U-STATISTICS

Notes for Statistics 200B, Winter 2003

Thomas S. Ferguson

1. Definitions. The basic theory of U-statistics was developed by W. Hoeffding (1948a). Detailed expositions of the general topic may be found in M. Denker (1985) and A. J. Lee (1990). See also Fraser (1957) Chapter 6, Serfling (1980) Chapter 5, and Lehmann (1999), Chapter 6.

Let $\mathcal{P}$ be a family of probability measures on an arbitrary measurable space. The problems treated here are nonparametric, which means that $\mathcal{P}$ will be taken to be a large family of distributions subject only to mild restrictions such as continuity or existence of moments. Let $\theta(P)$ denote a real-valued function defined for $P \in \mathcal{P}$. The first notion we need is that of an estimable parameter. (Hoeffding called these regular parameters.)

Definition 1. *We say that $\theta(P)$ is an* **estimable parameter** *within $\mathcal{P}$, if for some integer m there exists an unbiased estimator of $\theta(P)$ based on m i.i.d. random variables distributed according to P; that is, if there exists a real-valued measurable function $h(x_1, \ldots, x_m)$ such that*

$$\mathrm{E}_P(h(X_1, \ldots, X_m)) = \theta(P) \qquad \textit{for all } P \in \mathcal{P}, \tag{1}$$

when $X_1, \ldots, X_m$ are i.i.d. with distribution P. The smallest integer m with this property is called the **degree** *of $\theta(P)$.*

It should be noted that the function h may be assumed to be a symmetric function of its arguments. This is because if f is an unbiased estimator of $\theta(P)$, then the average of f applied to all permutations of the variables is still unbiased and is, in addition, symmetric. That is,

$$h(x_1, \ldots, x_m) = \frac{1}{m!} \sum_{\pi \in \Pi_m} f(x_{\pi_1}, \ldots, x_{\pi_m}), \tag{2}$$

where the summation is over the group Π_m of all permutations of an m-vector, is obviously symmetric in its arguments, and has the same expectation under P as does f.

Definition 2. *For a real-valued measurable function, $h(x_1, \ldots, x_m)$ and for a sample, $X_1, \ldots, X_n$, of size $n \geq m$ from a distribution P, a* **U-statistic** *with* **kernel** *h is defined as*

$$U_n = U_n(h) = \frac{(n-m)!}{n!} \sum_{\mathbf{P}_{m,n}} h(X_{i_1}, \ldots, X_{i_m}) \tag{3}$$

where the summation is over the set $\mathbf{P}_{m,n}$ of all $n!/(n-m)!$ permutations $(i_1, i_2, \ldots, i_m)$ of size m chosen from $(1, 2, \ldots, n)$. If the kernel, h, is symmetric in its arguments, U_n has the equivalent form

$$U_n = U_n(h) = \frac{1}{\binom{n}{m}} \sum_{\mathbf{C}_{m,n}} h(X_{i_1}, \ldots, X_{i_m}) \tag{4}$$

where the summation is over the set $\mathbf{C}_{m,n}$ of all $\binom{n}{m}$ combinations of m integers, $i_1 < i_2 < \ldots < i_m$ chosen from $(1, 2, \ldots, n)$.

If $\theta(P) = \mathrm{E}_P h(X_1, \ldots, X_m)$ exists for all $P \in \mathcal{P}$, then an obvious property of the U-statistic, U_n, is that it is an unbiased estimate of $\theta(P)$. Moreover it has the optimality property of being a best unbiased estimate of $\theta(P)$ if $\mathcal{P}$ is large enough, for example if it contains all distributions, P, for which $\theta(P)$ is finite. Then the order statistics form a complete sufficient statistic from $P \in \mathcal{P}$. And U_n, being a symmetric function of $X_1, \ldots, X_n$, is a function of the order statistics, and so is a best unbiased estimate of its expectation, due to the Hodges-Lehmann theorem. This means, for example, that no unbiased estimate of $\theta(P)$, based on $X_1, \ldots, X_n$, can have a variance smaller than the variance of U_n. We do not deal further with this subject since our interest here is in the asymptotic distribution of U_n.

2. Examples. *1. Moments.* If $\mathcal{P}$ is the set of all distributions on the real line with finite mean, then the mean, $\mu = \mu(P) = \int x\, dP(x)$, is an estimable parameter of degree $m = 1$, because $f(X_1) = X_1$ is an unbiased estimate of μ. The corresponding U-statistic is the sample mean, $U_n = \overline{X}_n = (1/n)\sum_1^n X_i$. Similarly, if $\mathcal{P}$ is the set of all distributions on the real line with finite kth moment, then the kth moment, $\mu_k = \int x^k\, dP(x)$ is an estimable parameter of degree 1 with U-statistic, $(1/n)\sum_1^n X_i^k$.

How about estimating the square of the mean, $\theta(P) = \mu^2$? Since $\mathrm{E}(X_1 X_2) = \mu^2$, it is also an estimable parameter with degree at most 2. It is easy to show it cannot have degree 1 (Exercise 1), so it has degree 2. The U-statistic U_n of (3) and (4) corresponding to $h(x_1, x_2) = x_1 x_2$ is

$$U_n = \frac{1}{n(n-1)} \sum_{i \neq j} X_i X_j = \frac{2}{n(n-1)} \sum_{i<j} X_i X_j. \tag{5}$$

If $\mathcal{P}$ is taken to be the set of all distributions on the real line with finite second moment, then the variance, $\sigma^2 = \mu_2 - \mu^2$, is also estimable of degree 2, since we can estimate μ_2 by X_1^2 and μ^2 by $X_1 X_2$:

$$\mathrm{E}(X_1^2 - X_1 X_2) = \sigma^2. \tag{6}$$

However the kernel, $f(x_1, x_2) = x_1^2 - x_1 x_2$, is not symmetric in x_1 and x_2. The corresponding symmetric kernel given by (2) is the average,

$$h(x_1, x_2) = \frac{1}{2}(f(x_1, x_2) + f(x_2, x_1)) = \frac{x_1^2 - 2x_1 x_2 + x_2^2}{2} = \frac{(x_1 - x_2)^2}{2}. \tag{7}$$

This leads to the U-statistic,

$$\begin{aligned} U_n &= \frac{2}{n(n-1)} \sum_{i<j} \frac{(X_i - X_j)^2}{2} \\ &= \cdots = s_x^2 = \frac{1}{n-1} \sum_1^n (X_i - \overline{X})^2. \end{aligned} \tag{8}$$

This is the unbiased sample variance.

It is easy to see that any linear combination of estimable parameters is estimable, and any product of estimable parameters is estimable (Exercise 2). Thus there are U-statistics for estimating all moments and all cumulants. (The cumulants are the coefficients of $(it)^k/k!$ in the power series expansion of $\log \phi(t)$, the logarithm of the characteristic function. They are polynomial functions of the moments.)

In the definition of estimable parameter and its corresponding U-statistic, no restriction is made on the space on which the distributions must lie. Thus each $P \in \mathcal{P}$ could be a distribution on the plane or in d-dimensions, and then the corresponding observations would be random vectors. One can construct U-statistics for estimating a covariance (Exercise 3) and higher cross moments.

2. The Wilcoxon Signed Rank Test. Let $\mathcal{P}$ be the family of continuous distributions on the real line. Consider the problem of testing the hypothesis, H_0, that the true distribution, P, is symmetric about the origin based on a sample $Z_1, \ldots, Z_n$ from P. (This problem arises most naturally from a paired comparison experiment based on random variables, (X_i, Y_i), when $Z_i = X_i - Y_i$. The hypothesis that X_i and Y_i are independent identically distributed leads to the hypothesis that Z_i is distributed symmetrically about the origin.)

Of course the sign test (reject H_0 if the number of positive Z_i is too large) can be used in this problem as a quick and dirty test, but if you have more time, a better choice is the Wilcoxon signed rank test. This test is based on the statistic

$$W_n^+ = \sum_{i=1}^n R_i^+ I(Z_i > 0) \tag{9}$$

where R_i^+ is the rank of $|Z_i|$ among $|Z_1|$, $|Z_2|$, $\ldots$, $|Z_n|$. Although it is not a U-statistic, one can show (Exercise 4) that W_n^+ is a linear combination of two U-statistics,

$$W_n^+ = \sum_i I(Z_i > 0) + \sum_{i<j} I(Z_i + Z_j > 0). \tag{10}$$

and writing it in this way gives some insight into its behavior. The first U-statistic is based on the kernel, $h(z) = I(z > 0)$. The U-statistic itself is $U_n^{(1)} = n^{-1} \sum_1^n I(Z_i > 0)$. This is the U-statistic used for the sign test. The second U-statistic is based on the kernel, $h(z_1, z_2) = I(z_1 + z_2 > 0)$, and the corresponding U-statistic is $U_n^{(2)} = \binom{n}{2}^{-1} \sum_{i<j} I(Z_i + Z_j > 0)$. Thus,

$$W_n^+ = nU_n^{(1)} + \binom{n}{2} U_n^{(2)}. \tag{11}$$

For large n the second term dominates the first, so asymptotically W_n^+ behaves like $n^2 U_n^{(2)}/2$. The Wilcoxon signed rank test rejects H_0 if W_n^+ is too large, and this is asymptotically equivalent to the test that rejects if $U_n^{(2)}$ is too large.

3. Testing Symmetry. In some situations, it is important to test for symmetry about an unknown center. Here is one method based of the observation that for a sample of size 3, X_1, X_2, X_3 from a continuous distribution, symmetric about a point ξ, $P(X_1 > (X_2 + X_3)/2) = P((X_1 - \xi) > ((X_2 - \xi) + (X_3 - \xi))/2) = 1/2$. Because of this, $f(X_1, X_2, X_3) = \mathrm{sgn}(2X_1 - X_2 - X_3)$ is an unbiased estimate of $\theta(P) = P(2X_1 > X_2 + X_3) - P(2X_1 < X_2 + X_3)$. Here, $\mathrm{sgn}(x)$ represents the sign function, which is 1 if $x > 0$, 0 if $x = 0$ and -1 if $x < 0$. When P is symmetric, $\theta(P)$ has value zero . The corresponding symmetric kernel is

$$h(x_1, x_2, x_3) = \frac{1}{3}[\mathrm{sgn}(2x_1 - x_2 - x_3) + \mathrm{sgn}(2x_2 - x_1 - x_3) + \mathrm{sgn}(2x_3 - x_1 - x_2)]. \quad (12)$$

This is an example of a kernel of degree 3. The hypothesis of symmetry is rejected if the corresponding U-statistic is too large in absolute value. One can easily show that

$$h(x_1, x_2, x_3) = \frac{1}{3}\mathrm{sgn}(\mathrm{median}(x_1, x_2, x_3) - \mathrm{mean}(x_1, x_2, x_3)). \quad (13)$$

Thus the validity of the test also follows from the observation that for a sample of size three from a symmetric distribution, the sample median is equally likely to be above the sample mean as below it.

4. Measures of Association. For continuous probability distributions in 2-dimensions, there are several measures of dependence, or association, the simplest of which is perhaps Kendall's tau. Two vectors (x_1, y_1) and (x_2, y_2), are said to be concordant if $x_1 < x_2$ and $y_1 < y_2$, or if $x_2 < x_1$ and $y_2 < y_1$; in other words, if the line joining the points has positive slope. If the line joining the points has negative slope, the points are said to be discordant.

Suppose (X_1, Y_1) and (X_2, Y_2) are independently distributed according to a distribution $F(x, y)$ in the plane. If the probability of concordance, $P(X_1 < X_2, Y_1 < Y_2) + P(X_2 < X_1, Y_2 < Y_1)$ is bigger than 1/2, there is a positive association between X and Y. If it is negative, there is negative association. This leads to a measure of association called Kendall's τ, defined as

$$\tau = 2[P(X_1 < X_2, Y_1 < Y_2) + P(X_2 < X_1, Y_2 < Y_1)] - 1 = 4\,\mathrm{P}(X_1 < X_2, Y_1 < Y_2) - 1. \quad (14)$$

Kendall's tau behaves like a correlation coefficient in that $-1 \le \tau \le 1$, $\tau = 0$ when X and Y are independent, and $\tau = +1$, (resp. $\tau = -1$), if an increase in X almost surely implies and increase (resp. decrease) in Y. The definition of Kendall's tau shows that it is an estimable parameter with kernel, $f((x_1, y_1), (x_2, y_2)) = 4\,\mathrm{I}(x_1 < x_2, y_1 < y_2) - 1$ of degree two, and a corresponding symmetric kernel, $h((x_1, y_1), (x_2, y_2)) = 2\,\mathrm{I}(x_1 < x_2, y_1 < y_2) + 2\,\mathrm{I}(x_2 < x_1, y_2 < y_1) - 1$. The corresponding U-statistic,

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} h((X_i, Y_i), (X_j, Y_j)), \quad (15)$$

is known as Kendall's coefficient of rank correlation. This was seen in Exercise 5.7 of Ferguson (1996) to have an asymptotically normal distribution, when suitably normalized, in the case where X and Y are independent. We will see that the asymptotic distribution is normal for general dependent X and Y.

Another measure of association in 2-dimensions is given by Spearman's rho, defined as

$$\rho = 12\,P(X_1 < X_2, Y_1 < Y_3) - 3, \tag{16}$$

where (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) are independently distributed according to F. It also has the properties of a correlation coefficient, being between zero and one and zero when the variables are independent. In fact, one can show that ρ is simply the correlation coefficient between the random variables $F(X, \infty)$ and $F(\infty, Y)$. It is clear that ρ is also an estimable parameter with kernel of degree 3, $h((x_1, y_1), (x_2, y_2), (x_3, y_3)) = 12\,\mathrm{I}(x_1 < x_2, y_1 < y_3) - 3$. The symmetrized version has 6 terms. The corresponding U-statistic is related to the rank statistic of Example 12.5 of Ferguson (1996), which was seen to have an asymptotically normal distribution under the hypothesis of independence.

3. The Asymptotic Distribution of U_n. For a given estimable parameter, $\theta = \theta(P)$, and corresponding symmetric kernel, $h(x_1, \ldots, x_m)$, we take $\mathcal{P}$ to be the class of distributions for which $\mathrm{Var}(h(X_1, \ldots, X_m)) < \infty$. Let us define a sequence of functions related to h. For $c = 0, 1, \ldots, m$, let

$$h_c(x_1, \ldots, x_c) = \mathrm{E}\,h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m) \tag{17}$$

where $X_{c+1}, \ldots, X_n$ are i.i.d. P. Then $h_0 = \theta$ and $h_m(x_1, \ldots, x_m) = h(x_1, \ldots, x_m)$. These functions are all have expectation θ,

$$\mathrm{E}\,h_c(X_1, \ldots, X_c) = \mathrm{E}\,h(X_1, \ldots, X_c, X_{c+1}, \ldots, X_m) = \theta, \tag{18}$$

but they cannot be called kernels since they may depend on P.

The variance of the U-statistic U_n of (4) depends on the variances of the h_c. For $c = 0, 1, \ldots, m$, let

$$\sigma_c^2 = \mathrm{Var}(h_c(X_1, \ldots, X_c)), \tag{19}$$

so that $\sigma_0^2 = 0$ and $\sigma_m^2 = \mathrm{Var}(h(X_1, \ldots, X_m))$.

To compute the variance of U_n of (4), we start out by

$$\begin{aligned} \mathrm{Var}(U_n) &= \mathrm{Var}\left(\binom{n}{m}^{-1} \sum_{\mathbf{i} \in \mathbf{C}_{m,n}} h(X_{i_1}, \ldots, X_{i_m})\right) \\ &= \binom{n}{m}^{-2} \sum_{\mathbf{i} \in \mathbf{C}_{m,n}} \sum_{\mathbf{j} \in \mathbf{C}_{m,n}} \mathrm{Cov}(h(X_{i_1}, \ldots, X_{i_m}), h(X_{j_1}, \ldots, X_{j_m})) \end{aligned} \tag{20}$$

The following lemma relates these covariances to the σ_c^2.

Lemma 1. *For $P \in \mathcal{P}$ and $(i_1, \ldots, i_m)$ and $(j_1, \ldots, j_m)$ in $\mathbf{C}_{m,n}$,*

$$\begin{aligned}&\mathrm{Cov}(h(X_{i_1}, \ldots, X_{i_m}), h(X_{j_1}, \ldots, X_{j_m})) \\ &\qquad = \mathrm{Cov}(h_c(X_1, \ldots, X_c), h(X_1, \ldots, X_m)) \\ &\qquad = \sigma_c^2,\end{aligned} \tag{21}$$

where c is the number of integers common to $(i_1, \ldots, i_m)$ and $(j_1, \ldots, j_m)$.

Proof. If $(i_1, \ldots, i_m)$ and $(j_1, \ldots, j_m)$ have c elements in common, then

$$\begin{aligned}&\mathrm{Cov}(h(X_{i_1}, \ldots, X_{i_m}), h(X_{j_1}, \ldots, X_{j_m})) \\ &\quad = \mathrm{E}[(h(X_1, \ldots, X_c, X_{c+1}, \ldots, X_m) - \theta)(h(X_1, \ldots, X_c, X'_{c+1} \ldots, X'_m) - \theta)),\end{aligned} \tag{22}$$

where $X_1, \ldots, X_m, X'_{c+1}, \ldots, X'_m$ are i.i.d. Conditionally, given $X_1, \ldots, X_c$, the two terms in this expectation are independent, so taking the expectation of the conditional expectation, we have

$$\begin{aligned}&\mathrm{Cov}(h(X_{i_1}, \ldots, X_{i_m}), h(X_{j_1}, \ldots, X_{j_m})) \\ &\qquad = \mathrm{E}[(h_c(X_1, \ldots, X_c) - \theta)(h_c(X_1, \ldots, X_c) - \theta)) \\ &\qquad = \sigma_c^2.\end{aligned} \tag{23}$$

The same argument shows $\mathrm{Cov}(h_c(X_1, \ldots, X_c), h(X_1, \ldots, X_m)) = \sigma_c^2$. ■

From this we see that $\sigma_c^2 \leq \sigma_m^2$ for all c because $\sigma_c^2 = \mathrm{Cov}(h_c, h) \leq \sigma_c \sigma_m$. The same argument shows that the σ_c^2 are nondecreasing: $\sigma_1^2 \leq \sigma_2^2 \leq \cdots \leq \sigma_m^2$.

Theorem 1. *For $P \in \mathcal{P}$,*

$$\mathrm{Var}(U_n) = \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2. \tag{24}$$

If $\sigma_m^2 < \infty$, then $\mathrm{Var}(U_n) \sim m^2 \sigma_1^2 / n$ for large n.

Proof. We continue (20) by separating out of the sum those terms with exactly c elements in common. The number of such pairs of m-tuples, $(i_1, \ldots, i_m)$ and $(j_1, \ldots, j_m)$, having exactly c elements in common is $\binom{n}{m}\binom{m}{c}\binom{n-m}{m-c}$, because there are $\binom{n}{m}$ ways of choosing $i_1, \ldots, i_m$, and then $\binom{m}{c}$ ways of choosing a subset of size c from them, and finally $\binom{n-m}{m-c}$ ways of choosing the remaining $m - c$ elements of $j_1, \ldots, j_m$ from the remaining $n - m$ numbers. Therefore,

$$\begin{aligned}\mathrm{Var}(U_n) &= \binom{n}{m}^{-2} \sum_{c=0}^{m} \binom{n}{m} \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2 \\ &= \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2.\end{aligned} \tag{25}$$

If $\sigma_m^2 < \infty$, then $\sigma_i^2 < \infty$ for $i < m$. For large n, the first term of the sum dominates since it is the largest order. The coefficient of σ_1^2 is $m\binom{n-m}{m-1}/\binom{n}{m} \sim m^2/n$. ■

In the example of estimating a variance with kernel (7), $h(x_1, x_2) = (x_1 - x_2)^2/2$, we find $h_1(x_1) = \mathrm{E}(X - x_1)^2/2 = \sigma^2/2 + (x_1 - \mu)^2/2$. Then $\sigma_1^2 = \mathrm{Var}(h_1(X_1)) = \mathrm{Var}((X-\mu)^2/2) = (\mu_4 - \sigma^4)/4$, and $\sigma_2^2 = \mathrm{Var}((X_1 - X_2)^2/2) = (\mu_4 - \sigma^4)/2$. From this we find

$$\mathrm{Var}(U_n) = \frac{2}{n(n-1)}[2(n-2)\sigma_1^2 + \sigma_2^2] = (\mu_4 - \sigma^4)/n. \tag{26}$$

Theorem 2. *If $\sigma_m^2 < \infty$, then $\sqrt{n}(U_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, m^2\sigma_1^2)$.*

Proof. Let

$$U_n^* = \frac{m}{n}\sum_{k=1}^{n}(h_1(X_i) - \theta). \tag{27}$$

Then since $m(h_1(X_i) - \theta)$ are i.i.d. with mean 0 and variance $m^2\sigma_1^2$, the central limit theorem implies that $\sqrt{n}U_n^* \xrightarrow{\mathcal{L}} \mathcal{N}(0, m^2\sigma_1^2)$. We complete the proof by showing that $\sqrt{n}(U_n - \theta)$ and $\sqrt{n}U_n^*$ are asymptotically equivalent and so have the same limiting distribution. For this it suffices to show that $n\mathrm{E}(U_n^* - (U_n - \theta))^2 \to 0$.

$$n\mathrm{E}(U_n^* - (U_n - \theta))^2 = n\mathrm{Var}(U_n^*) - 2n\mathrm{Cov}(U_n^*, U_n) + n\mathrm{Var}(U_n) \tag{28}$$

The first term on the right is equal to $m^2\sigma_1^2$ and the last term converges to $m^2\sigma_1^2$ from Theorem 1, so we will be finished when we show $n\mathrm{Cov}(U_n^*, U_n)$ is equal to $m^2\sigma_1^2$.

$$n\mathrm{Cov}(U_n^*, U_n) = \frac{m}{\binom{n}{m}}\sum_{k=1}^{n}\sum_{\mathbf{j}\in\mathbf{C_{m,n}}}\mathrm{Cov}(h_1(X_k), h(X_{j_1}, \ldots, X_{j_m})). \tag{29}$$

The inside covariance is zero if k is not equal to one of the j_i, and it is σ_1^2 otherwise, from Lemma 1. For fixed k the number of sets $\{i_1, \ldots, i_m\}$ containing k is $\binom{n-1}{m-1}$ and since there are n such k,

$$n\mathrm{Cov}(U_n^*, U_n) = \frac{m}{\binom{n}{m}}n\binom{n-1}{m-1}\sigma_1^2 = m^2\sigma_1^2. \; ■ \tag{30}$$

Application. As an application of this theorem, consider the U-statistic, $U_n^{(2)}$ with kernel, $h(x_1, x_2) = I(x_1 + x_2 > 0)$ of degree $m = 2$, associated with the Wilcoxon signed rank test. The parameter estimated is $\theta = \mathrm{E}h(X_1, X_2) = \mathrm{P}(X_1 + X_2 > 0)$. From Lemma 1, we have

$$\sigma_1^2 = \mathrm{Cov}(h(X_1, X_2), h(X_1, X_3)) = \mathrm{P}(X_1 + X_2 > 0, X_1 + X_3 > 0) - \theta^2. \tag{31}$$

Under the null hypothesis that the distribution P is symmetric about 0, we have $\theta = 1/2$ and $\mathrm{P}(X_1 + X_2 > 0, X_1 + X_3 > 0) = \mathrm{P}(X_1 > -X_2, X_1 > -X_3) = \mathrm{P}(X_1 > X_2, X_1 >$

$X_3) = 1/3$, since this is just the probability that of three i.i.d. random variables, the first is the largest. Therefore, under the null hypothesis, $\sigma_1^2 = (1/3) - (1/2)^2 = 1/12$, and since $m = 2$, Theorem 2 gives

$$\sqrt{n}(U_n^{(2)} - 1/2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1/3). \tag{32}$$

This test of the null hypothesis based on $U_n^{(2)}$ is consistent only for alternatives P for which $\theta(P) \neq 1/2$. In Exercise 5, you are to find a test that is consistent against all alternatives.

Under the general hypothesis, $\sqrt{n}(U_n^{(2)} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\sigma_1^2)$. This may be used to find a confidence interval for θ. For this purpose though, we need an estimate of σ_1^2. Why not use a U-statistic? One can estimate $\mathrm{P}(X_1 + X_2 > 0, X_1 + X_3 > 0)$ by the U-statistic associated with the kernel, $f(x_1, x_2, x_3) = \mathrm{I}(x_1 + x_2 > 0, x_1 + x_3 > 0)$, or its symmetrized counterpart, $h(x_1, x_2, x_3) = (1/3)[f(x_1, x_2, x_3) + f(x_2, x_1, x_3) + f(x_3, x_2, x_1)]$.

4. Two-Sample Problems. The important extention to k-sample problems for $k \geq 2$ has been made by Lehmann (1951). The basic ideas are contained in the 2-sample case which is discussed here. Here $\mathcal{P}$ is a family of pairs of probability measures, (F, G).

Consider independent samples, $X_1, \ldots, X_{n_1}$ from $F(x)$ and $Y_1, \ldots, Y_{n_2}$ from $G(y)$. Let $h(x_1, \ldots, x_{m_1}, y_1, \ldots, y_{m_2})$ be a kernel, and let $\mathcal{P}$ be the set of all pairs such that the expectation

$$\theta = \theta(F, G) = \mathrm{E}_{F_1, F_2} h(X_1, \ldots, X_{m_1}, Y_1, \ldots, Y_{m_2}) \tag{33}$$

is finite. As before we may assume without loss of generality that h is symmetric under independent permutations of $x_1, \ldots, x_{m_1}$ and $y_1, \ldots, y_{m_2}$. The corresponding U-statistic is

$$U_{n_1, n_2} = U(h) = \frac{1}{\binom{n_1}{m_1}\binom{n_2}{m_2}} \sum h(X_{i_1}, \ldots, X_{i_{m_1}}, Y_{j_1}, \ldots, Y_{j_{m_2}}), \tag{34}$$

where the sum is over all $\binom{n_1}{m_1}\binom{n_2}{m_2}$ sets of subscripts such that $1 \leq i_1 < \cdots < i_{m_1} \leq n_1$ and $1 \leq j_1 < \cdots < j_{m_2} \leq n_2$. Again it is clear that U is an unbiased estimate of θ.

Examples. There are various two-sample tests based on U-statistics of the hypothesis of equality of distributions, $H_0 : F = G$. They differ in their behavior against various alternative hypotheses.

1. *A two-sample comparison of means.* Taking F and G to be distributions on the real line with finite variances, let $h(x_1, y_1) = x_1 - y_1$, a kernel of degree $(m_1, m_2) = (1, 1)$. Then $\theta = \mathrm{E}X - \mathrm{E}Y$. The corresponding U-statistic is

$$U_{n_1, n_2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (X_i - Y_j) = \overline{X}_{n_1} - \overline{Y}_{n_2}. \tag{35}$$

2. *The Wilcoxon (1945), Mann-Whitney (1947), two-sample test.* Take F and G to be continuous distributions on the real line, and let the kernel be $h(x, y) = \mathrm{I}(y < x)$, with expectation $\theta = P(Y < X)$. The corresponding U-statistic is

$$U_{n_1, n_2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(X_i, Y_j) = \frac{W}{n_1 n_2} \tag{36}$$

where W is the number of pairs, (X_i, Y_j), with $X_i > Y_j$. The corresponding test of the hypothesis $F = G$ (or $\theta = 1/2$) is equivalent to the rank-sum test. It is consistent only against alternatives (F, G) for which $P_{F,G}(X > Y) \neq 1/2$.

3. *A test consistent against all alternatives.* With F and G continuous as before, let $h(x_1, x_2, y_1, y_2) = \mathrm{I}(x_1 < y_1, x_2 < y_1) + \mathrm{I}(y_1 < x_1, y_2 < x_1)$. (The symmetrized version would have four terms.) The expectation is

$$\begin{aligned}\theta &= P(X_1 < Y, X_2 < Y) + P(Y_1 < X, Y_2 < X) \\ &= \frac{2}{3} + \int (F(x) - G(x))^2 d(F(x) + G(x))/2.\end{aligned} \tag{37}$$

(See Exercise 6.) The hypothesis that $F = G$ is equivalent to the hypothesis $\theta = 2/3$. The test that rejects this hypothesis if the corresponding U-statistic is too large is consistent against all alternatives.

Asymptotic Distribution. Corresponding to theorems 1 and 2, we have the following. Let

$$\begin{aligned}\sigma_{ij}^2 = \mathrm{Cov}[&h(X_1, \ldots, X_i, X_{i+1}, \ldots, X_{m_1}, Y_1, \ldots, Y_j, Y_{j+1}, \ldots, Y_{m_2}), \\ &h(X_1, \ldots, X_i, X'_{i+1}, \ldots, X'_{m_1}, Y_1, \ldots, Y_j, Y'_{j+1}, \ldots, Y'_{m_2})]\end{aligned} \tag{38}$$

where the X's and Y's are independently distributed according to F and G respectively.

Theorem 3. *For $P \in \mathcal{P}$,*

$$\mathrm{Var}(U_{n_1,n_2}) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{\binom{m_1}{i}\binom{n_1-m_1}{m_1-i}}{\binom{n_1}{m_1}} \frac{\binom{m_2}{j}\binom{n_2-m_2}{m_2-j}}{\binom{n_2}{m_2}} \sigma_{ij}^2. \tag{39}$$

Moreover, if $\sigma_{m_1 m_2}^2$ is finite, and if $n_1/N \to p \in (0,1)$ as $N = (n_1 + n_2) \to \infty$, then

$$\sqrt{N}(U_{n_1,n_2} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2), \quad \textit{where} \quad \sigma^2 = \frac{m_1^2}{p}\sigma_{10}^2 + \frac{m_2^2}{1-p}\sigma_{01}^2. \tag{40}$$

As an application of this theorem, let us derive the asymptotic distribution of the Wilcoxon two-sample test of Example 2. We have $h(x, y) = \mathrm{I}(y < x)$ and $\theta = \mathrm{P}(Y < X)$. To find σ^2, we have $m_1 = m_2 = 1$ so we need σ_{10}^2 and σ_{01}^2.

$$\sigma_{10}^2 = \mathrm{Cov}(\mathrm{I}(Y < X), \mathrm{I}(Y' < X)) = \mathrm{P}(Y < X, Y' < X) - \mathrm{P}(Y < X)^2, \tag{41}$$

and similarly, $\sigma_{01}^2 = \mathrm{P}(Y < X, Y < X') - \mathrm{P}(Y < X)^2$. Under the null hypothesis that $F = G$, we have $\theta = 1/2$ and $\sigma_{10}^2 = \sigma_{01}^2 = 1/3 - 1/4 = 1/12$, so that $\sigma^2 = 1/(12p(1-p))$. Then p may be replaced by n_1/N resulting in

$$\sqrt{N}(U - 1/2) \approx \mathcal{N}(0, N^2/(12 n_1 n_2)). \tag{42}$$

5. Degeneracy. When using U-statistics for testing hypotheses, it occasionally happens that at the null hypothesis, the asymptotic distribution has variance zero. This is a degenerate case, and we cannot use Theorem 2 to find approximate cutoff points. The general definition of degeneracy for a U-statistic of order m and variances, $\sigma_1^2 \leq \sigma_2^2 \leq \cdots \leq \sigma_m^2$ given by (19) is as follows.

Definition 3. *We say that a U-statistic has a degeneracy of order k if $\sigma_1^2 = \cdots = \sigma_k^2 = 0$ and $\sigma_{k+1}^2 > 0$.*

To present the ideas, we restrict attention to kernels with degeneracy of order 1, for which $\sigma_1^2 = 0$ and $\sigma_2^2 > 0$.

Example 1. Consider the kernel, $h(x_1, x_2) = x_1 x_2$, used in (5). Then, $h_1(x_1) = E(x_1 X_2) = x_1 E(X_2) = x_1 \mu$, and $\sigma_1^2 = \mathrm{Var}(h_1(X_1)) = \mu^2 \sigma^2$, where $\sigma^2 = \mathrm{Var}(X_1)$. So from Theorem 2,

$$\sqrt{n}(U_n - \mu^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\mu^2\sigma^2). \tag{43}$$

But suppose that $\mu = E(X_1) = 0$ under the null hypothesis. Then the limiting variance is zero, so that this theorem is useless for finding cutoff points for a test of the null hypothesis.

But, assuming $\sigma^2 > 0$, we have $\sigma_2^2 = \mathrm{Var}(X_1 X_2) = \sigma^4 > 0$, so that the degeneracy is of order 1. To find the asymptotic distribution of $U_n = \binom{n}{2}^{-1} \sum_{i<j} X_i X_j$ for a sample $X_1, X_2, \ldots$ from a distribution with mean 0 and variance σ^2, we rewrite U_n as follows.

$$\begin{aligned} U_n &= \frac{1}{n(n-1)} \sum\sum_{i \neq j} X_i X_j = \frac{1}{n(n-1)}((\sum_{i=1}^n X_i)^2 - \sum_{i=1}^n X_i^2) \\ &= \frac{1}{n-1}((\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i)^2 - \frac{1}{n} \sum_{i=1}^n X_i^2) \end{aligned} \tag{44}$$

From the central limit theorem we have $\frac{1}{\sqrt{n}} \sum_1^n X_i \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$, and from the law of large numbers we have $\frac{1}{n} \sum_1^n X_i^2 \xrightarrow{\mathcal{L}} \sigma^2$. Therefore by Slutsky's Theorem, we have

$$nU_n \xrightarrow{\mathcal{L}} (Z^2 - 1)\sigma^2 \qquad \text{where } Z \in \mathcal{N}(0, 1). \tag{45}$$

As a slight generalization of Example 1, consider the kernel, $h(x_1, x_2) = f(x_1) f(x_2)$ for some real-valued function $f(x)$ for which $\mathrm{E}f(X_1) = 0$ and $\sigma^2 = \mathrm{E}f(X_1)^2 > 0$. Then the above analysis implies that

$$nU_n = \frac{1}{(n-1)} \sum\sum_{i \neq j} f(X_i) f(X_j) \xrightarrow{\mathcal{L}} (Z^2 - 1)\sigma^2 \tag{46}$$

as well.

Example 2. Suppose now that $h(x_1, x_2) = af(x_1)f(x_2) + bg(x_1)g(x_2)$, where $f(x)$ and $g(x)$ are orthonormal functions of mean zero; that is, $\mathrm{E}f(X)^2 = \mathrm{E}g(X)^2 = 1$, $\mathrm{E}f(X)g(X) = 0$ and $\mathrm{E}f(X) = \mathrm{E}g(X) = 0$. Then, $h_1(x_1) = \mathrm{E}h(x_1, X_2) \equiv 0$, so that $\sigma_1^2 = 0$, and

$$\begin{aligned} \sigma_2^2 &= a^2 \mathrm{Var}(f(X_1)f(X_2)) + 2ab\mathrm{Cov}(f(X_1)f(X_2), g(X_1)g(X_2)) + b^2 \mathrm{Var}(g(X_1)g(X_2)) \\ &= a^2 + b^2 \end{aligned} \tag{47}$$

so the degeneracy is of order 1 (assuming $a^2+b^2>0$). To find the asymptotic distribution of U_n, we perform an analysis as in Example 1.

$$\begin{aligned}(n-1)U_n &= \frac{1}{n}\sum\sum_{i\neq j}[af(X_i)f(X_j)+bg(X_i)g(X_j)] \\ &= a[(\frac{1}{\sqrt{n}}\sum f(X_i))^2-\frac{1}{n}\sum f(X_i)^2]+b[(\frac{1}{\sqrt{n}}\sum g(X_i))^2-\frac{1}{n}\sum g(X_i)^2] \\ &\xrightarrow{\mathcal{L}} a(Z_1^2-1)+b(Z_2^2-1)\end{aligned} \tag{48}$$

where Z_1 and Z_2 are independent $\mathcal{N}(0,1)$.

The General Case. Example 2 is indicative of the general result for kernels with degeneracy of order 1. This is due to a result from the Hilbert-Schmidt theory of integral equations: For given i.i.d. random variables, X_1 and X_2, any symmetric, square integrable function, $A(x_1,x_2)$, ($A(x_1,x_2)=A(x_2,x_1)$ and $\mathrm{E}A(X_1,X_2)^2<\infty$), admits a series expansion of the form,

$$A(x_1,x_2)=\sum_{k=1}^{\infty}\lambda_k\varphi_k(x_1)\varphi_k(x_2) \tag{49}$$

where the λ_k are real numbers, and the φ_k are an orthonormal sequence,

$$\mathrm{E}\varphi_j(X_1)\varphi_k(X_1)=\begin{cases}1 & \text{if } j=k,\\ 0 & \text{if } j\neq k.\end{cases} \tag{50}$$

The λ_k are the eigenvalues, and the $\varphi_k(x)$ are corresponding eigenfunctions of the transformation, $g(x)\to \mathrm{E}A(x,X_1)g(X_1)$. That is, for all k,

$$\mathrm{E}A(x,X_2)\varphi_k(X_2)=\lambda_k\varphi_k(x). \tag{51}$$

Equation (49) is to be understood in the L_2 sense, that

$$\sum_{k=1}^{n}\lambda_k\varphi_k(X_1)\varphi_k(X_2)\xrightarrow{q.m.}A(X_1,X_2). \tag{52}$$

Stronger conditions on A are required to obtain convergence a.s.

In our problem, we take $A(x_1,x_2)=h(x_1,x_2)-\theta$, where $\theta=\mathrm{E}h(X_1,X_2)$. This is a symmetric square integrable kernel, but we are also assuming $\sigma_1^2=\mathrm{Var}\,h_1(X)=0$, where $h_1(x)=\mathrm{E}h(x,X_2)$. Note $\mathrm{E}h_1(X)=\theta$, but since $\mathrm{Var}\,h_1(X)=0$, we have $h_1(x)\equiv\theta$ a.s. Now replace x in (51) by X_1 and take expectations on both sides. We obtain

$$\begin{aligned}\lambda_k\mathrm{E}(\varphi_k(X_1)) &= \mathrm{E}[(h(X_1,X_2)-\theta)\varphi_k(X_2)] \\ &= \mathrm{E}[\mathrm{E}(h(X_1,X_2)-\theta|X_2)\varphi_k(X_2)] \\ &= \mathrm{E}[(h_1(X_2)-\theta)\varphi_k(X_2)]=0.\end{aligned} \tag{53}$$

Thus all eigenfunctions corresponding to nonzero eigenvalues have mean zero. Now we can apply the method of Example 2, to find the asymptotic distribution of $n(U_n-\theta)$.

Theorem 4. *Let U_n be the U-statistic associated with a symmetric kernel of degree 2, degeneracy of order 1, and expectation θ. Then $n(U_n - \theta) \xrightarrow{\mathcal{L}} \sum_1^\infty \lambda_j (Z_j^2 - 1)$, where $Z_1, Z_2, \ldots$ are independent $\mathcal{N}(0,1)$ and $\lambda_1, \lambda_2, \ldots$ are the eigenvalues satisfying (49) with $A(x_1, x_2) = h(x_1, x_2) - \theta$.*

For h having degeneracy of order 1 and arbitrary degree $m \geq 2$, the corresponding result gives the asymptotic distribution of $n(U_n - \theta)$ as $\binom{m}{2} \sum_1^\infty \lambda_j (Z_j^2 - 1)$, where the λ_i are the eigenvalues for the kernel $h_2(x_1, x_2) - \theta$. (See Serfling (1980) or Lee (1990).)

Computation. To obtain the asymptotic distribution of U_n in a specific case requires computation of the eigenvalues, λ_i, each taken with its multiplicity. In general, there may be an infinite number of these. However, for many kernels, there are just a finite number of nonzero eigenvalues. This occurs, for example, when $h(x, y)$ is a polynomial in x and y, or more generally, when $h(x, y)$ is given in the form, $h(x, y) = \sum_1^p f_i(x) g_i(y)$, for some functions f_i and g_i. See Exercise 8 for an indication of how the λ_i are found for such kernels.

Exercises.

1. Let $\mathcal{P}$ be the set of all distributions on the real line with finite first moment. Show that there does not exist a function $f(x)$ such that $\mathrm{E}_P f(X) = \mu^2$ for all $P \in \mathcal{P}$, where μ is the mean of P, and X is a random variable with distribution P.

2. Let g_1 and g_2 be estimable parameters within $\mathcal{P}$ with respective degrees m_1 and m_2. (a) Show $g_1 + g_2$ is an estimable parameter with degree $\leq \max\{m_1, m_2\}$. (b) Show $g_1 \cdot g_2$ is an estimable parameter with degree at most $m_1 + m_2$.

3. Let $\mathcal{P}$ be the class of distributions of two-dimensional vectors, $\boldsymbol{V} = (X, Y)$, with finite second moments. Find a kernel, $h(\boldsymbol{V}_1, \boldsymbol{V}_2)$ of degree 2, for estimating the covariance. Show that the corresponding U-statistic is the (unbiased) sample covariance, $s_{xy} = \frac{1}{n-1} \sum_1^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)$.

4. Derive Equation (10).

5. A continuous distribution, $F(x)$, on the real line is symmetric about the origin if, and only if, $1 - F(x) = F(-x)$ for all real x. This suggests using the parameter,

$$\begin{aligned}
\theta(F) &= \int (1 - F(x) - F(-x))^2 \, dF(x) \\
&= \int (1 - F(-x))^2 \, dF(x) - 2 \int (1 - F(-x)) F(x) \, dF(x) + \int F(x)^2 \, dF(x)
\end{aligned}$$

as a nonparametric measure of departure from symmetry. Find a kernel, h, of degree 3, such that $E_F h(X_1, X_2, X_3) = \theta(F)$ for all continuous F. Find the corresponding U-statistic. (This provides another test for the problem of Example 2. It has the advantage of being consistent against all alternatives to the hypothesis of symmetry about the origin.)

6. (a) In the two-sample problem with samples $X_1, \ldots, X_{n_1}$ from F and $Y_1, \ldots, Y_{n_2}$ from G, what is the U-statistic with kernel $h(x_1, x_2, y_1) = \mathrm{I}(x_1 < y_1, x_2 < y_1)$?

(b) What is its asymptotic distribution as $n_1 + n_2 \to \infty$ and $n_1/(n_1 + n_2) \to p \in (0, 1)$?

(c) What is the asymptotic distribution under the hypothesis $H_0 : F = G$? (Give numerical values for the mean and variance.)

7. Suppose the distribution of X is symmetric about the origin, with variance $\sigma^2 > 0$ and $\mathrm{E}X^4 < \infty$. Consider the kernel, $h(x, y) = xy + (x^2 - \sigma^2)(y^2 - \sigma^2)$.

(a) Show the problem is degenerate of order 1.

(b) Find λ_1, λ_2, and $\varphi_1(x)$ and $\varphi_2(x)$ orthonormal, so that $h(x, y) = \lambda_1\varphi_1(x)\varphi_1(y) + \lambda_2\varphi_2(x)\varphi_2(y)$.

(c) Find the asymptotic distribution of nU_n.

8. Suppose the distribution of X is symmetric about the origin, with variance $\sigma^2 > 0$ and $\mathrm{E}X^6 < \infty$. Consider the kernel, $h(x, y) = xy(1 + x^2y^2)$.

(a) Show the problem is degenerate of order 1.

(b) Using (51) with $A = h$, show that any eigenfunction with nonzero eigenvalue must be of the form, $\varphi(x) = ax^3 + bx$, for some a and b.

(c) Specializing to the case where X has a $\mathcal{N}(0, 1)$ distribution ($\mathrm{E}X^2 = 1$, $\mathrm{E}X^4 = 3$ and $\mathrm{E}X^6 = 15$), find the linear equations for a and b by equating coefficients of x and x^3 in (51).

(d) Find the two nonzero eigenvalues (no need to find the eigenfunctions).

(e) What is the asymptotic distribution of nU_n?

References.

Denker, Manfred (1985) *Asymptotic Distribution Theory in Nonparametric Statistics*, Fr. Vieweg & Sohn, Braunschweig, Wiesbaden.

Ferguson, T. S. (1996) *A Course in Large Sample Theory*, Chapman-Hall, New York.

Fraser, D. A. S. (1957) *Nonparametric Methods in Statistics*, John Wiley & Sons, New York.

Hoeffding, W. (1948a) A class of statistics with asymptotically normal distribution *Ann. Math. Statist.* **19**, 293-325.

Hoeffding, W. (1948b) A non-parametric test for independence, *Ann. Math. Statist.* **19**, 546-557.

Lee, A. J. (1990) *U-Statistics*, Marcel Dekker Inc., New York.

Lehmann, E. L. (1951) Consistency and unbiasedness of certain nonparametric tests *Ann. Math. Statist.* **22**, 165-179.

Lehmann, E. L. (1999) *Elements of Large Sample Theory*, Springer.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York.