CHAPTER 5

Fisher's Exact P-Values for Completely Randomized Experiments

5.1 INTRODUCTION

As discussed in Chapter 2, Fisher appears to have been the first to grasp fully the importance of physical randomization for credibly assessing causal effects (1925, 1936). A few years earlier, Neyman (1923) had introduced the language and the notation of potential outcomes, using this notation to define causal effects *as if* the assignments were determined by random draws from an urn, but he did not take the next logical step of appreciating the importance of actually randomizing. It was instead Fisher who made this leap.

Given data from a completely randomized experiment, Fisher was intent on assessing the *sharp null hypothesis* (or *exact null hypothesis*, Fisher, 1935) of no effect of the active versus control treatment, that is, the null hypothesis under which, for each unit in the experiment, both values of the potential outcomes are identical. In this setting, Fisher developed methods for calculating "p-values." We refer to them as *Fisher Exact P-values* (FEPs), although we use them more generally than Fisher originally proposed. Note that Fisher's null hypothesis of no effect of the treatment versus control whatsoever is distinct from the possibly more practical question of whether the *typical* (e.g., average) treatment effect across all units is zero. The latter is a weaker hypothesis, because the average treatment effect may be zero even when for some units the treatment effect is positive, as long as for some others the effect is negative. We discuss the testing of hypotheses on, and inference for, average treatment effects in Chapter 6. Under Fisher's null hypothesis, and under sharp null hypotheses more generally, for units with either potential outcome observed, the other potential outcome is known; and so, under such a sharp null hypothesis, both potential outcomes are "known" for each unit in the sample – being either directly observed or inferred through the sharp null hypothesis.

Consider any test statistic T: a function of the stochastic assignment vector, $\mathbf{W}$; the observed outcomes, $\mathbf{Y}^{\text{obs}}$; and any pre-treatment variables, $\mathbf{X}$. As we discuss in more detail shortly, the fact that the null hypothesis is sharp allows us to determine the distribution of T, generated by the complete randomization of units across treatments. The test statistic is stochastic solely through the stochastic nature of the assignment vector. We refer to the distribution of the statistic determined by the randomization as the *randomization distribution* of the test statistic T. Using this distribution, we can compare

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

the actually observed value of the test statistic, T^{obs}, against the distribution of T under the null hypothesis. An observed value that is "very unlikely," given the null hypothesis and the induced distribution for the test statistic, will be taken as evidence against the null hypothesis in what is, essentially, a stochastic version of the mathematician's "proof by contradiction."

How unusual the observed value is under the null hypothesis will be measured by the probability that a value as extreme or more extreme (in practice, as large or larger) would have been observed – the significance level or p-value. Hence, the FEP approach entails two steps: (*i*) the choice of a sharp null hypothesis (in Fisher's original version, always the null hypothesis of no effect whatsoever, but easily generalized to any sharp null hypothesis, that is, a null hypothesis that allows us to infer all the missing potential outcomes from the observed potential outcomes), and (*ii*) the choice of test statistic. The scientific nature of the problem should govern these choices. In particular, although in Fisher's analysis the null hypothesis was always the one with no treatment effect whatsoever, in general the null hypothesis should follow from the substantive question of interest. The statistic should then be chosen to be sensitive to the difference between the null and some alternative hypothesis that the researcher wants to assess for its scientific interest. That is, the statistic should be chosen to have, what is now commonly referred to as, *statistical power* against a scientifically interesting alternative hypothesis.

An important characteristic of this approach is that it is truly nonparametric, in the sense that it does not rely on a model specified in terms of a set of unknown parameters. In particular, we do not model the distribution of the outcomes: the vectors of potential outcomes $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ are regarded as fixed but *a priori* unknown quantities. The only reason that the observed outcomes, $\mathbf{Y}^{\text{obs}}$, and thus the statistic, T^{obs}, are random is that a stochastic assignment mechanism determines which of the two potential outcomes we observe for each unit. This assignment mechanism is, by definition, known for a classical randomized experiment. In addition, given the null hypothesis, all potential outcomes are known. Thus, we do not need modeling assumptions to calculate the randomization distribution of any test statistic; instead, the assignment mechanism completely determines the randomization distribution of the test statistic. The validity of any resulting p-value is therefore not dependent on assumptions concerning the distribution of the potential outcomes. This freedom from reliance on modeling assumptions 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 distribution of the p-value when the null hypothesis is false (i.e., the statistical power of the test). They will not, however, affect the validity of the test, which depends solely on the randomized assignment mechanism.

The remainder of this chapter begins with a brief description of the data that we will use to illustrate this approach. The data set is from a completely randomized evaluation of the effect of honey on nocturnal cough and resulting sleep quality for coughing children. Next, in Section 5.3, we start with a simple example using data from only six of the seventy-two children in the experiment. After that follows a detailed discussion of the two choices necessary for calculating FEPs: in Section 5.4 we discuss the choice of the null hypothesis, and in Section 5.5 we discuss the choice of the test statistic. In Section 5.6 we carry out a small simulation study to illustrate the properties of the method. Next, in Section 5.7 we discuss how the FEP approach can be extended to construct interval estimates. We then continue in Section 5.8 with a discussion of how to estimate,

Table 5.1. *Summary Statistics for Observed Honey Data*

Variable	Mean	(S.D.)	Mean Controls	Mean Treated
Cough frequency prior to treatment (`cfp`)	3.86	(0.92)	3.73	4.00
Cough frequency after treatment (`cfa`)	2.47	(1.61)	2.81	2.11
Cough severity prior to treatment (`csp`)	3.99	(1.03)	3.97	4.00
Cough severity after treatment (`csa`)	2.54	(1.74)	2.86	2.20

rather than calculate exactly, the p-value – the level of significance associated with a given observed value of the test statistic – when N is so large that such exact calculations are tedious at best and possibly infeasible. Next, in Section 5.9, we discuss how to use covariates to refine the choice of statistic. In Section 5.10, we expand the analysis to apply this approach to the full sample in which a random subset of the group of seventy-two children was given honey as a cough treatment. Section 5.11 concludes.

5.2 THE PAUL ET AL. HONEY EXPERIMENT DATA

The data used in this chapter are from a randomized experiment by Paul et al. (2007) on the evaluation of the effect of three treatments on nocturnal cough and sleep difficulties associated with childhood upper respiratory tract infections. The three treatments are (*i*) a single dose of buckwheat honey; (*ii*) a single dose of honey-flavored dextromethorphan, an over-the-counter drug; and (*iii*) no active treatment. The subjects were 105 children between two and eighteen years of age. Here we only use data on the $N = 72$ children receiving buckwheat honey ($N_t = 35$) or no active treatment ($N_c = 37$). The authors measure six different outcomes. We focus on two of them, cough frequency afterwards (`cfa`), and cough severity afterwards (`csa`), referring to measures of cough frequency and severity the night after being randomly assigned or not to the administration of the treatment. Both outcomes are measured on a scale from zero ("not at all frequent/severe") to six ("extremely frequent/severe"). We also use two covariates, measured on the night prior to the randomized assignment: cough frequency prior (`cfp`) and cough severity prior (`cfp`), both measured on the same scale as the outcomes.

Table 5.1 presents some summary statistics (means and standard deviations, and means by treatment status) for the four observed variables (`cfp`, `cfa`, `csp`, `csa`), for the 72 children receiving honey or no active treatment in this study. In Table 5.2 we also present cumulative frequencies for the two outcomes variables (`cfa` and `csa`) by treatment group for the seven levels of the outcome scale.

5.3 A SIMPLE EXAMPLE WITH SIX UNITS

Initially let us consider, for relative ease of exposition and data display, a subsample from the honey data set, with six children. Table 5.3 gives the observed data on cough frequency for these six children in the potential outcome form. A key part of the table is the pair of columns listing the potential outcomes, observed and missing. The first child (unit 1) was assigned to the (buckwheat honey) treatment group ($W_1 = 1$). Hence we

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

Table 5.2. ***Cumulative Distribution Functions for Cough Frequency and Severity after Treatment Assignment for the Honey Study***

Value	cfa		csa	
	Controls	Treated	Controls	Treated
0	0.14	0.14	0.16	0.17
1	0.19	0.40	0.22	0.46
2	0.32	0.63	0.35	0.54
3	0.73	0.83	0.59	0.77
4	0.89	0.91	0.86	0.91
5	0.92	0.97	0.95	0.94
6	1.00	1.00	1.00	1.00

Table 5.3. ***Cough Frequency for the First Six Units from the Honey Study***

Unit	Potential Outcomes				
	Cough Frequency (cfa)		Observed Variables		
	$Y_i(0)$	$Y_i(1)$	W_i	X_i (cfp)	Y_i^{obs} (cfa)
1	?	3	1	4	3
2	?	5	1	6	5
3	?	0	1	4	0
4	4	?	0	4	4
5	0	?	0	1	0
6	1	?	0	5	1

observe $Y_1^{\text{obs}} = Y_1(1)$ (equal to 3 for this child). We do not observe $Y_1(0)$, and in the table this missing potential outcome is represented by a question mark. The second child was also assigned to the treatment ($W_2 = 1$), and again we observe $Y_2^{\text{obs}} = Y_2(1)$ (equal to 5), and we do not observe $Y_2(0)$ (represented again by a question mark). Table 5.3 directly shows the fundamental problem of causal inference: many of the potential outcomes (in this particular case exactly half) are missing.

Using this subset of the honey data, we first calculate the p-value for the sharp null hypothesis that the treatment had absolutely no effect on coughing outcomes, that is:

$$H_0: \quad Y_i(0) = Y_i(1) \quad \text{for } i = 1, \ldots, 6.$$

Under this null hypothesis, for each child, the missing potential outcomes, Y_i^{mis} are identical to the observed outcomes for the same child, Y_i^{obs}, or $Y_i^{\text{mis}} = Y_i^{\text{obs}}$ for all $i = 1, \ldots, N$. Thus, we can fill in all six of the missing entries in Table 5.3 using the observed data; Table 5.4 lists the fully expanded data set under Fisher's sharp null hypothesis. This step is the first key insight of the FEP approach; under the sharp null hypothesis, all the missing values can be inferred from the observed ones.

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

Table 5.4. ***Cough Frequency for the First Six Units from Honey Study with Missing Potential Outcomes in Parentheses Filled in under the Null Hypothesis of No Effect of the Treatment***

Unit	Potential Outcomes					
	Cough Frequency (cfa)		Observed Variables			
	$Y_i(0)$	$Y_i(1)$	Treatment	X_i	Y_i^{obs}	rank(Y_i^{obs})
1	(3)	3	1	4	3	4
2	(5)	5	1	6	5	6
3	(0)	0	1	4	0	1.5
4	4	(4)	0	4	4	5
5	0	(0)	0	1	0	1.5
6	1	(1)	0	5	1	3

We use the absolute value of the difference in average outcomes by treatment status as our test statistic:

$$T(\mathbf{W}, \mathbf{Y}^{\text{obs}}) = T^{\text{dif}} = \left| \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} \right|,$$

where $\overline{Y}_{\text{t}}^{\text{obs}} = \sum_{i:W_i=1} Y_i^{\text{obs}}/N_{\text{t}}$ and $\overline{Y}_{\text{c}}^{\text{obs}} = \sum_{i:W_i=0} Y_i^{\text{obs}}/N_{\text{c}}$ are the average of the observed outcomes in the treatment and control groups, respectively, and $N_{\text{c}} = \sum_{i=1}^{N} (1 - W_i)$ and $N_{\text{t}} = \sum_{i=1}^{N} W_i$ are the number of units in the control and treatment groups respectively. This test statistic is likely to be sensitive to deviations from the null hypothesis corresponding to a constant additive effect of the treatment. For the observed data in Table 5.3, the value of the test statistic is

$$\begin{aligned} T^{\text{obs}} &= T(\mathbf{W}, \mathbf{Y}^{\text{obs}}) = |\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}| \\ &= |(Y_1^{\text{obs}} + Y_2^{\text{obs}} + Y_3^{\text{obs}})/3 - (Y_4^{\text{obs}} + Y_5^{\text{obs}} + Y_6^{\text{obs}})/3| = |8/3 - 5/3| = 1.00. \end{aligned}$$

Under the null hypothesis, we can calculate the value of this statistic under each vector of treatment assignments, $\mathbf{W}$. Suppose for example, that instead of the observed assignment vector $\mathbf{W}^{\text{obs}} = (1, 1, 1, 0, 0, 0)$, the assignment vector had been $\tilde{\mathbf{W}} = (0, 1, 1, 0, 0, 1)$. That would *not* have changed any of the values of the observed outcomes Y_i^{obs}, because under the null hypothesis, for each unit, $Y_i(0) = Y_i(1) = Y_i^{\text{obs}}$, but it *could* have changed the value of the test statistic because different units would have been assigned to the treatment and control groups. For example, under the assignment vector, $\tilde{\mathbf{W}} = (0, 1, 1, 0, 1, 0)$, the test statistic would have been $T(\tilde{\mathbf{W}}, \mathbf{Y}^{\text{obs}}) = |(Y_2^{\text{obs}} + Y_3^{\text{obs}} + Y_5^{\text{obs}})/3 - (Y_1^{\text{obs}} + Y_4^{\text{obs}} + Y_6^{\text{obs}})/3| = |6/3 - 7/3| = 0.33$, different from $T^{\text{obs}} = 1.00$. We can repeat this calculation for each possible assignment vector. Given that we have a population of six children with three assigned to treatment, there are $\binom{6}{3} = 20$ different possible assignment vectors. Table 5.5 lists all twenty possible assignment vectors for these six children. For the moment, focus on the first unit, $i = 1$. For all assignment vectors, Y_1^{obs} remains the same, but given our null hypothesis of no effect, Y_1^{obs} is associated with $Y_1(0)$ for those assignment vectors with $W_1 = 0$, and is associated

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

Table 5.5. ***Randomization Distribution for Two Statistics for the Honey Data from Table 5.3***

						Statistic: Absolute Value of Difference in Average	
W_1	W_2	W_3	W_4	W_5	W_6	Levels (Y_i)	Ranks (R_i)
0	0	0	1	1	1	−1.00	−0.67
0	0	1	0	1	1	−3.67	−3.00
0	0	1	1	0	1	−1.00	−0.67
0	0	1	1	1	0	−1.67	−1.67
0	1	0	0	1	1	−0.33	0.00
0	1	0	1	0	1	2.33	2.33
0	1	0	1	1	0	1.67	1.33
0	1	1	0	0	1	−0.33	0.00
0	1	1	0	1	0	−1.00	−1.00
0	1	1	1	0	0	1.67	1.33
1	0	0	0	1	1	−1.67	−1.33
1	0	0	1	0	1	1.00	1.00
1	0	0	1	1	0	0.33	0.00
1	0	1	0	0	1	−1.67	−1.33
1	0	1	0	1	0	−2.33	−2.33
1	0	1	1	0	0	0.33	0.00
1	1	0	0	0	1	1.67	1.67
1	1	0	0	1	0	1.00	0.67
1	1	0	1	0	0	3.67	3.00
1	1	1	0	0	0	**1.00**	**0.67**

Note: Observed values in boldface (R_i is rank(Y_i)). Data based on cough frequency for first six units from honey study.

with $Y_1(1)$ for those assignment vectors with $W_1 = 1$; likewise for the other units. Thus the value of the corresponding statistics $T(\mathbf{W}, \mathbf{Y}^{\text{obs}})$ varies with $\mathbf{W}$.

For each vector of assignments, we calculate the corresponding value of the statistic. The last row of Table 5.5 lists the actual assignment vector, corresponding to the data in Table 5.4. In this case, $T^{\text{obs}} = 1.00$; in the sample of six children, the measure of the average cough frequency for the three children who had been given honey differs by one unit of measurement from the average for the three children who had not been given any active treatment for their coughing. The other rows list the value of the statistic under the alternative values of the assignment vector for the expanded data of Table 5.4. Under random assignment, each assignment vector has prior probability $1/20$. Thus we can derive the prior probabilities for each of the twenty values of the test statistic under Fisher's null hypothesis.

Given the distribution of the test statistic, we can ask the following question: How unusual or extreme is the observed absolute average difference between children who had been given honey versus nothing (the number 1.00) assuming the null hypothesis is true? That is, how unusual is this observed difference, assuming that there is, in fact, absolutely no causal effect of giving honey on cough frequency? One way to implement

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

this calculation is to ask how likely it is, according to the randomization distribution, to observe a value of the test statistic that is as large as the one actually observed, or even larger. This calculation clearly underestimates the likelihood of the observed result because it bundles it with all rarer events. Simply counting from Table 5.5 we see that there are sixteen assignment vectors with at least a difference in absolute value of 1.00 between children in the treated and control groups, out of a set of twenty possible assignment vectors. This corresponds to a p-value of $16/20 = 0.80$ for the given combination of the sharp null hypothesis and the test statistic. Under the null hypothesis of absolutely no effect of administering honey, the observed difference could, therefore, well be due to chance. If there were no effect of giving honey at all, we could have seen an effect as large as, or larger than, the one we actually observed for eighty out of every hundred times that we randomly assigned the honey. Note that, with three children out of six receiving the treatment, the most extreme p-value that we could have for this statistic for any values of the data is $2/20 = 0.10$; if $T = t$ is a possible value for the test statistic, then t will also be the value of the test statistic obtained by using the opposite assignment vector. Hence the sample of size six is generally too small to be able to assess, with any reasonable certainty, the existence of some effect of honey versus nothing – the sample size is not sufficient to have adequate statistical power to reach any firm conclusion.

In the next three sections we go over these three steps, specifying the null hypothesis, choosing the statistic, and measuring the extremeness, in more detail and generality.

5.4 THE CHOICE OF NULL HYPOTHESIS

The first choice that arises when calculating the FEP is the choice of null hypothesis. Fisher himself only focused on what is arguably the most obvious sharp null hypothesis, that of no effect whatsoever of the active treatment:

$$H_0 : Y_i(0) = Y_i(1), \quad \text{for } i = 1, \ldots, N. \tag{5.1}$$

We need not necessarily believe such a null hypothesis, but we may wish to see how strongly the data can speak against it. Note again that this sharp null hypothesis of no effect whatsoever is very different from the null hypothesis that the *average* effect of the treatment in the sample of N units is zero. This "average null" hypothesis is *not* a sharp null hypothesis, because it does not allow the researcher to infer values for all potential outcomes in the sample. The "average null" therefore does not fit into the framework that originates with Fisher, or its direct extensions. This does not imply that the average null hypothesis is less relevant than the hypothesis that the treatment effect is zero for all units. As we will see in Chapter 6, Neyman, whose approach focused on estimating the average effect of the treatment, was criticized, perhaps unfairly, by Fisher for his (Neyman's) questioning of the relative importance of the sharp null of absolutely no effect that was the focus of Fisher's analysis, compared to the null hypothesis of no average effect.

Although Fisher's approach cannot accommodate a null hypothesis of an average treatment effect of zero, it can accommodate sharp null hypotheses other than the null

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

hypothesis of no effect whatsoever. Fisher did not actually take this step, but it is a natural one. An obvious alternative to the null hypothesis of no effect whatsoever, is the hypothesis that there is a constant additive treatment effect, $Y_i(1) = Y_i(0) + C$, possibly after some transformation of the outcomes, (e.g., by taking logarithms, so that the null hypothesis is that $Y_i(1)/Y_i(0) = C$ for all units) for some pre-specified value C. Once we depart from the world of no effect, however, we encounter several possible complications, among them, why the treatment effect should be additive in levels rather than in logarithms, or after some other transformation of the basic outcome.

The most general case that fits into the FEP framework is the null hypothesis that $Y_i(1) = Y_i(0) + C_i$ for some set of pre-specified treatment effects C_i for $i = 1, \ldots, N$. In practice, however, it is rare to have a meaningful and interesting null hypothesis precise enough to specify individual treatment effects for each unit, without these treatment effects being identical for all units (again, possibly after some transformation).

Although the FEP approach can allow for general sharp null hypotheses, we focus in the following discussion on the implementation of the case where the null hypothesis is that of no effect whatsoever, $Y_i(1) = Y_i(0)$ for all $i = 1, \ldots, N$, thereby implying that $Y_i^{\text{mis}} = Y_i^{\text{obs}}$. This limitation is without essential loss of generality.

5.5 THE CHOICE OF STATISTIC

The second decision in the FEP approach, the choice of test statistic, is typically more difficult than the choice of the null hypothesis. First let us formally define a statistic:

Definition 5.1 (Statistic)
A statistic T is a known, real-valued function $T(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X})$ of: the vector of assignments, $\mathbf{W}$; *the vector of observed outcomes,* $\mathbf{Y}^{\text{obs}}$ *(itself a function of* $\mathbf{W}$ *and the potential outcomes* $\mathbf{Y}(0)$ *and* $\mathbf{Y}(1)$*); and the matrix of pre-treatment variables,* $\mathbf{X}$.

Any statistic that satisfies this definition can be used in the FEP approach in the sense that we can calculate its exact distribution under the null hypothesis. When such a statistic is scalar and used to find a p-value, we call it a "test statistic." However, not all statistics are sensible. We also want the test statistic to have the ability to distinguish between the null hypothesis and an interesting alternative hypothesis. Using the statistical term already introduced, we want the resulting test statistic to have *power* against alternatives, that is, to be likely to have a value, when the null hypothesis is false, that would be unusually large if the null hypothesis were true. Our desire for statistical power is complicated by the fact that there may be many alternative hypotheses of interest, and it is typically difficult, or even impossible, to specify a single test statistic that has substantial power against all interesting alternatives. We therefore look for statistics that lead to tests that have power against those alternative hypotheses that are viewed as the most interesting from a substantive point of view. Let us now introduce some test statistics and then return to the question of choosing among them.

The most popular choice of test statistic, although not necessarily the most highly recommended, is the one we also used in Section 5.3, the absolute value of the difference

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

in average outcomes by treatment status:

$$T^{\text{dif}} = \left| \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} \right| = \left| \frac{\sum_{i:W_i=1} Y_i^{\text{obs}}}{N_{\text{t}}} - \frac{\sum_{i:W_i=0} Y_i^{\text{obs}}}{N_{\text{c}}} \right|. \tag{5.2}$$

This test statistic is relatively attractive if the most interesting alternative hypothesis corresponds to an additive treatment effect, and the frequency distributions of $Y_i(0)$ and $Y_i(1)$ have few outliers.

This particular test statistic, without the absolute value, also has an interpretation as an "unbiased" estimator for the average effect of the treatment under any alternative hypothesis, as we shall discuss in detail in the next chapter. However, this is somewhat coincidental and largely irrelevant here. In general, the test statistic need not have a direct interpretation in terms of estimating causal effects. Such an interpretation may be an attractive property, but it is not essential, and in this FEP approach, focusing only on such statistics can at times divert attention from generally more powerful test statistics.

Before discussing alternative statistics, we should add one note of caution. Although there are many choices for the statistic, the validity of the FEP approach and its p-value hinges on using one statistic and its p-value only. If one calculates multiple statistics and their corresponding p-values, the probability of observing at least one p-value less than a fixed value of p, say p^*, is larger than p^*. We return to this issue of multiple comparisons in Section 5.5.7.

5.5.1 Transformations

An obvious alternative to the simple difference in average outcomes by treatment status in (5.2) is to transform the outcomes before comparing average differences between treatment levels. This procedure would be an attractive option if a plausible alternative hypothesis corresponds to an additive treatment effect after such a transformation. For example, it may be interesting to consider a constant multiplicative effect of the treatment. In that case, the treatment effect would be an additive constant after taking logarithms, and so we might compare the average difference on a logarithmic scale by treatment status using the following test statistic:

$$T^{\log} = \left| \frac{\sum_{i:W_i=1} \ln(Y_i^{\text{obs}})}{N_{\text{t}}} - \frac{\sum_{i:W_i=0} \ln(Y_i^{\text{obs}})}{N_{\text{c}}} \right|. \tag{5.3}$$

Such a transformation could also be sensible if the raw data have skewed distributions, which is typically the case for positive variables such as earnings or wealth, or levels of a pathogen, and treatment effects are more likely to be multiplicative than additive, although one needs to take care in case there are units with zero values. In such a case, the test statistic based on taking the average difference, after transforming to logarithms, would likely be more powerful than the test based on the simple average difference, as we illustrate later.

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

5.5.2 Quantiles

Motivated by the same concerns that led to test statistics based on logarithms, one may be led to test statistics based on trimmed means or other "robust" estimates of location, which are not sensitive to outliers. For example, one could use the absolute value of the difference in medians in the two samples,

$$T^{\text{median}} = \left| \text{med}_{\text{t}}(Y_i^{\text{obs}}) - \text{med}_{\text{c}}(Y_i^{\text{obs}}) \right|, \tag{5.4}$$

where $\text{med}_{\text{t}}(Y_i^{\text{obs}})$ and $\text{med}_{\text{c}}(Y_i^{\text{obs}})$ are the observed sample medians of the subsamples with $W_i = 0$, $\{Y_i^{\text{obs}} : W_i = 0\}$, and $W_i = 1$, $\{Y_i^{\text{obs}} : W_i = 1\}$, respectively. Other test statistics based on robust estimates of location include the average in each subsample after trimming (i.e., deleting) the lower and upper 5% or 25% of the two subsamples. Another way of generalizing the statistic based on the difference in medians is to use differences in other quantiles:

$$T^{\text{quant}} = \left| q_{\delta,\text{t}}(Y_i^{\text{obs}}) - q_{\delta,\text{c}}(Y_i^{\text{obs}}) \right|, \tag{5.5}$$

where $q_{\delta,\text{t}}(Y_i^{\text{obs}})$ and $q_{\delta,\text{c}}(Y_i^{\text{obs}})$, for $\delta \in (0,1)$, are the δ quantiles of the empirical distribution of Y_i^{obs} in the subsample with $W_i = 0$ and $W_i = 1$ respectively, so that, $\sum_{i:W_i=0} \mathbf{1}_{Y_i^{\text{obs}} \leq q_{\delta,\text{c}}(Y_i^{\text{obs}})}/N_{\text{c}} \geq \delta$, and $\sum_{i:W_i=0} \mathbf{1}_{Y_i^{\text{obs}} < q_{\delta,\text{c}}(Y_i^{\text{obs}})}/N_{\text{c}} < \delta$. Here $\mathbf{1}_E$ is the indicator function, equal to 1 if the event E is true and equal to 0 otherwise.

5.5.3 T-Statistics

Another choice for the test statistic is the conventional t-statistic for the test of the null hypothesis of equal means, with unequal variances in the two groups,

$$T^{\text{t-stat}} = \left| \frac{\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}}{\sqrt{s_{\text{c}}^2/N_{\text{c}} + s_{\text{t}}^2/N_{\text{t}}}} \right|, \tag{5.6}$$

where $s_{\text{c}}^2 = \sum_{i:W_i=0} (Y_i^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}})^2/(N_{\text{c}} - 1)$ and $s_{\text{t}}^2 = \sum_{i:W_i=1} (Y_i^{\text{obs}} - \overline{Y}_{\text{t}}^{\text{obs}})^2/(N_{\text{t}} - 1)$. Note that, in the approach of this chapter, we do not compare this test statistic to a student-t or normal distribution. Rather, we use the randomization distribution to obtain the exact distribution of the test statistic $T^{\text{t-stat}}$ under the null hypothesis given the potential outcomes. In many cases, the conventional normal or student-t approximation may be excellent in moderate to large samples, but in small samples, and with thick-tailed or skewed distributions for the potential outcomes, these approximations can be poor, and generally there is no need to rely on them in our era of fast computing, as we illustrate in Section 5.8.

5.5.4 Rank Statistics

An important class of test statistics involves transforming the outcomes to *ranks* before considering differences by treatment status. Such a transformation is particularly attractive when the raw outcomes have a distribution with a substantial number of outliers.

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

Assuming no ties, the rank of unit i, for $i = 1, \ldots, N$, is defined as the number of units, out of the sample of size N, with an observed outcome less than or equal to Y_i^{obs}. Without ties, the rank will take on all integer values from 1 to N, with a discrete uniform distribution, irrespective of the observed potential outcomes. This transformation leads to inferences that are insensitive to outliers, without requiring consideration of which continuous transformation would lead to a well-behaved distribution of potential outcomes. Formally the basic definition of rank in the absence of ties is

$$\tilde{R}_i = \tilde{R}_i(Y_1^{\text{obs}}, \ldots, Y_N^{\text{obs}}) = \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\text{obs}} \leq Y_i^{\text{obs}}}.$$

We often subtract $(N+1)/2$ from each rank to obtain a normalized rank that has average value equal to zero in the sample:

$$\dot{R}_i = \tilde{R}_i(Y_1^{\text{obs}}, \ldots, Y_N^{\text{obs}}) - \frac{N+1}{2} = \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\text{obs}} \leq Y_i^{\text{obs}}} - \frac{N+1}{2}.$$

When there are ties in outcomes within the sample, the definition is typically modified, for instance, by averaging all possible ranks across the tied observations. Suppose we have two units with outcomes both equal to y; if there are L units with outcomes smaller than y, the two possible ranks for these two units are $L+1$ and $L+2$. Hence we assign each of these units the average rank $(L+1)/2 + (L+2)/2 = L + 3/2$. More generally, if there are M observations with the same outcome value, and L observations with a strictly smaller value, the rank for the M observations with the same outcome value is $L + (1+M)/2$. Formally, after again subtracting the mean rank, we use the following definition for the normalized rank:

$$R_i = R_i(Y_1^{\text{obs}}, \ldots, Y_N^{\text{obs}}) = \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\text{obs}} < Y_i^{\text{obs}}} + \frac{1}{2}\left(1 + \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\text{obs}} = Y_i^{\text{obs}}}\right) - \frac{N+1}{2}.$$

Given the N ranks R_i, $i = 1, \ldots, N$, an obvious test statistic is the absolute value of the difference in average ranks for treated and control units:

$$T^{\text{rank}} = \left|\overline{R}_{\text{t}} - \overline{R}_{\text{c}}\right| = \left|\frac{\sum_{i:W_i=1} R_i}{N_{\text{t}}} - \frac{\sum_{i:W_i=0} R_i}{N_{\text{c}}}\right|, \tag{5.7}$$

where $\overline{R}_{\text{t}}$ and $\overline{R}_{\text{c}}$ are the average rank in the treatment and control group respectively. In the absence of ties, the p-value for this test statistic is closely related to that based on the Wilcoxon rank sum test statistic, which is defined as $T^{\text{wilcoxon}} = \sum_{i=1}^{N} \tilde{R}_i$, because T^{rank} is a simple transformation of T^{wilcoxon}:

$$T^{\text{rank}} = \left|\frac{T^{\text{wilcoxon}} - N(N+1)/2}{N_{\text{t}}} - \frac{N(N-1)/2 - T^{\text{wilcoxon}}}{N_{\text{c}}}\right|.$$

Let us return to the first six units from the honey data in Table 5.3. The observed cough frequency for the first child is 3. There are three units with a smaller value for the outcome, so the rank for the first child's value of the outcome is 4. The second child

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

has an observed outcome equal to 5, which is the largest observed value, so the rank for this child's value is 6. The cough frequency for the third child is zero, tied for the smallest value with one other child, so that the non-normalized rank is $(1 + 2)/2 = 1.5$. The ranks for all six units are reported in Table 5.4. We then calculate the test statistic as the average difference in rank between the three treated and the three control units, which leads to a test statistic of 0.67. To obtain the FEP for this test statistic, we count the number of times we get a test statistic equal to, or larger than, 0.67, across all randomized allocations. With all values reported in Table 5.5, this number is 16, so that the p-value is $16/20 = 0.80$.

Unlike the simple difference in means, or the difference in logarithms, the rank-based statistics do not have a direct interpretation as a meaningful treatment effect. Nevertheless, rank-based statistics can in practice lead to more powerful tests than statistics that have an interpretation as an estimated causal effect, due to their insensitivity to thick-tailed or skewed distributions. We will illustrate this feature when we look at an example with real data.

5.5.5 Model-Based Statistics

A rich class of possible test statistics with a form very different from a simple difference of averages outcomes, possibly after some transformation, is motivated by parametric models of the potential outcomes. Other uses of such models will be discussed in greater detail in Chapter 8. Here we briefly discuss their role in motivating statistics in the FEP approach.

Suppose we have two models, one for the distribution of the potential control outcomes $Y_i(0)$ and the other for the distribution of the potential treated outcomes $Y_i(1)$, governed by unknown parameters θ_c and θ_t respectively, where both θ_c and θ_t generally are vectors. For ease of exposition, let us assume that both models have a common functional form so that θ_c and θ_t have the same number of components. Let us estimate θ_c using the observed outcomes from the units assigned to the control group and denote the estimator by $\hat{\theta}_c$. We can use a variety of methods for estimation here, for example, method of moments, least squares, or maximum likelihood estimation. Similarly, let us estimate the parameter θ_t using outcomes from the units assigned to the treatment group, with estimator $\hat{\theta}_t$. Now, take any scalar function of the resulting estimates, say the difference in one of the components of the two vectors $\hat{\theta}_c$ and $\hat{\theta}_t$, or the sum of the squared differences between elements of the vectors $\hat{\theta}_c$ and $\hat{\theta}_t$. Because $\hat{\theta}_c$ and $\hat{\theta}_t$ are functions of the observed data $(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X})$, they are statistics according to Definition 5.1. Hence any scalar function of the estimated parameters $\hat{\theta}_c$ and $\hat{\theta}_t$ is a test statistic that can be used to obtain a p-value for a sharp null hypothesis.

Although these test statistics are motivated by statistical models, the validity of an FEP based on any one of them does not rely on the validity of these models. In fact, these models are purely descriptive given that the potential outcomes are considered fixed quantities. The reason such models may be useful, however, is that they may provide good descriptive approximations to the sample distribution of the potential outcomes under some alternative hypothesis. If so, the models can suggest a test statistic that is relatively powerful against such alternatives.

Let us consider two examples. First, suppose the model for $Y_i(0)$ is normal with mean μ_c and variance σ_c^2. Similarly, suppose the model for $Y_i(1)$ is also normal but with a generally different mean μ_t and variance σ_t^2. Thus, $\theta_c = (\mu_c, \sigma_c^2)$, and $\theta_t = (\mu_t, \sigma_t^2)$. The natural estimates for μ_c and μ_t are the two subsample means by treatment status $\hat{\mu}_c = \overline{Y}_c^{\text{obs}}$ and $\hat{\mu}_t = \overline{Y}_t^{\text{obs}}$. Hence, if we use the statistic

$$T^{\text{model}} = \left|\hat{\mu}_t - \hat{\mu}_c\right| = \left|\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}\right| = T^{\text{dif}},$$

we return to the familiar territory of using the difference in averages by treatment status for the test statistic.

Second, suppose that the model for $Y_i(0)$ is a normal distribution with mean μ_c and variance σ_c^2, censored from above at C, and similarly that $Y_i(1)$ has a normal distribution with mean μ_t and variance σ_t^2, also censored from above at a known value C, so that again, $\theta_c = (\mu_c, \sigma_c^2)$, and $\theta_t = (\mu_t, \sigma_t^2)$. We can estimate the parameters μ_c, μ_t, σ_c^2, and σ_t^2 by maximum likelihood as $\hat{\mu}_{\text{ml},c}$, $\hat{\mu}_{\text{ml},t}$, $\hat{\sigma}^2_{\text{ml},c}$, and $\hat{\sigma}^2_{\text{ml},t}$ respectively, or by the method of moments. There are no analytic solutions for the maximum likelihood estimates in this case, but the FEP based on a test statistic using such estimates, for example, $T^{\text{model}} = \left|\hat{\mu}_{\text{ml},t} - \hat{\mu}_{\text{ml},c}\right|$, is still valid.

5.5.6 The Kolmogorov-Smirnov Statistic

The test statistics discussed so far focus on difference in particular features of the outcome distributions between treated and control units. Initially this was the difference in averages, and later we considered differences in averages after taking transformations of outcomes, including ranks. Focusing on a single, or even multiple, features of these distributions may lead the researcher to miss differences in other aspects. For example, suppose we focus on the difference in average outcomes by treatment status. If the true distribution for the potential outcomes given treatment is normal with mean zero and unit variance, and the true distribution for the potential outcome given no treatment is normal with the same mean, zero, but a different variance, say, two, focusing solely on the average difference will not generate extreme p-values very often, even in large samples, despite the null hypothesis not holding. Formally, the test based on the difference in averages will have little power against an alternative hypothesis with different variances. We may, therefore, be interested in test statistics that would be able to detect, given sufficiently large samples, any differences in distributions between treated and control units. An example of such a test statistic is the Kolmogorov-Smirnov statistic.

Let $\hat{F}_c(y)$ and $\hat{F}_t(y)$ be the empirical distribution functions based on units with treatment $W_i = 0$ and $W_i = 1$, respectively:

$$\hat{F}_c(y) = \frac{1}{N_c} \sum_{i:W_i=0} \mathbf{1}_{Y_i^{\text{obs}} \leq y}, \quad \text{and} \quad \hat{F}_t(y) = \frac{1}{N_t} \sum_{i:W_i=1} \mathbf{1}_{Y_i^{\text{obs}} \leq y},$$

for all $-\infty < y < \infty$. Then the Kolmogorov-Smirnov test statistic is

$$T^{\text{ks}} = \sup_y \left|\hat{F}_t(y) - \hat{F}_c(y)\right| = \max_{i=1,\ldots,N} \left|\hat{F}_t\left(Y_i^{\text{obs}}\right) - \hat{F}_c\left(Y_i^{\text{obs}}\right)\right|. \tag{5.8}$$

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

This is a more complicated test statistic than, say, the average T^{dif}. Nevertheless, because it is a scalar function of the vector of assignments and the vector of observed outcomes, it is a valid test statistic. Therefore, we use exactly the same procedure as with the simpler statistics: calculate its exact finite-sample distribution generated by the randomization and then calculate the associated exact p-value.

5.5.7 Statistics with Multiple Components

The validity of the FEP approach depends on an *a priori* (i.e., before seeing the data) commitment to a specific pair: a null hypothesis and a test statistic. The corresponding p-values are valid for each pair considered in isolation, but the p-values are not independent across pairs. Specifically, consider two possible test 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 $T^{1,\text{obs}}$ and $T^{2,\text{obs}}$. This situation may arise in a number of ways. First, it may be that there are multiple alternative hypotheses of interest. For example, under one alternative hypothesis the mean of the outcome distribution may shift (suggesting a test statistic based on the difference in means by treatment status), whereas under another alternative hypothesis the dispersion may change (suggesting a test statistic based on the ratio of sample variances by treatment status). Second, it may be that the researcher has two outcomes for each unit. In the honey study, there are, for example, measures on both cough frequency and cough severity. In that case, one statistic could be the difference in average cough frequency by treatment status and the other difference in average cough severity by treatment status. Under any sharp null hypothesis, one can calculate p-values for each of the tests, for example,

$$p_1 = \Pr(T^1 \geq T^{1,\text{obs}} | \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1), H_0) \quad \text{and} \quad p_2 = \Pr(T^2 \geq T^{2,\text{obs}} | \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1), H_0).$$

These p-values are valid for each test in isolation, but using the minimum of p_1 and p_2 as an overall p-value for the null hypothesis is not valid, nor is using the average of p_1 and p_2 for this purpose.

The simplest way to obtain a valid p-value with multiple test statistics is to combine the two (or more) test statistics into a single test statistic. One can do this directly, by defining the test statistic as a function of the two original test statistics,

$$T^{\text{comb}} = g(T^1, T^2),$$

for some scalar function $g(\cdot, \cdot)$. Choices for T^{comb} could include a (weighted) average of the two statistics, or the minimum or maximum of the two statistics. Alternatively, T^{comb} could be a function of the two p-values, for example, the minimum or the average. Because T^1 and T^2 (or p_1 and p_2) are functions of $(\mathbf{W}, \mathbf{Y}, \mathbf{X})$, it follows that T^{comb} is a function of these vectors and thus a valid scalar test statistic according to our definition. Hence, its randomization distribution can be calculated, and the corresponding p-value would equal

$$p_g = \Pr(g(T^1, T^2) \geq g(T^{1,\text{obs}}, T^{2,\text{obs}}) | \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1), H_0).$$

As an example, suppose we have for, each unit, two outcome measures, Y_{i1}^{obs} and Y_{i2}^{obs}. These may be distinct measurements (e.g., in the honey study, the cough frequency and

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

cough severity, both post-treatment), or one could be a transformation of the other. For each outcome we could calculate the statistics based on the t-statistic:

$$T^{\text{t-stat,1}} = \left| \frac{\overline{Y}^{\text{obs}}_{\text{t1}} - \overline{Y}^{\text{obs}}_{\text{c1}}}{\sqrt{s^2_{\text{c1}}/N_{\text{c}} + s^2_{\text{t1}}/N_{\text{t}}}} \right|, \quad \text{and} \quad T^{\text{t-stat,2}} = \left| \frac{\overline{Y}^{\text{obs}}_{\text{t2}} - \overline{Y}^{\text{obs}}_{\text{c2}}}{\sqrt{s^2_{\text{c2}}/N_{\text{c}} + s^2_{\text{t2}}/N_{\text{t}}}} \right|.$$

Then we could choose for our test statistic

$$T^{\text{comb}} = \max\left(T^{\text{t-stat,1}}, T^{\text{t-stat,2}}\right).$$

In this case, a slightly more natural test statistic is based on Hotelling's T^2 statistic for the difference in vector of means. For $j = 1, 2$ let $\overline{Y}^{\text{obs}}_{\text{c},j} = \sum_{i:W_i=0} Y^{\text{obs}}_{i,j}/N_{\text{c}}$ and $\overline{Y}^{\text{obs}}_{\text{t},j} = \sum_{i:W_i=1} Y^{\text{obs}}_{i,j}/N_{\text{t}}$. Then let $\hat{V}_{\text{c}}/N_{\text{c}} + \hat{V}_{\text{t}}/N_{\text{t}}$ be an estimator for the covariance matrix of $(\overline{Y}_{\text{t},1} - \overline{Y}_{\text{c},1}, \overline{Y}_{\text{t},2} - \overline{Y}_{\text{c},2})'$, where

$$\hat{V}_{\text{c}} = \frac{1}{N_{\text{c}} - 1} \sum_{i:W_i=0} \begin{pmatrix} Y^{\text{obs}}_{i,1} - \overline{Y}^{\text{obs}}_{\text{c},1} \\ Y^{\text{obs}}_{i,2} - \overline{Y}^{\text{obs}}_{\text{c},2} \end{pmatrix} \cdot \begin{pmatrix} Y^{\text{obs}}_{i,1} - \overline{Y}^{\text{obs}}_{\text{c},1} \\ Y^{\text{obs}}_{i,2} - \overline{Y}^{\text{obs}}_{\text{c},2} \end{pmatrix}',$$

and

$$\hat{V}_{\text{t}} = \frac{1}{N_{\text{t}} - 1} \sum_{i:W_i=1} \begin{pmatrix} Y^{\text{obs}}_{i,1} - \overline{Y}^{\text{obs}}_{\text{t},1} \\ Y^{\text{obs}}_{i,2} - \overline{Y}^{\text{obs}}_{\text{t},2} \end{pmatrix} \cdot \begin{pmatrix} Y^{\text{obs}}_{i,1} - \overline{Y}^{\text{obs}}_{\text{t},1} \\ Y^{\text{obs}}_{i,2} - \overline{Y}^{\text{obs}}_{\text{t},2} \end{pmatrix}'.$$

Then a natural test statistic is

$$T^{\text{Hotelling}} = \begin{pmatrix} \overline{Y}^{\text{obs}}_{\text{t},1} - \overline{Y}^{\text{obs}}_{\text{c},1} \\ \overline{Y}^{\text{obs}}_{\text{t},2} - \overline{Y}^{\text{obs}}_{\text{c},2} \end{pmatrix}' \left(\hat{V}_{\text{c}}/N_{\text{c}} + \hat{V}_{\text{t}}/N_{\text{t}} \right)^{-1} \begin{pmatrix} \overline{Y}^{\text{obs}}_{\text{t},1} - \overline{Y}^{\text{obs}}_{\text{c},1} \\ \overline{Y}^{\text{obs}}_{\text{t},2} - \overline{Y}^{\text{obs}}_{\text{c},2} \end{pmatrix}, \tag{5.9}$$

which measures the Mahalanobis squared distance between the averages in the treatment group and the control group.

5.5.8 Choosing a Test Statistic

Given the wide variety of test statistics introduced here, let us now return to the question of how to choose one among them to calculate the one valid p-value. In principle, the choice should be governed by considering both plausible alternative hypotheses and the approximate distribution of the potential outcomes under both null and alternative hypotheses. Suppose one suspects the effect of the treatment to be multiplicative; in that case, a natural test statistic for assessing the null hypothesis of no effect would be the differences in the average logarithms of the outcomes between the treatment groups. If the null hypothesis does not hold because the effect is in fact multiplicative, such a test statistic will be more sensitive to this alternative hypothesis than the simple difference in averages, thus leading to greater power in the FEP. Similarly, if we expect the treatment to increase the dispersion of the outcomes but to leave the location unchanged, we can use the difference in or ratio of estimates of measures of dispersion, such as the sample variances or the interquartile ranges, for our test statistic. If the treatment does

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

increase the dispersion but does not alter the location, such a test statistic will lead to more power when using the FEP than would a test statistic based on the difference in average outcomes by treatment status.

A second consideration concerns the distribution of the values of the observed potential outcomes. If the empirical distributions of the observed potential outcomes have some outliers, calculating average differences by treatment status may lead to an FEP with low power against an alternative that corresponds to a constant and additive treatment effect. In that case it may be possible to use a test statistic that measures the difference in the centers of the two observed potential outcome distributions, not affected by a few extreme values, such as the medians, trimmed means, ranks, or even maximum likelihood estimates of locations based on long-tailed distributions, such as the family of t-distributions. In practice, using the average difference in ranks is an attractive test statistic that has decent power in a wide range of settings.

5.6 A SMALL SIMULATION STUDY

To illustrate how the different statistics perform in a known setting, we conducted a small simulation study. The study was designed to see how much power various statistics had against different (e.g., additive versus multiplicative) alternatives under various distributions of the outcomes. Although we look here at multiple statistics, one must remember that the p-value retains its properties only for a single statistic: one cannot look at multiple p-values and choose the "best," as we discussed in Section 5.5.7.

In the basic simulation setting, the population distribution for $Y_i(0)$ is normal with mean zero and unit variance, $\mathcal{N}(0, 1)$. The treatment effect is τ for all units, so that $Y_i(1) = Y_i(0) + \tau \sim \mathcal{N}(\tau, 1)$. In each replication, we draw a random sample of size $N = 2000$ with $N_c = 1000$ assigned to the control group and $N_t = 1000$ assigned to the treatment group. We calculate p-values for the sharp null hypothesis that $Y_i(1) = Y_i(0)$ for all units. We carry out the calculations using three different test statistics. First, the absolute value of the simple difference in means for treated and controls, T^{dif} given in Equation (5.2). Second, we take the absolute value of the difference in medians T^{median} given in (5.4). Third, we take the absolute value of the difference in average ranks, T^{rank} given in (5.7). In all three cases, we calculate the p-value as the probability under the null hypothesis of getting a test statistic as large as the observed test statistic, or larger.

We repeat this process by repeatedly drawing random samples and approximating the corresponding p-values by simulation. We then compute the power of the tests for each test statistic as the proportion of p-values less than or equal to 0.10. We do this simulation for a range of values of $\tau > 0$. Figure 5.1 reports the proportions for the three different test statistics that generate p-values less than 0.1, as a function of τ. The solid line corresponds to the mean, the dashed line to the median, and the dotted line corresponds to the rank statistic. We see that the FEP-based rank and mean test statistics have similar performances, whereas the FEP based on the median has less power in this situation.

We then modify the basic data-generating process by changing the distribution of $Y_i(0)$. We add a binary random variable U_i to the normal components with $\Pr(U_i = 0) = 0.8$ and $\Pr(U_i = 5) = 0.2$, which leads to a distribution with 20% outliers. We again consider additive alternatives where $Y_i(1) = Y_i(0) + \tau$. In Figure 5.2 we present the

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

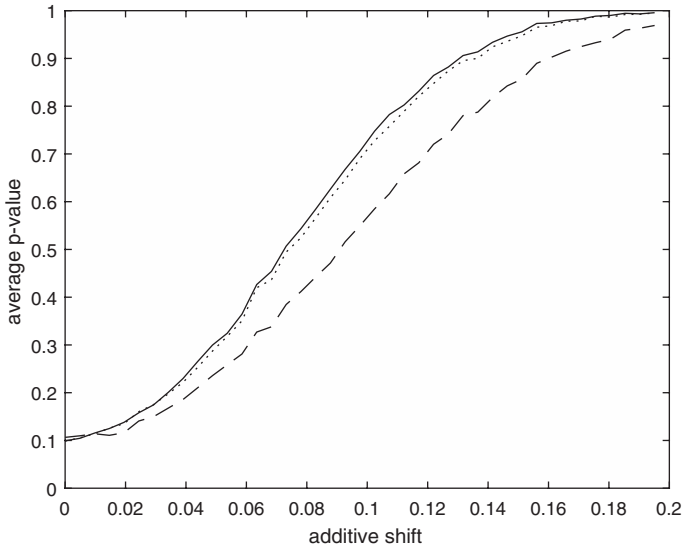

Figure 5.1. Additive model with normal outcomes T^{dif} (solid), T^{median} (dashed), T^{rank} (dotted)

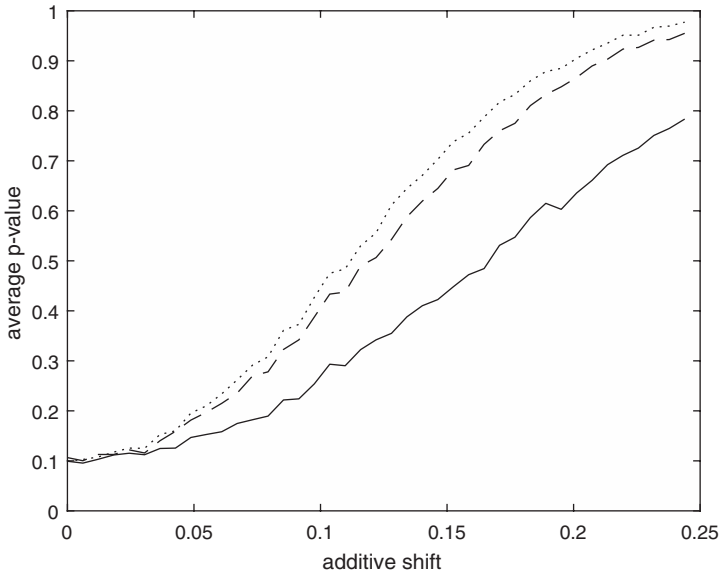

Figure 5.2. Additive model with outliers T^{dif} (solid), T^{median} (dashed), T^{rank} (dotted)

power functions for the same three statistics. The rank-based and the median-based FEP's are superior here. The mean-based FEP has substantially worse power due to the presence of outliers.

In the third part, we change the distribution of $Y_i(0)$ so that the logarithm of $Y_i(0)$ has a normal distribution with mean zero and unit variance, and make the treatment effect multiplicative: $Y_i(1) = Y_i(0) \cdot \exp(\tau)$ for a range of values of τ. Exploiting the fact that the outcomes are positive in this case, we include a test statistic based on the difference

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

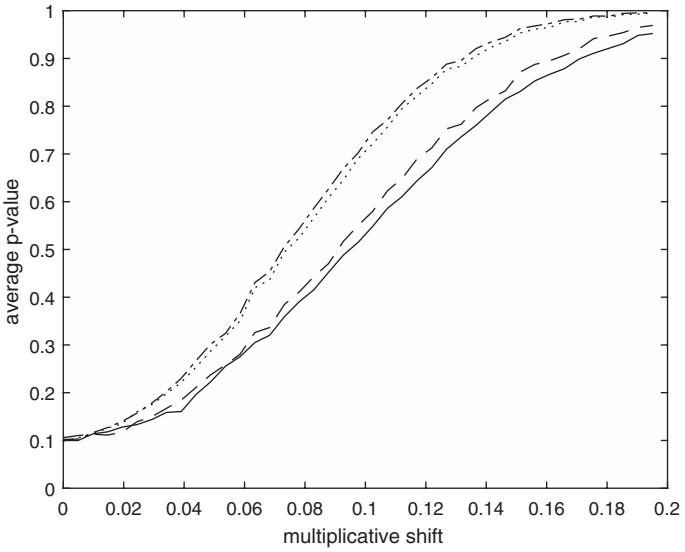

Figure 5.3. Multiplicative model T^{dif} (solid), T^{median} (dashed), T^{rank} (dotted), $T^{\log}$ (dash-dot)

in average logarithms of the basic outcome, $T^{\log}$ given in (5.3). Figure 5.3 presents the results. Again the solid line corresponds to the mean, the dashed line to the median, and the dotted line corresponds to the rank statistic, and now the dash-dot line corresponds to the statistic based on the difference in average logarithms. The logarithm-based FEP and rank-based FEP both have superior power in this case compared to the mean-based FEP and median-based FEP.

Overall, these simulations suggest that the rank-based statistic is an attractive choice in a range of settings. It has relatively good power in all three settings considered, whereas the other choices for the test statistics performed well only in settings that play to their advantages, at the expense of relatively poor power in other settings.

5.7 INTERVAL ESTIMATES BASED ON FISHER P-VALUE CALCULATIONS

Earlier we discussed how we can use FEP calculations for null hypotheses other than that of absolutely no effect of the treatment, even if this was never considered in the original proposals by Fisher. Suppose, for example, we wish to assess the null hypothesis that for all units the effect of the treatment is an increase in test score equal to $C = 0.5$: $Y_i(1) = Y_i(0) + 0.5$. This assumption is itself a sharp null hypothesis and allows us to fill in all of the missing outcomes; Table 5.6 lists the full set of potential outcomes for the first six observations in the honey data set based on this null hypothesis. Given this complete knowledge, we can again calculate the randomization distribution of any test statistic and the corresponding p-value of any observed test statistic.

Let us now do this for a range of values of a postulated effect τ. The second column of Table 5.7 lists, for the full honey data set, the FEPs associated with a constant treatment

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

Table 5.6. ***First Six Observations from Data from Honey Study with Missing Data in Parentheses under the Null Hypothesis of a Constant Effect of Size 0.5. Missing Potential Outcomes in Parentheses***

Unit	Potential Outcomes		Actual Treatment	Observed Outcome
	$Y_i(0)$	$Y_i(1)$		
1	(2.5)	3.0	1	3.0
2	(4.5)	5.0	1	5.0
3	(−0.5)	0.0	1	0.0
4	4.0	(4.5)	0	4.0
5	0.0	(0.5)	0	0.0
6	1.0	(1.5)	0	1.0

Note: Data based on cough frequency for first six units from honey study.

effect, C, for $C \in \{-3, -2.75, -2.50, \ldots, 1.00\}$. Here the test statistic is the absolute value of the difference in average outcomes for treated and control units minus C, and the p-value is the proportion of draws of the assignment vector leading to a test statistic at least as large as the observed value of that test statistic. From Table 5.7 we see that, for very negative values of C ($C < -1.50$) or very positive values ($C > 0.25$), the p-value is more extreme (smaller) than 0.05. Between these values there is a region where the C-based null hypothesis leads to p-values larger than 0.05. At the lower end of the range, we find that we obtain p-values less than 0.05 with a null hypothesis of a constant additive effect of -1.5, but not a constant additive effect of -1.25. The set of values where we get p-values larger than 0.05 is $[-1.44, 0.06]$, which provides a 95% "Fisher" interval for a common additive treatment effect, in the spirit of Fisher's exact p-values.

In the third column of Table 5.7, we do the same for a rank-based test. To be clear here, let us be explicit about the calculation of the statistic and the p-value. If the null hypothesis is that the treatment effect is $Y_i(1) - Y_i(0) = C$, then we first calculate for each unit the implied value of $Y_i(0)$. For units with $W_i = 0$, we have $Y_i(0) = Y_i^{\text{obs}}$, and for units with $W_i = 1$, we have $Y_i(0) = Y_i^{\text{obs}} - C$ under the null hypothesis. Then we convert these $Y_i(0)$ to ranks R_i. Note that this rank is not the rank of Y_i^{obs}; rather it is, under the null hypothesis, the rank of $Y_i(0)$ (or, equivalently, under the null hypothesis, the rank of $Y_i(1)$). Next, we calculate the statistic as the average rank for the treated minus the average rank for the controls, $T = |\overline{R}_{\text{t}} - \overline{R}_{\text{c}}|$. Finally, we calculate the p-value for this test statistic, under the randomization distribution, as the proportion of values of the test statistic under the randomization distribution that are larger than or equal to the realized value of the test statistic. The set of values where we get p-values equal to or larger than 0.05 is $[-2.00, -0.00]$, which provides a 95% "Fisher" interval for the treatment effect.

5.8 COMPUTATION OF P-VALUES

The p-value calculations presented so far, other than those in the simulations in Section 5.6, have been exact; we have been able to calculate precisely in how many randomizations the test statistic T would be more extreme than our observed value

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

Table 5.7. ***P-Values for Tests of Constant Treatment Effects (Full Honey Data Set from Table 5.1, with Cough Frequency as Outcome)***

Hypothesized Treatment Effect	P-Value (level)	P-Value (rank)
−3.00	0.000	0.000
−2.75	0.000	0.000
−2.50	0.000	0.000
−2.25	0.000	0.000
−2.00	0.001	0.000
−1.75	0.006	0.078
−1.50	0.037	0.078
−1.44	0.050	0.078
−1.25	0.146	0.078
−1.00	0.459	0.628
−0.75	0.897	0.428
−0.50	0.604	0.428
−0.25	0.237	0.429
0.00	0.067	0.043
0.06	0.050	0.043
0.25	0.014	0.001
0.50	0.003	0.000
0.75	0.000	0.001
1.00	0.000	0.000

Note: The level statistic is the absolute value of the difference in treated and control averages minus the hypothesized value, and the p-value is based on the proportion of statistics at least as large as the observed value. The rank-based statistic is the difference in average ranks for the treated and control units, of the value of the potential outcome under the null treatment.

of T. We could do these calculations exactly because the samples were small. In general, however, with N_t units assigned to the treatment group and N_c units assigned to the control group, the number of distinct values of the assignment vector is $\binom{N_c+N_t}{N_t}$, which, as we saw in Table 4.1 in Chapter 4, can grow very quickly with N_c and N_t. With both N_c and N_t sufficiently large, it may be infeasible to calculate the test statistic for every value of the assignment vector, even with current advances in computing. This does not mean, however, that it is difficult to calculate an accurate p-value associated with a test statistic, because we can rely on numerical approximations to the p-value.

It is typically very easy to obtain an accurate approximation of the p-value associated with a specific test statistic and null hypothesis. To do this, instead of calculating the statistic for every single value of the assignment vector $\mathbf{W} \in \mathbb{W}^+$, we calculate it for only a randomly chosen subset of possible assignment vectors. Let $T^{\text{dif,obs}}$ be the observed value of the test statistic. Then, randomly draw an N-dimensional vector with N_c zeros and N_t ones from the set of possible assignment vectors. For each draw from this set,

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

Table 5.8. ***P-Values Estimated through Simulation for Honey Data from Table 5.1 for Null Hypothesis of Zero Effects***

Number of Simulations	P-Value	$\widehat{(\text{s. e.})}$
100	0.010	(0.010)
1,000	0.044	(0.006)
10,000	0.044	(0.002)
100,000	0.042	(0.001)
1,000,000	0.043	(0.000)

Note: Statistic is absolute value of difference in average ranks of treated and control cough frequencies. P-value is proportion of draws at least as large as observed statistic.

the probability of being drawn is $1 \Big/ \binom{N_c+N_t}{N_t}$. Calculate the statistic for the first draw, say $T^{\text{dif},1} = \overline{Y}_{t,1} - \overline{Y}_{c,1}$. Repeat this process $K-1$ times, in each instance drawing a new vector of assignments and calculating the statistic $T^{\text{dif},k} = \overline{Y}_{t,k} - \overline{Y}_{c,k}$, for $k = 2, \ldots, K$. We then approximate the p-value for our test statistic by the fraction of these K statistics that are as extreme as, or more extreme than, the observed value $T^{\text{dif,obs}}$,

$$\hat{p} = \frac{1}{K} \sum_{k=1}^{N} \mathbf{1}_{T^{\text{dif},k} \geq T^{\text{dif,obs}}}.$$

If we were to draw the assignment vectors without replacement, and we sampled $\binom{N_c+N_t}{N_t}$ assignment vectors, we would have calculated the statistic for all assignment vectors, and we would obtain the exact p-value. In practice, if K is large, the p-value based on a random sample will be quite accurate. For this approximation, it does not matter whether we sample with or without replacement. The latter will lead to slightly more precise p-values for modest values of K, but both will lead to accurate p-values with K large enough because each assignment vector has the same probability of being drawn with or without replacement. The accuracy of this approximation is, therefore, entirely within the researcher's control. One can determine the number of independent draws required for a given degree of accuracy. Given a true p-value of p^*, and K draws from the set of possible assignment vectors, the large-sample standard error of the p-value is $\sqrt{p^*(1-p^*)/K}$. The maximum value for the standard error is achieved at $p^* = 1/2$, in which case the standard error of the estimated p-value is $1/(2\sqrt{K})$. Hence, if we want to estimate the p-value accurately enough that its standard error is less than 0.001, it suffices to use $K = 250,000$ draws, which is computationally entirely feasible unless the calculation of the test statistic is itself tedious (which it rarely is, although it can be, for example, when the test statistic is based on a model without closed-form estimates).

To illustrate this approach, we now analyze the full data set from the Honey Study for which the summary statistics are presented in Table 5.1. Table 5.8 reports the p-value for the null hypothesis of no effect, and using for our approximated p-values, $K = 100$, $K = 1{,}000$, $K = 10{,}000$, $K = 100{,}000$, and $K = 1{,}000{,}000$. The statistic used is the absolute value of the difference between average ranks for treated and control, and the p-value

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

reported is the proportion of assignment vectors that leads to a value for the test statistic at least as large as the observed value of the test statistic.

5.9 FISHER EXACT P-VALUES WITH COVARIATES

Thus far, all of the statistics considered have ignored the presence of any pre-treatment variables. Their presence greatly expands the set of possible test statistics. Here we discuss a few additional statistics that are feasible exploiting the presence of covariates.

First, one can use the pre-treatment variables to transform the observed outcome. For instance, if the pre-treatment variable is analogous to the outcome but measured prior to assignment to treatment or control (for instance, a pre-test score), it can be useful to subtract this variable from the potential outcomes and then carry out the test on the transformed outcomes, commonly referred to as *gain scores*. Thus, define

$$Y'_i(w) = Y_i(w) - X_i,$$

for each level of the treatment w, and define the realized transformed outcome as

$$Y_i'^{,\text{obs}} = Y_i^{\text{obs}} - X_i = \begin{cases} Y'_i(0) & \text{if } W_i = 0, \\ Y'_i(1) & \text{if } W_i = 1. \end{cases}$$

Such gain scores are often used in educational research. One should resist the temptation, though, to interpret the gain $Y_i'^{,\text{obs}}$ as a causal effect of the program for a treated unit i. Such an interpretation requires that $Y_i(0)$ is equal to X_i, which is generally not warranted.

The unit-level causal effect on the modified outcome Y' is $Y'_i(1) - Y'_i(0)$. Substituting $Y'_i(w) = Y_i(w) - X_i$ shows that this causal effect is identical to the unit-level causal effect on the original outcome Y_i, $Y_i(1) - Y_i(0)$. Hence the null hypothesis that $Y_i(0) = Y_i(1)$ for all units is identical to the null hypothesis that $Y'_i(1) = Y'_i(0)$ for all units. However, the FEP based on $Y_i'^{,\text{obs}}$ generally differs from the FEP based on Y_i^{obs}. A natural test statistic, based on average differences between treated and control units, measured in terms of the transformed outcome is

$$\begin{aligned} T^{\text{gain}} &= \frac{\sum_{i:W_i=1} Y_i'^{,\text{obs}}}{N_{\text{t}}} - \frac{\sum_{i:W_i=0} Y_i'^{,\text{obs}}}{N_{\text{c}}} \qquad (5.10) \\ &= \frac{\sum_{i:W_i=1} \left(Y_i^{\text{obs}} - X_i\right)}{N_{\text{t}}} - \frac{\sum_{i:W_i=0} \left(Y_i^{\text{obs}} - X_i\right)}{N_{\text{c}}} \\ &= \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - (\overline{X}_{\text{t}} - \overline{X}_{\text{c}}), \end{aligned}$$

where $\overline{X}_{\text{c}} = \sum_{i:W_i=0} X_i/N_{\text{c}}$ and $\overline{X}_{\text{t}} = \sum_{i:W_i=1} X_i/N_{\text{t}}$ are the average value of the covariate in the control and treatment group respectively. Compare this test statistic with the statistic based on the simple difference in average outcomes, $T^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$. The difference between the two statistics is equal to the difference in pre-treatment averages by treatment group, $\overline{X}_{\text{t}} - \overline{X}_{\text{c}}$. This difference is, on average (i.e., averaged over all assignment vectors), equal to zero by the randomization, but typically it is different from zero for any particular assignment vector. The distribution of the

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

test statistic $T^{\text{gain}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - (\overline{X}_{\text{t}} - \overline{X}_{\text{c}})$ will therefore generally differ from that of $T^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$, and thus so will be the associated p-value.

An alternative transformation involving the pre-test score is to use the proportional change from baseline, so that

$$Y_i''(w) = \frac{Y_i(w) - X_i}{X_i}, \quad \text{for } w = 0, 1,$$

and

$$Y_i''^{,\text{obs}} = \frac{Y_i^{\text{obs}} - X_i}{X_i}.$$

Here the implicit causal effect being estimated for unit i is

$$\frac{Y_i(1) - X_i}{X_i} - \frac{Y_i(0) - X_i}{X_i} = \frac{Y_i(1) - Y_i(0)}{X_i}.$$

A natural test statistic is now

$$T^{\text{prop-change}} = \overline{Y''}_{\text{t}} - \overline{Y''}_{\text{c}} = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} \frac{Y_i^{\text{obs}} - X_i}{X_i} - \frac{1}{N_{\text{c}}} \sum_{i:W_i=0} \frac{Y_i^{\text{obs}} - X_i}{X_i}. \tag{5.11}$$

Both the gain score and the proportional change from baseline statistics are likely to lead to more powerful tests if the covariate X_i is a good proxy for $Y_i(0)$. Such a situation often arises if the covariate is a lagged value of the outcome, for example, a pre-test score in an educational testing example, or lagged earnings in a job-training example.

Both T^{gain} and $T^{\text{prop-change}}$ use the covariates in a very specific way: transforming the original outcome using a known, pre-specified function. Such transformations make sense if one has a clear prior notion about the relationship between the potential outcomes and the covariate. Often, however, one may think that the covariate is highly correlated with the potential outcomes, but their scales may be different, for example, if X_i is a health index and Y_i is post-randomization medical complications for unit i. In that case, it is useful to consider a more general way to exploit the presence of covariates.

Recall that any scalar function $T = T(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X})$ can be used in the FEP framework. One possibility is to calculate a more complicated transformation that involves the values of both outcomes and pre-treatment variables for all units. For instance, let $(\hat{\beta}_0, \hat{\beta}_X, \hat{\beta}_W)$ be the least squares coefficients in a regression of Y_i^{obs} on a constant, X_i, and W_i:

$$\left(\hat{\beta}_0, \hat{\beta}_X, \hat{\beta}_W\right) = \arg \min_{\beta_0, \beta_X, \beta_W} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \beta_0 - \beta_X \cdot X_i - \beta_W \cdot W_i\right)^2.$$

These least squares coefficients are obviously functions of $(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X})$. An alternative choice for the test statistic is then

$$T^{\text{reg-coef}} = \hat{\beta}_W. \tag{5.12}$$

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

Table 5.9. ***P-Values for Honey Data from Table 5.1, for Null Hypothesis of Zero Effects Using Various Statistics***

Test Statistic	Statistic	P-Value
T^{dif}	−0.697	0.067
T^{quant} $(\delta = 0.25)$	−1.000	0.440
T^{quant} $(\delta = 0.50)$	−1.000	0.637
T^{quant} $(\delta = 0.75)$	−1.000	0.576
$T^{\text{t-stat}}$	−1.869	0.065
T^{rank}	−9.785	0.043
T^{ks}	0.304	0.021
$T^{\text{Hotelling}}$	3.499	0.182
T^{gain}	−0.967	0.006
$T^{\text{reg-coef}}$	−0.911	0.008

Note: Outcome is cough frequency (`cfa`) with the exception of $T^{\text{Hotelling}}$, which is based on cough frequency and cough severity (`cfa` and `csa`). The p-value is proportion of draws at least as large as observed statistic.

This statistic is likely to be more powerful than those based on simple differences in observed outcomes if the covariates are powerful predictors of the potential outcomes.

As before, the validity of a test based on only one such statistic does not rely on the regression model being correctly specified. However, the increases in power will be especially realized when the model provides a reasonable approximation to the distribution of values of the potential outcomes in both treatment conditions.

5.10 FISHER EXACT P-VALUES FOR THE HONEY DATA

Now we return to the full honey data set with all seventy-two observations. Table 5.9 lists ten test statistics and corresponding p-values, with the p-values estimated using 1,000,000 draws from the randomization distribution. The p-values are based on the post-treatment cough frequency (`cfa`) and the post-treatment cough severity (`csa`). Again, here we report multiple p-values, although, in theory, only one is valid, the one specified *a priori*, and in practice, one should do only one, or adjust the p-values as discussed in Section 5.5.7.

First we report the p-values when the statistic is the absolute value of the simple difference in average cough frequency by treatment status, $T^{\text{dif}} = |\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}|$. This leads to a p-value of 0.067. Next we report three quantile-based statistics, T^{quant} given in (5.5), for the quartiles $\delta = 0.25$, $\delta = 0.5$, and $\delta = 0.75$. Note that, due to the discrete nature of the outcome variable used here, cough frequency after the treatment, the observed values of the statistic are the same for all three choices of δ, although the implied p-values differ. The quantile-based p-values are considerably higher compared to those based on the difference-in-means statistic, illustrating that with discrete outcomes, quantile-based statistics can have low statistical power. Fifth, we use the conventional t-statistic, $T^{\text{t-stat}}$ given in (5.6). The p-value for this test is similar to that for the simple difference in

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

means. Note that the p-value based on the normal approximation to the distribution of this statistic is 0.062, fairly close to the p-value based on the randomization distribution because the sample size is reasonably large. Next, we use the difference in average ranks, taking account of ties, using the statistic T^{rank} given in (5.7). This leads to a smaller p-value, equal to 0.042. Then we use the Kolmogorov-Smirnov-based test statistic, given in (5.8). The maximum difference observed between the cumulative distribution functions is 0.304. As can be seen from Table 5.2, this maximum difference occurs at $y = 2$, where $\hat{F}_{\text{t}}(2) = 0.63$ and $\hat{F}_{e}(2) = 0.32$. The p-value using the Kolmogorov-Smirnov-based statistic is 0.021.

The eighth p-value uses both outcomes, cough frequency and cough severity. The test statistic is based on Hotelling's T^2 statistic, $T^{\text{Hotelling}}$ in (5.9). The last two p-values involve the pre-treatment variable `cfp`. First we calculate the statistic based on the absolute value of the difference in gains scores, T^{gain}, as given in (5.10). The last test uses the estimated regression coefficient as the test statistic, $T^{\text{reg}-\text{coef}}$, as given in (5.12). Both lead to substantially lower p-values than the statistics that do not exploit the pre-treatment variables. This reflects the strong correlation between the prior cough frequency and *ex post* cough frequency (the unconditional correlation is 0.41 in the full sample).

5.11 CONCLUSION

The FEP approach is an excellent one for simple situations when one is willing to assess the premise of a sharp null hypothesis. It is also a very useful starting point, prior to any more sophisticated analysis, to investigate whether a treatment does indeed have some effect on outcomes of interest. For this purpose, an attractive approach is to use the test statistic equal to the absolute value of the difference in average ranks by treatment status, and to calculate the p-value as the probability, under the null hypothesis of absolutely no effect of the treatment, of the test statistic being as large as, or larger than, the realized value of the test statistic. In most situations, however, researchers are not solely interested in obtaining p-values for sharp null hypotheses. Simply being confident that there is some effect of the treatment for some units is not sufficient to inform policy decisions. Instead researchers often wish to obtain estimates of the average treatment effect without being concerned about variation in the effects. In such settings the FEP approach does not immediately apply. In the next chapter, we discuss a framework for inference developed by Neyman (1923) that does directly apply in such settings, at least asymptotically, while maintaining a randomization perspective.

NOTES

As stated here, what we call "Fisher interval" was not actually proposed by Fisher, but may be close to what Fisher would have called a "fiducial interval."

Extensive work on exact inference using the randomization distribution, considerably extending Fisher's work in this area, has been done by Kempthorne and in

https://doi.org/10.1017/CBO9781139025751.006 Published online by Cambridge University Press

the recent literature by Rosenbaum. See among others, Kempthorne (1952, 1955), Rosenbaum (1984a, 1988, 1989b, 2002), and Imbens and Rosenbaum (2004). Rosenbaum's work also focuses on interval estimation using randomization inference. Surveys of this work include Rosenbaum (2002, 2009). Randomization tests based on residuals from regression analyses are discussed in Gail, Tian, and Piantadosi (1988). An interesting application of randomization inference to the California recall election is presented in Ho and Imai (2006).

A Bayesian approach to the analysis of randomized experiments is developed in Rubin (1978). We will discuss a closely related model-based approach in Chapter 8. Rubin (1990a) provides a general discussion of modes of inference for causal effects, relating randomization-based inference to other modes of inference, such as those discussed in Chapters 6, 7, and 8.

The Wilcoxon rank sum test was originally developed for equal-sized treatment and control groups in Wilcoxon (1945). Generalizations were developed in Mann and Whitney (1947); see also Lehman (1975) and Rosenbaum (2000).