Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Optimal tightening of the KWW joint confidence region for a ranking

Tommy Wright[1,2]

Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, 20233, United States of America

ARTICLE INFO

Keywords:
Independence (Šidák) correction
Official statistics
Ranking
Tightness

ABSTRACT

Klein, Wright, and Wieczorek (2020), hereafter KWW, constructs a simple novel measure of uncertainty for an estimated ranking using a joint confidence region for the true ranking of K populations. In this current paper, our proposed framework permits some control over the amount of uncertainty and tightness in various portions of the estimated ranking with an optimal allocation of sample among the K populations.

1. Introduction

The KWW joint confidence region for the unknown overall true ranking, in Klein et al. (2020), is constructed as follows: by observing how K known joint confidence intervals for K means overlap or not, by obtaining a confidence set for each population rank, and ultimately obtaining the joint confidence region for the overall true ranking. We tighten (Definition 1) this joint confidence region with an exact optimal allocation of sample among the K populations.

Assume $K(\geq 2)$ disjoint finite populations where N_k is the known number of units in population k for $k = 1, 2, \ldots, K$. Let $N = \sum_{k=1}^{K} N_k$, and let Y_{ki} be the fixed unknown value of interest for the ith unit in population k, where the kth population mean is $\bar{Y}_k = \sum_{i=1}^{N_k} Y_{ki}/N_k$ and variance is $S_k^2 = \sum_{i=1}^{N_k}(Y_{ki} - \bar{Y}_k)^2/(N_k - 1)$. The desire is to rank the K populations from smallest to largest based on the unknown values of the ordered population means.

For any unit in the population (sample) setting, we represent its value of interest using upper (lower) case Y (y). If $y_{k1}, \ldots, y_{kn_k}$ are the observed values in a simple random sample of size n_k from the kth population where the sample mean $\bar{y}_k = \sum_{i=1}^{n_k} y_{ki}/n_k$ is an estimator of $\bar{Y}_k$, we rank the K populations based on the estimated ranking of values, $\bar{y}_1, \ldots, \bar{y}_K$, i.e.,

$$\bar{y}_{(1)} \leq \bar{y}_{(2)} \leq \cdots \leq \bar{y}_{(k)} \leq \cdots \leq \bar{y}_{(K)}. \tag{1}$$

KWW uses a collection (family) of K joint confidence intervals for $\bar{Y}_1, \ldots, \bar{Y}_K$ to form the basis for the uncertainty measure presented. For KWW methodology, we summarize in Section 2, present 3 new properties (Section 2.3), and summarize 4 uses (Section 2.5). In Section 3, we define tightness for a joint confidence region. In Section 4, we optimize an objective function with an algorithm aimed at tightening the KWW joint confidence region.

Selected Ranking Literature. KWW (2020) cites three different approaches (with references) relating to ranks and rankings from the literature: (a) frequentist; (b) Bayesian; and (c) the bootstrap. KWW (2020) goes beyond answering the question, "How good is the estimated individual rank of a specific state?" to answering the question, "How good is the overall estimated ranking, hereafter estimated ranking, of K populations?" KWW presents a frequentist (as well as how to adapt for Bayesian inference) quantification of this uncertainty via a joint confidence region for the true ranking (Section 2). In the Appendix (Concluding Remarks: Remark 5.5),

E-mail address: tommy.wright@census.gov.

[1] Adjunct faculty at Georgetown University.

[2] The author is very grateful for the useful comments from the AE and two referees. The views expressed are those of the author and not those of the U.S. Census Bureau.

https://doi.org/10.1016/j.spl.2024.110288
Received 16 February 2024; Received in revised form 15 October 2024; Accepted 17 October 2024

we note related work for a "simultaneous confidence set" for a ranking by Mogstad et al. (2024). As noted in KWW (2020), Wright (2024), and Hall and Miller (2009), uncertainty (due to sampling variability) is often less at the extremes of the estimated ranking as compared with middle portion. We show how to tighten (Definition 1) the joint confidence region using an optimal allocation of an overall sample size n among the K populations, especially the middle portion.

2. Main results of KWW

The unknown true ranking is $(r_1, \ldots, r_K)$, where r_k, the rank of kth population, is

$$r_k = \sum_{j=1}^{K} I(\bar{Y}_j \le \bar{Y}_k) = 1 + \sum_{\{j\,:\,j\neq k\}} I(\bar{Y}_j \le \bar{Y}_k), \quad \text{for} \quad k = 1, \ldots, K. \tag{2}$$

The estimated ranking is $(\hat{r}_1, \ldots, \hat{r}_K)$, where $\hat{r}_k$, estimated rank of kth population, is

$$\hat{r}_k = 1 + \sum_{\{j\,:\,j\neq k\}} I(\bar{y}_j \le \bar{y}_k), \quad \text{for} \quad k = 1, 2, \ldots, K. \tag{3}$$

The values of $\bar{Y}_1, \ldots, \bar{Y}_K$ are unknown, and KWW assumes for each k that we know real numbers $L_k < U_k$ such that $\bar{Y}_k \in (L_k, U_k)$. When KWW constructs the joint confidence region shown in Section 2.4, KWW replaces the intervals (L_k, U_k) with joint confidence intervals and the Main Result (Section 2.1) is then used to obtain an uncertainty (or probability) statement. See Wright (2024) for an example clarifying the basic KWW underlying concept.

Notation. For each $k \in \{1, 2, \ldots, K\}$, and $j \in I_k = \{1, 2, \ldots, K\}\backslash\{k\}$, KWW defines three sets

1. $j \in \Lambda_{Lk}$ if and only if (L_j, U_j) lies to the left of (L_k, U_k);
2. $j \in \Lambda_{Rk}$ if and only if (L_j, U_j) lies to the right of (L_k, U_k);
3. $j \in \Lambda_{Ok}$ if and only if $(L_j, U_j) \cap (L_k, U_k) \neq \emptyset$.

It follows that Λ_{Lk}, Λ_{Rk}, and Λ_{Ok} are mutually exclusive, and $\Lambda_{Lk} \cup \Lambda_{Rk} \cup \Lambda_{Ok} = I_k$.

2.1. Main result

Where $|A|$ is the size of finite set A, KWW shows for $k \in \{1, 2, \ldots, K\}$, that

$$r_k \in \left\{|\Lambda_{Lk}| + 1, |\Lambda_{Lk}| + 2, |\Lambda_{Lk}| + 3, \ldots, |\Lambda_{Lk}| + |\Lambda_{Ok}| + 1\right\}. \tag{4}$$

2.2. Joint confidence region for an overall true ranking

Henceforth, KWW assumes that $\{\ (L_1, U_1),\ (L_2, U_2),\ \ldots,\ (L_K, U_K)\ \}$ is a collection (or family) of confidence intervals for the unknown parameters $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_K$, respectively, and that the joint (or familywise) coverage probability of these intervals is greater than or equal to $1 - \alpha$. Thus,

$$\left\{(r_1, \ldots, r_K) : r_k \in \left\{|\Lambda_{Lk}| + 1, |\Lambda_{Lk}| + 2, |\Lambda_{Lk}| + 3, \ldots, |\Lambda_{Lk}| + |\Lambda_{Ok}| + 1\right\} \text{ for } k = 1, \ldots, K\right\} \tag{5}$$

is a **joint confidence region** (or set) for the true ranking $(r_1, \ldots, r_K)$ having joint coverage probability of at least $1 - \alpha$. We refer to the following set as a **marginal confidence set** for r_k: $\{|\Lambda_{Lk}| + 1, |\Lambda_{Lk}| + 2, |\Lambda_{Lk}| + 3, \ldots, |\Lambda_{Lk}| + |\Lambda_{Ok}| + 1\}$.

If $(L_1, U_1), \ldots, (L_K, U_K)$ are constructed such that the estimator $\bar{y}_k \in (L_k, U_k)$ for all k with probability 1 (which is the case with the KWW approach), then the estimated ranking $(\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_K)$ is contained in the joint confidence region (5) with probability 1.

In general, the joint confidence region in (5) contains more than one possible true ranking. However, if the values of $\bar{Y}_k$ are sufficiently different from each other such that $(L_k, U_k) \cap (L_{k'}, U_{k'}) = \emptyset$ for all $k \neq k'$ and $k = 1, 2, \ldots, K$, then it follows immediately that the joint confidence region contains only one possible true ranking, and it is the estimated ranking $(\hat{r}_1, \ldots, \hat{r}_K)$; when this happens, we would have the "tightest" (see Definition 1) possible joint confidence region.

2.3. Properties of joint confidence region and a marginal confidence set

Proofs of three new properties are immediate and straightforward (MOE means margin of error).

P-1: If $|\Lambda_{Ok}| = 0$, the marginal confidence set for r_k is $\{|\Lambda_{Lk}| + 1\}$.

P-2: If $|\Lambda_{Ok}| = 0\ \forall\ k$, the joint confidence region only contains the estimated ranking.

P-3: If $|\bar{y}_j - \bar{y}_k| \ge MOE_j + MOE_k$ for all $j \neq k$, the marginal confidence set for r_k is $\{|\Lambda_{Lk}| + 1\}$. Here, $MOE_j = z_{\frac{\gamma}{2}} SE_j$, and $z_{\frac{\gamma}{2}}$ is $1 - \frac{\gamma}{2}$ quantile of standard normal.

Table 1
Computation Details for 90% Joint Confidence Region for True Ranking (Y = Travel Time To Work Data in Minutes).
Data Source: 1-Year American Community Survey (2011), Ranking Table R0801, U.S. Census Bureau.

$\hat{r}_k$	State (k)	$\bar{y}_k$	SE_k	90% Joint Confidence Intervals for $\bar{Y}_k$'s	90% Joint Confidence Region for True Ranking
9	$Maryland\ (MD)$	32.2	0.1	(31.9, 32.5)	{9}
8	$New\ York\ (NY)$	31.5	0.1	(31.2, 31.8)	{8}
7	$New\ Jersey\ (NJ)$	30.5	0.1	(30.2, 30.8)	{6, 7}
6	$District\ of\ Columbia\ (DC)$	30.1	0.3	(29.3, 30.9)	{6, 7}
5	$Illinois\ (IL)$	28.2	0.1	(27.9, 28.5)	{3, 4, 5}
4	$Massachusetts\ (MA)$	28.0	0.1	(27.7, 28.3)	{3, 4, 5}
3	$Virginia\ (VA)$	27.7	0.1	(27.4, 28.0)	{2, 3, 4, 5}
2	$Georgia\ (GA)$	27.1	0.2	(26.6, 27.6)	{1, 2, 3}
2	$California\ (CA)$	27.1	0.1	(26.9, 27.3)	{1, 2}

Table 1a
State k Details for $r_k \in \{|\Lambda_{Lk}|+1, |\Lambda_{Lk}|+2, |\Lambda_{Lk}|+3, \ldots, |\Lambda_{Lk}|+|\Lambda_{Ok}|+1\}$.

$\hat{r}_k$	k	Λ_{Lk}	Λ_{Ok}	90% Joint Confidence Region for True Ranking
9	MD	$\{NY, NJ, DC, IL, MA, VA, GA, CA\}$	$\emptyset$	{9}
8	NY	$\{NJ, DC, IL, MA, VA, GA, CA\}$	$\emptyset$	{8}
7	NJ	$\{IL, MA, VA, GA, CA\}$	$\{DC\}$	{6, 7}
6	DC	$\{IL, MA, VA, GA, CA\}$	$\{NJ\}$	{6, 7}
5	IL	$\{GA, CA\}$	$\{MA, VA\}$	{3, 4, 5}
4	MA	$\{GA, CA\}$	$\{IL, VA\}$	{3, 4, 5}
3	VA	$\{CA\}$	$\{IL, MA, GA\}$	{2, 3, 4, 5}
2	GA	$\emptyset$	$\{VA, CA\}$	{1, 2, 3}
2	CA	$\emptyset$	$\{GA\}$	{1, 2}

2.4. KWW joint confidence region construction

KWW (2020) constructs a joint confidence region with an Independence (or Šidák (1967) Correction, assumes $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_K$ are independently distributed where $\bar{y}_k \sim N(\bar{Y}_k, SE_k)$ for k with $\bar{Y}_k$ unknown and standard error SE_k known by estimation.

With the framework of Section 4.1 in this paper, SE_k is computed by

$$SE_k = \sqrt{\left(\frac{N_k - n_k}{N_k}\right)\frac{s_k^2}{n_k}} \quad \text{and} \quad s_k^2 = \sum_{i=1}^{n_k}(y_{ki} - \bar{y}_k)^2/(n_k - 1).$$

where $y_{k1}, y_{k2}, \ldots, y_{kn_k}$ are the observed values in a simple random sample from population k.

To construct a $100(1-\alpha)\%$ joint confidence region, KWW considers joint confidence intervals whose joint coverage is shown to equal $1-\alpha$, where $\gamma = 1-(1-\alpha)^{1/K}$:

$$\left(\bar{y}_k - z_{\frac{\gamma}{2}} SE_k, \bar{y}_k + z_{\frac{\gamma}{2}} SE_k\right), \quad \text{for} \quad k = 1, 2, \ldots, K. \tag{6}$$

Table 1 shows the Independence corrected 90% joint confidence intervals for $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_9$ as given by (6) with overall $\alpha = 0.10$. (In some cases, it is convenient to show that k ranges over the names of the states rather than over the integers $1, \ldots, K$.) To construct the joint confidence region, we "correct" the level α for each state so that our "overall" level for the $K = 9$ joint confidence intervals is $\alpha = 0.10$. We proceed as follows using the Y = travel time to work data for 9 states in columns 1–4 of Table 1. For $\alpha = 0.10$ and $K = 9$, the z value that bounds the inner $(1-\alpha)^{1/9} = 0.9884$ of the N(0,1) is $z_{\frac{\gamma}{2}} = 2.523$. Thus, for $k = VA$

$$|\Lambda_{L,VA}| = |\{CA\}| = 1 \quad \text{and} \quad |\Lambda_{O,VA}| = |\{IL, MA, GA\}| = 3.$$

Using (4), the marginal confidence set for r_{VA} is $\{1+1, 1+2, 1+3, 1+3+1\} = \{2, 3, 4, 5\}$. The other rows of Table 1 are obtained similarly. Table 1a gives the details for obtaining each confidence set in the last columns of Tables 1 and 1a.

Note 1: The values in Table 1 for $\bar{y}_k$ and SE_k are from the actual published official estimates (U.S. Census Bureau) and their computations are more complex than we assume throughout the remainder of this paper. This complexity is due to several reasons: complex sampling design, adjustments for nonresponse, and the methodology used for computing $\bar{y}_k$ and SE_k. Hereafter in this paper, we have defined $\bar{y}_k$ and SE_k as would be used under classical simple random sampling, independently sampling among the K populations, for simplicity in presenting basic concepts.

RANK										*One Possible True Ranking*	*Another Possible True Ranking*
9									**MD**	**MD**	**MD**
8								**NY**		**NY**	**NY**
7						*DC*	**NJ**			**NJ**	*DC*
6						**DC**	*NJ*			**DC**	*NJ*
5			*VA*	*MA*	**IL**					**IL**	*MA*
4			*VA*	**MA**	*IL*					**MA**	*IL*
3		*GA*	**VA**	*MA*	*IL*					**VA**	**VA**
2	**CA**	**GA**	*VA*							**CA**	**GA**
1	*CA*	*GA*								*GA*	*CA*
	CA	*GA*	*VA*	*MA*	*IL*	*DC*	*NJ*	*NY*	*MD*		

STATES ORDERED BY ESTIMATES

Fig. 1. KWW 90% Joint Confidence Region for Tables 1 and 1a & Two Other Possible True Rankings.

Fig. 1 provides a visual of the 90% joint confidence region for the true ranking of the 9 states using the travel time to work data, and it follows from the last columns of Tables 1 and 1a. In the joint confidence region, highlighted states in bold show the estimated ranking, and the non-bold states show uncertainty in the estimated ranking. This joint confidence region contains other possible true rankings, and two examples are shown in the last two columns of Fig. 1.

Fig. 1 reveals two other aspects of a joint confidence region. First, each specific row of the joint confidence region shows (with at least 90% confidence) which states could occupy the associated rank. For example, rank $r = 5$ could be occupied with 90% confidence by states VA, MA, or IL. Second, each specific column of the joint confidence region shows (with at least 90% confidence) which ranks the associated state could occupy, i.e., a marginal confidence set for r_k. For example, GA could occupy ranks 1, 2, or 3; while NY can, with at least 90% confidence, only occupy rank 8.

2.5. KWW summary

The uncertainty shown in a KWW joint confidence region reveals:
(*i*) uncertainty in the estimated ranking $(\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_K)$;
(*ii*) other possible true rankings $(r_1, r_2, \ldots, r_K)$ beyond the estimated ranking;
(*iii*) for population (column) k, a marginal confidence set (set of ranks r) for r_k; and
(*iv*) for rank (row) r, a marginal confidence set (set of populations k) for r.

Note 2: Fig. 1 is a 9 × 9 grid with 81 positions: 9 positions with bold letters for the estimated ranking, 12 positions with non-bold letters for the uncertainty of the estimated ranking, and 60 (= 81-9-12) positions that are blank. The joint confidence region consists of the 21 (= 9+12) positions for the estimated ranking and its uncertainty. More (Less) non-bold letters indicate more (less) uncertainty in the estimated ranking.

3. Optimizing (tightening) a KWW joint confidence region

Definition 1. The **tightness** of a specific joint confidence region for an estimated ranking of K populations is defined as $T = 1 - \frac{OP}{K^2}$, where OP is the total number of occupied positions (bold and non-bold letters) in the specified joint confidence region out of the total number of positions K^2. Note that $OP = K + \sum_k |\Lambda_{Ok}|$.

Example. From Fig. 1, note that $OP = 2 + 3 + 4 + 3 + 3 + 2 + 2 + 1 + 1 = 21$. Hence the tightness of the joint confidence region in Fig. 1 is $T = 1 - \frac{21}{9^2} = \frac{60}{81}$. Note that $0 \le T \le \frac{K-1}{K}$. If $T = 0$, there is no tightness. If $T = \frac{K-1}{K}$, there is complete tightness; joint confidence region only contains the estimated ranking; and we have strong evidence it is the true ranking.

It is clear from (5) or Definition 1 that we can tighten a joint confidence region if $|\Lambda_{Ok}|$ can be made small for as many values of k as possible; that is, if we can minimize the overlap of the joint confidence intervals in the family of joint confidence intervals. Given the KWW approach and the desire that the confidence level be the same for each confidence interval in the family, one straightforward way lies at the sample size value of n_k for population k. Wright (2024) demonstrates empirically how changing values of α (and γ) can affect the tightness of a joint confidence region, and he gives another example illustrating how tightness increases by increasing n_k to $2n_k$ and $2n_k$ to $3n_k$, as shown in Figs. 2, 3, and 4. For Fig. 2, $T = 60/81$; for Fig. 3, $T = 62/81$; for Fig. 4, $T = 68/81$.

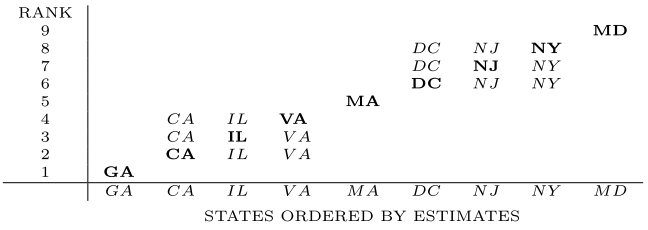

Fig. 2. KWW 90% Joint Confidence Region for the True Ranking of 9 States for n_k.

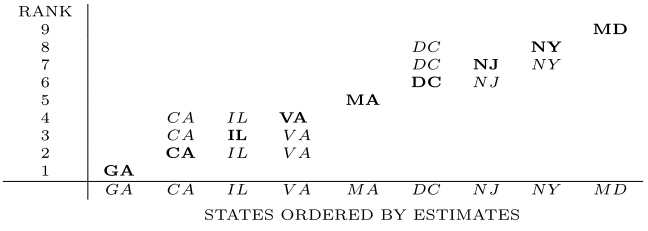

Fig. 3. KWW 90% Joint Confidence Region for the True Ranking of 9 States for $2n_k$.

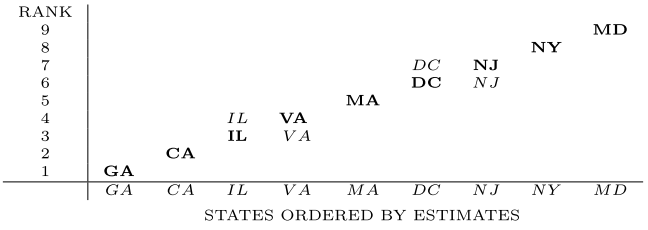

Fig. 4. KWW 90% Joint Confidence Region for the True Ranking of 9 States for $3n_k$.

4. A plan for optimal tightening of KWW joint confidence region

4.1. Framework

Sampling independently among the K populations, assume a simple random sample of n_k units selected from population k with sample mean $\bar{y}_k$ which is an estimator of population k mean $\bar{Y}_k$. The overall sample size for the K populations is $n = \sum_{k=1}^{K} n_k$. The sampling variance of $\bar{y}_k$ is Cochran (1977), Lohr (2022) $Var(\bar{y}_k) = \left(\frac{N_k - n_k}{N_k}\right)\frac{S_k^2}{n_k}$.

Consider objective function (7) given N_k and S_k^2

$$f(n_1, n_2, \ldots, n_K) = \sum_{k=1}^{K} N_k^2 Var(\bar{y}_k) = \sum_{k=1}^{K} N_k^2 \left(\frac{N_k - n_k}{N_k}\right)\frac{S_k^2}{n_k} \tag{7}$$

and seek to find optimal choice of n_k (for all k) that minimize $f(n_1, n_2, \ldots, n_K)$ subject to the overall sample size n being fixed and the following additional constraints. Note that the estimator of $\sqrt{Var(\bar{y}_k)}$, which is SE_k, is how we aim to control the length of each $100(1-\gamma)\%$ joint confidence interval given in (6). Clearly for given N_k and S_k^2 (or an estimate), we can carefully choose n_k with the aim of making SE_k small and hence *decrease* the number of overlaps, i.e. $|\Lambda_{Ok}|$, which is key to "tightening" the KWW 90% joint confidence region, as shown in Wright (2024).

Additional Constraints: Wright (2017, 2020) assume the desire that sample size n_k be bounded below and above by specified positive integers a_k and b_k, respectively, that is,

$$1 \le a_k \le n_k \le b_k \le N_k, \quad \text{for each } k. \tag{8}$$

As a result, n, the overall sample size, will be at least $\sum_{k=1}^{K} a_k$ and no more than $\sum_{k=1}^{K} b_k$.

The objective function in (7) is appealing in this framework where we aim to tighten a joint confidence region because $SE_k^2 = \left(\frac{N_k - n_k}{N_k}\right)\frac{s_k^2}{n_k}$ is an unbiased estimator of $\left(\frac{N_k - n_k}{N_k}\right)\frac{S_k^2}{n_k}$; $2MOE_k = 2\big(2.523(SE_k)\big)$ is the width of the kth joint confidence interval for $\bar{Y}_k$ when $K = 9$; and it is clear that we can decrease $2MOE_k$ if we increase n_k, or, more precisely, allocate more of n to those populations k^* where the values of $S_{k^*}^2$ (almost equivalently $s_{k^*}^2$) are larger.

4.2. Optimization

Without loss of generality, assume $\frac{N_1^2 S_1^2}{a_1(a_1+1)} \geq \cdots \geq \frac{N_K^2 S_K^2}{a_K(a_K+1)}$.

Where $\sum_{k=1}^{K} N_k(N_k - a_k)S_k^2\frac{1}{a_k}$ is the value of $f(n_1, n_2, \ldots, n_K)$ when $n_k = a_k$ for population k, a decomposition of the objective function $f(n_1, n_2, \ldots, n_K)$ is Wright (2017, 2020)

$$
\begin{aligned}
f(n_1, n_2, \ldots, n_K) \quad = \quad & \sum_{k=1}^{K} N_k(N_k - a_k)\frac{S_k^2}{a_k} \\
& + \left(-\frac{N_1^2 S_1^2}{a_1(a_1+1)} - \frac{N_1^2 S_1^2}{(a_1+1)(a_1+2)} - \cdots - \frac{N_1^2 S_1^2}{(n_1-1)(n_1)}\right) \\
& \qquad \vdots \\
& + \left(-\frac{N_k^2 S_k^2}{a_k(a_k+1)} - \frac{N_k^2 S_k^2}{(a_k+1)(a_k+2)} - \cdots - \frac{N_k^2 S_k^2}{(n_k-1)(n_k)}\right) \\
& \qquad \vdots \\
& + \left(-\frac{N_K^2 S_K^2}{a_K(a_K+1)} - \frac{N_K^2 S_K^2}{(a_K+1)(a_K+2)} - \cdots - \frac{N_K^2 S_K^2}{(n_K-1)(n_K)}\right).
\end{aligned}
\tag{9}
$$

With the decomposition (9), the constraint (8) is satisfied, and the objective function (7) is minimized whenever we stop sequential sample size assignment among K populations if we use the following *Algorithm* (Wright, 2017, 2020), discussed in Remark 5.1 of Appendix: this Algorithm is more efficient than Neyman allocation (Neyman, 1934).

OPTIMAL SAMPLE SIZE ASSIGNMENT Algorithm

Step 1: First, note a_k units are to be selected for the sample from population k, for all k.

Step 2: For additional sample units, compute the array of *squared priority values*:

$$
\begin{array}{lcccc}
\text{Population } 1 & \frac{N_1^2 S_1^2}{a_1(a_1+1)} & \frac{N_1^2 S_1^2}{(a_1+1)(a_1+2)} & \cdots & \frac{N_1^2 S_1^2}{(b_1-1)(b_1)} \\
 & & \vdots & & \\
\text{Population } k & \frac{N_k^2 S_k^2}{a_k(a_k+1)} & \frac{N_h^2 S_h^2}{(a_k+1)(a_k+2)} & \cdots & \frac{N_k^2 S_k^2}{(b_k-1)(b_k)} \\
 & & \vdots & & \\
\text{Population } K & \frac{N_K^2 S_K^2}{a_K(a_K+1)} & \frac{N_K^2 S_K^2}{(a_K+1)(a_K+2)} & \cdots & \frac{N_K^2 S_K^2}{(b_K-1)(b_K)}
\end{array}
\tag{10}
$$

Step 3: From the kth row (population), at least a_k units and no more than b_k units are to be included in the sample. So from *Step 2*, select the largest values sequentially until sample size assignment is stopped. Each population is assigned an additional sample unit each time one of its squared priority values is among the largest assigned values. (Whenever sample size assignment stops, $f(n_1, n_2, \ldots, n_K)$ will be minimized for overall n at that point (Wright, 2020).)

Appendix A. Supplementary Materials

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2024.110288.

Data availability

No data was used for the research described in the article.

References

Cochran, W.G., 1977. Sampling Techniques, third ed. John Wiley & Sons, New York, NY.
Hall, P., Miller, H., 2009. Using the bootstrap to quantify the authority of an empirical ranking. Ann. Statist. 37 (6B), 3929–3959.
Klein, M., Wright, T., Wieczorek, J., 2020. A Joint Confidence Region for an overall ranking of populations. J. R. Stat. Soc. Ser. C. 69 (3), 589–606.
Lohr, S.L., 2022. Sampling: Design and Analysis, third ed. CRC Press, Boca Raton, FL.
Mogstad, M., Romano, J., Shaikh, A., Wilhelm, D., 2024. Inferences for ranks with applications to mobility across neighbourhoods and academic achievement across countries. Rev. Econ. Stud. 91, 476–518.
Neyman, J., 1934. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. J. R. Stat. Soc. 97, 558–606.
Šidák, Z.K., 1967. Rectangular Confidence Regions for the means of multivariate normal distributions. J. Amer. Statist. Assoc. 62 (318), 626–633.
Wright, T., 2017. Exact optimal sample allocation: More efficient than Neyman. Statist. Probab. Lett. 129, 50–57.
Wright, T., 2020. A general exact optimal sample allocation algorithm: With bounded cost and bounded sample sizes. Statist. Probab. Lett. 165, 108829.
Wright, T., 2024. Understanding and optimal tightening of the KWW joint confidence region for a ranking, Research Report Series (Statistics #2024-01). In: Center for Statistical Research & Methodology. U.S. Census Bureau, Washington, D.C.