Stat 463/853-2021 Lecture 18 OCT 27, 2021

Parameter estimation: method of moments

In Statistics, one always starts with the observed values of random variables, or **data**,

$$X_1 = x_1, ..., X_n = x_n. \tag{1}$$

Based on the data, a statistician often wants to fit a distribution to the given sample. A rough preliminary idea on what kind of distribution could be used, may be based on the *histogram, or block plot,* of the data.

For instance, if the **block plot** of the data looks roughly **symmetric**, one may think of a fitting a **normal distribution**, with some parameters μ and σ^2. After a normal distribution has been selected, one would still have to estimate its parameters, μ and σ^2.

If the data is **positive and skewed** to the right, one could go for an *exponential distribution* $\mathcal{E}(\lambda)$, or a *gamma* $\Gamma(\alpha, \beta)$.

If data are supported by a **bounded interval**, one could opt for a *uniform distribution* $\mathcal{U}[a, b]$ on an interval $[a, b]$, and then estimate the end-points of this interval, a and b. If the interval $[a, b]$ coincides with $[0, 1]$, one could fit a more general *beta distribution* $\mathrm{B}(\alpha, \beta)$. Then, the unknown parameters α and β need to be estimated.

If data were **discrete**, one could think of a *Poisson distribution* $\mathcal{P}(\lambda)$, or a *geometric distribution* $\mathcal{G}(p)$. Sometimes, the available data can make us think of fitting a *Bernoulli, or a binomial, or a multinomial,* distributions.

In each case, there will be some unknown parameters to estimate based on the available data. Depending on the type of distribution, these parameters may have different meaning, like in following distributions:

$$\mathcal{N}(\mu, \sigma^2), \quad \mathcal{E}(\lambda), \quad \Gamma(\alpha, \lambda), \quad \mathcal{B}(\alpha, \beta), \quad \mathcal{U}[a, b],$$

$$\mathcal{B}(p), \quad \mathcal{M}(p_1, p_2, p_3), \quad \mathcal{P}(\lambda), \quad G(p), \quad \textit{etc.}$$

So, the problem arises as to how to choose these parameters; or, as statisticians say, **estimate** them, based on the available data. There are two classical methods of estimation, each of them having its own advantage. We will first discuss the so-called *method of moments* for estimation of unknown parameters.

The method of moments. Assume for simplicity, first, that there is only one unknown parameter to be estimated. Let us call this unknown parameter θ. Thus, our data comes from i.i.d. random variables, with a given pdf/pmf,

$$X_1, ..., X_n \quad \textbf{i.i.d.} \quad \sim \quad f(x|\theta),$$

where θ is a single *unknown* parameter, and we want an **estimator** for θ based on the given data (1). By an estimator, we mean *any function* of the data,

$$\bar{\theta}_n = \bar{\theta}_n(x_1, ..., x_n).$$

When the data is given, the value of such a function is fixed, or non-random. However, often we are interested in its expectation, or its mean squared error. In all such cases, we view our estimator as *a realization* of the corresponding *random variable*,

$$\bar{\theta}_n = \bar{\theta}_n(x_1, ..., x_n) = \bar{\theta}_n(X_1, ..., X_n).$$

In some cases, the parameter θ may *coincide* with the mean value, $\theta = \mu = \mathbf{E}X_i$, like in the cases of **normal, exponential, Bernoulli, or Poisson** distributions. For all such cases, we already know from the previous lecture, how to construct an estimator and a corresponding confidence interval. Indeed, a *consistent and unbiased* estimator of $\mu = \theta$ is given by

$$\bar{\theta}_n = \bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \quad \overset{p}{\to} \quad \mu = \theta,$$

and a $(1-\alpha)100\%$ CI for $\mu = \theta$ is given by

$$\bar{\theta}_n \pm z(\alpha/2)\sqrt{\frac{\hat{\sigma}_n^2}{n}},$$

where $\hat{\sigma}_n^2$ is any consistent estimator of the variance σ^2 (if it is unknown, which is typically the case).

Although, generally, μ does not necessarily coincide with parameter θ, it is always a function of θ,

$$\mu = \mathbf{E}X_i = h(\theta), \tag{2}$$

which can be found explicitly. For instance, we know that in the case of *geometric distribution*, with unknown parameter $\theta = p$,

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

Then, we can *express* the unknown parameter θ in terms of the mean,

$$\theta = h^{-1}(\mu) := g(\mu). \tag{3}$$

Thus, in the case of geometric distribution,

$$\theta = g(\mu) = \frac{1}{\mu}.$$

Of course, here the true value of μ is still unknown, as is the parameter θ. However, for μ we always have a *consistent estimator*, $\bar{X}_n$. By replacing the mean value μ in (3) by its consistent estimator $\bar{X}_n$, we obtain the **method of moments estimator (MME)** of θ,

$$\bar{\theta}_n = g(\bar{X}_n). \tag{4}$$

For instance, in the case of geometric distribution, $\bar{\theta}_n = 1/\bar{X}_n$.

Function $\mu = h(\theta)$ and its inverse function $\theta = g(\mu)$, connecting the mean value μ to the unknown parameter θ, will play a central role in our discussion. In the discrete case,

$$\mu = h(\theta) = \sum_x x f(x|\theta),$$

while in the continuous case,

$$\mu = h(\theta) = \int x f(x|\theta)\, dx.$$

In most cases of interest, the function $h(\theta)$ is *invertible*. This is guaranteed if, for instance, $h'(\theta) > 0$ (i.e., $h(\theta)$ is strictly increasing), or if $h'(\theta) < 0$ ($h(\theta)$ is strictly decreasing). Then, by the so-called *inverse function theorem*, there is a function $g(\mu)$ such that

$$\underline{\theta = g(\mu)} \quad \text{and} \quad \mu = h(\theta) = h(g(\mu)).$$

Moreover, function $g(\mu)$ is differentiable, and

$$\underline{g'(\mu) = \frac{1}{h'(\theta)}.} \tag{5}$$

Indeed, one has

$$1 = \frac{d\mu}{d\mu} = h'(g(\mu)) \cdot g'(\mu) = h'(\theta) \cdot g'(\mu). \qquad \square$$

Let us take a closer look at our MME estimator (4). In studying it, we will use all what we have already learned about different modes of convergence of random variables.

1. The MME estimator $\bar{\theta}_n$ is always *consistent*. Indeed, since $g(\mu)$ is a continuous function, by the **Service theorem 1**,

$$\bar{\theta}_n = g(\bar{X}_n) \quad \overset{p}{\to} \quad g(\mu) = \theta.$$

2. Denote $\mathbf{Var} X_i = \sigma^2 = \sigma^2(\theta)$. By the CLT,

$$\bar{X}_n \quad \overset{d}{\approx} \quad \mathcal{N}\left(\mu, \frac{\sigma^2(\theta)}{n}\right).$$

Hence, by the δ-method and (5),

$$\hat{\theta}_n = g(\bar{X}_n) \quad \overset{d}{\approx} \quad \mathcal{N}\left(g(\mu), \frac{(g'(\mu))^2\sigma^2(\theta)}{n}\right) = \mathcal{N}\left(\theta, \frac{\sigma^2(\theta)}{(h'(\theta))^2 n}\right). \tag{6}$$

We see that the *asymptotic variance* of our MME $\bar{\theta}_n$ essentially is determined by the function

$$AV(\theta) = \frac{\sigma^2(\theta)}{(h'(\theta))^2},$$

so that

$$\bar{\theta}_n \quad \overset{d}{\approx} \quad \mathcal{N}\left(\theta, \frac{AV(\theta)}{n}\right).$$

Precisely, this means that

$$\frac{\bar{\theta}_n - \theta}{\sqrt{\frac{\sigma^2(\theta)}{n(h'(\theta))^2}}} \quad \xrightarrow{d} \quad Z.$$

3. In order to construct corresponding confidence intervals, one can use the **plug-in method.** Assuming that both functions $\sigma^2(\theta)$ and $h'(\theta)$ are continuous, by the *Slutsky theorem* and *Service theorem 2*,

$$\frac{\bar{\theta}_n - \theta}{\sqrt{\frac{\sigma^2(\bar{\theta}_n)}{n(h'(\bar{\theta}_n))^2}}} = \frac{\sqrt{\frac{\sigma^2(\theta)}{(h'(\theta))^2}}}{\sqrt{\frac{\sigma^2(\bar{\theta}_n)}{(h'(\bar{\theta}_n))^2}}} \cdot \frac{\bar{\theta}_n - \theta}{\sqrt{\frac{\sigma^2(\theta)}{n(h'(\theta))^2}}} \quad \xrightarrow{d} \quad Z.$$

This leads, in the usual way, to the **approximate** $(1-\alpha)100\%$ **CI**,

$$\mathbf{P}\left(\bar{\theta}_n + z(\alpha/2)\sqrt{\frac{\sigma^2(\bar{\theta}_n)}{n(h'(\bar{\theta}_n))^2}} \le \theta \le \bar{\theta}_n + z(\alpha/2)\sqrt{\frac{\sigma^2(\bar{\theta}_n)}{n(h'(\bar{\theta}_n))^2}}\right) \quad \to \quad 1-\alpha, \tag{7}$$

where, as always $z(\alpha)$ is the critical value such that

$$\mathbf{P}(Z \ge z(\alpha)) = \alpha.$$

The above confidence interval can be written is shorter forms as

$$\left[\bar{\theta}_n - z(\alpha/2)\sqrt{\frac{AV(\bar{\theta}_n)}{n}}, \bar{\theta}_n + z(\alpha/2)\sqrt{\frac{AV(\bar{\theta}_n)}{n}}\right], \quad \text{or as} \quad \bar{\theta}_n \pm z(\alpha/2)\sqrt{\frac{AV(\bar{\theta}_n)}{n}}.$$

Students need to learn well formulas (6)–(7).

The generalized method of moments. The ideas and methods leading to the MME are, in fact, much more general than what immediately meets the eye. Suppose that – for whatever reason – we don't want or can't use the observations

X_i themselves, but prefer to use instead some other random variables, $Y_i = u(X_i)$, based on X_i. Then instead of X_i, we consider the moments of Y_i,

$$\mu = \mathbf{E}Y_i = \mathbf{E}u(X_i) = h(\theta), \qquad \mathbf{Var}Y_i = \mathbf{Var}\,u(X_i) = \sigma^2(\theta), \qquad \theta = h^{-1}(\mu) = g(\mu).$$

Most of the standard textbooks, consider only the case $Y_i = u(X_i) = X_i^k$, where $h(\theta) = \mathbf{E}X_i^k$ is the so-called *k-th order moment* of X_i. This is the *classical* method of moments. In fact, we can consider *any* function $Y_i = u(X_i)$, and we will call $h(\theta) = \mathbf{E}u(X_i)$ a *generalized moment.*

Of course, in that case, the sample mean $\bar{X}_n$ will be replaced by the *generalized sample moment*

$$\bar{Y}_n = \frac{u(X_1) + \cdots + u(X_n)}{n}.$$

In the particular case $u(X_i) = X_i^k$, $\bar{Y}_n$ coincides with the k-th order sample moment,

$$\bar{Y}_n = \frac{X_1^k + \cdots + X_n^k}{n}.$$

Notice, that nothing of importance has essentially changed, only now, instead of X_i, we have used the transformed random variables $Y_i = u(X_i)$. Of course, here $\sigma^2 = \mathbf{Var}\,Y_i = \mathbf{Var}\,u(X_i)$. The corresponding *generalized MME* is then

$$\bar{\theta}_n = g(\bar{Y}_n).$$

One of the advantages of the generalized method of moments is that we can choose *any function* $u(x)$ which is more convenient, or easier to deal with. The method always works, with the only exception when $h'(\theta) \equiv 0$, or $\mu = h(\theta) = \text{const}$. The meaning of this limitation is clear. Indeed, if $\mu = h(\theta) = \text{const}$, then, even if we knew the value of μ precisely, it would tell us nothing about the true value of θ!

Later on, we will touch on the issue of the most "efficient" choice of the function $u(x)$. For now, consider two illustrating examples. In both of them, we will have an i.i.d. sample X_i from the so-called *double exponential, or Laplace,* distribution.

Example: double exponential distribution. Let for $-\infty < x < \infty$,

$$f(x|\lambda) = \frac{\lambda}{2} e^{-\lambda|x|},$$

where $\lambda > 0$ if the **unknown parameter**. Here, due to the symmetry of the pdf,

$$\mu = h(\lambda) = \mathbf{E}X = \frac{\lambda}{2} \int_{-\infty}^{\infty} x e^{-\lambda|x|}\, dx = 0.$$

(Recall the geometric meaning of the definite integral as the algebraic sum – with signs – of the areas, enclosed between the integrand and the real axis!) So, we cannot use the first moment $\mu = \mathbf{E}X_i$, since it does not tell us anything about the true value of λ!

Still, there are plenty of other choices; for instance,

$$u_1(x) = |x|, \quad \text{or} \quad u_2(x) = x^2.$$

We will derive the MME's for these two functions, and then decide which of the two resulting estimators is actually better. The corresponding *(generalized) sample moments* are

$$\bar{Y}_1 = \frac{\sum_{i=1}^n |X_i|}{n} \quad \text{and} \quad \bar{Y}_2 = \frac{\sum_{i=1}^n X_i^2}{n}.$$

We will need the familiar *gamma integral*,

$$\int_0^\infty x^{\alpha-1} e^{-\lambda x}\, dx = \frac{\Gamma(\alpha)}{\lambda^\alpha}.$$

In particular,

$$\int_0^\infty x^{n-1} e^{-\lambda x}\, dx = \frac{\Gamma(n)}{\lambda^n} = \frac{(n-1)!}{\lambda^n}.$$

First, let us calculate the corresponding means $\mu = h(\lambda)$ and their derivatives $h'(\lambda)$.

$$h_1(\lambda) = \mathbf{E}Y_1 = \mathbf{E}|X| = \frac{\lambda}{2}\int_{-\infty}^\infty |x| e^{-\lambda|x|}\, dx = \lambda \int_0^\infty x^{2-1} e^{-\lambda x}\, dx = \lambda \cdot \frac{\Gamma(2)}{\lambda^2} = \frac{1}{\lambda}, \tag{8}$$
$$h_1'(\lambda) = -\frac{1}{\lambda^2},$$

$$h_2(\lambda) = \mathbf{E}Y_2 = \mathbf{E}X^2 = \frac{\lambda}{2}\int_{-\infty}^\infty x^2 e^{-\lambda|x|}\, dx = \lambda \int_0^\infty x^{3-1} e^{-\lambda x}\, dx = \lambda \cdot \frac{\Gamma(3)}{\lambda^3} = \frac{2}{\lambda^2}, \tag{9}$$
$$h_2'(\lambda) = -\frac{4}{\lambda^3}.$$

Next, from equations (8), (9), one can express parameter of interest λ in terms of the generalized moments:

$$\lambda = \frac{1}{\mathbf{E}Y_1},$$

and

$$\lambda = \sqrt{\frac{2}{\mathbf{E}Y_2}}.$$

Replacing the generalized moments by the corresponding sample moments leads to the corresponding generalized MME's:

$$\bar{\lambda}_1 = \frac{1}{\bar{Y}_1},$$

and

$$\bar{\lambda}_2 = \sqrt{\frac{2}{\bar{Y}_2}}.$$

As we already know, both estimators are **consistent!** To construct the corresponding CI's, we need $\sigma_1^2(\lambda) = \mathbf{Var}\, Y_1$ and $\sigma_2^2(\lambda) = \mathbf{Var}\, Y_2$. Note that by (9),

$$\mathbf{E}Y_1^2 = \mathbf{E}X^2 = \frac{2}{\lambda^2}.$$

Hence,

$$\sigma_1^2(\lambda) = \mathbf{E}\, Y_1^2 - (\mathbf{E}Y_1)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

This also gives

$$AV_1(\lambda) = \frac{\sigma_1^2(\lambda)}{(h_1'(\lambda))^2} = \frac{\frac{1}{\lambda^2}}{\frac{1}{\lambda^4}} = \lambda^2.$$

Thus, the approximate $(1-\alpha)100\%$ CI, based on the MME $\bar{\lambda}_1$, is

$$\bar{\lambda}_1 \pm z(\alpha/2)\sqrt{\frac{AV_1(\bar{\lambda}_1)}{n}} = \bar{\lambda}_1 \pm z(\alpha/2)\sqrt{\frac{\bar{\lambda}_1^2}{n}} = \bar{\lambda}_1 \pm z(\alpha/2)\frac{\bar{\lambda}_1}{\sqrt{n}} = \bar{\lambda}_1\left(1 \pm \frac{z(\alpha/2)}{\sqrt{n}}\right).$$

Now, let us calculate $\sigma_2^2(\lambda) = \mathbf{Var}Y_2$. We have already found $\mathbf{E}Y_2 = \frac{2}{\lambda^2}$. Next, from the gamma integral,

$$\mathbf{E}Y_2^2 = \mathbf{E}X^4 = \frac{\lambda}{2}\int_{-\infty}^{\infty} x^4 e^{-\lambda|x|}\,dx = \lambda\int_0^{\infty} x^{5-1}e^{-\lambda x}\,dx = \lambda\cdot\frac{\Gamma(5)}{\lambda^5} = \frac{4!}{\lambda^4} = \frac{24}{\lambda^4}.$$

Thus,

$$\sigma_2^2(\lambda) = \mathbf{Var}\, Y_2 = \mathbf{E}Y_2^2 - (\mathbf{E}Y_2)^2 = \frac{24}{\lambda^4} - \frac{4}{\lambda^4} = \frac{20}{\lambda^4},$$

and

$$AV_2(\lambda) = \frac{\sigma_2^2(\theta)}{(h_2'(\theta))^2} = \frac{\frac{20}{\lambda^4}}{\frac{16}{\lambda^6}} = \frac{5}{4}\lambda^2.$$

The approximate $(1-\alpha)100\%$ CI, based on the MME $\bar{\lambda}_2$, is

$$\bar{\lambda}_2 \pm z(\alpha/2)\sqrt{\frac{AV_2(\bar{\lambda}_2)}{n}} = \bar{\lambda}_2 \pm z(\alpha/2)\sqrt{\frac{5\bar{\lambda}_2^2}{4n}} = \bar{\lambda}_2 \pm z(\alpha/2)\frac{\sqrt{5}\ \bar{\lambda}_2}{2\sqrt{n}} = \bar{\lambda}_2\left(1 \pm \frac{\sqrt{5}}{2\sqrt{n}}z(\alpha/2)\right).$$

Assume that the relative accuracy of the two estimators is determined by the length of the corresponding confidence intervals. Note that since both estimators are consistent, $\bar{\lambda}_1 \approx \bar{\lambda}_2 \approx \lambda$. Thus, the length of the second CI is approximately $\sqrt{5}/2 \approx 1.12$ times that of the first CI. We can conclude that the CI based on $\bar{\lambda}_1$ is more accurate (smaller by approximately 12%, for large n)!