

Estimation of Unknown Parameters

In statistics, we take data and draw conclusions (called “statistical inferences”). Typically, we assume that the data are “observed values” of random variables whose distributions are known in advance *except* for the unknown values of certain parameters.

Example 1: Suppose we run an experiment involving n “trials”, each resulting in “success” or “failure”. Let X = observed number of successes. Assuming that $P(\text{success}) = p$, the same on every trial, we know that

$$X \sim \text{Binomial}(n, p).$$

There are 2 parameters of this distribution, n and p , but usually we know n . So the “unknown parameter” is p .

Q: If we observe $X = x$, what can we infer about p ?

A: One common kind of inference is to *estimate* p . To do this, we need to define an *estimator* $T = T(x)$.

Since p is the probability of success, it is natural to estimate p by the observed frequency of success – i.e.

$$T(x) = \frac{x}{n}.$$

Example 2: Suppose we run an experiment whose outcome is a set of observed values of random variables iid + normally distributed:

$$X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2),$$

where μ and σ^2 are both unknown parameters. If we want to estimate μ , a natural choice is

$$T_1 = T_1(X_1, \dots, X_n) = \bar{X}_n = \frac{X_1 + \dots + X_n}{n},$$

the “sample mean”. If we want to estimate σ^2 , a natural choice is

$$T_2 = T_2(X_1, \dots, X_n) = \frac{\sum (X_i - \bar{X}_n)^2}{n},$$

the “sample variance”. (Note! Sometimes the term “sample variance” refers to

$$T_3 = \frac{\sum (X_i - \bar{X}_n)^2}{n-1}.$$

That ambiguity will be discussed later.)

Def: The k^{th} *sample moment* of a sample x_1, \dots, x_n (not necessarily distinct numbers), is the k^{th} moment of the *sample distribution* – i.e. the distribution giving probability $\frac{1}{n}$ to each x_i .

In general, if we observe random variables X_1, X_2, \dots, X_n whose joint density function (discrete or continuous) depends on unknown parameters $\Theta_1, \dots, \Theta_k$, then a good estimator $T_1 = T_1(x_1, \dots, x_n)$ of Θ_1 (say), based on observed values $X_1 = x_1, \dots, X_n = x_n$, is a T_1 such that the

$$\text{estimation error} = T_1 - \Theta_1$$

is “small” – that is, T_1 is close to Θ_1 , at least most of the time.

Note: One definition of “closeness” (to be discussed later) is the so-called “mean square error”

$$\text{MSE}(\Theta_1) = E_{\Theta_1} (T_1 - \Theta_1)^2,$$

where “ E_{Θ_1} ” means that the expectation is computed assuming that Θ_1 is the true value of the (first) unknown parameter. In the binomial example, $\Theta_1 = p$ is the unknown parameter, and if $T_1 = \frac{X}{n}$ is the estimator, then the mean square error (MSE) is

$$\begin{aligned} E_p \left(\frac{X}{n} - p \right)^2 &= E_p \left(\frac{X}{n} - E_p \frac{X}{n} \right)^2 = \text{Var}_p \left(\frac{X}{n} \right) \\ &= \frac{1}{n^2} \text{Var}_p(X) = \frac{npq}{n^2} = \frac{pq}{n}. \end{aligned}$$

So the MSE is different for different p in this example.

Q: Are there good general methods for choosing estimators?

A: Yes, quite a few.

Q: Is there a *best* method?

A: No, for reasons to be discussed later.

One general method, the so-called “Method of Moments” (Sec 8.4 of Rice) is based essentially on the idea: “estimate a parameter that is the true mean (expectation) of iid rv’s X_1, \dots, X_n by using the sample mean \bar{X}_n ”. Similarly, the true variance is estimated by the sample variance, and parameters that can be written as a function of the true mean or variance or k^{th} moment should be estimated by the *same* function of the corresponding *sample moments*.

Comment: We’re skipping the Method of Moments. For the last 75 years or so, it has been used mainly as a “quick and dirty” method to derive estimates. A frequently better method is called “Maximum Likelihood Estimation”.

Q: What is “likelihood”?

A: If X_1, \dots, X_n have joint density function

$$f_{\Theta}(x_1, \dots, x_n),$$

where Θ is a single unknown parameter or $\Theta = (\Theta_1, \dots, \Theta_n)$, a vector of unknown parameters, then for observed values $X_1 = x_1, \dots, X_n = x_n$ the *likelihood function* is

$$L_{x_1, \dots, x_n}(\Theta) = f_{\Theta}(x_1, \dots, x_n).$$

Note: The likelihood function is

– a function of the unknown parameter or parameters

and

– it's a *random function*, in the sense that it depends on the values of the random variables X_1, \dots, X_n that happen to be observed.

Def: A *maximum likelihood estimator* (mle) of Θ is a value $\hat{\Theta} = \hat{\Theta}(x_1, \dots, x_n)$ such that

$$L_{x_1, \dots, x_n}(\hat{\Theta}) = \max_{\Theta} L_{x_1, \dots, x_n}(\Theta).$$

Example 3: In Example 1 (binomial) after observing $X = x$ we can compute the likelihood function – a function of the unknown parameter p – as

$$L_x(p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

To find \hat{p} that maximizes this, take logarithms,

$$\log L_x(p) = \log \binom{n}{x} + x \log p + (n-x) \log(1-p)$$

and find \hat{p} maximizing this by taking

$$\frac{\partial}{\partial p} \log L_x(p) = \frac{x}{p} - \frac{n-x}{1-p}. \quad (1)$$

If $x = 0$, this is negative, so $\log L_x(p)$ is decreasing and $\hat{p} = 0$. If $x = n$, $\hat{p} = 1$ by similar reasoning and $\hat{p} = 0$. If $x \in \{1, \dots, n-1\}$, then (1) is positive for p near 0 and negative for p near 1. Hence $\log L_x(p)$ as p goes from 0 to 1 is initially increasing and finally decreasing, changing sign of the derivative at \hat{p} where (1) is zero – i.e.

$$\frac{x}{\hat{p}} = \frac{n-x}{1-\hat{p}},$$

which leads to $\hat{p} = \frac{x}{n}$. This formula gives the right answer for $x = 0$ and $x = n$, too.

Example 4: In Example 2 (iid Normal) the likelihood function is, after taking logarithms

$$\log L_{x_1, \dots, x_n}(\mu, \sigma) = -\frac{n}{2} \log 2\pi - n \log \sigma - \frac{1}{2} \sum \frac{(x_i - \mu)^2}{\sigma^2}. \quad (2)$$

1. If σ is known and μ is unknown, then regardless of what the value of σ is, it is clear that (2) is maximized by *minimizing* $\sum (x_i - \mu)^2 = \sum x_i^2 - 2\mu \sum x_i + n\mu^2$. This yields

$$\hat{\mu} = \bar{X}_n.$$

2. If both σ and μ are unknown, then the values $\hat{\mu}, \hat{\sigma}$ that maximize (2) are obtainable as follows: first maximize over μ for fixed $\sigma > 0$ and note that the answer (in 1) is $\hat{\mu} = \bar{X}_n$, which doesn't depend on σ . So the maximum over μ and σ will certainly have that value of $\hat{\mu}$, and the maximizing value, $\hat{\sigma}$, can be found by differentiating (2) with respect to σ (with $\mu = \hat{\mu}$) to yield

$$-\frac{n}{\sigma} + \frac{\sum (x_i - \bar{x})^2}{\sigma^3},$$

which changes sign from positive to negative at

$$\widehat{\sigma^2} = \frac{\sum (x_i - \bar{x})^2}{n} \quad \text{or} \quad \hat{\sigma} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}.$$

A standard way of evaluating the performance of estimators is the following.

Def: If $T = T(X_1, \dots, X_n)$ is being used to estimate $g(\Theta)$, where g is a given function of the unknown parameter(s), then the *mean squared error* of T is

$$\text{MSE}_T(\Theta) = E_{\Theta}(T - g(\Theta))^2.$$

The idea here is that we square the "error", which is $T - g(\Theta)$, and then take the expectation, assuming Θ is true. Sometimes $\text{MSE}_T(\Theta)$ is constant in Θ , but often it is not.

Example 1: If X_1, \dots, X_n are iid $N(\mu, \sigma^2)$ then the mean squared error of the sample mean, \bar{X}_n , as an estimator of μ is

$$\text{MSE}_{\bar{X}_n}(\mu) = E_{\mu, \sigma^2} (\bar{X}_n - \mu)^2.$$

One can evaluate this by expanding the square first, but a more insightful way to do it is to note that since $E_{\mu, \sigma^2} \bar{X}_n = \mu$, the MSE equals

$$\text{Var}_{\mu, \sigma^2} (\bar{X}_n) = \frac{\sigma^2}{n}.$$

This depends on σ , obviously, but is constant as a function of μ .

Example 2: Suppose X_1, \dots, X_n are iid exponentially distributed random variables with

$$f_{\lambda}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then $g(\lambda) = \frac{1}{\lambda}$ is the mean, $E_\lambda X$, and a natural estimator (also the mle) of $\frac{1}{\lambda}$ is \bar{X}_n . Its mean squared error is

$$\begin{aligned} E_\lambda \left(\bar{X}_n - \frac{1}{\lambda} \right)^2 &= E_\lambda (\bar{X}_n - E_\lambda \bar{X}_n)^2 = \text{Var}_\lambda (\bar{X}_n) \\ &= \frac{1}{n} \text{Var}_\lambda (X_1) = \frac{1}{n} \cdot \frac{1}{\lambda^2} = \boxed{\frac{1}{n\lambda^2}}. \end{aligned}$$

Here the MSE is not a constant function.

There is another property that estimators have that can be important.

Def: If $T = T(X_1, \dots, X_n)$ is being used to estimate $g(\Theta)$, then its *bias* is defined as

$$b_T(\Theta) = E_\Theta (T - g(\Theta)).$$

Note that this could be called its "mean error" (but it isn't).

An estimator T is called an *unbiased estimator* of $g(\Theta)$ if $b_T(\Theta) \equiv 0$, i.e. if

$$E_\Theta T = g(\Theta) \text{ for all } \Theta.$$

Note that \bar{X}_n , for example, is always an unbiased estimator of the true mean of the population from which the X 's are sampled. A simple and interesting way to express the MSE (though not necessarily the easiest way to calculate it) is this:

$$\begin{aligned} \text{MSE}_T(\Theta) &= \text{Var}_T(\Theta) + E_\Theta (T - g(\Theta))^2 \\ &= \text{Var}_T(\Theta) + (b_T(\Theta))^2. \end{aligned} \tag{3}$$

In other words, the mean square error of an estimator equals its variance plus the square of its bias.

It is tempting to infer that unbiased estimators, since they minimize the last term in (3), tend to have smaller (hence better) MSE's than biased estimators. This is not true!

Example: Suppose X_1, \dots, X_n are iid \sim uniform on $[0, \Theta]$, where $\Theta > 0$ is unknown. The mle (as shown in lecture) is

$$\hat{\Theta} = \max(X_1, \dots, X_n).$$

This is *clearly* not an unbiased estimator of Θ since the X 's are less than Θ (being uniformly distributed on $[0, \Theta]$) and therefore the error $\hat{\Theta} - \Theta$ is negative 100% of the time! But this example is typical of many problems in which a slight modification of an estimator can be made to make it unbiased. The density of $\max(X_1, \dots, X_n)$ is, for $0 < t < \Theta$

$$\begin{aligned} \frac{d}{dt} P_\Theta (\max(X_1, \dots, X_n) \leq t) &= \frac{d}{dt} [P_\Theta (X_1 \leq t)]^n \\ &= \frac{d}{dt} \left(\frac{t}{\Theta} \right)^n = \frac{n}{\Theta} \left(\frac{t}{\Theta} \right)^{n-1} \end{aligned}$$

(Note that Θ is a scale parameter.) We get

$$E_{\Theta} \max(X_1, \dots, X_n) = \int_0^{\Theta} \frac{n}{\Theta^n} t^n dt = \frac{n}{n+1} \Theta.$$

So an unbiased estimator is

$$T = \frac{n+1}{n} \max(X_1, \dots, X_n) = \frac{n+1}{n} \hat{\Theta}.$$

To compare its mean squared error with that of $\hat{\Theta}$ - and with all estimators of the form $T = C_n \max(X_1, \dots, X_n)$ we can compute

$$\begin{aligned} \text{MSE}_{T_n} &= E_{\Theta} (c_n \max(X_1, \dots, X_n) - \Theta)^2 \\ &= c_n^2 E_{\Theta} (\max)^2 - 2c_n E_{\Theta} \max + \Theta^2 \end{aligned}$$

We know $E_{\Theta} \max = \frac{n}{n+1} \Theta$ and calculate

$$E_{\Theta} (\max)^2 = \int_0^{\Theta} \frac{n}{\Theta^n} t^{n+1} dt = \frac{n}{n+2} \Theta^2.$$

So

$$\begin{aligned} \text{MSE}_{T_n}(\Theta) &= c_n^2 \cdot \frac{n}{n+2} \Theta^2 - 2c_n \frac{n}{n+1} \Theta^2 + \Theta^2 \\ &= \left(c_n^2 \frac{n}{n+2} - 2c_n \frac{n}{n+1} + 1 \right) \Theta^2, \end{aligned}$$

and we notice that all of these estimators have mean squared error of the form "constant times Θ^2 " (this happens because Θ is a scale parameter). It is clear, then, that we can compare any two of these estimators and will prefer the one that makes

$$c_n^2 \frac{n}{n+2} - 2c_n \frac{n}{n+1} + 1$$

smaller. Here are some choices

c_n	Value of (4)
$\frac{1}{n+1}$ (mle)	$\frac{2}{(n+1)(n+2)}$
$\frac{n+1}{n}$	$\frac{1 + \frac{1}{n+2}}{(n+1)^2}$
$\frac{n+2}{n+1}$	$\frac{1}{(n+1)^2}$

So we see that the unbiased estimator can be improved upon (in the sense of slightly better mean squared error) by an estimator that has a slight bias. One of the homework problems concerns a similar comparison, where the improvement is larger.

In general?

Unbiasedness is considered too restrictive a condition to impose universally on the selection of estimators, but generally we want the "bias term" to be small compared to the "variance term" in the mean squared error.

[Typeset by Keegan McAllister (keegan@caltech.edu). Last updated 2006-02-23.]