RANDOMIZATION TEST

Junshi Wang, Supervised by Qingyuan Zhao

香 港 大 學

THE UNIVERSITY OF HONG KONG

The University of Hong Kong
Department of Statistics and Actuarial Science

August 24, 2022

Oversea Research Fellowship at University of Cambridge, Technical Report

Contents

1 Introduction

In this report, we discuss how randomization test, which is closely linked to causal relation, can be formulated with the potential outcomes framework developed by Neyman (Holland, 1986). If binary outcome assumption is made, the whole population could be stratified into four groups and (partially) sharp null hypothesis can be established. Based on such notion, randomization tests motivated by Fisher's exact test are proposed. Simulation is carried out to examine how these tests performs against each alternative hypothesis. A surprising conjecture that the test constructed for sharp null is actually valid for Neyman's null is made. The asymptotic case is proven to be true by finite population Central Limit Theorem. In order to understand what finite population CLT means, we finally review the related papers and discuss how it is validated and what limitations it has.

In Section 2, we clarify the idea of randomization versus permutation test and conditional versus unconditional test. After differentiate intervention from observation, we formulate randomization test for (partially) sharp null hypothesis based on fisher's exact test.

In Section 3, we present the power function which is generated through simulations in 3D graphs. Surprisingly, power of test against sharp null for alternatives which fulfill partially sharp null is smaller than alpha. This implies that this test maybe valid for testing partially sharp null as well and naturally lead to our conjecture in the next section.

In Section 4, conjecture is formalized, and several attempts including asymptotic methods and sensitivity analysis, etc. Finally, in Section 5, we review the theorems used to proved the asymptotic normality of standardized test statistics in finite population scenario(and explain what it means to be finite population CLT).

2 Methodology

2.1 Randomization test

Randomization, as argued by many, is a crucial part of statistical reasoning (Fisher, 1936) and is indispensable if causal relation is to be identified. Randomization test, as its name suggests, requires a physical act of randomization. To be more specific, with a finite fixed population, the treatment assignment is random, according to some distribution. A very similar concept is permutation test, which have almost the same computational procedure as the former. However, these two are different conceptually. While the randomness of randomization test comes from randomized assignment, that of permutation test roots in random sampling from a population (Lehmann et al., 2005). Also, the latter does not require the physical act of randomization, and serves more or less as an algorithm. In fact, Fisher's exact test is the simplest randomization test.

2.2 Conditional and Unconditional Test

Before the formulation, it is beneficial to introduce conditional and unconditional test. Conditional test means that the rejection region depends on some random variable that we condition on, while unconditional test has a fixed rejection region. Fisher's exact test, out motivating example is a conditional test. Although it is an exact test, but due to the discreteness of distribution, it is usually more conservative (Little, 1989). As a matter of fact, conditioning also causes conservativeness in comparison to Barnard's test, an unconditional test. However, whether it is too conservative depends on whether conditioning on the margins can be justified. It can be argued that if the margins are ancillary and do not contain information about the parameter we concern about, then conditional test is advisable.

2.3 Notations and Contingency Table

In this section we establish the notations and show the tables that will facilitate comprehension. For the 2 by 2 contingency table that is observed, we denote the figure by N_{ij} and for the number of subjected assigned to treatment and control group in each strat, M_{i1} and M_{i0} are used. $M_{.0}$ is the total number of treatment assigned and $M_{..}$ is the sample size.

	$Y=0$	$Y=1$	
$A=0$	N_{00}	N_{01}	$M_{.0}$
$A=1$	N_{10}	N_{11}	$M_{.1}$
	r	s	$M_{..}$

Table 1: Contingency table with counterfactuals

Under sharp null, $M_{30} = M_{20} = M_{21} = M_{31} = 0$. Considering Fisher's exact test which rejects the null H_F when N_{00} is extreme(too large or small), We can build a randomization test based on the Fisher's exact test by conditioning on $M_{1.}, M_{2.}, M_{3.}, M_{4.}, M_{.0}$ and $M_{.1}$.

$Y(0)$	$Y(1)$	Total	$A=0$	$A=1$
1	1	$M_{1.}$	M_{10}	M_{11}
1	0	$M_{2.}$	M_{20}	M_{21}
0	1	$M_{3.}$	M_{30}	M_{31}
0	0	$M_{4.}$	M_{40}	M_{41}
		$M_{..}$	$M_{.0}$	$M_{.1}$

Table 2: Contingency table with counterfactuals

$$
\begin{aligned}
N_{00} &= M_{30} + M_{40} \\
N_{01} &= M_{10} + M_{20} \\
N_{10} &= M_{21} + M_{41} \\
N_{11} &= M_{11} + M_{31}
\end{aligned}
$$

1: Expressed in counterfactuals

The random variables $M_{10}, M_{20}, M_{30}, M_{40}$ follows an multivariate hypergeometric distribution, which can be calculated by computer. Under null, this degenerates to a univariate hypergeometric distribution where $M_{20} = M_{30} = M_{2.} = M_{3.} = 0$.

In particular, the density can be wrote as below,

$$
\begin{aligned}
f(x_1, x_2, x_3, x_4) &= \mathbf{1}\{\sum_{i=1}^{4} x_i = M_{.0}\} \binom{M_{1.}}{x_1} \binom{M_{2.}}{x_2} \binom{M_{3.}}{x_3} \binom{M_{4.}}{x_4} \prod_{i=1}^{4} \mathbf{1}\{0 \le x_i \le M_{i0}\} \\
\mathbf{P}(M_{10} = x_1, M_{20} = x_2, M_{30} = x_3, M_{40} = x_4) &= \frac{f(x_1, x_2, x_3, x_4)}{\sum_{x_1', x_2', x_3', x_4'} f(x_1', x_2', x_3', x_4')}
\end{aligned}
$$

2: Probability Density of Constrained Multinomial

This probability distribution can be obtained through Bayesian computation, e.g. direct sampling, MCMC.

2.4 Formulation of Sharp Null Hypothesis

With the background knowledge, we proceed to give a formal definition of our proposed test. First, classify the population into four strata(doomed, helped, hurt, immune) by their counterfactual schedules(restricted to simple discrete case $Z \in \{0,1\}$ and $A \in \{0,1\}$).

Instead of going into the calculations directly, it may be worthwhile discussing different abstract models that can occur (Barnard, 1947).

Type	Y(0)	Y(1)	Proportion
Doomed	1	1	p_1
Helped	1	0	p_2
Hurt	0	1	p_3
Immune	0	0	p_4

Table 3: Four Strata

1. All margins are fixed: Lay $N = m+n$ white and black balls on the table together, then write 0/1 on r/s of the balls.

2. Row margins are fixed: Lay $N = m+n$ white and black balls on the table separately, then write 0/1 on each ball randomly.

3. Total is fixed: Pick N balls from a urn containing four types of balls.

4. Nothing is fixed: Observe the occurrence of events during a period of time.

Furthermore, there may also be different types of hypothesis. The following two hypothesis are of great interest. Suppose we have a sample of size n.

1. $H_N : \bar{Y}(0) - \bar{Y}(1) = 0$ is a partially sharp null hypothesis.

2. $H_F : Y_i(0) - Y_i(1) = 0, \forall i$ is sharp null. The hypothesis is very strong and even considered to be uninformative by some.

We can also differentiate interventions from observations.

1. *Intervention with Random subjects*: Subjects are considered to be randomly sampled from the population. Then we randomly assign subjects to treatment, in order to eliminate confounding effect.

2. *Intervention without Random subjects*: Subjects are considered to be constant. Then we randomly assign subjects to treatment, in order to eliminate confounding effect.

3. *Observation*: without random assignment, prone to confounders, need adjustments.

With so many different scenarios given, only some of them are of interest in terms of randomization test. For instance, randomization test concerns *Intervention without Random subjects* and the "Row margins are fixed"(number of treatments assigned are fixed).

In order to do randomization test, we must be able to impute the potential outcome schedule $W = (W_i), W_i = (Y_i(0), Y_i(1))^T$ from the observation $Y = (Y_i)$. H_F obviously fulfills such requirement. Given that we observe

$$N = \begin{pmatrix} N_{00} & N_{01} \\ N_{10} & N_{11} \end{pmatrix}$$

and assume H_F, then the imputed counterfactual schedule is $M_{1.}* = N_{01} + N_{11}, M_{2.}* = M_{3.}* = 0, M_{4.}* = N_{00} + N_{10}, M_{.0}* = M_{.0} = N_{00} + N_{01}$ and $M_{10}* = N_{01}, M_{20}* = M_{30}* = 0, M_{40}* = N_{00}$. But so that a randomization test to be computable for H_N, we need to impose stronger assumptions such as monotonicity (Chiba, 2015) or form the p-value as the supremum of p-value of sharp null hypotheses (Caughey et al., 2021).

As briefly mentioned in the last section, we shall start with defining the case for H_F and *Intervention with Random subjects*. Under the assumption that individual effect is zero, we are able to impute the full potential outcome schedule $W = (Y_i(0), Y_i(1)) = (Y_i, Y_i)$. Suppose we observe the 2 by 2 table $N_{ij}, i, j \in \{0,1\}$ and define a random vector Z as the treatment assignment. If the i^{th} subject is assigned to treatment group, then $Z_i = 1$. Analagous to Fisher's exact test, the test statistic is N_{00}, let $N*_{00} = (1-Z)^T(1-Y)$. Suppose we stick to a two-sided test, hypothesis H_F will be rejected if,

$$P(N*_{00} \leq \min(N_{00}, 2 \cdot \text{mean} - N_{00}) \text{ or } N*_{00} \geq \min(N_{00}, 2 \cdot \text{mean} - N_{00}))$$

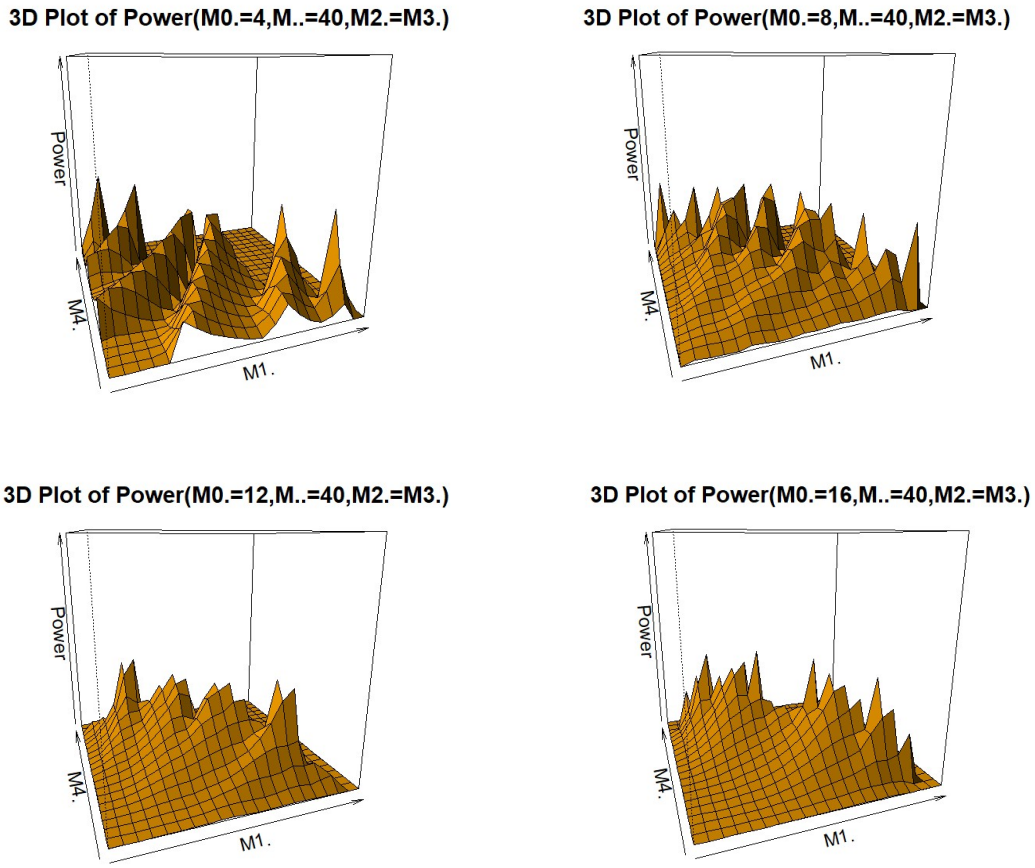
where $\text{mean} = [\max(0, N_{00} - N_{11}) + \min(N_{00} + N_{10}, N_{00} + N_{01})]/2$.

2.5 Beyond Sharp Null

Fisher's sharp null, though providing the nice result of exact Type I error, is rather restrictive and not informative enough. One may be interested in testing the partial sharp null hypothesis that $M_{2.} = M_{3.}$ or sample average mean effect is zero. It is possible to start with $H'_{0m} : M_{2.} = M_{3.} = m$ where m is some constant, resulting in a family of p-values p_m. Then extend the test by adopting the form $\sup_{0 \leq m \leq M_{..}/2} \{p_m : M_{20} = m_{20}, M_{30} = m_{30}, M_{21} = m_{21}, M_{31} = m_{31}, m_{20} + m_{21} = m_{30} + m_{31}\}$ (Caughey et al., 2021). There are alternatives for testing H_N with randomization tests. For example, if we impose the monotonicity assumption where $M_{3.} = 0$ then H_N and H_F are actually equivalent (Chiba, 2015). Chiba pointed out unconditional randomization test could also be used for testing H_N if exchangeability is assumed.

3 Simulation Study

There are two ways to simulate the test and calculate(approximately) its power. The first one involves Noniterative Monte Carlo methods. We generate the data which follows roughly the target distribution $f(x)$ through rejection sampling. This is preferred towards Markov Chain Monte Carlo chain methods (e.g., Metropolitan-Hasting) because the latter methods converge much slower and do not avail in the case of a rather simple discrete distribution (Carlin and Louis, 2008). Since sampling is used, we can not expect the power to be exact. However, it will tend to the true power function as $N \to \infty$. The second method is more straight-forward and less computing power(at least when N is small) are required. It expands the mass function by conditioning on x_2 and x_3. This allows us to calculate the exact power and utilize the existing Fisher.test easily(Note that Fisher.test does not provide continuity adjustment and neither do we trouble ourselves with such fine details).Graphs below depict how power function of the proposed test changes with $M_{0.}, M_{4.}$ when $M_{2.} = M_{3.}$.



We observe that under all possible realization of $(M_{1.}, M_{2.}, M_{3.}, M_{4.})$ with $(M_{.0}, M_{..})$ given under the constraint $M_2 = M_3$, the power function is smaller than alpha. This means the proposed test for Fisher's null is actually valid for testing Neymanian null. Such impression leads us to the following conjecture.

4 Conjecture

After simulation study, we found that the test in Section 2.4 always has power smaller than α when the $M_{2.} = M_{3.}$. This lead to the speculation that this test is also valid for $M_{2.} = M_{3.}$, not just for $M_{2.} = M_{3.} = 0$. Suppose we have sample size of $M_{..}$, the first $M_{1.}$ of them are type 1, and the subsequent $M_{i.}$ subjects are type i. The treatment assignment is $\mathbf{Z} = (\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{Z_3}, \mathbf{Z_4})^T$, where $\mathbf{Z_i}$ is for type i subjects. Subsequently we have $X = (X_1, X_2, X_3, X_4), X_i = \mathbf{1}^T Z_i$, which is the number of people assigned to control group for each type, following distribution $\mathbf{P}_{(M_{1.}, M_{2.}, M_{3.}, M_{4.}, M_{.0}, M)}$.

And the observed data is denoted as

$$N = \begin{pmatrix} N_{00} & N_{01} \\ N_{10} & N_{11} \end{pmatrix} = \begin{pmatrix} X_3 + X_4 & X_1 + X_2 \\ M_{2.} + M_{4.} - X_2 - X_4 & M_{1.} + M_{3.} - X_1 - X_3 \end{pmatrix} \tag{1}$$

To facilitate the calculation of p-value, we define the following distribution for every possible value of X.

$$\tilde{\mathbf{P}}_{\mathbf{X}} = \mathbf{P}_{(N_{01}+N_{11}, 0, 0, N_{00}+N_{10}, N_{00}+N_{01}, M)}, N_{01}(\tilde{X}) \sim Hypergeometric(M_{..}, M_{1.} + M_{3.} + X_2 - X_3, M_{.0}), \tag{2}$$

where $N_{00} = N_{00}(X)$.

With $\tilde{X} \sim \tilde{\mathbf{P}}_{\mathbf{X}}$, we can calculate p-value as follows,

$$PV(X) = \tilde{\mathbf{P}}_{\mathbf{X}}\{T(\tilde{X}) \leq T(X))\} = \tilde{\mathbf{P}}_{\mathbf{X}}\{N_{01}(\tilde{X}) \text{ at least as extreme as } N_{01}(X))\}$$

In order to show that the test is indeed valid, only need to prove $\mathbf{P}\{PV(X) \leq \alpha\} \leq \alpha$. Considering that $\tilde{P}_X = \tilde{P}_{X_2, X_3}$ is determined entirely by X_2 and X_3, we condition on X_2, X_3.

$$\mathbf{P}\{PV(X) \leq \alpha\} = \sum_{x_2, x_3 \in Support} \mathbf{P}\{PV(X) \leq \alpha | X_2 = x_2, X_3 = x_3\} \mathbf{P}\{X_2 = x_2, X_3 = x_3\} \tag{3}$$

If we are able to show that $\mathbf{P}\{PV(X) \leq \alpha | X_2 = x_2, X_3 = x_3\} \leq \alpha$, where $(X_1, X_4) | (X_2 = x_2, X_3 = x_3) \sim Hypergeometric(M_{1.} + M_{4.}, M_{1.}, M_{.0} - x_2 - x_3), \forall x_2, x_3$, the proof is complete. However it is unlikely to be true. One counterexample is letting $M = (0, 6, 6, 12), M_{.0} = 12, (x_2, x_3) = (0, 0), (0, 1), (5, 6), (6, 6)$.

Indeed, even if this is true, it would be hard to prove, owing to the fact that pascal's triangle sum does not have a nice closed form(a mathematical expression that uses a finite number of standard operations) (Graham et al., 1989; Spivey, 2011).

4.1 Attempts

Listed below are a few attempts made to prove the conjecture.

4.1.1 Asymptotic distribution

It is possible to approximate the cumulative distribution function of hypergeometric distribution by normal distribution. Since we now only have a small sample size and $X \sim$

$\mathbf{P}(M_{1.}, M_{2.}, M_{3.}, M_{4.}, M_{.0})$. Then it is natural to think of a sequence of $\{X_i\}$ such that $X_i \sim \mathbf{P}(M_{1.}, M_{2.}, M_{3.}, M_{4.}, M_{.0}, M_{..})$,where the variables tends to infinity in some fashion, with our test denoted as ϕ_i. In fact, under our setting, we are able to bypass the complicated formula 3 by proving Fisher's exact test is asymptotically equivalent to Neyman's test(see Section 4.3.1 for more information).

4.1.2 Neyman Pearson

The parameter space $\Omega = \{(M_{1.}, M_{2.}, M_{3.}, M_{4.}, M_{.0}, M_{..}) \in \mathbf{N}^6 : M_{..} \geq M_{.0} > 0, \sum_{i=1}^{4} M_{i.} = M_{..}\}$. For each $X \sim \mathbf{P}_M, M \in \Omega$, with support $R(X) = \{x \in \mathbf{N}^4 : \sum_{i=1}^{4} x_i = M_{.0}, x_i \leq M_{i.}, \forall i\}$, our test can be used to test against $H_0 : M \in \Omega_0 = \{M \in \Omega : M_{2.} = M_{3.} = 0\}$. We want to show that it is also a valid test for $H_1 : M \in \Omega_1 = \{M \in \Omega : M_{2.} = M_{3.}\}$. Considering the fact that fisher's conditional test for independence in two binomial setting is valid and is similar to testing the average treatment effect, namely $M_{2.} = M_{3.}$, there might be some food for thought. However, there's no obvious connection, owing to the fact that mass function 2 is not in the exponential family (and does not satisfy the regularity assumption, its support is changing).

4.1.3 Upper bound

It is Observed in the experiments that for most of the time, $\mathbf{P}\{PV(X) \leq \alpha\}$ is much smaller than α. It might be possible to find an upper bound in some $\Omega_2 \subset \Omega_1$. But simplifying the hypergeometric functions is hard. This lead us to sensitivity analysis.

4.2 Relationship with Sensitivity Analysis

Sensitivity analysis should be executed when we are not sure whether all the assignments are equally likely, or whether some hidden bias that affect assignments exist. Formally, for subject j and k that has the same(or close) covariate x, their propensity score $\pi_i = P(Z_i = 1)$ may be different. To be more specific, $\frac{1}{r} \leq \frac{\pi_j(1-\pi_k)}{\pi_k(1-\pi_j)} \leq r$. Based on the concept of arrangement increasing, distributive lattice, isotonic function, **The FKG Inequality** and **Holley's Inequality**(Observational study, Page....), Rosenbaum found the lower and upper bound of p-value(this is unknown since the propensity score is not uniform and unspecified) of randomization test. Such randomization test are usually based on sign statistics that are arrangement-increasing for **sharp null hypothesis**.

From first sight, our problem has many similarities to that encountered in sensitivity analysis. First of all, we may formulate the original problem in the proceeding way: $M_2 = M_3 = k, k/M \leq u$. When $u = 0$, it is Fisher's null; $u = 1$, it is Neymanian null. Secondly, in order to show that the conjecture is true, we need to find an upper bound for the power function(ideally, the maximum should be attained at $M_2 = M_3 = 0$). However, discrepancies in the settings leads to complications. The potential outcome schedule is not fixed in our case and every treatment assignment is equally likely, as opposed to the fixed potential outcome schedule and varied propensity in sensitivity analysis (Rosenbaum et al., 2010). This leads to the futility of the brilliant theorem on distributive lattice and we are forced

to tackle the technical problem directly, which is not at all easy. Considering this, relevant resources are sought after and attempts are made to solve the asymptotic case.

4.3 Related Literature Asymptotic cases

When it comes to asymptotic study in randomization test, extra caution should be given to its true meaning and definition. We could imagine a sequence of population with sample size N, $\Phi_N = w_1, w_2, ..., w_i = (Y(0), Y(1))$. For each population, there is a power function $pw()$. To show that the conjecture holds asymptotically, it is to calculate and find the upper bound of $pw(\theta)$, $\theta = (M1, k, k, M4)$, as $N \to \infty$. The proportion of each M_i, as N tends to infinity holds the key to solution. To facilitate thinking, it is beneficial to start with related literature.

4.3.1 A Paradox From Randomization-Based Causal Inference (Ding, 2017)

H_F is a more stronger hypothesis than H_N and one would expect that the rejection of H_N with test ϕ_N will imply the rejection of H_F with ϕ_F. Ding Peng pointed out that this does not necessarily hold. In the paper, this problem was tackled by explicitly writing out the studentized test statistcs and analyzing the denominators(since the nominators are the same), where $t_F = \frac{\hat{\tau}-\tau}{\sqrt{Var_F}}$ and $t_N = \frac{\hat{\tau}-\tau}{\sqrt{Var_N}}$. Both t_F and t_N can be proved to be asymptotically normal under certain conditions by finite population CLT(details in Section 5). An interesting question is asked, how does a test motivated by Fisher's exact test for only sharp null perform on the Neyman's null, which is also our focus. For general non-binary cases, such test may not be valid for Neyman's null. However, further investigations on the binary case is of great importance to us. Results are shown that as sample size tends to infinity, ϕ_F and ϕ_N are equivalent.

Theorem 4.1. *Under Neyman's null $\tau = 0$, Fisher's test and Neyman's test are asymptotically equivalent, thus Fisher's test is valid for testing Neyman's null asymptotically.*

We shall examine further what this means in Section 5.2, namely, what conditions are imposed and whether we can make the same claim in our framework. Ding Peng's reasoning focused entirely on the test statistics and attributes all paradox to the pooled and unpooled method of estimating variance. We shall see next that this train of thought is not complete and the approximating distribution contributes as well.

4.3.2 An Apparant Paradox Explained (Loh et al., 2017)

This paper refutes the conclusion that the difference in power is contributed by test statistic, instead it is the result of approximating distribution we adopt when calculating p-value for these statistics.

Some notations: observation x^o, test statistic function $r(.)$, reference distribution for test statistic $r^o = r(x^o)$ and p-value $pv(r, m, \theta; x^o) = Pr_m(R \geq r^o)$. A test is anticonservative if $Pr_\theta[pv(r, m, \theta; X) \leq \alpha] \geq \alpha$. Let $m_Z(.)$ and $m_F(.)$ be standard normal distribution and randomization distribution.

In balanced design and small samples, the paradox is wholly due to the fact that p-values based on the reference distribution m_Z are anticonservative in finite samples. This can be shown by a experiment where N_F is true(citation).

For asymptotic analysis, we should separate the discussion of balanced and unbalanced design. When experiment is balanced, if $\tau = N^{-1/2}$(Pitman's local alternative), then the probability that two tests disagree converges to zero, else if τ is of order 1, two tests will both have asymptotic power 1. When experiment is unbalanced, it is possible that two tests disagree with limiting positive probability. Moreover, even if N_F is false, Fisher's test may not have asymptotic power 1. It is natural to think whether the test that is valid for N_F has asymptotic power 1 under all nonlocal alternatives. This property should be fulfilled by the test (r^*, m_F) with r^* a member of the class of generalized Kolmogorov–Smirnov [KS] test statistics (Præstgaard, 1995).

4.3.3 Exact tests for the weak causal null hypothesis in randomized trials (Chiba, 2015)

1. As the title suggests, the paper is dedicated to the construction of tests that are exact for weak null hypothesis, in contrast to Neyman's test which is only valid asymptotically.

2. Both conditional(e.g., Fisher's exact test) and unconditional test(e.g., Bernard's test) are considered. The test are valid for weak null when monotonicity or exchangeability is assumed. Those tests are constructed from supremum of p-value, which is a convention already. This idea has been used in many papers (Caughey et al., 2021; Zhang and Zhao, 2022). Their test is very similar to ours in the sense of using potential outcome framework. But the difference is still significant. For example, Several assumptions are made for weak null to be equivalent to sharp null, while our conjecture is that randomization test for sharp null is valid for weak null.

3. One example(Figure 2) is presented for illustration, where p-value at $M_{2\cdot} = M_{3\cdot} = 0$ is not the largest. But it does not contradict our conjecture, since our claim is that power function attains maximum at $M_{2\cdot} = M_{3\cdot} = 0$ among all $M_{2\cdot} = M_{3\cdot}$, not for every realization of p-value.

5 Review the finite population CLT

Central limit theorem, corner stone of statistical asymptotic analysis, is crucial in our previous analysis. Here we discuss what does central limit theorem stands for in terms of randomization and finite population, how it is derived and some caveats. First, we shall start with the fundamental theorem, Lindberg CLT.

5.1 Lindberg CLT

Theorem 5.1. *If $X_1, ..., X_n$ are independent, identically distributed with expectation $E(X_i) = \mu$ and finite variance $Var(X_i) = \sigma^2$, then $\sqrt{n}(\bar{X}_n - \mu)$ is asymptotically normally distributed with distribution $N(0, \sigma^2)$.*

This result is the widely known standard CLT. But the iid condition makes such result unrealistic and stringent. The next result, Lindberg CLT gives a much stronger theorem (Feller, 1971). For the sake of simplicity suppose that the variables $X_1, X_2, \ldots$ are discrete, with X_1 taking on values $a_1, a_2, \ldots$;X_2 taking on values $b_1, b_2, \ldots$, etc.

Theorem 5.2. *Let $X_1, ..., X_n$ be independent, with means $E(X_i) = \mu_i m$ and variances $Var(X_i) = \sigma_i^2$, and let S_n^* be the standardized sum of the X's,*

$$S_n^* = \frac{\sum_{i=1}^n (X_i - \mu_i)}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \tag{4}$$

Then the limit distribution of S_n^* is the standard normal distribution $N(0,1)$ provided for any $t > 0$

$$\frac{\tau_{n1}^2(t) + \cdots + \tau_{nn}^2(t)}{\sigma_1^2 + \cdots + \sigma_n^2} \to 0 \; n \to \infty \tag{5}$$

where $\tau_{n1}^2(t) = \sum \mathbf{P}(X_i = a_j)(a_j - \mu_i)$, $\tau_{ni}^2(t)$ are defined analogously.

5.2 Finite Population CLT

Here we formally introduce the CLT for finite population(or randomization distribution). We no longer constrain ourselves to a single super-population, but an increasing sequence of population Φ_N with sample size N where subjects have value $v_{N1}, \ldots, v_{NN}$. $n = n(N)$ treatments are **randomly** assigned(which give rise to the random vector Z) to Φ_N. Let the n values in this group be denoted as $A_{N1}, \ldots, A_{Nn}$ and sample mean of such treated group is $S_N = \sum_{i=1}^n A_{Ni}$ while population mean is $v_{N\cdot} = \frac{v_{N1} + \cdots + v_{NN}}{N}$.

The following theorem about the asymptotic normality of $S_N^* = [S_n - E(S_N)]/\sqrt{Var(S_N)}$ can be proved based on Lindberg CLT, which is give in Theorem 6, Appendix 4 (Lehmann and D'Abrera, 1975).

Theorem 5.3. *A sufficient condition for the the standardized variables S_N^* to be asymptotically normally distributed according to $N(0,1)$ is that n and $m = N-n$ both tend to infinity and that*

$$\frac{max(v_{Ni} - v_{N.})^2}{\sum(v_{Nj} - vN.)^2} max(\frac{m}{n}, \frac{n}{m}) \to 0 \text{ as } N \to \infty \tag{6}$$

5.3 General Finite Population CLT

The previous result could be extended to muti-level treatments, where subjects are assigned to Q groups randomly (Li and Ding, 2017).

Theorem 5.4. *Let $(\bar{y_{S1}}, \ldots, \bar{y_{SQ}})$ be the Q sample averages of a random partition of sizes $(n_1, ..., n_Q)$ for a finite population, $\Pi_N = y_{N1}, ..., y_{NN}$. As $N \to \infty$, if (i) $cov(t_N)$ where t_N is the standardized sample averages has a limiting value $V \in R^{QQ}$, and (ii)*

$$\frac{1}{min_{1 \le q \le Q} n_q} \frac{m_N}{v_N} \to 0, \tag{7}$$

then $t_N \xrightarrow{d} (0, V)$.

5.4 Application of the theorem

Fisher's exact test can be shown to be equivalent to mean difference test, since the test statistic for mean difference is **monotone** in N_{00} which is the test statistic for Fisher's exact test. Therefore, we can turn to focus on mean difference

$$\begin{aligned} t_{md} &= \frac{N_{11}}{N - M_{.0}} - \frac{N_{01}}{M_{.0}} = \frac{\sum_{i=1}^N Z_i Y_i(1)}{N - M_{.0}} - \frac{\sum_{i=1}^N (1 - Z_i) Y_i(0)}{M_{.0}} \\ &= \sum_{i=1}^N Z_i \frac{Y_i(1)}{N - M_{.0}} - \sum_{i=1}^N (1 - Z_i) \frac{Y_i(0)}{M_{.0}} = \sum_{i=1}^N Z_i [\frac{Y_i(1)}{N - M_{.0}} + \frac{Y_i(0)}{M_{.0}}] - \sum_{i=1}^N \frac{Y_i(0)}{M_{.0}}. \end{aligned}$$

The previously mentioned finite population CLT(Theorem 5.3) can be applied if we let $v_{Ni} = \frac{Y_i(1)}{N - M_{.0}} + \frac{Y_i(0)}{M_{.0}}$. usually the results holds if H_F or H_N is true.

However, unless the conditions listed for finite population CLT are fulfilled, the standardized statistic will not follow an asymptotic normal distribution necessarily. For instance, suppose $M_1 = 1, M_2 = M_3 = 0, M_4 = N$, $\frac{max(v_{Ni} - v_{N.})^2}{\sum(v_{Nj} - vN.)^2} max(\frac{m}{n}, \frac{n}{m})$ actually tends to infinity. This condition is rather special and not realistic, and we can in fact show that under regular conditions where the proportion of each type tends to a fixed ratio, the theorem holds.

Corollary 5.4.1. *Suppose that $\frac{1}{N}(M_{.0}, M_{1.}, M_{2.}, M_{3.}, M_{4.}) \to (p_0, p_1, p_2, p_3, p_4)$ and $p_i < 1$, then standardized statistic is asymptotically normal.*

Proof. Only need to show $\frac{max(v_{Ni}-v_{N.})^2}{\sum(v_{Nj}-vN.)^2} max(\frac{m}{n}, \frac{n}{m}) \to 0$ as $N \to \infty$. In fact we have,

$$\frac{max(v_{Ni}-v_{N.})^2}{\sum(v_{Nj}-vN.)^2} max(\frac{m}{n}, \frac{n}{m}) \to \frac{1}{N} \frac{max(c_i)}{\sum_{i=1}^{4} p_i c_i} max(p_0, 1-p_0), \tag{8}$$

where $c_i = (v - v_{N.})^2$ for v from group i and c_i is bounded. Hence, the previous equation tends to 0 and standardized mean difference tends to a normal distribution. □

6 Discussion and Conclusion

In summary, the paper proposed a randomization test for Fisher's null and Neyman's null based on potential outcome framework and Fisher's exact test. Simulation was carried out for visualizing the power of such test against difference alternatives. It turned out that the randomization test for Fisher's null may also be valid for Neyman's null. This conjecture can be proved for asymptotic cases, by showing that Fisher's exact test is equivalent to Neyman's test. However, for finite sample size, the conjecture is not yet proved due to technical difficulties. Several attempts including exploration on the link with sensitivity analysis were made.

There are several interesting questions open for discussion. Firstly, whether the conjecture is true for small sample size and under what conditions such conjecture is true. It is appealing to extend simulation to cases beyond binary outcomes and check whether similar results exist. Secondly, formulation of new tests for Neyman's null. Randomization test can be easily defined for Fisher's null, but such null hypothesis can be unrealistic and of little practical usage. But different tests for Neyman's null or other weak null have been established and comparison of their performance may be impactful. Thirdly, study the performance of our proposed test under local and non-local alternatives. It was shown that Fisher's exact test could not obtain asymptotic power 1 even for some non-local alternatives. Investigation on why this happens and how this issue could be fixed is intriguing.

References

Barnard, G. (1947). Significance tests for 2× 2 tables. *Biometrika*, *34*(1/2), 123–138.

Carlin, B. P., & Louis, T. A. (2008). *Bayesian methods for data analysis*. CRC Press.

Caughey, D., Dafoe, A., Li, X., & Miratrix, L. (2021). Randomization inference beyond the sharp null: Bounded null hypotheses and quantiles of individual treatment effects. *arXiv preprint arXiv:2101.09195*.

Chiba, Y. (2015). Exact tests for the weak causal null hypothesis on a binary outcome in randomized trials. *Journal of Biometrics and Biostatistics*, *6*, 244.

Ding, P. (2017). A paradox from randomization-based causal inference. *Statistical science*, 331–345.

Feller, W. (1971). An introduction to probability theory and its applications. *Vols. I & II, Wiley I*, *968*.

Fisher, R. A. (1936). Design of experiments. *British Medical Journal*, *1*(3923), 554.

Graham, R. L., Knuth, D. E., Patashnik, O., & Liu, S. (1989). Concrete mathematics: A foundation for computer science. *Computers in Physics*, *3*(5), 106–107.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960.

Lehmann, E. L., & D'Abrera, H. J. (1975). *Nonparametrics: Statistical methods based on ranks*. Holden-day.

Lehmann, E. L., Romano, J. P., & Casella, G. (2005). *Testing statistical hypotheses* (Vol. 3). Springer.

Li, X., & Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, *112*(520), 1759–1769.

Little, R. J. (1989). Testing the equality of two independent binomial proportions. *The American Statistician*, *43*(4), 283–288.

Loh, W. W., Richardson, T. S., & Robins, J. M. (2017). An apparent paradox explained. *Statistical Science*, *32*(3), 356–361.

Præstgaard, J. T. (1995). Permutation and bootstrap kolmogorov-smirnov tests for the equality of two distributions. *Scandinavian Journal of Statistics*, 305–322.

Rosenbaum, P. R., Rosenbaum, P., & Briskman. (2010). *Design of observational studies* (Vol. 10). Springer.

Spivey, M. (2011). Partial sum of rows of pascal's triangle (Mathematics StackExchange, Ed.) [Accessed: 2022-07-30].

Zhang, Y., & Zhao, Q. (2022). What is a randomization test? *arXiv preprint arXiv:2203.10980*.