Hájek Projection and U-Statistics

Bryan S. Graham, UC - Berkeley & NBER

September 29, 2020

Hájek projection

One important method of deriving the asymptotic distribution of a sequence of statistics, say U_N, is show the sequence is asymptotically equivalent to a second sequence, say U_N^*, the large sample properties of which are well-understood (cf., van der Vaart, 1998, Chapter 11).

The asymptotic properties of sums of independent random variables, appropriately scaled, are especially well-understood (since Central Limit Theorems (CLTs) are generally applicable to them). With this observation in mind, let $X_1, X_2, \ldots, X_N$ be independent $K \times 1$ random vectors. Let $\mathcal{L}$ be the linear subspace containing of all functions of the form

$$\sum_{i=1}^{N} g_i(X_i) \tag{1}$$

for $g_i : \mathbb{R}^K \to \mathbb{R}$ arbitrary with $\mathbb{E}\left[g_i(X_i)^2\right] < \infty$ for $i = 1, \ldots, N$.

Next let Y be an arbitrary random variable with finite variance, but unknown distribution. We can use the Projection Theorem to approximate the statistic Y with one composed of a sum of independent random functions. Such a sum, by appeal to a CLT, may be well-described by a normal distribution. If the projection is also a very good approximation of Y, then the hope is that Y may be well-described by a normal distribution as well.

The projection of Y onto $\mathcal{L}$, called the **Hájek Projection**, equals

$$\Pi(Y|\mathcal{L}) = \sum_{i=1}^{N} \mathbb{E}[Y|X_i] - (N-1)\,\mathbb{E}[Y]. \tag{2}$$

To verify (2) it suffices to check the necessary and sufficient orthogonality condition of the

Projection Theorem. Before doing so, it is helpful to observe that, for $j \neq i$,

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[Y\middle|X_i\right]\middle|X_j\right] &= \mathbb{E}\left[\mathbb{E}\left[Y\middle|X_i\right]\right] \\ &= \mathbb{E}\left[Y\right], \end{aligned} \tag{3}$$

due to independence of X_i and X_j and the law of iterated expectations. In contrast, if $j = i$, then

$$\mathbb{E}\left[\mathbb{E}\left[Y\middle|X_i\right]\middle|X_i\right] = \mathbb{E}\left[Y\middle|X_i\right]. \tag{4}$$

The orthogonality condition to verify, for $U = Y - \Pi\left(Y\middle|\mathcal{L}\right)$, is

$$\begin{aligned} 0 =& \mathbb{E}\left[U\left(\sum_{j=1}^{N} g_j\left(X_j\right)\right)\right] \\ =& \sum_{j=1}^{N} \mathbb{E}\left[U g_j\left(X_j\right)\right] \\ =& \sum_{j=1}^{N} \mathbb{E}\left[\mathbb{E}\left[U\middle|X_j\right] g_j\left(X_j\right)\right] \end{aligned}$$

Next observe that, using (3) and (4),

$$\begin{aligned}\mathbb{E}\left[U\middle|X_j\right] &= \mathbb{E}\left[Y\middle|X_j\right] - \sum_{i=1}^{N} \mathbb{E}\left[\mathbb{E}\left[Y\middle|X_i\right]\middle|X_j\right] + \left(N-1\right)\mathbb{E}\left[Y\right] \\ &= \mathbb{E}\left[Y\middle|X_j\right] - \mathbb{E}\left[Y\middle|X_j\right] - \left(N-1\right)\mathbb{E}\left[Y\right] + \left(N-1\right)\mathbb{E}\left[Y\right] \\ &= 0, \end{aligned}$$

for $j = 1, \ldots, N$. Hence the require orthogonality condition does hold when $\Pi\left(Y\middle|\mathcal{L}\right)$ is of the form asserted in (2).

If, in addition to independence, we have that (i) $\{X_i\}_{i=1}^N$ are identically distributed and (ii) $Y = h\left(X_1, \ldots, X_N\right)$ is a permutation symmetric function of $\{X_i\}_{i=1}^N$, then

$$\begin{aligned}\mathbb{E}\left[Y\middle|X_i = x\right] &= \mathbb{E}\left[Y\middle|X_1 = x\right] \\ &= \mathbb{E}\left[h\left(x, X_2 \ldots, X_N\right)\right] \\ &\overset{def}{\equiv} \bar{h}_1\left(x\right) \end{aligned}$$

for all $i = 1, \ldots, N$. Since $h_1\left(x\right)$ does not depend on i it follows that (2) simplifies, in this

case, to

$$\Pi\left(Y\middle|\mathcal{L}\right)=\sum_{i=1}^{N}\bar{h}_1\left(X_i\right)-\left(N-1\right)\mathbb{E}\left[Y\right]. \tag{5}$$

Projection (5) is important, as we shall see later, for the analysis of U-Statistics.

A fact that will be helpful for what follows is that average of Y and its projection coincide:

$$\begin{aligned}\mathbb{E}\left[\Pi\left(Y\middle|\mathcal{L}\right)\right]=&\sum_{i=1}^{N}\mathbb{E}\left[\mathbb{E}\left[Y\middle|X_i\right]\right]-\left(N-1\right)\mathbb{E}\left[Y\right]\\=&N\mathbb{E}\left[Y\right]-\left(N-1\right)\mathbb{E}\left[Y\right]\\=&\mathbb{E}\left[Y\right]\end{aligned}$$

Note also that $\Pi\left(Y\middle|1\right)=\mathbb{E}\left[Y\right]$.

Hájek projection and large sample theory

Next let $\{Y_N\}$ be a sequence of statistics indexed by the sample size and $\mathcal{L}_N$ a corresponding sequence of linear subspaces of form (1). The goal is to use the limiting distribution of $\sqrt{N}\left(\Pi\left(Y_N\middle|\mathcal{L}_N\right)-\Pi\left(Y_N\middle|1\right)\right)$ to approximate that of $\sqrt{N}\left(Y_N-\Pi\left(Y_N\middle|1\right)\right)$. Such an approach will be (asymptotically) valid if these two statistics converge in mean square (to one another). It is attractive because in the main cases of interest the asymptotic sampling distribution of $\sqrt{N}\left(\Pi\left(Y_N\middle|\mathcal{L}_N\right)-\Pi\left(Y_N\middle|1\right)\right)$ is straightforward to derive, whereas that of $\sqrt{N}\left(Y_N-\Pi\left(Y_N\middle|1\right)\right)$ may be ex ante non-obvious.

The "Analysis of Variance" decomposition for projections gives, after some re-arrangement

$$N\left\|Y_N-\Pi\left(Y_N\middle|\mathcal{L}_N\right)\right\|^2=N\left\|\Pi\left(Y_N\middle|\mathcal{L}_N\right)-\Pi\left(Y_N\middle|1\right)\right\|^2-N\left\|Y_N-\Pi\left(Y_N\middle|1\right)\right\|^2.$$

Or, invoking the covariance inner product, that $\Pi\left(Y\middle|1\right)=\mathbb{E}\left[Y\right]$, as well as the definition of variance

$$N\mathbb{E}\left[\left(Y_N-\Pi\left(Y_N\middle|\mathcal{L}_N\right)\right)^2\right]=N\mathbb{V}\left(Y_N\right)-N\mathbb{V}\left(\Pi\left(Y_N\middle|\mathcal{L}_N\right)\right). \tag{6}$$

Hence if the limits of $N\mathbb{V}\left(Y_N\right)$ and $N\mathbb{V}\left(\Pi\left(Y_N\middle|\mathcal{L}_N\right)\right)$ coincides as $N\to 0$ we have that $\sqrt{N}\left(Y_N-\Pi\left(Y_N\middle|\mathcal{L}_N\right)\right)$ converges in mean square to zero. This means that $\sqrt{N}Y_N$ and $\sqrt{N}\Pi\left(Y_N\middle|\mathcal{L}_N\right)$ will have identical limit distributions.

In the next section we apply the above ideas to study the asymptotic properties of U-Statistics.

U-Statistics

Let $\{X_i\}_{i=1}^N$ be a simple random sample from some population of interest. Let $h(X_{i_1}, \ldots, X_{i_m})$ be a symmetric *kernel* function. The assumption of symmetry is without loss of generality since we can always replace $h(X_{i_1}, \ldots, X_{i_m})$ with its average across permutations. A U-statistic is an average of the kernel $h(X_{i_1}, \ldots, X_{i_m})$ over all possible m-tuples of observations in the sample.

$$U_N = \binom{N}{m}^{-1} \sum_{\mathbf{i} \in C_{m,N}} h(X_{i_1}, \ldots, X_{i_m})$$

where $C_{m,N}$ denotes the set of all unique combinations of indices of size m drawn from the set $\{1, 2, \ldots, N\}$.

The parameter of interest is

$$\theta = \mathbb{E}[U_N] = \mathbb{E}[h(X_1, \ldots, X_m)],$$

where the expectation is over m independent random draws from the target population.

Variance of U_N

For $s = 1, \ldots, m$ let

$$\bar{h}_s(x_1, \ldots, x_s) = \mathbb{E}[h(x_1, \ldots, x_s, X_{s+1}, \ldots, X_m)]$$

be the average over the last $m - s$ elements of $h(\cdot)$ holding the first s elements fixed. Note that since X_{i_k} is independent of X_{i_l} for all $k \neq l$ we have

$$\mathbb{E}[h(X_1, \ldots, X_s, X_{s+1}, \ldots, X_m) | (X_1, \ldots, X_s) = (x_1, \ldots, x_s)] = \mathbb{E}[h(x_1, \ldots, x_s, X_{s+1}, \ldots, X_m)].$$

It is also useful to observe that

$$\mathbb{E}\left[\bar{h}_s(X_1, \ldots, X_s)\right] = \mathbb{E}[h(X_1, \ldots, X_m)] = \theta.$$

The variance of U_N has a special structure. Define, for $s = 1, \ldots, m$,

$$\delta_s^2 = \mathbb{V}\left(\bar{h}_s(X_1, \ldots, X_s)\right).$$

Applying the variance operator to U_N yields

$$\begin{aligned}\mathbb{V}\left(U_N\right) &= \mathbb{V}\left(\binom{N}{m}^{-1}\sum_{\mathbf{i}\in C_{m,N}} h\left(X_{i_1},\ldots,X_{i_m}\right)\right)\\ &= \binom{N}{m}^{-2}\sum_{\mathbf{i}\in C_{m,N}}\sum_{\mathbf{j}\in C_{m,N}}\mathbb{C}\left(h\left(X_{i_1},\ldots,X_{i_m}\right),h\left(X_{j_1},\ldots,X_{j_m}\right)\right). \qquad (7)\end{aligned}$$

The form of the covariances in (7) depends on the number of indices in common. Let s be the number of indices in common in $X_{i_1},\ldots,X_{i_m}$ and $X_{j_1},\ldots,X_{j_m}$:

$$\begin{aligned}\mathbb{C}\left(h\left(X_{i_1},\ldots,X_{i_m}\right),h\left(X_{j_1},\ldots,X_{j_m}\right)\right) &= \mathbb{E}\left[\left(h\left(X_1,\ldots,X_s,X_{s+1},\ldots,X_m\right)-\theta\right)\right.\\ &\quad\left.\times\left(h\left(X_1,\ldots,X_s,X'_{s+1},\ldots,X'_m\right)-\theta\right)\right] \qquad (8)\end{aligned}$$

Conditional on $X_1,\ldots,X_s$ the two terms in (8) are independent so that, using the Law of Iterated Expectations,

$$\begin{aligned}\mathbb{C}\left(h\left(X_{i_1},\ldots,X_{i_m}\right),h\left(X_{j_1},\ldots,X_{j_m}\right)\right) =& \mathbb{E}\left[\left(\bar{h}_s\left(X_1,\ldots,X_s\right)-\theta\right)\left(\bar{h}_s\left(X_1,\ldots,X_s\right)-\theta\right)\right]\\ =& \delta_s^2.\end{aligned}$$

Using the same argument yields

$$\mathbb{C}\left(\bar{h}_s\left(X_1,\ldots,X_s\right),h\left(X_1,\ldots,X_m\right)\right)=\delta_s^2.$$

By the Cauchy-Schwartz Inequality we have

$$\frac{\mathbb{C}\left(\bar{h}_s\left(X_1,\ldots,X_s\right),h\left(X_1,\ldots,X_m\right)\right)}{\delta_s\delta_m}\leq 1$$

and hence

$$\delta_s^2\leq\delta_m^2.$$

Continuing with this type of reasoning we get the weak ordering

$$\delta_1^2\leq\delta_2^2\leq\ldots\leq\delta_m^2.$$

In what follows we will assume that $\delta_m^2<\infty$.

To use these results to get an expression for $\mathbb{V}\left(U_N\right)$ begin by observing that the number of

pairs of m-tuples $(i_1, \ldots, i_m)$ and $(j_1, \ldots, j_m)$ having exactly s elements in common is

$$\binom{N}{m}\binom{m}{s}\binom{N-m}{m-s}.$$

This follows since $\binom{N}{m}$ equals the number of ways of choosing $(i_1, \ldots, i_m)$ from the set $\{1, \ldots, N\}$. For each unique m-tuple there are $\binom{m}{s}$ ways of choosing a subset of size s from it. Having fixed the s indices in common there are then $\binom{N-m}{m-s}$ ways of choosing the $m-s$ non-common elements of $(j_1, \ldots, j_m)$ from the $N-m$ integers not already present in $(i_1, \ldots, i_m)$.

We therefore have

$$\begin{aligned}
\mathbb{V}(U_N) &= \binom{N}{m}^{-2}\sum_{s=0}^{m}\binom{N}{m}\binom{m}{s}\binom{N-m}{m-s}\delta_s^2 \\
&= \binom{N}{m}^{-1}\sum_{s=1}^{m}\binom{m}{s}\binom{N-m}{m-s}\delta_s^2. \\
&= \sum_{s=1}^{m}\frac{m!^2}{s!\,(m-s)!^2}\frac{(N-m)\,(N-m-1)\cdots(N-2m+s+1)}{N\,(N-1)\cdots(N-m+1)}\delta_s^2. \qquad (9)
\end{aligned}$$

To understand this expression note that each of the covariances in (7) above have $s = 0, \ldots, m$ elements in common. The coefficients on the δ_s^2 in (9) give the number of covariances with s elements in common. Also note that $\delta_0^2 = 0$.

The coefficient on δ_1^2 is

$$\begin{aligned}
\frac{m!^2}{1!\,(m-1)!^2}\frac{(N-m)\,(N-m-1)\cdots(N-2m+1+1)}{N\,(N-1)\cdots(N-m+1)} &= m^2\frac{\overbrace{(N-m)\,(N-m-1)\cdots(N-2m+2)}^{\text{m-1 terms}}}{\underbrace{N\,(N-1)\cdots(N-m+1)}_{\text{mterms}}} \\
&\simeq \frac{m^2}{N}.
\end{aligned}$$

The coefficient on δ_2^2 is $O\left(N^{-2}\right)$ etc. We therefore have

$$\mathbb{V}(U_N) = \frac{m^2}{N}\delta_1^2 + O\left(N^{-2}\right)$$

and also that $\mathbb{V}\left(\sqrt{N}\,(U_N - \theta)\right) \to m^2\delta_1^2$ as $N \to \infty$.

If $\delta_1 = 0$ we say that U_N is a degenerate U-Statistic with degeneracy of order 1. I will not consider the properties of degenerate U-Statistics here.

First projection of U_N

The arguments outlined so far provide expressions for the mean and variance of U_N. To conduct inference we need an asymptotic normality result. While U_N is constructed from independently and identically distributed random variables, not all elements of its summand are independent of one another. We cannot apply a standard central limit theorem (CLT). To show asymptotic normality of $\sqrt{N}U_N$ we will therefore proceed in the way used to motivate our introduction of the Hájek projection earlier. Let $\mathcal{L}$ the linear subspace containing of all functions of the form

$$\sum_{i=1}^{N} g_i\left(X_i\right) \tag{10}$$

for $g_i : \mathbb{R}^K \to \mathbb{R}$ arbitrary with $\mathbb{E}\left[g_i\left(X_i\right)^2\right] < \infty$ for $i = 1, \ldots, N$. The Hájek Projection of U_N onto $\mathcal{L}$ equals, from (2) above,

$$\Pi\left(U_N \middle| \mathcal{L}_N\right) = \sum_{i=1}^{N} \mathbb{E}\left[U_N \middle| X_i\right] - (N-1)\,\mathbb{E}\left[U_N\right]. \tag{11}$$

To simplify the argument assume that $m = 2$. The L^2 projection of U_N onto just the first observation X_1 is

$$\begin{aligned}
\mathbb{E}\left[U_N \middle| X_1\right] &= \tbinom{N}{2}^{-1} \sum_{i=1}^{N} \sum_{i<j} \mathbb{E}\left[h\left(X_i, X_j\right) \middle| X_1\right] \\
&= \tbinom{N}{2}^{-1} (N-1)\,\bar{h}_1\left(X_1\right) + \tbinom{N}{2}^{-1} \left(\tbinom{N}{2} - (N-1)\right)\theta \\
&= \frac{2}{N}\left\{\bar{h}_1\left(X_1\right) - \theta\right\} + \theta.
\end{aligned} \tag{12}$$

The second equality follows because $\mathbb{E}\left[h\left(X_i, X_j\right) \middle| X_1\right] = \bar{h}_1\left(X_1\right)$ if either i or j equals 1 (which occurs $N-1$ times). In all other cases, by random sampling, $\mathbb{E}\left[h\left(X_i, X_j\right) \middle| X_1\right] = \mathbb{E}\left[h\left(X_i, X_j\right)\right] = \theta$ (which occurs $\binom{N}{2} - (N-1)$ times). Substituting (12) into (11) yields

$$\Pi\left(U_N - \theta \middle| \mathcal{L}_N\right) = \frac{2}{N} \sum_{i=1}^{N} \left\{\bar{h}_1\left(X_1\right) - \theta\right\}.$$

For the general $m \geq 2$ case a similar calculation gives

$$\Pi\left(U_N - \theta \middle| \mathcal{L}_N\right) = \frac{m}{N} \sum_{i=1}^{N} \left\{\bar{h}_1\left(X_1\right) - \theta\right\}.$$

Since $\Pi\left(U_N-\theta|\,\mathcal{L}_N\right)$ is a sum of i.i.d. random variables with $\mathbb{V}\left(\bar{h}_1\left(X_1\right)-\theta\right)=\delta_1^2$, a CLT gives

$$\sqrt{N}\Pi\left(U_N-\theta|\,\mathcal{L}_N\right)\xrightarrow{D}\mathcal{N}\left(0,m^2\delta_1^2\right).$$

Our (combinatoric) variance calculations gave

$$\mathbb{V}\left(\sqrt{N}\left(U_N-\theta\right)\right)\rightarrow m^2\delta_1^2$$

as $N\rightarrow\infty$. Therefore $N\mathbb{V}\left(U_N\right)-N\mathbb{V}\left(\Pi\left(U_N|\,\mathcal{L}_N\right)\right)\rightarrow 0$ as $N\rightarrow\infty$, in turn implying that $\sqrt{N}\left(U_N-\theta\right)$ convergences in mean square to $\sqrt{N}\Pi\left(U_N-\theta|\,\mathcal{L}_N\right)$ and hence that

$$\sqrt{N}\left(U_N-\theta\right)\xrightarrow{D}\mathcal{N}\left(0,m^2\delta_1^2\right)$$

as needed.

Bibliographic notes

Hoeffding (1948) developed the basic theory of U-Statistics. van der Vaart (1998, Chapters 11 & 12) is a standard, and highly recommended, textbook reference. The exposition in these notes draws, in part, of unpublished lectures notes by Ferguson (2005).

References

Ferguson, T. S. (2005). U-statistics. University of California - Los Angeles.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19(3), 293 – 325.

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

