Reading: VDV Chapter 12

Outline:

- U-Statistics (VDV Chapter 12)
  - Definitions
  - Examples
  - Variance calculation

1 U-Statistics

1.1 Definitions

Suppose I have \( h : X^r \to \mathbb{R} \) and want to estimate \( \theta = E[h(X_1, ..., X_r)] \), where the \( X_i \) are independent. Given a sample \((X_1, ..., X_n)\), how should I estimate \( \theta \)?

**Example:**

Observe that

\[
\text{Var}(X) = E[X_1^2] - E[X_1X_2] = \frac{1}{2} E[(X_1 - X_2)^2].
\]

So,

\[ h(X_1, X_2) = \frac{1}{2} (X_1 - X_2)^2 \]

**Remark** Without loss of generality, we assume \( h \) is symmetric, i.e it is invariant under any permutation of its arguments.

I should estimate \( \theta \) with with U-Statistics (Hoeffding 1940s). It allows us to
(1) abstract away annoying details and still perform inference, and
(2) develop statistics and tests that do not depend on parametric assumptions (non-parametric) making our inference more "robust".

**Definition 1.1** (U-Statistics). For \( X_i \overset{i.i.d.}{\sim} P \), denote \( \theta(P) = E_P[h(X_1, ..., X_r)] \). A U-statistic is a random variable of the form

\[
U_n := \frac{1}{\binom{n}{r}} \sum_{|\beta| = r, \beta \subseteq \{1,...,n\}} h(X_\beta)
\]
where \( h : X^r \to \mathbb{R} \) is a symmetric (kernel) function, \( \beta \) ranges over all size \( r \) subsets of \([n] := \{1, \ldots, n\}\), and \( X_\beta := (X_{i_1}, \ldots, X_{i_r}) \) for \( \beta = (i_1, \ldots, i_r) \).

**Remark** The U in "U-statistics" is because \( \mathbb{E}_P[U_n] = \theta(P) = \mathbb{E}[h(X_1, \ldots, X_r)] \), so \( U_n \) is unbiased.

Why use a U-statistic at all? Why not use 
\[
h(X_1, X_2, \ldots, X_r)
\]

or
\[
\frac{1}{\binom{n}{r}} \sum_{\ell=1}^{\binom{n}{r}} h(X_{\ell(r-1)+1}, \ldots, X_{\ell r})?
\]

Let \( \{X_1, \ldots, X_n\} \) be the sample with “index” information removed. (e.g. Order Statistics. Generally a histogram. In EE terminology, called “type” of the sample.) Then, under \( X_i \overset{i.i.d.}{\sim} P \), \( \{X(i)\}_{i=1}^n \) is a sufficient statistic. Observe that
\[
\mathbb{E}\{h(X_1, \ldots, X_r) | X_1, \ldots, X_n\} = U_n := \frac{1}{\binom{n}{r}} \sum_{|\beta|=r, \beta \subseteq [n]} h(X_\beta)
\]

By Rao-Blackwellization, we know that for any convex (loss) function \( L \) and any r.v. \( Z_n \) such that \( \mathbb{E}[Z_n(X(i))_{1\leq i \leq n}] = U_n \),
\[
\mathbb{E}[L(Z_n)] \geq \mathbb{E}[L(U_n)].
\]

### 1.2 Examples

**Example** (Sample Variance): Consider \( h(x, y) = \frac{1}{2} (x - y)^2 \). Then \( \mathbb{E}[h(X_1, X_2)] = \frac{1}{2} (\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2]) - \mathbb{E}[X_1, X_2] = \text{Var}(X) \). When we have more than two samples, we use

\[
U_n = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \frac{1}{2} (X_i - X_j)^2
\]

\[
= \frac{1}{2n(n-1)} \sum_{i,j} (X_i - X_j)^2
\]

\[
= \frac{1}{2n(n-1)} \sum_{i,j} ((X_i - \bar{X}_n) - (X_j - \bar{X}_n))^2
\]

\[
= \frac{1}{2n(n-1)} \sum_{i,j} ((X_i - \bar{X}_n)^2 + (X_j - \bar{X}_n)^2)
\]

\[
= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2
\]

♣

**Example** (Gini’s Mean-Difference): \( h(x, y) = |x - y| \) and \( \mathbb{E}[U_n] = \mathbb{E}[|X_1 - X_2|] \). ♣

**Example** (Quantiles):
\[ \theta(P) = P(X \leq t) = \int_{-\infty}^{t} dp \] and \( h(X) = 1 \{ X \leq t \} \)

This is a first order U-statistic.

**Example** (Signed Rank Statistic): Suppose we want to know whether the central location of \( P \) is 0. Then we can use

\[ \theta(P) := P(X_1 + X_2 > 0), \]

even when \( \mathbb{E}[X] \) isn’t well-defined.

This means \( h(x,y) = 1 \{ x + y > 0 \} \) and \( U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} 1 \{ X_i + X_j > 0 \} \).

**Definition 1.2** (Two-sample U-Statistic). Given two samples \( \{X_1, ..., X_n\} \) and \( \{Y_1, ..., Y_n\} \), a two-sample U-statistic is a random variable of the form

\[ U = \frac{1}{\binom{n}{r} \binom{m}{s}} \sum_{|\alpha|=s, \alpha \subseteq [m]} \sum_{|\beta|=r, \beta \subseteq [n]} h(X_{\beta}, Y_{\alpha}) \]

where \( h : X^r \times Y^s \to \mathbb{R} \). \( h \) is symmetric in its first \( r \) arguments and in its last \( s \) arguments.

**Example** (Mann-Whitney Statistic): Do \( X \) and \( Y \) have the same location? We can consider

\[ \theta(P) = P(X \leq Y), \]

\[ h(X,Y) = 1 \{ X \leq Y \}, \]

\[ U_{n,m} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} 1 \{ X_i \leq Y_j \}, \]

which should be close to \( \frac{1}{2} \) when \( X \) and \( Y \) have the same location.

**Example:** Here’s another motivating example for two-sample U-statistics. Suppose we have \( X_i \overset{i.i.d.}{\sim} P \) and \( Y_i \overset{i.i.d.}{\sim} Q \). Are \( P \) and \( Q \) different?

The null in this two-sample problem is: \( P = Q \). This is a huge null: \( P \) is unknown and could be anything. We approximate the null by looking at the distribution of \( h(Z_A) \), where \( Z = \{X_1, ..., X_n, Y_1, ..., Y_n\} \) and \( A \) ranges over all possible index sets of size \( |A| = r + s \). We use that under the null,

\[ h(Z_A) \overset{\text{dist}}{=} h(Z_B) \]

for any \( A, B \in [n] \) such that \( |A| = |B| = r + s \).
1.3 Variance of U-Statistics

This is a precursor to asymptotic normality because "1st order terms" dominate everything else.

**Definition 1.3.** Assume that $E \left[ \left| h \right|^2 \right] < \infty$ for any $c < r$. Define

$$h_c(X_1,...,X_c) := E \left[ h \left( X_1,...,X_c,X_{c+1},...,X_r \right) \right].$$

**Remark**

1. $h_0 = E \left[ h(X_1,...,X_r) \right] = \theta(P)$
2. $E \left[ h_c(X_1,...,X_c) \right] = E \left[ h(X_1,...,X_r) \right] = \theta(P)$

**Definition 1.4.**

$$\hat{h}_c := h_c - E \left[ h_c \right] = h_c - \theta(P)$$

$$E \left[ \hat{h}_c \right] = 0$$

Then define

$$\zeta_c := \text{Var}(h_c(X_1,...,X_c)) = E \left[ \hat{h}_c^2 \right]$$

(Note that $\zeta_0 = 0$.)

**Goal:** Write $\text{Var}[U_n]$ in terms of $\zeta'_c$s for $c=1,2,...,r$.

**Lemma 1.** If $\alpha, \beta \subseteq [n]$, $S = \alpha \cap \beta$, $c = |S|$, then

$$E \left[ \hat{h}(X_\alpha)\hat{h}(X_\beta) \right] = \zeta_c.$$

**Proof** Using the symmetry of $h$,

$$E \left[ \hat{h}(X_\alpha)\hat{h}(X_\beta) \right] = E \left[ \hat{h}(X_\alpha \setminus S,X_S)\hat{h}(X_\beta \setminus S,X_S) \right]$$

$$= E \left[ E[\hat{h}(X_\alpha \setminus S,X_S) | X_S] \cdot E[\hat{h}(X_\beta \setminus S,X_S) | X_S] \right]$$

(since $X_\alpha \setminus S$, $X_\beta \setminus S$ indep.)

$$= E \left[ \hat{h}_c(X_S) \cdot \hat{h}_c(X_S) \right]$$

$$= \zeta_c.$$

**Theorem 2.** Let $U_n$ be an $r^{th}$ order U-statistic. Then

$$\text{Var}U_n = \frac{r^2}{n} \zeta_1 + O(n^{-2}).$$
There are $\binom{n}{r} \binom{n-r}{r-c}$ ways to select a pair of subsets of $[n]$, each of size $r$, with $c$ common elements. Hence,

$$U_n - \theta = \binom{n}{r}^{-1} \sum_{|\beta|=r} \hat{h}(X_\beta),$$

$$\text{Var} U_n = \binom{n}{r}^{-2} \sum_{|\alpha|=r} \sum_{|\beta|=r} \mathbb{E} \left[ \hat{h}(X_\alpha) \hat{h}(X_\beta) \right]$$

$$= \binom{n}{r}^{-2} \sum_{c=1}^{r} \binom{n}{r} \binom{r}{c} \binom{n-r}{r-c} \zeta_c$$

$$= \sum_{c=1}^{r} \frac{r!^2}{c!(r-c)!^2} \frac{(n-r)(n-r-1) \ldots (n-2r+c+1)}{n(n-1) \ldots (n-r+1)} \zeta_c.$$

For fixed $c$, $\frac{(n-r)(n-r-1) \ldots (n-2r+c+1)}{n(n-1) \ldots (n-r+1)}$ has $r-c$ terms in the numerator and $r$ terms in the denominator. Hence,

$$\text{Var} U_n = r^2 \frac{(n-r)(n-r-1) \ldots (n-2r+2)}{n(n-1) \ldots (n-r+1)} \zeta_1 + \sum_{c=2}^{r} \frac{O \left( \frac{n^{r-c}}{n^r} \right)}{n^2} \zeta_c$$

$$= r^2 \left[ \frac{1}{n} + O(n^{-2}) \right] \zeta_1 + O(n^{-2})$$

$$= \frac{r^2}{n} \zeta_1 + O(n^{-2}).$$

With this theorem, we know that the variance of U-statistics behaves like the variance of a sample mean plus high-order errors.

**New Goal:** Show that $U_n$ is asymptotically normal by projecting out all high-order interactions.