DISTRIBUTION FREE TWO-SAMPLE METHODS FOR JUDGMENT POST-STRATIFIED DATA

Omer Ozturk

The Ohio State University

Supplementary Material

S1 Proof of Lemma 2

We first observe that the conditional expectation of T_{hq}, given the judgment ranks $\boldsymbol{R} = (R_1, \cdots, R_n)$ and $\boldsymbol{W} = (W_1, \cdots, W_m)$, is

$$E(T_{hq}|\boldsymbol{R}, \boldsymbol{W}) = N_h M_q \int (1 - G_{[q]}(y)) dF_{[h]}(y) = N_h M_q \tau_{[hq]}(F, G).$$

Using the iterative expectation, we obtain

$$\begin{aligned} E(T) &= E(ET|\boldsymbol{R}, \boldsymbol{W}) = E(\frac{1}{d_n d_m} \sum_{h=1}^{H} \sum_{q=1}^{Q} \frac{I_{hx} I_{qy} N_n M_q}{N_h M_q} \int (1 - G_{[q]}(y) df_{[h]}(y), \qquad \text{(S1.1)} \\ E(T) &= \frac{1}{HQ} \sum_{h=1}^{H} \sum_{q=1}^{Q} \tau_{[hq]}(F, G) = \tau_{[..]}(F, G) = \int (1 - G(y)) dF(y) = \int F(y) dG(y). \end{aligned}$$

This completes the proof of the expectation. For the proof of the variance, the conditional variance yields that

$$var(T) = Var(E(T|\boldsymbol{R}, \boldsymbol{W})) + E(var(T|\boldsymbol{R}, \boldsymbol{W})) = A_{n,m,[H,Q]}(F, G) + B_{n,m,[H,Q]}(F, G).$$

Note that from equation (S1.1) we have

$$\begin{aligned} A_{n,m,[H,Q]}(F, G) &= var(\frac{1}{d_n d_m} \sum_{h=1}^{H} \sum_{q=1}^{Q} I_{hx} I_{qy} \tau_{[hq]}(F, G)) \\ &= \sum_{h=1}^{H} \sum_{q=1}^{Q} \sum_{h'=1}^{H} \sum_{q'=1}^{Q} \tau_{[hq]}(F, G) \tau_{[h'q']}(F, G) cov(\frac{I_{hx} I_{qy}}{d_n d_m}, \frac{I_{h'x} I_{q'y}}{d_n d_m}). \end{aligned}$$

This sum can be partitioned into four different parts

$$\begin{aligned}
A_{n,m,[H,Q]}(F,G) &= \sum_{h=1}^{H}\sum_{q=1}^{Q}\tau^2_{[hq]}(F,G)cov(\frac{I_{hx}I_{qy}}{d_nd_m},\frac{I_{hx}I_{qy}}{d_nd_m}) \\
&+ \sum_{h=1}^{H}\sum_{q=1}^{Q}\sum_{q'\neq q}^{Q}\tau_{[hq]}(F,G)\tau_{[hq']}(F,G)cov(\frac{I_{hx}I_{qy}}{d_nd_m},\frac{I_{hx}I_{q'y}}{d_nd_m}) \\
&+ \sum_{h=1}^{H}\sum_{q=1}^{Q}\sum_{h'\neq h}^{H}\tau_{[hq]}(F,G)\tau_{[h'q]}(F,G)cov(\frac{I_{hx}I_{qy}}{d_nd_m},\frac{I_{h'x}I_{qy}}{d_nd_m}) \\
&+ \sum_{h=1}^{H}\sum_{q=1}^{Q}\sum_{q'\neq q}^{Q}\sum_{h'\neq h}^{H}\tau_{[hq]}(F,G)\tau_{[h'q']}(F,G)cov(\frac{I_{hx}I_{qy}}{d_nd_m},\frac{I_{h'x}I_{q'y}}{d_nd_m}).
\end{aligned}$$

Using the fact that I_{hx}/d_n, $h=1,\cdots,H$, and I_{qy}/d_m, $q=1,\cdots,Q$, are identically distributed, after some simplifications, $A_{n,m,[H,Q]}(F,G)$ can be written as

$$\begin{aligned}
A_{n,m,[H,Q]}(F,G) &= \sum_{h=1}^{H}\sum_{q=1}^{Q}\tau^2_{[hq]}(F,G)\left\{E(\frac{I_{1x}^2}{d_n^2})E(\frac{I_{1y}^2}{d_m^2})-\frac{1}{H^2Q^2}\right\} \\
&+ \left\{\sum_{h=1}^{H}\tau^2_{[h.]}(F,G)-\sum_{h=1}^{H}\sum_{q=1}^{Q}\tau^2_{[hq]}(F,G)\right\}\left\{E(\frac{I_{1x}^2}{d_n^2})E(\frac{I_{1y}I_{2y}}{d_m^2})-\frac{1}{H^2Q^2}\right\} \\
&+ \left\{\sum_{q=1}^{Q}\tau^2_{[.q]}(F,G)-\sum_{h=1}^{H}\sum_{q=1}^{Q}\tau^2_{[hq]}(F,G)\right\}\left\{E(\frac{I_{1y}^2}{d_m^2})E(\frac{I_{1x}I_{2x}}{d_n^2})-\frac{1}{H^2Q^2}\right\} \\
&+ \left\{\tau^2_{[..]}(F,G)-\sum_{q=1}^{Q}\tau^2_{[.q]}(F,G)-\sum_{h=1}^{H}\tau^2_{[h.]}(F,G)+\sum_{h=1}^{H}\sum_{q=1}^{Q}\tau^2_{[hq]}(F,G)\right\} \\
&\times \left\{E(\frac{I_{1x}I_{2x}}{d_n^2})E(\frac{I_{1y}I_{2y}}{d_m^2})-\frac{1}{H^2Q^2}\right\}
\end{aligned}$$

which completes the proof of $A_{n,m,[H,Q]}(F,G)$.

For the proof $B_{n,m,[H,Q]}(F,G)$, without loss of generality we assume that T is centered and write

$$T=\sum_{h=1}^{H}\sum_{q=1}^{Q}c_{hq}T_{hq}, \quad T_{hq}=\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{I(X_i\leq Y_j)-\tau_{[hq]}\right\}I(R_i=h)I(W_j=q),$$

where $c_{hq}=\frac{I_{hx}I_{qy}}{N_hM_qd_nd_m}$. To compute $B_{n,m,[H,Q]}(F,G)$, we first consider

$$var(T|\boldsymbol{W},\boldsymbol{R})=var\{\sum_{h=1}^{H}\sum_{q=1}^{Q}c_{hq}T_{hq}|\boldsymbol{W},\boldsymbol{R}\}=\sum_{h=1}^{H}\sum_{q=1}^{Q}\sum_{h'=1}^{H}\sum_{q'=1}^{Q}c_{hq}c_{h'q'}cov(T_{hq},T_{h'q'})|\boldsymbol{W},\boldsymbol{R}$$

We again partition this sum into four pieces

$$\begin{aligned}
var(T|\boldsymbol{W},\boldsymbol{R}) &= \sum_{h=1}^{H}\sum_{q=1}^{Q} c_{hq}^2 var(T_{hq}|\boldsymbol{R},\boldsymbol{W}) \\
&+ \sum_{h=1}^{H}\sum_{q=1}^{Q}\sum_{q'\neq q}^{Q} c_{hq}c_{hq'}cov(T_{hq},T_{hq'})|\boldsymbol{W},\boldsymbol{R} \\
&+ \sum_{h=1}^{H}\sum_{q=1}^{Q}\sum_{h'\neq h}^{H} c_{hq}c_{h'q}cov(T_{hq},T_{h'q})|\boldsymbol{W},\boldsymbol{R} \\
&+ \sum_{h=1}^{H}\sum_{q=1}^{Q}\sum_{h'\neq h}^{H}\sum_{q'\neq q}^{Q} c_{hq}c_{h'q'}cov(T_{hq},T_{h'q'})|\boldsymbol{W},\boldsymbol{R} \\
&= B_{1,n,m} + B_{2,n,m} + B_{3,n,m} + 0.
\end{aligned}$$

The last term in the above equation is zero since T_{hq}, $T_{h'q'}$ are conditionally independent given $\boldsymbol{R}$ and $\boldsymbol{W}$. Let $K_{ij}(h,q) = I(X_{[h]i} \le Y_{[q]j}) - \tau_{[hq]}$. Note that I_{hx}, $h = 1,\cdots,H$, and I_{qy}, $q = 1,\cdots,Q$, are identically distributed. We then simplify $B_{1,n,m}$

$$\begin{aligned}
E(B_{1,n,m}) &= E\left(\frac{I_{1x}^2 I_{1y}^2}{d_n^2 d_m^2 N_1 M_1}\right)\sum_{h=1}^{H}\sum_{q=1}^{Q}\left\{\tau_{[hq]}(F,G)\{1-\tau_{[hq]}(F,G)\}\right\} \\
&- E\left(\frac{I_{1x}^2 I_{1y}^2}{d_n^2 d_m^2 N_1 M_1}\right)\left(\xi_{[QH]}(G,F) - \xi_{[HQ]}(F,G)\right) \\
&+ E\left(\frac{I_{1x}^2 I_{1y}^2}{d_n^2 d_m^2 N_1}\right)\xi_{[QH]}(G,F) + E\left(\frac{I_{1x}^2 I_{1y}^2}{d_n^2 d_m^2 M_1}\right)\xi_{[HQ]}(F,G) \\
&= E\left(\frac{I_{1x}^2 I_{1y}^2}{d_n^2 d_m^2 N_1 M_1}\right)\left\{\tau_{..}(F,G) - \gamma_{HQ}(F,G) - \xi_{[QH]}(G,F) - \xi_{[HQ]}(F,G)\right\} \\
&+ E\left(\frac{I_{1x}^2 I_{1y}^2}{d_n^2 d_m^2 N_1}\right)\xi_{[QH]}(G,F) + E\left(\frac{I_{1x}^2 I_{1y}^2}{d_n^2 d_m^2 M_1}\right)\xi_{[HQ]}(F,G).
\end{aligned}$$

With similar argument the expected value of $B_{2,n,m}$ reduces to

$$E(B_{2,n,m}) = E\left(\frac{I_{1x}^2 I_{1y} I_{2y}}{d_n^2 d_m^2 N_1}\right)\sum_{h=1}^{H}\sum_{q=1}^{Q}\sum_{q'\neq q}^{Q} E(K_{12}(h,q)K_{13}(h,q'))$$

$$\begin{aligned}
&= E\left(\frac{I_{1x}^2 I_{1y} I_{2y}}{d_n^2 d_m^2 N_1}\right)\sum_{h=1}^{H}\sum_{q=1}^{Q}\sum_{q'\neq q}^{Q}\int\{1-F_{[q]}(y)-\tau_{[hq]}(F,G)\}\{1-F_{[q']}(y)-\tau_{[hq']}(F,G)\}dF_{[h]}(y) \\
&= E\left(\frac{I_{1x}^2 I_{1y} I_{2y}}{d_n^2 d_m^2 N_1}\right)\left\{Q^2 H\int(1-G(y))^2 dF(y) - \gamma_{[H.]}(F,G) - \xi_{[QH]}(G,F)\right\}. \\
&= E\left(\frac{I_{1x}^2 I_{1y} I_{2y}}{d_n^2 d_m^2 N_1}\right)\left\{Q^2 H\theta(G,F) - \gamma_{[H.]}(F,G) - \xi_{[QH]}(G,F)\right\}.
\end{aligned}$$

With similar computations, we also obtain

$$E(B_{3,n,m}) = E\left(\frac{I_{1y}^2 I_{1x} I_{2x}}{d_n^2 d_m^2 M_1}\right)\left\{H^2 Q\theta(F,G) - \gamma_{[.Q]}(F,G) - \xi_{[HQ]}(F,G)\right\}.$$

The proof is completed by combining the expressions in $B_{1,n,m}$, $B_{2,n,m}$ and $B_{3,n,m}$.

S2 Proof of Corollary 2

We first show that $(n+m)a_k(n,m,H,Q)$, $k=1,\cdots,4$, in Lemma 2 are asymptotically negligible. Let n_0 be the minimum of n and m. Consider

$$
\begin{aligned}
&\lim_{n_0\to\infty}(n+m)a_1(n,m,H,Q) = \lim_{n_0\to\infty}(n+m)\left\{E(\frac{I_{1x}I_{2x}}{d_n^2})E(\frac{I_{1y}I_{2y}}{d_m^2}) - \frac{1}{H^2Q^2}\right\} \\
= &\lim_{n_0\to\infty}(n+m)\left[\left\{\frac{1}{H^2} - \frac{1}{H^2(H-1)}\sum_{h=1}^{H-1}(\frac{h}{H})^{n-1}\right\}\left\{\frac{1}{Q^2} - \frac{1}{Q^2(Q-1)}\sum_{q=1}^{Q-1}(\frac{q}{Q})^{m-1}\right\} - \frac{1}{Q^2H^2}\right] \\
= &-\frac{1}{H^2Q^2(Q-1)}\sum_{q=1}^{Q-1}\lim_{n_0\to\infty}(\frac{q}{Q})^{m-1}(n+m) \\
&-\frac{1}{Q^2H^2(H-1)}\sum_{h=1}^{H-1}\lim_{n_0\to\infty}(\frac{h}{H})^{n-1}(n+m) \\
&+\frac{1}{H^2(H-1)Q^2(Q-1)}\sum_{q=1}^{Q-1}\lim_{n_0\to\infty}(\frac{q}{Q})^{m-1}\sum_{h=1}^{H-1}\lim_{n_0\to\infty}(\frac{h}{H})^{n-1}(n+m) = 0.
\end{aligned}
$$

In a similar fashion, it can be shown that $(n+m)a_k(n,m,H,Q)$, $k=2,\cdots,4$, also converge to zero as the minimum of n and m goes to infinity. Hence, we proved that $A_{n,m,[H,Q]}(F,G)$ is asymptotically zero.

In expression $B_{n,m,[H.Q]}(F,G)$, we first show that the $(n+m)b_k(n,m,H,Q)$, $k=1,2,3$, converge to zero. Note that I_{1x}, I_{1y}, d_n, d_m, N_1 and M_1 are positive random variables. It is then justified to interchange the limit with the expectation in the following expression

$$
\lim_{n_0\to\infty}E(\frac{(n+m)I_{1x}^2}{d_n^2N_1}) = E\left\{\lim_{n_0\to\infty}\frac{n+m}{n}\lim_{n_0\to\infty}\frac{I_{1x}^2}{d_n^2}\lim_{n_0\to\infty}\frac{n}{N_1}\right\} = \frac{H}{\lambda H^2}.
$$

Using Lemma 1, we show that

$$
\begin{aligned}
\lim_{n_0\to\infty}E(\frac{I_{1x}^2}{d_n^2}) &= \lim_{n_0\to\infty}\left\{\frac{1}{H^2}(1+\sum_{h=1}^{H-1}(\frac{h}{H})^{n-1})\right\} = 1/H^2, \\
\lim_{n_0\to\infty}E\left\{\frac{I_{1x}I_{2x}}{d_n^2}\right\} &= \lim_{n_0\to\infty}\left\{\frac{1}{H^2} - \frac{1}{H^2(H-1)}\sum_{q=1}^{H-1}(\frac{h}{H})^{n-1}\right\} = 1/H^2, \\
\lim_{n_0\to\infty}E(\frac{I_{1x}^2}{d_x^2N_1}) &= E\left\{\lim_{n_0\to\infty}\frac{I_{1x}^2}{d_n^2}\lim_{n_0\to\infty}\frac{n}{N_1}\lim_{n_0\to\infty}\frac{1}{n}\right\} = 0.
\end{aligned}
$$

Similar results can be established for the limits of the expected values of Y-sample sample sizes. Only difference is that the λ and H in the above equations will be replaced with $1-\lambda$ and Q in the Y-sample sample sizes. Using these limits, we show that

$$
\begin{aligned}
\lim_{n_0\to\infty}(n+m)b_k(n,m,H,Q) &= 0, \quad k=1,2,3, \\
\lim_{n_0\to\infty}(n+m)b_4(n,m,H,Q) &= \frac{1}{(1-\lambda)QH^2}
\end{aligned}
$$

and

$$
\lim_{n_0\to\infty}(n+m)b_5(n,m,H,Q) = \frac{1}{\lambda Q^2H}.
$$

This completes the proof.

S3 Proof of Lemma 3

Without loss of generality, we consider the centered version of T

$$T = \sum_{h=1}^{H} \sum_{q=}^{Q} \frac{I_{hx} I_{qy}}{d_n d_m N_h M_q} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\{ I(X_i \leq Y_j) - \tau_{[h,q]}(F,F) \right\} I(R_i = h) I(W_j = q).$$

Let $\psi_1(x, R_i = h) = E(T|X_i = x, R_i = h, d_n, N_h, I_{hx})$ and $\psi_2(y, W_j = q) = E(T|Y_j = y, W_j = q, d_m, M_q, I_{qy}) - 1/2$. Then the projection of T, T_P, is given by

$$T_p = \sqrt{n+m} \left\{ \sum_{h=1} \sum_{i=1}^{n} \psi_1(X_i, R_i = h) + \sum_{q=1} \sum_{j=1}^{n} \psi_1(Y_j, W_j = q) \right\},$$

where

$$\begin{aligned} \psi_1(X_i, R_i = h) &= \frac{I_{hx}}{d_n N_h} (1 - F(X_i) - \bar{\tau}_{[h.]}(F,F)) I(R_i = h) \\ \psi_2(Y_j, W_j = q) &= \frac{I_{qy}}{d_m M_q} (F(Y_j) - \bar{\tau}_{[.q]}(F,F)) I(W_j = q), \end{aligned}$$

and $\bar{\tau}_{[h.]} = \sum_{q=1}^{Q} \tau_{[hq]}(F,F)/Q$. We finish the proof by observing $E(\sqrt{n+m}(T - T_p)^2 = var(\sqrt{n+m}(T) - var(\sqrt{n+m}T_p)$ goes to zero as n_0 approaches to infinity.

Let $\bar{\boldsymbol{\psi}}_{\mathbf{1}} = (\bar{\psi}_{1,1}, \ldots, \bar{\psi}_{1,H})^\top$ and $\bar{\boldsymbol{\psi}}_{\mathbf{2}} = (\bar{\psi}_{2,1}, \ldots, \bar{\psi}_{2,Q})^\top$, where

$$\bar{\psi}_{1,h} = \frac{\sum_{i=1}^{n} \psi_1(X_i, R_i = h)}{\sqrt{N_h}}, \text{ and } \bar{\psi}_{2,q} = \frac{\sum_{j=1}^{m} \psi_2(X_i, W_j = q)}{\sqrt{M_q}}.$$

Using Theorem 3.2 in Gutts (2005, p. 347), or modifying the proof of Theorem 1 in Ozturk (2014) to a two sample problem, one can show that $\bar{\boldsymbol{\psi}}_{\mathbf{1}}$ and $\bar{\boldsymbol{\psi}}_{\mathbf{2}}$ converge to H- and Q-dimensional normal random vectors with mean zero and variances $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, respectively, where

$$\mathbf{\Sigma}_1 = diag\left(var(1 - F(X_{[h]}) - \bar{\tau}_{[h.]}(F,F)))\right) \text{ and } \mathbf{\Sigma}_2 = diag\left(var(F(Y_{[q]}) - \bar{\tau}_{[.q]}) \right).$$

Let $\boldsymbol{U}(\boldsymbol{N}, \boldsymbol{I}_x) = (\frac{I_{1x}\sqrt{n}}{d_n\sqrt{N_1}}, \cdots, \frac{I_{Hx}\sqrt{n}}{d_n\sqrt{N_H}})$ and $\boldsymbol{U}(\boldsymbol{M}, \boldsymbol{I}_y) = (\frac{I_{1y}\sqrt{m}}{d_m\sqrt{M_1}}, \cdots, \frac{I_{Qy}\sqrt{m}}{d_m\sqrt{M_Q}})$. For large n and m, it is easy to observe that $\boldsymbol{U}(\boldsymbol{N}, \boldsymbol{I}_x)$ and $\boldsymbol{U}(\boldsymbol{M}, \boldsymbol{I}_y)$ converge in probability to $(1/\sqrt{H}, \cdots, 1/\sqrt{H})$ and $(1/\sqrt{Q}, \cdots, 1/\sqrt{Q})$, respectively. As n_0 goes to infinity we observe that

$$\sqrt{\frac{n+m}{n}} \boldsymbol{U}(\boldsymbol{N}, \boldsymbol{I}_x) \bar{\boldsymbol{\psi}}_{\mathbf{1}} + \sqrt{\frac{n+m}{m}} \boldsymbol{U}(\boldsymbol{M}, \boldsymbol{I}_y) \bar{\boldsymbol{\psi}}_{\mathbf{2}}$$

converges to a normal distribution with mean zero and variance $\sigma^2_{H,Q}$, where

$$\sigma^2_{H,Q} = \frac{1}{\lambda} \left\{ 1/3 - \frac{1}{H} \sum_{h=1} \left(\int F(y) dF_{[h]}(y) \right)^2 \right\} + \frac{1}{(1-\lambda)} \left\{ 1/3 - \frac{1}{Q} \sum_{q=1} \left(\int F(y) dF_{[q]}(y) \right)^2 \right\}.$$

S4 Proof of Lemma 4

Let $A_{1,k-1,t}$ be the event that $N_1 = n_1$, $M_1 = m_1$ and there exist exactly $k-1$ matching non-empty judgment classes in X- and Y-samples ($N_{i_2} > 0, \cdots, N_{i_k} > 0; M_{i_2} > 0, \cdots, M_{i_k} > 0$) and t non-matching non-empty judgment classes in X-samples ($N_{i_{k+1}} > 0, \cdots, N_{i_{k+t}} > 0; M_{i_{k+1}} = 0, \cdots, M_{i_{k+t}} = 0$)

$$A_{1,k-1,t} = \left\{ \begin{array}{l} N_1 = n_1, M_1 = m_1, N_{i_2} > 0, \cdots, N_{i_k} > 0 \cdots, N_{i_{k+t}} > 0, \text{ and} \\ M_{i_2} > 0 \cdots, M_{i_k} > 0, M_{i_{k+1}} = 0, \cdots, M_{i_{k+t}} = 0, M_{i_{k+t+1}} \geq 0 \cdots, M_{i_H} \geq 0 \end{array} \right\},$$

where $i_2, \cdots, i_H$ is a permutation of integers $(2, \cdots, H)$ and $1 \leq k \leq k^*$, $k^* = min(n, m, H)$. Note that $N_h, h = 1, \cdots, H$, and $M_h, h = 1, \cdots, H$, are identically distributed. The probability of the event $A_{1,k-1,t}$ can be computed by considering all possible combinations yielding the event in set $A_{1,k-1,t}$

$$P(A_{1,k-1,t}) = C_{H,k,t} P \left(\begin{array}{l} N_1 = n_1, M_1 = m_1, N_j > 0, j = 2, \cdots, k+t, M_i > 0, i = 2, \cdots, k \\ M_r = 0, r = k+1, \cdots, k+t; M_z \geq 0, z = k+t+1, \cdots, H \end{array} \right),$$

where

$$C_{H,k,t} = \binom{H-1}{k-1} \binom{H-k}{t}$$

and $0 \leq t \leq t^*$ with $t^* = min(n-1, H-k)$. Let

$$N_{[a:b]} = \{N_a > 0, \cdots, N_b > 0\}, \quad N^*_{[a:b]} = \{N_a = 0, \cdots, N_b = 0\}, \quad N^+_{[a,b]} = \{N_a \geq 0, \cdots, N_b \geq 0\}.$$

We also use equivalent definitions for $M_{[a:b]}$, $M^*_{[a:b]}$ and $M^+_{[a,b]}$. Since $(N_1, \cdots, N_H)$ and $(M_1, \cdots, M_H)$ are independent, we have

$$\begin{aligned} P(A_{1,k-1,t}) &= \binom{H-1}{k-1} \binom{H-k}{t} P(N_1 = n_1; N_{[2:k+t]}; N^*_{[k+t+1:H]}) \\ &\times \quad P(M_1 = m_1; M_{[2:k]}; M^*_{[k+1:k+t]}; M^+_{[k+t+1:H]}). \end{aligned}$$

Since M_h, $h = 1, \cdots, H$, are identically distributed, we can rearrange the subscript in sets M^+ and M^* as follows

$$P(M_1 = m_1; M_{[2:k]}; M^*_{[k+1:k+t]}; M^+_{[k+t+1:H]}) = P(M_1 = m_1; M_{[2:k]}; M^+_{[k+1:H-t]}; M^*_{[H-t+1:H]}).$$

By conditioning on the given value of $N_1 = n_1$ we write

$$\begin{aligned} P(P(N_1 &= n_1; N_{[2:k+t]}; N^*_{[k+t+1:H]}) = P(N_{[2:k+t]}; N^*_{[k+t+1:H]} | N_1 = n_1) P(N_1 = n_1) \\ &= \sum_{n_{[2:k+t]}} \binom{n-n_1}{n_2, \cdots, n_{k+t}} \binom{n}{n_1} \frac{1}{H^n}, \quad 1 \leq n_1 \leq n-k-t+1 \\ &= \sum_{j=1}^{k+t-1} (-1)^{j-1} \binom{k+t-1}{j-1} (k+t-j)^{n-n_1} \binom{n}{n_1} H^{-n}, 1 \leq n_1 \leq n-k-t+1, \end{aligned}$$

where the notation $\sum_{n_{[2:k+t]}}$ indicates the sum over the index set $\{n_2 > 0, \cdots, n_{k+r} > 0\}$.

We now derive a similar expression for the probabilities in Y-sample sample size vector. Again we condition on the given value of $M_1 = m_1$ to write

$$\begin{aligned}
& P(M_1 = m_1; M_{[2:k]}; M^{+}_{[k+1:H-t]}; M^{*}_{[H-t+1:H]}) \\
= \; & P(M_{[2:k]}; M^{+}_{[k+1:H-t]}; M^{*}_{[H-t+1:H]} | M_1 = m_1) P(M_1 = m_1) \\
= \; & \sum_{u=0}^{H-k-t} \binom{H-k-t}{u} P(M_{[2:k+u]}; M^{*}_{[k+u+1:H]} | M_1 = m_1) P(M_1 = m_1), \quad 1 \le m_1 \le m-k-u+1 \\
= \; & \sum_{u=0}^{u^*} \binom{H-k-t}{u} \sum_{m_{[2:k+u]}} \binom{m-m_1}{m_2, \cdots, m_{k+u}} \binom{m}{m_1} \frac{1}{H^m}, \quad 1 \le m_1 \le m-k-u+1 \\
= \; & \sum_{u=0}^{u^*} \binom{H-k-t}{u} \sum_{i=1}^{k+u-1} (-1)^{i-1} \binom{k+u-1}{i-1} \frac{\binom{m}{m_1}}{H^m} (k+u-i)^{m-m_1}, \quad 1 \le m_1 \le m-k-u+1,
\end{aligned}$$

where $u^* = min(m-1, H-k-t)$. The expected value $\frac{I_{1x} I_{1y} J^a_{1x} J^b_{1y}}{d^2_{nm}}$ can be computed by using $P(A_{1,k-1,t})$ and appropriate limits in the summation indexes

$$\begin{aligned}
& E(\frac{I_{1x} I_{1y} J^a_{1x} J^b_{1y}}{d^2_{nm}}) = \sum_{k=1}^{k^*} \sum_{t=0}^{tm} \sum_{n_1=1} \sum_{m_1=1} \frac{1}{k^2 n_1^a m_1^b} P(A_{1,k-1,t}) \\
= \; & \sum_{k=1}^{k^*} \sum_{t=0}^{t^*} \frac{\binom{H-1}{k-1} \binom{H-k}{t}}{H^{n+m} k^2} \left\{ \sum_{n_1=1}^{n-k-t+1} \sum_{j=1}^{k+t-1} (-1)^{j-1} \binom{k+t-1}{j-1} \binom{n}{n_1} (k+t-j)^{n-n1} / n_1^a \right. \\
\times \; & \left. \sum_{u=0}^{u^*} \binom{H-k-t}{u} \sum_{m_1=1}^{m-k-u+1} \sum_{i=1}^{k+u-1} (-1)^{i-1} \binom{k+u-1}{i-1} \binom{m}{m_1} (k+u-i)^{m-m_1} / m_1^b \right\}.
\end{aligned}$$

S5 Proof of Lemma 5

First we consider the expected value

$$\begin{aligned}
E(T_\omega) \; &= \; E\{E(T_\omega | \boldsymbol{R}, \boldsymbol{W})\} = E\left\{ \sum_{h=1}^{H} \frac{\omega_h I_{hx} I_{hy} N_h M_h}{N_h M_h} \int \{1 - G_{[h]}(y)\} dF_{[h]}(y) \right\} \\
&= \; \sum_{h=1}^{H} E\left(\omega_h I_{hx} I_{hy}\right) \int \{1 - G_{[h]}(y)\} dF_{[h]}(y) = E\left(\omega_1 I_{1x} I_{1y}\right) \sum_{h=1}^{H} \int \{1 - G_{[h]}(y)\} dF_{[h]}(y).
\end{aligned}$$

The last equality follows form the fact that $\omega_h I_{hx} I_{hy}$, $h = 1, \cdots, H$, are identically distributed. Let $a = E(\omega_h I_{1x} I_{1y}) = \cdots = E(\omega_H I_{Hx} I_{Hy})$. We have from the equation below

$$Ha = E\left\{ \sum_{h=1}^{H} \omega_h I_{hx} I_{hy} \right\} = 1$$

that $a = \frac{1}{H}$. This completes the proof of the expectation.

For the proof of the variance we consider the conditional variance formula

$$V(T_\omega) = V(E(T_\omega | \boldsymbol{R}, \boldsymbol{W}) + E(V(T_\omega | \boldsymbol{R}, \boldsymbol{W})).$$

The first term in the above equation is zero since the conditional expectation is constant. In the second term, the conditional variance is the variance of the Mann-Whitney Wilcoxon rank-sum test statistic

$$V(T_\omega|\boldsymbol{R},\boldsymbol{W}) = \sum_{h=1}^{H} \frac{\omega_h^2 I_{hx}^2 I_{hy}^2 N_h M_h (N_h + M_h + 1)}{12 N_h^2 M_h^2}.$$

Since $I_{hx}^2 = I_{hx}$ the variance of T_ω becomes

$$V(T_\omega) = \frac{H}{12}\left\{ E\left(\frac{\omega_1^2 I_{1x} I_{1y}}{N_1}\right) + E\left(\frac{\omega_1^2 I_{1x} I_{1y}}{M_1}\right) + E\left(\frac{\omega_1^2 I_{1x} I_{1y}}{N_1 M_1}\right)\right\}.$$

Bibliography

Gut, A. (2005). Probability: *A Graduate Course*, Springer, New York.

Ozturk, O. (2014). Statistical inference for population quantiles and variance in judgment post-stratified samples. *Computational Statistics and Data Analysis* **77**, 188-205.