12

U-Statistics

One-sample U-statistics can be regarded as generalizations of means. They are sums of dependent variables, but we show them to be asymptotically normal by the projection method. Certain interesting test statistics, such as the Wilcoxon statistics and Kendall's τ-statistic, are one-sample U-statistics. The Wilcoxon statistic for testing a difference in location between two samples is an example of a two-sample U-statistic. The Cramér–von Mises statistic is an example of a degenerate U-statistic.

12.1 One-Sample U-Statistics

Let $X_1, \ldots, X_n$ be a random sample from an unknown distribution. Given a known function h, consider estimation of the "parameter"

$$\theta = \mathrm{E}h(X_1, \ldots, X_r).$$

In order to simplify the formulas, it is assumed throughout this section that the function h is permutation symmetric in its r arguments. (A given h could always be replaced by a symmetric one.) The statistic $h(X_1, \ldots, X_r)$ is an unbiased estimator for θ, but it is unnatural, as it uses only the first r observations. A *U-statistic with kernel h* remedies this; it is defined as

$$U = \frac{1}{\binom{n}{r}} \sum_{\beta} h(X_{\beta_1}, \ldots, X_{\beta_r}),$$

where the sum is taken over the set of all unordered subsets β of r different integers chosen from $\{1, \ldots, n\}$. Because the observations are i.i.d., U is an unbiased estimator for θ also. Moreover, U is permutation symmetric in $X_1, \ldots X_n$, and has smaller variance than $h(X_1, \ldots, X_r)$. In fact, if $X_{(1)}, \ldots, X_{(n)}$ denote the values $X_1, \ldots, X_n$ stripped from their order (the order statistics in the case of real-valued variables), then

$$U = \mathrm{E}\big(h(X_1, \ldots, X_r) \mid X_{(1)}, \ldots, X_{(n)}\big).$$

Because a conditional expectation is a projection, and projecting decreases second moments, the variance of the U-statistic U is smaller than the variance of the naive estimator $h(X_1, \ldots, X_r)$.

In this section it is shown that the sequence $\sqrt{n}(U-\theta)$ is asymptotically normal under the condition that $\mathrm{E}h^2(X_1,\ldots,X_r)<\infty$.

12.1 ***Example.*** A U-statistic of degree $r=1$ is a mean $n^{-1}\sum_{i=1}^n h(X_i)$. The asserted asymptotic normality is then just the central limit theorem. □

12.2 ***Example.*** For the kernel $h(x_1,x_2)=\frac{1}{2}(x_1-x_2)^2$ of degree 2, the parameter $\theta=\mathrm{E}h(X_1,X_2)=\operatorname{var}X_1$ is the variance of the observations. The corresponding U-statistic can be calculated to be

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

Thus, the sample variance is a U-statistic of order 2. □

The asymptotic normality of a sequence of U-statistics, if $n\to\infty$ with the kernel remaining fixed, can be established by the projection method. The projection of $U-\theta$ onto the set of all statistics of the form $\sum_{i=1}^n g_i(X_i)$ is given by

$$\hat{U}=\sum_{i=1}^n\mathrm{E}(U-\theta\mid X_i)=\frac{r}{n}\sum_{i=1}^n h_1(X_i),$$

where the function h_1 is given by

$$h_1(x)=\mathrm{E}h(x,X_2,\ldots,X_r)-\theta.$$

The first equality in the formula for $\hat{U}$ is the Hájek projection principle. The second equality is established in the proof below.

The sequence of projections $\hat{U}$ is asymptotically normal by the central limit theorem, provided $\mathrm{E}h_1^2(X_1)<\infty$. The difference between $U-\theta$ and its projection is asymptotically negligible.

12.3 ***Theorem.*** *If* $\mathrm{E}h^2(X_1,\ldots,X_r)<\infty$, *then* $\sqrt{n}(U-\theta-\hat{U})\xrightarrow{\mathrm{P}}0$. *Consequently, the sequence* $\sqrt{n}(U-\theta)$ *is asymptotically normal with mean* 0 *and variance* $r^2\zeta_1$, *where, with* $X_1,\ldots,X_r,X_1',\ldots,X_r'$ *denoting i.i.d. variables,*

$$\zeta_1=\operatorname{cov}\bigl(h(X_1,X_2,\ldots,X_r),h(X_1,X_2',\ldots,X_r')\bigr).$$

Proof. We first verify the formula for the projection $\hat{U}$. It suffices to show that $\mathrm{E}(U-\theta\mid X_i)=h_1(X_i)$. By the independence of the observations and permutation symmetry of h,

$$\mathrm{E}\bigl(h(X_{\beta_1},\ldots,X_{\beta_r})-\theta\mid X_i=x\bigr)=\begin{cases}h_1(x) & \text{if } i\in\beta\\ 0 & \text{if } i\notin\beta.\end{cases}$$

To calculate $\mathrm{E}(U-\theta\mid X_i)$, we take the average over all β. Then the first case occurs for $\binom{n-1}{r-1}$ of the vectors β in the definition of U. The factor r/n in the formula for the projection $\hat{U}$ arises as $r/n=\binom{n-1}{r-1}/\binom{n}{r}$.

The projection $\hat{U}$ has mean zero, and variance equal to

$$\begin{aligned}\operatorname{var}\hat{U} &= \frac{r^2}{n}\mathrm{E}h_1^2(X_1)\\ &= \frac{r^2}{n}\int \mathrm{E}\big(h(x, X_2, \ldots, X_r) - \theta\big)\mathrm{E}h(x, X_2', \ldots, X_r')\, dP_{X_1}(x) = \frac{r^2}{n}\zeta_1.\end{aligned}$$

Because this is finite, the sequence $\sqrt{n}\,\hat{U}$ converges weakly to the $N(0, r^2\zeta_1)$-distribution by the central limit theorem. By Theorem 11.2 and Slutsky's lemma, the sequence $\sqrt{n}(U - \theta - \hat{U})$ converges in probability to zero, provided $\operatorname{var} U/\operatorname{var}\hat{U} \to 1$.

In view of the permutation symmetry of the kernel h, an expression of the type $\operatorname{cov}\big(h(X_{\beta_1}, \ldots, X_{\beta_r}), h(X_{\beta_1'}, \ldots, X_{\beta_r'})\big)$ depends only on the number of variables X_i that are common to $X_{\beta_1}, \ldots, X_{\beta_r}$ and $X_{\beta_1'}, \ldots, X_{\beta_r'}$. Let ζ_c be this covariance if c variables are in common. Then

$$\begin{aligned}\operatorname{var} U &= \binom{n}{r}^{-2}\sum_{\beta}\sum_{\beta'}\operatorname{cov}\big(h(X_{\beta_1}, \ldots, X_{\beta_r}), h(X_{\beta_1'}, \ldots, X_{\beta_r'})\big)\\ &= \binom{n}{r}^{-2}\sum_{c=0}^{r}\binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}\zeta_c.\end{aligned}$$

The last step follows, because a pair (β, β') with c indexes in common can be chosen by first choosing the r indexes in β, next the c common indexes from β, and finally the remaining $r - c$ indexes in β' from $\{1, \ldots, n\} - \beta$. The expression can be simplified to

$$\operatorname{var} U = \sum_{c=1}^{r}\frac{r!^2}{c!(r-c)!^2}\,\frac{(n-r)(n-r-1)\cdots(n-2r+c+1)}{n(n-1)\cdots(n-r+1)}\zeta_c.$$

In this sum the first term is $O(1/n)$, the second term is $O(1/n^2)$, and so forth. Because n times the first term converges to $r^2\zeta_1$, the desired limit result $\operatorname{var} U/\operatorname{var}\hat{U} \to 1$ follows. ■

12.4 ***Example (Signed rank statistic).*** The parameter $\theta = \mathrm{P}(X_1 + X_2 > 0)$ corresponds to the kernel $h(x_1, x_2) = 1\{x_1 + x_2 > 0\}$. The corresponding U-statistic is

$$U = \frac{1}{\binom{n}{2}}\sum\sum_{i<j} 1\{X_i + X_j > 0\}.$$

This statistic is the average number of pairs (X_i, X_j) with positive sum $X_i + X_j > 0$, and can be used as a test statistic for investigating whether the distribution of the observations is located at zero. If many pairs (X_i, X_j) yield a positive sum (relative to the total number of pairs), then we have an indication that the distribution is centered to the right of zero.

The sequence $\sqrt{n}(U - \theta)$ is asymptotically normal with mean zero and variance $4\zeta_1$. If F denotes the cumulative distribution function of the observations, then the projection of $U - \theta$ can be written

$$\hat{U} = -\frac{2}{n}\sum_{i=1}^{n}\big(F(-X_i) - \mathrm{E}F(-X_i)\big).$$

This formula is useful in subsequent discussion and is also convenient to express the asymptotic variance in F.

The statistic is particularly useful for testing the null hypothesis that the underlying distribution function is continuous and symmetric about zero: $F(x) = 1 - F(-x)$ for every x. Under this hypothesis the parameter θ equals $\theta = 1/2$, and the asymptotic variance reduces to 4 var $F(X_1) = 1/3$, because $F(X_1)$ is uniformly distributed. Thus, under the null hypothesis of continuity and symmetry, the limit distribution of the sequence $\sqrt{n}(U - 1/2)$ is normal $N(0, 1/3)$, independent of the underlying distribution. The last property means that the sequence U_n is *asymptotically distribution free* under the null hypothesis of symmetry and makes it easy to set critical values. The test that rejects H_0 if $\sqrt{3n}(U - 1/2) \geq z_\alpha$ is asymptotically of level α for every F in the null hypothesis.

This test is asymptotically equivalent to the *signed rank test* of Wilcoxon. Let $R_1^+, \ldots, R_n^+$ denote the *ranks* of the absolute values $|X_1|, \ldots, |X_n|$ of the observations: $R_i^+ = k$ means that $|X_i|$ is the kth smallest in the sample of absolute values. More precisely, $R_i^+ = \sum_{j=1}^n 1\{|X_j| \leq |X_i|\}$. Suppose that there are no pairs of tied observations $X_i = X_j$. Then the signed rank statistic is defined as $W^+ = \sum_{i=1}^n R_i^+ 1\{X_i > 0\}$. Some algebra shows that

$$W^+ = \binom{n}{2} U + \sum_{i=1}^n 1\{X_i > 0\}.$$

The second term on the right is of much lower order than the first and hence it follows that $n^{-3/2}(W^+ - \mathrm{E}W^+) \rightsquigarrow N(0, 1/12)$. □

12.5 ***Example (Kendall's τ).*** The U-statistic theorem requires that the observations $X_1, \ldots, X_n$ are independent, but they need not be real-valued. In this example the observations are a sample of bivariate vectors, for convenience (somewhat abusing notation) written as $(X_1, Y_1), \ldots, (X_n, Y_n)$. *Kendall's τ-statistic* is

$$\tau = \frac{4}{n(n-1)} \sum\sum_{i<j} 1\{(Y_j - Y_i)(X_j - X_i) > 0\} - 1.$$

This statistic is a measure of dependence between X and Y and counts the number of concordant pairs (X_i, Y_i) and (X_j, Y_j) in the observations. Two pairs are *concordant* if the indicator in the definition of τ is equal to 1. Large values indicate positive dependence (or concordance), whereas small values indicate negative dependence. Under independence of X and Y and continuity of their distributions, the distribution of τ is centered about zero, and in the extreme cases that all or none of the pairs are concordant τ is identically 1 or -1, respectively.

The statistic $\tau + 1$ is a U-statistic of order 2 for the kernel

$$h\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}\right) = 21\{(y_2 - y_1)(x_2 - x_1) > 0\}.$$

Hence the sequence $\sqrt{n}\big(\tau + 1 - 2\mathrm{P}\big((Y_2 - Y_1)(X_2 - X_1) > 0\big)\big)$ is asymptotically normal with mean zero and variance $4\zeta_1$. With the notation $F^l(x, y) = \mathrm{P}(X < x, Y < y)$ and $F^r(x, y) = \mathrm{P}(X > x, Y > y)$, the projection of $U - \theta$ takes the form

$$\hat{U} = \frac{4}{n} \sum_{i=1}^n \big(F^l(X_i, Y_i) + F^r(X_i, Y_i) - \mathrm{E}F^l(X_i, Y_i) - \mathrm{E}F^r(X_i, Y_i)\big).$$

If X and Y are independent and have continuous marginal distribution functions, then $\mathrm{E}\tau = 0$ and the asymptotic variance $4\zeta_1$ can be calculated to be $4/9$, independent of the

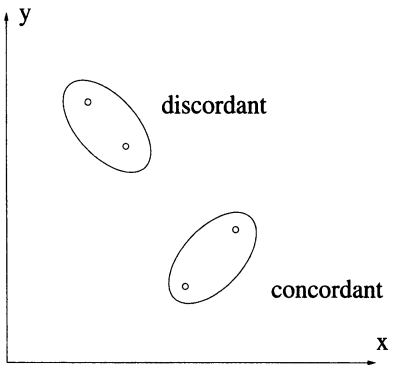

Figure 12.1. Concordant and discordant pairs of points.

marginal distributions. Then $\sqrt{n}\tau \rightsquigarrow N(0, 4/9)$ which leads to the test for "independence": Reject independence if $\sqrt{9n/4}\,|\tau| > z_{\alpha/2}$. □

12.2 Two-Sample U-statistics

Suppose the observations consist of two independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, i.i.d. within each sample, from possibly different distributions. Let $h(x_1, \ldots, x_r, y_1, \ldots, y_s)$ be a known function that is permutation symmetric in $x_1, \ldots, x_r$ and $y_1, \ldots, y_s$ separately. A *two-sample U-statistic* with kernel h has the form

$$U = \frac{1}{\binom{m}{r}\binom{n}{s}} \sum_{\alpha} \sum_{\beta} h(X_{\alpha_1}, \ldots, X_{\alpha_r}, Y_{\beta_1}, \ldots, Y_{\beta_s}),$$

where α and β range over the collections of all subsets of r different elements from $\{1, 2, \ldots, m\}$ and of s different elements from $\{1, 2, \ldots, n\}$, respectively. Clearly, U is an unbiased estimator of the parameter

$$\theta = \mathrm{E}h(X_1, \ldots, X_r, Y_1, \ldots, Y_s).$$

The sequence $U_{m,n}$ can be shown to be asymptotically normal by the same arguments as for one-sample U-statistics. Here we let both $m \to \infty$ and $n \to \infty$, in such a way that the number of X_i and Y_j are of the same order. Specifically, if $N = m + n$ is the total number of observations we assume that, as $m, n \to \infty$,

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

To give an exact meaning to $m, n \to \infty$, we may think of $m = m_\nu$ and $n = n_\nu$ indexed by a third index $\nu \in \mathbb{N}$. Next, we let $m_\nu \to \infty$ and $n_\nu \to \infty$ as $\nu \to \infty$ in such a way that $m_\nu / N_\nu \to \lambda$.

The projection of $U - \theta$ onto the set of all functions of the form $\sum_{i=1}^{m} k_i(X_i) + \sum_{j=1}^{n} l_j(Y_j)$ is given by

$$\hat{U} = \frac{r}{m} \sum_{i=1}^{m} h_{1,0}(X_i) + \frac{s}{n} \sum_{j=1}^{n} h_{0,1}(Y_j),$$

where the functions $h_{1,0}$ and $\hat{h}_{0,1}$ are defined by

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

This follows, as before, by first applying the Hájek projection lemma, and next expressing $\mathrm{E}(U \mid X_i)$ and $\mathrm{E}(U \mid Y_j)$ in the kernel function.

If the kernel is square-integrable, then the sequence $\hat{U}$ is asymptotically normal by the central limit theorem. The difference between $\hat{U}$ and $U - \theta$ is asymptotically negligible.

12.6 ***Theorem.*** *If* $\mathrm{E}h^2(X_1, \ldots, X_r, Y_1, \ldots, Y_s) < \infty$, *then the sequence* $\sqrt{N}(U - \theta - \hat{U})$ *converges in probability to zero. Consequently, the sequence* $\sqrt{N}(U - \theta)$ *converges in distribution to the normal law with mean zero and variance* $r^2\zeta_{1,0}/\lambda + s^2\zeta_{0,1}/(1-\lambda)$, *where, with the* X_i *being i.i.d. variables independent of the i.i.d. variables* Y_j,

$$\zeta_{c,d} = \operatorname{cov}\big(h(X_1, \ldots, X_r, Y_1, \ldots, Y_s),$$
$$h(X_1, \ldots, X_c, X'_{c+1}, \ldots, X'_r, Y_1, \ldots, Y_d, Y'_{d+1}, \ldots, Y'_s)\big).$$

Proof. The argument is similar to the one given previously for one-sample U-statistics. The variances of U and its projection are given by

$$\operatorname{var}\hat{U} = \frac{r^2}{m}\zeta_{1,0} + \frac{s^2}{n}\zeta_{0,1}$$
$$\operatorname{var}U = \frac{1}{\binom{m}{r}^2\binom{n}{s}^2}\sum_{c=0}^{r}\sum_{d=0}^{s}\binom{m}{r}\binom{r}{c}\binom{m-r}{r-c}\binom{n}{s}\binom{s}{d}\binom{n-s}{s-d}\zeta_{c,d}.$$

It can be checked from this that both the sequence $N\operatorname{var}\hat{U}$ and the sequence $N\operatorname{var}U$ converge to the number $r^2\zeta_{1,0}/\lambda + s^2\zeta_{0,1}/(1-\lambda)$. ■

12.7 ***Example (Mann-Whitney statistic).*** The kernel for the parameter $\theta = \mathrm{P}(X \le Y)$ is $h(x, y) = 1\{X \le Y\}$, which is of order 1 in both x and y. The corresponding U-statistic is

$$U = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}1\{X_i \le Y_j\}.$$

The statistic mnU is known as the *Mann-Whitney statistic* and is used to test for a difference in location between the two samples. A large value indicates that the Y_j are "stochastically larger" than the X_i.

If the X_i and Y_j have cumulative distribution functions F and G, respectively, then the projection of $U - \theta$ can be written

$$\hat{U} = -\frac{1}{m}\sum_{i=1}^{m}\big(G_-(X_i) - \mathrm{E}G_-(X_i)\big) + \frac{1}{n}\sum_{j=1}^{n}\big(F(Y_i) - \mathrm{E}F(Y_i)\big).$$

It is easy to obtain the limit distribution of the projections $\hat{U}$ (and hence of U) from this formula. In particular, under the null hypothesis that the pooled sample $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ is i.i.d. with continuous distribution function $F = G$, the sequence $\sqrt{12mn/N}(U - 1/2)$

converges to a standard normal distribution. (The parameter equals $\theta = 1/2$ and $\zeta_{0,1} = \zeta_{1,0} = 1/12$.)

If no observations in the pooled sample are tied, then $mnU + \frac{1}{2}n(n+1)$ is equal to the sum of the ranks of the Y_j in the pooled sample (see Chapter 13). Hence the latter statistic, the *Wilcoxon two-sample statistic*, is asymptotically normal as well. □

*12.3 Degenerate U-Statistics

A sequence of U-statistics (or, better, their kernel function) is called *degenerate* if the asymptotic variance $r^2\zeta_1$ (found in Theorem 12.3) is zero. The formula for the variance of a U-statistic (in the proof of Theorem 12.3) shows that $\operatorname{var} U$ is of the order n^{-c} if $\zeta_1 = \cdots = \zeta_{c-1} = 0 < \zeta_c$. In this case, the sequence $n^{c/2}(U_n - \theta)$ is asymptotically tight. In this section we derive its limit distribution.

Consider the Hoeffding decomposition as discussed in section 11.4. For a U-statistic U_n with kernel h of order r, based on observations $X_1, \ldots, X_n$, this can be simplified to

$$U_n = \sum_{c=0}^{r} \sum_{|A|=c} \frac{1}{\binom{n}{r}} \sum_{\beta} P_A h(X_{\beta_1}, \ldots, X_{\beta_r}) = \sum_{c=0}^{r} \binom{r}{c} U_{n,c} \qquad \text{(say)}.$$

Here, for each $0 \le c \le r$, the variable $U_{n,c}$ is a U-statistic of order c with kernel

$$h_c(X_1, \ldots, X_c) = P_{\{1,\ldots,c\}} h(X_1, \ldots, X_r).$$

To see this, fix a set A with c elements. Because the space H_A is orthogonal to all functions $g(X_j : j \in B)$ (i.e., the space $\sum_{C \subset B} H_C$) for every set B that does not contain A, the projection $P_A h(X_{\beta_1}, \ldots, X_{\beta_r})$ is zero unless $A \subset \beta = \{\beta_1, \ldots, \beta_r\}$. For the remaining β the projection $P_A h(X_{\beta_1}, \ldots, X_{\beta_r})$ does not depend on β (i.e., on the $r - c$ elements of $\beta - A$) and is a fixed function h_c of $(X_j : j \in A)$. This follows by symmetry, or explicitly from the formula for the projections in section 11.4. The function h_c is indeed the function as given previously. There are $\binom{n-c}{r-c}$ vectors β that contain the set A. The claim that $U_{n,c}$ is a U-statistic with kernel h_c now follows by simple algebra, using the fact that $\binom{n-c}{r-c}/\binom{r}{c}\binom{n}{r} = 1/\binom{n}{c}$.

By the defining properties of the space $H_{\{1,\ldots,c\}}$, it follows that the kernel h_c is degenerate for $c \ge 2$. In fact, it is *strongly degenerate* in the sense that the conditional expectation of $h_c(X_1, \ldots, X_c)$ given any strict subset of the variables $X_1, \ldots, X_c$ is zero. In other words, the integral $\int h(x, X_2, \ldots, X_c)\, dP(x)$ with respect to any single argument vanishes. By the same reasoning, $U_{n,c}$ is uncorrelated with every measurable function that depends on strictly fewer than c elements of $X_1, \ldots, X_n$.

We shall show that the sequence $n^{c/2}U_{n,c}$ converges in distribution to a limit with variance $c!\,\mathrm{E}h_c^2(X_1, \ldots, X_c)$ for every $c \ge 1$. Then it follows that the sequence $n^{c/2}(U_n - \theta)$ converges in distribution for c equal to the smallest value such that $h_c \not\equiv 0$. For $c \ge 2$ the limit distribution is not normal but is known as *Gaussian chaos*.

Because the idea is simple, but the statement of the theorem (apparently) necessarily complicated, first consider a special case: $c = 3$ and a "product kernel" of the form

$$h_3(x_1, x_2, x_3) = f_1(x_1) f_2(x_2) f_3(x_3).$$

A U-statistic corresponding to a product kernel can be rewritten as a polynomial in sums of the observations. For ease of notation, let $\mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(X_i)$ (the empirical measure), and let $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f$ (the empirical process), for P the distribution of the observations $X_1, \ldots, X_n$. If the kernel h_3 is strongly degenerate, then each function f_i has mean zero and hence $\mathbb{G}_n f_i = \sqrt{n}\mathbb{P}_n f_i$ for every i. Then, with (i_1, i_2, i_3) ranging over all triplets of three different integers from $\{1, \ldots, n\}$ (taking position into account),

$$\begin{aligned}\frac{3!}{n^{3/2}}\binom{n}{3}U_{n,3} &= \frac{1}{n^{3/2}}\sum\nolimits_{(i_1,i_2,i_3)} f_1(x_{i_1})f_2(x_{i_2})f_3(x_{i_3}) \\ &= \mathbb{G}_n f_1 \mathbb{G}_n f_2 \mathbb{G}_n f_3 - \mathbb{P}_n(f_1 f_2)\mathbb{G}_n f_3 \\ &\quad - \mathbb{P}_n(f_1 f_3)\mathbb{G}_n f_2 - \mathbb{P}_n(f_2 f_3)\mathbb{G}_n f_1 + 2\frac{\mathbb{P}_n(f_1 f_2 f_3)}{\sqrt{n}}.\end{aligned}$$

By the law of large numbers, $\mathbb{P}_n \to Pf$ almost surely for every f, while, by the central limit theorem, the marginal distributions of the stochastic processes $f \mapsto \mathbb{G}_n f$ converge weakly to multivariate Gaussian laws. If $\{\mathbb{G}f : f \in L_2(\mathcal{X}, \mathcal{A}, P)\}$ denotes a Gaussian process with mean zero and covariance function $\mathrm{E}\mathbb{G}f\mathbb{G}g = Pfg - PfPg$ (a *P-Brownian bridge process*), then $\mathbb{G}_n \rightsquigarrow \mathbb{G}$. Consequently,

$$n^{3/2}U_{n,3} \rightsquigarrow \mathbb{G}f_1\mathbb{G}f_2\mathbb{G}f_3 - P(f_1 f_2)\mathbb{G}f_3 - P(f_1 f_3)\mathbb{G}f_2 - P(f_2 f_3)\mathbb{G}f_1.$$

The limit is a polynomial of order 3 in the Gaussian vector $(\mathbb{G}f_1, \mathbb{G}f_2, \mathbb{G}f_3)$.

There is no similarly simple formula for the limit of a general sequence of degenerate U-statistics. However, any kernel can be written as an infinite linear combination of product kernels. Because a U-statistic is linear in its kernel, the limit of a general sequence of degenerate U-statistics is a linear combination of limits of the previous type.

To carry through this program, it is convenient to employ a decomposition of a given kernel in terms of an orthonormal basis of product kernels. This is always possible. We assume that $L_2(\mathcal{X}, \mathcal{A}, P)$ is separable, so that it has a countable basis.

12.8 ***Example (General kernel).*** If $1 = f_0, f_1, f_2, \ldots$ is an orthonormal basis of $L_2(\mathcal{X}, \mathcal{A}, P)$, then the functions $f_{k_1} \times \cdots \times f_{k_c}$ with $(k_1, \ldots, k_c)$ ranging over the nonnegative integers form an orthonormal basis of $L_2(\mathcal{X}^c, \mathcal{A}^c, P^c)$. Any square-integrable kernel can be written in the form $h_c(x_1, \ldots, x_c) = \sum a(k_1, \ldots, k_c) f_{k_1} \times \cdots \times f_{k_c}$ for $a(k_1, \ldots, k_c) = \langle h_c, f_{k_1} \times \cdots \times f_{k_c}\rangle$ the inner products of h_c with the basis functions. □

12.9 ***Example (Second-order kernel).*** In the case that $c = 2$, there is a choice that is specially adapted to our purposes. Because the kernel h_2 is symmetric and square-integrable by assumption, the integral operator $K : L_2(\mathcal{X}, \mathcal{A}, P) \mapsto L_2(\mathcal{X}, \mathcal{A}, P)$ defined by $Kf(x) = \int h_2(x, y) f(y)\, dP(y)$ is self-adjoint and Hilbert-Schmidt. Therefore, it has at most countably many eigenvalues $\lambda_0, \lambda_1, \ldots$, satisfying $\sum \lambda_k^2 < \infty$, and there exists an orthonormal basis of eigenfunctions $f_0, f_1, \ldots$. (See, for instance, Theorem VI.16 in [124].) The kernel h_2 can be expressed relatively to this basis as

$$h_2(x, y) = \sum_{k=0}^{\infty} \lambda_k f_k(x) f_k(y).$$

For a degenerate kernel h_2 the function 1 is an eigenfunction for the eigenvalue 0, and we can take $f_0 = 1$ without loss of generality.

The gain over the decomposition in the general case is that only product functions of the type $f \times f$ are needed. □

The (nonnormalized) *Hermite polynomial* H_j is a polynomial of degree j with leading coefficient x^j such that $\int H_i(x)H_j(x)\phi(x)\,dx = 0$ whenever $i \neq j$. The Hermite polynomials of lowest degrees are $H_0 = 1$, $H_1(x) = x$, $H_2(x) = x^2 - 1$ and $H_3(x) = x^3 - 3x$.

12.10 ***Theorem.*** *Let $h_c : \mathcal{X}^c \mapsto \mathbb{R}$ be a permutation-symmetric, measurable function of c arguments such that $\mathrm{E}h_c^2(X_1, \ldots, X_c) < \infty$ and $\mathrm{E}h_c(x_1, \ldots, x_{c-1}, X_c) \equiv 0$. Let $1 = f_0, f_1, f_2, \ldots$ be an orthonormal basis of $L_2(\mathcal{X}, \mathcal{A}, P)$. Then the sequence of U-statistics $U_{n,c}$ with kernel h_c based on n observations from P satisfies*

$$n^{c/2}U_{n,c} \rightsquigarrow \sum_{k=(k_1,\ldots,k_c)\in\mathbb{N}^c} \langle h_c, f_{k_1} \times \cdots \times f_{k_c}\rangle \prod_{i=1}^{d(k)} H_{a_i(k)}\big(\mathbb{G}\psi_i(k)\big).$$

Here $\mathbb{G}$ is a P-Brownian bridge process, the functions $\psi_1(k), \ldots, \psi_{d(k)}(k)$ are the different elements in $f_{k_1}, \ldots, f_{k_c}$, and $a_i(k)$ is number of times $\psi_i(k)$ occurs among $f_{k_1}, \ldots, f_{k_c}$. The variance of the limit variable is equal to $c!\,\mathrm{E}h_c^2(X_1, \ldots, X_c)$.

Proof. The function h_c can be represented in $L_2(\mathcal{X}^c, \mathcal{A}^c, P^c)$ as the series $\sum_k \langle h_c, f_{k_1} \times \ldots, f_{k_c}\rangle f_{k_1} \times \cdots \times f_{k_c}$. By the degeneracy of h_c the sum can be restricted to $k = (k_1, \ldots, k_c)$ with every $k_j \geq 1$. If $U_{n,c}h$ denotes the U-statistic with kernel $h(x_1, \ldots, x_c)$, then, for a pair of degenerate kernels h and g,

$$\mathrm{cov}(U_{n,c}h, U_{n,c}g) = \frac{c!}{n(n-1)\cdots(n-c+1)}\,P^c hg.$$

This means that the map $h \mapsto n^{c/2}\sqrt{c!}\,U_{n,c}h$ is close to being an isometry between $L_2(P^c)$ and $L_2(P^n)$. Consequently, the series $\sum_k \langle h_c, f_{k_1} \times \cdots \times f_{k_c}\rangle U_{n,c} f_{k_1} \times \cdots \times f_{k_c}$ converges in $L_2(P^n)$ and equals $U_{n,c}h_c = U_{n,c}$. Furthermore, if it can be shown that the finite-dimensional distributions of the sequence of processes $\{U_{n,c} f_{k_1} \times \cdots \times f_{k_c} : k \in \mathbb{N}^c\}$ converge weakly to the corresponding finite-dimensional distributions of the process $\big\{\prod_{i=1}^{d(k)} H_{a_i(k)}\big(\mathbb{G}\psi_i(k)\big) : k \in \mathbb{N}^c\big\}$, then the partial sums of the series converge, and the proof can be concluded by approximation arguments.

There exists a polynomial $\hat{P}_{n,c}$ of degree c, with random coefficients, such that

$$\frac{c!}{n^{c/2}}\binom{n}{c} U_{n,c} f_{k_1} \times \cdots \times f_{k_c} = \hat{P}_{n,c}\big(\mathbb{G}_n f_{k_1}, \ldots, \mathbb{G}_n f_{k_c}\big).$$

(See the example for $c = 3$ and problem 12.13). The only term of degree c in this polynomial is equal to $\mathbb{G}_n f_{k_1}\mathbb{G}_n f_{k_2} \cdots \mathbb{G}_n f_{k_c}$. The coefficients of the polynomials $\hat{P}_{n,c}$ converge in probability to constants. Conclude that the sequence $n^{c/2}c!\,U_{n,c} f_{k_1} \times \cdots \times f_{k_c}$ converges in distribution to $P_c\big(\mathbb{G}f_{k_1}, \ldots, \mathbb{G}f_{k_c}\big)$ for a polynomial P_c of degree c with leading term, and only term of degree c, equal to $\mathbb{G}f_{k_1}\mathbb{G}f_{k_2} \cdots \mathbb{G}f_{k_c}$. This convergence is simultaneous in sets of finitely many k.

It suffices to establish the representation of this limit in terms of Hermite polynomials. This could be achieved directly by algebraic and combinatorial arguments, but then the occurrence of the Hermite polynomials would remain somewhat mysterious. Alternatively,

the representation can be derived from the definition of the Hermite polynomials and covariance calculations. By the degeneracy of the kernel $f_{k_1} \times \cdots \times f_{k_c}$, the U-statistic $U_{n,c} f_{k_1} \times \cdots \times f_{k_c}$ is orthogonal to all measurable functions of $c-1$ or fewer elements of $X_1, \ldots, X_n$, and their linear combinations. This includes the functions $\prod_i (\mathbb{G}_n g_i)^{a_i}$ for arbitrary functions g_i and nonnegative integers a_i with $\sum a_i < c$. Taking limits, we conclude that $P_c(\mathbb{G} f_{k_1}, \ldots, \mathbb{G} f_{k_c})$ must be orthogonal to every polynomial in $\mathbb{G} f_{k_1}, \ldots, \mathbb{G} f_{k_c}$ of degree less than $c-1$. By the orthonormality of the basis f_i, the variables $\mathbb{G} f_i$ are independent standard normal variables. Because the Hermite polynomials form a basis for the polynomials in one variable, their (tensor) products form a basis for the polynomials of more than one argument. The polynomial P_c can be written as a linear combination of elements from this basis. By the orthogonality, the coefficients of base elements of degree $< c$ vanish. From the base elements of degree c only the product as in the theorem can occur, as follows from consideration of the leading term of P_c. ■

12.11 ***Example.*** For $c = 2$ and a basis $1 = f_0, f_1, \ldots$ of eigenfunctions of the kernel h_2, we obtain a limit of the form $\sum_k \langle h_2, f_k \times f_k \rangle H_2(\mathbb{G} f_k)$. By the orthonormality of the basis this variable is distributed as $\sum_k \lambda_k (Z_k^2 - 1)$ for $Z_1, Z_2, \ldots$ a sequence of independent standard normal variables. □

12.12 ***Example (Sample variance).*** The kernel $h(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$ yields the sample variance S_n^2. Because this has asymptotic variance $\mu_4 - \mu_2^2$ (see Example 3.2), the kernel is degenerate if and only if $\mu_4 = \mu_2^2$. This can happen only if $(X_1 - \alpha_1)^2$ is constant, for $\alpha_1 = \mathrm{E} X_i$. If we center the observations, so that $\alpha_1 = 0$, then this means that X_1 only takes the values $-\sigma$ and $\sigma = \sqrt{\mu_2}$, each with probability $1/2$. This is a very degenerate situation, and it is easy to find the limit distribution directly, but perhaps it is instructive to apply the general theorem. The kernels h_c take the forms (See section 11.4),

$$\begin{aligned} h_0 &= \mathrm{E} \tfrac{1}{2}(X_1 - X_2)^2 = \sigma^2, \\ h_1(x_1) &= \mathrm{E} \tfrac{1}{2}(x_1 - X_2)^2 - \sigma^2, \\ h_2(x_1, x_2) &= \tfrac{1}{2}(x_1 - x_2)^2 - \mathrm{E} \tfrac{1}{2}(x_1 - X_2)^2 - \mathrm{E} \tfrac{1}{2}(X_1 - x_2)^2 + \sigma^2. \end{aligned}$$

The kernel is degenerate if $h_1 = 0$ almost surely, and then the second-order kernel is $h_2(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2 - \sigma^2$. Because the underlying distribution has only two support points, the eigenfunctions f of the corresponding integral operator can be identified with vectors $\big(f(-\sigma), f(\sigma)\big)$ in $\mathbb{R}^2$. Some linear algebra shows that they are $(1, 1)$ and $(-1, 1)$, corresponding to the eigenvalues 0 and $-\sigma^2$, respectively. Correspondingly, under degeneracy the kernel allows the decomposition

$$h_2(x_1, x_2) = \tfrac{1}{2}\big(x_1^2 + x_2^2\big) - \sigma^2 - x_1 x_2 = -\sigma^2 \Big(\frac{x_1}{\sigma}\Big)\Big(\frac{x_2}{\sigma}\Big).$$

We can conclude that the sequence $n(S_n^2 - \mu_2)$ converges in distribution to $-\sigma^2(Z_1^2 - 1)$. □

12.13 ***Example (Cramér–von Mises).*** Let $\mathbb{F}_n(x) = n^{-1} \sum_{i=1}^n 1\{X_i \le x\}$ be the empirical distribution function of a random sample $X_1, \ldots, X_n$ of real-valued random variables. The *Cramér–Von Mises statistic* for testing the (null) hypothesis that the underlying cumulative

distribution is a given function F is given by

$$n\int(\mathbb{F}_n - F)^2\,dF = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\int\bigl(1_{X_i\le x} - F(x)\bigr)\bigl(1_{X_j\le x} - F(x)\bigr)\,dF(x).$$

The double sum restricted to the off-diagonal terms is a U-statistic, with, under H_0, a degenerate kernel. Thus, this statistic converges to a nondegenerate limit distribution. The diagonal terms contribute the constant $\int F(1-F)\,dF$ to the limit distribution, by the law of large numbers. If F is uniform, then the kernel of the U-statistic is

$$h(x, y) = \tfrac{1}{2}x^2 + \tfrac{1}{2}y^2 - x\vee y + \tfrac{1}{3}.$$

To find the eigenvalues of the corresponding integral operator K, we differentiate the identity $Kf = \lambda f$ twice, to find the equation $\lambda f'' + f = \int f(s)\,ds$. Because the kernel is degenerate, the constants are eigenfunctions for the eigenvalue 0. The eigenfunctions corresponding to nonzero eigenvalues are orthogonal to this eigenspace, whence $\int f(s)\,ds = 0$. The equation $\lambda f'' + f = 0$ has solutions $\cos ax$ and $\sin ax$ for $a^2 = \lambda^{-1}$. Reinserting these in the original equation or utilizing the relation $\int f(s)\,ds = 0$, we find that the nonzero eigenvalues are equal to $j^{-2}\pi^{-2}$ for $j \in \mathbb{N}$, with eigenfunctions $\sqrt{2}\cos \pi jx$. Thus, the Cramér–Von Mises statistic converges in distribution to $1/6 + \sum_{j=1}^{\infty} j^{-2}\pi^{-2}(Z_j^2 - 1)$. For another derivation of the limit distribution, see Chapter 19. □

Notes

The main part of this chapter has its roots in the paper by Hoeffding [76]. Because the asymptotic variance is smaller than the true variance of a U-statistic, Hoeffding recommends to apply a standard normal approximation to $(U - \mathrm{E}U)/\operatorname{sd} U$. Degenerate U-statistics were considered, among others, in [131] within the context of more general linear combinations of symmetric kernels. Arcones and Giné [2] have studied the weak convergence of "U-processes", stochastic processes indexed by classes of kernels, in spaces of bounded functions as discussed in Chapter 18.

PROBLEMS

1. Derive the asymptotic distribution of *Gini's mean difference*, which is defined as $\binom{n}{2}^{-1}\sum\sum_{i<j}|X_i - X_j|$.
2. Derive the projection of the sample variance.
3. Find a kernel for the parameter $\theta = \mathrm{E}(X - \mathrm{E}X)^3$.
4. Find a kernel for the parameter $\theta = \operatorname{cov}(X, Y)$. Show that the corresponding U-statistic is the sample covariance $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})/(n-1)$.
5. Find the limit distribution of $U = \binom{n}{2}^{-1}\sum\sum_{i<j}(Y_j - Y_i)(X_j - X_i)$.
6. Let U_{n1} and U_{n2} be U-statistics with kernels h_1 and h_2, respectively. Derive the joint asymptotic distribution of (U_{n1}, U_{n2}).
7. Suppose $\mathrm{E}X_1^2 < \infty$. Derive the asymptotic distribution of the sequence $n^{-1}\sum\sum_{i\neq j} X_iX_j$. Can you give a two line proof without using the U-statistic theorem? What happens if $\mathrm{E}X_1 = 0$?
8. **(Mann's test against trend.)** To test the null hypothesis that a sample $X_1, \ldots, X_n$ is i.i.d. against the alternative hypothesis that the distributions of the X_i are stochastically increasing in i, Mann

suggested to reject the null hypothesis if the number of pairs (X_i, X_j) with $i < j$ and $X_i < X_j$ is large. How can we choose the critical value for large n?

9. Show that the U-statistic U with kernel $1\{x_1 + x_2 > 0\}$, the signed rank statistic W^+, and the positive-sign statistic $S = \sum_{i=1}^n 1\{X_i > 0\}$ are related by $W^+ = \binom{n}{2}U + S$ in the case that there are no tied observations.

10. A *V-statistic* of order 2 is of the form $n^{-2}\sum_{i=1}^n \sum_{j=1}^n h(X_i, X_j)$ where $h(x, y)$ is symmetric in x and y. Assume that $\mathrm{E}h^2(X_1, X_1) < \infty$ and $\mathrm{E}h^2(X_1, X_2) < \infty$. Obtain the asymptotic distribution of a V-statistic from the corresponding result for a U-statistic.

11. Define a V-statistic of general order r and give conditions for its asymptotic normality.

12. Derive the asymptotic distribution of $n(S_n^2 - \mu_2)$ in the case that $\mu_4 = \mu_2^2$ by using the delta-method (see Example 12.12). Does it make a difference whether we divide by n or $n - 1$?

13. For any $(n \times c)$ matrix a_{ij} we have

$$\sum_i a_{i_1,1} \cdots a_{i_c,c} = \sum_{\mathcal{B}} \prod_{B \in \mathcal{B}} (-1)^{|B|-1}(B| - 1)! \sum_{i=1}^n \prod_{j \in B} a_{ij}.$$

Here the sum on the left ranges over all ordered subsets $(i_1, \ldots, i_c)$ of different integers from $\{1, \ldots, n\}$ and the first sum on the right ranges over all partitions $\mathcal{B}$ of $\{1, \ldots, c\}$ into nonempty sets (see Example [131]).

14. Given a sequence of i.i.d. random variables $X_1, X_2, \ldots$, let $\mathcal{A}_n$ be the σ-field generated by all functions of $(X_1, X_2, \ldots)$ that are symmetric in their first n arguments. Prove that a sequence U_n of U-statistics with a fixed kernel h of order r is a reverse martingale (for $n \geq r$) with respect to the filtration $\mathcal{A}_r \supset \mathcal{A}_{r+1} \supset \cdots$.

15. (Strong law.) If $\mathrm{E}|h(X_1, \cdots, X_r)| < \infty$, then the sequence U_n of U-statistics with kernel h converges almost surely to $\mathrm{E}h(X_1, \cdots, X_r)$. (For $r > 1$ the condition is not necessary, but a simple necessary and sufficient condition appears to be unknown.) Prove this. (Use the preceding problem, the martingale convergence theorem, and the Hewitt-Savage 0-1 law.)