Anna Janicka

Mathematical Statistics 2021/2022
Lecture 4

1. Estimator properties – introduction

In lecture 3, we studied the different methods of point estimation – i.e. methods of providing data-based assessments (values) of unknown values of parameters of data distributions. Immediately, some basic questions arise: are any of these methods of estimating better than others? In what terms can we describe the properties of the introduced estimators? Do we really estimate what we want, when we are using a particular method? Aren't the errors we are making too large? During this lecture and the following lectures 5 and 6, we will define basic estimator characteristics and properties, and provide the tools to determine which methods prove best under given sets of assumptions.

Obviously, if we want to assess the quality of an approximation, the intuitive mechanism is to look at the error we are making, and base our judgement on this aspect. In the case of estimators, there are two basic problems with such an approach. First, due to the fact that the data an estimator is based on are assumed to be random variables, the errors will also be random, and thus impossible to predict with certainty. Therefore, instead of looking at precise values of errors, we will need to look at *expected* values of errors (averaged over all the possible outcomes) for an estimator. Second, the error we will make will depend on the true value of the parameter we wish to estimate, which we do not know. This is a drawback that we will need to get used to – the properties will need to be described in terms of functions (of the unknown parameter), rather than values.

2. Bias

The most basic property of an estimator is it's bias, i.e. the difference between the (expected) value of the estimator and the value that was to be estimated.

Definition 1. *The* **bias** *of an estimator $\hat{\theta}(X)$ of the value θ is*

$$b(\theta) = \mathbb{E}_\theta(\hat{\theta}(X) - \theta) = \mathbb{E}_\theta\hat{\theta}(X) - \theta,$$

and of an estimator $\hat{g}(X)$ of the value $g(\theta)$ is

$$b(\theta) = \mathbb{E}_\theta(\hat{g}(X) - g(\theta)) = \mathbb{E}_\theta\hat{g}(X) - g(\theta).$$

By adding the subscript $_\theta$ to the expected value we want to underline the fact that the expected value is calculated for a given value of the parameter, and what is averaged are the possible experiment outcomes for this given value. Obviously, the most desirable case is the one where the bias is zero:

Definition 2. *An estimator is* **unbiased**, *if $\forall\theta \in \Theta$ we have $b(\theta) = 0$.*

Example: Let us assume that $X_1, X_2, \ldots, X_n$ are random variables from a normal distribution $\mathcal{N}(\mu, \sigma^2)$. Let us first estimate the value of μ. Let us consider the following estimators:

(1) $\mu_1 = \bar{X}$ is an unbiased estimator of μ, since

$$\mathbb{E}_{\mu,\sigma}(\bar{X}) = \mathbb{E}_{\mu,\sigma}\frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{n \cdot \mu}{n} = \mu;$$

(2) $\mu_2 = X_1$ is also an unbiased estimator of μ;
(3) $\mu_3 = 5$ is a biased estimator of μ: $b(\mu) = \mathbb{E}_{\mu,\sigma}(5 - \mu) = 5 - \mu \neq 0$ for all values of $\mu \neq 5$.

The same conclusions hold for any other distribution (i.e. non-normal) with a mean μ, for the equivalent estimators of the mean.

Let us now revert to the normal distribution and estimate the variance. The empirical-distribution-based estimator of the variance, i.e.

$$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is a biased estimator of σ^2:

$$\mathbb{E}_{\mu,\sigma}\hat{S}^2(X) = \mathbb{E}_{\mu,\sigma}\frac{1}{n}\sum(X_i-\bar{X})^2 = \frac{1}{n}\mathbb{E}_{\mu,\sigma}\left(\sum X_i^2 - n\bar{X}^2\right) = \frac{1}{n}\left(n(\mu^2+\sigma^2) - n(\mu^2+\frac{\sigma^2}{n})\right) = \sigma^2 - \frac{\sigma^2}{n} \neq \sigma^2,$$

where we have used the property that for an IID sample from a $\mathcal{N}(\mu,\sigma^2)$ distribution, the average has a known distribution: $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$. The bias of the $\hat{S}^2$ estimator is equal to $b(\sigma) = -\frac{\sigma^2}{n}$, which means that the estimator gives assessments of the true value of the variance which are systematically too small (the bias is negative).

However, this biased formula for the variance may easily be transformed to provide an unbiased estimator S^2, if we divide the sum of squares of the differences from the average by $n-1$ rather than by n:

$$\mathbb{E}_{\mu,\sigma}S^2(X) = \mathbb{E}_{\mu,\sigma}\frac{1}{n-1}\sum(X_i - \bar{X})^2 = \ldots = \sigma^2.$$

The same calculations hold for any other distribution with a variance σ^2: the $\hat{S}^2$ estimator (with division by n) is biased, while the S^2 estimator (with division by $n-1$) is unbiased. This is why, especially if we have a small sample and are interested in a precise assessment of σ^2, we may want to use the estimator with division by $n-1$. Let us note, however, that the bias of the $\hat{S}^2$ estimator will be very small if the sample size is large; we have $b(\sigma) = -\frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0$. We will come back to this observation later on.

3. Mean Square Error and Variance

Although unbiasedness is a welcome property, as we have seen above, there may exist more than one unbiased estimator. How can we therefore choose the best one from among them? In order to be able to provide a criterion, we will go back to one of the questions asked at the beginning of this lecture, namely to the question about the error of the estimator. Bias may be seen as a type of error, but it is not the only possible component: we may imagine an estimator which is not biased, but which is always far off from the value it is supposed to estimate (i.e., the "too large" and "too small" estimates are equally common, but they are indeed too small and too large – too often). This high dispersion is not a welcome property (it is not comforting to know that on average our estimator is right, if we also know that in each case we are going to be very far off from the real value). The variance is a characteristic that describes the variability in a set of outcomes. We will introduce a similar concept:

Definition 3. *Let $\hat{\theta}(X)$ be an estimator of θ. The* **Mean Square Error (MSE)** *of the estimator $\hat{\theta}(X)$ is the function*

$$MSE(\theta,\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta}(X) - \theta)^2.$$

If $\hat{g}(X)$ is an estimator of $g(\theta)$, then the MSE of the estimator $\hat{g}(X)$ is the function

$$MSE(\theta,\hat{g}) = \mathbb{E}_\theta(\hat{g}(X) - g(\theta))^2.$$

The MSE will measure how far, on average, are the values of the estimator from the true value of the parameter we are aiming for. Note that the choice of the type of averaged function is arbitrary; instead of the square of the difference of the value of the estimator from the theoretical counterpart we could have used the absolute value or a different power. Throughout this lecture, however, we will not consider the other choices of these other so-called score functions.

The formula in the definition of the MSE may be rearranged; simple calculations lead to the conclusion that we can write

$$MSE(\theta, \hat{\theta}) = b^2(\theta) + \mathrm{Var}(\hat{\theta}),$$

which means that the quality of an estimator can be decomposed into two factors: the variability of the estimator, and a function of the bias. For unbiased estimators, the MSE is equal to the estimator variance.

We will want to use the MSE as a criterion of comparison of estimators; we will prefer the estimator with smaller errors, i.e. with lower MSE. We need to be careful, however, because the MSEs are functions of unknown parameters, which means that they may intersect. This means that for some values of parameters one of the estimators may be better, while for some different values – the other one. In this case, the estimators are incomparable. In order to be able to say that one estimator is indeed (strictly) better than another one, we will need to make sure that the MSE of the first one is always at least as low as for the second one (and strictly lower for at least one value of the parameter).

Examples: If $X_1, X_2, \ldots, X_n$ are an IID sample from a distribution with mean μ and variance σ^2, and these two parameters are unknown, then

(1) The MSE of $\mu_1 = \bar{X}$ – an unbiased estimator of the mean – is equal to

$$MSE(\mu, \sigma, \bar{X}) = \mathbb{E}_{\mu,\sigma}(\bar{X} - \mu)^2 = \mathrm{Var}_{\mu,\sigma}(\bar{X}) = \frac{\sigma^2}{n};$$

(2) The MSE of $\mu_2 = X_1$ – an unbiased estimator of the mean – is equal to

$$MSE(\mu, \sigma, X_1) = \mathbb{E}_{\mu,\sigma}(X_1 - \mu)^2 = \mathrm{Var}_{\mu,\sigma}(X_1) = \sigma^2;$$

(3) The MSE of $\mu_3 = 5$ – a biased estimator of the mean – is equal to

$$MSE(\mu, \sigma, 5) = \mathbb{E}_{\mu,\sigma}(5 - \mu)^2 = (5 - \mu)^2 = (b(\mu))^2.$$

Note that the variance of the μ_3 estimator, which always returns 5, is equal to 0 (no variability).

(4) In the normal model, the MSE of S^2 – the unbiased estimator of the variance – is equal to

$$MSE(\mu, \sigma, S^2) = \mathbb{E}_{\mu,\sigma}(S^2 - \sigma^2)^2 = \mathrm{Var}_{\mu,\sigma}(S^2) = \frac{2\sigma^4}{n-1};$$

(5) In the normal model, the MSE of $\hat{S}^2$ – the biased estimator of the variance – is equal to

$$MSE(\mu, \sigma, \hat{S}^2) = \mathbb{E}_{\mu,\sigma}(\hat{S}^2 - \sigma^2)^2 = b^2(\sigma) + \mathrm{Var}_{\mu,\sigma}(\hat{S}^2) = \frac{\sigma^4}{n^2} + \frac{(n-1)^2}{n^2}\frac{2\sigma^4}{n-1} = \frac{2n-1}{n^2}\sigma^4.$$

Therefore, we can see that in terms of the MSE, the estimator μ_1 is better than the estimator μ_2 (it has a lower variance if only the available sample size is greater than 1). Unfortunately, the intuitively worse estimator μ_3 is incomparable to μ_1 and μ_2, because for some values of the parameters – if it happens that the true value of μ is equal to 5 – the MSE of μ_3 will be equal to 0, and lower than in the case of the other two estimators which have non-zero variances. This shows that the MSE criterion is not a perfect one – based on this criterion alone, we can't reject an estimator that is obviously "stupid"!

On the other hand, if we compare the two estimators of the variance, S^2 and $\hat{S}^2$, we can see that the biased estimator always has a lower MSE (i.e. it dominates the unbiased estimator). What happens here is that when we introduce a change of scale which corrects the bias of the $\hat{S}^2$ estimator, in effect we inflate the variance of the estimator. What needs to be stressed at this point is that although the precise values of the MSEs of the two estimators depend on the distribution (for non-normal distributions, the formulae will be different), the conclusion that the biased estimator has a lower MSE stands regardless of the distribution (provided it has a variance). What also needs to be underlined at this point is that the choice whether we will use the $\hat{S}^2$ or the S^2 estimator in a particular situation depends on whether the criterion

which is more important to us is the estimator behaving better overall (having smaller MSE) or the estimator giving us results which are on average correct (being unbiased). This may change in different situations.

The examples above show that it may be pointless to try to compare all estimators, since if we look at the natural criterion – the estimator error – we will always be able to provide a "stupid" constant estimator like μ_3 which will be better than all other estimators for one particular value of the estimated parameter (while being much worse for all other values). Therefore, it is worthwhile to constrain our quest for the best estimator to a search within unbiased estimators only. In these cases, the estimator with lower variance will have lower MSE.

Immediately, the question of the *best possible* estimator arises: is there a limit to how small the variance may be (how good an estinmator may be)? In the next lecture, we will show that indeed there is; for now, we will limit ourselves to introducing the following definition:

Definition 4. *$g^*(X)$ is a* **Minimum Variance Unbiased Estimator (MVUE)***, for $g(\theta)$, if:*

(1) *$g^*(X)$ is an unbiased estimator of $g(\theta)$;*
(2) *for any unbiased estimator $\hat{g}(X)$ we have $\forall\theta \in \Theta$:*

$$\mathrm{Var}_\theta g^* \leq \mathrm{Var}_\theta \hat{g}.$$

How can we verify whether an estimator is MVUE? In general, it is not possible to freely minimize the variance of unbiased estimators – for many statistical models there exists a limit of variance minimization. This limit depends on the underlying distribution and on the sample size. We will introduce the necessary distribution properties in the next lecture. At this point, we will just note that the condition of unbiasedness is a crucial one – as the μ_3 example shows, it is not a problem to construct a biased estimator with zero variance – any constant (i.e. an estimation rule "regardless of the data, we always say the same thing") has such a property.