Stat210B: Theoretical Statistics **Lecture Date: May 1, 2007**

Functional Delta Method and Bootstrap

Lecturer: Michael I. Jordan *Scribe: Arash Ali Amini*

1 Functional Delta Method

Example 1 (quantile function continued)**.** Recall the definition of the quantile function $F^{-1}(p) = \inf\{x : F(x) \geq p\}$. Let's assume for simplicity that $p = FF^{-1}(p)$. To obtain the influence function we need to differentiate implicitly using this relation. We have done this in the previous lecture and obtained

$$\mathrm{IF}_F(x) = -\frac{1_{[x,+\infty)}\big(F^{-1}(p)\big) - p}{f\big(F^{-1}(p)\big)} \tag{1}$$

where $f(\cdot)$ is the density associated with F. In deriving (1), we have also assumed differentiability of F and positivity of f at the quantile. The graph of the influence function is given in Figure 1. It is easy to show

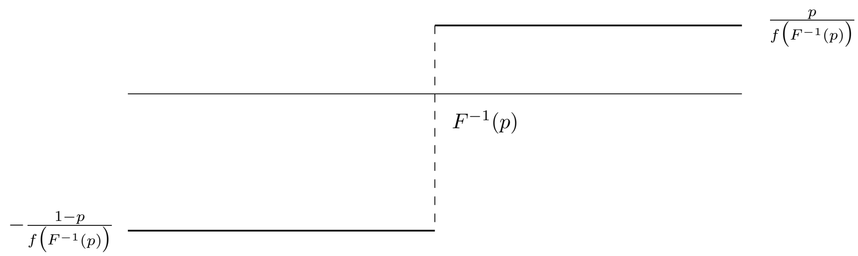

Figure 1: Influence function of the pth quantile

that $E\big(\mathrm{IF}_F(X)\big) = 0$. Let's compute the variance:

$$\begin{aligned}
\gamma^2(F) = \mathrm{Var}\big(\mathrm{IF}_F(X)\big) &= \int \left(\frac{1_{[x,+\infty)}\big(F^{-1}(p)\big) - p}{f\big(F^{-1}(p)\big)} \right)^2 dF(x) \\
&= \frac{1}{f^2\big(F^{-1}(p)\big)} \int \left[1_{[x,+\infty)}\big(F^{-1}(p)\big) - 2p \cdot 1_{[x,+\infty)}\big(F^{-1}(p)\big) + p^2 \right] dF(x) \\
&= \frac{1}{f^2\big(F^{-1}(p)\big)} \left[F\big(F^{-1}(p)\big) - 2p\, F\big(F^{-1}(p)\big) + p^2 \right]
\end{aligned}$$

where the last equality follows because $1_{[x,+\infty)}(y) = 1\{x \leq y\} = 1_{(-\infty,y]}(x)$. Reusing our assumption $p = FF^{-1}(p)$, we get

$$\gamma^2(F) = \frac{p\,(1-p)}{f^2\big(F^{-1}(p)\big)}$$

which (correctly) suggests that

$$\sqrt{n}\Big(\mathbb{F}_n{}^{-1}(p) - F^{-1}(p)\Big) \xrightarrow{d} N\left(0, \frac{p\,(1-p)}{f^2\big(F^{-1}(p)\big)}\right). \tag{2}$$

Let's now derive this result rigorously using Hadamard-differentiability. We will need the following lemma which we state without proof:

Lemma 2. *(van der Vaart, 1998, Lemma 21.3, p. 306) Let F be differentiable at a point $\xi_p \in (a,b)$ such that $F(\xi_p) = p$ with $F'(\xi_p) > 0$. Then $\phi(F) = F^{-1}(p)$ is Hadamard-differentiable at F tangentially to the set of functions h that are continuous at ξ_p, with derivative*

$$\phi'_F(h) = -\frac{h(\xi_p)}{F'(\xi_p)}.$$

Using a variation of the functional delta method, namely the second part of Thm 20.8 form van der Vaart (1998, p. 297), we conclude that $\sqrt{n}\Big(\mathbb{F}_n{}^{-1}(p) - F^{-1}(p)\Big)$ is asymptotically equivalent to ϕ'_F evaluated at $\sqrt{n}\big(\mathbb{F}_n - F\big)$. Using the lemma, this means

$$\sqrt{n}\Big(\mathbb{F}_n{}^{-1}(p) - F^{-1}(p)\Big) = \frac{-\sqrt{n}\big(\mathbb{F}_n - F\big)(\xi_p)}{f\big(F^{-1}(p)\big)} + o_p(1).$$

Expanding $\mathbb{F}_n$ and rearranging, we get

$$\begin{aligned}\sqrt{n}\Big(\mathbb{F}_n{}^{-1}(p) - F^{-1}(p)\Big) &= \frac{-\sqrt{n}\big(\frac{1}{n}\sum_i 1_{\{X_i \le F^{-1}(p)\}} - p\big)}{f\big(F^{-1}(p)\big)} + o_p(1) \\ &= -\frac{1}{\sqrt{n}}\sum_i\left(\frac{1_{\{X_i \le F^{-1}(p)\}} - p}{f\big(F^{-1}(p)\big)}\right) + o_p(1).\end{aligned}$$

Note that the influence function appears again, i.e. we have got the expansion $(1/\sqrt{n})\sum_i \mathrm{IF}_F(X_i)$. It only remains to apply CLT (and Slutsky's lemma) to get (2).

2 Bootstrap

Bootstrap is a plug-in methodology, introduced by Brad Efron, for estimating performance measures associated with statistics. The bootstrap estimate (of a performance measure) is obtained by replacing every occurrence of the true (unknown) distribution F with the empirical distribution $\mathbb{F}_n$ (e.g. $E_F(\cdot)$ is replaced by $E_{\mathbb{F}_n}(\cdot)$ and $\phi(F)$ is replaced by $\phi(\mathbb{F}_n)$) and by replacing the original sample $\{X_i\}_{i=1}^n$ with bootstrap sample $\{X_i^*\}_{i=1}^n$ obtained by resampling (with replacement) from $\mathbb{F}_n$. In practice, computing expectations w.r.t to $\mathbb{F}_n$ is difficult (if not impossible). Instead, the bootstrap estimate is usually obtained by computer simulation, i.e. by generating multiple bootstrap samples and approximating $E_{\mathbb{F}_n}(\cdot)$ by the (frequentist's) average $(1/B)\sum_{b=1}^B \big(\,\cdot\,\big)$, where B is the number of bootstrap samples. For example, a bootstrap estimate of the variance of the median could be

$$\frac{1}{B}\sum_{b=1}^{B}\left(\hat\theta_b^* - \frac{1}{B}\sum_{i=1}^{B}\hat\theta_b^*\right)^2$$

where $\hat\theta_b^*$ is the sample meadian of the b-th bootstrap sample.

We will demonstrate the idea by several examples. First, let's consider a toy example to show that the bootstrap estimate may be computed only based on the knowledge of the original sample, without any resampling.

Example 3. Let $\phi(F) = E_F(X)$ and $n = 2$. Also, let $X_{(1)} = c < X_{(2)} = d$ be the order statstics of the original sample. Then, the bootstrap sample (X_1^*, X_2^*) can take on one of the following values: (c, c), (c, d), (d, c), (d, d), each with probability $1/4$. It follows that $\hat{\theta}_n^*$ (the bootstrap sample mean) takes on values

$$c,\ \frac{1}{2}(c + d),\ d \quad \text{w.p.} \quad \frac{1}{4}, \frac{1}{2}, \frac{1}{4}.$$

Thus, performance measures (e.g. variance, bias, etc.) can be computed and no sampling is needed.

Example 4 (bias of the median)**.** Let $\theta = \phi(F) =$ median of F. Define the performance measure $\lambda_n(F) = E(\hat{\theta}_n) - \theta$ where $\hat{\theta}_n$ is the sample median. Bootstrap replaces F with $\mathbb{F}_n$, which gives the following estimate of the bias:

$$\lambda_n(\mathbb{F}_n) = E_{\mathbb{F}_n}(\hat{\theta}^*) - \hat{\theta}_n$$

where $\hat{\theta}^*$ is the sample median of the bootstrap sample. Also note that we have replaced $\theta = \phi(F)$ with $\hat{\theta}_n = \phi(\mathbb{F}_n)$.

Now let's consider, e.g., the case $n = 3$. Let $X_{(1)} = b$, $X_{(2)} = c$ and $X_{(3)} = d$. Then (X_1^*, X_2^*, X_3^*) can take on 27 different values. The resulting distribution on $(X_{(1)}^*, X_{(2)}^*, X_{(3)}^*)$ is given by

bbb	bbc	bbd	bcc	bcd	$\cdots$	ddd
$\frac{1}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{6}{27}$	$\cdots$	$\frac{1}{27}$

One concludes that $X_{(2)}^*$ is distributed according to

$$b,\ c,\ d \quad \text{w.p.} \quad \frac{7}{27}, \frac{13}{27}, \frac{7}{27}$$

which gives the following formula for bootstrap estimate of the bias

$$\begin{aligned}
\lambda_n(\mathbb{F}_n) &= E_{\mathbb{F}_n}(X_{(2)}^*) - X_{(2)} \\
&= \left(\frac{7}{27} X_{(1)} + \frac{13}{27} X_{(2)} + \frac{7}{27} X_{(3)} \right) - X_{(2)} \\
&= \frac{14}{27} \left(\frac{X_{(1)} + X_{(3)}}{2} - X_{(2)} \right).
\end{aligned}$$

Again note that we got the bias estimate in terms of the original sample without any resampling. This hopefully illustrates the idea of bootstrap. In general, bootstrap may be summarized as 'plug-in + Monte Carlo integration', but the second step is not always required .

Now, we show by an example that bootstrap does not always work, i.e. our estimate of the bias may not converge to the true value of the bias. In other words, we may not always get consistency.

Example 5 (U-statistic)**.** In this example, $\lambda_n(\mathbb{F}_n)$ is and is not consistent depending on the assumptions. Let $\lambda_n(F) = E\big(\sqrt{n}\,\hat{\theta}_n\big)^2$, where $\hat{\theta}_n$ is a U-statistic:

$$\hat{\theta}_n = \frac{1}{n(n-1)} \sum_{i \neq j} \Psi(X_i, X_j).$$

We have calculated moments of U-statistics before (cf. the final part of Thm 12.3 of van der Vaart (1998)):

$$\begin{aligned}\lambda_n(F) &= \frac{4(n-2)}{n-1}\,\gamma_1^2 + \frac{2}{n-1}\,\gamma_2^2,\\ \gamma_1^2 &= E\Psi(X_1, X_2)\Psi(X_1, X_3),\\ \gamma_2^2 &= E\Psi^2(X_1, X_2).\end{aligned}$$

It follows that $\lambda_n(F) \to \lambda(F) = 4\gamma_1^2$. We also obtain

$$\begin{aligned}\lambda_n(\mathbb{F}_n) &= \frac{4(n-2)}{n-1}\,\gamma_1^{*2} + \frac{2}{n-1}\,\gamma_2^{*2},\\ \gamma_1^{*2} &= \frac{1}{n^3}\sum_i\sum_j\sum_k \Psi(X_i, X_j)\Psi(X_i, X_k),\\ \gamma_2^{*2} &= \frac{1}{n^2}\sum_i\sum_j \Psi^2(X_i, X_j).\end{aligned}$$

Note that $\hat{\theta}_n$ depends on $\Psi(X_i, X_j)$ for $i \neq j$, but $\lambda_n(\mathbb{F}_n)$ depends also on $\Psi(X_i, X_i)$. The situation thus depends on the diagonal of the kernel:

- We obtain consistency when γ_1^2, γ_2^2 and γ_3^2 are all finite, where

  $$\gamma_3^2 = E\Psi^2(X_i, X_i).$$

  Indeed, because of finiteness assumption, we can use the law of large numbers[1] (LLN) to conclude $\gamma_1^{*2} \to \gamma_1^2$ and $\gamma_2^{*2} \to \gamma_2^2$, which in turn imply $\lambda_n(\mathbb{F}_n) \to 4\lambda_1^2$.

- But what if $\gamma_3^2 = \infty$? Note that

  $$\gamma_2^{*2} = \frac{1}{n^2}\sum_{i\neq j} \Psi(X_i, X_j)^2 + \frac{1}{n^2}\sum_i \Psi^2(X_i, X_i)$$

  The first term converges in probability to γ_2^2. But, the second term may diverge or converge. An example of a setting in which the second term diverges is:

  $$\begin{aligned}&X_i \overset{\text{iid}}{\sim} \text{Un}(0,1),\\ &|\Psi(x,y)| < M, \quad \forall x \neq y,\\ &\Psi(x,x) = e^{1/x}.\end{aligned}$$

References

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

[1] We are using both the usual LLN and also the LLN for U-statistics.