Maestría en Economía
MEC5011 Econometrics 1
Universidad de Montevideo

Replication of "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results".

Valentina Roballo
Horacio Castellanos
Belén Puga

January 31, 2023

1 Introduction

In this paper we decided to replicate the results of (Young, 2019) using the novel R package designed by (Olivares-González and Sarmiento-Barbieri, 2020) because we wanted to test the validity of permutation tests vs classical econometric theory regarding Average Treatment Effects (ATE), using simulated data and also data sets from papers. In (Olivares-González and Sarmiento-Barbieri, 2020) we are provided by a robust test for testing treatment effects as we can test difference in means in data sets coming from different probability distributions and our goal is to test permutation tests vs classical econometric theory.

(Young, 2019) uses 53 data sets from different papers and checks the validity of the significance of individual and joint test of multiple treatment effects presented on the papers results, and finds that many of those have overestimated the statistical significance of their impact estimates, attributing these discrepancies to a concentration of leverage in a few observations, that makes coefficients and standard errors extremely volatile and generate t-statistics distributions with large tail probabilities, since their values get dependant with the realization of small number of residuals.

(Canay, 2022) establishes that classical permutation tests approach may not lead to a valid test and could over reject in finite samples, so we need to adequate them taking this into account, using the test defined in (Chung and Romano, 2013) to estimate ATE.

The paper is organized as follows: First, we briefly state some definitions of Basic Abstract Algebra needed to understand how permutation tests work. Then, we explain how permutation tests work and provide two examples. In section IV we present the data used and finally, in the section V we present the results found and the conclusions to which we arrived.

2 Basic Abstract Algebra

Definition 2.1. *Let G be a set. A* binary operation *is a map of sets:*

$$* : G \times G \to G$$

For simplicity we will write $*(x, y)$ as $x * y$. for all $x, y \in G$.

Definition 2.2. *A* group *is a set G together with a binary operation $*$ such that the following hold:*

- *(Associativity): $(a * b) * c = a * (b * c)$ for all $a, b, c \in G$*
- *(Existence of identity) $\exists e \in G$ such that $a * e = a$ for all $a \in G$*
- *(Existence of inverses) Given $a \in G; \exists b \in G$ such that $a * b = b * a = e$*

Definition 2.3. *A group $(G, *)$ is called* abelian *if it also satisfies*

$$a * b = b * a; \forall a, b \in G$$

The classic example of an abelian group is $(\mathbb{Z}, +)$ the set of integers together with the sum. One example of a non abelian group is the set of invertible $n \times n$ matrices together with matrix multiplication.

Definition 2.4. *Let $(G, *)$ be a group. A* subgroup *of G is a subset $H \subset G$ such that*

- $e \in H$
- $x, y \in H \implies x * y \in H$
- $x \in H \implies x^{-1} \in H$

As the reader can verify, a subgroup is a subset that also has the same group structure, it also has the same identity element.

Theorem 2.1. *Let $(G, *)$ be a group, then $(H, *)$ is a subgroup of $(G, *)$ if and only if*

- *H is non-empty*
- *For all a,b $\in H$ $a * b^{-1} \in H$*

This theorem is called the one-step subroup test.

Theorem 2.2. *Let $(G, *)$ be a group with identity element e. Then the trivial subgroup $(\{e\}, *)$ is a subgroup of $(G, *)$*

Proof. We can use the one-step subgroup test

- $e \in \{e\}$ so e is non-empty
- $e * e^{-1} = e \in \{e\}$

□

Definition 2.5. *Let X be a set and $(G, *)$ a group. An* action of G over X *is a function $\mu \colon G \times X \to X$ such that $\mu(e, x) = x$ and $\mu(g, \mu(h, x)) = \mu(g * h, x)$. If such function exists, we say that G acts over X and X is a G-set. Usually, if the context is clear, we use the notation $\mu(g, x) = gx$.*

Proposition 2.1. *Let $(G, *)$ be a group and $g \in G$. Then $G = G * g$.*

Proof. Since G is closed under the binary operation, then $G * g^{-1} \subseteq G$. So we have $(G * g^{-1}) * g \subseteq G * g$. On the other hand, we have that since G is closed, $G * g \subseteq G$.

□

Definition 2.6. *Let X be a G-set and $x \in X$, the subset $G_x := \{g \in G \mid g * x = x\}$ is called the stabilizer subgroup of x or the isotropy subgroup of x. If the stabilizer subgroup corresponds to the trivial subgroup for all $x \in X$, the action is called a free action.*

Lemma 2.1. *The stabilizer subgroup is a subgroup of $(G, *)$.*

Proof. We have that for all $g, s \in G_x$ $(sg)x = s(gx) = sx = x$. Equally we have if $g \in G_x$ we have $gx = x$ so $(g^{-1}g)x = g^{-1}x = x$ which implies that $g^{-1} \in G_x$. So we can conclude G_x is a subgroup of $(G, *)$.

□

Definition 2.7. *The subset $Gx := \{g * x \in X \mid g \in G\}$ of X, is called the G-orbit of x.*

Lemma 2.2. *Let X be a G-set. The relation in X given by*

$$x \sim y \text{ if and only if } y \in Gx$$

is an equivalence relation. The equivalence class of an element x with respect to this equivalence relation is the orbit Gx.

Proof.

1. (Reflexivity) We have that for all $x \in X$, $e * x = x$. Therefore $x \in Gx$
2. (Symmetry) If $y \in Gx$ then exists $g \in G$ such tat $y = gx$. Therefore $g^{-1}(gx) = g^{-1}y$, thus $x = g^{-1}y$. i.e, exists $g^{-1} \in G$ such that $x = g^{-1}y$ therefore $x \in Gy$.
3. (Transitivity) Let $y \in Gx$ and let $z \in Gy$. Therefore, exists $g_1 \in G$ such that $y = g_1 x$. In the same manner, exits $g_2 \in G$ such that $z = g_2 y$. Therefore, we have $z = g_2(g_1 x)$ which is equal to $z = (g_1 g_2)x$. That is $z \in Gx$

□

Definition 2.8. *Let S be a set. A* permutation *of S is a bijection $f : S \to S$.*

Lemma 2.3. *Let S be a set.*

1. *Let f and g be two permutations of S. Then the composition of f and g is a permutation of S.*
2. *Let f be a permutation of S. Then the inverse of f is a permutation of S.*

The proof follows easily from the properties of bijective functions.

Lemma 2.4. *Let S be a set. The set of all permutations under the composition of functions forms a group $A(S)$.*

Proof. By 2.3 we have that permutations are closed under composition. Let $i : S \to S$ be the identity function, it is clear that the identity is a bijection and that $f \circ i = f$ and $i \circ f = f$ for all $f \in A(S)$. Since permutations are bijections and bijections by definition are invertible, therefore it exist f^{-1} such that $f \circ f^{-1} = f^{-1} \circ f = i$. So we can conclude that A(S) is a group.

□

Lemma 2.5. *Let S be a finite set with n elements, then $A(S)$ has $n!$ elements.*

The proof is omitted but it is a well known result in abstract algebra. If the set S is finite and has n elements, normally the permutation group of S is denoted as S_n.

3 Permutation Tests

Consider a data set X taking values in a sample space $\mathcal{X}$ and we are interested in testing the null hypothesis H_0 that the probability distribution P generating X belongs to a family of distributions Ω. Let G be a finite group of permutations $g : \mathcal{X} \to \mathcal{X}$. Usually in statistics literature, the word used for describing g is "transformation" in order to include reflections and rotations, and also differentiate between "permutation" in a combinatorial sense i.e, re-ordering, but as the reader can verify, a transformation is a function between a set and itself so in order to have group structure, the functions need to be bijections, to avoid confusions we will simply call $g : \mathcal{X} \to \mathcal{X}$ a permutation. Following closely (Romano and Lehman, 2010) and (Canay, 2022) we will define the main idea behind permutation tests that allows us to test H_0.

Definition 3.1. *(Randomization Hyphotesis) Under H_0 the distribution of X is invariant under permutations in G, that is, for every $g \in G$, gX and X have the same distribution whenever X has a distribution $P \in \Omega$.*

In order to clarify the last idea, let's consider two independent samples $(Y_1, ..., Y_m)$ and $(Z_1, ..., Z_n)$. Under the H_0 both samples are generated from the same probability distribution, so if H_0 is true, the observations can be permuted or assigned at random to either of the two groups and the distribution of the permuted data is the same as the distribution of the original samples. Now, we will describe how permutation tests are built. Let $T(X)$ be any real valued test statistic for testing H_0 and let's suppose G has M elements and $X = x$ and let

$$T(x)^{(1)} \leq T(x)^{(2)} \leq ... \leq T(x)^{(M)}$$

Be the ordered values of $T(gX)$ as g varies in G. Let's fix a nominal level $\alpha, 0 < \alpha < 1$ and let $k = M - [M\alpha]$ where $[C]$ is the greatest integer less than or equal to C and α is the desired type I error rate.

Let's define

$$M^+(x) = \#\{g \in G : T(gx) > T^k(x)\}$$

$$M^0(x) = \#\{g \in G : T(gx) = T^k(x)\}$$

$$a(x) = \frac{M\alpha - M^+}{M^0}$$

So we can define the test function

$$\phi(x) = 1\{T(x) > T^k(x)\} + a(x)1\{T(x) = T^k(x)\}$$

Since we have that $M^+(x) \leq M - k \leq M\alpha$ and $M^+(x) + M^0(x) \geq M - k + 1 > M\alpha$ we can conclude that $0 \leq a(x) < 1$. According to (Hoeffding, 1952) the test function $\phi(x)$ corresponds to the probability that H_0 is rejected when $X = x$.

Theorem 3.1. *Suppose that X has a distribution P on $\mathcal{X}$ and the problem is to test the null hypothesis $P \in \Omega$. Let G be a finite group of permutations $g : \mathcal{X} \to \mathcal{X}$. Suppose that the randomization hypothesis holds. Given a test statistic $T(X)$ and $\phi(x)$ the randomization test, then*

$$E_P(\phi(X)) = \alpha, \textit{for all } P \in \Omega$$

Proof. By 2.1 we have

$$(T^{(1)}(X), ..., T^{(M)}(X)) = (T^{(1)}(gX), ..., T^{(M)}(gX))$$

So we have that $T^{(k)}(X) = T^{(k)}(gX)$, $M^0(X) = M^0(gX)$ and $a(X) = a(gX)$. Therefore

$$\sum_{g \in G} \phi(gX) =$$

$$\sum_{g \in G} 1\{T(gX) > T^{(k)}(gX)\} + a(gX)1\{T(gX) = T^{(k)}(gX)\} =$$

$$\sum_{g \in G} 1\{T(X) > T^{(k)}(X)\} + a(X)1\{T(X) = T^{(k)}(X)\}$$

Which by construction equals

$$M^+(X) + a(X)M^0(X) = M\alpha$$

So we can conclude that under H_0 , for every $g \in G$, $\phi(X) = \phi(gX)$. Therefore

$$E_P(\phi(X)) = \frac{1}{M} E_P(\sum_{g \in G} \phi(gX)) = \alpha$$

.

□

A conclusion that we can draw from the previous proof is that permutation tests are exact. Following (Hoeffding, 1952) and (Hemerik, 2013) we have that under H_0, $P(rejectH_0) = E_P(\phi(X)) = \alpha$. Another important remark is that the previous proof is only possible relying on the group structure of G.

Example 3.1. *(Two sample problem) Suppose that $(Y_1, ..., Y_m)$ are i.i.d observations from a probability distribution P_Y and $(Z_1, ..., Z_n)$ are i.i.d observations from a probability distribution P_Z. In this case we have $X = (Y_1, ..., Y_m, Z_1, ..., Z_n)$ and the sample space is $\mathcal{X} = \mathbb{R}^N$ where $N = m + n$. We are interested in testing*

$$H_0 : P_Y = P_Z \; vs \; H_1 : P_Y \neq P_Z$$

So, for $x = (x_1, ..., x_N) \in \mathbb{R}^N$ we define G such that for all $g \in G$, $gx \in \mathbb{R}^N$ is defined as $gx = (x_{\pi(1)}, ..., x_{\pi(N)})$, where $(\pi(1), ..., \pi(N))$ is a permutation of $\{1, ..., N\}$, so in this case we have $M = N!$. In this case when $P_Y = P_Z$, gX and X have the same distribution. In this case, each transformation $g \in G$ produce a new data set gx and the first m elements are used as the Y sample and the remaining n as the Z sample.

Example 3.2. *Let's consider a practical example taken from (Hemerik, 2013). We have two types of soil, Type A and Type B, and we want to know if the type of soil has any influence in the lenght of the plants. So we define an experiment, where we growth 10 plants in Type A soil and 10 in Type B soil. So we have the following data $X = (X_1, ..., X_{20})$ where $(X_1, ..., X_{10})$ represents the height of plants growth in type A soil and $(X_{11}, ..., X_{20})$ represents the height of plants growth in type B soil. Following this, we can define a test statistic T:*

$$T(X) = \mid \sum_{i=1}^{10} X_i - \sum_{i=11}^{20} X_i \mid$$

Where high values clearly indicates that the type of soil has an effect on plant growth. Other conclusion that we can draw from the test is that we can reorder the 20 plants in 20! *different ways, so we can perform* 20! *different tests, each corresponding to a different permutation of the data set. If we define the null hypothesis H_0 to be that the type of soil doesn't have any influence in plant growth, in other words, it means that all X_i are identically distributed. We can test the null hypothesis in the following manner, let's consider the ordered test values:*

$$T^{(1)}(X) \leq ... \leq T^{(20!)}(X)$$

So, in this case we have $M = 20!$. *Suppose that we want a probability of type I error* $\alpha = 0.05$, *so we have* $k = 20! - [0.05(20!)]$. *In the same manner we could easily built* $\phi(x)$, *the randomization test and reject or not reject the null hyphotesis.*

One of the main disadvantages of permutation tests is that it can be computationally intensive, for testing 20 data points we have $20! = 2432902008176640000$ possible tests. There are some approaches that can be used to deal with this issue, one is to select randomly a subset of the 20! permuted tests, others involve taking a more abstract algebraic approach, the interested reader can consult (Hemerik, 2013) for more details, as this is not going to be covered in this paper.

3.1 Asymptotic Behaviour of Permutation Tests

Now we will consider a sequence $X = X^n = (X_1, ..., X_n)$, $P = P_n$, $\mathcal{X} = \mathcal{X}_n$, $G = G_n$, $T = T_n$, etc. We are interested in the behaviour of permutation tests as $n \to \infty$.

Definition 3.2. *Let* $\hat{R}_n$ *denote the* randomization distribution *of* T_n, *it is defined as* $\hat{R}_n(t) = M_n^{-1} \sum_{g \in G_n} 1\{T_n(gX^n) \leq t\}$.

Definition 3.3. *The* $1 - \alpha$ *quantile of* $\hat{R}_n$ *is defined as* $\hat{r}_n(1-\alpha) = inf\{t : \hat{R}_n(t) \geq 1 - \alpha\}$.

Let G'_n be a random variable whose values are the M_n elements $g \in G_n$, each having the same probability $\frac{1}{M_n}$ i.e G'_n is uniform in G_n. Let's observe that $E(\hat{R}_n(t)) = P(T_n(G'_n X^n) \leq t)$. So if the randomization hypothesis holds $G_n X^n$ and X^n have the same distribution, so $E_{P_n}(\hat{R}_n(t)) = P(T_n(X^n) \leq t)$. So, if T_n converges in distribution to a c.d.f $R()$ which is continuous at t, then it follows that $E_{P_n}(\hat{R}_n(t)) \to R(t)$. In order to prove $\hat{R}_n(t) \xrightarrow{p} R(t)$ and $\hat{r}_n(1-\alpha) \xrightarrow{p} r(1-\alpha)$, the reader may recall that unbiasedness + vanishing variance implies convergence in mean-square; convergence in mean-square implies convergence in probability, so we need to show $Var_{P_n}(\hat{R}_n(t)) \to 0$. The details of this proof are omitted but the reader can consult them in (Romano and Lehman, 2010).

3.2 Permutation Tests For Parameters

Suppose that $(X_1, X_2, ..., X_n)$ are i.i.d observations from a probability distribution P and $(Y_1, Y_2, ..., Y_m)$ are i.i.d from a probability distribution Q, and also suppose that we are interested in testing $H_0 : \mu(P) = \mu(Q)$ vs $H_1 : \mu(P) \neq \mu(Q)$ where $\mu(P)$ denotes the mean of P. If we assume that $P = Q$ we could define a usual permutation test, but this is not normally the case. If $P \neq Q$ holds, it is proved in (Romano, 1990) that if the variances of P and Q are equal, the permutation test for means is asymptotically robust, but if the variances are unknown, then the test might not work. Fortunately in (Chung and Romano, 2013) a

test for $H_0 : \mu(P) = \mu(Q)$ vs $H_1 : \mu(P) \neq \mu(Q)$ without the necessity of $\sigma^2(Q) = \sigma^2(P)$ is defined, in fact, the test defined in (Chung and Romano, 2013) is valid for any parameter that holds certain properties.

Definition 3.4. *The notation $o_P(1)$ is short for a sequence of random vectors that converge in probability to 0. The notation The notation $O_P(1)$ denotes a sequence of random vectors that is bounded in probability. Let R_n be a sequence of random vectors then $X_n = o_P(R_n)$ means $X_n = Y_n R_n$ and $Y_n \xrightarrow{p} 0$. $X_n = O_P(R_n)$ means $X_n = Y_n R_n$ and $Y_n = O_P(1)$.*

Consider $(X_1, X_2, ..., X_n)$ are i.i.d observations from a probability distribution P and $(Y_1, Y_2, ..., Y_m)$ are i.i.d from a probability distribution Q. Let $\theta()$ be a real valued parameter and suppose we are interested in testing:

$$H_0 : \theta(P) = \theta(Q) \text{ vs } H_1 : \theta(P) \neq \theta(Q) \tag{1}$$

Now suppose that there exists an estimator $\hat{\theta}_n$ such that under P and Q is asymptotically linear, i.e. satisfies

$$n^{1/2}[\hat{\theta}_n - \theta(P)] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f_P(X_i) + o_P(1) \tag{2}$$

Similarly under Q we have

$$m^{1/2}[\hat{\theta}_m - \theta(Q)] = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} f_Q(Y_i) + o_Q(1) \tag{3}$$

Where f_Q and f_P are functions such that $E_P(f_P) = 0$ and $E_Q(f_Q) = 0$. We will also need the assumption that 2 holds for the mixture distribution $\tilde{P} = pP + (1-p)Q$.

Theorem 3.2. *Assume $X_1, ..., X_n$ are i.i.d P and $Y_1, ..., Y_m$ are i.i.d Q. Let $N = m + n$ and consider testing the null hypothesis 1 based on a test statistic of the form*

$$T_{m,n} = N^{1/2}[\hat{\theta}_m - \hat{\theta}_n]$$

Where the estimators are asymptotically linear and $0 < E_p(f_P^2) = \sigma^2(P) < \infty$(the same also holds for Q). Let $m \to \infty$, $n \to \infty$ and $p_n = n/N$, $q_m = m/n$ with $p_n \to p \in (0,1)$ such that $p_n - p = O(N^{-1/2})$. Assume that 2 holds for $\tilde{P} = pP + (1-p)$ then the permutation distribution of $T_{m,n}$ as defined in 2.3 satisfies

$$sup_t \mid \hat{R}_{m,n} - \phi(t/\tau^2(\tilde{P}) \mid \xrightarrow{p} 0$$

Where $\tau^2(\tilde{P}) = \frac{1}{p(1-p)}\sigma^2(\tilde{P})$.

The proof of the last theorem can be consulted in (Chung and Romano, 2013), also it can be consulted that under H_0 the true sampling distribution of $T_{m,n}$ is asymptotically normal with mean 0 and variance $\frac{1}{p}\sigma^2(P) + \frac{1}{1-p}\sigma^2(Q)$ which in general does not equal $\tau^2(\tilde{P})$. This disparity is corrected in (Chung and Romano, 2013) defining a studentized test.

Theorem 3.3. *Assume the same conditions as in 3.2 and assume that σ_n and σ_m are consistent estimators of $\sigma(P)$ and $\sigma(Q)$ respectively. Define the test statistic*

$$S_{m,n} = \frac{T_{m,n}}{V_{m,n}}$$

Where

$$V_{m,n} = \sqrt{\frac{N}{m}\hat{\sigma}_m^2 + \frac{N}{n}\hat{\sigma}_n^2}$$

And consider the permutation distribution of $S_{m,n}$, $\hat{R}^S_{m,n}$ then

$$sup_t \mid \hat{R}^S_{m,n} - \phi(t) \mid \xrightarrow{p} 0$$

Thus the permutation distribution is asymptotically standard normal as the true sampling distribution of $S_{m,n}$.

Proposition 3.1. *The sample mean is asymptotically linear.*

Proof. Let $X_1, ...X_n$ be i.i.d observations from a probability distribution P, and let $\bar{X}$ be the sample mean and μ the population mean. Then $n^{1/2}[\bar{X} - \mu] = \frac{1}{\sqrt{n}}\sum_{i=1}^n f_P(X_i)$, where $f_P = X_i - \mu$. Note that $E(f_P^2) = \sigma^2(P)$ so it also holds 3.2.

□

So, finally (!!!) following (Chung and Romano, 2013) we can define the studentized permutation test for means when the underlying distributions have different variances and may have different sample sizes as:

$$S_{m,n} = \frac{N^{1/2}(\bar{X}_n - \bar{Y}_m)}{\sqrt{NS_X^2/n + NS_Y^2/m}} \tag{4}$$

Where $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ and $S_Y^2 = \frac{1}{m-1}\sum_{i=1}^m (Y_i - \bar{Y})^2$.

3.3 Permutation Tests and ATE

Following (Canay, 2022), suppose that we observe a random sample $\{(Y_1, D_1), ..., (Y_n, D_n)\}$ from a randomized control trial where

$$Y = Y(1)D + (1 - D)Y(0)$$

is the observed outcome and $D \in \{0, 1\}$ is the exogenous treatment assigned. Under classical permutation tests we would define hypothesis test

$$H_0 : Q_0 = Q_1 \text{ vs } Q_0 \neq Q_1$$

Where Q_0 is the probability distribution of Y(0) and Q_1 is the probability distribution of Y(1). Under the null $\{(Y_1, D_1), ..., (Y_n, D_n)\}$ and $\{(Y_1, D_{\pi(1)}), ..., (Y_n, D_{\pi(n)})\}$ have the same distribution. However, in causal inference we are interested in the average treatment effect, i.e.

$$H_0 : E(Y(1)) = E(Y(0)) \text{ vs } H_1 : E(Y(1)) \neq E(Y(0))$$

As we already have seen and according to (Canay, 2022), classical permutation tests may not be valid, so we need an adequate test, taking this into account we propose the test defined in (Chung and Romano, 2013) to estimate ATEs. In this manner we will recreate (Young, 2019) using the studentized permutation test for means.

4 Data presentation

In order to check the validity of permutation tests, we will implement a robust test that comes in an R package designed by (Olivares-González and Sarmiento-Barbieri, 2020) called RATest, that allows to use this test defined in (Chung and Romano, 2013).

We used some public data sets available on R database and we also used simulated numerical data sets to recreate (Young, 2019), since many of the papers found that had a data set available did not meet all the criteria established in the paper.

We used five simulated data sets using different combinations of distributions and means. The first one uses a normal distribution with big mean differences. The second one with a small mean difference and a normal distribution. The third one has the same mean but using two different distributions (normal and poisson). The fourth one is using the same mean, different standard deviation and a normal distribution. And finally, the last data set is using a poisson distribution with a small mean difference.

Table 1: Simulated data sets

Definition			
Data Set	Obs	Mean	Std. Dev.
1	Normal	Big difference	No difference
2	Normal	Small difference	No difference
3	Normal and Poisson	No difference	No difference
4	Normal	No difference	Big difference
5	Poisson	Small difference	No difference

We also used 8 public data sets from different papers available for R, that are describe in Table 3.

Table 2: Real World data sets

Data Set	Description	Paper
U.S. Job Corps	Experimental study with information on the participation of disadvantaged youths in (academic and vocational) training in the first and second year after program assignment.	(Schochet, 2001)
Student's Sleep	Effect of two soporific drugs (increase in hours of sleep compared to control) on 10 patients.	(Scheffé, 1959)
Lalonde	National Supported Work Demonstration. This program randomly assigned applicants to the job training program (or out of the job training program).	(Lalonde, 1986)
General Social Survey	Public data set from the General Social Survey (GSS). Randomized controlled trial where individuals were asked about their thoughts on government spending on the social safety net.	(Smith, 2016)
Social Pressure and Voter Turnout.	A large-scale field experiment involving several registered voters used a series of mailings promising to publicize their turnout to their household or their neighbors to gauge these effects.	(Gerber, 2008).
Acute myeloid leukemia	A clinical trial to evaluate the efficacy of maintenance chemotherapy for acute myelogenous leukaemia at Stanford University.	(Embury S, 1977)
Health Evaluation and Linkage to Primary Care	The HELP study was a clinical trial for adult inpatients recruited from a detoxification unit. Patients with no primary care physician were randomized to receive a multidisciplinary assessment and a brief motivational intervention or usual care, with the goal of linking them to primary medical care.	(Samet J, 2003)
Video vs. Standard Laryngoscope	This study tested the hypothesis that intubation with the Pentax AWS would be easier and faster than with a standard Macintosh laryngoscope	(Abdallah R, 2011)

5 Results and conclusions

Our main results are that both, in the simulated data sets as well as in the real world data sets (except for one), the use of permutation tests didn't provide a rejection of a result already published in a paper. We used the simulated data sets as a sensitivity test for both the permutation test and the linear regression. In all cases, linear regression didn't provide false positive results, but one thing we observed is that permutation tests might be slightly more sensitive to tiny difference in means.

Our conclusion is that permutation tests are a powerful tool and a great addition to causal inference, but did not invalidate any of the previous published results. The only exception we had is in (Scheffé, 1959), but this might be with the fact that there are too few observations in the data set and the paper originally didn't use linear regression. The explanation we provide is that (Young, 2019) in its paper tested the sharp null hypothesis, i.e.

$$H_0 : Q_0 = Q_1 \text{ vs } Q_0 \neq Q_1$$

But as we can recall from (Canay, 2022) this hypothesis is too strong, as it is testing if a treatment is valid for all i in a data set, and this is of little interest in programe evaluations.

References

Abdallah R, Galway U, Y. J. K. A. S. D. D. D. (2011). A randomized comparison between the pentax aws video laryngoscope and the macintosh laryngoscope in morbidly obese patients. *Anesthesia Analgesia.*, 113:1082–7.

Canay, I. A. (2022). *Econ 480-3 Introduction to Econometrics*. Northwestern University, USA.

Canay, I. A., Romano, J. P., and Shaikh, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, 85(3).

Chung, E. and Romano, J. (2013). Exact and asymptotically robust permutation tests. *Annals of Statistics*, 41(2).

Embury S, Elias L, H. P. H. E. G. P. S. S. (1977). Remission maintenance therapy in acute myelogenous leukaemia. *Western Journal of Medicine*, 126:267–272.

Gerber, A., G. D. . L. C. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, 102(1):33–48.

Hemerik, J. (2013). *Permutation Tests and Multiple Testing*. Universiteit Leiden, Netherlands.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, 23(2).

Imai, K. (2013). *Statistical Hypothesis Tests*. Department of Politics, Princeton University, USA.

Judson, T. W. (2010). *Abstract Algebra Theory and Applications*. Stephen F. Austin State University, USA.

Lalonde, R. (1986). Evaluating the econometric evaluations of training programs. *American Economic Review*, 76:604–620.

Olivares-González, M. and Sarmiento-Barbieri, I. (2020). Ratest package. *https://cran.r-project.org/web/packages/RATest/index.html*.

Romano, J. and Lehman, E. (2010). *Testing Statistical Hypothesis*. Springer Texts in Statistics 4th edition, Switzerland.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumptions. *Journal of the American Statistical Association*, 85(411).

Samet J, Larson M, H. N. D. K. W. M. S. R. (2003). Linking alcohol and drug-dependent adults to primary medical care: A randomized controlled trial of a multi-disciplinary health intervention in a detoxification unit. *Addiction*, 98(4):509–516.

Scheffé, H. (1959). The analysis of variance.

Schochet, P. Z., B. J. G. S. (2001). National job corps study: The impacts of job corps on participants' employment and related outcomese. *Mathematica Policy Research*.

Smith, W. (2016). The general social surveys. *GSS Project Report No. 32*.

Wasserman, L. (2020). *36-705 Intermediate Statistics. Lecture notes*. Carnegie Mellon University, USA.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quaterly Journal of Economics*, 134(2).