Taylor Approximation and the Delta Method

Alex Papanicolaou*

April 28, 2009

1 Taylor Approximation

1.1 Motivating Example: Estimating the odds

Suppose we observe $X_1, \ldots, X_n$ independent Bernoulli(p) random variables. Typically, we are interested in p but there is also interest in the parameter $\frac{p}{1-p}$, which is known as the *odds*. For example, if the outcomes of a medical treatment occur with $p = 2/3$, then the odds of getting better is $2:1$. Furthermore, if there is another treatment with success probability r, we might also be interested in the *odds ratio* $\frac{p}{1-p}/\frac{r}{1-r}$, which gives the relative odds of one treatment over another.

If we wished to estimate p, we would typically estimate this quantity with the observed success probability $\hat{p} = \sum_i X_i/n$. To estimate the odds, it then seems perfectly natural to use $\frac{\hat{p}}{1-\hat{p}}$ as an estimate for $\frac{p}{1-p}$. But whereas we know the variance of our estimator $\hat{p}$ is $p(1-p)$ (check this be computing $\text{var}(\hat{p})$), what is the variance of $\frac{\hat{p}}{1-\hat{p}}$? Or, how can we approximate its sampling distribution?

The Delta Method gives a technique for doing this and is based on using a Taylor series approximation.

1.2 The Taylor Series

Definition: *If a function $g(x)$ has derivatives of order r, that is $g^{(r)}(x) = \frac{d^r}{dx^r} g(x)$ exists, then for any constant a, the* Taylor polynomial of order r about a *is*

$$T_r(x) = \sum_{k=0}^{r} \frac{g^{(k)}(a)}{k!}(x-a)^k.$$

While the Taylor polynomial was introduced as far back as beginning calculus, the major theorem from Taylor is that the *remainder* from the approximation, namely $g(x) - T_r(x)$, tends to 0 faster than the highest-order term in $T_r(x)$.

Theorem: *If $g^{(r)}(a) = \frac{d^r}{dx^r} g(x)|_{x=a}$ exists, then*

$$\lim_{x \to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0.$$

*The material here is almost word for word from pp. 240-245 of *Statistical Inference* by George Casella and Roger L. Berger and credit is really to them.

For the purposes of the Delta Method, we will only be considering $r = 1$. Furthermore, we will not be concerned with the remainder term since, (1), we are interested in approximations, and (2), we will have a nice convergence result that says from a probabilistic point of view, the remainder will vanish.

1.3 Applying the Taylor Theorem

Let's now put the first-order Taylor polynomial to use from a statistical point of view: Let $T_1, \ldots, T_k$ be random variables with means $\theta_1, \ldots, \theta_k$, and define $T = (T_1, \ldots, T_k)$ and $\theta = (\theta_1, \ldots, \theta_k)$. Suppose there is a differentiable function $g(T)$ (say, an estimator of some parameter. In our motivating example, $T = p$ and $g(p) = \frac{p}{1-p}$) for which we want an estimate of variance. Define the partial derivatives as

$$g_i'(\theta) = \frac{\partial}{\partial t_i} g(t)|_{t_1=\theta_1,\ldots,t_k=\theta_k},$$

where $t = (t_1, \ldots, t_k)$ is just any k-dimensional coordinates. The first-order Taylor series expansion (this is actually coming from the multivariate version of the Taylor series which shall be addressed later) of g about θ is

$$g(t) = g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(t_i - \theta_i) + \text{Remainder}.$$

So far, we have done nothing special. Now, let's turn this into a statistical approximation by bringing in T and dropping the remainder. This gives

$$g(T) \approx g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(T_i - \theta_i). \tag{1}$$

Continuing, let's take expectations on both sides (noticing that everything but the T_i terms on the right-hand side are non-random) to get

$$\begin{aligned} \mathbf{E}g(T) &\approx g(\theta) + \sum_{i=1}^{k} g_i'(\theta)\mathbf{E}(T_i - \theta_i) \\ &= g(\theta). \end{aligned} \tag{2}$$

We can also approximate the variance of $g(T)$ by

$$\begin{aligned} \text{Var}g(T) &\approx \mathbf{E}[(g(T) - g(\theta))^2] && \text{From Eq. (2).} \\ &\approx \mathbf{E}\left(\left(\textstyle\sum_{i=1}^{k} g_i'(\theta)(T_i - \theta_i)\right)^2\right) && \text{From Eq. (1).} \\ &= \sum_{i=1}^{k} g_i'(\theta)^2 \text{Var}T_i + 2\sum_{i>j} g_i'(\theta) g_j'(\theta) \text{Cov}(T_i, T_j), && (3) \end{aligned}$$

where the last equality comes from expanding the square and using the definitions of variance and covariance. Notice that we have approximated the variance to our estimator $g(T)$ using only the variances and covariances of the T_i, which if the problem is set up well, are not terribly difficult to compute or estimate. Let's now put this to work.

1.4 Continuation: Estimating the Odds

Recall that we wanted to gather some properties about $\frac{\hat{p}}{1-\hat{p}}$ as an estimate of $\frac{p}{1-p}$, where p is a binomial success probability. Using the notation described in the previous section, we take $g(p) = \frac{p}{1-p}$ so that $g'(p) = \frac{1}{(1-p)^2}$ (this is a univariate this case, so $k = 1$ and thus there is only one derivative) and

$$\begin{aligned}\text{Var}\left(\frac{\hat{p}}{1-\hat{p}}\right) &\approx g'(p)^2 \text{Var}(\hat{p}) \\ &= \left(\frac{1}{(1-p)^2}\right)^2 \frac{p(1-p)}{n} = \frac{p}{n(1-p)^3},\end{aligned}$$

giving us an approximation for the variance of our estimator. ‖

1.5 Example: Approximate Mean and Variance

Suppose X is a random variable with $\mathbf{E}X = \mu \neq 0$. If we want to estimate a function $g(\mu)$, a first-order approximation like before would give us

$$g(X) = g(\mu) + g'(\mu)(X - \mu).$$

Thus, if we use $g(X)$ as an estimator of $g(\mu)$, we can say that approximately

$$\begin{aligned}\mathbf{E}g(X) &\approx g(\mu), \\ \text{Var}g(X) &\approx g'(\theta)^2 \text{Var}X.\end{aligned}$$

What is the purpose of this? Well, suppose we take $g(\mu) = 1/\mu$ with μ unknown. If we estimate $1/\mu$ with $1/X$, then we can say

$$\begin{aligned}\mathbf{E}\left(\frac{1}{X}\right) &\approx \frac{1}{\mu}, \\ \text{Var}\left(\frac{1}{X}\right) &\approx \left(\frac{1}{\mu}\right)^4 \text{Var}X.\end{aligned}$$

As we have seen, we can use these Taylor series approximations to estimate the mean and variance estimators. As mentioned earlier, we can generalize this into a convergence result akin to the Central Limit Theorem. This result is known as the *Delta Method*.

2 The Delta Method

2.1 Slutsky's Theorem

Before we address the main result, we first state a useful result, named after Eugene Slutsky.

Theorem: (Slutsky's Theorem) *If $W_n \to W$ in distribution and $Z_n \to c$ in probability, where c is a non-random constant, then*

$$\begin{aligned}W_n Z_n &\to cW \text{ in distribution.} \\ W_n + Z_n &\to W + c \text{ in distribution.}\end{aligned}$$

The proof is omitted.

2.2 Delta Method: A Generalized CLT

Theorem: *Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathcal{N}(0, \sigma^2)$ in distribution. For a given function and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then,*

$$\sqrt{n}(g(Y_n) - g(\theta)) \to \mathcal{N}(0, \sigma^2 g'(\theta)^2) \text{ in distribution.}$$

Proof: The Taylor expansion of $g(Y_n)$ around $Y_n = \theta$ is

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \text{Remainder},$$

where the remainder $\to 0$ as $Y_n \to \theta$. From the assumption that Y_n satisfies the standard CLT, we have $Y_n \to \theta$ in probability, so it follows that the remainder $\to 0$ in probability as well. Rearranging terms, we have

$$\sqrt{n}(g(Y_n) - g(\theta)) = g'(\theta)\sqrt{n}(Y_n - \theta) + \text{Remainder}.$$

Applying Slutsky's Theorem with $W_n = g'(\theta)\sqrt{n}(Y_n - \theta)$ and Z_n as the remainder, we have the right-hand side converging to $\mathcal{N}(0, \sigma^2 g'(\theta)^2)$, and thus the desired result follows. □

2.3 Continuation: Approximate Mean and Variance

Before, we considered the case of just estimating $g(\mu)$ with $g(X)$. Suppose now we have took an i.i.d. random sample of a population to get $X_1, \ldots, X_n$ to get a sample mean $\widehat{X}_n$[1] For $\mu \neq 0$, from the Delta Method we have

$$\sqrt{n}(\frac{1}{\widehat{X}} - \frac{1}{\mu}) \to \mathcal{N}\left(0, \left(\frac{1}{\mu}\right)^4 \text{Var}X_1\right)$$

in distribution.

This is pretty good! But what if we don't know the variance of X_1? Furthermore, we're trying to estimate $1/\mu$ and the variance on the right-hand side requires knowledge of μ. This actually poses no major problem since we shall just estimate everything to get the approximate variance

$$\widehat{\text{Var}}\left(\frac{1}{\widehat{X}}\right) \approx \left(\frac{1}{\widehat{X}}\right) S^2,$$

where $\hat{\sigma}^2$ is an estimate of the variance of X_1, say the sample variance. Now, we know that both $\widehat{X}$ and $\hat{\sigma}^2$ are consistent estimators in that $\widehat{X} \to \mu$ and $\hat{\sigma}^2 \to \sigma^2$ in probability. Thus, $\left(\frac{1}{\widehat{X}}\right)^2 \hat{\sigma} \to \left(\frac{1}{\mu}\right)^2 \sigma$ in probability. This allows us to apply Slutsky's Theorem to get

$$\frac{\sqrt{n}\left(\frac{1}{\widehat{X}} - \frac{1}{\mu}\right)}{\left(\frac{1}{\widehat{X}}\right)^2 \hat{\sigma}} = \frac{\left(\frac{1}{\mu}\right)^2 \sigma}{\left(\frac{1}{\widehat{X}}\right)^2 \hat{\sigma}} \cdot \frac{\sqrt{n}\left(\frac{1}{\widehat{X}} - \frac{1}{\mu}\right)}{\left(\frac{1}{\mu}\right)^2 \sigma} \to \mathcal{N}(0, 1)$$

in distribution. It bears pointing out that the written form of the convergence has changed since now our parameters that were once in the limiting distribution are estimates dependent on n. It would not make much sense to have convergence of $\sqrt{n}\left(\frac{1}{\widehat{X}} - \frac{1}{\mu}\right)$ to a distribution with variance dependent on n.

[1] The statement of the Delta Method allows for great generality of sequences Y_n satisfying the CLT. This is because there are multiple forms of the CLT. Typically, the sample mean is used in these types of approximations and from elementary probability, we know the sample mean is one such sequence of random variables that satisfies the CLT.

3 Second-Order Delta Method

A natural question to ask is, in all the above work, what happens if $g'(\theta) = 0$? To answer this, we go back to the Taylor expansion. Using the notation from the Delta Method theorem, we add the second-order term to get

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \frac{g''(\theta)}{2}(Y_n - \theta)^2 + \text{Remainder}.$$

Since $g'(\theta) = 0$, this gives

$$g(Y_n) - g(\theta) = \frac{g''(\theta)}{2}(Y_n - \theta)^2 + \text{Remainder}.$$

Now, just like $\sqrt{n}(Y_n - \theta)/\sigma^2 \to \mathcal{N}(0, 1)$ in distribution, we also have

$$\frac{n(Y_n - \theta)^2}{\sigma} \to \chi_1^2$$

in distribution where χ_1^2 is a *chi-squared* random variable with 1 *degree of freedom*. This new convergence is all very natural because we are now dealing with a second-order term. The first-order approximation converged to a Gaussian random variable so we could reasonably guess that the second-order term would converge to the square of a Gaussian, which just so happens to be a chi-squared random variable. In precise terms, we give the Second-Order Delta Method:

Theorem: (Second-Order Delta Method) *Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathcal{N}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta) = 0$ and $''(\theta)$ exists and is not 0. Then*

$$n(g(Y_n) - g(\theta)) \to \sigma^2 \frac{g''(\theta)}{2}\chi_1^2 \text{ in distribution.}$$

4 Multivariate Delta Method

We have actually already seen the multivariate precursor to the multivariate extension to the Delta Method. We use an example to illustrate the usage.

4.1 Moments of a Ratio Estimator

Suppose X and Y are random variables with nonsero means μ_X and μ_Y, respectively. The parametric function to be estimated is $g(\mu_X, \mu_Y) = \mu_X/\mu_Y$. It is straightforward to calculate

$$\frac{\partial}{\partial \mu_X} g(\mu_X, \mu_Y) = \frac{1}{\mu_Y}$$

and

$$\frac{\partial}{\partial \mu_Y} g(\mu_X, \mu_Y) = -\frac{\mu_X}{\mu_Y^2}.$$

Using Equations (2) and (3), we have

$$\mathbf{E}\left(\frac{X}{Y}\right) \approx \frac{\mu_X}{\mu_Y}$$

and

$$\begin{aligned}\mathrm{Var}\left(\frac{X}{Y}\right) &\approx \frac{1}{\mu_Y^2}\mathrm{Var}X + \frac{\mu_X^2}{\mu_Y^4}\mathrm{Var}Y - 2\frac{\mu_X}{\mu_Y^3}\mathrm{Cov}(X,Y) \\ &= \left(\frac{\mu_X}{\mu_Y}\right)^2\left(\frac{\mathrm{Var}X}{\mu_X^2} + \frac{\mathrm{Var}Y}{\mu_Y^2} - 2\frac{\mathrm{Cov}(X,Y)}{\mu_X\mu_Y}\right).\end{aligned}$$

We now have an approximation that can be built up from estimates of the means, variances, and covariances of the individual random variables X and Y.‖

We now come to the finale, the multivariate version of the Delta Method. First, we set out a few definitions. Define the random vector $X = (X_1, \ldots, X_p)$ with mean $\mu = (\mu_1, \ldots, \mu_p)$ and covariances $\mathrm{Cov}(X_i, X_j) = \sigma_{ij}$. We shall observe n i.i.d. random ramples of the population of X and denote these samples as $X^{(1)}, \ldots, X^{(n)}$. Furthermore, we shall call the sample means for each element of the vector $\widehat{X}_i = \sum_{k=1}^n X_i^{(k)}, i = 1, \ldots, p$ and $\widehat{X}$ as the vector of sample means. Lastly, we consider the multivariate function $g : \mathbb{R} \mapsto \mathbb{R}$ with $g(x) = g(x_1, \ldots, x_p)$ and use (1) to write

$$g(\widehat{X}_1, \ldots, \widehat{X}_p) \approx g(\mu_1, \ldots, \mu_p) + \sum_{i=1}^p g_i'(\mu_i)(\widehat{X}_i - \mu_i).$$

In vector notation, this is

$$g(\widehat{X}) \approx g(\mu) + \nabla^T g(\mu)(\widehat{X} - \mu),$$

with the abuse of notation that $\nabla^T g(\mu) = (\nabla^T g(x))|_{x=\mu}$.

Theorem: (Multivariate Delta Method) *Let $X^{(1)}, \ldots, X^{(n)}$ be a random sample with $\mathbf{E}X_i^{(k)} = \mu_i$ and $\mathrm{Cov}(X_i^{(k)}, X_j^{(k)}) = \sigma_{ij}$. For a given function g with continuous first partial derivatives and a specific value of $\mu = (\mu_1, \ldots, \mu_p)$ for which $\tau^2 = \sum_i \sum_j \sigma_{ij} g_i'(\mu) g_j'(\mu) > 0$,*

$$\sqrt{n}(g(\widehat{X}^{(1)}, \ldots, \widehat{X}^{(n)}) - g(\mu_1, \ldots, \mu_p)) \to \mathcal{N}(0, \tau^2) \text{ in distribution.}$$

In vector form, this is

Theorem: (Multivariate Delta Method in Vector Form) *Let $X^{(1)}, \ldots, X^{(n)}$ be a random sample with $\mathbf{E}X^{(i)} = \mu$ and covariance matrix $\mathbf{E}(X^{(i)} - \mu)(X^{(i)} - \mu)^T = \Sigma$. For a given function g with continuous first partial derivatives and a specific value of μ for which $\tau^2 = \nabla^T g(\mu) \Sigma \nabla g(\mu) > 0$,*

$$\sqrt{n}(g(\widehat{X}) - g(\mu)) \to \mathcal{N}(0, \tau^2) \text{ in distribution.}$$