A U-statistic-based test of treatment effect heterogeneity

Maozhu Dai and Hal S. Stern
Department of Statistics, University of California – Irvine, Irvine, CA, USA

ABSTRACT
Many studies include a goal of determining whether there is treatment effect heterogeneity across different subpopulations. In this paper, we propose a U-statistic-based nonparametric test of the null hypothesis that the treatment effects are identical in different subgroups. The proposed test provides more power than the standard parametric test when the underlying distribution assumptions of the latter are violated. We apply the method to data from an economic study of programme effectiveness and find that there is treatment effect heterogeneity in different subpopulations.

1. Introduction
Treatment effect heterogeneity is of great importance, as the average treatment effect across the whole population may neglect important variability of the treatment effect across subpopulations. In health care, the concept of personalised medicine is attracting a great deal of attention (Ginsburg and Willard 2009; Jain 2015; Chan and Ginsburg 2011), as it promises a way to provide treatment recommendations with greater precision based on a patient’s baseline characteristics. In social sciences, people are using similar approaches to assess localised effects in order to comprehensively evaluate a policy or a campaign strategy (Bitler, Gelbach, and Hoynes 2006; Feller and Holmes 2009). In one criminology study, arrest can result in effects of opposite sign on recidivism rates for different kinds of criminals (Pate and Hamilton 1992). In all areas, subgroup analysis (Dixon and Simon 1991) can be conducted to assess treatment effects at the subpopulation level. Both randomised studies and observational studies can provide insight into defining relevant subpopulations. In randomised clinical trials, patients can be assigned to different subgroups based on one or several baseline factors. In observational studies, one approach to identify subpopulation is through subclassification on propensity scores (Xie, Brand, and Jann 2012). In either case, reliable inference about whether there is heterogeneity in treatment effects is needed before undertaking an approach based on subpopulations.

There is a great deal of literature on exploring heterogeneity of treatment effects. The published results focus on different aspects of heterogeneity in that they examine different null hypotheses. Some focus on testing conditional stochastic dominance (Delgado and
Escanciano 2013; Hsu 2017). Other focus on testing whether the treatment effects in sub-populations are of the same sign (Gail and Simon 1985; Chang et al. 2015). Crump, Hotz, Imbens, and Mitnik (2008) and Chang et al. (2015) are interested in testing whether the treatment effects are consistently equal to zero. Ding, Feller, and Miratrix (2016) focus on the null hypothesis that all individuals have the same treatment effects. In this paper, we focus on testing whether the average stratum-specific treatment effects are constant across different strata.

A common approach to identifying whether the stratum-specific average treatment effects are equal across different strata is through parametric statistical tests. For example, we can test the interaction term between treatment assignment and effect modifiers in multiple linear regressions (Allison 1977; Byar 1985). We can also use the likelihood ratio test (LRT) proposed by Gail and Simon (1985). These approaches are widely used but rely on the parametric assumptions of outcome distributions being correct. When the assumptions are not correct, the inference is invalid or the power of the test is impacted.

Nonparametric approaches exist as well. Crump et al. (2008) created a nonparametric approach based on a particular series estimator for treatment effect introduced by Imbens, Newey, and Ridder (2006). Sant’Anna (2021) generalised (Crump et al. 2008) by allowing censored data and endogenous treatment selections. In this paper, we propose a U-statistic-based approach to test whether the stratum treatment effects are homogeneous without having to estimate the stratum treatment effects. Our approach relies on an unconfoundedness assumption in each stratum. When there are confounders, methods that create appropriate strata are needed such as subclassification on propensity scores (as proposed by Xie et al. 2012).

U-statistics have been widely used to create distribution-free tests. Examples include the signed rank test and Mann-Whitney test (Van der Vaart 2000). Compared to their parametric counterparts assuming normal distributions, U-statistic-based tests usually have higher power when the distributions are far from normal. When the normality assumption is satisfied, the parametric test has slightly higher power, but when the U-statistic-based test is more powerful, the advantage can be significant (Hodges and Lehmann 1956; Lehmann and D’Abrera 1975; Zimmerman 1998). The nonparametric heterogeneity test we propose here is also based on U-statistics. We use U-statistics to compute a test statistic comparing treatment effects across pairs of strata. The overall test statistic is a combination of the pairwise test statistics. Its performance will be compared to the LRT proposed by Gail and Simon (1985).

The remainder of the paper is structured as follows. In Section 2, we provide a review of U-statistics. In Section 3, we introduce our proposed U-statistic-based nonparametric test for treatment effect heterogeneity in detail. In Section 4, some simulation studies demonstrate the validity of the test and its comparison with the LRT under several different circumstances. In Section 5, we apply the proposed method to a randomised study of programme effectiveness in labour economics. Additional discussion of this approach can be found in Section 6.

2. Background: review of U-statistics

U-statistics are a class of statistics widely used to construct nonparametric unbiased estimators of estimable parameters with minimum variance. The asymptotic normality property
of U-statistics (under some mild conditions) makes it very popular as a nonparametric testing tool.

We start with a review of one-sample U-statistics (Van der Vaart 2000). Let $X_1, \ldots, X_n$ be a random sample from $F(x)$, and assume there is a symmetric function $\phi(x_1, \ldots, x_m)$ ($m \leq n$) such that $E[\phi(X_1, \ldots, X_m)] = \theta$, where $\theta$ is the parameter of interest. Then the U-statistic for the parameter $\theta$ created by kernel $\phi$ is

$$U(X_1, \ldots, X_n) = \frac{1}{\binom{n}{m}} \sum_{\beta \in B} \phi(X_{\beta_1}, \ldots, X_{\beta_m}),$$

where $B$ contains all $\binom{n}{m}$ ordered subsets $\beta = (\beta_1, \ldots, \beta_m)$ of $m$ integers chosen without replacement from the set $\{1, \ldots, n\}$ with $1 \leq \beta_1 < \cdots < \beta_m \leq n$.

The signed rank statistic is an example of a one-sample U-statistic where $\theta = E(I(X_1 + X_2 > 0))$. This can be used to test whether the location (median) of a symmetric distribution is equal to 0 via testing whether $\theta = \frac{1}{2}$. With the symmetric kernel $\phi(x_1, x_2) = I(x_1 + x_2 > 0)$, the corresponding U-statistic estimator of $\theta$ is

$$U = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} I(X_i + X_j > 0).$$

A key property of the one-sample U-statistic is that it has an asymptotic normal distribution. Denote $\zeta = \text{Var}(h(X_1))$ with $h(X_1) = E(\phi(X_1, \ldots, X_m)|X_1) - \theta$. If $E\phi^2(X_1, \ldots, X_m) < \infty$ and $\zeta > 0$, then

$$\sqrt{n}(U - \theta) \xrightarrow{D} N(0, \sigma^2),$$

where $\sigma^2 = m^2 \zeta$.

The multi-sample U-statistic is a natural extension of the one-sample U-statistic. Let $\{X_{1\alpha}, \alpha = 1, \ldots, n_1\}, \ldots, \{X_{c\delta}, \delta = 1, \ldots, n_c\}$ be $c$ independent random samples from distribution functions $F_1(x), \ldots, F_c(x)$ respectively, and $\phi(x_{11}, \ldots, x_{1m_1}; \ldots; x_{c1}, \ldots, x_{cm_c})$ be a symmetric function within each set of variables $\{x_{j1}, \ldots, x_{jm_j}\}$ ($j = 1, \ldots, c$) with $E(\phi)$ equal to the parameter of interest $\theta$, where $m_1 \leq n_1, \ldots, m_c \leq n_c$. Then the corresponding $c$-sample U-statistic is

$$U = \left[ \prod_{j=1}^c \binom{n_j}{m_j} \right]^{-1} \sum_{\alpha_1} \cdots \sum_{\alpha_c} \phi(X_{1\alpha_1,1}, \ldots, X_{1\alpha_1,m_1}; \ldots; X_{c\alpha_c,1}, \ldots, X_{c\alpha_c,m_c})$$

where the summation is over all possible sets of subscripts $\alpha_j = (\alpha_{j,1}, \ldots, \alpha_{j,m_j})$ such that $1 \leq \alpha_{j,1} < \cdots < \alpha_{j,m_j} \leq n_j$ for each of the $c$ samples (i.e. $j = 1, \ldots, c$).

The Mann-Whitney statistic is an example of two-sample U-statistic with $m_1 = m_2 = 1$. The parameter of interest is $\theta = E(I(X_{11} < X_{21}))$. Under the assumption that the distributions of $X_{1i}$ and $X_{2j}$ are the same up to location shift, $\theta = \frac{1}{2}$ indicates that there is no location shift. The U-statistic with respect to the parameter of interest is

$$U = \frac{1}{\binom{n_1}{1} \binom{n_2}{1}} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(X_{1i} < X_{2j}).$$
The Mann-Whitney statistic is the test statistic for the Mann-Whitney test (Mann and Whitney 1947), which is a consistent test for the null hypothesis that $\theta = \frac{1}{2}$ versus the alternative hypothesis that $\theta \neq \frac{1}{2}$.

There is also an asymptotic normality property for multi-sample U-statistics, even for a vector of several multi-sample U-statistics defined upon the same sets of mutually independent samples with different kernel functions. Lehmann (1963) showed that if there are $r$ multi-sample U-statistics $U^{(1)}, \ldots, U^{(r)}$, each defined as in (4), with corresponding kernel functions $\phi^{(1)}, \ldots, \phi^{(r)}$ such that $E[\phi^{(k)}] = \theta^{(k)}$, $E[(\phi^{(k)})^2] < \infty$, $\text{Var}(\phi^{(k)}|X_{j1}) > 0$ for $k \in \{1, \ldots, r\}, j \in \{1, \ldots, c\}$, and if there also exists positive constants $\lambda_j$ ($0 < \lambda_j < 1$) such that $\frac{n_j}{N} \to \lambda_j$ as $N = \sum_{j=1}^{c} n_j \to \infty$ for $j \in \{1, \ldots, c\}$, then

$$
\sqrt{N} \left( \begin{array}{c}
U^{(1)} - \theta^{(1)} \\
U^{(2)} - \theta^{(2)} \\
\vdots \\
U^{(r)} - \theta^{(r)}
\end{array} \right) \xrightarrow{D} N(0, \Sigma),
$$

(6)

where $\Sigma$ is the asymptotic covariance matrix of $\sqrt{N}(U^{(1)}, U^{(2)}, \ldots, U^{(r)})$.

In order to apply (6) to hypothesis testing with regard to the parameter $\theta = (\theta^{(1)}, \ldots, \theta^{(r)})$, we need to identify the form of $\Sigma$. This can be addressed using the Hájek projection principle (Hájek 1968) to derive the asymptotic normality property of the U-statistics.

For one $c$-sample U-statistic of degree $(m_1, \ldots, m_c)$, if $E(\phi^2) < \infty$, the Hájek projection of $U - \theta$ onto the space $V = \{V|V = \sum_{i=1}^{m_1} f_1(X_{1i}) + \cdots + \sum_{i=1}^{m_c} f_c(X_{ci})\}$ where $f_j$ ($j \in \{1, \ldots, c\}$) are some real-valued functions) is

$$
\hat{U} = \frac{m_1}{n_1} \sum_{i=1}^{n_1} h_1(X_{1i}) + \cdots + \frac{m_c}{n_c} \sum_{i=1}^{n_c} h_c(X_{ci}),
$$

(7)

where the $h$ functions are defined as

$$
h_j(x) = E[\phi(X_{11}, \ldots, X_{1m_1}; \ldots; X_{1c}, \ldots, X_{cm_c})|X_{j1} = x] - \theta, \quad j \in \{1, \ldots, c\}.
$$

(8)

Then it can be proved (Korolyuk and Borovskich 2013) that

$$
\sqrt{N}(U - \theta - \hat{U}) \xrightarrow{P} 0 \quad \text{as } N \to \infty.
$$

(9)

This shows that $U - \theta$ and $\hat{U}$ have the same asymptotic distribution. By the Central Limit Theorem,

$$
\sqrt{N} \hat{U} \xrightarrow{d} N \left( 0, \frac{m_1^2}{\lambda_1} \text{Var}(h_1(X_{1i})) + \cdots + \frac{m_c^2}{\lambda_c} \text{Var}(h_c(X_{ci})) \right) \quad \text{as } N \to \infty,
$$

(10)

provided the variance terms are finite. Thus we have

$$
\sqrt{N}(U - \theta) \xrightarrow{D} N \left( 0, \frac{m_1^2}{\lambda_1} \text{Var}(h_1(X_{1i})) + \cdots + \frac{m_c^2}{\lambda_c} \text{Var}(h_c(X_{ci})) \right) \quad \text{as } N \to \infty.
$$

(11)
With the list of U-statistics \((U^{(1)}, \ldots, U^{(r)})\), there is a list of Hájek projection \((\hat{U}^{(1)}, \ldots, \hat{U}^{(r)})\) corresponding to each of them with

\[
\hat{U}^{(k)} = \frac{m_1^{(k)}}{n_1} \sum_{i=1}^{n_1} h_1^{(k)}(X_{1i}) + \cdots + \frac{m_c^{(k)}}{n_c} \sum_{i=1}^{n_c} h_c^{(k)}(X_{ci}) \quad \text{for } k \in \{1, \ldots, r\},
\]

where \(h_j^{(k)}(x) = E[\phi^{(k)}(X_{1i}, \ldots, X_{1m}; \ldots; X_{ci}, \ldots, X_{cn} | X_{1j} = x] - \theta^{(k)} \) for \(j \in \{1, \ldots, c\}\).

By the multidimensional Central Limit Theorem, we know

\[
\sqrt{N} \left( \begin{array}{c}
\hat{U}^{(1)} \\
\hat{U}^{(2)} \\
\vdots \\
\hat{U}^{(r)}
\end{array} \right) \xrightarrow{D} N \left( \begin{array}{c} 
0, \frac{1}{\lambda_1} \Sigma_1 + \cdots + \frac{1}{\lambda_c} \Sigma_c 
\end{array} \right)
\]

where \(\Sigma_j = \text{Cov}[m_j^{(1)} h_j^{(1)}(X_j), \ldots, m_j^{(r)} h_j^{(r)}(X_j)]\) for \(j \in \{1, \ldots, c\}\). Since

\[
\sqrt{N} \left( \begin{array}{c}
U^{(1)} \theta^{(1)} - \hat{U}^{(1)} \\
U^{(2)} \theta^{(2)} - \hat{U}^{(2)} \\
\vdots \\
U^{(r)} \theta^{(r)} - \hat{U}^{(r)}
\end{array} \right) \xrightarrow{P} 0 \quad \text{as } N \to \infty,
\]

we have

\[
\sqrt{N} \left( \begin{array}{c}
U^{(1)} \theta^{(1)} \\
U^{(2)} \theta^{(2)} \\
\vdots \\
U^{(r)} \theta^{(r)}
\end{array} \right) \xrightarrow{D} N \left( \begin{array}{c} 
0, \frac{1}{\lambda_1} \Sigma_1 + \cdots + \frac{1}{\lambda_c} \Sigma_c 
\end{array} \right).
\]

3. Testing for treatment effect heterogeneity

Suppose we are focussed on a study population comprised of \(S\) strata. For each stratum \(s\), \(s \in \{1, \ldots, S\}\), let \(Y_t^s\) denote the outcomes of subjects in the treatment group where \(Y_t^s = \{Y_{si}^t, i = 1, \ldots, n_t^s\}\), and \(Y_c^s\) denotes the outcomes in the control group where \(Y_c^s = \{Y_{ci}^c, i = 1, \ldots, n_c^s\}\). Define \(N_s = n_t^s + n_c^s\) as the total sample size in strata \(s\), and \(N = \sum_{s=1}^{S} N_s\) as the overall sample size across all strata. We develop a nonparametric U-statistic-based test (U test) for the null hypothesis of no treatment effect heterogeneity against the alternative hypothesis that not all treatment effects are equal. In the derivation and in our studies, we focus on an assumed additive treatment effect. Alternative methods of the treatment effect can be considered, they would require alternative choices for the U-statistic kernel functions.

The technique to be discussed here relies on two assumptions: (1) \(Y_t^1, \ldots, Y_t^S, Y_c^1, \ldots, Y_c^S\) are mutually independent; (2) There exist constants \(\lambda_s^\omega \in (0, 1)\) such that \(\frac{n_s^\omega}{N} \to \lambda_s^\omega\) for all \(s \in \{1, \ldots, S\}\) and \(\omega \in \{t, c\}\).
3.1. Comparing treatment effects between the first two strata

We start by constructing a U-statistic comparing the treatment effects of the first two strata. The hypotheses we focus on are

\[ H_0 : P(Y^t_1 - Y^c_1 < Y^t_2 - Y^c_2) + \frac{1}{2}P(Y^t_1 - Y^c_1 = Y^t_2 - Y^c_2) = \frac{1}{2} \]

\[ H_a : P(Y^t_1 - Y^c_1 < Y^t_2 - Y^c_2) + \frac{1}{2}P(Y^t_1 - Y^c_1 = Y^t_2 - Y^c_2) \neq \frac{1}{2}. \]

The term \( \frac{1}{2}P(Y^t_1 - Y^c_1 = Y^t_2 - Y^c_2) \) is used to account for possible ties for discrete distributions. Under the null hypothesis \( H_0 \), \( E[\lambda^o_s] \leq 1 \) and under three assumptions: (1) the outcomes \( (Y^t_1, Y^t_2, Y^c_1, Y^c_2) \) are mutually independent; (2) there exist positive constants \( 0 < \lambda^o_s < 1 \) for every \( s \in \{1, 2\} \) and \( \omega \in \{t, c\} \) such that \( \frac{n^o_s}{N} \to \lambda^o_s \) as \( N \to \infty \); and (3) \( 0 < \text{Var}(h_s^{o,(1,2)}(Y_s^o)) \to \infty \) for \( s \in \{1, 2\} \) and \( \omega \in \{t, c\} \), we have

\[ \sqrt{N}(U^{(1,2)} - \theta^{(1,2)}) \xrightarrow{D} N(0, \sigma^2_{1,2}), \quad \text{when } N \to \infty, \]

where

\[ \sigma^2_{1,2} = \frac{1}{\lambda^o_1} \text{Var}(h^{t,(1,2)}_1(Y^t_1)) + \frac{1}{\lambda^o_1} \text{Var}(h^{c,(1,2)}_1(Y^c_1)) + \frac{1}{\lambda^o_2} \text{Var}(h^{t,(1,2)}_2(Y^t_2)) + \frac{1}{\lambda^o_2} \text{Var}(h^{c,(1,2)}_2(Y^c_2)) \in (0, \infty), \]

\[ h_s^{o,(1,2)}(x) = E[\phi^{(1,2)}(Y^t_1; Y^c_1, Y^t_2; Y^c_2)|Y^o_s = x] - \theta^{(1,2)} \quad \text{for } s \in \{1, 2\}, \omega \in \{t, c\}. \]

Hypothesis testing based on this asymptotic result is consistent for hypotheses \( H_0 \) vs \( H_a \) in (16).

To apply this method, we first estimate \( h_s^{o,(1,2)}(x) \) (\( s \in \{1, 2\} \) and \( \omega \in \{t, c\} \)), an expectation, by the method of moments. For instance, \( h^{t,(1,2)}_1(x) \) is estimated by the sample mean \( \hat{h}_1^{t,(1,2)}(x) = \frac{1}{n^t_1 n^c_1 n^t_2 n^c_2} \sum_{i=1}^{n^t_1} \sum_{j=1}^{n^t_2} \sum_{k=1}^{n^c_1} \sum_{l=1}^{n^c_2} I(x - Y^c_{ij} < Y^t_{2k} - Y^c_{2l}) \). Note that this calculation is repeated with each data value \( Y^t_{ij} \) (\( i = 1, \ldots, n^t_1 \)) taking the place of \( x \). Likewise for other \( h \) terms. Then we estimate \( \text{Var}(h_s^{o,(1,2)}(Y^o_s)) \) by the sample variance of \( \hat{h}_s^{o,(1,2)}(Y^o_s) \) as \( \frac{1}{n^o_s - 1} \sum_{i=1}^{n^o_s} \left[ \hat{h}_s^{o,(1,2)}(Y^o_{si}) - \frac{1}{n^o_s} \sum_{j=1}^{n^o_s} \hat{h}_s^{o,(1,2)}(Y^o_{sj}) \right]^2 \), for \( s \in \{1, 2\} \) and \( \omega \in \{t, c\} \), and take the weighted sum of them to approximate \( \sigma^2_{1,2} \).
3.2. Testing treatment effect heterogeneity across multiple strata

With S strata, the hypotheses we focus on are

\[ H_0 : P(Y_p^t - Y_p^c < Y_q^t - Y_q^c) + \frac{1}{2} P(Y_p^t - Y_p^c = Y_q^t - Y_q^c) = \frac{1}{2} \quad \text{for any } 1 \leq p < q \leq S \]

\[ H_a : \text{the equation does not hold for at least one pair of } (p, q). \] (19)

For any pair of strata, we can construct a test statistic like (17). Denote \( U_{(p,q)} \) with \( p < q \) as the U-statistic comparing strata \( p \) and \( q \) with kernel \( \phi_{(p,q)}(Y_p^t, Y_p^c, Y_q^t, Y_q^c) = I(y_p^t - y_p^c < y_q^t - y_q^c) + \frac{1}{2} I(y_p^t - y_p^c = y_q^t - y_q^c) \) and expectation \( \theta_{(p,q)} \). By applying the Hájek projection principle to a vector of multi-sample U-statistics as in Section 2, we have

\[
\sqrt{N} \begin{pmatrix}
U_{(1,2)} - \theta_{(1,2)} \\
U_{(1,3)} - \theta_{(1,3)} \\
\vdots \\
U_{(S-1,S)} - \theta_{(S-1,S)}
\end{pmatrix} \xrightarrow{D} N(0, \Sigma)
\] (20)

where \( \Sigma = \frac{1}{\lambda_1} \Sigma_1^t + \frac{1}{\lambda_1} \Sigma_1^c + \cdots + \frac{1}{\lambda_S} \Sigma_S^t + \frac{1}{\lambda_S} \Sigma_S^c \) and \( \Sigma_s^{\omega} = \text{Cov}(\hat{h}_s^{\omega,(1,2)}(Y_s^\omega), \ldots, \hat{h}_s^{\omega,(1,S)}(Y_s^\omega), \hat{h}_s^{\omega,(2,3)}(Y_s^\omega), \ldots, \hat{h}_s^{\omega,(S-1,S)}(Y_s^\omega)) \) for all \( s \in \{1, \ldots, S\} \) and \( \omega \in \{t, c\} \). Here

\[
\hat{h}_s^{\omega,(p,q)}(x) = \begin{cases} 
\hat{h}_s^{\omega,(p,q)}(x) = E[\phi_{(p,q)}(Y_p^t, Y_p^c, Y_q^t, Y_q^c)|Y_s^\omega = x] - \theta_{(p,q)} & \text{if } s = \text{por } s = q, \\
o.w. & \text{otherwise.}
\end{cases}
\]

for \( \{(p, q)|1 \leq p < q \leq S\} \). Under the null hypothesis \( H_0 \) in (19), all \( \theta \)'s are equal to \( \frac{1}{2} \).

Estimation of \( \Sigma \) is carried out using a similar approach as described for estimation of \( \sigma_{1,2}^2 \) in Section 3.1. We first construct an empirical estimate for each \( h \) function (as in the paragraph below (17)) and then use the sample covariance matrix of \( \{\hat{h}_s^{\omega,(1,2)}(Y_s^\omega), \ldots, \hat{h}_s^{\omega,(1,S)}(Y_s^\omega), \hat{h}_s^{\omega,(2,3)}(Y_s^\omega), \ldots, \hat{h}_s^{\omega,(S-1,S)}(Y_s^\omega)\} \) with \( h \)'s replaced by their corresponding estimates to get \( \hat{\Sigma}_s^{\omega} \). Then \( \hat{\Sigma} \), the estimate of \( \Sigma \), is the sum of the \( \hat{\Sigma}_s^{\omega} \) over \( s \in \{1, \ldots, S\} \) and \( \omega \in \{t, c\} \) with each term weighted by \( \frac{1}{\lambda_s} \).

The vector of pairwise test statistics \( U = (U_{(1,2)}, U_{(1,3)}, \ldots, U_{(S-1,S)})^T \) can be combined into a single overall test statistic using any function of \( U \). Here we focus on \( U_h = N \cdot \sum_{1 \leq p < q \leq S} |U_{(p,q)} - \frac{1}{2}|^2 \). The asymptotic distribution of \( U_h \) is not available in analytic form, but a simulation approach can be used to assess \( U_h \). A large number of independent samples of \( \sqrt{N}(U - \frac{1}{2}) \) are generated from the null \( N(0, \Sigma) \) distribution, and \( U_h \) is computed for each sample to generate the empirical null distribution of \( U_h \). For \( \alpha \) level test, we reject \( H_0 \) when \( U_h \) is greater than or equal to the \( 100(1 - \alpha) \) percentile of the empirical null distribution. Note that other test statistics are also possible, e.g. \( \sqrt{N} \cdot \max_{1 \leq p < q \leq S} |U_{(p,q)} - \frac{1}{2}| \), and simulation is always an option for deriving the reference distribution. In our simulation study, we use the statistic \( U_h \) because it proved reliable. Hypothesis testing based on \( U_h \) is consistent for the hypotheses in (19).

Another test statistic that might seem natural is \( T = N(U - \frac{1}{2}) \cdot \hat{\Sigma} - (U - \frac{1}{2})^T \) whose reference distribution is \( \chi^2_k \), where \( k \) is the rank of \( \hat{\Sigma} \). \( \chi^2 = \sum_{i=1}^k \frac{1}{\alpha_i} q_i q_i^T \) is a generalised inverse of \( \hat{\Sigma} \), where \( \{\alpha_i, i = 1, \ldots, k\} \) are the non-zero eigenvalues of \( \hat{\Sigma} \) and \( \{q_i, i = 1, \ldots, k\} \) are the corresponding eigenvectors.
1, ..., k] are the corresponding orthogonal eigenvectors. However, as \( U \) is a vector of all of the pairwise U-statistics, the determinant of its covariance matrix can be very close to 0. Then \( \Sigma \) can have an eigenvalue \( \alpha \) very close to 0, for which a tiny rounding error would have a large impact on \( \Sigma^{-} \) and thus on \( T \). So even though the reference distribution of the test statistic \( T \) has known distribution, we prefer using \( U_h \).

### 3.3. Three specific cases to consider

The approach described above does not make any specific assumption about the shapes of the distributions of \( Y_{1t}, \ldots, Y_{St}, Y_{1c}, \ldots, Y_{Sc} \) or the relationship among them other than independence. Our U-statistic is testing whether the probability \( P(Y_{pt} < Y_{qt} \mid \text{alternative}) + \frac{1}{2} P(Y_{pt} = Y_{qt} \mid \text{null}) \) is equal to one half for all \( p, q \). This does not however provide much insight into the feature of outcomes that is being tested. When there is no further assumptions regarding the outcome distributions, confusion may arise as to the meaning of a significant result. We describe three cases here, with a set of assumptions for each of them, under which the interpretation of the test is clear. Many real-life problems may fit into those cases.

**Case A:** We assume all outcomes in different strata and different treatment groups follow a common distribution \( F \) up to a location shift, which is comprised of the additive treatment effects \( \tau_s \) within each stratum and assumed additive stratum effects \( \Delta_s \) for \( s \in \{1, \ldots, S\} \). In this case, \( P(Y_{pt} < Y_{qt} \mid \text{alternative}) + \frac{1}{2} P(Y_{pt} = Y_{qt} \mid \text{null}) \) is equal to a half if and only if the within-stratum location shifts \( (\tau_p, \tau_q) \) are the same. It can then be easily shown that our test is consistent with identifying whether \( \tau_s \) is the same across all strata. Since the shift can be considered as the difference in means or difference in any fixed percentile between two different treatment groups, the test is consistent with respect to the alternative hypothesis that the difference of means (or percentiles) are unequal for at least one pair of strata. This case is equivalent to testing the interaction in a two-factor factorial design when one factor has two levels; in our case, one factor is stratum and the other is treatment which takes two levels. The ANOVA F-test can be used to address this scenario, and nonparametric tests have been proposed as an alternative (Patel and Hoel 1973; De Neve and Thas 2017). The two approaches proposed by De Neve and Thas (2017) are similar to our approach in that they use the same basic U-statistic that we do. They focus on testing interaction terms for factorial designs. A limitation is that one of these approaches does not allow comparison among more than two strata and the other approach only applies to this case (Case A) and not to Case B or Case C that are described next.

**Case B:** We assume that all outcomes in the treatment groups follow a common distribution \( F^t \) up to a stratum-specific location shift \( \Delta^t_s \) and all outcomes in the control groups follow a possibly different common distribution \( F^c \) up to a stratum-specific location shift \( \Delta^c_s \) for all \( s \in \{1, \ldots, S\} \). Then the within-stratum treatment effect can be depicted as the difference of the location shift \( \Delta^t_s - \Delta^c_s \) plus the difference of \( F^t \) and \( F^c \). No matter what metric is used to describe the difference of \( F^t \) and \( F^c \), it is consistent across all strata. So in this case, \( P(Y_{pt} < Y_{qt} \mid \text{alternative}) + \frac{1}{2} P(Y_{pt} = Y_{qt} \mid \text{null}) \) is equal to \( \frac{1}{2} \) if and only if \( \Delta^t_p - \Delta^c_p = \Delta^t_q - \Delta^c_q \). So our test is identifying whether the difference of location shifts \( \Delta^t_s - \Delta^c_s \) are consistent across strata, which can also be considered as the consistency of
the difference of means or percentiles between different treatment groups. We can easily prove the test is consistent with the alternative hypothesis that the difference of means (or percentiles) between the treatment and control groups are not all identical across strata.

**Case C:** We assume that the outcomes in the same strata $s \in \{1, \ldots, S\}$ follow the same distribution $F_s$ up to a stratum-specific location shift $\tau_s$, the additive treatment effect. Similarly, in this case, our method is a consistent test with respect to the alternative hypothesis that there are at least two strata such that the difference of means (or percentiles) between the treatment and control groups are different.

**Case A** can be applied in practice whenever we think the factor we use to stratify subjects, as well as the treatment, can only shift the location of the outcome distributions but cannot change the shape. This is a common assumption. **Case B** can be applied when the treatment changes the shape of the outcome distributions, whereas the factor used to stratify subjects would only shift the location. Suppose we are studying whether the effect of a welfare reform policy on household income differs across several different geographic regions. Usually a welfare reform plan has different magnitude of impact on families who are on different economic levels. Thus we may expect the distribution of income after the reform to differ from the control condition. If the shape of income distribution is similar across geographic regions or just differs by scale among different regions, then this scenario could fit in our **Case B**. **Case C** can be applied when the treatment effects are constant within each stratum, but the distribution of outcomes differs across strata. Back to the welfare reform example. If we stratify subjects according to their economic levels, then income distributions would be expected to vary among different strata, but it is possible that treatment effects could be constant.

### 4. Simulation study

We demonstrate the U-statistic test of treatment effect heterogeneity via a simulation study. There are some computational challenges that are addressed first, then the simulation study is described and results are provided. Simulations compare the U test to the LRT in a range of scenarios, some of which match the assumptions in the parametric test and others do not. Additionally, we show the relationship among power, sample size and effect size in two scenarios.

#### 4.1. Computational issues

When the sample sizes are large, the computation of the U-statistic is computationally expensive. Let’s take $U^{(1,2)}$ as an example. We need to compute the average over all the combinations of $I(Y^t_{1i} - Y^c_{1j} < Y^t_{2k} - Y^c_{2l}) + \frac{1}{2}I(Y^t_{1i} - Y^c_{1j} = Y^t_{2k} - Y^c_{2l})$, denoted by $\phi^{(1,2)}(i, j, k, l)$, which includes $n_1^t \times n_1^c \times n_2^t \times n_2^c$ terms. As this computation can be done in parallel for different $(i, j, k, l)$, it should not be a big problem when applying the method for a single data set. In simulations, we need to generate thousands of data sets and compute U-statistics for each of them. So in the simulation study, instead of computing the average through exhaustive enumeration, we generate approximate U-statistics by randomly selecting some of the combinations with replacement and use this average to approximate the U-statistic. We randomly selected $M = 10^3 N$ samples with replacement from each treatment subgroup as $\{y^t_{1i}, y^c_{1i}, y^t_{2i}, y^c_{2i}, i = 1, \ldots, M\}$ to approximate $U^{(1,2)}$, here $N$ is the total
sample size of the two strata. The sample size $M$ was determined by considering a range of scenarios and estimating the variance of the test statistic $U_h = N \cdot \sum_{1 < i < j \leq S} (U(i,j) - \frac{1}{2})^2$ constructed by approximate U-statistics within each scenario via simulation. The variance increases as $N$ increases. In the simulation, as $N$ ranged from 60 to 3000, the maximum estimated variances for this choice of $M$ ranged from 0.0058 to 0.104, which was judged to provide sufficient precision. One requirement of this sampling is that all subjects have to be selected at least once, because we also used the sampled indicators to estimate $h_{10,12}(s)$.

For instance, the estimate of $h_{10,12}(s)$ is computed as the average of all selected $\phi(1,2)(i,j,k,l)$ with $Y_{1i}^s$ equal to $Y_{1i}$. Though this requirement is not a challenge due to the large sampling size $M$, in the rare events that it occurs we would need to redo the sampling procedure. As for the empirical reference distribution of the test statistic $U_h$, we generated $10^5$ random samples $r_i = (r_{i1,2}, \ldots, r_{i(S-1),S})$ ($i = 1, \ldots, 10^5$) from the multivariate normal distribution $N(0, \hat{\Sigma})$ for each simulation, and got the distribution of $||r_i||^2$ as the empirical reference distribution under $H_0$. The empirical $p$-value is the percentage of generated $||r_i||^2$ greater than $U_h$. Fixing the type I error as $\alpha = 0.05$, we reject $H_0$ when the $p$-value is smaller than $\alpha$. We determined the sample size $10^5$ by considering a range of scenarios and estimating the variance associated with a simulation-based 95th percentile. The sample size of $10^5$ in the various cases makes the variance less than 0.01.

4.2. Review of the likelihood ratio test for treatment effect heterogeneity

The likelihood ratio test for treatment effect heterogeneity was developed by Gail and Simon (1985). Let $\tau_s = E[Y_{1i}^s] - E[Y_{ci}^s]$ denote the treatment effect in subgroup $s$ ($s \in \{1, \ldots, S\}$). The test assesses the null hypotheses

$$H_0' : \tau_1 = \cdots = \tau_S$$

$$H_a' : \text{at least two of them are not equal.}$$

Under the assumption that $\hat{\tau}_s(s \in \{1, \ldots, S\})$ follows a normal distribution with

$$\hat{\tau}_s \text{ indep } N(\tau_s, \sigma_s^2), \quad s \in \{1, \ldots, S\},$$

we have heterogeneity test statistic

$$H = \sum_{s=1}^{S} (\hat{\tau}_s - \bar{\tau})^2 / s_s^2 \sim \chi_{S-1}^2$$

where $\bar{\tau} = \left( \sum_{s=1}^{S} \hat{\tau}_s / s_s^2 \right) / \left( \sum_{s=1}^{S} 1 / s_s^2 \right)$, and $s_s^2$ is a consistent estimator of $\sigma_s^2$.

With fixed type I error $\alpha$, we reject the null hypotheses $H_0'$ when the test statistic $H$ is greater than or equal to the $100(1 - \alpha)$th percentile of $\chi_{S-1}^2$.

For additive treatment effects, the treatment effect estimates $\hat{\tau}_s$ can be the difference between the sample means of the two treatment groups within strata $s$. When the subgroup sample sizes are large, according to the Central Limit Theorem, $\hat{\tau}_s$ will approximate
to a normal distribution. However, when the distributions of the outcomes differ from normality and the validity of the test relies on large sample sizes, the power of the test will be impacted as with other parametric tests (Lehmann 1999).

4.3. Simulation study design

Assuming we have three strata, we generated \( n^\omega_s \) random samples from treatment subgroup \( \omega \) within strata \( s \) from a distribution \( F^\omega_s \) \((s \in 1, 2, 3; \omega \in \{t, c\})\). The choices of \( n^\omega_s \) and \( F^\omega_s \) are described below. The hypotheses in (19) and (22) were tested via our nonparametric U test and LRT respectively. For all scenarios to be described below, the two pairs of hypotheses are equivalent. So the comparison between the two tests is fair. For each simulation scenario (choices of \( n^\omega_s \) and \( F^\omega_s \)), we generate \( L = 2000 \) data sets and carry out the tests on each. This yields rejection rates and the empirical distribution of \( p \)-values.

We developed 17 different scenarios for the choice of the distributions \( F^\omega_s \). These scenarios are listed in Table 1. They are organised according to the three application cases outlined in Section 3.3. For each scenario, the true treatment effects were varied to provide the null and alternative instances. The upper half of Table 1 lists the null cases and the lower half lists the alternative cases.

For Case A, outcomes in the three strata and two treatment groups follow a common distribution \( F \) up to location shifts. The first scenario (A1) is that \( F = N(0, 1) \). For the next two scenarios (A2 and A3), \( F \) are still symmetric distributions, but with tails lighter (A2) or heavier (A3) than normal distribution. Next we consider three skewed distributions \( \chi^2_1 \), \( \chi^2_2 \) (labelled as A4–A6) with their skewnesses decreasing in that order. The support of these distributions are all positive. We are generally more interested in comparing the scales of the treatment and control groups instead of location shifts in these scenarios. So we suppose there are constants \( c^\omega_s \) such that \( \frac{\chi^\omega_s}{c^\omega_s} \sim F \) \((s \in 1, 2, 3; \omega \in \{t, c\})\). Here we use the logarithm of those outcomes in our test statistic \( U_h \) instead of using the original outcomes directly. Now the problem of testing the consistency of ratio of scales is changed into a problem of testing the consistency of the location shift \( \log(c^t_s) - \log(c^c_s) \) \((s = 1, \ldots, S)\). The final Case A example is a bimodal distribution \( 0.5N(-5, 1) + 0.5N(5, 1) \) (labelled as A7).

In Case B, all outcomes in the treatment groups follow a common distribution \( F^t \) up to a location shift and all outcomes in the control group follow a different common distribution \( F^c \) up to a location shift. To create examples here, we choose two of the distributions used in Case A whose supports are the whole real line (A1–A3, A7) and randomly assign them to the treatment and control group. We try all \( \binom{3}{2} \) combinations, they are labelled as B1–B6.

In Case C where all outcomes in the same stratum follow a common distribution \( F_s \) \((s = 1, \ldots, S)\) up to a location shift, we select three distribution from Case A with support on the whole real line and randomly assign them to the three strata. So we have \( \binom{3}{3} \) combinations and they are labelled as scenarios C1–C4.

For each scenario described above, we vary the true treatment effects to get the null and alternative cases, and also consider a range of different sample sizes. In each scenario, there are stratum–specific location shifts \((\Delta_1 = 0, \Delta_2 = 1, \Delta_3 = 2 \) for scenarios in Case A and Case B). For all null cases, the stratum treatment effects are the same across the three strata. For alternative cases, the treatment effects \((\tau_1, \tau_2, \tau_3)\) form an arithmetic series with \( \tau_2 = \tau_1 + \Gamma \) and \( \tau_3 = \tau_1 + 2\Gamma \), where \( \Gamma > 0 \). Simulations were carried out for a range of
| Scenario | $F_1^*$ | $F_1^*$ | $F_2^*$ | $F_2^*$ | $F_3^*$ | $F_3^*$ |
|----------|---------|---------|---------|---------|---------|---------|
| **Null cases** | | | | | | |
| A1 | $N(0,1)$ | $N(0,1)$ | $N(0,1) + 1$ | $N(0,1)$ | $N(0,1) + 1$ | $N(0,1) + 1$ |
| A2 | $U(-2,2)$ | $U(-2,2)$ | $U(-2,2) + 1$ | $U(-2,2)$ | $U(-2,2) + 2$ | $U(-2,2) + 1$ |
| A3 | $t_4$ | $t_4 - 1$ | $t_4 + 1$ | $t_4$ | $t_4 + 1$ | $t_4 + 1$ |
| A4 | $e^1 \cdot X_i^1$ | $X_i^1$ | $e^3 \cdot X_i^1$ | $e^1 \cdot X_i^1$ | $e^1 \cdot X_i^1$ | $e^1 \cdot X_i^1$ |
| A5 | $e^1 \cdot \text{Exp}(1)$ | $\text{Exp}(1)$ | $e^3 \cdot \text{Exp}(1)$ | $e^1 \cdot \text{Exp}(1)$ | $e^3 \cdot \text{Exp}(1)$ | $e^3 \cdot \text{Exp}(1)$ |
| A6 | $e^3 \cdot X_i^3$ | $X_i^3$ | $e^3 \cdot X_i^3$ | $X_i^3$ | $X_i^3$ | $X_i^3$ |
| A7 | $0.5N(-5,1) + 0.5N(5,1)$ | $0.5N(-5,1) + 0.5N(5,1) - 1$ | $0.5N(-5,1) + 0.5N(5,1) + 1$ | $0.5N(-5,1) + 0.5N(5,1) + 2$ | $0.5N(-5,1) + 0.5N(5,1) + 1$ | $0.5N(-5,1) + 0.5N(5,1) + 1$ |
| **Alternative cases** | | | | | | |
| A1 | $N(0,1)$ | $N(0,1)$ | $N(0,1)$ | $0.5N(-5,1) + 0.5N(5,1) + 1 - 1$ | $N(0,1)$ | $N(0,1)$ |
| A2 | $U(-2,2)$ | $U(-2,2)$ | $U(-2,2)$ | $1 + 1.1$ | $U(-2,2)$ | $U(-2,2)$ |
| A3 | $t_4$ | $t_4 - 1$ | $t_4 + 1$ | $t_4 + 1$ | $t_4 + 1$ | $t_4 + 1$ |
| A4 | $e^1 \cdot X_i^1$ | $X_i^1$ | $e^3 \cdot X_i^1$ | $e^1 \cdot X_i^1$ | $e^3 \cdot X_i^1$ | $e^3 \cdot X_i^1$ |
| A5 | $e^1 \cdot \text{Exp}(1)$ | $\text{Exp}(1)$ | $e^3 \cdot \text{Exp}(1)$ | $e^1 \cdot \text{Exp}(1)$ | $e^3 \cdot \text{Exp}(1)$ | $e^3 \cdot \text{Exp}(1)$ |
| A6 | $e^3 \cdot X_i^3$ | $X_i^3$ | $e^3 \cdot X_i^3$ | $X_i^3$ | $X_i^3$ | $X_i^3$ |
| A7 | $0.5N(-5,1) + 0.5N(5,1)$ | $0.5N(-5,1) + 0.5N(5,1) - 1$ | $0.5N(-5,1) + 0.5N(5,1) + 1$ | $0.5N(-5,1) + 0.5N(5,1) + 2$ | $0.5N(-5,1) + 0.5N(5,1) + 1$ | $0.5N(-5,1) + 0.5N(5,1) + 1$ |

Table 1. Simulation scenarios.
Table 2. Rejection rates of null cases under various settings.

\[ n_1^\omega = n_1, n_2^\omega = n_2, n_3^\omega = n_1 \]

| Test | U test | LRT | U test | LRT | U test | LRT | U test | LRT | U test | LRT |
|------|--------|-----|--------|-----|--------|-----|--------|-----|--------|-----|
| Scenario | | | | | | | | | | |
| A1 | 0.075 | 0.068 | 0.048 | 0.045 | 0.048 | 0.044 | 0.053 | 0.052 | 0.054 | 0.048 | 0.054 |
| A2 | 0.071 | 0.070 | 0.048 | 0.046 | 0.055 | 0.054 | 0.051 | 0.051 | 0.054 | 0.060 | 0.049 |
| A3 | 0.078 | 0.065 | 0.056 | 0.059 | 0.056 | 0.051 | 0.051 | 0.056 | 0.064 | 0.055 | 0.052 |
| A4 | 0.058 | 0.052 | 0.044 | 0.045 | 0.051 | 0.047 | 0.049 | 0.051 | 0.052 | 0.053 | 0.053 |
| A5 | 0.067 | 0.064 | 0.060 | 0.056 | 0.056 | 0.056 | 0.052 | 0.053 | 0.052 | 0.051 | 0.054 |
| A6 | 0.070 | 0.064 | 0.045 | 0.044 | 0.057 | 0.054 | 0.046 | 0.044 | 0.052 | 0.054 | 0.048 |
| A7 | 0.074 | 0.073 | 0.049 | 0.052 | 0.047 | 0.051 | 0.060 | 0.056 | 0.055 | 0.052 | 0.048 |
| B1 | 0.082 | 0.076 | 0.052 | 0.051 | 0.052 | 0.055 | 0.057 | 0.053 | 0.061 | 0.062 |
| B2 | 0.069 | 0.066 | 0.049 | 0.048 | 0.046 | 0.044 | 0.044 | 0.049 | 0.058 | 0.053 | 0.046 |
| B3 | 0.070 | 0.089 | 0.042 | 0.052 | 0.044 | 0.044 | 0.046 | 0.046 | 0.057 | 0.063 | 0.053 |
| B4 | 0.082 | 0.074 | 0.054 | 0.051 | 0.057 | 0.056 | 0.059 | 0.058 | 0.052 | 0.048 | 0.054 |
| B5 | 0.079 | 0.090 | 0.056 | 0.052 | 0.049 | 0.052 | 0.060 | 0.053 | 0.052 | 0.052 | 0.054 |
| B6 | 0.066 | 0.079 | 0.056 | 0.054 | 0.044 | 0.048 | 0.043 | 0.048 | 0.052 | 0.054 | 0.053 |
| C1 | 0.076 | 0.070 | 0.060 | 0.060 | 0.046 | 0.044 | 0.052 | 0.055 | 0.052 | 0.046 | 0.054 |
| C2 | 0.072 | 0.073 | 0.062 | 0.062 | 0.051 | 0.051 | 0.054 | 0.054 | 0.050 | 0.048 | 0.052 |
| C3 | 0.066 | 0.073 | 0.066 | 0.067 | 0.050 | 0.053 | 0.051 | 0.045 | 0.052 | 0.049 | 0.072 |
| C4 | 0.064 | 0.060 | 0.044 | 0.045 | 0.052 | 0.052 | 0.054 | 0.054 | 0.059 | 0.059 | 0.052 |

values of \( \Gamma \). If \( \Gamma \) is too large, then both tests always reject the null hypothesis for almost any sample size. Results are presented for a representative choice of \( \Gamma \) where this does not occur. All simulation scenarios were investigated with six assumptions regarding sample sizes. First, all subgroup sample sizes are the same \((n_1^\omega = n_2^\omega = n_3^\omega = n_1)\) with \( n_1 \) equal to 10, 50, 100 and 500. Second, the sample sizes of treatment and control within each stratum are the same, but sample size varies across strata. We tried \((n_1^\omega, n_2^\omega, n_3^\omega) (\omega \in \{t, w\}) \) equal to \((50, 100, 150)\) and \((150, 100, 50)\). The former corresponds to the case that the strata with higher effect sizes have larger sample sizes. The latter is the opposite case.

4.4. Simulation study results

The rejection rates of both the U test and the LRT assuming \( \alpha = 0.05 \) for all null cases of the different scenarios are provided in Table 2. The first column of Table 2 shows the labels of all scenarios. Each of the remaining columns corresponds to one sample size setting for all scenarios. As expected all rejection rates are close to the 0.05 level. Since empirical type I errors were computed by generating data for \( L = 2000 \) times, the standard error for each is approximately 0.005. The table shows that when \( n_1^\omega = 10 \) \((s \in \{1, 2, 3\}; \omega \in \{t, w\})\), the type I errors of both tests are a bit too high in all scenarios except for \( A_4 \) and \( C_4 \). For scenario \( C_3 \), when \((n_1^\omega, n_2^\omega, n_3^\omega) \) is equal to \((50, 50, 50)\) or \((150, 100, 50)\), type I errors of both tests are a bit too high. In all other settings, the type I errors are well controlled for both tests.

We then compare the power of the two tests by comparing their rejection rates for the alternative cases in all sample size settings. Figures 1, 2 and 3 show the results for scenarios in Case A, Case B and Case C separately. Each figure is comprised of two subfigures (a) and (b). Subfigure (a) shows rejection rates for all cases where sample sizes \( n_1^\omega \) \((s \in \{1, 2, 3\})\) are equal. Subfigure (b) focuses on the three cases whose stratum-specific sample sizes can
Figure 1. Rejection rates and their 95% confidence intervals for alternative cases in Case A, where all distributions assumed same up to location/treatment effects. (a) Cases where sample sizes are consistent across strata and (b) Cases where sample sizes can change across strata.

vary. Within each subfigure, there is a set of panes, each of which corresponds to a scenario. Within each pane, the vertical axis indicates the rejection rate and the horizon axis indicates the sample size setting. For each type of test, we plot a point showing the empirical rejection rate and a line showing the corresponding 95% confidence interval. The red ones are for our proposed U test, and the blue ones are for the LRT.
Figure 2. Rejection rates and their 95% confidence intervals of alternative cases in Case B, where distributions in same treatment groups assumed same up to location/treatment effects. (a) Cases where sample sizes are consistent across strata and (b) Cases where sample sizes can change across strata.

For Case A, Figure 1(a) shows that when stratum-specific sample sizes are equal, the powers of both tests increase as \( n \) increases. When the common distribution \( F \) is normal (A1) or \( F \) is symmetric with tails lighter than normal (A2), the power of the LRT is a bit higher than the U test. When \( F \) is symmetric with tails heavier than normal (A3), or \( F \) is skewed (A4–A6) or bimodal (A7), the U test is more powerful than the LRT. Also as \( F \)
Figure 3. Rejection rates and their 95% confidence intervals of alternative cases in Case C, where distributions in same strata assumed same up to location/treatment effects. (a) Cases where sample sizes are consistent across strata and (b) Cases where sample sizes can change across strata.

departs more from the Gaussian distribution, the advantage of the U test over the LRT is more substantial. Figure 1(b) shows the cases when subgroup sample sizes average 100 but vary across the strata. Compared to the cases with equal sample sizes, the power of both the U test and the LRT drop, and the U test power drops a bit more than the LRT.
Figure 2 shows the rejection rates of the U test and the LRT for Case B. The results of cases with equal sample sizes across strata are in Figure 2(a). As the sample size increases, the powers of both tests increase in all scenarios. When the distributions in both treatment groups are close to normal (B1), the LRT is more powerful, otherwise the U test is more powerful. When one of the distributions is very far from normal (B3, B5 and B6), the advantage of the U test over the LRT is large. Next we compare the cases with average subgroup sample sizes all equal to 100 but where sample sizes can vary across strata (Figure 2(b)). As with Case A, the results with different stratum-specific sample sizes indicate less power than the case with equal stratum-specific sample sizes for both the U test and the LRT.

The empirical power of the two tests for Case C are displayed in Figure 3. When the stratum-specific sample sizes are equal (Figure 3(a)), the power of both tests increase as sample size increases. In all scenarios, the U test outperforms the LRT, and when there is a bimodal distribution (C2, C3, C4), the advantage of the U test is substantial. Figure 3(b) shows the effect of varying sample sizes across strata. When the distributions of the data in the three strata have similar variances and none of them are too far from normal (C1), the comparison result is similar to that in Case A and Case B. The setting with consistent stratum-specific sample sizes has largest power for both the U test and the LRT, and the difference of powers between balanced sample size setting and unbalanced sample size setting is larger for the U test than the LRT. When one of the strata follows a distribution that is mixture normal (0.5N(−5, 1) + 0.5N(5, 1)) which has a lot larger variance than the other two distributions and also departs more from the normal, the performances of the two tests are very different. For the U test, if the stratum with large-variance distribution has the smallest sample size, it is least powerful. For the LRT, when the stratum with distribution very far from normal has the largest sample size, it is least powerful.

4.5. Investigating the power of the tests

The results in Section 4.4 focus on only a single non-null example for each scenario. This section investigate the power as a function of sample size for different treatment effects. We use the scenarios A3 and A4, and carried out simulations as described in the previous section.

For scenario A3, we generated three strata with \( n_3^\omega = n \ (s \in \{1, 2, 3\}, \omega \in \{t, c\}) \), and generated random samples from \( t_4 \) distribution with location shifts comprised of strata effects \( \Delta_1 = 0, \Delta_2 = 1, \Delta_3 = 2 \) and additive treatment effects \( \tau_1 = 1, \tau_2 = 1 + \Gamma \) and \( \tau_3 = 1 + 2\Gamma \). Here the sequence of the treatment effects \( \{\tau_1, \tau_2, \tau_3\} \) is arithmetic and we treat \( \Gamma \) as the effect size. For each fixed effect size, we explore the relationship between the sample size \( n \) and the rejection rates for both tests, and the results are shown in Figure 4 with \( \Gamma \) ranging from 0 to 0.5, and \( n \) ranging from 10 to 1000. In alternative cases when \( \Gamma > 0 \), with each fixed \( \Gamma \), as \( n \) increases, the rejection rates of both tests increase, and the power of the U test is always higher than the LRT for each \( n \).

For scenario A4, again we generated three strata with \( n_3^\omega = n \ (s \in \{1, 2, 3\}, \omega \in \{t, c\}) \). With \( \frac{Y_{\omega}^s}{\sqrt{c_\omega}} \sim \chi_1^2 \), we took the logarithm of \( Y_{\omega}^s \) as the outcome. So the treatment effect is defined as \( \tau_s = \log(c_\omega) - \log(c_\omega^s) \). We focus on the case where the sequence of the treatment effects \( \{\tau_1, \tau_2, \tau_3\} \) are arithmetic with \( \tau_2 = \tau_1 + \Gamma \) and \( \tau_3 = \tau_1 + 2\Gamma \), and \( \Gamma \) is the effect size.
Figure 4. Relationship between rejection rates and sample sizes for t4 distribution.

size. By fixing $\log(c_1^e) = 0$, $\log(c_2^e) = 1$, $\log(c_3^e) = 2$ and $\log(c_1^t) = 1$, we have $\tau_1 = 1$. Then we can get different values of $\Gamma$ by changing the values of $c_2^t$ and $c_3^t$. Then for each $\Gamma$, we can explore the relationship between the rejection rates of our U test and the LRT as the sample size $n$ varies. Figure 5 shows the rejection rates with $n$ ranging from 10 to 1000 when $\Gamma$ ranging from 0 to 1. As we would expect, with fixed $n$, larger effect size leads to larger rejection rates. The rejection rates of the U test is always larger than the LRT for each fixed $n$ and $\Gamma$.

5. Case study

In this section, we apply our proposed U test to a randomised data set from a programme evaluation study in labour economics, an evaluation of the National Supported Work (NSW) Demonstration. The NSW is a labour training programme conducted in the mid-1970s aiming at providing work experience to people with economic difficulties. Please refer to LaLonde (1986) and Dehejia and Wahba (1999) for details about the study and the data. We use a subset of the LaLonde (1986) data that was created and used by Dehejia and Wahba (1999). The data are available at https://users.nber.org/rdehejia/data/.nswdata2.html. These data described results from male participants with earnings information available for 1974. Earnings in 1978 were treated as the outcome, and several pretreatment variables were recorded. There are 185 subjects in treatment group and 260 subjects in control group.
In this randomised study, pretreatment variables should have the same distribution between the treatment and control groups. So we can directly compare the distributions of the outcome, 1978 earnings, for the treatment and control groups to get the treatment effect. The outcome distributions of the treatment and control groups are shown in Figure 6. Both of them are heavily right-skewed and have an excess of 0 values. Because the distributions are far from a normal distribution, a nonparametric test is more appropriate than parametric test assuming normality. The $p$-value of Mann-Whitney test is 0.01, and the test statistic is 0.43. Here the expectation of the test statistic is the probability that a random outcome in the treatment group is smaller than a random outcome in the control group. The result indicates that there is a positive treatment effect.

Next we construct strata based on two important pretreatment variables, age and 1974 earnings, separately, and then apply our proposed U test to identify whether there is treatment effect heterogeneity across the strata. We first split all subjects by quartiles of age. The subgroup sample sizes are in Table 3, and the pairwise U-statistics are in Table 4. Here the expectation of $U^{(p,q)}$ is the probability that the difference between treatment and control outcomes in stratum $p$ are smaller than the difference in stratum $q$. As the $U$ values are greater then 0.5, the treatment effects in younger strata are generally smaller than those in older strata. However, the $p$-value of our proposed heterogeneity test is 0.58, so the observed heterogeneity is not statistically significant.

Then we explore whether the treatment effect differs between participants with and without positive incomes in 1974. The first stratum is for participants without income in
Figure 6. Distribution of 1978 earnings in the treatment and control groups.

1974 and the second stratum is for those with positive income. The subgroup sample sizes are shown in Table 5. The U-statistic comparing their treatment effects $U_{1,2}^{(1,2)}$ is 0.409, and the $p$-value of our heterogeneity test is 0.032, which indicates this programme has greater
Table 3. Sample sizes in different treatment and age groups.

| Stratum | 1 [17, 20] | 2 (20, 24) | 3 (24, 28) | 4 (28, 55) |
|---------|-------------|------------|------------|------------|
| Age     | Treatment   | Control    | Treatment  | Control    |
|         | 47          | 41         | 49         | 48         |
|         | 83          | 56         | 60         | 61         |

Table 4. Pairwise U-statistics comparing treatment effects between age groups.

|     | \(U^{(1,2)}\) | \(U^{(1,3)}\) | \(U^{(1,4)}\) | \(U^{(2,3)}\) | \(U^{(2,4)}\) | \(U^{(3,4)}\) |
|-----|----------------|----------------|----------------|----------------|----------------|----------------|
|     | 0.52           | 0.55           | 0.57           | 0.53           | 0.55           | 0.51           |

Table 5. Sample sizes of different treatment and income groups.

| Stratum | 1974 Income | 1974 Income | 2 1974 Income |
|---------|-------------|-------------|--------------|
|         | Treatment   | Control     | Treatment    |
| 1974 Income = 0 | 131         | 54          | 65          |
| 1974 Income > 0 | 195         |             |             |

impact for participants who did not have any income in 1974 than those who had some income.

6. Discussion

Identifying the existence of treatment effect heterogeneity is a key element of attempts to provide more precise treatment recommendations for individuals. We have described a U-statistic-based approach to formally test the hypothesis of homogeneous treatment effects without assuming a particular parametric form of the outcome distributions, and compared its performance with the LRT. The LRT requires the distribution of treatment effect estimates to be normal, which can be satisfied if the outcomes are normal or the sample sizes are large (by the Central Limit Theorem). Our results show that, as expected, when the outcome distributions are close to normal, the power of the LRT is a little better than the U test. However when at least one of the outcome distributions departs substantially from the normal distribution, the power of our nonparametric test can be significantly larger than the LRT. As the departure increases, the advantage of the U test increases. This observation is similar to the comparison between Mann-Whitney test and t-test (Lehmann 1999).

Though our approach can be applied to any situations as long as the two assumptions addressed in the second paragraph of Section 3 are satisfied, we do need to be cautious about understanding what aspect of the distribution is being compared by the U test. The U test is most easily interpreted when we believe that one of the scenarios described in Section 3.3 applies. For those scenarios (Case A, B and C), the test is comparing locations of distributions assumed to come from common families.

One limitation of this method is that it requires the distributions of all confounding variables are the same between the treatment group and the control group within stratum, which is true in randomised experiments. However, in observational studies, we will need to adjust for confounding variables. If the confounding variables are all discrete, then one approach is to apply the test within strata defined by the confounders. But that will not generally be the case. If there are continuous confounders, then one possibility is to
apply our approach on strata created based on estimated propensity scores as suggested by Xie et al. (2012). There is no guarantee that such an approach will work as some further adjustments for remaining covariate imbalance may be needed.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**References**

Allison, P.D. (1977), ‘Testing for Interaction in Multiple Regression’, *American Journal of Sociology*, 83(1), 144–153.

Bitler, M.P., Gelbach, J.B., and Hoynes, H.W. (2006), ‘What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments’, *American Economic Review*, 96(4), 988–1012.

Byar, D.P. (1985), ‘Assessing Apparent Treatment-Covariate Interactions in Randomized Clinical Trials’, *Statistics in Medicine*, 4(3), 255–263.

Chan, I.S., and Ginsburg, G.S. (2011), ‘Personalized Medicine: Progress and Promise’, *Annual Review of Genomics and Human Genetics*, 12, 217–244.

Chang, M., Lee, S., and Whang, Y.-J. (2015), ‘Nonparametric Tests of Conditional Treatment Effects with an Application to Single-sex Schooling on Academic Achievements’, *The Econometrics Journal*, 18(3), 307–346.

Crump, R.K., Hotz, V.J., Imbens, G.W., and Mitnik, O.A. (2008), ‘Nonparametric Tests for Treatment Effect Heterogeneity’, *The Review of Economics and Statistics*, 90(3), 389–405.

De Neve, J., and Thas, O. (2017), ‘A Mann–Whitney Type Effect Measure of Interaction for Factorial Designs’, *Communications in Statistics-Theory and Methods*, 46(22), 11243–11260.

Dehejia, R.H., and Wahba, S. (1999), ‘Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs’, *Journal of the American Statistical Association*, 94(448), 1053–1062.

Delgado, M.A., and Escanciano, J.C. (2013), ‘Conditional Stochastic Dominance Testing’, *Journal of Business & Economic Statistics*, 31(1), 16–28.

Ding, P., Feller, A., and Miratrix, L. (2016), ‘Randomization Inference for Treatment Effect Variation’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3), 655–671.

Dixon, D.O., and Simon, R. (1991), ‘Bayesian Subset Analysis’, *Biometrics*, 47(3), 871–881.

Feller, A., and Holmes, C.C. (2009), ‘Beyond Toptlines: Heterogeneous Treatment Effects in Randomized Experiments’, Unpublished manuscript, Oxford University.

Gail, M., and Simon, R. (1985), ‘Testing for Qualitative Interactions Between Treatment Effects and Patient Subsets’, *Biometrics*, 41(2), 361–372.

Ginsburg, G.S., and Willard, H.F. (2009), ‘Genomic and Personalized Medicine: Foundations and Applications’, *Translational Research*, 154(6), 277–287.

Hájek, J. (1968), ‘Asymptotic Normality of Simple Linear Rank Statistics Under Alternatives’, *The Annals of Mathematical Statistics*, 39(2), 325–346.

Hodges, J.L., and Lehmann, E.L. (1956), ‘The Efficiency of Some Nonparametric Competitors of the t-Test’, *The Annals of Mathematical Statistics*, 27(2), 324–335.

Hsu, Y.-C. (2017), ‘Consistent Tests for Conditional Treatment Effects’, *The Econometrics Journal*, 20(1), 1–22.

Imbens, G.W., Newey, W.K., and Ridder, G. (2006), ‘Mean-Square-Error calculations for average treatment effects’, IRP discussion paper.

Jain, K.K. (2015). *Textbook of Personalized Medicine* (2nd ed.), New York, NY: Humana Press.

Korolyuk, V.S., and Borovskich, Y.V. (2013), *Theory of U-Statistics* (Vol. 273), Berlin: Springer Science & Business Media.

LaLonde, R.J. (1986), ‘Evaluating the Econometric Evaluations of Training Programs with Experimental Data’, *The American Economic Review*, 76(4), 604–620.
Lehmann, E.L. (1963), ‘Robust Estimation in Analysis of Variance’, *The Annals of Mathematical Statistics*, 34(3), 957–966.
Lehmann, E.L. (1999), *Elements of Large-Sample Theory*, New York: Springer.
Lehmann, E.L., and D’Abrera, H.J. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco, CA: Holden-Day.
Mann, H.B., and Whitney, D.R. (1947), ‘On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other’, *The Annals of Mathematical Statistics*, 18(1), 50–60.
Pate, A.M., and Hamilton, E.E. (1992), ‘Formal and Informal Deterrents to Domestic Violence: The Dade County Spouse Assault Experiment’, *American Sociological Review*, 57(5), 691–697.
Patel, K.M., and Hoel, D.G. (1973), ‘A Nonparametric Test for Interaction in Factorial Experiments’, *Journal of the American Statistical Association*, 68(343), 615–620.
Sant’Anna, P.H. (2021), ‘Nonparametric Tests for Treatment Effect Heterogeneity with Duration Outcomes’, *Journal of Business & Economic Statistics*, 39(3), 816–832.
Van der Vaart, A.W. (2000), *Asymptotic Statistics* (Vol. 3), Cambridge University Press.
Xie, Y., Brand, J.E., and Jann, B. (2012), ‘Estimating Heterogeneous Treatment Effects with Observational Data’, *Sociological Methodology*, 42(1), 314–347.
Zimmerman, D.W. (1998), ‘Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions’, *The Journal of Experimental Education*, 67(1), 55–68.