Stat210B: Theoretical Statistics **Lecture Date: January 23, 2007**

Lecture 3

Lecturer: Michael I. Jordan *Scribe: Sahand Negahban*

1 U-statistics

U-statistics are a useful tool. The beauty of the U-statistics framework is that by abstracting away some details, can have a general representation of various meaningful quantities. The theory of U-statistics was initially developed by Hoeffding, one of the pioneers of non-parametric statistics.

Definition 1 (U-statistic). Let $X_i \overset{\text{i.i.d.}}{\sim} F$, $h(x_1, x_2, \ldots, x_r)$ be a symmetric kernel function, and $\theta(F) = E[h(X_1, X_2, \ldots, X_r)]$. A U-statistic U_n is defined as

$$U_n = \frac{1}{\binom{n}{r}} \sum_{\beta} h(X_1, X_2, \ldots, X_r) \tag{1}$$

where β ranges over all subsets of size r chosen from $\{1, 2, \ldots, n\}$. $E[U_n] = \theta(F)$ (i.e. U-statistics are unbiased).

Example 2 (Sample Variance). Let $\theta(F) = \sigma^2 = \int (X - \mu)^2 dF$ where $\mu = \int x dF(x)$.

$$\begin{aligned}
\theta(F) &\overset{(a)}{=} \int \left(x_1 - \int x_2 dF(x_2) \right)^2 dF(x_1) \\
&= \int x_1^2 dF(x_1) - 2 \int x_1 dF(x_1) \int x_2 dF(x_2) + \left(\int x_2 dF(x_2) \right)^2 \\
&= \int x_1^2 dF(x_1) - \left(\int x_2 dF(x_2) \right)^2 \\
&= \frac{1}{2} \int x_1^2 dF(x_1) + \frac{1}{2} \int x_2^2 dF(x_2) - \int x_1 x_2 dF(x_1) dF(x_2) \\
&= \frac{1}{2} \int (x_1 - x_2)^2 dF(x_1) dF(x_2) \\
&\Rightarrow h(X_1, X_2) = \frac{1}{2}(X_1 - X_2)^2.
\end{aligned}$$

Where (a) follows by expanding μ to $\int x_2 dF(x_2)$. Thus, the U-statistic for the variance is

$$\begin{aligned}
U_n &= \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \frac{1}{2}(X_i - X_j)^2 \\
&= \frac{1}{n(n-1)} \frac{1}{2} \sum_i \sum_j (X_i - X_j)^2
\end{aligned}$$

where the last equality follows because taking the sum over all indices results in double counting the $(X_i - X_j)^2$ terms. Continuing the simplification shows that

$$\begin{aligned} U_n &= \frac{1}{2n(n-1)} \sum_i \sum_j \left[(X_i - \bar{X}) - (X_j - \bar{X})\right]^2 \\ &= \frac{1}{2n(n-1)} \sum_i \sum_j (X_i - \bar{X})^2 + (X_j - \bar{X})^2 \\ &= \frac{1}{n-1} \sum_i (X_i - \bar{X})^2 \\ &= s_n^2. \end{aligned}$$

Thus, s_n^2 is the U-statistic for the variance of a set of samples. Unbiasedness of this statistic follows immediately from the unbiasedness of U-statistics.

1.1 Novel U-statistics

Example 3 (Gini's mean difference).

$$\theta(F) = \int |x_1 - x_2| dF(x_1) dF(x_2) \tag{2}$$

and the corresponding U-statistic is

$$U_n = \frac{2}{n(n-1)} \sum_{i<j} |X_i - X_j|. \tag{3}$$

Example 4 (Quantile statistic.).

$$\theta(F) = \int_{-\infty}^{t} dF(x). \tag{4}$$

$$U_n = \frac{1}{n} \sum 1_{X_i \le t} = F_n(t) \tag{5}$$

where

$$h(x) = 1_{x \le t}. \tag{6}$$

Example 5 (Signed rank statistic). The following statistic can be used in testing whether the location of the samples is 0.

$$\theta(F) = P(X_1 + X_2 > 0) \tag{7}$$

$$U_n = \frac{2}{n(n-1)} \sum_{i<j} 1_{X_1 + X_2 > 0} \tag{8}$$

where

$$h(x_1, x_2) = 1_{x_1 + x_2 > 0} \tag{9}$$

Definition 6 (Two-sample U-statistics). Given $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ define

$$U_n = \frac{1}{\binom{m}{r}\binom{n}{s}} \sum_\alpha \sum_\beta h(\overbrace{\underbrace{X_{\alpha_1}, \ldots, X_{\alpha_r}}_{\text{symmetric}}, \underbrace{Y_{\beta_1}, \ldots, Y_{\beta_s}}_{\text{symmetric}}}^{\text{not symmetric}}) \tag{10}$$

where $h(\cdot, \cdot)$ is symmetric in $x_1, \ldots, x_r$ and $y_1, \ldots, y_s$, but not across both sets of inputs.

Example 7 (Mann-Whitney statistic). This statistic is "used to test for a difference in location between the two samples" (van der Vaart, 1998).

$$U_n = \frac{1}{n_1 n_2} \sum_i \sum_j 1_{X_i \leq Y_j} \tag{11}$$

1.2 Variance of U-statistics

The analysis was first done by Hoeffding.
Assume $E[h] < \infty$ and $X_i \overset{\text{i.i.d}}{\sim} F$. Define $h_c(x_1, \ldots, x_c)$ for $c < r$ as

$$h_c(x_1, \ldots, x_c) \quad = E[h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_r)] \tag{12}$$

Remark 8. The following facts follow from the above definition:

1. $h_0 = \theta(F)$
2. $E[h_c(X_1, \ldots, X_c)] = E[h(X_1, \ldots, X_c, X_{c+1}, \ldots, X_r)] = \theta(F)$.

Let $\widehat{h}_c = h_c - E[h_c] = h_c - \theta(F)$, which follows from remark 8. Thus, $E[\widehat{h}_c] = 0$. Define ζ_c as

$$\zeta_c = \text{Var}(h_c(X_1, \ldots, X_c)) = E[\widehat{h}_c^2(X_1, \ldots, X_c)]. \tag{13}$$

Let $B = \{\beta_1, \ldots, \beta_r\}$ and $B' = \{\beta'_1, \ldots, \beta'_r\}$ be two subsets of $\{1, \ldots, n\}$. Let c be the number of integers in common between each of the sets. Let $S = B \cap B'$, $S_1 = S \setminus B$, and $S_2 = S \setminus B'$, which implies that $|S| = c$, and $|S_1| = |S_2| = r - c$. For some subset $A = \{\alpha_1, \ldots, \alpha_r\}$ of $\{1, \ldots, n\}$ let $X_A = \{X_{\alpha_1}, \ldots, X_{\alpha_r}\}$. Thus,

$$\begin{aligned}
E[\widehat{h}(X_{\beta_1}, \ldots, X_{\beta_r})\widehat{h}(X_{\beta'_1}, \ldots, X_{\beta'_r})] &= E[\widehat{h}(X_B)\widehat{h}(X_{B'})] && (14)\\
&= E[\widehat{h}(X_S, X_{S_1})\widehat{h}(X_S, X_{S_2})] && (15)\\
&= E[E[\widehat{h}(X_S, X_{S_1})\widehat{h}(X_S, X_{S_2})|X_S]] && (16)\\
&= E[\widehat{h}_c^2(X_S)] && (17)\\
&= \zeta_c \ , && (18)
\end{aligned}$$

where the second equality follows because h is a symmetric kernel function and the third and fourth equalities follow from iterated expectations, the fact that each X_i is i.i.d., and the definition of h_c.
Remark 9. The number of distinct choices for two sets having c elements in common is

$$\binom{n}{r}\binom{r}{c}\binom{n-r}{r-c} \tag{19}$$

From the definitions

$$U_n - \theta(F) = \frac{1}{\binom{n}{r}} \sum_\beta \widehat{h}(X_{\beta_1}, \ldots, X_{\beta_r}). \tag{20}$$

Thus,

$$\begin{aligned}
\text{Var}(U_n) &= \binom{n}{r}^{-2} \sum_\beta \sum_{\beta'} E[\widehat{h}(X_{\beta_1}, \ldots, X_{\beta_r})\widehat{h}(X_{\beta'_1}, \ldots, X_{\beta'_r})] && (21a)\\
&= \binom{n}{r}^{-2} \sum_{c=0}^{r} \binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}\zeta_c \quad \text{and } \zeta_c = 0 && (21b)\\
&= \sum_{c=1}^{r} \frac{r!^2}{c!(r-c)!^2} \frac{(n-r)(n-r-1)\cdots(n-2r+c+1)}{n(n-1)\cdots(n-r+1)}\zeta_c, && (21c)
\end{aligned}$$

where the term in the summation corresponding to some c is $O(\frac{1}{n^c})$. Thus,

$$\text{Var}(U_n) = O(\frac{1}{n}) + O(\frac{1}{n^2}) + \cdots + O(\frac{1}{n^r}). \tag{22}$$

Example 10 (Sampling variance of the variance). Let $\theta(F) = \sigma^2$. Thus by example 2

$$h(X_1, X_2) = \frac{1}{2}(X_1 - X_2)^2 \text{ and } r = 2. \tag{23}$$

Therefore,

$$\widehat{h}(X_1, X_2) = \frac{1}{2}(X_1 - X_2)^2 - \sigma^2,$$

$$h_1(X_1) = \frac{1}{2}(X_1^2 - 2X_1\mu + \sigma^2 + \mu^2),$$

so

$$\widehat{h}_1(X_1) = \frac{1}{2}((X_1 - \mu)^2 - \sigma^2)$$

$$\begin{aligned} E[h^2(X_1, X_2)] &= \frac{1}{4}E[((X_1 - \mu) - (X_2 - \mu))^4] \\ &= \frac{1}{4}\sum_{j=0}^{4}\binom{4}{j}E[(X_1 - \mu)^j]E[(X_2 - \mu)^{4-j}] \\ &= \frac{1}{4}(2\mu_4 + 6\sigma^4), \end{aligned}$$

where the final equality follows because $E[(X - \mu)^4] = \mu_4$, $E[(X - \mu)^2] = \sigma^2$, and $E[(X - \mu)] = 0$. Thus, the following equalities follow:

$$\begin{aligned} \zeta_2 &= E[h^2] - \sigma^4 = \frac{\mu_4}{2} + \frac{\sigma^4}{2} \\ \zeta_1 &= E[\widehat{h}_1^2] = \frac{1}{4}\text{Var}((X_1 - \mu)^2) = \frac{1}{4}(\mu_4 - \sigma^4). \end{aligned}$$

Applying equation 21 yields:

$$\begin{aligned} \text{Var}(s_n^2) &= \binom{n}{2}^{-1}(2(n-2)\zeta_1 + \zeta_2) \\ &= \frac{2}{n(n-1)}\left[2(n-1)\zeta_1 - 2\zeta_1 + \zeta_2\right] \\ &= \frac{4\zeta_1}{n} - \frac{4\zeta_1}{n(n-1)} + \frac{2\zeta_2}{n(n-1)} \\ &= \frac{\mu_4 - \sigma^4}{n} + \frac{2\sigma^4}{n(n-1)} = \frac{\mu_4 - \sigma^4}{n} + O(n^{-2}). \end{aligned}$$

The variance is asymptotically the same as what was found in homework 1. However, using the above method also gives the exact value of all higher order terms.

The variance of U-statistics is known, however the question of whether or not U-statistics are asymptotically normal has yet to be answered. *Hájek projections* will help prove that U-statistics do indeed asymptotically go to Gaussians.

References

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