

CHAPTER 4: FISHER'S EXACT TEST IN COMPLETELY RANDOMIZED EXPERIMENTS

Fisher (1925, 1926) was concerned with testing hypotheses regarding the effect of treatments. Specifically, he focused on testing *sharp null hypotheses*, that is, null hypotheses under which all potential outcomes are known exactly. Under such null hypotheses all unknown quantities in Table 4 in Chapter 1 are known—there are no missing data anymore. As we shall see, this implies that we can figure out the distribution of any statistic generated by the randomization. Fisher's great insight concerns the value of the physical randomization of the treatments for inference.

Fisher's classic example is that of the tea-drinking lady:

“A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. ... Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject in random order. ... Her task is to divide the cups into two sets of 4, agreeing, if possible, with the treatments received. ... The element in the experimental procedure which contains the essential safeguard is that the two modifications of the test beverage are to be prepared “in random order.” This is in fact the only point in the experimental procedure in which the laws of chance, which are to be in exclusive control of our frequency distribution, have been explicitly introduced. ... it may be said that the simple precaution of randomisation will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged.”

The approach is clear: an experiment is designed to evaluate the lady's claim to be able to discriminate whether the milk or tea was first poured into the cup. The null hypothesis of interest is that the lady has no such ability. In that case, when confronted with the task in

the experiment she will randomly choose four cups out of the set of eight. Choosing four objects out of a set of eight can be done in seventy different ways. When the null hypothesis is right, therefore, the lady will choose the correct four cups with probability $1/70$. If the lady chooses the correct four cups, then either a rare event took place – the probability of which is $1/70$ – or the null hypothesis that she cannot tell the difference is false. The significance level, or p-value, of correctly classifying the four cups is $1/70$.

Fisher describes at length the impossibility of ensuring that the cups of tea are identical in all aspects other than the order of pouring tea and milk. Some cups will have been poured earlier than others. The cups may differ in their thickness. The amounts of milk may differ. Although the researcher may, and in fact would be well advised to minimize such differences, they can never be completely eliminated. By the physical act of randomization, however, the possibility of a systematic effect of such factors is controlled in the formal sense defined above. Fisher was not the first to use randomization to eliminate systematic biases. For example, Peirce and Jastrow (18??, reprinted in Stigler, 1980, p75-83) used randomization in a psychological experiment with human subjects to ensure that “any possible guessing of what changes the [experimenter] was likely to select was avoided”, seemingly concerned with the human aspect of the experiment. Fisher, however, made it a cornerstone of this approach to inference in experiments, on human or other units.

In our setup with N units assigned to one of two levels of a treatment, it is the combination of (i), knowledge of all potential outcomes implied by a sharp null hypothesis with (ii), knowledge of the the distribution of the vector of assignments in classical randomized experiments that allows the enumeration of the distribution of any any *statistic*, that is, any function of observed outcomes and assignments.

Definition 1 (STATISTIC)

A statistic T is a known function $T(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X})$ of known assignments, \mathbf{W} , observed out-

comes, \mathbf{Y}^{obs} , and pretreatment variables, \mathbf{X} .

Comparing the value of the statistic for the realized assignment, say T^{obs} , with its distribution over the assignment mechanism leads to an assessment of the probability of observing a value of T as extreme as, or more extreme than what was observed. This allows the calculation of the p-value for the maintained hypothesis of a specific null set of unit treatment values.

This argument is similar to that of proof by contradiction in mathematics. In such a proof, the mathematician assumes the statement to be proved false is true and works out its implications until the point that a contradiction appears. Once a contradiction appears, either there was a mistake in the mathematical argument, or the original statement was false.

With the Fisher randomization test, the argument is modified to allow for chance, but otherwise very similar. The null hypothesis of no treatment effects (or of a specific set of treatment effects) is assumed to be true. Its implication in the form of the distribution of the statistic is worked out. The observed value of the statistic is compared with this distribution, and if the observed value is very unlikely to occur under this distribution, it is interpreted as a contradiction at that level of significance, implying a violation of the basic assumption underlying the argument, that is, a violation of the null hypothesis.

An important feature of this approach is that it is truly *nonparametric*, in the sense of not relying on a model specified in terms of a set of unknown parameters. We do *not* model the distribution of the outcomes. In fact the potential outcomes $Y_i(0)$ and $Y_i(1)$ are regarded as fixed quantities. The only reason that the observed outcome Y_i^{obs} , and thus the statistic T , is random, is that there is a stochastic assignment mechanism that determines which of the two potential outcomes is revealed for each unit. Given that we have a randomized experiment, the assignment mechanism is known, and given that all potential outcomes are known under the null hypothesis, we do not need to make any modelling assumptions

to calculate distributions of any statistics, that is, functions of the assignment vector, the observed outcomes and the pretreatment variables. The distribution of the test statistic is induced by this assignment mechanism, through the *randomization distribution*. The validity of the p-values is therefore not dependent on the values or distribution of the potential outcomes. This does not mean, of course, that the values of the potential outcomes do not affect the properties of the test. These values will certainly affect the *power* of the test, that is, the expected p-value when the null hypothesis is false. They will not, however, affect the *validity* which depends only on the randomization.

A second important feature is that once the experiment has been designed, there are only three choices to be made by the researcher. The researcher only has to choose the null hypothesis to be tested, the statistic used to test the hypothesis, and the definition of “at least as or more extreme than”. These choices should be governed by the scientific nature of the problem and involve judgements concerning what are interesting hypothesis and alternatives. Especially the choice of statistic is very important for the power of the testing procedure.

4.1 A SIMPLE EXAMPLE WITH TWO UNITS

Let us consider a simple example with 2 units in a completely randomized experiment. Suppose that the first unit got assigned $W_1 = 1$ and so we observed $Y_1(1) = y_1$. The second unit therefore was assigned $W_2 = 1 - W_1 = 0$ and we observed $Y_2(0) = y_2$. We are interested in the null hypothesis that there is no effect whatsoever of the treatment. Under that hypothesis the two unobserved potential outcomes become known for each: $Y_1(0) = Y_1(1) = y_1$ and $Y_2(1) = Y_2(0) = y_2$. Now consider a statistic. The most obvious choice is difference between the observed value under treatment and the observed value under control:

$$T = W_1 \cdot (Y_1(1) - Y_2(0)) + (1 - W_1) \cdot (Y_2(1) - Y_1(0)),$$

(here we use the fact that $W_2 = 1 - W_1$ because we have a completely randomized experiment

with two units). The value of this statistic given the actual assignment, $w_1 = 1$, is

$$T^{\text{obs}} = Y_1(1) - Y_2(0) = y_1 - y_2.$$

Now consider the set of all possible assignments which in this case is simple: there are only two possible values for $\mathbf{W} = (W_1, W_2)'$. The first is the actual assignment vector $\mathbf{W}_1 = (1, 0)$, and the second is the reverse, $\mathbf{W}_2 = (0, 1)$. For each vector of assignments we can calculate what the value of the statistic would have been. In the first case

$$T^{\mathbf{W}_1} = y_1 - y_2,$$

and in the second case

$$T^{\mathbf{W}_2} = y_2 - y_1.$$

Both assignment vectors have equal probability, that is probability $1/2$, so the distribution of the statistic, maintaining the null hypothesis, is

$$Pr(T = t) = \begin{cases} 1/2 & \text{if } t = y_1 - y_2, \text{ or } t = y_2 - y_1, \\ 0 & \text{otherwise.} \end{cases}$$

So, under the null hypothesis we know the entire distribution of the statistic T . We also know the value of T given the actual assignment, $T^{\text{obs}} = T^{\mathbf{W}^{\text{obs}}}$.

In this case the p-value is $1/2$ (and this is always the case for this two-unit experiment, irrespective of the outcome), so the outcome does not appear extreme; with only two units, irrespective of the null hypothesis, no outcome is unusual.

To illustrate the potential for the level of “unusualness” to increase with additional units, suppose we have a completely randomized experiment with $2N$ units, N of whom are to be assigned to treatment, and the remaining N to control. Now there are $\binom{2N}{N}$ possible values for the assignment vector, whereas with 2 units there were only two possible values.

Hence, whereas before the only possible p-value was $1/2$, now the p-value can be as small as $1/\binom{2N}{N}$. See for illustrative calculations Table 1 in Chapter 3.

4.2 A SIMPLE EXAMPLE WITH SIX UNITS

Table 1 presents observations from a randomized experiment to evaluate the effect of an educational television program on reading skills. The unit is a class of students. The outcome of interest is the average reading score in the class. Half of the classes were shown the program of interest, and the other half did not have access to the program. The characteristics measured were in indicator for the treatment, an indicator for whether the class was in Fresno or Youngstown, the grade of the class, a pre-test score, and the post-test score. (ETS report)

Initially, we shall only analyze the first six observations from this data set. The null hypothesis of preliminary interest is that the program has no effect on reading scores whatsoever, that is, $Y_i(0) = Y_i(1)$ for all $i = 1, \dots, 6$. Under the null hypothesis, we can fill in the missing entries in Table 1 using the observed outcomes. See Table 2 for the layout with all the missing and observed potential outcomes. Now consider testing this null hypothesis using the difference in the outcomes by treatment status as the test statistic:

$$T_1 = \sum_{i=1}^6 W_i \cdot Y_i^{\text{obs}} / 3 - \sum_{i=1}^6 (1 - W_i) \cdot Y_i^{\text{obs}} / 3 = \bar{y}_1 - \bar{y}_0,$$

where

$$\bar{y}_w = \sum_{i=1}^N 1\{W_i = w\} \cdot Y_i^{\text{obs}} / \sum_{i=1}^N 1\{W_i = w\}.$$

Under the null hypothesis we can calculate the value of this statistic under each vector of treatment assignments, \mathbf{W} . There are $\binom{6}{3} = 20$ such assignments. In Table 3 we list all 20 different assignments with three classes exposed to the experimental television program and three not exposed. For each vector of assignments we calculate the value of the statistic. The first row in this table corresponds to the actual vector of assignments. On average the

reading score for the three exposed classes is 5.1 points higher than for the not-exposed classes. How unusual is it that we would have observed as big a gain for the exposed classes as we did if there was no effect of the television reading program? Counting from Table 3 we see that there are six (out of twenty) vectors of assignments that would lead to as large a difference between exposed and not-exposed classes as we in fact found, leading to a p-value of $6/20 = 0.30$, suggesting that under the null hypothesis the observed difference could well be due to chance. Fifteen out of twenty vectors of assignments have a difference between exposed and not-exposed classes that is smaller than or equal to the one based on the observed statistic, with p-value for this test is therefore $15/20 = 0.75$. Finally, twelve out of the twenty have a difference in average treated and control outcomes that is in absolute value at least as large as the difference we found, leading to a p-value of $12/20 = 0.60$.

4.3 THE CHOICE OF NULL HYPOTHESIS

There are three choices to be made by the researcher, the null hypothesis, the statistic, and the measure of extremeness. We shall consider each of these choices in more detail.

The first is the choice of null hypothesis. Typically the most interesting sharp null hypothesis is the hypothesis that there is no effect at all of the treatment, and $Y_i(0) = Y_i(1)$ for all units. We do not necessarily believe that this null hypothesis is correct, but wish to see how strongly the data can reject this hypothesis. Note that this is distinctly different from the null hypothesis that the *average* treatment effect is zero. This second, or “average null” hypothesis is not a sharp null hypothesis because it does not specify values for all potential outcomes under the null hypothesis, and therefore does not fit into the framework outlined by Fisher. This does not imply it is more or less interesting than the hypothesis that the treatment effect is zero for all units. Neyman, who, as we shall see in Chapter 4, focussed on the average null, was attacked by Fisher (1930) in a sharp exchange:

fisher-neyman exchange in discussion of Neyman paper

**

Although Fisher's approach cannot accommodate Neyman's null hypothesis, it can accommodate any sharp null hypothesis. An alternative to the sharp null of no effects whatsoever may be the hypothesis that there is a constant additive treatment effect, $Y_i(1) = Y_i(0) + c$, or that there is a constant multiplicative treatment effect, $Y_i(1) = c \cdot Y_i(0)$ for some prespecified value of the treatment effect c . Once we depart from the world of absolutely no effects, however, it becomes more difficult to argue why the treatment effect should be additive in levels, rather than in logarithms or any other transformation of the basic outcome.

The most general case is one where the null hypothesis is $Y_i(1) = Y_i(0) + c_i$ for some set of prespecified treatment effects c_i . It is rare in practice, however, to have a hypothesis precise enough to specify treatment effects for all units without these treatment effects being identical for all units.

4.4 THE CHOICE OF STATISTIC

The choice of null hypothesis is dictated by the substantive aspect of the analysis and is often obvious. The second decision to be made by the researcher is the choice of statistic and this is typically more difficult. A standard choice is often the difference in average outcomes by treatment status minus the average effect under the null hypothesis:

$$T_1 = \frac{\sum W_i \cdot Y_i(1)}{\sum W_i} - \frac{\sum (1 - W_i) \cdot Y_i(0)}{\sum 1 - W_i} - \frac{1}{N} \sum_{i=1}^N c_i = \bar{y}_1 - \bar{y}_0 - \bar{c},$$

or its absolute value. An obvious alternative is to first transform the outcomes, for example to logarithms if the outcomes are positive with a skewed distribution, such as, for example, incomes.

$$T_2 = \frac{\sum W_i \cdot \ln Y_i(1)}{\sum W_i} - \frac{\sum (1 - W_i) \cdot \ln Y_i(0)}{\sum 1 - W_i} - \frac{1}{N} \sum_{i=1}^N \tilde{c}_i,$$

where \tilde{c}_i is the difference $\ln Y_i(1) - \ln Y_i(0)$ under the null hypothesis. The last column in Table 3 presents the distribution of this statistic under the null hypothesis of no effect. The statistic under the actual vector of assignments is 0.079. A simple count shows that the number of values of the statistic larger than, or equal to, this is seven, leading to a p-value of 0.35. Note that this differs slightly from 0.30, which was the p-value under the previous statistic, the difference in treatment and control averages. In general the test statistics based on differences in average levels and logarithms (or any other monotone, not dramatically different transformation) are likely to give similar, but not identical, answers. Note that a p-value has its pristine interpretation only once—one cannot do two and take the smallest p-value. In general the p-value of a function of two statistics (e.g., the minimum of the two) is not equal to the function of the two p-values.

The choice of statistic is not, however, limited to such simple averages. One could compare the median outcome for the treated with the median outcome for the controls, or even other quantiles. So far the statistics considered have distributions that are centered around zero under the null hypothesis. Even this need not be true in general. One can calculate the 75th percentile for the treated and subtract the 25th percentile for the controls. Whatever the statistic is, the randomization distribution guarantees the validity of the induced p-value.

Given this bewildering choice of statistics, the question arises as to how to choose among them? In principal, the choice of statistic should be governed by thinking about plausible alternative hypotheses. If one expects the effect of the treatment to be additive or multiplicative, one should use differences in average levels or logarithms respectively. On the other hand, if one expects the treatment to increase the dispersion of the outcomes, one may use the differences in the interquartile range for the treated and controls as the statistic of interest. In addition, if the empirical distribution of the outcomes has some outliers, calculating average differences by treatment status may lead to a test with very low power. It may be

possible in that case to construct more powerful tests using robust estimates of the center of the outcome distributions by treatment status such as the median or trimmed means or rank tests. More generally, we can use a model-based statistic such as a maximum likelihood estimator.

4.5 DEFINITIONS OF EXTREMENESS

The final choice confronting the researcher is the operationalization of the measure of extremeness. Given a null hypothesis and a test statistic T , we can derive the distribution of the statistic under the null and calculate the observed value of the statistic. The probability of a random draw of T from the distribution under the null hypothesis being exactly equal to the observed value of the test statistic is typically extremely small, though never zero. This measure of extremeness is therefore not very useful. Instead we often use the probability of a random draw from the distribution being at least as large as the value of the observed statistic. An alternative is to calculate the probability of such a draw being at least as small as the observed statistic. A third possibility is to consider the minimum of the two earlier probabilities, the minimum of the probability of a random draw from the distribution being at least as large as the value of the observed statistic and the probability of such a draw being at least as small as the observed statistic.

To illustrate this, let us use Fisher's approach to testing for different values under the null hypothesis. In Table 4 we report for a set of treatment effects c the corresponding p-value for testing the null hypothesis $Y_i(1) - Y_i(0) = c$ against the alternative $Y_i(1) - Y_i(0) \neq c$ for at least one unit. The statistic is the absolute value of the difference in average treated and control units minus c , and the p-value is the proportion of draws of the assignment vector leading to statistics at least as large as the observed value of the statistic.

One limitation that has to be kept in mind when choosing between tests is that the validity of the test depends on the commitment to a null hypothesis, statistic and measure of

extremeness. The p-values are valid for each triple separately, but they are not independent across triples. Specifically, consider two statistics, $T_1(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X})$, and $T_2(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X})$, with realized values T_{1a} and T_{2a} . Under any null hypothesis sharp null hypothesis one can calculate the p-values for each of the tests,

$$p_1 = \min\{Pr(T_1 \leq T_{1a}), Pr(T_1 \geq T_{1a})\},$$

and

$$p_2 = \min\{Pr(T_2 \leq T_{2a}), Pr(T_2 \geq T_{2a})\}.$$

These p-values are valid for each test separately, but one cannot consider the minimum of p_1 and p_2 and use that as a p-value for the null hypothesis.

4.6 COMPUTATION

The p-value calculations presented so far have been exact. When the p-value for the null hypothesis of no effect, based on the statistic of the difference in treatment and control averages was 0.30, it meant that under that null hypothesis, the probability of getting a statistic larger than or equal to, in absolute value, 5.1, was exactly 0.30. We could do these exact calculations because the sample was so small. In general, with N units and M subject to treatment, the number of distinct values of the assignment vector is $\binom{N}{M}$. With both N and M large it may not be easy to calculate the statistic for every value of the assignment vector. However, it is typically very easy to obtain a very accurate approximation to the p-value. Instead of calculating the statistic for every single value of the assignment vector, we can calculate it for randomly chosen values of the assignment vector.

Formally, randomly draw an N -dimensional vector with $N - M$ zeros and M ones from the population of such vectors. For each member of this population the probability of being drawn is $1 / \binom{N}{M}$. Calculate the statistic for this draw, and denote it by T_1 . Repeat this process $K - 1$ times, each time drawing another vector of assignments with or without

replacement and calculating the statistic T_k , for $k = 2, \dots, K$. Then approximate the p-value by the fraction of these K statistics larger than the actual statistic T_a in absolute value. If K is large, the p-value will be very accurate. Note that the accuracy of the approximation is entirely within the control of the researcher. For a given degree of accuracy, one can determine the number of independent draws required.

To illustrate this we now analyze the full data set from the Children's Television Workshop Experiment presented in Table 1, separately for the 23 observations from Fresno and the 15 observations from Youngstown. Table 5 reports the p-value for the null hypothesis of no effect for $K = 100$, $K = 1000$, $K = 10,000$ and $K = 100,000$ draws from the distribution of assignment vectors. The statistic is the absolute value of the difference of average treated and control outcomes. The p-value measures the proportion of assignment vectors that leads to a statistic at least as large as the observed statistic.

Note that now there are far too many distinct values for the assignment vector, $\binom{23}{12}$ for the Fresno data and $\binom{15}{8}$ for the Youngstown data (over 32 million), to carry out exact calculations. In addition to reporting the estimates for the p-value, we also report standard errors for these estimates. These standard errors reflect the fact that we did not calculate the exact p-values, but instead estimated them by simulating K assignment vectors and calculating the frequency of statistics exceeding (or not exceeding) the observed value of the statistic. Given a p-value of p , and K draws from the assignment vector, the standard error is calculated as $\sqrt{p(1-p)/K}$.

Consider the Youngstown results. Suppose we wish to test the null hypothesis of zero effects at the 10% level. With only 100 draws from the distribution of the assignment vector we would not be able to tell whether we should reject the null hypothesis or not. Only by the time we use 10,000 draws is the standard error for the p-value small enough that we can be confident that the null hypothesis should not be rejected. For the Fresno data 100 draws

is sufficient to conclude that we should not reject the null hypothesis.

Table 1: DATA FROM CHILDREN'S TELEVISION WORKSHOP EXPERIMENT

Unit	Treatment	Fresno/Youngstown	Pre-test Score	Post-test Score
1	0	F	12.3	55.0
2	1	F	16.5	70.0
3	0	F	18.7	72.0
4	1	F	51.4	66.0
5	0	F	18.7	72.7
6	1	F	19.4	78.9
7	1	F	13.8	48.9
8	0	F	17.7	80.4
9	1	F	18.5	89.7
10	0	F	11.5	47.0
11	1	F	8.8	44.2
12	0	F	16.4	70.0
13	1	F	15.3	77.5
14	0	F	16.8	74.1
15	1	F	15.0	84.7
16	0	F	18.2	97.3
17	1	F	15.0	86.8
18	0	F	15.4	74.1
19	1	F	11.8	60.8
20	0	F	18.7	76.3
21	1	F	16.4	75.7
22	0	F	17.1	84.5
23	1	F	16.2	95.1
24	0	Y	12.9	54.6
25	1	Y	12.0	60.6
26	0	Y	15.1	56.5
27	1	Y	12.3	55.5
28	0	Y	16.8	75.2
29	1	Y	17.2	84.8
30	0	Y	15.8	75.6
31	1	Y	18.9	101.9
32	0	Y	13.9	55.3
33	1	Y	15.3	70.6
34	0	Y	14.5	59.3
35	1	Y	16.6	78.4
36	0	Y	17.0	87.0
37	1	Y	16.0	84.2
38	1	Y	20.1	108.6

Table 2: FIRST SIX OBSERVATIONS FROM CHILDREN'S TELEVISION WORKSHOP EXPERIMENT DATA WITH MISSING DATA IN BRACKETS FILLED IN UNDER THE NULL HYPOTHESIS OF NO EFFECT

Unit	Potential Outcomes		Actual Treatment	Observed Outcome
	$Y_i(0)$	$Y_i(1)$		
1	55.0	(55.0)	0	55.0
2	(70.0)	70.0	1	70.0
3	72.0	(72.0)	0	72.0
4	(66.0)	66.0	1	66.0
5	72.7	(72.7)	0	72.7
6	(78.9)	78.9	1	78.9

Table 3: RANDOMIZATION DISTRIBUTION FOR TWO STATISTICS FOR CHILDREN'S TELEVISION WORKSHOP DATA

W_1	W_2	W_3	W_4	W_5	W_6	$\left(\frac{\sum W_i \cdot Y_i^{\text{obs}}}{\sum (1 - W_i) \cdot Y_i^{\text{obs}}}\right)/3$	$\left(\frac{\sum W_i \cdot \ln Y_i^{\text{obs}}}{\sum (1 - W_i) \cdot \ln Y_i^{\text{obs}}}\right)/3$
0	0	0	1	1	1	5.1	0.079
0	0	1	0	1	1	6.9	0.104
0	0	1	1	0	1	9.5	0.143
0	0	1	1	1	0	0.9	0.024
0	1	0	0	1	1	6.4	0.097
0	1	0	1	0	1	9.1	0.137
0	1	0	1	1	0	0.5	0.018
0	1	1	0	0	1	10.9	0.162
0	1	1	0	1	0	2.3	0.043
0	1	1	1	0	0	4.9	0.082
1	0	0	0	1	1	-4.9	-0.082
1	0	0	1	0	1	-2.3	-0.043
1	0	0	1	1	0	-10.9	-0.162
1	0	1	0	0	1	-0.5	-0.018
1	0	1	0	1	0	-9.1	-0.137
1	0	1	1	0	0	-6.4	-0.097
1	1	0	0	0	1	-0.9	-0.024
1	1	0	0	1	0	-9.5	-0.143
1	1	0	1	0	0	-6.9	-0.104
1	1	1	0	0	0	-5.1	-0.079

Table 4: FIRST SIX OBSERVATIONS FROM DATA FROM CHILDREN'S TELEVISION WORKSHOP EXPERIMENT DATA WITH MISSING DATA IN BRACKETS UNDER THE NULL HYPOTHESIS OF A CONSTANT EFFECT OF SIZE 12

Unit	Potential Outcomes		Actual Treatment	Observed Outcome
	$Y_i(0)$	$Y_i(1)$		
1	55.0	(67.0)	0	55.0
2	(58.0)	70.0	1	70.0
3	72.0	(84.0)	0	72.0
4	(54.0)	66.0	1	66.0
5	72.7	(84.7)	0	72.7
6	(66.9)	78.9	1	78.9

Table 5: P-VALUES FOR TESTS OF CONSTANT TREATMENT EFFECTS USING ABSOLUTE VALUE OF DIFFERENCE IN TREATED AND CONTROLS AVERAGES MINUS THE HYPOTHESIZED VALUE, BASED ON PROPORTION OF STATISTICS AT LEAST AS LARGE AS OBSERVED STATISTIC

Treatment Effect	p-value	Treatment Effect	p-value
-20	0.05	11	0.30
-19	0.05	12	0.25
-18	0.05	13	0.25
-17	0.05	14	0.30
-16	0.05	15	0.30
-15	0.05	16	0.20
-14	0.05	17	0.20
-13	0.10	18	0.20
-12	0.10	19	0.20
-11	0.10	20	0.20
-10	0.10	21	0.20
-9	0.10	22	0.20
-8	0.10	23	0.20
-7	0.10	24	0.10
-6	0.30	25	0.10
-5	0.25	26	0.10
-4	0.35	27	0.10
-3	0.35	28	0.10
-2	0.50	29	0.10
-1	0.60	30	0.10
0	0.60	31	0.10
1	0.65	32	0.10
2	0.70	33	0.10
3	0.75	34	0.10
4	0.75	35	0.05
5	1.00	36	0.05
6	1.00	37	0.05
7	0.80	38	0.05
8	0.80	39	0.05
9	0.60	40	0.05
10	0.60	41	0.05

Table 6: SIMULATED P-VALUES FOR FRESNO AND YOUNGSTOWN DATA. NULL HYPOTHESIS OF ZERO EFFECTS FOR ALL UNITS. STATISTIC IS ABSOLUTE VALUE OF DIFFERENCE IN AVERAGE TREATED AND CONTROL OUTCOMES. P-VALUE IS PROPORTION OF DRAWS AT LEAST AS LARGE AS OBSERVED STATISTIC.

Number of Simulations	Fresno (N=23)		Youngstown (N=15)	
	p-value	(s.e.)	p-value	(s.e.)
100	0.980	(0.014)	0.060	(0.024)
1,000	0.987	(0.004)	0.117	(0.010)
10,000	0.980	(0.001)	0.109	(0.003)
100,000	0.981	(0.000)	0.113	(0.001)