Stat 200: Introduction to Statistical Inference **Winter 2023**

Lecture 16: Delta method, plug-in, bootstrap part I

Lecturer: Dominik Rothenhäusler *March 13*

Note: *These lecture notes were written by Art Owen. If you like the material, he gets the credit! These notes have not been subjected to the usual scrutiny reserved for formal publications. They are meant as a memory aid for students who took stat 200 at Stanford University. They may be distributed outside this class only with the permission of Art Owen. Also, Stanford University holds the copyright.*

Abstract

These notes are mnemonics about what was covered in class. They don't replace being present or reading the book. Reading ahead in the book is very effective.

16.1 Delta method

Also called δ-method and 'propagation of error' in Rice Ch 4.6.

Suppose that $Y = g(X)$ and we know the distribution of X but want the distribution of Y. That can often be done in closed form and cases where it can be done in closed form are very well represented in text books. Other times we cannot easily get a closed form answer but might be satisfied with an approximation. Sometimes we might only know some moments of X and we might want some moments of Y. Then the moments of Y are not at all determined by just a few moments of Y but we may still get a useful approximation.

Similarly, we might know moments of X and Y but want moments of $Z = g(X, Y)$. For instance, we often study ratios $Z = Y/X$ with both numerator and denominator random. Carrying on from one to two to many, we might know moments of $X_1, \dots, X_r$ and want moments of $Z = g(X_1, \dots, X_r)$.

16.2 Approximating the mean

We make a Taylor approximation

$$Y = g(X) \doteq g(\mu_x) + (X - \mu_x)g'(\mu_x) + \frac{1}{2}(X - \mu_x)^2 g''(\mu_x)$$

around $\mu_x = \mathbb{E}(X)$. Using RHS to denote the right hand side above we find

$$\mathbb{E}(\text{RHS}) = g(\mu_x) + \frac{1}{2}\sigma_x^2 g''(\mu_x)$$

where σ_x^2 is the variance of X. The delta method approximation for $\mathbb{E}(Y)$ is

$$\mathbb{E}(Y) \approx g(\mu_x) + \frac{1}{2}\sigma_x^2 g''(\mu_x).$$

If all $g''(x) > 0$ then the second order Taylor approximation lies completely above the first order one which has expectation $g(\mu_x)$. [I sketched that in class.] If only $g''(\mu_x) > 0$ then the second order approximation is

locally above the first order one. These remarks explain why $\mathbb{E}(\text{RHS}) - g(\mu_x)$ has the same sign as $g''(\mu_x)$. Similarly if σ_x is small then there is less room for curvature to make a difference.

It is important to remember that we are getting the ***exact mean*** of an ***approximate g(X)***, namely the RHS. Our error comes from the Taylor approximation. This δ-method is most useful when X fluctuates only a little around μ_x for then the Taylor approximation is most accurate. In applications X might be the average of many observations or some other statistic with small variance. Then the delta method can be extremely accurate.

Going the other way, if $Y = 1/X$ and $X \sim U(0,1)$, then doing the integral would show us that $\mathbb{E}(Y) = \infty$ but the delta method will give a finite answer.

For $Z = g(X,Y)$ we did a second order Taylor approximation of g around (μ_x, μ_y) and found that the expected value of the RHS was

$$g(\mu_x,\mu_y) + \frac{1}{2}\Big[\sigma_x^2\frac{\partial^2}{\partial\mu_x^2}g(\mu_x,\mu_y) + \sigma_y^2\frac{\partial^2}{\partial\mu_y^2}g(\mu_x,\mu_y) + 2\rho\sigma_x\sigma_y\frac{\partial^2}{\partial\mu_x\partial\mu_y}g(\mu_x,\mu_y)\Big] \tag{16.1}$$

where $\rho = \text{Corr}(X,Y)$. In class I used σ_{xy} for $\text{Cov}(X,Y) = \rho\sigma_x\sigma_y$ and you will often see that notation in use.

For $Z = g(X_1,\dots,X_r)$ the same steps give us

$$g(\mu_1,\dots,\mu_r) + \frac{1}{2}\sum_{j=1}^{r}\sum_{k=1}^{r}\sigma_{jk}\frac{\partial^2}{\partial\mu_j\partial\mu_k}g(\mu_1,\dots,\mu_r) \tag{16.2}$$

where $\sigma_{jk} = \text{Cov}(X_j, X_j)$. In the special case $j = k$, we get $\sigma_{jj} = \sigma_j^2$. Equation (16.1) has $2\rho\sigma_x\sigma_y$ which is $2\sigma_{xy}$. We don't see that 2 in (16.2). It is there because $\sigma_{jk} = \sigma_{kj}$, so the corresponding mixed partial derivative comes in twice. We could write

$$g(\mu_1,\dots,\mu_r) + \frac{1}{2}\sum_{j=1}^{r}\sigma_j^2\frac{\partial^2}{\partial\mu_j^2}g(\mu_1,\dots,\mu_r) + \sum_{j=1}^{r-1}\sum_{k=j+1}^{r}\sigma_{jk}\frac{\partial^2}{\partial\mu_j\partial\mu_k}g(\mu_1,\dots,\mu_r) \tag{16.3}$$

which is how we might prefer to compute it by hand. To my way of thinking (16.2) is less likely to lead to mistakes in theory or implementation.

In class we did the worked example (also in Rice) with $Z = Y/X$ and

$$\mathbb{E}(Z) \approx \frac{\mu_y}{\mu_x} + \frac{\sigma_x^2\mu_y - \rho\sigma_x\sigma_y\mu_x}{\mu_x^3}.$$

You really want $\mathbb{E}(X)^3 = \mu_x^3$ in that computation and ***not*** $\mathbb{E}(X^3)$. The reason is that the Taylor approximation is around the point (μ_x,μ_y) and derivatives there are functions of μ_x and μ_y and not functions of X and Y.

Commonly X and Y are both positive random variables. Then we see that positive ρ tends to reduce Z and negative ρ tends to increase it. For the latter case think of Y growing large just when X approaches zero.

16.3 Delta method variance

Returning to our trio: $Y = g(X)$, $Z = g(X,Y)$ and $Z = g(X_1,\dots,X_r)$, we might want to approximate the variance. Once again, what we do is approximate g first by a Taylor approximation and then get the ***exact*** expectation of that approximation to g.

In the first case

$$Y = g(X) \doteq g(\mu_x) + (X - \mu_x)g'(\mu_x)$$

and then

$$\mathrm{Var}(\mathrm{RHS}) = g'(\mu_x)^2\sigma_x^2,$$

so the delta method approximation is

$$\mathrm{Var}(Y) \approx g'(\mu_x)^2\sigma_x^2.$$

Notice that now we are getting the second moment of a first order Taylor approximation while for the mean we got a first moment of a second order Taylor approximation. If X is usually very close to μ_x (as it often is when we use the delta method) then terms $(X-\mu_x)^k$ for $k \geqslant 2$ become negligible so stopping at a first order approximation is appropriate.

There is a corner case: if $g'(\mu_x) = 0$ then the delta method does not give a useful approximation to the variance.

Similarly for $Y = 1/X$ and $X \sim U(0,1)$ the delta method will give a finite variance because the Taylor approximation has a finite variance, even though $\mathbb{E}(Y^2) = \infty$.

For $Z = g(X, Y)$,

$$\mathrm{Var}(Z) \approx \Big[\frac{\partial}{\partial\mu_x}g(\mu_x,\mu_y)\Big]^2\sigma_x^2 + \Big[\frac{\partial}{\partial\mu_y}g(\mu_x,\mu_y)\Big]^2\sigma_y^2 + 2\Big[\frac{\partial}{\partial\mu_x}g(\mu_x,\mu_y)\frac{\partial}{\partial\mu_y}g(\mu_x,\mu_y)\Big]\rho\sigma_x\sigma_y$$

and similarly for $Z = g(X_1, \ldots, X_r)$,

$$\mathbb{E}(Z) \approx \sum_{j=1}^{r}\sum_{k=1}^{r}\frac{\partial}{\partial\mu_j}g(\mu_1,\ldots,\mu_r)\frac{\partial}{\partial\mu_k}g(\mu_1,\ldots,\mu_r)\sigma_{jk}$$

for $\sigma_{jk} = \mathrm{Cov}(X_j, X_j) = \rho_{jk}\sigma_j\sigma_k$.

We worked out that for $Z = Y/X$,

$$\mathrm{Var}(Z) \approx \frac{1}{\mu_x^2}\Big[\frac{\mu_y^2\sigma_x^2}{\mu_x^2} + \sigma_y^2 - 2\rho\mu_y\frac{\sigma_x\sigma_y}{\mu_x}\Big].$$

Note also how positive ρ decreases $\mathrm{Var}(Z)$ (or its approximation), which makes sense because X and Y moving up and down together leaves a more stable ratio Y/X than if they move in opposite directions. We also see that small μ_x gives a large $\mathrm{Var}(Z)$.

16.4 Ratio again

Now suppose that we have independently sampled (X_i, Y_i) pairs for $i = 1, \ldots, n$. Each X_i has correlation ρ with its corresponding Y_i but no correlation with a Y value from a different pair. We let $Z = \bar{Y}/\bar{X}$ and want the delta method approximations to the mean and variance of Z.

This is a good use case. Both $\bar{X}$ and $\bar{Y}$ have variances that decrease as n increases. So if n is large we can reasonably expect the Taylor expansions to be accurate. We write $Z = g(\bar{X}, \bar{Y})$ for the same g we used for Y/X and replace the moments of X and Y by those of $\bar{X}$ and $\bar{Y}$.

The means are μ_x and μ_y. The variances are σ_x^2/n and σ_y^2/n. We also need $\mathrm{Corr}(\bar{X}, \bar{Y})$. A good way to refresh your knowledge of the early chapters is to write out a proof that $\mathrm{Corr}(\bar{X}, \bar{Y}) = \mathrm{Corr}(X_1, Y_1) = \rho$.

We found that

$$\mathbb{E}\Big(\frac{\bar{Y}}{\bar{X}}\Big) \approx \frac{\mu_y}{\mu_x} + \frac{1}{n}\frac{\sigma_x^2\mu_y - \rho\sigma_x\sigma_y\mu_x}{\mu_x^3}.$$

The new ingredient is the $1/n$. Indeed taking $n = 1$ gives us back the formula we had for Y/X (which it should).

Similarly, replacing σ_x^2 by $\sigma_{\bar{x}}^2 = \sigma_x^2/n$ and σ_y^2 by $\sigma_{\bar{y}}^2 = \sigma_y^2/n$ in the formula for $\mathrm{Var}(Y/X)$ gives

$$\mathrm{Var}\Big(\frac{\bar{Y}}{\bar{X}}\Big) \approx \frac{1}{n}\frac{1}{\mu_x^2}\Big[\frac{\mu_y^2\sigma_x^2}{\mu_x^2} + \sigma_y^2 - 2\rho\mu_y\frac{\sigma_x\sigma_y}{\mu_x}\Big]$$

which once again gets us back to the original for $n = 1$.

16.5 Plug-in

A common trope in statistics is to work out a formula for something we want, like a variance, allowing that formula to include some unknown quantity. Then when it comes time to compute it we plug in some good guess for the unknown. That ignores how the guess fluctuates around the unknown but when the guess is very accurate (as it commonly is for large n) then the fluctuations we ignore may be negligible.

Here is how it works for the t statistic in linear regression. We work out that

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - \beta_1}{\sigma\sqrt{n s_{xx}}}.$$

If the responses are normally distributed then $Z \sim N(0,1)$. Otherwise Z may be approximately $N(0,1)$ by the CLT. Unfortunately we usually don't know σ.

Suppose we have an estimte of σ, call it $\hat{\sigma}$ such as s that we used in regression. A consistent estimate will approach the true value σ as $n \to \infty$. Now for any $\epsilon > 0$

$$\Pr(\sigma - \epsilon \leqslant \hat{\sigma} \leqslant \sigma + \epsilon) \to 1, \quad \text{as } n \to \infty.$$

Let ϵ be small enough that replacing σ by some $\hat{\sigma}$ with $|\hat{\sigma} - \sigma| < \epsilon$ makes a negligible difference to Z. For large enough n the difference $\hat{\sigma} - \sigma$ will be negligible with probability nearly one and then the plug-in approximation will be reliable.

A more advanced course would work out the approximation errors described here in much greater detail.

What the t distribution gives us is the exact distribution of that ratio assuming a normal distribution. If the data are not normal (hint: they're not normal) then the t test still works out because of the plug-in principle.

16.6 Bootstrap

Suppose that our data are $X_1, \ldots, X_n$ IID from some F but maybe F isn't one of our favorite parametric families (such as the one in Appendix A of Rice). We compute some statistic $\hat{\theta}$ from the data. We might then want to approximate $\mathrm{Var}(\hat{\theta})$ or the bias $\mathbb{E}(\hat{\theta}) - \theta$ where θ is the quantity being estimated by $\hat{\theta}$ or we might want a confidence interval $\Pr(L \leqslant \theta \leqslant U) \approx 1 - \alpha$. We want these things without having to assume a parametric model. We are willing to spend computer time to get them.

At a high level, we do something to the data getting a statistic $\hat{\theta} = T(X_1, \ldots, X_n)$ for some function T (data in, statistic out).

We can usually write that statistic as a function of the empirical CDF

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leqslant x\}.$$

So $\hat{\theta} = \mathcal{T}(\hat{F})$. We use $\mathcal{T}$ to designate a slightly more complicated function that usual (distribution goes in, number comes out). Now let's think of the true value of θ as $\mathcal{T}(F)$. In other words the true value of our statistic is the exact same function $\mathcal{T}$ to the true distribution F instead of $\hat{F}$.

For example, if $\mathcal{T}(F) = \mathbb{E}(X; X \sim F)$ then the true θ is $\mathbb{E}(X)$ and the estimate is $\bar{X}$. If $\mathcal{T}(F) = \Pr(X > 0; X \sim F)$ then the sample value is the fraction of positive X_i. If $\mathcal{T}(F) = \text{median}(X; X \sim F)$ then $\mathcal{T}(\hat{F})$ is the sample median.

So $\theta = \mathcal{T}(F)$ and $\hat{\theta} = \mathcal{T}(\hat{F})$. It is a plug-in estimate.

Now lets consider the variance of $\hat{\theta}$. It is also a function of F. Let's write it as

$$\mathcal{V}(F) = \text{Var}(\hat{\theta}; X_i \sim F) = \text{Var}(\mathcal{T}(\hat{F}); X_i \sim F).$$

That would be a daunting n-fold integral for continuous F and a daunting n-fold sum for discrete F. The bootstrap idea is to plug in $\hat{F}$ and use $\mathcal{V}(\hat{F})$. Imagine sampling n observations from $\hat{F}$. We call them $X_1^*, \ldots, X_n^*$ so that we keep them separate from $X_1, \ldots, X_n$. Let $\hat{F}^*$ be their empirical CDF. Then the boostrap plug-in idea $\mathcal{V}(F) \approx \mathcal{V}(\hat{F})$ translates to

$$\text{Var}(\mathcal{T}(\hat{F}); X_i \sim F) \approx \text{Var}(\mathcal{T}(\hat{F}^*); X_i^* \sim \hat{F}).$$

We cannot ordinarily get a convenient formula for the RHS above. We can however simulate sampling from $\hat{F}$ and that proves to be good enough. To sample $X^* \sim \hat{F}$, set $X^* = X_i$ with probability $1/n$ for $i = 1, \ldots, n$. The CDF of X^* is then

$$\Pr(X^* \leqslant x) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leqslant x\} = \hat{F}(x).$$

So X^* really has distribution $\hat{F}$. We can get a whole sample $X_1^*, \ldots, X_n^*$ from $\hat{F}$ by sampling X_i^* from $X_1, \ldots, X_n$ independently, n times. This is called a ***bootstrap sample***. From it we can compute $\hat{\theta}^* = T(X_1^*, \ldots, X_n^*) = \mathcal{T}(\hat{F}^*)$.

We repeat this a large number B of times independently getting $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$. Now we get the estimate

$$\text{Var}(\hat{\theta}) = \text{Var}(\mathcal{T}(\hat{F}); X_i \sim F) \approx \text{Var}(\mathcal{T}(\hat{F}^*); X_i^* \sim \hat{F}) \approx \frac{1}{B}\sum_{b=1}^{B}\big(\hat{\theta}^{*b} - \bar{\hat{\theta}^*}\big)^2 \tag{16.4}$$

where $\bar{\hat{\theta}^*} = (1/B)\sum_{b=1}^{B}\hat{\theta}^{*b}$. We can of course use $1/(B-1)$ instead of $1/B$ in the variance estimate.

The first approximation in (16.4) is a plug-in approximation based on $\hat{F} \approx F$ (for large n) and then $\mathcal{V}(\hat{F}) \approx \mathcal{V}(F)$. The second approximation is just plain law of large numbers for the randomly sampled variance estimate (for large B). We can usually make B extremely large if we want.

For $b = 1, \ldots, B$
  For $i = 1, \ldots, n$
    $i^* \sim U\{1, 2, \ldots, n\}$
    $X_i^{*b} = X_{i^*}$
  $\hat{\theta}^{*b} = T(X_1^{*b}, \ldots, X_n^{*b})$
$\widehat{\mathrm{Var}}(\hat{\theta}) = (1/B)\sum_{b=1}^{B}\big(\hat{\theta}^{*b} - \bar{\hat{\theta}^*}\big)^2$

Above is some pseudocode to compute the bootstrap variance. You would probably just use X_i^* over and over instead of keeping all $n \times B$ resampled data values. Notice that the key step is sampling i uniformly from integer values 1 through n and using that to index into the sample.

16.7 Bootstrap or delta?

The bootstrap is doing for you by computer what the delta method does by calculus. So it saves you all that calculus and it automatically adapts to your problem if you change your mind about what statistic you're interested in. While it takes away that mathematical chore, figuring out when and why it works is another much harder mathematical task (beyond this course). It frees you up to consider statistics that are more complicated than the ones we can handle in closed form. It also gets us away from having to assume that the data come from one of our favorite distributional families.

We do lose something. If we bootstrapped $\bar{Y}/\bar{X}$ we might not notice the role of ρ or the issue about μ_x being small. The delta method gave us some interpretable insights.

You could quibble that your answer is now random based on the bootstrap sampling and somebody else bootstrapping the very same data values would get a different answer. This is true but for super large B the differences are unlikely to be consequential.

Sometimes having to do B simulations is a problem. That might happen if your B simulations are the inner loop of something else. People can and do bootstrap the bootstrap (but not that often).

16.8 Brad Efron

The bootstrap was invented by Stanford's Brad Efron in 1977. In 2018 he was awarded the "International Prize in Statistics" in recognition of the bootstrap. Here is a quote from him at that time:

> I remember when going into statistics that first year I thought, 'This will be pretty easy, I've dealt with math and that's supposed to be hard.' But statistics was much harder for me at the beginning than any other field. It took years before I felt really comfortable.