Solutions for the Section on U-statistics

1. We must have \( \int h(x) \, dP(x) = \mu^2 \) for all distributions on the real line with mean \( \mu \). If \( P \) is degenerate at a point \( y \), this implies that \( h(y) = y^2 \) for all \( y \). But if \( P \) has mean zero and is not degenerate, then \( \int h(x) \, dP(x) = \int x^2 \, dP(x) > 0 \), which is a contradiction.

2. Suppose \( g_1 \) and \( g_2 \) are estimable parameters within \( P \) with respective degrees \( m_1 \) and \( m_2 \). Let \( h_1(x_1, \ldots, x_{m_1}) \) and \( h_2(x_1, \ldots, x_{m_2}) \) be the corresponding kernels.
   (a) Let \( m = \max\{m_1, m_2\} \) and let \( h(x_1, \ldots, x_m) = h_1(x_1, \ldots, x_{m_1}) + h_2(x_1, \ldots, x_{m_2}) \). Then \( E h(X_1, \ldots, X_m) = g_1 + g_2 \), showing that \( g_1 + g_2 \) is an estimable parameter with degree at most \( m \).
   (b) Let \( m = m_1 + m_2 \) and let \( h(x_1, \ldots, x_m) = h_1(x_1, \ldots, x_{m_1}) h_2(x_{m_1+1}, \ldots, x_{m_1+m_2}) \). Then \( E h(X_1, \ldots, X_m) = g_1 g_2 \), showing that \( g_1 g_2 \) is an estimable parameter with degree at most \( m \).

3. The kernel, \( h((x_1, y_1), (x_2, y_2)) = x_1 y_1 - x_1 y_2 \), has the correct expectation because \( E h((X_1, Y_1), (X_2, Y_2)) = E(XY) - E(X)E(Y) = \text{Cov}(X, Y) \). We could symmetrize this, but we might as well just use the permutation form of the U-statistic,
   \[
   U_n = \frac{1}{n(n-1)} \sum_{i \neq j} [X_i Y_i - X_i Y_j]
   \]
   \[
   = \frac{1}{n(n-1)} \left[ (n-1) \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \left( \sum_{j=1}^{n} Y_j - Y_i \right) \right]
   \]
   \[
   = \frac{1}{n(n-1)} \left[ n \sum X_i Y_i - \left( \sum X_i \right) \left( \sum Y_i \right) \right]
   \]
   \[
   = \frac{n}{n-1} \left[ \frac{1}{n} \sum X_i Y_i - \bar{X}_n \bar{Y}_n \right]
   \]
   \[
   = \frac{n}{n-1} \left[ \frac{1}{n} \sum (X_i - \bar{X}_n)(Y_i - \bar{Y}_n) \right] = s_{xy}
   \]

4. The positive rank of \( Z_i \) may be written \( R_i^+ = 1 + \sum_{j=1}^{n} I(|Z_j| < |Z_i|) \). So, \( W_n^+ = \sum_{i=1}^{n} (1 + \sum_{j=1}^{n} I(|Z_j| < |Z_i|) I(Z_i > 0) = \sum_i I(Z_i > 0) + \sum_i \sum_{i \neq j} I(|Z_j| < |Z_i|, Z_i > 0). \)

We must show
\[
\sum_{i < j} \sum I(Z_i + Z_j > 0) = \sum_{i \neq j} I(|Z_j| < |Z_i|, Z_i > 0).
\]
Now note that \( I(|Z_j| < |Z_i|, Z_i > 0) + I(|Z_i| < |Z_j|, Z_j > 0) = I(Z_i + Z_j > 0). \) So the double sum on the right is
\[
\sum_{i < j} I(|Z_j| < |Z_i|, Z_i > 0) + \sum_{i > j} I(|Z_j| < |Z_i|, Z_i > 0)
\]
\[
= \sum_{i < j} \left[ I(|Z_j| < |Z_i|, Z_i > 0) + I(|Z_i| < |Z_j|, Z_j > 0) \right] = \sum_{i < j} I(Z_i + Z_j > 0).
\]
5. Write for independent $X_1$, $X_2$, and $X_3$,

$$\theta(F) = \int P(X_1 > -x, X_2 > -x) dF(x) - 2 \int P(X_1 > -x, X_2 < x) dF(x) + 1/3$$

$$= P(X_1 + X_3 > 0, X_2 + X_3 > 0) - 2P(X_1 + X_3 > 0, -X_2 + X_3 > 0) + 1/3.$$ 

This leads to the unbiased estimate of $\theta$, $f(x_1, x_2, x_3) = I(x_1 + x_3 > 0, x_2 + x_3 > 0) - 2I(x_1 + x_3 > 0, -x_2 + x_3 > 0) + 1/3$. This is not symmetric in its arguments, so the symmetrized version has six terms,

$$h(x_1, x_2, x_3) = [f(x_1, x_2, x_3) + f(x_1, x_3, x_2) + f(x_2, x_1, x_3)$$

$$+ f(x_2, x_3, x_1) + f(x_3, x_1, x_2) + f(x_3, x_2, x_1)]/6$$

The corresponding U-statistic is

$$U_n = \left(\begin{array}{c} n \\ 3 \end{array}\right)^{-1} \sum_{i_1 < i_2 < i_3} h(X_{i_1}, X_{i_2}, X_{i_3}).$$

6. (a)

$$U_{n_1, n_2} = \frac{2}{n_1(n_1 - 1)n_2} \sum_{i_1 < i_2} \sum_j I(X_{i_1} < Y_j, X_{i_2} < Y_j).$$

(b) $\sqrt{n_1 + n_2}(U_{n_1, n_2} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (4\sigma_{10}^2/p) + (\sigma_{01}^2/(1 - p)))$, where

$$\theta = P(X_1 < Y, X_2 < Y)$$

$$\sigma_{10}^2 = \text{Cov}(h(X_1, X_2, Y), h(X_1, X_2', Y')) = P(X_1 < Y, X_2 < Y, X_1 < Y', X_2 < Y') - \theta^2$$

$$\sigma_{01}^2 = \text{Cov}(h(X_1, X_2, Y), h(X_1', X_2', Y')) = P(X_1 < Y, X_2 < Y, X_1' < Y, X_2' < Y) - \theta^2$$

with $X_1, X_2, X_1', X_2', Y, Y'$ independent.

(c) Assume $H_0$. Then $\theta = 1/3$ and $P(X_1 < Y, X_2 < Y, X_1' < Y, X_2' < Y) = 1/5$, since this is just the probability that $Y$ will be the largest of a sample of size 5. Thus $\sigma_{01}^2 = 1/5 - 1/9 = 4/45$. Computation of $P(X_1 < Y, X_2 < Y, X_1 < Y', X_2' < Y')$ is more lengthy. All $5! = 120$ orderings of these 5 variables are equally likely. Of the 24 with $X_1$ the smallest, there are exactly 6 satisfying the condition. Of the 24 with $X_1$ the next to smallest, there are exactly 6 satisfying the condition. Of the 24 with $X_1$ in the center, there are exactly 4 satisfying the condition. The condition prevents $X_1$ being largest or next to largest. So $P(X_1 < Y, X_2 < Y, X_1 < Y', X_2' < Y') = 16/120 = 2/15$, and $\sigma_{10}^2 = 2/15 - 1/9 = 1/45$. Therefore, $\sqrt{n_1 + n_2}(U_{n_1, n_2} - 1/3) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (4/45p) + ((4/45)/(1 - p))) = \mathcal{N}(0, 4/(45p(1 - p))).$

7. (a) Since the distribution of $X$ is symmetric about 0, we have $EX = EX^3 = 0$, so $\theta = Eh(X, Y) = EXEY + E(X^2 - \sigma^2)E(Y^2 - \theta^2) = 0$. Moreover, $h_1(x) = Eh(x, Y) =$
xEY + (x^2 - \sigma^2)E(Y^2 - \sigma^2) \equiv 0$, so $\sigma_1^2 = \text{Var} h_1(X) = 0$. Since $h(X, Y)$ is not degenerate, $\sigma_2^2 > 0$ so the degeneracy is of order 1.

(b) $X$ and $X^2 - \sigma^2$ are orthogonal since $EX(X^2 - \sigma^2) = EX^3 - \sigma^2 EX = 0$. Since $EX^2 = \sigma^2$, $\varphi_1(x) = x/\sigma$ has variance 1; similarly, since $E(X^2 - \sigma^2)^2 = \mu_4 - \sigma^4$, $\varphi_2(x) = (x^2 - \sigma^2)/\sqrt{\mu_4 - \sigma^4}$ has variance 1. Then, $h(x, y) = \sigma^2 \varphi_1(x)\varphi_1(y) + (\mu_4 - \sigma^4)\varphi_2(x)\varphi_2(y)$, so the eigenvalues are $\lambda_1 = \sigma^2$ and $\lambda_2 = \mu_4 - \sigma^4$.

(c) $nU_n \xrightarrow{\mathcal{L}} \sigma^2(Z_1^2 - 1) + (\mu_4 - \sigma^4)(Z_2^2 - 1)$, where $Z_1$ and $Z_2$ are independent standard normals.

8. (a) Since the distribution of $X$ is symmetric about 0, we have $h_1(x) = Eh(x, Y) \equiv 0$, so $\sigma_1^2 = \text{Var} h_1(X) = 0$ and $\theta = Eh_1(X) = 0$. Since $h(X, Y)$ is not degenerate, the degeneracy is of order 1.

(b) From (51), we have $E[(xY + x^3Y^3)\varphi(Y)] = xEY\varphi(Y) + x^3EY^3\varphi(Y) = \lambda\varphi(x)$. Thus we see that $\varphi(x)$ must be a polynomial of degree 3 of the form, $\varphi(x) = ax^3 + bx$.

(c) Equating the coefficient of $x^3$, we have $\lambda a = aEY^6 + bEY^4$. Equating the coefficient of $x$, we have $\lambda b = aEY^4 + bEY^2$. Substituting $EY^2 = 1$, $EY^4 = 3$ and $EY^6 = 15$, we have the equations,

$$(15 - \lambda)a + 3b = 0$$
$$(1 - \lambda)b = 0$$

(d) These equations have a nontrivial solution if and only if the determinant is zero. This equation is $(15 - \lambda)(1 - \lambda) - 9 = 0$, or $\lambda^2 - 16\lambda + 6 = 0$. The roots are $\lambda_1 = 8 + \sqrt{58}$ and $\lambda_2 = 8 - \sqrt{58}$.

(e) $nU_n \xrightarrow{\mathcal{L}} (8 + \sqrt{58})(Z_1^2 - 1) + (8 - \sqrt{58})(Z_2^2 - 1)$, where $Z_1$ and $Z_2$ are independent standard normals.