Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Mann-Whitney Test for Two-phase Stratified Sampling

Takumi Saegusa*

[1]Department of Mathematics, University of Maryland, Maryland, USA

Correspondence
*Takumi Saegusa, Department of Mathematics, University of Maryland, College Park, MD USA. Email: tsaegusa@umd.edu

Summary

We consider the Mann-Whitney test for two-phase stratified sampling. In this design, the i.i.d. sample is obtained at the first phase and then stratified based on auxiliary variables. At the second phase, stratified subsamples are obtained without replacement to collect variables of interest. The resultant data are biased and dependent sample due to stratification and sampling without replacement. This setting is different from the one considered by the existing method called the van Elteren test that considers multiple i.i.d. samples from non-overlapping subsets of the infinite population. We propose the inverse probability weighted Mann-Whitney statistic and study its asymptotic properties. The proposed test is shown to have a correct size and power 1 in the limit. Simulation study is presented to illustrate the finite sample performance of our test.

KEYWORDS:
hypothesis testing, rank test, sampling without replacement, stratified sampling

1 | INTRODUCTION

We consider nonparametric testing of the hypothesis

$$H_0 : F(x) = G(x), \quad \text{for all } x \in \mathbb{R},$$

for two-phase stratified sampling (Breslow & Chatterjee 1999; Breslow, Lumley, Ballantyne, Chambless, & Kulich 2009b; Breslow, McNeney, & Wellner 2003) where F and G are cumulative distribution functions of the continuous random variable X in the first and second groups respectively. In this sampling design, an independent and identically distributed (i.i.d.) sample is obtained from the infinite population at the first phase. At the second phase, this sample is stratified and then subsamples are collected without replacement. Stratification is expected to correlate with the target variable X for effective sampling. The final sample is a biased and dependent sample due to stratification and sampling without replacement. This property of the sample poses a serious statistical challenge for our two-sample hypothesis testing. Although all X_i are independent and generated from either F or G, two groups in our setting are dependent and do not properly represent the populations from F and G. Hence we need a novel approach to the choice of a test statistic, the derivation of its limiting distribution, and the construction of the rejection region of a test.

In this paper, we extend the Mann-Whitney two-sample test to two-phase stratified sampling. The Mann-Whitney test, equivalent to the Wilcoxon rank sum test, is one of the most popular nonparametric testing procedures for comparing two independent samples. This test was extended by van Elteren (1960) to stratified samples by computing the sum of the Mann-Whitney statistics within strata. The primary setting of this test known as the van Elteren test is an experimental setting such as clinical trials where treatment assignments and stratification defined by experimental conditions are under the control of a study investigator. This is formalized as the stratification of the infinite population and the i.i.d. sampling from each stratum. Its null hypothesis is that two groups are distributionally equivalent in all of strata. This null hypothesis is not suitable in our setting because our stratification is allowed to depend on the partial information of the target variable X. In this case, the null hypothesis of the van Elteren test may not hold even if our null hypothesis holds. Moreover, theoretical derivations of the van Elteren test is no longer valid in our setting because of the lack of the i.i.d. structure within stratum in the two-phase stratified sample.

The main contribution of this paper is two-fold. First, we propose a new inverse probability weighted statistic and derive its asymptotic properties in the null and alternative hypotheses. The proposed modified Mann-Whitney test statistic is the inverse probability weighted U-statistic whose inverse probability weights must be based on the simultaneous selection of two observations. Instead, we write the proposed statistic as the function of two empirical distributions which can be inverse probability weighted based on each observation. Because the limiting distribution depends on random functions, we apply the special empirical process theory for a biased and dependent sample developed by Breslow and Wellner (2007) and Saegusa and Wellner (2013). The second contribution is the novel method to construct the rejection region of the proposed test. Unlike the i.i.d. setting, the limiting distribution of the proposed test statistic depends on unknown parameters even under the null hypothesis, which prevents analytical computation of the rejection region. Our approach is to separate sources of randomness into sampling from the infinite population and subsequent stratified sampling. The former randomness is parameter-free. The latter is estimated by the bootstrap method specialized in the finite population sampling (Bickel & Freedman 1984; Gross 1980). We show that the proposed test achieves the correct size and its power converges to 1 in the limit. The finite sample performance is shown in the simulation study and data example.

2 | SETTING AND TEST STATISTIC

We first give the formal description of two-phase stratified sampling. Let X be a target continuous random variable and $S \in \{1,2\}$ be an indicator of group assignment with $P(S=1) = q \in (0,1)$. Let $F(t) = P(X \le t | S=1)$ and $G(t) = P(X \le t | S=2)$ be the cumulative distribution functions for the group 1 and group 2, respectively. Let $(X_i, S_i), i = 1, \dots, N$, be the i.i.d. copy of (X,S). Sample sizes for each group are denoted as $N^{(1)}$ and $N^{(2)}$. Unlike the i.i.d. setting, we do not observe all of X_i. Instead, we observe a random vector V for all subjects. The component of $V \in \mathcal{V}$ includes random variables correlated with X such as S as well as auxiliary variables. The variable V helps stratification of the sample. We divide the sample space $\mathcal{V}$ into a partition $\{\mathcal{V}_j\}_{j=1}^J$. We say the subject belongs to stratum j if $V \in \mathcal{V}_j$. The stratum membership probability is denoted as $\nu_j = P(V \in \mathcal{V}_j)$.

Two-phase stratified sampling is carried out as follows. At the first phase, we obtain the i.i.d. sample of $V_1, \dots, V_N$. The values of V_i determines stratum membership. Denote $N_j = \#\{i : V_i \in \mathcal{V}_j\}$ as the sample size of stratum j. At the second phase, a subsample of size n_j is collected without replacement from stratum j. The sampling probability is $\pi_i = \sum_{j=1}^J (n_j/N_j) I\{V_i \in \mathcal{V}_j\}$. The sampling indicator $R_i \in \{0,1\}$ is 1 if the ith subject is selected at the second phase and 0 otherwise. For the selected subjects, information on X_i is collected. Because n_j is under the control of the study investigator, we assume that the sampling fraction n_j/N_j converges almost surely to $p_j > 0$. The above mathematical formulation of two-phase stratified sampling is standard in the literature (Breslow, Lumley, Ballantyne, Chambless, & Kulich 2009a; Breslow & Wellner 2007; Saegusa & Wellner 2013).

Next, we propose the weighted Mann-Whitney test statistic. To this end, we first describe the Mann-Whitney test statistic in the i.i.d. setting. For simplicity, we assume observations $i = 1, \dots, N^{(1)}$ are in group 1 and observations $i = N^{(1)}+1, \dots, N$ are in group 2. This statistic counts how many observations in the second group exceed observations in the first group. With our notation, the statistic is written as the U-statistic given by

$$U_N = \sum_{i=1}^{N^{(1)}} \sum_{j=N^{(1)}+1}^{N} I\{X_i \le X_j\}.$$

This representation is not suitable for the extension to two-phase stratified sampling as well as other missing data problems. The standard technique to address missing data is inverse probability weighting of individual observations. Asymptotic theory for inverse probability weighted averages is available for our setting. For this statistic, however, inverse probability weights must be computed for the pair of observations. Theory to address this type of inverse probability weights does not exist in our setting.

Another representation of the Mann-Whitney statistic is based on the empirical distributions for two groups given by

$$\mathbb{F}_{N^{(1)}}(x) = \frac{1}{N^{(1)}} \sum_{i=1}^{N^{(1)}} I\{X_i \le t\}, \quad \mathbb{G}_{N^{(2)}}(x) = \frac{1}{N^{(2)}} \sum_{j=N^{(1)}+1}^{N} I\{X_j \le t\}.$$

We write U_N as the function of two empirical distributions given by

$$\frac{U_N}{N^{(1)}N^{(2)}} = \int_{-\infty}^{\infty} \mathbb{F}_{N^{(1)}}(x) d\mathbb{G}_{N^{(2)}}(x).$$

An advantage of this presentation is that U_N is now computed by sums of individual observations, which admits inverse probability weighting. Let

$$\mathbb{F}^{\pi}_{N^{(1)}}(x) = \frac{1}{N^{(1)}} \sum_{i=1}^{N^{(1)}} \frac{R_i}{\pi_i} I\{X_i \le t\}, \quad \mathbb{G}^{\pi}_{N^{(2)}}(x) = \frac{1}{N^{(2)}} \sum_{j=N^{(1)}+1}^{N} \frac{R_i}{\pi_i} I\{X_j \le t\}.$$

be inverse probability weighted empirical distributions. The proposed test statistic U_N^π is

$$U_N^\pi \equiv N^{(1)}N^{(2)} \int_{-\infty}^{\infty} \mathbb{F}^\pi_{N^{(1)}}(x) d\mathbb{G}^\pi_{N^{(2)}}(x).$$

Note that the computation of U_N^π does require the knowledge on $N^{(1)}$ and $N^{(2)}$ due to cancelation with $\mathbb{F}^\pi_{N^{(1)}}(x)$ and $\mathbb{G}^\pi_{N^{(1)}}(x)$. This property is useful because $N^{(1)}$ and $N^{(2)}$ are unknown when S_i is not included in V_i collected at the first phase.

The derivation of the asymptotic distribution of the proposed statistic is challenging because it consists of the random functions and because our dependent biased sample does not admit the standard central limit theorem. Our approach is to use the functional delta method and empirical process theory for two-phase stratified sampling. In the i.i.d. setting, the map of the two empirical distributions to the Mann-Whitney statistic is shown to be Hadamard differentiable so that its asymptotic distribution reduces to the weak convergence of two empirical distribution functions to Brownian bridge processes through the functional delta method. Because the map of $\mathbb{F}^\pi_{N^{(1)}}$ and $\mathbb{G}^\pi_{N^{(2)}}$ to U_N^π is also Hadamard differentiable with an appropriate domain, we can apply the functional delta method to our setting.

To complete the application of the functional delta method, we need to develop the weak convergence of $\mathbb{F}^\pi_{N^{(1)}}$ and $\mathbb{G}^\pi_{N^{(2)}}$. We carry out this task with the help of empirical process theory for two-phase stratified sampling. Note that one cannot simply apply the empirical process result because we divide sums by random sample sizes $N^{(1)}$ and $N^{(2)}$, which require the conditional argument with additional care. Proofs for the following lemma as well as other results are deferred to Appendix.

Lemma 1. Let P_j be the conditional probability measure given the membership in stratum j. As $N \to \infty$,

$$\begin{aligned}
&\sqrt{N^{(1)}}(\mathbb{F}^\pi_{N^{(1)}} - F) \rightsquigarrow \mathbb{G}_{F,0} + \sum_{j=1}^J \sqrt{\frac{\nu_j}{q}}\sqrt{\frac{1-p_j}{p_j}}\mathbb{G}_{F,j},\\
&\sqrt{N^{(1)}}(\mathbb{F}^\pi_{N^{(1)}} - \mathbb{F}_{N^{(1)}}) \rightsquigarrow \sum_{j=1}^J \sqrt{\frac{\nu_j}{q}}\sqrt{\frac{1-p_j}{p_j}}\mathbb{G}_{F,j},\\
&\sqrt{N^{(2)}}(\mathbb{G}^\pi_{N^{(2)}} - G) \rightsquigarrow \mathbb{G}_{F,0} + \sum_{j=1}^J \sqrt{\frac{\nu_j}{1-q}}\sqrt{\frac{1-p_j}{p_j}}\mathbb{G}_{G,j},\\
&\sqrt{N^{(2)}}(\mathbb{G}^\pi_{N^{(2)}} - \mathbb{G}_{N^{(1)}}) \rightsquigarrow \sum_{j=1}^J \sqrt{\frac{\nu_j}{1-q}}\sqrt{\frac{1-p_j}{p_j}}\mathbb{G}_{G,j},
\end{aligned}$$

where $\mathbb{G}_{F,j}, \mathbb{G}_{G,j}, j = 0, 1, \ldots, J$ are independent zero-mean Gaussian processes with covariance functions given by

$$\begin{aligned}
\rho_{F,0}(x,y) &= P(X \le x \wedge y | S = 1) - P(X \le x | S = 1)P(X \le y | S = 1),\\
\rho_{F,j}(x,y) &= P_j(X \le x \wedge y, S = 1) - P_j(X \le x, S = 1)P_j(X \le y, S = 1),\\
\rho_{G,0}(x,y) &= P(X \le x \wedge y | S = 2) - P(X \le x | S = 2)P(X \le y | S = 2),\\
\rho_{G,j}(x,y) &= P_j(X \le x \wedge y, S = 2) - P_j(X \le x, S = 2)P_j(X \le y, S = 2).
\end{aligned}$$

The functional delta method now yields the following theorem.

Theorem 1. Let Var_j be the conditional variance given membership in stratum j. Under H_0,

$$\sqrt{\frac{N^{(1)}N^{(2)}}{N}}\left(\frac{U_N^\pi}{N^{(1)}N^{(2)}} - \frac{1}{2}\right) \to_d Z_0 \sim N\left(0, \sigma_0^2\right)$$

where

$$\sigma_0^2 = \frac{1}{12} + \sum_{j=1}^J \nu_j \frac{1-p_j}{p_j}\left\{\frac{q}{1-q}\mathrm{Var}_j\left(\int F d\mathbb{G}_{F,j}\right) + \frac{1-q}{q}\mathrm{Var}_j\left(\int \mathbb{G}_{F,j} dF\right)\right\}.$$

In the i.i.d. setting, the Mann-Whitney statistic U_N has asymptotic variance $\sigma_{0,1}^2 = 1/12$ under H_0. Additional stratified sampling increases the randomness to the weighted Mann-Whitney statistic in our setting. We call this additional variance the phase II variance $\sigma_{0,2}^2 = \sigma_0^2 - \sigma_{0,1}^2$. This phase II variance contains unknown parameters such as F, ν_j, p_j, q, and P_j. Among them, the conditional probability measure P_j given stratum membership is the greatest obstacle to constructing the rejection region based on U_N^π. This conditional distribution involves conditional variance Var_j for the stochastic integrals and the covariance function of the Gaussian process $\mathbb{G}_{F,j}$ in the integral. To propose a valid hypothesis testing procedure, we need to address the unknown phase II variance under the null hypothesis.

3 | METHODOLOGY

We propose the test with the rejection region of the form

$$\sqrt{\frac{N^{(1)}N^{(2)}}{N}}\left|\frac{U_N^\pi}{N^{(1)}N^{(2)}}-\frac{1}{2}\right|\geq q_N.$$

Because the statistic U_N^π is a consistent estimator of the probability that the random variable from the group 1 is larger than the random variable from the second group, the deviation from 1/2 indicates $F\neq G$. The valid hypothesis testing should have the probability of the rejection region is the prescribed significance level $\alpha\in(0,1)$ at least asymptotically. As seen in Theorem 1, the limiting null distribution of U_N^π has asymptotic variance which contains unknown parameters. In general, the presence of the unknown asymptotic variance does not cause a serious obstacle to hypothesis testing, because a consistent estimator of the asymptotic variance (e.g. the plug-in estimator or bootstrap estimator) is easily accommodated into q_N. In our setting, however, the asymptotic variance involves the variance of stochastic integrals with the unknown Gaussian processes with respect to the unknown distribution. It may not be tractable to simplify this variance for obtaining a plug-in estimator. Bootstrap is not available either for two-phase stratified sample that correctly quantify randomness due to sampling from the infinite population and subsequent sampling from strata without replacement at the same time.

The proposed approach is simple and easy to implement. First, we observe that the asymptotic variance can be decomposed into the phase I variance and phase II variance. The phase I variance is 1/12 which coincides with the asymptotic variance of the Mann-Whitney statistic in the i.i.d. setting. This indicates that 1/12 corresponds to sampling from the infinite population at the first phase. Thus, the phase II variance which contains unknown parameters corresponds to randomness from the subsequent stratified sampling without replacement. This suggests that estimation of the asymptotic variance reduces to estimating a data generating mechanism of stratified sampling conditionally on sampling from the infinite population. The main idea of our approach is to carry out this task by adopting the bootstrap method specialized in the finite population proposed by Gross (1980) and Bickel and Freedman (1984). Originally aimed for survey sampling, this bootstrap method quantifies randomness in R only from stratified sampling, not randomness in X.

This bootstrap method mimics sampling without replacement based on selected observations. For an illustrative example, consider selecting 5 observations from 10 people. In this case, each sampled observation corresponds to 2 (=10/5) observations in the original sample of size 10. Thus, the bootstrap population of size 10 is constructed by creating two copies of sampled observations. Sampling 5 observations from this bootstrap population now yields a bootstrap sample of size 5. A mathematically formal definition of this procedure is as follows. There are two cases depending on whether or not N_j/n_j is an integer. For stratum j, suppose $k_j=N_j/n_j$ is an integer. The bootstrap population is then created as the k_j copies of selected observations at the second phase. That is, an observation i with $V_i\in\mathcal{V}_j$ and $R_i=1$ is duplicated k_j times. This bootstrap population of size N_j consists only of selected observations at the second phase. From this bootstrap population, a bootstrap sample of size n_j is collected without replacement. If $N_j=n_jk_j+r_j$ and r_j is a remainder for division of N_j by n_j, we construct two bootstrap populations, one containing k_j copies of selected observations and the other containing k_j+1 copies. We choose the first population with probability $s_j=(1-r_j/N_j)\{1-r_j/(n_j-1)\}$ or the second with probability $1-s_j$. Then sample of size n_j is collected without replacement as above. We perform this procedure for all strata and then compute the corresponding weighted Mann-Whitney statistic $\hat{U}_N^\pi$ based on the bootstrap stratified samples.

For the estimation of the phase II variance, we compute the bootstrap weighted Mann-Whitney statistics $\hat{U}_{N,1}^\pi,\dots,\hat{U}_{N,B}^\pi$ for each bootstrap sample. We use the sample variance of bootstrap weighted Mann-Whitney statistics as the estimator of phase II variance denoted by $\hat{\sigma}_{N,2,B}^2$. The variance estimator of the entire asymptotic variance is then $\hat{\sigma}_{N,B}^2=1/12+\hat{\sigma}_{N,2,B}^2$. Now we determine the rejection by setting

$$q_{N,B}=q_{1-\alpha/2}\hat{\sigma}_{N,B}$$

where $q_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal random variable.

In our setting described in Section 2, information on group assignment S is collected at the first phase so that stratification may depend on S. In practice, information on group assignment may not be available so that group sizes $N^{(1)}$ and $N^{(2)}$ are unknown quantities. Our proposed test is also valid when S is only available at the second phase. In this case, one can estimate $N^{(1)}$ and $N^{(2)}$ by

$$\hat{N}^{(1)}=\sum_{i=1}^N\frac{R_i}{\pi_i}I\{S_i=1\},\quad \hat{N}^{(2)}=\sum_{i=1}^N\frac{R_i}{\pi_i}I\{S_i=2\}.$$

These are consistent estimators in the sense that $\hat{N}^{(1)}/N^{(1)}\to_p 1$ and $\hat{N}^{(2)}/N^{(2)}\to_p 1$. Then the rejection region becomes

$$\sqrt{\frac{\hat{N}^{(1)}\hat{N}^{(2)}}{N}}\left|\frac{U_N^\pi}{\hat{N}^{(1)}\hat{N}^{(2)}}-\frac{1}{2}\right|\geq q_{1-\alpha/2}\hat{\sigma}_{N,B}.$$

We omit proofs for this case due to their similarity to the case where $N^{(1)}$ and $N^{(2)}$ are known.

The following theorem shows that the proposed test has correct size and asymptotic power of 1 in the limit.

Theorem 2. Under the null hypothesis H_0, as $N, B \to \infty$,

$$P\left(\sqrt{\frac{N^{(1)}N^{(2)}}{N}}\left|\frac{U_N^\pi}{N^{(1)}N^{(2)}}-\frac{1}{2}\right| \geq q_{N,B\cdot}\right) \to \alpha.$$

Under the alternative hypothesis that $F \neq G$ with $\int_{-\infty}^{\infty} F(x)dG(x) \neq 1/2$, as $N, B \to \infty$,

$$P\left(\sqrt{\frac{N^{(1)}N^{(2)}}{N}}\left|\frac{U_N^\pi}{N^{(1)}N^{(2)}}-\frac{1}{2}\right| \geq q_{N,B\cdot}\right) \to 1.$$

Note that for independent random variables Y and $\tilde{Y}$ with cumulative distribution functions F and G respectively, we have $P(Y \leq \tilde{Y}) = \int_{-\infty}^{\infty} F(x)dG(x)$.

4 | SIMULATION

To evaluate the finite sample performance of the proposed Mann-Whitney test for two-phase stratified sampling, we conducted a simulation study for exponential and normal random variables respectively. For both cases, we consider three strata $\mathcal{V}_j, j = 1, 2, 3$, with stratum membership probability $\nu_1 = \nu_2 = \nu_3 = 1/3$ characterized by conditional distributions $F_j(x) = P(X \leq x|V \in \mathcal{V}^{(j)}, S = 1)$ and $G_j(x) = P(X \leq x|V \in \mathcal{V}^{(j)}, S = 2)$. Sampling probabilities are .5, .3, .2 in the first setting and .1, .1, .1 in the second setting for each stratum. The final sample size is $n = n_1 + n_2 + n_3 = 333$ for $N = 1000$ and $n = 666$ for $N = 2000$ in the first setting and $n = 100$ for $N = 1000$ and $n = 200$ for $N = 2000$ in the second setting on average. The treatment assignment is a Bernoulli distribution with parameter 1/2. For the exponential case, we consider the mixture of exponential distributions where $F = (F_1 + F_2 + F_3)/3$ and $G = (G_1 + G_2 + G_3)/3$. Under the null hypothesis $H_0 : F = G$, the parameters $\lambda_{j,1}$ and $\lambda_{j,2}$ for F_j and G_j are

$$(\lambda_{1,1}, \lambda_{2,1}, \lambda_{3,1}) = (e^{1/2}, e, e^2), \quad (\lambda_{1,2}, \lambda_{2,2}, \lambda_{3,2}) = (e^2, e, e^{1/2}).$$

Under the alternative hypothesis $H_a : F \neq G$,

$$(\lambda_{1,1}, \lambda_{2,1}, \lambda_{3,1}) = (e^{1/2}, e, e^2), \quad (\lambda_{1,2}, \lambda_{2,2}, \lambda_{3,2}) = (e^{1/2}, e^{3/2}, e^{3/2}).$$

For the normal case, we consider the mixture of normal distributions with the same variance 1 and stratum mean

$$(\mu_{1,1}, \mu_{2,1}, \mu_{3,1}) = (1, -1, 0), \quad (\mu_{1,2}, \mu_{2,2}, \mu_{3,2}) = (0, -1, 1),$$

under H_0 and

$$(\mu_{1,1}, \mu_{2,1}, \mu_{3,1}) = (1, -1, 0), \quad (\mu_{1,2}, \mu_{2,2}, \mu_{3,2}) = (1, 0, 0),$$

under H_a.

Table 1 summarizes the simulation results. We conducted the computation of the proposed rejection region based on 1000 bootstrap samples in each case. The proposed test shows superior performance over the van Elteren test in all settings. Our test achieved the empirical type I error close to the nominal level of 5 percent especially for large sample size. Because the difference between the settings 1 and 2 lies only in the sampling probability, we see the better empirical type I error in the first setting where sampling probabilities are large. The van Elteren test, on the other hand, is sensitive to sampling probabilities. This test rejected the null hypothesis very frequently in the first setting, and almost never rejected the null hypothesis in the second setting. This finite sample performance may be explained by the difference in the null hypotheses between two tests. The null hypothesis for the van Elteren test is the equivalence of two samples in all of strata but the null hypothesis in this simulation allows difference in each stratum while $F = G$. The results on the empirical power show the same tendency in empirical type I error. Our test increases the empirical power when the final sample sizes increase by changing $N = 1000$ to $N = 2000$ and/or by changing the setting from the second to the first. Moreover, the proposed test showed higher empirical power than the van Elteren test. Unlike the null hypothesis, the alternative hypothesis in this simulation is also the alternative hypothesis for the van Elteren test. Thus, this result indicates the better performance of the proposed test in the setting of consideration. One possible reason is that our proposed test statistic accounts for sampling design well by inverse probability weighting while the van Elteren test simply combines the stratum-wise Mann-Whitney statistics.

5 | DATA ANALYSIS

We applied the proposed testing procedure to the national Wilms tumor study (D'Angio et al. 1989). Wilms tumor is a rare kidney cancer for children. Our interest lies in the prediction of relapse. We conducted two-phase stratified sampling based on strata with deceased patients ($N_1 = 444$),

			H_0		H_a	
	Setting	Method	$N = 1000$	$N = 2000$	$N = 1000$	$N = 2000$
Exponential	1	Ours	0.044	0.050	0.611	0.892
		van Elteren	0.920	0.999	0.544	0.868
	2	Ours	0.044	0.044	0.298	0.501
		van Elteren	0.003	0.001	0.248	0.428
Normal	1	Ours	0.044	0.050	0.756	0.946
		van Elteren	0.628	0.926	0.688	0.932
	2	Ours	0.037	0.060	0.348	0.597
		van Elteren	0.012	0.012	0.260	0.536

TABLE 1 Empirical type I error and power of the proposed test.

living patients with unfavorable histology measured at the hospital ($N_2 = 235$), and the all the rest ($N_3 = 3236$) with sampling probabilities 100%, 50%, and 10%, respectively. We collected 886 subjects without replacement from the entire cohort of size $N = 3915$. For selected patients, tumor diameter was measured. We applied our method to test the null hypothesis that tumor diameters are equally distributed between patients with and without relapse We generated 1000 bootstrap in our procedure to obtain $\hat{q}_N = 1.21$. Because the proposed test statistic is 1.71, we reject the null hypothesis. The van Elteren test, on the other hand, failed to reject the null hypothesis for this data set.

ACKNOWLEDGMENTS

The author is supported by NSF grant DMS 2014971.

Author contributions

The author is responsible for the preparation of the manuscript, theoretical derivations, the implementation of the simulation studies and data analysis.

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

SUPPORTING INFORMATION

Not applicable.

How to cite this article: T. Saegusa (2016), Mann-Whitney Test for Two-phase Stratified Sampling, , *2020;00:1–6.*

APPENDIX

A PROOFS

Proof of Lemma 1. Rewrite the inverse probability weighted empirical distribution functions by

$$\mathbb{F}^{\pi}_{N^{(1)}}(x) = \frac{1}{N^{(1)}}\sum_{i=1}^{N}\frac{R_i}{\pi_i}I\{X_i \le x, S_i = 0\},$$
$$\mathbb{G}^{\pi}_{N^{(2)}}(x) = \frac{1}{N^{(2)}}\sum_{i=1}^{N}\frac{R_i}{\pi_i}I\{X_i \le x, S_i = 1\}.$$

We use the double subscript to denote e.g. a random variable X for the ith observation in stratum j by X_{ji}. Let $Y_i(x) \equiv I\{X_i \le x, S_i = 0\}$. Then for a fixed $x \in \mathbb{R}$, the decomposition of $\mathbb{F}^{\pi}_{N^{(1)}}(x) - F(x)$ yields

$$\begin{aligned}
&\sqrt{N^{(1)}}(\mathbb{F}^{\pi}_{N^{(1)}}(x) - F(x)) \\
&= \sqrt{N^{(1)}}(\mathbb{F}_{N^{(1)}}(x) - F(x)) + \sqrt{N^{(1)}}(\mathbb{F}^{\pi}_{N^{(1)}}(x) - \mathbb{F}_{N^{(1)}}(x)) \\
&= \sqrt{N^{(1)}}\left(\frac{1}{N^{(1)}}\sum_{i=1}^{N} I\{X_i \le x, S_i = 1\} - F(x)\right) \\
&\quad + \sum_{j=1}^{J}\sqrt{\frac{N}{N^{(1)}}}\sqrt{\frac{N_j}{N}}\frac{N_j}{n_j}\sqrt{N_j}\left(\frac{1}{N_j}\sum_{i=1}^{N_j} R_{ji}Y_{ji}(x) - \frac{n_j}{N_j}\frac{1}{N_j}\sum_{i=1}^{N_j} Y_{ji}(x)\right).
\end{aligned}$$

Note that the first term concerns the group specific average of Y_i. Since $N^{(1)} \to \infty$ almost surely as $N \to \infty$, it follows that conditionally on S_i the standard central limit theorem yields

$$\sqrt{N_1}\left(\frac{1}{N^{(1)}}\sum_{i=1}^{N} I\{X_i \le x, S_i = 1\} - F(x)\right) \to_d Z_{F,0} \sim N\left(0, F(x)\{1 - F(x)\}\right)$$

from which we obtain the unconditional convergence to the same limiting variable. Viewing this term as the stochastic process indexed by $\mathbb{R}$, the Donsker theorem yields

$$\sqrt{N_1}\left(\frac{1}{N^{(1)}}\sum_{i=1}^{N} I\{X_i \le \cdot, S_i = 1\} - F(\cdot)\right) \rightsquigarrow \mathbb{G}_{F,0}(\cdot)$$

where $\rightsquigarrow$ means the weak convergence in the class of cadlag functions on $\mathbb{R}$. Because the first term and the second term are uncorrelated and Gaussian, the limiting process $\mathbb{G}_{F,0}$ is independent of the limiting process for the second term. The second term consists of stratum-wise sums of Y_i which are uncorrelated. As in Breslow and Wellner (2007), we apply the bootstrap Donsker theorem conditionally to obtain the unconditional weak convergence

$$\sqrt{\frac{N_j}{N}}\frac{N_j}{n_j}\sqrt{N_j}\left(\frac{1}{N_j}\sum_{i=1}^{N_j} R_{ji}Y_{ji}(\cdot) - \frac{n_j}{N_j}\frac{1}{N_j}\sum_{i=1}^{N_j} Y_{ji}(\cdot)\right) \rightsquigarrow \sqrt{\nu_j}\sqrt{\frac{1-p_j}{p_j}}\mathbb{G}_{F,j}(\cdot).$$

Because $N^{(1)}/N \to q$ by the law of large numbers, combining above results yields the first result. The second result also follows from the second term in the decomposition. □

Theorem 3. As $N \to \infty$,

$$\sqrt{\frac{N^{(1)}N^{(2)}}{N}}\left(\frac{U^{\pi}_N}{N^{(1)}N^{(2)}} - \int_{-\infty}^{\infty} F(x)dG(x)\right) \to_d Z \sim N\left(0, \sigma^2\right)$$

where $\sigma^2 = \sigma_1^2 + \sigma^2$ with

$$\begin{aligned}
\sigma_1^2 &\equiv q\mathrm{Var}(F(X)|S = 1) + (1-q)\mathrm{Var}(G(X)|S = 2), \\
\sigma_2^2 &\equiv \sum_{j=1}^{J}\frac{1-p_j}{p_j}\nu_j\left\{\frac{q}{1-q}\mathrm{Var}_j\left(\int F d\mathbb{G}_{G,j}\right) + \frac{1-q}{q}\mathrm{Var}_j\left(\int \mathbb{G}_{F,j}dG\right)\right\}.
\end{aligned}$$

Proof of Theorems 1 and 3. In the i.i.d. setting, the limiting distribution of the Mann-Whitney statistic U_N can be obtained by the application of the functional delta method based on the weak convergence of $\mathbb{F}_{N^{(1)}}$ and $\mathbb{G}_{N^{(2)}}$ (see e.g. 3.9.4.1 of van der Vaart and Wellner (1996)). The same argument applies to U^{π}_N when replacing $\mathbb{F}_{N^{(1)}}$ and $\mathbb{G}_{N^{(2)}}$ by $\mathbb{F}^{\pi}_{N^{(1)}}$ and $\mathbb{G}^{\pi}_{N^{(2)}}$ with the help of the weak convergence result in Lemma 1. The map considered for U_N in the i.i.d. setting is $(A, B) \mapsto \int A dB$ where A is a cadlag function and B is a function of bounded variation. Because $\mathbb{F}^{\pi}_{N^{(1)}}$ and

$\mathbb{G}^{\pi}_{N^{(2)}}$ are a cadlag function and a function of bounded variation, respectively, the functional delta method is valid in our setting. Now, it follows from the functional delta method and Lemma 1 that

$$\begin{aligned}
&\sqrt{\frac{N^{(1)}N^{(2)}}{N}}\{U_N/(N^{(1)}N^{(2)}) - \int F dG\} = \sqrt{\frac{N^{(1)}N^{(2)}}{N}}\left(\int \mathbb{F}_{N^{(1)}} d\mathbb{G}_{N^{(2)}} - \int F dG\right)\\
&\rightsquigarrow \sqrt{q}\int F d\mathbb{G}_{G,0} + \sqrt{1-q}\int \mathbb{G}_{F,0} dG\\
&\quad + \sum_{j=1}^{J}\left(\sqrt{\frac{\nu_j q}{1-q}}\sqrt{\frac{1-p_j}{p_j}}\int F d\mathbb{G}_{G,j} + \sqrt{\frac{\nu_j(1-q)}{q}}\sqrt{\frac{1-p_j}{p_j}}\int \mathbb{G}_{F,j} dG\right)\\
&\sim N\left(0,\sigma^2\right).
\end{aligned}$$

Note that the first and second terms $\sqrt{q}\int F d\mathbb{G}_{G,0} + \sqrt{1-q}\int \mathbb{G}_{F,0} dG$ in the limiting variable are exactly the same as the limiting variable itself for U_N in the i.i.d. setting. This asymptotically normal random variable has mean zero and variance $q\mathrm{Var}(F(X)|S=2) + (1-q)\mathrm{Var}(G(X)|S=1)$ which is 1/12 when $F = G$. □

Proof of Theorem 2. The weak convergence of the corresponding bootstrap empirical process was established in Lemma 4.1 of Saegusa (2015). Applying the functional delta method for bootstrap (Theorem 3.9.13 of van der Vaart and Wellner (1996)) yields that under the null hypothesis H_0

$$\sqrt{\frac{N^{(1)}N^{(2)}}{N}}\left(\frac{\hat{U}^{\pi}_N}{N^{(1)}N^{(2)}} - \frac{U_N}{N^{(1)}N^{(2)}}\right) \to_d Z_{0,2} \sim N\left(0,\sigma^2_{0,2}\right).$$

Because the bootstrap estimate $\hat{\sigma}^2_{N,2,B}$ of σ^2_2 is consistent as $N, B \to \infty$, the estimator $\hat{\sigma}^2_{N,B}$ is consistent for σ^2_0. The desired result follows from Theorem 1 and this observation under H_0. Under the alternative hypothesis H_a, the same reasoning applies to obtain

$$\sqrt{\frac{N^{(1)}N^{(2)}}{N}}\left(\frac{\hat{U}^{\pi}_N}{N^{(1)}N^{(2)}} - \frac{U_N}{N^{(1)}N^{(2)}}\right) \to_d Z_2 \sim N\left(0,\sigma^2_2\right).$$

As $N, B \to \infty$ $\sigma^2_{N,B}$ converges in probability to a finite number $1/12 + \sigma^2_2$. Suppose $1/2 > \int F dG$. It follows from Theorem 3 that

$$\begin{aligned}
&P\left(\sqrt{\frac{N^{(1)}N^{(2)}}{N}}|U_N/(N^{(1)}N^{(2)}) - 1/2| \geq q_{N,B}\right)\\
&\geq P\left(\sqrt{\frac{N^{(1)}N^{(2)}}{N}}\left\{1/2 - \int F dG\right\} \geq q_{N,B} + \sqrt{\frac{N^{(1)}N^{(2)}}{N}}\left\{U_N/(N^{(1)}N^{(2)}) - \int F dG\right\}\right) \to 1.
\end{aligned}$$

The other case $1/2 < \int F dG$ is similar. □

References

Bickel, P. J., & Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Ann. Statist.*, *12*(2), 470–482. doi: 10.1214/aos/1176346500

Breslow, N. E., & Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *48*(4), 457–468. doi: 10.1111/1467-9876.00165

Breslow, N. E., Lumley, T., Ballantyne, C., Chambless, L., & Kulich, M. (2009a). Improved horvitz-thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Stat. Biosc.*, *1*, 32-49.

Breslow, N. E., Lumley, T., Ballantyne, C., Chambless, L., & Kulich, M. (2009b). Using the whole cohort in the analysis of case-cohort data. *American J. Epidemiol.*, *169*, 1398-1405.

Breslow, N. E., McNeney, B., & Wellner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Ann. Statist.*, *31*(4), 1110–1139.

Breslow, N. E., & Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand. J. Statist.*, *34*(1), 86–102. doi: 10.1111/j.1467-9469.2006.00523.x

D'Angio, G. J., Breslow, N., Beckwith, J. B., Evans, A., Baum, H., deLorimier, A., … Kelalis, P. (1989, Jul). Treatment of Wilms' tumor. Results of the Third National Wilms' Tumor Study. *Cancer*, *64*(2), 349–360.

Gross, S. (1980). Median estimation in sample surveys. In *Proceedings of the section on survey research methods, american statistical association* (pp. 181–184).

Saegusa, T. (2015). Variance estimation under two-phase sampling. *Scandinavian Journal of Statistics*, *42*(4), 1078–1091. doi: 10.1111/sjos.12152

Saegusa, T., & Wellner, J. A. (2013). Weighted likelihood estimation under two-phase sampling. *Ann. Statist.*, *41*(1), 269–295. doi: 10.1214/12-AOS1073

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer-Verlag.

van Elteren, P. (1960). On the combination of independent two sample test of Wilcoxon. *Bull. Inst. Internat. Statist.*, *37*(3), 351–361.