than the MSE of the MVUE. In fact, sometimes a small increase in bias is associated with a large decrease in variance, overall decreasing the MSE.

The question now is how to find the MVUE (when it exists). To answer this question, different techniques exist in the literature. One of these techniques is based on a variance inequality known as the ***Cramér-Rao bound***. Before

presenting this method, we need to introduce the concept of ***Fisher information*** which plays a crucial role in statistical inference.