Statistics 3858 : Construction of Two Common Types of Estimators

Estimators are statistics used to estimate parameters or functions of parameters for a statistical model.

Here we consider finite parameter models. Let Θ be the parameter space.

Consider the typical setting of the random data $X_1, X_2, \ldots, X_n$ being iid from a distribution that belongs to a statistical model. That is X_i are iid with distribution $f(\cdot, \theta)$ where $\theta \in \Theta$, but unknown.

The first method we discuss is the method of moments. The other commonly used method is the method of maximum likelihood, and is discussed in a later handout.

1 Method of Moments

Let $\mu_k = \mathrm{E}(X^k)$. Notice this expectation depends on the parameter θ, and so we write $\mu_k(\theta) = \mathrm{E}_\theta(X^k)$. The subscript θ is denote the dependence on the parameter θ.

Aside To help clarify this notation and idea recall, in the continuous r.v. case that we define, when finite,

$$\mathrm{E}(X^k) = \int_{-\infty}^{\infty} x^k f(x) dx \ . \tag{1}$$

In the case of a statistical model with parameter space Θ possible choices for the pdf f are $f(\cdot;\theta)$. Thus in (1) we have a possibly different value for this integral for each possible choice of θ. Thus we have

$$\mathrm{E}_\theta(X^k) = \int_{-\infty}^{\infty} x^k f(x;\theta) dx \ .$$

which gives a mapping $\theta \mapsto \mathrm{E}_\theta(X) \equiv \mu_k(\theta)$.

In the case of a discrete r.v. we have a similar property.

End of Aside

Notice this gives a mapping $\theta \mapsto \mu_k(\theta)$, typically a many to one mapping. However we can often find a mapping

$$\theta \mapsto (\mu_1(\theta), \mu_2(\theta), \ldots, \mu_K(\theta))$$

so that this mapping (or function) is a one to one function. Specifically this gives a function

$$h : \Theta \mapsto R^K \ .$$

When we restrict the range appropriately this is a 1 to 1 function. Therefore, when using the domain Θ and suitable range, the function h has an inverse h^{-1}. Usually $K = d$ where $d = \text{dimension}(\Theta)$, but sometimes we need to modify this to obtain a 1 to 1 function. Specifically we have

$$\theta = h^{-1}\left(\mu_1(\theta), \mu_2(\theta), \ldots, \mu_K(\theta)\right) \ .$$

Examples : (fill in details)

Exponential , Bernoulli , Poisson , Normal, Gamma, $N(0, \sigma^2)$

Bernoulli and Binomial

If $X \sim \text{Bernoulli}(\theta)$ then $E_\theta(X) = \theta$. Thus for a given value of $\mu_1 = \text{E}(X)$ we can determine $\theta = \mu_1$.

If $X \sim \text{Binom}(m, \theta)$ then $E_\theta(X) = m\theta$. Thus for a given value of $\mu_1 = \text{E}(X)$ we can determine $\theta = \frac{\mu_1}{m}$.

Notice that if $X_1, \ldots, X_m$ are iid Bernoulli(θ) then $Y = X_1 + X_2 + \ldots + X_m \sim \text{Binom}(m, \theta)$.

End of Example

Exponential

If $X \sim \text{exponential}, \theta$ then X has pdf

$$f(x;\theta) = \theta e^{-\theta x} \text{I}(x \geq 0)$$

The parameter θ belongs to the parameter space $\Theta = (0, \infty)$. $\text{E}_\theta(X) = \frac{1}{\theta}$ and $E_\theta(X^2) = \frac{2}{\theta^2}$. Thus we have a mapping

$$\theta \mapsto \text{E}_\theta(X) = \frac{1}{\theta} \ .$$

This produces a mapping $h : \Theta \mapsto R^+$ given by $h(\theta) = \frac{1}{\theta}$. This has an inverse mapping $h^{-1} : R^+ \mapsto \Theta$ given by $h^{-1}(x) = \frac{1}{x}$. Thus for a given value $\mu = \text{E}(X)$ we can determine

$$\theta = h^{-1}(\mu) = \frac{1}{\mu} \ .$$

We do not need to use the second moment, since stopping at the first moment produces a 1 to 1 mapping between Θ and the set of possible first moments.

End of Example

Gamma Example

If $X \sim \text{Gamma}(\alpha, \lambda)$ then X has pdf

$$f(x;\alpha,\lambda) = \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} \text{I}(x > 0)$$

and

$$\mu_1(\alpha,\lambda) = \text{E}_\theta(X) = \frac{\alpha}{\lambda} \ , \ \mu_2(\alpha,\lambda) = \text{E}_\theta(X^2) = \frac{\alpha}{\lambda^2} + \frac{\alpha^2}{\lambda^2} = \frac{\alpha(\alpha+1)}{\lambda^2}$$

In this example it is not possible to find a 1 to 1 mapping between Θ and the possible values of the first moments. This is because Θ is of dimension 2 and the set of possible first moments is R^+, a set of dimension 1. Thus we next have to consider first and second moments.

The above formulae gives us a mapping $h : \Theta \mapsto B$, where

$$B = \{(\mu_1, \mu_2) : \mu_1 > 0, \mu_2 > 0, \mu_2 - \mu_1^2 > 0\}$$

This formula can be inverted, that is an inverse map can be found.

Given $(\mu_1, \mu_2) \in B$, that is such that $\mu_2 - \mu_1^2 > 0$, we can then solve

$$\begin{aligned} \frac{\alpha}{\lambda} &= \mu_1 \\ \frac{\alpha(\alpha+1)}{\lambda^2} &= \mu_2 \end{aligned}$$

The solution is

$$\begin{aligned} \lambda &= \frac{\mu_1}{\mu_2 - \mu_1^2} \\ \alpha &= \frac{\mu_1^2}{\mu_2 - \mu_1^2} \end{aligned}$$

Thus knowing the parameter $\theta = (\alpha, \lambda)$ is equivalent to knowing the value $(\mu_1, \mu_2) \in B$.

End of Example

Normal mean 0

This is an example where the construction of the method of moments estimators does not start at μ_1.

$X \sim N(0, \theta)$, with $\theta \in \Theta = (0, \infty)$. Then

$$\mathrm{E}_\theta(X) = 0 \ , \ \mathrm{E}_\theta(X^2) = \theta \ .$$

Notice if we have a possible value of μ_1 in this case, the only possible value is $\mu_1 = 0$ and then we cannot find an inverse map to determine what is the corresponding value of θ. However for a given possible value of μ_2 then we can determine (that is find the corresponding value) θ from

$$\mu_2 = \theta \ .$$

End of Example

Let

$$\hat{\mu}_{k,n} = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

be the k-th sample moment. Let

$$\tilde{\theta}_n = h^{-1}\left(\hat{\mu}_{1,n}(\theta), \hat{\mu}_{2,n}(\theta), \ldots, \hat{\mu}_{K,n}(\theta)\right)$$

We call $\tilde{\theta}_n$ the method of moments estimator. It is also sometimes called the moment matching method, but the usual statistical terminology is the method of moments.

Examples : (fill in details)

Exponential

The method of moments estimator for θ is obtained by solving

$$\hat{\mu}_{1,n} = \frac{1}{\theta}$$

Recall also that $\hat{\mu}_{1,n} = \bar{X}_n$ so therefore we obtain the method of moments estimator

$$\tilde{\theta}_n = \frac{1}{\bar{X}_n}$$

In this particular example we can calculate the expected value of $\tilde{\theta}_n$.

If $X_1, \ldots, X_n$ are iid exponential, θ, then $T_n = X_1 + \ldots + X_n$ has Gamma(n, θ) distribution. The student should verify this at home. The simplest method is to use moment generating functions.

Therefore

$$\begin{aligned}
\mathrm{E}_\theta\left(\frac{1}{\bar{X}_n}\right) &= n\mathrm{E}_\theta\left(\frac{1}{T_n}\right) \\
&= n\int_0^\infty \frac{1}{x}\frac{\theta^n x^{n-1}}{\Gamma(n)} e_{-\theta x} dx \\
&= \frac{n\Gamma(n-1)\theta}{\Gamma(n)} \int_0^\infty \frac{\theta^{n-1} x^{(n-1)-1}}{\Gamma(n-1)} e^{-\theta x} dx \\
&= \frac{n\Gamma(n-1)\theta}{\Gamma(n)} \\
&= \frac{n(n-2)!\theta}{(n-1)!} \\
&= \frac{n}{n-1}\theta \ .
\end{aligned}$$

We can therefore determine that the method of moments estimator is not unbiased. However we can construct another estimator from this that does happen to be unbiased. What is it?

We can also determine the distribution of $\frac{1}{\bar{X}_n} = \frac{n}{T_n}$. Let f_T be the pdf of T_n. It is the Gamma(n, θ) pdf. Set $W = \frac{n}{T_n}$. The pdf of W is the of course the pdf of $\frac{1}{\bar{X}_n}$, and we are just using more convenient notation of the type used earlier in Chapter 2. The from the methods in Chapter 2 we obtain the pdf f_W of W

$$f_W(w) = f_T(\frac{n}{w}) \mid -\frac{n}{w^2} \mid$$

For $w \le 0$, $f_W(w) = 0$. For $w > 0$

$$\begin{aligned}
f_W(w) &= f_T(\frac{n}{w})\frac{n}{w^2} \\
&= \frac{\theta^n \left(\frac{n}{w}\right)^{n-1} n}{\Gamma(n) w^2} e^{-\theta(\frac{n}{w})} \\
&= \frac{n^n \theta^n}{\Gamma(n) w^{n+1}} e^{-(\frac{n\theta}{w})}
\end{aligned}$$

End of Example

Gamma Example

X_i, $i = 1, \ldots, n$ are iid random variables from the Gamma model. The method of moments estimator is then

$$\begin{aligned} \tilde{\lambda}_n &= \frac{\hat{\mu}_{1,n}}{\hat{\mu}_{2,n} - \hat{\mu}_{1,n}^2} \\ \tilde{\alpha}_n &= = \tilde{\lambda}_n \hat{\mu}_{1,n} = \frac{\hat{\mu}_{1,n}\hat{\mu}_{1,n}}{\hat{\mu}_{2,n} - \hat{\mu}_{1,n}^2} \end{aligned}$$

End of Example

Normal mean 0

We have a 1 to 1 mapping $\Theta \mapsto R^+$ given by

$$\theta \mapsto \mathrm{E}_\theta(X^2) = \mu_2 = \theta \ .$$

Thus the method of moments estimator is

$$\tilde{\theta}_n = \hat{\mu}_{2,n} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

This is an example where we do not use $\hat{\mu}_{1,n} = \bar{X}_n$ to construct our method of moments estimator.

End of Example

Return to the exponential θ example. We also have

$$E_\theta(X^2) = \frac{2}{\theta^2}$$

so the method of moments also gives an estimator based on solving

$$\frac{2}{\theta^2} = \hat{\mu}_{2,n}$$

yielding an estimator

$$\tilde{\tilde{\theta}}_n = \frac{\sqrt{2}}{\sqrt{\hat{\mu}_{2,n}}}$$

We now have two estimators $\tilde{\theta}_n$ and $\tilde{\tilde{\theta}}_n$. Which estimator is better?

We can also obtain a third estimator, using the property that $\mathrm{Var}(X) = \frac{1}{\theta^2}$. This suggests, as an estimator of θ based on the same general idea as the method of moments estimator, the following :

$$\tilde{\theta}_{3,n} = \frac{1}{\sqrt{S_n^2}}$$

where S_n^2 is the sample variance.

Which is the best estimator?

To answer this we need to think about what this question might mean and how we can proceed to compare estimators.

The paradigm for determining properties of estimators is to to use the sampling distribution, for each possible $\theta \in \Theta$, of an estimator. When possible we do this for each relevant sample size n, especially if we can determine the sampling distribution of the estimator, say $\hat{\theta}_n$. Often however the sampling distribution of $\hat{\theta}_n$ is not easy to calculate. In this case we often use one of (i) asymptotic or approximate sampling distribution for n large or (ii) for an appropriate and manageable set of values $\theta_1, \ldots, \theta_k$ use simulation or numerical methods or (iii) use non-parametric methods. (i) is often based on an approximation derived from some methods related to the central limit theorem. (ii) typically uses Monte Carlo simulation methods and is sometimes called the parametric bootstrap. (iii) there are some special methods such as the Wilcoxon rank test and various so called permutation tests; the non-parametric bootstrap.

The non-parametric bootstrap is a very useful method in practice. However before trying to understand some of the intuition behind the bootstrap method it is very helpful to understand parametric estimation and the parametric bootstrap.

To compare two estimators, say $\hat{\theta}_{1,n}$ and $\hat{\theta}_{2,n}$ we use properties of their sampling distributions. Ideally we would say $\hat{\theta}_{1,n}$ is a better estimator of θ that $\hat{\theta}_{2,n}$ if it is closer in probability, or more specifically if

$$P_\theta\left(\mid \hat{\theta}_{1,n} - \theta \mid < a\right) > P_\theta\left(\mid \hat{\theta}_{2,n} - \theta \mid < a\right) \tag{2}$$

for every number $a > 0$. This is generally not the case, but it gives us way of contemplating *better* in a probabilistic sense.

See the R script file for this example to see how we can use computer simulation to evaluate this, at least for one specific parameter value.

In general this is not easy to do. However if $\hat{\theta}_{1,n}$ and $\hat{\theta}_{2,n}$ are unbiased estimators of θ and are approximately normally distributed this idea is easier to implement.

The R script file shows this for a sample size $n = 40$ for two estimators for an iid sample from an exponential parameter λ family. For $\hat{\lambda}_{1,n} = \frac{1}{\bar{X}_n}$ the delta method applies, giving

$$\sqrt{n}(\hat{\lambda}_{1,n} - \lambda) \to N(0, \lambda^2)$$

in distribution as $n \to \infty$. The R script shows this approximation is reasonable for $n = 40$. The same is true for the estimator $\hat{\lambda}_{2,n} = 1/\sqrt{S_n^2}$, but we need further properties about convergence to verify this.

In this case

$$\sqrt{n}(\hat{\theta}_{1,n} - \theta) \to W_1 \sim N(0, \gamma_1^2)$$

in distribution as $n \to \infty$. In the above example the variance for the limit distribution would be λ^2.

Therefore

$$\begin{aligned} P_\theta\left(\mid \hat{\theta}_{1,n} - \theta \mid < a \right) &= P_\theta\left(\frac{\sqrt{n} \mid \hat{\theta}_{1,n} - \theta \mid}{\gamma_1} < \frac{a\sqrt{n}}{\gamma_1} \right) \\ &\approx P(|Z| < \frac{a\sqrt{n}}{\gamma_1}) \\ &= P(|W_1| < a\sqrt{n}) \end{aligned}$$

Similarly

$$\sqrt{n}(\hat{\theta}_{2,n} - \theta) \to W_2 \sim N(0, \gamma_2^2)$$

in distribution as $n \to \infty$. Therefore

$$\begin{aligned} P_\theta\left(\mid \hat{\theta}_{2,n} - \theta \mid < a \right) &= P_\theta\left(\frac{\sqrt{n} \mid \hat{\theta}_{2,n} - \theta \mid}{\gamma_1} < \frac{a\sqrt{n}}{\gamma_2} \right) \\ &\approx P(|Z| < \frac{a\sqrt{n}}{\gamma_2}) \\ &= P(|W_2| < a\sqrt{n}) \end{aligned}$$

Thus we can implement the idea behind (2) by comparing these two approximate normal probabilities. In particular if $\text{Var}(\hat{\theta}_{1,n}) < \text{Var}(\hat{\theta}_{1,n})$ then estimator $\hat{\theta}_{1,n}$ is closer to θ than is $\hat{\theta}_{2,n}$ in this probability sense.

If estimators have an approximate normal distribution, and are unbiased, then we can decide on the better estimator by comparing variances. This single number, variance, replaces an in principle more complicated comparison (2).

What happens if the estimators are not unbiased?

$$\begin{aligned} \mathrm{E}_\theta\left((\hat{\theta}_n - \theta)^2 \right) &= \mathrm{E}_\theta\left((\hat{\theta}_n - \mathrm{E}_\theta(\hat{\theta}_n) + \mathrm{E}_\theta(\hat{\theta}_n) - \theta)^2 \right) \\ & \text{add and subtract the mean of the estimator} \\ &= \mathrm{E}_\theta\left((\hat{\theta}_n - \mathrm{E}_\theta(\hat{\theta}_n))^2 \right) + 2\mathrm{E}_\theta\left((\hat{\theta}_n - \mathrm{E}_\theta(\hat{\theta}_n)) \right) \times \left(\mathrm{E}_\theta\left(\hat{\theta}_n \right) - \theta \right) + \left(\mathrm{E}_\theta\left(\hat{\theta}_n \right) - \theta \right)^2 \\ &= \mathrm{E}_\theta\left((\hat{\theta}_n - \mathrm{E}_\theta(\hat{\theta}_n))^2 \right) + \left(\mathrm{E}_\theta\left(\hat{\theta}_n \right) - \theta \right)^2 \\ &= \text{Var}(\hat{\theta}_n) + \left(\text{Bias}(\hat{\theta}_n) \right)^2 \end{aligned}$$

We call this term the Mean Square Error (MSE) of $\hat{\theta}_n$

$$\mathrm{MSE}(\hat{\theta}_n) = \mathrm{Var}(\hat{\theta}_n) + \Big(\mathrm{Bias}(\hat{\theta}_n)\Big)^2$$

Thus to compare two estimators that have a normal or approximate normal distribution, we can compare them by comparing their MSE. The estimator with the smaller MSE is the better estimator.

We return to this idea in Rice Chapter 8.7.