Delta Method

Often estimators are functions of other random variables, for example in the method of moments. These functions of random variables can sometimes inherit a normal approximation from the underlying random variables. Earlier in the course we obtained a result where a continuous function of a sequence of consistent estimators also inherited the property of being consistent estimators. The delta method allows a normal approximation (a normal central limit type or result, that is convergence in distribution to a normal distribution) for a continuous and differentiable function of a sequence of r.v.s that already has a normal limit in distribution.

Example : Method of Moments for Exponential Distribution.

$X_i, i = 1, 2, \ldots, n$ are iid exponential, λ with pdf

$$f(x; \lambda) = \lambda e^{-\lambda x} \mathrm{I}(x > 0)$$

The first moment is then $\mu_1(\lambda) = \frac{1}{\lambda}$. The the method of moments estimator is

$$\hat{\lambda}_n = \frac{1}{\bar{X}_n}$$

Notice this is of the form $\hat{\lambda}_n = g(\bar{X})$ where $g : R^+ \mapsto R^+$ with $g(x) = \frac{1}{x}$.

Theorem 1 (Delta Method) *Suppose $\bar{X}_n$ has an asymptotic normal distribution, that is*

$$\sqrt{n}(\bar{X}_n - \mu) \to N(0, \gamma^2)$$

in distribution as $n \to \infty$. Suppose g is a function that is continuous and also has a derivative g' at μ, and that $g'(\mu) \neq 0$. Then

$$\sqrt{n}\left(g(\bar{X}_n) - g(\mu)\right) \to N(0, (g'(\mu))^2\gamma^2)$$

Remark : The condition $g'(\mu) \neq 0$ is actually only needed so that

$$\frac{\sqrt{n}\left(g(\bar{X}_n) - g(\mu)\right)}{|g'(\mu)|\gamma} \to N(0,1)$$

in distribution as $n \to \infty$.

Remark : Theorem 1 is called the delta method.

Proof (Outline)

The first order Taylor approximation of g about the point μ, and evaluated at the random variable $\bar{X}_n$ is

$$g(\bar{X}_n) \approx g(\mu) + g'(\mu)\left(\bar{X}_n - \mu\right)$$

Subtract $g(\mu)$ from both sides and multiply by $\sqrt{n}$ gives

$$\sqrt{n}\left(g(\bar{X}_n) - g(\mu)\right) \approx g'(\mu)\sqrt{n}\left(\bar{X}_n - \mu\right) \to N(0, (g'(\mu))^2\gamma^2)$$

Remark : A more careful study of Taylor's formula with remainder is needed to justify all steps in this approximation. For our purposes in this course these details are not needed.

Remark : Notice this delta method is an extension of the idea used earlier to approximate moments, specifically to approximate means and variances.

Example Continued : For the exponential example we have

$$\sqrt{n}\left(\bar{X}_n - \frac{1}{\lambda}\right) \to N(0, \frac{1}{\lambda^2})$$

Using the function $g(x) = 1/x$ expanded about $\mu = \frac{1}{\lambda}$, and noting

$$\hat{\lambda}_n = g(\bar{X}_n) = \frac{1}{\bar{X}_n} \text{ and } g(\mu) = \lambda$$

we thus have as an application of the delta method (Theorem 1)

$$\sqrt{n}\left(\hat{\lambda}_n - \lambda\right) \approx -\frac{1}{\frac{1}{\lambda^2}}\sqrt{n}\left(\bar{X}_n - \frac{1}{\lambda}\right) \to N\left(0, \frac{\lambda^4}{\lambda^2}\right) = N(0, \lambda^2)$$

In our discussion of the Central Limit Theorem (CLT) we studied in the case of iid sampling from a distribution with 2 moments that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\sigma^2}}$$

converges in distribution to $N(0,1)$. The student should refer back to that section of the text and the notes to review the notation and conditions and the statement of the CLT. We also discussed, but without proof, that one can replace the population variance σ^2 in the denominator with the r.v. S_n^2 (the sample variance) and still obtain convergence in distribution to $N(0,1)$. This is due to S_n^2 being a consistent estimator of σ^2, that is $S_n^2 \to \sigma^2$ in probability as $n \to \infty$.

One can replace the population variance σ^2 in the denominator with any other consistent estimator. Therefore in the exponential example above we have two random variables that may be used

$$\begin{aligned} Z_{1,n} &= \frac{\sqrt{n}(\bar{X}_n - \frac{1}{\lambda})}{\sqrt{\frac{1}{\hat{\lambda}_n^2}}} \\ Z_{1,n} &= \frac{\sqrt{n}(\bar{X}_n - \frac{1}{\lambda})}{\sqrt{S_n^2}} \end{aligned}$$

both of which converge in distribution to $N(0,1)$. Both of these may be used to construct an approximate or asymptotic confidence interval for $\mu = \mathrm{E}(X)$. This will be discussed later in the course.

There is also a delta method for random vectors. It is described in the same fashion. For simplicity of writing and since we only apply it in our course for dimension 2, we write it in this form. For this we need a preliminary theorem.

Theorem 2 *Suppose $X_i, i \geq 1$ are iid with finite 4-th moments. Write*

$$\hat{\mu}_{1,n} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } \hat{\mu}_{2,n} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

and let μ_k be the k-th population moments. Then

$$\sqrt{n}\,(\hat{\mu}_{1,n} - \mu_1, \hat{\mu}_{2,n} - \mu_2) \to \ BVN(0, A)$$

in distribution as $n \to \infty$, where BVN means bivariate normal with mean vector 0 and variance (or covariance) matrix

$$A = \begin{pmatrix} \mu_2 - \mu_1^2 & \mu_3 - \mu_1\mu_2 \\ \mu_3 - \mu_1\mu_2 & \mu_4 - \mu_2^2 \end{pmatrix}$$

Remark : This is a consequence of a multivariate central limit theorem, and is not proven in the course.

In the Theorem above the matrix A is a variance matrix. What are some properties of this matrix that we can use below?

Suppose that b_1, b_2 are two constants. Set $b = (b_1, b_2)$, a 1 by 2 matrix (row vector). Suppose also that $(X, Y)^T$ is a 2 by 1 matrix (column vector) with components the r.v.s X, Y and variance matrix A. Then

$$b \begin{pmatrix} X \\ Y \end{pmatrix} = b_1 X + b_2 Y \ .$$

Then

$$\begin{aligned} \mathrm{Var}(b_1 X + b_2 Y) &= b_1^2 \mathrm{Var}(X) + 2 b_1 b_2 \mathrm{Cov}(X, Y) + b_2^2 \mathrm{Var}(X) \\ &= bAb^T \ . \end{aligned}$$

The student should write out the matrix multiplication and verify this.

Similarly if B is a matrix of appropriate size then $W = B(X, Y)^T$ is a random vector. In this case the random vector W has variance matrix

$$BAB^T \ .$$

End of Remark

Theorem 3 (below) is the delta method applied to a function of $(\hat{\mu}_{1,n}, \hat{\mu}_{2,n})$. We state this rather than the general delta method to avoid more complicated notation. The idea is the same as used in Theorem 1, but is based on working with bivariate normal distributions, and more generally with multivariate normal distributions.

Theorem 3 *Suppose the conditions of Theorem 2. Suppose g is a function of two variables mapped to two variables, that is continuous and also has a derivative g' at (μ_1, μ_2), and that $g'(\mu_1, \mu_2)$ is non zero. Then*

$$\sqrt{n}\,(g(\hat{\mu}_{1,n}, \hat{\mu}_{2,n}) - g(\mu_1, \mu_2)) \to N(0, g'(\mu_1, \mu_2) A g'(\mu_1, \mu_2)^T)$$

Remark :

If g maps pairs into R, then we interpret $g'(\mu_1, \mu_2)$ as a row vector of length 2. In this case

$$g'(\mu_1, \mu_2) A g'(\mu_1, \mu_2)^T$$

is a $(1 \times 2) \times (2 \times 2) \times (2 \times 1) = 1 \times 1$ matrix, that is a real number.

If g maps pairs into $R \times R$, then we interpret $g'(\mu_1, \mu_2)$ as a 2 by 2 matrix.

$$g'(\mu_1, \mu_2) A g'(\mu_1, \mu_2)^t$$

is a product of a 2 by 2 matrix g', a 2 by 2 matrix A and 2 by 2 matrix $(g')^T$ (the transpose of the first 2 by 2 matrix). Since variance matrices are positive definite this resulting matrix is also positive definite, that is it is a variance matrix.

End of Remark

For an application of this result, see the rainfall data example and the method of moments for that example. Part of this example is discussed in more detail later in this handout. In that example we use the following fact, that for a bivariate normal distribution the marginal distribution of each component is normal. For example for the first component we have a normal distribution with variance given by the (1,1) component of the variance matrix.

When we discussed the central limit theorem (CLT) we stated without proof, that one can replace the population variance σ^2 with a consistent estimator of σ^2, in that case s_n^2 the sample variance, and still retain the convergence in distribution to N(0,1). This same property carries over more generally. In our delta method this corresponding result allows one to replace (the first two for $g(\bar{X}_n)$ and the last two for $g(\hat{\mu}_{1,n}, \hat{\mu}_{2,n})$)

- $\sigma^2 = \text{Var}(X)$ by s_n^2
- $g'(\mu)$ by $g'(\bar{X}_n)$
- The matrix $A = \text{Var}(X, X^2)$ by

$$\hat{A}_n = \begin{pmatrix} \hat{\mu}_{2,n} - \hat{\mu}_{1,n}^2 & \hat{\mu}_{3,n} - \hat{\mu}_{1,n}\hat{\mu}_{2,n} \\ \hat{\mu}_{3,n} - \hat{\mu}_{1,n}\hat{\mu}_{2,n} & \hat{\mu}_{4,n} - \hat{\mu}_{2,n}^2 \end{pmatrix}$$

- $g'(\mu_1, \mu_2)$ by $g'(\hat{\mu}_{1,n}, \hat{\mu}_{2,n})$

We do not write this in a technically proper manner in this course.

Example : Delta Method applied to Gamma method of moment estimator

Recall method of moments estimator for the Gamma parameters are

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

Thus we have

$$\tilde{\lambda}_n = g(\hat{\mu}_{1,n}, \hat{\mu}_{2,n})$$

where the function g is given by

$$g(x, y) = \frac{x}{y - x^2} \ .$$

Thus we find

$$\begin{aligned}\frac{\partial g}{\partial x} &= \frac{y + x^2}{(y - x^2)^2} \\ \frac{\partial g}{\partial y} &= -\frac{x}{(y - x^2)^2}\end{aligned}$$

Thus the matrix (or row vector) for the partial derivative of g is

$$\begin{aligned}g'(x, y) &= \frac{\partial g(x, y)}{\partial (x, y)} \\ &= \left(\frac{y + x^2}{(y - x^2)^2} \ , \ -\frac{x}{(y - x^2)^2} \right) \ .\end{aligned}$$

Aside If one writes the partial derivative as a column vector

$$g'(x, y) = \begin{pmatrix} \frac{y+x^2}{(y-x^2)^2} \\ -\frac{x}{(y-x^2)^2} \end{pmatrix}$$

then one has to be careful to use the appropriate transposes in the matrix multiplications above. It does not matter as long as one is consistent. In this handout we will use the row vector.

End of Aside

Recall $\lambda = g(\mu_1, \mu_2)$, by the construction of our method of moments estimator.

The first order Taylor's formula approximation for $\tilde{\lambda}_n$ is then

$$\begin{aligned} \tilde{\lambda}_n - \lambda &= g(\hat{\mu}_{1,n}, \hat{\mu}_{2,n}) - g(\mu_1, \mu_2) \\ &\approx g'(\mu_1, \mu_2) \begin{pmatrix} \hat{\mu}_{1,n} - \mu_1 \\ \hat{\mu}_{2,n} - \mu_2 \end{pmatrix} \end{aligned}$$

Then

$$\begin{aligned} \sqrt{n}(\tilde{\lambda}_n - \lambda) &\approx g'(\mu_1, \mu_2)\sqrt{n} \begin{pmatrix} \hat{\mu}_{1,n} - \mu_1 \\ \hat{\mu}_{2,n} - \mu_2 \end{pmatrix} \\ &\Rightarrow g'(\mu_1, \mu_2)W \end{aligned}$$

as $n \to \infty$, and where W has the bivariate normal distribution with mean vector 0 and covariance matrix A (see earlier section). The symbol $\Rightarrow$ is used her for shorthand *convergence in distribution*. Thus

$$\begin{aligned} \sqrt{n}(\tilde{\lambda}_n - \lambda) &\Rightarrow g'(\mu_1, \mu_2)W \\ &\sim N(0, g'(\mu_1, \mu_2)Ag'(\mu_1, \mu_2)^T) \ . \end{aligned}$$

In particular we have that for large n, the r.v. $\sqrt{n}(\tilde{\lambda}_n - \lambda)$ has an approximate normal distribution. Thus our estimator has an asymptotic normal distribution approximation.

In particular we can use this to construct confidence intervals for λ.

There are a few additional ideas that are needed to make use of the delte method, Theorem 3, in practice. Recall that it gives a limit normal distribution

$$\sqrt{n}\,(g(\hat{\mu}_{1,n}, \hat{\mu}_{2,n}) - g(\mu_1, \mu_2)) \to N(0, g'(\mu_1, \mu_2) A g'(\mu_1, \mu_2)^T)$$

so that

$$g'(\mu_1, \mu_2) A g'(\mu_1, \mu_2)^T \tag{1}$$

is the variance matrix for the limiting normal distribution. We also need a sample estimate of this; this will play the role of the sample variance in our simpler 1 dimensional standardized sample mean

$$t = \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{S^2}} \ . \tag{2}$$

If the X_i are iid normal then (2) has a student's t distribution. If not, but the sample size n is large, then (2) has approximately a standard normal distribution.

In (1) we estimate A by using the sample momengts

$$\hat{A} = \begin{pmatrix} \hat{\mu}_2 - \hat{\mu}^2 & \hat{\mu}_3 - \hat{\mu}_1\hat{\mu}_2 \\ \hat{\mu}_3 - \hat{\mu}_1\hat{\mu}_2 & \hat{\mu}_4 - \hat{\mu}_2^2 \end{pmatrix}$$

and estimate $g'(\mu_1, \mu_2)$ by also using the sample moments

$$g'(\hat{\mu}_1, \hat{\mu}_2) \ .$$

Here the estimate of the sample moment with data $x_1, x_2, \ldots, x_n$ is

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

Earlier when we needed to keep track of the sample size in our notation we used $\hat{\mu}_{k,n}$, and as we noted earlier we drop the extra subscript n except where we need to pay attention to this sample size.