# Homework Assignment 4

Due in class, Monday March 26th

SDS 384-11 Theoretical Statistics

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 $Ef(X) = \mu^2$ for all $P \in \mathcal{P}$ where $\mu$ is the mean of $P$, and $X$ is a random variable with distribution $P$. We must have $h(x)dP(x) = \mu^2$ for all distributions on the real line with mean . If $P$ is degenerate at a point $y$, this implies that $h(y) = y^2$ for all $y$. But if $P$ has mean zero ($\mu = 0$) and is not degenerate, then $h(x)dP(x) = x^2 dP(x) > 0 = \mu^2$. which is a contradiction.

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 $\leq \max(m_1, m_2)$. 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. Let $m = maxm_1, m_2$ and let $h(x_1, ..., x_m) = h_1(x_1, ..., x_{m_1}) + h_2(x_1, ..., x_{m_2})$. Then $Eh(X_1, ..., X_m) = g_1 + g_2$, showing that $g_1 + g_2$ is an estimable parameter with degree at most $m$.

   (b) Show $g_1 g_2$ is an estimable parameter with degree at most $m_1 + m_2$.
   Let $m = m_1 + m_2$ and let $h(x_1, ..., x_m) = h_1(x_1, ..., x_{m_1}) h_2(x_{m_1+1}, ..., x_{m_1+m_2})$. Then $Eh(X_1, ..., X_m) = g_1 g_2$, showing that $g_1 g_2$ ia an estimable parameter with degree at most $m$.

3. A continuous distribution with CDF $F(x)$, 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,

$$\theta(F) = \int (1 - F(x) - F(-x))^2 dF(x) \tag{1}$$

$$= \int ((1 - F(-x))^2 dF(x) - 2 \int (1 - F(-x)) F(x) dF(x) + \int F(x)^2 dF(x) \tag{2}$$

   as a nonparametric measure of how asymmetric the distribution is. 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.

   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 $U_n = \frac{1}{\binom{n}{3}} \sum_{i_1<i_2<i_3} h(X_{i_1}, X_{i_2}, X_{i_3})$.

**Many of you also expanded the last term out as $P(X_1 \leq X_3, X_2 \leq X_3)$. But note that since we have i.i.d random variables, this quantity is 1/3.**

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

   (a) Show that the corresponding U statistic is degenerate of order 1, i.e. $\xi_1 = 0$, but $\xi_2 > 0$. 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, $h_1(x) = Eh(x, Y) = xEY + (x^2 - \sigma^2)E(Y^2 - \sigma^2) = 0$, so $\xi_1 = Varh_1(X) = 0$. Since $h(X, Y)$ is not degenerate, $\xi_2 > 0$, so the degeneracy is of order 1.

   (b) Find the asymptotic distribution of $nU$.
   **Thanks to Mauricio, Jack, Rohit, Jiangang, Rimli, Yanxin.**
   Plugging $h(X_1, X_2)$, we have

$$U = \frac{\sum_{i \neq j} [X_i X_j + (X_i^2 - \sigma^2)(X_j^2 - \sigma^2)]}{n(n-1)}$$

   From the class note, we know

$$U_1 \equiv \frac{\sum_{i \neq j} X_i X_j}{n(n-1)} = \frac{(\sum X_i)^2 - \sum_i X_i^2}{n(n-1)} = \frac{(\sqrt{n}\bar{X}_n)^2 - \frac{\sum_i X_i^2}{n}}{n-1}$$

   Note $(\sqrt{n}\bar{X}_n)^2$ converges in distribution to $\sigma^2 Z_1^2$, where $Z_1$ is a standard normal distribution. $\frac{\sum_i X_i^2}{n}$ converges in probability to $\sigma^2$. Then by Slutsky theorem, we have $nU_1 \xrightarrow{d} \sigma^2(Z_1^2 - 1)$. Similarly, we have

$$U_2 \equiv \frac{\sum_{i \neq j} (X_i^2 - \sigma^2)(X_j^2 - \sigma^2)}{n(n-1)} = \frac{(\sum_i (X_i^2 - \sigma^2))^2 - \sum_i (X_i^2 - \sigma^2)^2}{n(n-1)}$$

   Note $\frac{(\sum_i (X_i^2 - \sigma^2))^2}{n}$ converges in distribution to $(\mu_4 - \sigma^4) Z_2^2$, where $Z_2$ is a standard normal distribution and $\mu_4 = E[X_i^4]$. $\frac{\sum_i (X_i^2 - \sigma^2)^2}{n}$ converges in probability to $\mu_4 - \sigma^4$. Then by Slutsky theorem, we have $nU_2 \xrightarrow{d} (\mu_4 - \sigma^4)(Z_2^2 - 1)$.
   Furthermore, by multivariate CLT, with $Y_i := X_i^2 - \sigma^2$

$$(\sqrt{n}\bar{X}_n, \sqrt{n}\bar{Y}_n) \xrightarrow{d} N(\mu, \Sigma)$$

   where $\mu_1 = 0, \mu_2 = 0$ and $\Sigma = \text{diag}(\sigma^2, \mu_4 - \sigma^4)$. Now a continuous mapping theorem, plus Slutsky's lemma gives us, $nU \xrightarrow{d} \sigma^2(Z_1^2 - 1) + (\mu_4 - \sigma^4)(Z_2^2 - 1)$ where $Z_1, Z_2$ are independent standard normals.

5. Look at the seminar paper "Probability Inequalities for Sums of Bounded Random Variables" by Wassily Hoeffding. It should be available via `lib.utexas.edu`. Read and reproduce the proof of equation 5.7 for large sample deviation of order $r$ U statistics.