Theoretical Statistics. Lecture 5.

Peter Bartlett

1. U-statistics.

Outline of today's lecture

We'll look at U-statistics, a family of estimators that includes many interesting examples. We'll study their properties: unbiased, lower variance, concentration (via an application of the bounded differences inequality), asymptotic variance, asymptotic distribution. (See Chapter 12 of van der Vaart.)

First, we'll consider the standard unbiased estimate of variance—a special case of a U-statistic.

Variance estimates

$$
\begin{aligned}
s_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \\
&= \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\left((X_i - \bar{X}_n)^2 + (X_j - \bar{X}_n)^2\right) \\
&= \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\left((X_i - \bar{X}_n) - (X_j - \bar{X}_n)\right)^2 \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}\left(X_i - X_j\right)^2 \\
&= \frac{1}{\binom{n}{2}}\sum_{i<j}\frac{1}{2}\left(X_i - X_j\right)^2 .
\end{aligned}
$$

Variance estimates

This is unbiased for i.i.d. data:

$$\begin{aligned}
\mathbf{E}s_n^2 &= \frac{1}{2}\mathbf{E}\left(X_1 - X_2\right)^2 \\
&= \frac{1}{2}\mathbf{E}\left((X_1 - \mathbf{E}X_1) - (X_2 - \mathbf{E}X_2)\right)^2 \\
&= \frac{1}{2}\mathbf{E}\left((X_1 - \mathbf{E}X_1)^2 + (X_2 - \mathbf{E}X_2)^2\right) \\
&= \mathbf{E}\left(X_1 - \mathbf{E}X_1\right)^2 .
\end{aligned}$$

U-statistics

Definition: A U**-statistic of order** r **with kernel** h is

$$U = \frac{1}{\binom{n}{r}} \sum_{i \subseteq [n]} h(X_{i_1}, \ldots, X_{i_r}),$$

where h is symmetric in its arguments.

[If h is not symmetric in its arguments, we can also average over permutations.]

"U" for "unbiased." Introduced by Wassily Hoeffding in the 1940s.

U-statistics

Theorem: [Halmos] θ (parameter, i.e., function defined on a family of distributions) admits an unbiased estimator (ie: for all sufficiently large n, some function of the i.i.d. sample has expectation θ) iff for some k there is an h such that

$$\theta = \mathbf{E}h(X_1, \ldots, X_k).$$

Necessity is trivial. Sufficiency uses the estimator

$$\hat{\theta}(X_1, \ldots, X_n) = h(X_1, \ldots, h_k).$$

U-statistics make better use of the sample than this, since they are a symmetric function of the data.

U-statistics: Examples

- s_n^2 is a U-statistic of order 2 with kernel $h(x, y) = (1/2)(x - y)^2$.
- $\bar{X}_n$ is a U-statistic of order 1 with kernel $h(x) = x$.
- The U-statistic with kernel $h(x, y) = |x - y|$ estimates the *mean pairwise deviation* or *Gini mean difference*.

  [The *Gini coefficient*, $G = \mathbf{E}|X - Y|/(2\mathbf{E}X)$, is commonly used as a measure of income inequality.]
- Third k-statistic,

$$k_3 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (X_i - \bar{X}_n)^3$$

  is a U-statistic that estimates the 3rd cumulant.

U-statistics: Examples

- The U-statistic with kernel $h(x, y) = (x - y)(x - y)^T$ estimates the variance-covariance matrix.
- Kendall's τ: For a random pair $P_1 = (X_1, Y_1), P_2 = (X_2, Y_2)$ of points in the plane,

$$\begin{aligned}\tau &= \Pr(P_1P_2 \text{ has positive slope}) - \Pr(P_1P_2 \text{ has negative slope}) \\ &= \mathbf{E}\left(1[P_1P_2 \text{ has positive slope}] - 1[P_1P_2 \text{ has negative slope}]\right),\end{aligned}$$

where P_1P_2 is the line from P_1 to P_2. It is a measure of correlation: $\tau \in [-1, 1]$, $\tau = 0$ for independent X, Y, $\tau = \pm 1$ for $Y = f(X)$ for monotone f. Clearly, τ can be estimated using a U-statistic of order 2.

U-statistics: Examples

The Wilcoxon one-sample rank statistic:

$$T^+ = \sum_{i=1}^{n} R_i 1[X_i > 0],$$

where R_i is the rank (position when $|X_1|, \ldots, |X_n|$ are arranged in ascending order). It's used to test if the distribution is symmetric about zero. Assuming the $|X_i|$ are all distinct, then we can write

$$R_i = \sum_{j=1}^{n} 1[|X_j| \leq |X_i|],$$

U-statistics: Examples

Hence

$$
\begin{aligned}
T^+ &= \sum_{i=1}^{n}\sum_{j=1}^{n} 1[|X_j| \le X_i] \\
&= \sum_{i<j} 1[|X_j| < X_i] + \sum_{i<j} 1[|X_i| < X_j] + \sum_i 1[X_i > 0] \\
&= \sum_{i<j} 1[X_i + X_j > 0] + \sum_i 1[X_i > 0] \\
&= \frac{1}{\binom{n}{2}} \sum_{i<j} \binom{n}{2} 1[X_i + X_j > 0] + \frac{1}{n}\sum_i n\, 1[X_i > 0] \\
&= \frac{1}{\binom{n}{2}} \sum_{i<j} h_2(X_i, X_j) + \frac{1}{n}\sum_i h_1(X_i),
\end{aligned}
$$

U-statistics: Examples

where

$$h_2(X_i, X_j) = \binom{n}{2} 1[X_i + X_j > 0],$$
$$h_1(X_i) = n\, 1[X_i > 0].$$

So it's a sum of U-statistics.

[Why is it not a U-statistic?]

Properties of U-statistics

- "U" for "unbiased": U is an unbiased estimator for $\mathbf{E}h(X_1, \ldots, X_r)$: $\mathbf{E}U = \mathbf{E}h(X_1, \ldots, X_r)$.
- U is a lower variance estimate than $h(X_1, \ldots, X_r)$, because U is an average over permutations. Indeed, since U is an average over permutations π of $h(X_{\pi(1)}, \ldots, X_{\pi(r)})$, we can write

$$U(X_1, \ldots, X_n) = \mathbf{E}\left[h(X_1, \ldots, X_r) | X_{(1)}, \ldots, X_{(n)}\right],$$

  where $(X_{(1)}, \ldots, X_{(n)})$ is the data in some sorted order. Thus, for $\mathbf{E}U = \theta$, we can write the variance as:

Properties of U-statistics

$$\begin{aligned}\mathbf{E}(U-\theta)^2 &= \mathbf{E}\left(\mathbf{E}[h(X_1,\ldots,X_r)-\theta|X_{(1)},\ldots,X_{(n)}]\right)^2 \\ &\leq \mathbf{E}\mathbf{E}\left[(h(X_1,\ldots,X_r)-\theta)^2|X_{(1)},\ldots,X_{(n)}\right] \\ &= \mathbf{E}(h(X_1,\ldots,X_r)-\theta)^2,\end{aligned}$$

by Jensen's inequality (for a convex function ϕ, we have $\phi(\mathbf{E}X) \leq \mathbf{E}\phi(X)$).

This is the Rao-Blackwell theorem: the mean squared error of the estimator $h(X_1,\ldots,X_r)$ is reduced by replacing it by its conditional expectation, given the sufficient statistic $(X_{(1)},\ldots,X_{(n)})$.

Recall: Bounded Differences Inequality

Theorem: Suppose $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the following **bounded differences inequality**:

for all $x_1, \ldots, x_n, x_i' \in \mathcal{X}$,

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq B_i.$$

Then

$$P\left(|f(X) - \mathbf{E}f(X)| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_i B_i^2}\right).$$

Bounded Differences Inequality

Consider a U-statistic of order 2.

$$U = \frac{1}{\binom{n}{2}} \sum_{i<j} h(X_i, X_j).$$

Theorem: If $|h(X_1, X_2)| \leq B$ a.s., then

$$P(|U - \mathbf{E}U| \geq t) \leq 2\exp(-nt^2/(8B^2)).$$

Bounded Differences Inequality

Proof:

For X, X' differing in a single coordinate, we have

$$\begin{aligned} |U - U'| &\leq \frac{1}{\binom{n}{2}} \sum_{i<j} |h(X_i, X_j) - h(X'_i, X'_j)| \\ &\leq \frac{2B(n-1)}{\binom{n}{2}} \\ &= \frac{4B}{n}. \end{aligned}$$

The bounded differences inequality implies the result.

Variance of U-statistics

Now we'll compute the asymptotic variance of a U-statistic. Recall the definition:

$$U = \frac{1}{\binom{n}{r}} \sum_{i \subseteq [n]} h(X_{i_1}, \ldots, X_{i_r}),$$

So [letting S, S' range over subsets of $\{1, \ldots, n\}$ of size r]:

$$\begin{aligned} \mathrm{Var}(U) &= \frac{1}{\binom{n}{r}^2} \sum_S \sum_{S'} \mathrm{Cov}(h(X_S), h(X_{S'})) \\ &= \frac{1}{\binom{n}{r}^2} \sum_{c=0}^{r} \binom{n}{r} \binom{r}{c} \binom{n-r}{r-c} \zeta_c, \end{aligned}$$

where $\binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}$ is the number of ways of choosing S and S' with an intersection of size c (first choose S, then choose the intersection from S, then choose the non-intersection for the rest of S').

Variance of U-statistics

Also, $\zeta_c = \text{Cov}(h(X_S), h(X_{S'}))$ depends only on $c = |S \cap S'|$. To see this, suppose that $S \cap S' = I$ with $|I| = c$,

$$\begin{aligned}
\zeta_c &= \text{Cov}(h(X_S), h(X_{S'})) \\
&= \text{Cov}(h(X_I, X_{S-I}), h(X_I X_{S'-I})) \\
&= \text{Cov}(h(X_1^c, X_{c+1}^r), h(X_1^c X_{r+1}^{2r-c})) \\
&= \text{Cov}\left(\mathbf{E}\left[h(X_1^c, X_{c+1}^r)\middle| X_1^c\right], \mathbf{E}\left[h(X_1^c X_{r+1}^{2r-c})\middle| X_1^c\right]\right) \\
&\qquad + \mathbf{E}\text{Cov}\left[h(X_1^c, X_{c+1}^r), h(X_1^c X_{r+1}^{2r-c})\middle| X_1^c\right] \\
&= \text{Var}\left(\mathbf{E}\left[h(X_1^c, X_{c+1}^r)\middle| X_1^c\right]\right).
\end{aligned}$$

Clearly, $\zeta_0 = 0$.

Variance of U-statistics

Now,

$$\begin{aligned}
\text{Var}(U) &= \frac{1}{\binom{n}{r}^2} \sum_{c=1}^{r} \binom{n}{r} \binom{r}{c} \binom{n-r}{r-c} \zeta_c \\
&= \frac{1}{\binom{n}{r}} \sum_{c=1}^{r} \binom{r}{c} \binom{n-r}{r-c} \zeta_c \\
&= \theta(n^{-r}) \sum_{c=1}^{r} \theta(n^{r-c}) \zeta_c \\
&= \sum_{c=1}^{r} \theta(n^{-c}) \zeta_c .
\end{aligned}$$

Variance of U-statistics

So if $\zeta_1 \neq 0$, the first term dominates:

$$n\text{Var}(U) \to \frac{nr!(n-r)!r(n-r)!}{n!(r-1)!(n-2r+1)!}\zeta_1 \to r^2\zeta_1.$$

If $r^2\zeta_1 = 0$, we say that U is *degenerate*.