第五讲　U 统计量

张伟平

统计与金融系

from Peter Bartlett

第五讲　U 统计量,

1.1 Statistical Functional

■ 在很多时候，感兴趣的参数往往是分布函数 F 的函数，记为 $\theta = T(F)$, 称为**统计泛函**.

- 均值: $T(F) = \int x dF(x)$, 方差 $T(F) = \int (x-\mu)^2 dF(x)$, 分位数 $T(F) = F^{-1}(p)$
- 相关系数 $\rho(t_1, t_2, t_3, t_4, t_5) = \frac{t_3 - t_1 t_2}{\sqrt{(t_4 - t_1^2)(t_5 - t_2^2)}}$，其中 $T_1(F) = \iint x dF(x,y), T_2(F) = \iint y dF(x,y), T_3(F) = \int xy dF(x,y), T_4(F) = \int x^2 dF(x,y), T_5(F) = \int y^2 dF(x,y)$

记 F_n 为经验分布函数，则统计泛函 $\theta = T(F)$ 的"Plug-in"估计为 $\hat{\theta} = T(F_n)$.

Definition

- **Linear statistical functional**: $T(F)=E_F\phi(X)$
- Note in this case that $T\{\alpha F_1+(1-\alpha)F_2\}=\alpha \mathrm{E}_{F_1}\phi(X)+(1-\alpha)\mathrm{E}_{F_2}\phi(X)=\alpha T(F_1)+(1-\alpha)T(F_2)$
- To generalize this idea, we consider a real-valued function taking more than one real argument, say $\phi(x_1,\ldots,x_a)$ for some $a>1$, and define $T(F)=\mathrm{E}_F\phi(X_1,\ldots,X_a)$
- We see that for any permutation π mapping $\{1,\ldots,a\}$ onto itself.
- since there are $a!$ such permutations, consider the function

$$\phi^*(x_1,\ldots,x_a)\stackrel{\text{def}}{=}\frac{1}{a!}\sum_{\text{all }\pi}\phi\left(x_{\pi(1)},\ldots,x_{\pi(a)}\right)$$

Definition For some integer $a \geq 1$, let $\phi : \mathbb{R}^a \to \mathbb{R}$ be a function symmetric in its a arguments. The expectation of $\phi(X_1, \ldots, X_a)$ under the assumption that $X_1, \ldots, X_a$ are independent and identically distributed from some distribution F will be denoted by $\mathrm{E}_F \phi(X_1, \ldots, X_a)$. Then the functional

$$T(F) = \mathrm{E}_F \phi(X_1, \ldots, X_a)$$

is called an expectation functional. If $a = 1$, then T is also called a linear functional.

- Expectation functionals are important because they are precisely the functionals that give rise to V-statistics and U-statistics.

1.2 U Statistcs

Definition Let $T(F) = \mathrm{E}_F \phi(X_1, \ldots, X_a)$ be an expectation functional, where $\phi : \mathbb{R}^a \to \mathbb{R}$ is a function that is symmetric in its arguments. In other words, $\phi(x_1, \ldots, x_a) = \phi\left(x_{\pi(1)}, \ldots, x_{\pi(a)}\right)$ for any permutation π of the integers 1 through a. Then ϕ is called the kernel function associated with $T(F)$.

- **V-Statistics**: (Von Mises) $V_n = T\left(\hat{F}_n\right) = \mathrm{E}_{\hat{F}_n} \phi(X_1, \ldots, X_a) = \frac{1}{n^a} \sum_{i_1=1}^{n} \cdots \sum_{i_a=1}^{n} \phi\left(X_{i_1}, \ldots, X_{i_a}\right)$
- since the bias in V_n is due to the duplication among the subscripts, we might sum instead over all possible subscripts satisfying $i_1 < \cdots < i_a$

Variance estimates

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

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

Variance estimates

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

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

U-statistics

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

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

where h is symmetric in its arguments.

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

U-statistics

Theorem:[Halmos, 1946] A parameter θ 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

$$\theta = \mathrm{E}\, h\left(X_1, \ldots, X_k\right)$$

Necessity is trivial. Sufficiency uses the estimator

$$\hat{\theta}\left(X_1, \ldots, X_n\right) = h\left(X_1, \ldots, X_k\right)$$

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 $h(x,y) = (1/2)(x-y)^2$
- $\overline{X}_n$ 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, $G = \mathrm{E}|X-Y|/(2\mathrm{E}X)$, is commonly used as a measure of income inequality.]
- Third k-statistic, $k_3 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(X_i - \overline{X}_n\right)^3$ is a U-statistic that estimates the 3rd cumulant $\kappa_3 = K^{(3)}(0)$, where $K(t) = \log \mathrm{E}\left[e^{tX}\right]$.

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,

$$\begin{aligned}\tau =& Pr\,(P_1P_2 \text{ has positive slope}) \;-Pr\,(P_1P_2 \text{ has negative slope })\\ &= E1[(X_1 - X_2)(Y_1 - Y_2) > 0] - E1[(X_1 - X_2)(Y_1 - Y_2) < 0]\\ &= 4P(X_1 < X_2, Y_1 < Y_2) - 1\end{aligned}$$

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:

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

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

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

Hence

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

U-statistics: Examples

where

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

$$h_1(X_i) = n1[X_i > 0]$$

So it’s a sum of U-statistics. [Why is it not a U-statistic?]

1.2.1 Properties of U-statistics

- U for unbiased: U is an unbiased estimator for $\mathrm{E}h\left(X_1,\ldots,X_r\right)$, $\mathrm{E}U = \mathrm{E}h\left(X_1,\ldots,X_r\right)$

- U is a lower variance estimate than $h\left(X_1,\ldots,X_r\right)$, 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

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

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

Properties of U-statistics

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

by Jensen's inequality (for a convex function ϕ, we have

$$\phi(\mathrm{E}X)\le \mathrm{E}\phi(X))$$

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

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

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

Bounded Differences Inequality

Consider a U-statistic of order 2.

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

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

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

Bounded Differences Inequality

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

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

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:

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

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

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

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

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

where $X_1^c = (X_1, \ldots, X_c)$. Clearly, $\zeta_0 = 0$.

Variance of U-statistics

Now,

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

Variance of U-statistics

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

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

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

Variance of U-statistics: Example

- Estimator of variance: $h(X_1, X_2) = (1/2)(X_1 - X_2)^2$

$$
\begin{aligned}
\zeta_1 &= \mathrm{Cov}\left(h(X_1, X_2), h(X_1, X_3)\right) \\
&= \mathrm{Var}\left(\mathbf{E}\left[h(X_1, X_2) \,|X_1\right]\right) + \mathbf{E}\left[\mathrm{Cov}\left(h(X_1, X_2), h(X_1, X_3) \,|X_1\right]\right. \\
&= \mathrm{Var}\left(\mathbf{E}\left[h(X_1, X_2) \,|X_1\right]\right) = \mathrm{Var}\left(\mathbf{E}\left[\frac{1}{2}(X_1 - X_2)^2 \,|X_1\right]\right) \\
&= \mathrm{Var}\left(\mathbf{E}\left[\frac{1}{2}(X_1 - \mu + \mu - X_2)^2 \,|X_1\right]\right) \\
&= \mathrm{Var}\left(\frac{1}{2}\left((X_1 - \mu)^2 + \sigma^2\right)\right) = \frac{1}{4}\left(\mu_4 - \sigma^4\right)
\end{aligned}
$$

where $|\mu_4 = \mathrm{E}\left((X_1 - \mu)^4\right)$ is the 4 th central moment. So $n\,\mathrm{Var}(U) \to \mu_4 - \sigma^4$. We'll see that $\sqrt{n}\left(U - \sigma^2\right) \rightsquigarrow N\left(0, \mu_4 - \sigma^4\right)$. (What if $\mu_4 - \sigma^4 = 0$?)

Variance of U-statistics: Example

Recall Kendall's τ : For a random pair $P_1 = (X_1, Y_1)\,, P_2 = (X_2, Y_2)$ of points in the plane, if X, Y are independent and continuous [and P_1P_2 is the line from P_1 to P_2]

$$h\,(P_1, P_2) = (1\,[P_1P_2 \text{ has positive slope }] - 1\,[P_1P_2 \text{ has negative slope }])$$

$$\zeta_1 = \operatorname{Cov}\left(h\,(P_1, P_2)\,, h\,(P_1, P_3)\right)$$

$$\ldots = 1/9$$

so $n\,\mathrm{Var}(U) \to 4/9$. We'll see that $\sqrt{n}U \rightsquigarrow N(0, 4/9)$. And this gives a test for independence.

1.2.2 Asymptotic distribution of U-statistics

How do we find the asymptotic distribution of a U-statistic?

We' ll appeal to this theorem:

Theorem:

$$X_n \rightsquigarrow X \text{ and } d\left(X_n, Y_n\right) \xrightarrow{P} 0 \Longrightarrow Y_n \rightsquigarrow X$$

In particular, we find another sequence $\hat{U}$ such that

- $\sqrt{n}(U - \theta - \hat{U}) \xrightarrow{P} 0$, and
- The asymptotics of $\hat{U}$ are easy to understand.

In this case, we find $\hat{U}$ of the form $\hat{U} = \sum_i f\left(X_i\right)$. Then the CLT gives the result.

Asymptotic distribution of U-statistics

1. Why do functions of a single variable suffice? Because the interactions are weak

2. How do we find suitable functions? By projecting: finding the element of the linear space of functions of single variables that captures most of the variance of U.

This leads us to the idea of **Hájek projections**.

Projection Theorem

Consider a random variable T and a linear space $\mathcal{S}$ of random variables, with $\mathrm{E}S^2 < \infty$ for all $S \in \mathcal{S}$ and $\mathrm{E}T^2 < \infty$. [Write $T \in L_2(P)$, $\mathcal{S} \subset L_2(P)$, the Hilbert space of finite variance random variables defined on a probability space.] A projection $\hat{S}$ of T on $\mathcal{S}$ is a minimizer over $\mathcal{S}$ of $\mathbf{E}(T-S)^2$.]

> Theorem: $\hat{S}$ is a projection of T on $\mathcal{S}$ iff $\hat{S} \in \mathcal{S}$ and, for all $S \in \mathcal{S}$, the error $T - \hat{S}$ is orthogonal to $\mathcal{S}$, that is,
>
> $$\mathbf{E}(T - \hat{S})S = 0$$
>
> If $\hat{S}_1$ and $\hat{S}_2$ are projections of T onto $\mathcal{S}$, then $\hat{S}_1 = \hat{S}_2$ a.s.

Projection Theorem

Notice that if $\mathcal{S}$ contains constants, then $S = 1 \in \mathcal{S}$ shows that

$$\mathbf{E}(T - \hat{S}) = 0, \quad \text{i.e., } \mathbf{E}T = \mathbf{E}\hat{S}$$

Also, for all $S \in \mathcal{S}$, $S - \mathbf{E}S \in \mathcal{S}$, so

$$\mathrm{Cov}(T - \hat{S}, S) = \mathbf{E}((T - \hat{S})(S - \mathbf{E}S)) = 0$$

Projection Theorem Proof

Theorem: 1 . $\hat{S} \in \mathcal{S}$ is a projection of T on $\mathcal{S}$ (minimizes $\mathbf{E}(T-S)^2$ iff , for all $S \in \mathcal{S}, \mathbf{E}(T-\hat{S})S = 0$ 2. If $\hat{S}_1$ and $\hat{S}_2$ are projections of T onto $\mathcal{S}$, then $\hat{S}_1 = \hat{S}_2$ as.

We can write the criterion, for any $S \in \mathcal{S}$ as

$$\begin{aligned}\mathbf{E}\left(T-S\right)^2 &= \mathbf{E}(T-\hat{S}+\hat{S}-S)^2 \\ &= \mathbf{E}(T-\hat{S})^2 + 2\mathbf{E}((T-\hat{S})(\hat{S}-S)) + (\hat{S}-S)^2\end{aligned}$$

If $\mathrm{E}(T-\hat{S})S = 0$, then this is $\mathrm{E}(T-\hat{S})^2 + (\hat{S}-S)^2$, which is minimized for $S = \hat{S}$, and strictly minimized unless $\mathrm{E}(\hat{S}-S)^2 = 0$, so $\hat{S}$ is unique.

Projection Theorem Proof

If $\hat{S}$ is a projection, then

$$\mathbf{E}(T - \hat{S} - \alpha S)^2 = \mathbf{E}(T - \hat{S})^2 - 2\alpha \mathbf{E}(T - \hat{S})S + \alpha^2 \mathbf{E} S^2$$

is at least $\mathbf{E}(T - \hat{S})^2$ for any $S \in \mathcal{S}$ and any α. And this implies that

$$\mathbf{E}(T - \hat{S})S = 0$$

Projection Theorem

- Pythagoras theorem:$\mathbf{E}(T)^2 = \mathbf{E}(T - \hat{S} + \hat{S})^2 = \mathbf{E}(T - \hat{S})^2 + \mathbf{E}(\hat{S})^2$

- If $\mathcal{S}$ contains constants, $\mathbf{E}(T) = \mathbf{E}(\hat{S})$ and $\mathrm{Var}(T) = \mathrm{Var}(T - \hat{S}) + \mathrm{Var}(\hat{S})$

- So if $\mathcal{S}$ contains constants and $\hat{S}$ and T have the same variance, then $\hat{S} = T$ a.s.

- A similar property holds asymptotically...

Projections and Asymptotics

Consider $\mathcal{S}_n$ a sequence of linear spaces of random variables that contain the constants and that have finite second moments.

Theorem: For T_n with projections $\hat{S}_n$ on $\mathcal{S}_n$,

$$\frac{\mathrm{Var}(T_n)}{\mathrm{Var}(\hat{S}_n)} \to 1 \quad \Longrightarrow \quad \frac{T_n - \mathrm{E}T_n}{\sqrt{\mathrm{Var}\,(T_n)}} - \frac{\hat{S}_n - \mathrm{E}\hat{S}_n}{\sqrt{\mathrm{Var}(\hat{S}_n)}} \xrightarrow{P} 0$$

Projections and Asymptotics: Proof

Define

$$Z_n = \frac{T_n - \mathbf{E}T_n}{\sqrt{\operatorname{Var}(T_n)}} - \frac{\hat{S}_n - \mathbf{E}\hat{S}_n}{\sqrt{\operatorname{var}\left(\hat{S}_n\right)}}$$

Clearly, $\mathbf{E}Z_n = 0$

$$\begin{aligned}
\operatorname{Var}(Z_n) &= 2 - 2\frac{\operatorname{Cov}\left(T_n, \hat{S}_n\right)}{\sqrt{\operatorname{Var}(T_n)}\sqrt{\operatorname{Var}\left(\hat{S}_n\right)}} \\
&= 2 - 2\frac{\sqrt{\operatorname{Var}\left(\hat{S}_n\right)}}{\sqrt{\operatorname{Var}(T_n)}} \to 0
\end{aligned}$$

where the second equality is because $\mathcal{S}$ contains constants, so $\operatorname{Cov}\left(T_n - \hat{S}_n, \hat{S}_n\right) = 0$, hence $\operatorname{Cov}\left(T_n, \hat{S}_n\right) = \operatorname{Var}\left(\hat{S}_n\right)$.

Linear Spaces

- What linear spaces should we project onto? We need a rich space, since we have to lose nothing asymptotically when we project.
- We' ll consider the space of functions of a single random variable. Then projection corresponds to computing conditional expectations Just as $\mathrm{E}X = \arg\min_{a\in\mathbb{R}} \mathrm{E}(X-a)^2$

$$\mathrm{E}[X|Y] = \arg\min_{g:\mathbb{R}\to\mathbb{R}} \mathrm{E}(X-g(Y))^2$$

This is the projection of X onto the linear space $\mathcal{S}$ of measurable functions of Y.

Conditional Expectations as Projections

The projection theorem says: for all measurable g,

$$\mathbf{E}(X - \mathbf{E}[X|Y])g(Y) = 0$$

Properties of $\mathrm{E}[X|Y]$:

- $\mathrm{E}X = \mathrm{E}\mathrm{E}[X|Y]$(consider g=1)
- For a joint density $f(x, y)$

$$\mathbf{E}[X|Y] = \int x \frac{f(x, Y)}{f(Y)} dx$$

- For independent X, Y, $\mathbf{E}(X - \mathrm{E}X)g(Y) = 0$, so $\mathrm{E}[X|Y] = \mathrm{E}X$

Conditional Expectations as Projections

Properties of $\mathrm{E}[X|Y]$:

- $E[f(Y)X|Y] = f(Y)\mathrm{E}[X|Y]$ (Because $\mathrm{E}[f(Y)X - f(Y)\mathrm{E}[X|Y]g(Y) = \mathrm{E}[X - \mathrm{E}[X|Y]f(Y)g(Y) = 0.$)
- $\mathrm{E}[\mathrm{E}[X|Y,Z]|Y] = \mathrm{E}[X|Y]$ (Because $\mathrm{E}(\mathrm{E}[X|Y,Z] - \mathrm{E}[X|Y])g(Y) = \mathrm{E}(\mathrm{E}[g(Y)X|Y,Z] - \mathrm{E}[g(Y)X|Y]) = 0.$)

Definition: For independent random vectors $X_1, \ldots, X_n$, the Hájek pro- jection of a random variable is its projection onto the set of sums

$$\sum_{i=1}^{n} g_i\left(X_i\right)$$

$\sum_{i=1}^{n} g_i\left(X_i\right)$ of measurable functions satisfying $\mathrm{E}g_i\left(X_i\right)^2 < \infty$

Hájek Projections

Theorem: [Hájek projection principle:] The Hájek projection of $T \in L_2(P)$ is

$$\hat{S} = \sum_{i=1}^{n} \mathrm{E}\left[T|X_i\right] - (n-1)\mathbf{E}T$$

Hájek Projections Principle: Proof

From the projection theorem, we need to check that $T - \hat{S}$ is orthogonal to each $g_i(X_i)$. It suffices if $\mathbf{E}[T|X_i] = \mathbf{E}\left[\hat{S}|X_i\right]$:

$$\mathrm{E}(T - \hat{S})g_i(X_i) = \mathrm{E}\left(\mathrm{E}\left[T - \hat{S}|X_i\right] g_i(X_i)\right)$$

But

$$\begin{aligned}
\mathbf{E}\left[\hat{S}|X_i\right] &= \mathbf{E}\left[\sum_{j=1}^{n} \mathbf{E}[T|X_j] - (n-1)\mathbf{E}T|X_i\right] \\
&= \mathbf{E}[T|X_i] + \sum_{j \neq i} \mathbf{E}[T|X_j]\,|X_i\Bigg] - (n-1)\mathbf{E}T \\
&= \mathbf{E}[T|X_i]
\end{aligned}$$

because the X_i are independent, so $T - \hat{S}$ is orthogonal to $\mathcal{S}$.

Asymptotic Normality of U-Statistics

Theorem: If $\mathrm{E}h^2 < \infty$, define $\hat{U}$ as the Hájek projection of $U - \theta$. Then

$$\hat{U} = \frac{r}{n} \sum_{i=1}^{n} h_1 \left(X_i \right),$$

with

$$h_1(x) = \mathrm{Eh} \ \left(x, X_2, \ldots, X_r \right) - \theta$$

$$\sqrt{n}(U - \theta - \hat{U}) \xrightarrow{P} 0,$$

$$\sqrt{n}(U - \theta) \rightsquigarrow N \left(0, r^2 \zeta_1 \right), \quad \text{where}$$

$$\zeta_1 = \mathrm{E} h_1^2 \left(X_1 \right)$$

Asymptotic Normality of U-Statistics: Proof

Recall:

$$U = \frac{1}{\binom{n}{r}} \sum_{j \subseteq [n]} h\left(X_{j_1}, \ldots, X_{j_r}\right)$$

By the Hájek projection principle, the projection of $U - \theta$ is

$$\hat{U} = \sum_{i=1}^{n} \mathrm{E}\left[U - \theta | X_i\right]$$

$$= \sum_{i=1}^{n} \frac{1}{\binom{n}{r}} \sum_{j \subseteq [n]} \mathbf{E}\left[h\left(X_{j_1}, \ldots, X_{j_r}\right) - \theta | X_i\right]$$

But

$$\mathbf{E}\left[h\left(X_{j_1}, \ldots, X_{j_r}\right) - \theta | X_i\right] = \begin{cases} h_1\left(X_i\right) & \text{if } i \in j \\ 0 & \text{otherwise} \end{cases}$$

Asymptotic Normality of U-Statistics: Proof

For each X_i, there are $\binom{n-1}{r-1}$ of the $\binom{n}{r}$ subsets that contain i. Thus,

$$\hat{U} = \sum_{i=1}^{n} \frac{r!(n-r)!(n-1)!}{n!(r-1)!(n-r)!} h_1\left(X_i\right) = \frac{r}{n} \sum_{i=1}^{n} h_1\left(X_i\right)$$

To see that $\hat{U}$ has the same asymptotics as U, notice that $\mathrm{E}\hat{U} = 0$ and so its variance is asymptotically the same as that of U :

$$\operatorname{var} \hat{U} = \frac{r^2}{n} \mathrm{E}\, h_1^2\left(X_1\right) = \frac{r^2}{n} \mathrm{E}\left(\mathrm{E}\left[h\left(X_1^r\right)|X_1\right] - \theta\right)^2$$

$$= \tfrac{r^2}{n} \operatorname{Var}\left(\mathrm{E}\left[h\left(X_1^r\right)|X_1\right]\right) = \tfrac{r^2}{n} \zeta_1$$

Asymptotic Normality of U-Statistics: Proof

CLT (and finiteness of $\mathrm{Var}(\hat{U})$) implies $\sqrt{n}\hat{U} \rightsquigarrow N\left(0, r^2\zeta_1\right)$
Also [recall that $n \,\mathrm{Var}\, U \to r^2\zeta_1, \mathrm{Var}\, \hat{U} / \mathrm{Var}\, U \to 1$, so

$$\frac{U-\theta}{\sqrt{\mathrm{Var}(U)}} - \frac{\hat{U}}{\sqrt{\mathrm{Var}(\hat{U})}} \xrightarrow{P} 0$$

which implies $\sqrt{n}(U-\theta-\hat{U}) \xrightarrow{P} 0$, and hence

$$\sqrt{n}(U-\theta) \rightsquigarrow N\left(0, r^2\zeta_1\right)$$

Asymptotic Normality of U-Statistics: Examples

Estimator of variance: $h(X_1, X_2) = (1/2)(X_1 - X_2)^2$:

$$\zeta_1 = \frac{1}{4}\left(\mu_4 - \sigma^4\right)$$

where $\mu_4 = \mathrm{E}\left((X_1 - \mu)^4\right)$ is the 4 th central moment. So $n\,\mathrm{Var}(U) \to \mu_4 - \sigma^4$, hence $\sqrt{n}\left(U - \sigma^2\right) \rightsquigarrow N\left(0, \mu_4 - \sigma^4\right)$

Asymptotic Normality of U-Statistics: Examples

Recall Kendall's τ : For a random pair $P_1 = (X_1, Y_1), P_2 = (X_2, Y_2)$ of points in the plane, if X, Y are independent and continuous [recall: $P_1 P_2$ is the line from P_1 to P_2]

$$h(P_1, P_2) = (1 [P_1 P_2 \text{ has positive slope }] - 1 [P_1 P_2 \text{ has negative slope }])$$

$$\mathrm{E}\tau = 0$$

$$\zeta_1 = \mathrm{Cov}(h(P_1, P_2), h(P_1, P_3)) = \frac{1}{9}$$

Thus $\sqrt{n}U \rightsquigarrow N(0, 4/9)$. And this gives a test for independence of X and Y :

$$\Pr\left(\sqrt{9n/4}|\tau| > z_{\alpha/2}\right) \to \alpha$$

Asymptotic Normality of U-Statistics: Examples

Recall Wilcoxon's one sample rank statistic:

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

$$
\begin{aligned}
h_2\left(X_i, X_j\right) &= \binom{n}{2} 1\left[X_i + X_j > 0\right] \\
h_1\left(X_i\right) &= n 1\left[X_i > 0\right]
\end{aligned}
$$

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.

Asymptotic Normality of U-Statistics: Examples

It's a sum of U-statistics. The first sum dominates the asymptotics. So consider

$$U = \frac{1}{\binom{n}{2}} \sum_{i<j} \binom{n}{2} 1\left[X_i + X_j > 0\right]$$

The Hájek projection of $U - \theta$ is

$$\hat{U} = \frac{2}{n} \sum_{i=1}^{n} h_1\left(X_i\right)$$

Asymptotic Normality of U-Statistics: Examples

and

$$
\begin{aligned}
h_1(x) &= \mathrm{Eh}\left(x, X_2\right) - \mathrm{Eh}\left(X_1, X_2\right) \\
&= \binom{n}{2}\left(P\left(x + X_2 > 0\right) - P\left(X_1 + X_2 > 0\right)\right) \\
&= -\binom{n}{2}\left(F(-x) - \mathrm{EF}\left(-X_1\right)\right)
\end{aligned}
$$

Asymptotic Normality of U-Statistics: Examples

For F symmetric about $0, (F(x) = 1 - F(-x))$, we have

$$\begin{aligned}\hat{U} &= -\frac{2\binom{n}{2}}{n}\sum_{i=1}^{n}\left(F\left(-X_i\right) - \mathrm{E}F\left(-X_i\right)\right) \\ &= \frac{2\binom{n}{2}}{n}\sum_{i=1}^{n}\left(F\left(X_i\right) - \mathrm{E}F\left(X_i\right)\right)\end{aligned}$$

But $F\left(X_i\right)$ is always uniform on $[0,1]$, and so $\mathrm{E}F\left(X_i\right) = 1/2$ and $\mathrm{Var}\,F\left(X_i\right) = 1/12$. Thus,

$$\mathrm{Var}(\hat{U}) = \frac{4\binom{n}{2}^2}{n}\mathrm{Var}\left(F\left(X_i\right)\right) = \frac{n(n-1)^2}{12}$$

Asymptotic Normality of U-Statistics: Examples

Thus, for symmetric distributions,

$$n^{-3/2}\left(T^+ - \frac{\binom{n}{2}}{2}\right) \rightsquigarrow N(0,1/12)$$

So we have a test for symmetry:

$$Pr\left(\sqrt{12}n^{-3/2}\left|T^+ - \frac{\binom{n}{2}}{2}\right| > z_{\alpha/2}\right) \to \alpha$$