Institute of Mathematics, EPFL MATH524

Chapter 4: U- and V-statistics

Lecturer: Rajita Chandak *Spring 2025*

Lets now zoom out and look at the larger picture we have tried to understand thus far. We have attempted to, in its most general form, study the functional

$$\theta = \theta(F), \quad F \in \mathcal{F}$$

where $\mathcal{F}$ is a class of distributions. So far we have been concerned with the non-parametric estimation of $\theta(F) = F$ and $\theta(F) = \mathbb{E}[F]$ in the form of our density and regression estimation problems respectively.

While we have assessed our non-parametric estimators through the MSE, there may be reasons to ask a different set of questions regarding these estimators. In particular we may want to find out the following:

1. Does there exist unbiased estimator $\hat{\theta}$ for θ for all $F \in \mathcal{F}$?.
2. If such an estimator exists, what is it? If several exist, which is the best?

In particular, are there alternatives to the kernel estimators that are unbiased in finite sample? Do these estimators have better asymptotic properties than kernel estimators? Or is there a trade-off between the finite-sample and asymptotic properties?

Lets start by investigating the first question of existence. We can say that a functional θ defined on $\mathcal{F}$ admits an unbiased estimator if and only if there is a function h of d variables, $\mathbf{x} = [x_1, \ldots, x_d]$ such that

$$\theta(F) = \int h(x_1, \ldots, x_d) F(d\mathbf{x})$$

for all $F \in \mathcal{F}$. WLOG h is assumed to be symmetric. Our regression set up fits into this definition as $\theta(F) = \mathbb{E}_F[h(X_1, \ldots, X_d)]$, for $[X_1, \ldots, X_d] \overset{i.i.d.}{\sim} F$.

1 V-statistics

Lets now take the plug-in regression estimator, with n i.i.d. samples

$$\theta(\hat{F}_n) = \frac{1}{n^d} \sum_{i_1, \ldots i_d} h(X_{i1}, \ldots, X_{id})$$

If $d = 2$, we can re-write the estimator as

$$\begin{aligned} V_n &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h\left(X_i, X_j\right) \\ &= \frac{2}{n^2} \sum_{i<j} h\left(X_i, X_j\right) + \frac{1}{n^2} \sum_{i=1}^{n} h\left(X_i, X_i\right) \end{aligned}$$

We are essentially splitting the double sum into a sum that only contains the diagonal elements and another sum containing the remaining entries. Then, taking expectations, we get

$$\mathbb{E}[V_n] = \frac{n-1}{n}\theta(F) + \frac{1}{n}\mathbb{E}[h(X_i, X_i)].$$

Here we see the advantage of splitting the original double-sum. As $n \to \infty$, the first term converges to the true functional and the second term, which is the bias, converges to zero.

2 U-statistics

The V-statistic clearly had a non-negligible bias in finite samples. So our next step is to construct an unbiased statistic that will not suffer from this bias. We should be able to convince ourselves that the following symmetric estimator is unbiased:

$$U_n(X_1, \ldots, X_n) = \binom{n}{r}^{-1} \sum_{1 \le i_1 < \ldots < i_r \le n} h(X_{i_1}, \ldots, X_{i_r})$$

This estimator, also known as a **U-statistic**, was first introduced by Hoeffding in 1948. h is typically called the kernel, with its order being determined by the function space it maps from. i.e., $h : \mathbb{R}^r \to \mathbb{R}$ is a symmetric function (kernel) of order r.

It turns out this estimator is the only symmetric estimator which is unbiased for all F for which $\theta(F)$ exists, and it can be shown to have smaller variance than any other such unbiased estimator (i.e., it is UMVU).

How does this estimator compare to the V-statistic we constructed? Actually, we can show that the V-statistic can be written as a function of the U-statistic. For simplicity we choose $r = 2$, but this relationship generalizes for any r. Note that the U-statistic is

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

This sum is exactly the off-diagonal term in the V-statistic expansion. So now we simply re-scale the U-statistic by $(n-1)/n$ and we get

$$V_n = \frac{1}{n^2} \sum_{i} h(X_i, X_i) + \frac{n-1}{n} U_n.$$

Asymptotically,

$$\sqrt{n}(V_n - \theta) = \frac{n-1}{n}\sqrt{n}(U_n - \theta) + \frac{\sqrt{n}}{n^2}\sum_{i=1}^{n}[h(X_i, X_i) - \theta].$$

Therefore, V_n and U_n are asymptotically equivalent. Importantly, note that U-statistics are unbiased while V-statistics are only asymptotically unbiased.

We now provide some concrete examples of U-statistics.

Example 1 (Mean)**.** *Take*

$$\theta(F) = \mathbb{E}[X_1]$$

with the plug-in estimator

$$U_n = \frac{1}{n}\sum_{i=1}^{n} X_i = U_n.$$

Here, $h(x) = x$.

Example 2 (Functions of moments)**.** *Suppose we want to estimate more generic functions of moments of the distribution.Consider for example, the squared mean:*

$$\theta = \mathbb{E}[X]^2$$

It turns out that there is no degree-1 kernel function that can be used for constructing a U-statistic for this problem. We have to use a degree-2 kernel: $h(x_1, x_2) = x_1 x_2$. The U-statistic is given by

$$U = \frac{1}{n(n-1)}\sum_{i<j} X_i X_j.$$

Example 3 (Variance)**.** *Now we may think of using the second moment and squared-mean U-statistics to generate an estimate for the variance.*

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Plugging in,

$$\mathbb{V}(X) = \mathbb{E}[X_1^2 - X_1 X_2].$$

However, we notice that this corresponds to $f(x_1, x_2) = x_1^2 - x_1 x_2$, which is not symmetric. We have to symmetrize the function:

$$h(x_1, x_2) = \frac{1}{2}[f(x_1, x_2) + f(x_2, x_1)] = \frac{x_1^2 - 2x_1 x_2 + x_2^2}{2} = \frac{(x_1 - x_2)^2}{2}.$$

Then, the estimator of the U-statistic is given by:

$$U_n = \binom{n}{2}^{-1} \sum_{i<j} \frac{(X_i - X_j)^2}{2} = \frac{1}{n-1} \sum_i (X_i - \bar{X}_n)^2.$$

Example 4 (Gini mean difference)**.** *If we consider the absolute deviation (also known as the Gini mean difference),*

$$\theta(F) = \mathbb{E}[|X_1 - X_2|],$$

the corresponding U-statistic is

$$U_n = \binom{n}{2}^{-1} \sum_{i<j} \left| X_i - X_j \right|,$$

with $h(x_1, x_2) = |x_1 - x_2|$.

2.1 Bounded differences

The following property of U-statistics can be very helpful in establishing tail bounds:

Theorem 1 (Bounded difference of U-statistics)
Suppose $r = 2$. *If* $|h(x_1, x_2)| \leq B$ *a.s. Then,*

$$\mathbb{P}(|U - \mathbb{E}[U]| \geq t) \leq 2 \exp\left(\frac{-nt^2}{8B^2}\right).$$

The proof of this theorem follows directly given the following inequality bound,

Lemma 1 (Bounded differences inequality)
Suppose $f : \mathcal{X}^n \to \mathbb{R}$ *such that 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,

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

and the observation that for x, x' differing in a single coordinate,

$$|U - U'| \leq \frac{1}{\binom{n}{2}} \sum_{i<j} |h(x_i, x_j) - h(x_i', x_j')| \leq \frac{2B(n-1)}{\binom{n}{2}} = \frac{4B}{n}.$$

3 Variance of U-statistics

Now, we want to understand the variance of U-statistics, in order to answer the question of the 'best' unbiased statistic. In order to compute the variance most efficiently we first need to understand projections and the Hoeffding decomposition. We will use the following vector space, known as the **Hilbert space** in the forthcoming analysis:

Definition 1 (Hilbert space)**.** *A vector space is also called the Hilbert space if it is a complete normed space with inner product* $< \cdot, \cdot >$*:*

$$||u||^2 =< u, u > \quad \textit{and} \quad < x + y, u + v >=< x, u > + < x, v > + < y, u > + < y, v > .$$

We will be using two Hilbert spaces: $\mathbb{R}^n$ with the standard inner product and $L^2(F) = \{f : \mathbb{R} \to \mathbb{R} | \int f^2 \mathrm{d}F(x) < \infty\}$ with inner product $< f, g >= \int fg\mathrm{d}F$.

3.1 Projections

Let S be a closed linear subspace of $\mathbb{R}^n$. For some vector $v \in \mathbb{R}^n$, define the projection

$$\pi_S(v) = \underset{s \in S}{\operatorname{argmin}} ||v - s||^2.$$

The following theorem is a very useful fact about projections.

Theorem 2
The projection $\pi_s(v)$ *exists and uniquely defined by*

$$< v - \pi_s(v), s >= 0, \textit{ for all } \ s \in S.$$

Example 5. $X \in S \subseteq L^2(F)$ *is the r.v. with* $\mathbb{E}[X^2] < \infty$*, closed under linear combinations. Then,* $\hat{X}$ *is the projection of an r.v.* T *onto* S *iff*

$$\mathbb{E}[(T - \hat{X})X] = 0, \textit{ for all } X \in S$$

Another very common example of projections is the conditional expectation function.

Example 6. *Consider a random variable* Y *and the space*

$$S = \{\textit{Linear span of } g(Y) : g \textit{ measurable }, \mathbb{E}[g(Y)^2] < \infty\}.$$

Then, the conditional expectation of $X \in L^2$ *given* Y *is the projection of* X *on* S*.*

The projection theorem implies the property that $\mathbb{E}[(X - \mathbb{E}[X|Y])g(Y)] = 0$ for all integrable g (and therefore also the tower property).

The following lemma projections presents the argument for why we want to use projections in computing the variance of U-statistics.

Lemma 2
Let S_k be a sequence of subspaces of L^2 and T_k be a sequence of random variables. Let $X_k = \pi_{S_k}(T_k)$. If $\mathbb{V}(T_k)/\mathbb{V}(X_k) \to 1$,

$$\frac{T_k - \mathbb{E}[T_k]}{\sqrt{\mathbb{V}(T_k)}} - \frac{X_k - \mathbb{E}[X_k]}{\sqrt{\mathbb{V}(X_k)}} \xrightarrow{p} 0.$$

This lemma is useful in the sense that if the right projection space S is chosen, there is asymptotically no information loss in using the projected estimator.

3.2 Hájek Projection

For U-statistics we will use the Hájek projection. This is also sometimes referred to as the *linearization* of an estimator.

Definition 2 (Hájek projection)**.** *For independent random vectors $X_1, \dots X_n$, the Hájek projection is the projection onto i.i.d. sums $\sum_i g_i(X_i)$ of measurable functions satisfying $\mathbb{E}[g_i(X_i)^2] < \infty$.*

The Hájek projection of the U-statistic $(U_n - \theta)$, for $r = 2$, is

$$\hat{U}_n = \sum_i \mathbb{E}[U_n - \theta | X_i] = \frac{2}{n} \sum_i (\mathbb{E}[h(X_1, X_2)|X_i = x_i] - \theta),$$

This implies that (for $r = 2$)

$$\hat{U}_n = \frac{2}{n} \sum_{i=1}^{n} \mathbb{E}[h(X_1, X_2)|X_i] + o_p(n^{-1/2})$$

This projection allows for analyzing U-statistics with any of our tools for sums of i.i.d. variables.

3.3 Hoeffding Decomposition

The Hoeffding decomposition is a recursive Hájek projection of the U-statistic that separates the higher-order terms in order to help understand the asymptotic behavior. Define

$$h_k(x_1, \dots, x_k) = \mathbb{E}[h(X_1, \dots, X_d) \mid X_1 = x_1, \dots, X_k = x_k]$$

and the corresponding centered version $\tilde{h}_k = h_k - \theta$. Then, define

$$\begin{aligned}
g_1(X_1) &\equiv \tilde{h}_1(X_1), \\
g_2(X_1, X_2) &\equiv \tilde{h}_2(X_1, X_2) - g_1(X_1) - g_1(X_2), \\
g_3(X_1, X_2, X_3) &\equiv \tilde{h}_3(X_1, X_2, X_3) - \sum_{j=1}^{3} g_1(X_j) - g_2(X_1, X_2) - g_2(X_1, X_3) - g_2(X_2, X_3),
\end{aligned}$$

and so on.

Then, the U-statistic using g_k, i.e., $U_n(g_k)$, instead of h is known as the Hoeffding decomposition. This is a simple translation:

$$\hat{U}_n(h) - \theta = \sum_{k=1}^{d} \binom{n}{d}^{-1} U_n(g_k)$$

To understand better exactly what this transformation means, lets look at how it impacts the sample mean U-statistic: Recall, we had $h(x) = x$ and

$$U_n = \frac{1}{n} \sum_i X_i$$

The Hoeffding decomposition is then given by

$$\hat{U}_n = \frac{1}{n} \sum_i (X_i - \theta).$$

The linearity of the Hájek projection gives the general formulation

$$\sqrt{n}(U_n - \theta) = \frac{1}{\sqrt{n}} \sum_i \varphi(X_i) + o_p(1),$$

where φ is called the **influence function**. This is because it is mean-zero and at the true parameter it quantifies how the estimator changes with small perturbations around the evaluation point. The influence function of a U-statistic is simply the first term of the Hájek projection.

3.4 Asymptotic normality

We conclude our study of U-statistics by establishing asymptotic normality.

Theorem 3
Let h be a symmetric kernel of order r with finite variance. Then,

$$\sqrt{n}(U_n - \theta - \hat{U}_n) \xrightarrow{p} 0,$$

and

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} \mathcal{N}(0, r^2 \mathbb{E}[h_1^2]).$$

Proof. For simplicity, we will prove the result for $r = 2$. The generalization follows directly. Lets start by computing the variance of U_n;

$$\mathbb{V}(U_n) = \mathbb{V}\left(\frac{2}{n(n-1)} \sum_{i<j} h(X_i, X_j) \right) = \frac{4}{n} \mathbb{C}\text{ov}(h(X_1, X_2), h(X_2, X_1)) = \frac{4}{n} \mathbb{V}(\tilde{h}_1(X_1)),$$

where the last equality follows by symmetry of h. Recall that $\mathbb{E}[\tilde{h}_1(X_1)] = 0$. Then,

$$\mathbb{V}(\tilde{h}_1(X_1)) = \mathbb{E}[\tilde{h}_1^2(X_1)].$$

Now we turn to the U-statistic estimator. By the Hájek projection, we know

$$\hat{U}_n = \sum_i \mathbb{E}[U_n - \theta | X_i] = \frac{2}{n} \sum_i (\mathbb{E}[h(X_1, X_2)|X_i = x_i] - \theta),$$

Note that $\mathbb{E}[\hat{U}_n] = 0$, and

$$\mathbb{V}(\hat{U}_n) = \frac{4}{n}\mathbb{E}[\tilde{h}_1^2(X_1)]$$

Then, by CLT and finiteness of $\mathbb{V}(\hat{U}_n)$,

$$\sqrt{n}\hat{U}_n \xrightarrow{d} \mathcal{N}(0, 4\mathbb{E}[\tilde{h}_1^2(X_1)]).$$

Therefore, we can see that the condition

$$\frac{\mathbb{V}(\hat{U}_n)}{\mathbb{V}(U_n)} \to 1$$

is satisfied. Thus,

$$\frac{U_n - \theta}{\sqrt{\mathbb{V}(U)}} - \frac{\hat{U}_n}{\sqrt{\mathbb{V}(\hat{U}_n)}} \xrightarrow{p} 0,$$

which implies $\sqrt{n}(U_n - \theta - \hat{U}_n) \xrightarrow{p} 0$, and thus

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} \mathcal{N}(0, 4\mathbb{E}[\tilde{h}_1^2(X_1)]).$$

■

4 Two-sample U-statistics

Consider now, the case where we have two different samples, from possibly two different distributions. That is, $\{X_i\}_{i=1}^n$ and $\{Y_j\}_{j=1}^m$ such that $X_i \overset{i.i.d.}{\sim} F_X$ and $Y_j \overset{i.i.d.}{\sim} F_Y$ with possibly $F_X \neq F_Y$. We can extend our notion of U-statistics to the two samples by considering a new type of kernel function h that is permutation symmetric in X and Y separately. That is,

$$h(x_{i_1}, \ldots, x_{i_r}, y_{j_1}, \ldots, y_{j_s})$$

for all permutations of $[r]$ and $[s]$. The U-statistic is then defined as

$$U = \binom{n}{r}^{-1}\binom{m}{s}^{-1}\sum h(X_{i_1},\ldots,X_{i_r},Y_{j_1},\ldots,Y_{j_s}).$$

The natural target parameter here is then given by

$$\theta = \mathbb{E}[h(X_{i_1},\ldots,X_{i_r},Y_{j_1},\ldots,Y_{j_s})].$$

Turns out we can establish the same type of consistency and asymptotic normality results as for the one sample U-statistic under bounded differences of h in X and Y separately and finite variance. An additional constraint is required for the results to go through: $m, n \to \infty$ with $O(m) \asymp O(n)$. This condition simply ensures that both X and Y sample sizes grow proportionally. This condition can also be formulated as for $N = n + m$,

$$\frac{m}{N} \to \lambda \qquad \text{and} \qquad \frac{n}{N} \to \lambda, \qquad\qquad 0 < \lambda < 1.$$

The Hoeffding decomposition for analysing the U-statistic is now done with projections of the form $\sum_i k_i(x_i) + \sum_j l_j(y_j)$. That is,

$$\hat{U} = \frac{r}{n}\sum_i h_{1,0}(x_i) + \frac{s}{n}\sum_j h_{0,1}(y_j),$$

where

$$h_{1,0}(x) = \mathbb{E}[h(x, X_2,\ldots,X_r,Y_1,\ldots,Y_s)] - \theta$$
$$h_{0,1}(y) = \mathbb{E}[h(X_1,\ldots,X_r,y,Y_2,\ldots,Y_s)] - \theta.$$

Note that this is just the Hajek projection determined by the inclusion of both conditional expecations, $\mathbb{E}[h|X_i]$ and $\mathbb{E}[h|Y_j]$. From this, $U - \theta - \hat{U} \xrightarrow{p} 0$ holds. Furthermore, if $\int h^2 \mathrm{d}(y, x) < \infty$, $\hat{U} \sim \mathcal{N}(0, V)$, for some bounded variance function V.

Let's now look at an example where we might wish to use this two-sample U-statistic.

Example 7 (Mann-Whitney test statistics)**.** *Consider the problem of $\theta = \mathbb{P}(X \leq Y)$. This is the notion of identifying stochastic dominance between two random variables. If the probablity is large, then we say that Y stochastically dominates X. The natural kernel for this parameter is*

$$h(x, y) = \mathbf{1}(x \leq y).$$

The kernel if of order 1 and is already symmetric with respect to x and y separately so we don't need to change anything here. Let's use this to define the U-statistics:

$$U = \frac{1}{mn}\sum_i\sum_j \mathbf{1}(X_i \leq X_j).$$

The estimator mnU is known as the Mann-Whitney statistic and is used in hypothesis testing problems that deal with difference of means (or locations). Here is a concrete example of such a hypothesis testing problem for which the Mann-Whitney test is the natural test to use. Suppose $X_i \sim F$ *and* $Y_j \sim G$*. Then,*

$$\hat{U} = U - \theta = -\frac{1}{m}\sum_i (G(X_i) - \mathbb{E}[G(X_i)]) + \frac{1}{n}\sum_j (F(Y_j) - \mathbb{E}[F(Y_j)]).$$

Under the null hypothesis that $F = G$*, i.e.,,*

$$H_0 : \{X_i, Y_j\} \overset{i.i.d.}{\sim} F = G \qquad (\mathbb{P}(X \leq Y) = 0.5 - \mathbb{P}(Y \leq X))$$
$$H_1 : F \neq G \qquad (\mathbb{P}(X \leq Y) \neq \mathbb{P}(Y \leq X)),$$

we can see that U *is a statistic that, under null, admits asymptotic normality:*

$$\sqrt{\frac{12mn}{m+n}}(U - \frac{1}{2}) \xrightarrow{d} \mathcal{N}(0, 1)$$

5 Degenerate U-statistics

We will now look at one last implication of U-statistics. Note that by default, in Theorem 3, we relied on the fact that $\mathbb{E}[h_1^2] > 0$. It is not always the case that the U-statistic will satisfy this condition. In fact, it is quite easy to construct a U-statistic with $\mathbb{E}[h_1^2] = 0$. We will refer to such statistics as (first-order) degenerate U-statistics.

Example 8 (Squared-mean estimation)**.** *Recall that for* $\theta = \mathbb{E}[X]^2 = \mu^2$*, we established the U-statistic with a second-order kernel,*

$$U = \frac{1}{n(n-1)}\sum_{i<j} X_i X_j.$$

Now, lets observe that $h_1(x_1) = \mathbb{E}[x_1 X_2] = x_1\mu$*. Furthermore,* $\sigma_1^2 = \mathbb{V}(h_1(x_1)) = \mu^2\mathbb{V}(X_2) = \mu^2\sigma^2$*. Then, by Theorem 3, we have*

$$\sqrt{n}(U_n - \mu^2) \xrightarrow{d} \mathcal{N}(0, 4\mu^2\sigma^2).$$

Now, if $\mu = 0$*, the limiting distribution is degenerate with 0 variance.*

If this is the case, then the result of Theorem 3 is no longer valid as it gives us a degenerate distribution.

Thankfully, there are still ways in which we can make meaningful progress to understand the behavior of such statistics. Instead of asymptotic normality, we can show convergence to a different distribution. The key here lies in the fact that $\mathbb{E}[h_1^2] = 0$ implies that the first term of the Hoeffding decomposition is exactly 0. Thus, the natural step would be to look at the second term, which will be come the dominant term and try to understand its asymptotic behaviour. Through this analysis, we can establish the following convergence results:

Theorem 4 (Asymptotic convergence of degenerate U-statistics)
Suppose $\mathbb{E}[h^2] < \infty$ *and that U is a first-order degenerate U-statistic (i.e.,* $\mathbb{V}(X_1X_2) > 0$*). Then,*

$$nU_n \xrightarrow{d} \sum_{k=1}^{\infty} \gamma_k(Z_k^2 - 1),$$

where $Z_k \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ *and* γ_k *are the eigenvalues from the decomposition of h.*

Note: *Using the given assumptions, it is possible to should that* $\sum \gamma_k < \infty$ *and so we don't need to impose this additional assumption.*

Note that the eigen-decomposition of h takes the form

$$h(x_1, x_2) = \sum_{k=1}^{\infty} \gamma_k \phi_k(x_1)\phi_k(x_2),$$

with γ_k being real-values and $\{\phi_k\}_k$ forming an orthonormal sequence, called the eigenfunctions.