Chapter 5

Donald B. Rubin

Yau Mathematical Sciences Center, Tsinghua University

June 30, 2021

Introduction

Fisher (1935) suggested a test of no effect of the active versus control treatment under the *sharp null hypothesis* or *exact null hypothesis* under a complete RCT.

Under the 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—there are no missing data for the units in this experiment.

The setup enabled Fisher to develop methods for calculating "p-values". We refer to them as *Fisher Exact P-values* (FEPs).

The Fisher's null hypothesis is distinct from the *typical*, question of an average treatment effect across all units being zero (see Chapter 6).

The latter is a weaker hypothesis, because the average treatment effect may be zero even if there exists unit specif effects.

Introduction

Consider any test statistic T—a function of $\mathbf{W}$, $\mathbf{Y}^{\mathrm{obs}}$, and any potential $\mathbf{X}$.

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 $\mathbf{W}$. The statistic determined by the randomization is referred to as the *randomization distribution* of the test statistic T.

Using this distribution, we can compare the actually observed value of the test statistic, T^{obs}, against the randomization distribution distribution of T.

A T^{obs} that is "very unlikely,"given the randomization distribution distribution of T will be taken as evidence against the null hypothesis. That is, a stochastic version of a "proof by contradiction."

Introduction

The FEP approach entails two steps:

1. the choice of a sharp null hypothesis
2. 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 choosen to have, what is now commonly referred to as, *statistical power* against a scientifically interesting alternative hypothesis.

Introduction

An important characteristic of this approach is that it is truly nonparametric.

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.

This freedom from reliance on modelling 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 (that is, the statistical power of the test).

The Honey Study

The data used to illustrate this approach 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

1. a single dose of buckwheat honey,
2. a single dose of honey-flavored dextromethorphan, an over-the-counter drug
3. no active treatment

Here we only use data on the $N = 72$ children receiving buckweat honey ($N_t = 35$) or no active treatment ($N_c = 37$).

We focus on two, of in total six outcomes: cough frequency afterwards (`cfa`), and cough severity afterwards (`csa`). Both measured on a scale from zero ("not at all frequent/severe") to six ("extremely frequent/severe").

The Honey Study

We also use two covariates, measured on the night prior to the randomized assignment: cough frequency prior (`cfp`) and cough severity prior (`cfp`).

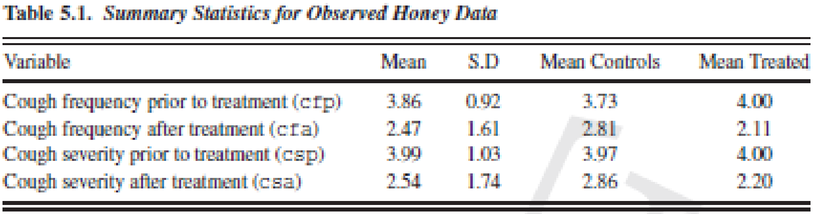


Table 5.1. ***Summary Statistics for Observed Honey Data***

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

The Honey Study

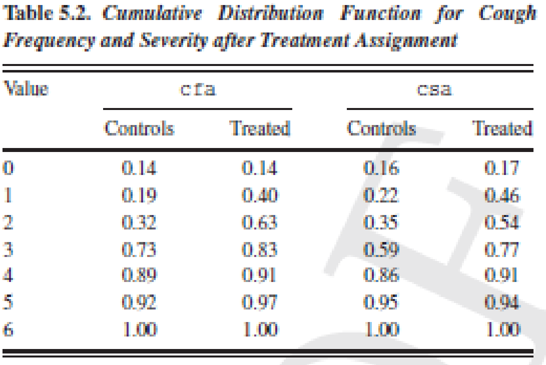


Table 5.2. ***Cumulative Distribution Function for Cough Frequency and Severity after Treatment Assignment***

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

A Simple Example with Six Units

Initially we consider a subsample from the honey data set, with six children.

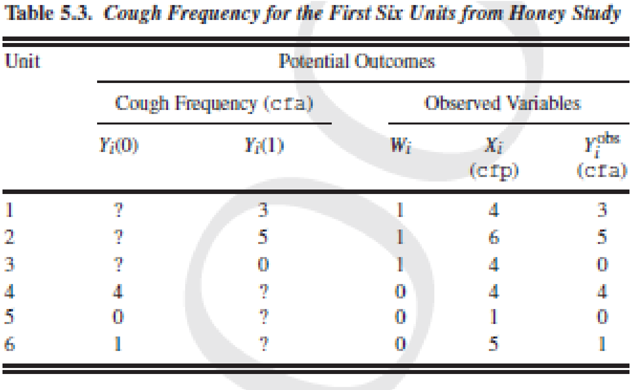


Table 5.3. *Cough Frequency for the First Six Units from Honey Study*

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

A Simple Example with Six Units

The sharp null hypothesis that the treatment had absolutely no effect on coughing otucomes, that is:

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

Under this null hypothesis, for each child, the missing potential outcomes, Y_i^{mis} are equal to the observed outcomes for the same child, Y_i^{obs}, or $Y_i^{\mathrm{mis}} = Y_i^{\mathrm{obs}}$ for all $i = 1, \ldots, N$.

This means that we can fill in all six of the missing entries in Table 5.3.

A Simple Example with Six Units

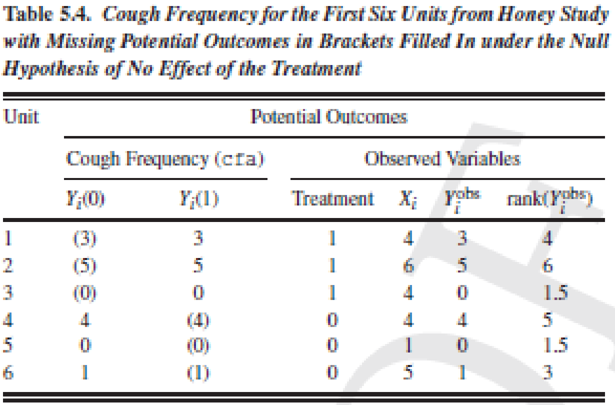


Table 5.4. ***Cough Frequency for the First Six Units from Honey Study with Missing Potential Outcomes in Brackets Filled In under the Null Hypothesis of No Effect of the Treatment***

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

A Simple Example with Six Units

To estimate the FEPs we use the test statistic:

$$T(\mathbf{W}, \mathbf{Y}^{\mathrm{obs}}) = T^{\mathrm{avg}} = \left| \overline{Y}_t^{\mathrm{obs}} - \overline{Y}_c^{\mathrm{obs}} \right|,$$

where $\overline{Y}_t^{\mathrm{obs}} = \sum_{i:W_i=1} Y_i^{\mathrm{obs}} / N_t$ and $\overline{Y}_c^{\mathrm{obs}} = \sum_{i:W_i=0} Y_i^{\mathrm{obs}} / N_c$, and $N_c = \sum_{i=1}^{N} (1 - W_i)$ and $N_t = \sum_{i=1}^{N} W_i$.

This test statistic is likely to be sensitive to deviations from the null hypothesis corresponding to a constant additive effect of the treatment.

A Simple Example with Six Units

The value of the test statistic is

$$T^{\rm obs} = T(\mathbf{W}, \mathbf{Y}^{\rm obs}) = |\overline{Y}_t^{\rm obs} - \overline{Y}_c^{\rm obs}|$$

$$= |(Y_1^{\rm obs} + Y_2^{\rm obs} + Y_3^{\rm obs})/3 - (Y_4^{\rm obs} + Y_5^{\rm obs} + Y_6^{\rm obs})/3| = |8/3 - 5/3| = 1.00.$$

Under the null hypothesis, we can calculate the value of this statistic under each vector of treatment assignments, $\mathbf{W}$.

Suppose for example, that the assignment vector would have been $\tilde{\mathbf{W}} = (0, 1, 1, 0, 1, 0)$, then the test statistic would have been $T(\tilde{\mathbf{W}}, \mathbf{Y}^{\rm obs}) = |(Y_2^{\rm obs} + Y_3^{\rm obs} + Y_5^{\rm obs})/3 - (Y_1^{\rm obs} + Y_4^{\rm obs} + Y_6^{\rm obs})/3| = |6/3 - 7/3| = 0.33.$

Given that we have a population of six children with three assigned to treatment, there are $\begin{pmatrix} 6 \\ 3 \end{pmatrix} = 20$ different assignment vectors. Table 5.5 lists all twenty possible assignment vectors for these six children.

A Simple Example with Six Units

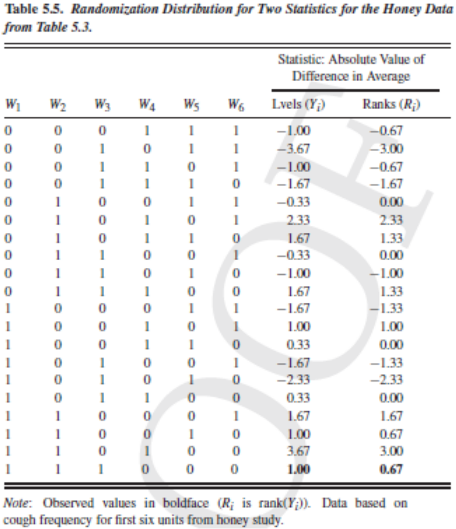


Table 5.5. ***Randomization Distribution for Two Statistics for the Honey Data from Table 5.3.***

						Statistic: Absolute Value of Difference in Average	
W_1	W_2	W_3	W_4	W_5	W_6	Lvels (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.

A Simple Example with Six Units

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 $16/20 = 0.80$. Thus, if there were no effect of giving honey at all, we could have seen an effect as large as 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 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 power to reach any firm conclusion.

The Choice of Null Hypothesis

Fisher himself only focused on what is arguably the most interesting sharp null hypothesis, that of no effect whatsoever of the treatment:

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

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.

Neyman, whose approach focused on estimating the average effect of the treatment, was critized, 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.

The Choice of Null Hypothesis

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 hypothesis of no effect whatsoever.

An obvious alternative 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 Choice of Null Hypothesis

Although the FEP approach can allow for general sharp null hypotheses, we focus on the case where the null hypothesis is that of no effect whatsoever, $Y_i(1) = Y_i(0)$ for all $i = 1, \dots, N$, hence implying that $Y_i^{\mathrm{mis}} = Y_i^{\mathrm{obs}}$.

This limitation is without essential loss of generality. If the null hypothesis is instead that the treatment effect for unit i is equal to C_i, we can transform the potential outcomes under the active treatment to $\tilde{Y}_i(1) = Y_i(1) - C_i$, and continue with testing the null hypothesis $\tilde{Y}_i(1) = Y_i(0)$.

If, under the nullhypothesis all $Y_i(0)$ and $Y_i(1)$ are known, then so are all $Y_i(0)$ and $\tilde{Y}_i(1)$.

The Choice of Statistic

The second decision, typically more difficult, is the choice of test statistic.

Definition

(Statistic)
A statistic T is a known, real-valued (scalar) 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 pretreatment variables, $\mathbf{X}$.

Any statistic that satisfies this definition is valid for use in the FEP approach.

When such a statistic is used to find p-values, we call it a "test-statistic."

However, not all such statistics are sensible. We also want the test statistic to have the ability to distinguish between the null hypothesis and an interesting alternative hypothesis, that is to have *power* against alternatives.

The Choice of Statistic

Our desire for statistical power is complicated by the fact that there may be many alternative hypotheses of interest, and it is typically difficult/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.

The most popular choice of test statistic is:

$$T^{\text{avg}} = \left| \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \right| = \left| \frac{\sum_{i:W_i=1} Y_i^{\text{obs}}}{N_t} - \frac{\sum_{i:W_i=0} Y_i^{\text{obs}}}{N_c} \right| . \tag{2}$$

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

The Choice of Statistic

Somewhat coincidental, and largely irrelevant here, T^{avg} has an interpretation as an "unbiased"estimator for the average effect of the treatment under any alternative hypothesis.

This 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.

Note: Although there are many choices for the statistic, the validity of the FEP approach and its p-values 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 is larger than p.

Transformations

An obvious alternative to (2) is to transform the outcomes before comparing average differences between treatment levels.

This procedure is attractive if a plausible alternative hypothesis corresponds to an additive treatment effect after such a transformation.

For example, if we consider a constant multiplicative effect of the treatment to be relevant, the following test statistic is relevant:

$$T^{\mathrm{log}} = \left| \frac{\sum_{i:W_i=1} \ln(Y_i^{\mathrm{obs}})}{N_t} - \frac{\sum_{i:W_i=0} \ln(Y_i^{\mathrm{obs}})}{N_c} \right| . \tag{3}$$

The Choice of Statistic

This transformation could also be sensible if the raw data have a quite skewed distribution (e.g. earnings or wealth, or levels of a pathogen and treatment effects are then also more likely to be multiplicative than additive), although one needs to take care in case there 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.

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:

$$T^{\mathrm{median}} = \left| \mathrm{med}_t(Y_i^{\mathrm{obs}}) - \mathrm{med}_c(Y_i^{\mathrm{obs}}) \right| , \tag{4}$$

where $\mathrm{med}_t(Y_i^{\mathrm{obs}})$ and $\mathrm{med}_c(Y_i^{\mathrm{obs}})$ are the observed sample medians of the subsamples with $W_i = 0$, $\{Y_i^{\mathrm{obs}} : W_i = 0\}$, and $W_i = 1$, $\{Y_i^{\mathrm{obs}} : W_i = 1\}$, respectively.

Quantiles

Other test statistics based on robust estimates of location include the average in each subsample after trimming the lower and upper, e.g. 5% of the two subsamples.

$$T^{\text{quant}} = \left| q_{\delta,t}(Y_i^{\text{obs}}) - q_{\delta,c}(Y_i^{\text{obs}}) \right|, \tag{5}$$

where $q_{\delta,t}(Y_i^{\text{obs}})$ and $q_{\delta,c}(Y_i^{\text{obs}})$, 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, $\sum_{i:W_i=0} \mathbf{1}_{Y_i^{\text{obs}} < q_{\delta,c}(Y_i^{\text{obs}})} / N_c < \delta$, and $\sum_{i:W_i=0} \mathbf{1}_{Y_i^{\text{obs}} \leq q_{\delta,c}(Y_i^{\text{obs}})} / N_c \geq \delta$, and similarity for $W_i = 1$.

Here $\mathbf{1}_E$ is the indicator function, equal to 1 if the event E is true and equal to 0 otherwise.

T-statistics

Another choice for the test statistic is

$$T^{\mathrm{t-stat}} = \left| \frac{\overline{Y}_t^{\mathrm{obs}} - \overline{Y}_c^{\mathrm{obs}}}{\sqrt{s_c^2/N_c + s_t^2/N_t}} \right|, \tag{6}$$

where $s_c^2 = \sum_{i:W_i=0}(Y_i^{\mathrm{obs}} - \overline{Y}_c^{\mathrm{obs}})^2/(N_c - 1)$ & $s_t^2 = \sum_{i:W_i=1}(Y_i^{\mathrm{obs}} - \overline{Y}_t^{\mathrm{obs}})^2/(N_t - 1)$.

Note that, here, we do not compare $T^{\mathrm{t-stat}}$ to a student t or normal distribution. Rather, we use the randomization distribution to obtain the exact distribution of $T^{\mathrm{t-stat}}$ 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.

Rank Statistics

An important class of test statistics involves transforming the outcomes to *ranks* before considering differences by treatment status.

This is particularly attractive when the raw outcomes have substantial number of outliers.

The rank of unit i, for $i = 1, \ldots, N$, is defined as the number of units 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.

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 observed potential outcomes.

Rank Statistics

Formally the basic definition of rank in the absence of ties is

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

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

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

Rank Statistics

When there are ties in outcomes 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 $(L+1)/2+(L+2)/2=L+3/2$.

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 $L+(1+M)/2$.

Rank Statistics

Formally, after again subtracting the mean rank, we use the following definition for the normalized rank:

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

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:

$$T^{\rm rank} = \left|\overline{R}_t - \overline{R}_c\right| = \left|\frac{\sum_{i:W_i=1} R_i}{N_t} - \frac{\sum_{i:W_i=0} R_i}{N_c}\right|, \qquad (7)$$

where $\overline{R}_t$ and $\overline{R}_c$ are the average rank in the treatment and control group respectively.

Rank Statistics

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}:

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

Let us return to the first six units from the honey data. The ranks for all six units are reported in Table 5.5.

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 assignment vectors (or allocations). This number is 16, why FEP$=16/20=0.80$

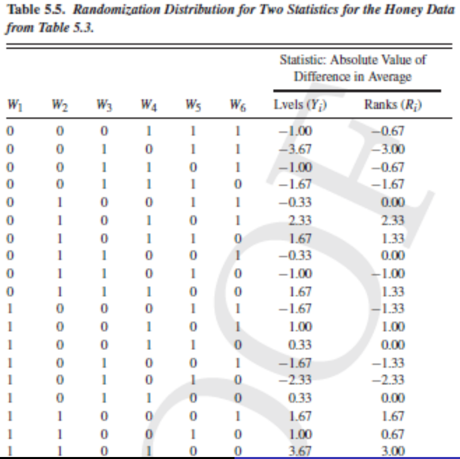


Table 5.5. ***Randomization Distribution for Two Statistics for the Honey Data from Table 5.3.***

						Statistic: Absolute Value of Difference in Average	
W_1	W_2	W_3	W_4	W_5	W_6	Lvels (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

Rank Statistics

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.

Model-based Statistics

A rich class of possible test statistics is motivated by parametric models of the potential outcomes. To be discussed extensively in chapter 8.

Here we briefly discuss their role in motivating statistics in the FEP approach.

Suppose we have two models, one for $Y_i(0)$ and the other for $Y_i(1)$, governed by unknown parameters θ_c and θ_t respectively, where both θ_c and θ_t generally are vectors.

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 and θ_t using the observed outcomes from the units assigned to the control and treatment groups, respectively.

Denote the estimators $\hat{\theta}_c$ and $\hat{\theta}_t$, respectively.

Model-based Statistics

We can use a variety of methods for estimation here, for example, method of moments (MM), least squares or maximum likelihood (ML) estimation.

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 $(\mathbf{W}, \mathbf{Y}^{\mathrm{obs}}, \mathbf{X})$, they are statistics according to Definition 1.

Hence any scalar function of the estimated parameters $\hat{\theta}_c$ and $\hat{\theta}_t$ is also a statistic that can be used to obtain a p-value for a sharp null hypothesis.

Model-based Statistics

As the models are purely descriptive given that the potential outcomes are considered fixed quantities, the validity of an FEP based on any one of them does not rely on the validity of these models.

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.

Two examples:

(1) Let $Y_i(0)$ be normal with mean μ_c and variance σ_c^2. Similarly, suppose the model for $Y_i(1)$ is also normal but with a possibly different mean μ_t and variance σ_t^2.

Model-based Statistics

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

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

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

Model-based Statistics

(2) Let $Y_i(0)$ be normal with mean μ_c and variance σ_c^2, censored from above at C, and similarly let $Y_i(1)$ be normal 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 ML as $\hat{\mu}_{\mathrm{ml},c}$, $\hat{\mu}_{\mathrm{ml},t}$, $\hat{\sigma}^2_{\mathrm{ml},c}$, and $\hat{\sigma}^2_{\mathrm{ml},t}$ respectively, or by the MM.

There are no analytic solutions for the ML in this case, but the FEP based on a test statistic using such estimates, *e.g.*, $T^{\mathrm{model}} = |\hat{\mu}_{\mathrm{ml},t} - \hat{\mu}_{\mathrm{ml},c}|$, is still valid.

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.

Focusing on a single, or even multiple, features of these distributions may lead the researcher to miss differences in other aspects.

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.

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:

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

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

$$T^{\mathrm{ks}} = T(\mathbf{W}, \mathbf{Y}^{\mathrm{obs}}) = \sup_y \left| \hat{F}_t(y) - \hat{F}_c(y) \right| = \mathrm{max}_{i=1,\ldots,N} \left| \hat{F}_t\left(Y_i^{\mathrm{obs}}\right) - \hat{F}_c\left(Y_i^{\mathrm{obs}}\right) \right|. \tag{8}$$

Because it is a function of the vector of assignments and the vector of observed outcomes, it is a valid 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.

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 of: a null hypothesis and a test statistic.

Sometimes there is an interest in testing more than one hypothesis. For instance when the researcher has more than one outcome for each unit or when testing for both mean and distribution shifts from the treatment.

Consider the honey study with 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 the other the difference in average cough severity by treatment status.

Consider the test statistics, $T^1(\mathbf{W}, \mathbf{Y}^{\mathrm{obs}}, \mathbf{X})$, and $T^2(\mathbf{W}, \mathbf{Y}^{\mathrm{obs}}, \mathbf{X})$, with realized values $T^{1,\mathrm{obs}}$ and $T^{2,\mathrm{obs}}$. For instance, testing for a mean effect on cough frequency and cough severity.

Statistics with Multiple Components

The corresponding p-values are valid for each pair considered in isolation, but the p-values are not independent across pairs.

Under any sharp null hypothesis, one can calculate p-values for each of the tests, for example,

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

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.

Statistics with Multiple Components

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:

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

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, e.g., the minimum or the average.

Statistics with Multiple Components

Because T^1 and T^2 (or p_1 and p_2) are functions of $(\mathbf{W}, \mathbf{Y}, \mathbf{X})$, it follows that $T^{\rm comb}$ 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

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

As an example, take the honey comb study. For each outcome we use:

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

Statistics with Multiple Components

Then we could choose for our test statistic

$$T^{\mathrm{comb}} = \max(|T^1|, |T^2|).$$

A more natural test statistic in this case is the Hotelling's T^2 statistic.

$$T^{\mathrm{Hotelling}} = \begin{pmatrix} \overline{Y}^{\mathrm{obs}}_{t,1} - \overline{Y}^{\mathrm{obs}}_{c,1} \\ \overline{Y}^{\mathrm{obs}}_{t,2} - \overline{Y}^{\mathrm{obs}}_{c,2} \end{pmatrix}' \left(\hat{V}_c/N_c + \hat{V}_t/N_t\right)^{-1} \begin{pmatrix} \overline{Y}^{\mathrm{obs}}_{t,1} - \overline{Y}^{\mathrm{obs}}_{c,1} \\ \overline{Y}^{\mathrm{obs}}_{t,2} - \overline{Y}^{\mathrm{obs}}_{c,2} \end{pmatrix}, \qquad (9)$$

where

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

Thus, $T^{\mathrm{Hotelling}}$ is the Mahalanobis squared distance between the averages in the treatment group and the control group.

Choosing a Test Statistic

In principle, the choice of statistic should be governed by considering plausible alternative hypotheses.

If we e.g. believe the treatment effect to be multiplicative a natural test statistic is the differences in the average logarithms of the outcomes between the treatment groups.

If the alternative hypothesis is correct the test will be more powerful than using differences in raw averages as the statistic.

If we expect the treatment to increase the dispersion of the outcomes, but leave the location unchanged and natural statistic is difference in or ratio of estimates of measures of dispersion (e.g. variance or the interquartile range), for our test statistic.

Again, if the alternative hypothesis is correct the test statistic will be more powerful then if e.g. using the difference in raw averages as our test statistic.

Choosing a Test Statistic

A second consideration concerns the distribution of the values of the potential outcomes.

If the empirical distributions of the potential outcomes have some outliers, calculating average differences by treatment status may lead to a FEP with low power against alternatives that correspond to constant additive treatment effects.

In that case, it may be possible to use test statistics that measure the difference in centers of the two potential outcome distributions, not affected by a few extreme values, such as the medians, trimmed means, ranks, or even ML 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.

A Small Simulation Study

Three data generating processes (DGPs). In the first 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.

The calculations of the FEP is based on three test statistics

$$\textbf{(i)}\ \ T^{\text{ave}} = \left|\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}\right| = \left|\frac{\sum_{i:W_i=1} Y_i^{\text{obs}}}{N_t} - \frac{\sum_{i:W_i=0} Y_i^{\text{obs}}}{N_c}\right|.$$

$$\textbf{(ii)}\ \ T^{\text{median}} = \left|\text{med}_t(Y_i^{\text{obs}}) - \text{med}_c(Y_i^{\text{obs}})\right|$$

A Small Simulation Study

$$\textbf{(iii)}\ \ T^{\text{rank}} = \left|\overline{R}_t - \overline{R}_c\right| = \left|\frac{\sum_{i:W_i=1} R_i}{N_t} - \frac{\sum_{i:W_i=0} R_i}{N_c}\right|$$

In all three cases, the p-values are calculated as the probability under the null (i.e. $Y_i(1) = Y_i(0)$ for all units) hypothesis of getting a test statistic as large as the observed test statistic, or larger.

This process is by repeated by drawing random samples and calculating the corresponding p-values. 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.

The simulation is conducted over a range of values of $\tau > 0$.

A Small Simulation Study

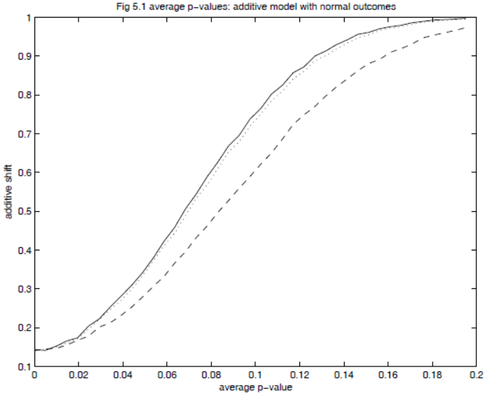


Figure: The solid line corresponds to the mean, the dashed line to the median, and the dotted line corresponds to the rank statistic

DGP two and three

In the second DGP a binary random variable U_i is added 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.

In the third DGP 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: $Y_i(1) = Y_i(0) \cdot \exp(\tau)$. Here also

$$\textbf{(iv)} \quad T^{\log} = \left| \frac{\sum_{i:W_i=1} \ln(Y_i^{\text{obs}})}{N_t} - \frac{\sum_{i:W_i=0} \ln(Y_i^{\text{obs}})}{N_c} \right|$$

is used in the calculation of the FEP.

Results second DGP

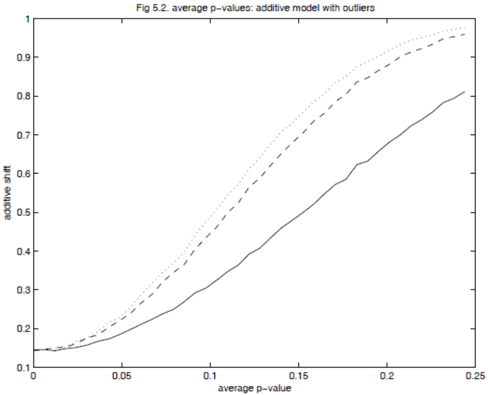


Figure: The solid line corresponds to the mean, the dashed line to the median, and the dotted line corresponds to the rank statistic

Results third DGP

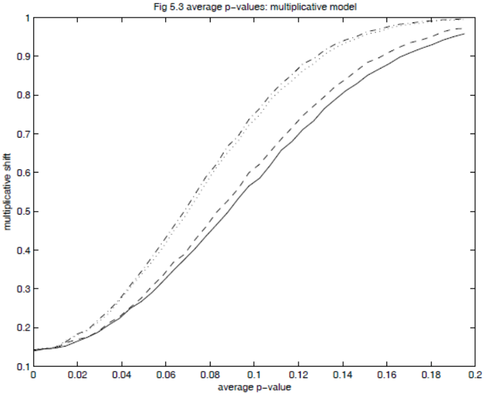


Figure: The solid line corresponds to the mean, the dashed line to the median, and the dotted line corresponds to the rank statistic, and dash-dot line corresponds to the statistic based on the difference in average logarithms

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.

If we e.g. wish to assess the null hypothesis that unit level effect of the 'treatment' on cough frequency is equal to $C = 0.5$: $Y_i(1) = Y_i(0) + 0.5$. This assumption is itself a sharp null hypothesis, and it allows us to fill in all of the missing outcomes

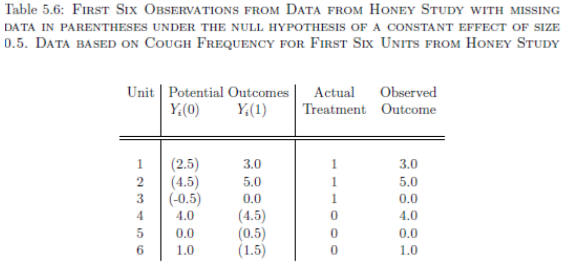


Table 5.6: First Six Observations from Data from Honey Study with missing data in parentheses under the null hypothesis of a constant effect of size 0.5. Data based on Cough Frequency for First Six Units from Honey Study

Unit	Potential Outcomes		Actual	Observed
	$Y_i(0)$	$Y_i(1)$	Treatment	Outcome
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

Interval Estimates Based on Fisher P-value Calculations

Given this complete knowledge, we can calculate the randomization distribution of any test statistic, and the corresponding p-value of any observed test statistic.

Table 5.7 display the results for the full honey data set for the FEP's associated with a constant treatment 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 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.

Interval Estimates Based on Fisher P-value Calculations

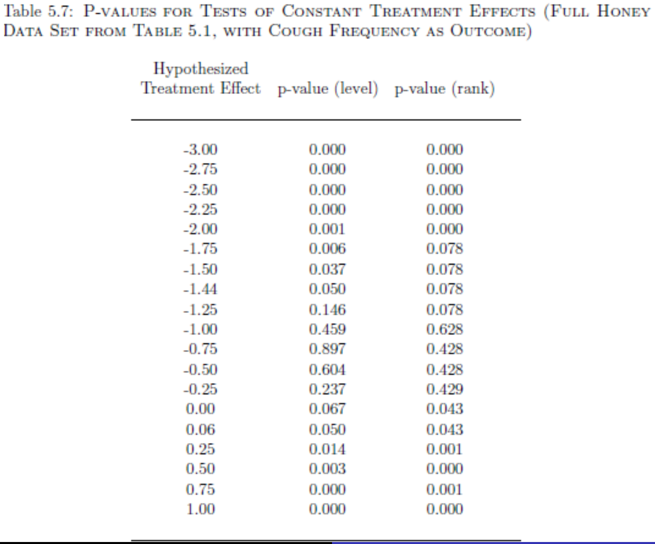


Table 5.7: P-values for Tests of Constant Treatment Effects (Full Honey Data Set from Table 5.1, with Cough Frequency as Outcome)

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

Interval Estimates Based on Fisher P-value Calculations

For $C < -1.50$ or $C > 0.25$, the p-value is more extreme (smaller) than 0.05.

The set of values where we get p-values larger than 0.05 is $[-1.44, 0.06]$, which provides a 90% "Fisher" interval for a common additive treatment effect, in the spirit of Fisher's exact p-values.

The rank-based test statistics

The statistic $T = |\overline{R}_t - \overline{R}_c|$, where *overlineR*$_t$ and $\overline{R}_c$ is the average rank of the treated and controls, respectively.

Given the null $Y_i(1) - Y_i(0) = C$ 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 $Y_i(0)$ is converted to ranks R_i.

Next, calculate $T^{obs} = |\overline{R}_t - \overline{R}_c|$.

Finally, the p-value is calculated as the proportion of values of T under the randomization distribution that are larger than or equal to T^{obs}.

The set of values where we get p-values larger than 0.05 is $[-2.00, -0.00]$, which provides a 90% "Fisher" interval for the treatment effect.

Computation of p-values

With N units, the number of distinct values of the assignment vector is

$$\mathbb{k} = \left(\begin{array}{c} N_c + N_t \\ N_t \end{array} \right).$$

With both N_c and N_t sufficiently large, it is 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.

Let $T_{\text{avg}}^{\text{obs}}$ 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, the probability of being drawn is $1/\mathbb{k}$.

Computation of p-values

Calculate the statistic for the first draw, say $T^{\mathrm{avg},1} = \overline{Y}_{t,1} - \overline{Y}_{c,1}$.

Repeat this process $K - 1$ times, in each instance drawing a new vector of assignments and calculating the statistic $T^{\mathrm{avg},k} = \overline{Y}_{t,k} - \overline{Y}_{c,k}$, 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 $T^{\mathrm{avg,obs}}$,

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

The exact p-value would have been obtained if we would have sampled all $\Bbbk$ assignment vectors (i.e. sampled without replacement). In practice, if K is large, the p-value based on a random sample will be quite accurate.

Computation of p-values

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.

Given a true p-value of p^*, the number of independent draws required for a given degree of accuracy can easily be obtained.

The large sample standard error of the p-value is $\sqrt{p^*(1-p^*)/K}$.

The maximum value for the standard error is thus $1/(2\sqrt{K})$, i.e. at $p^* = 1/2$.

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.

This may be a problem when the test statistic is based on a model without closed form.

Computation of p-values

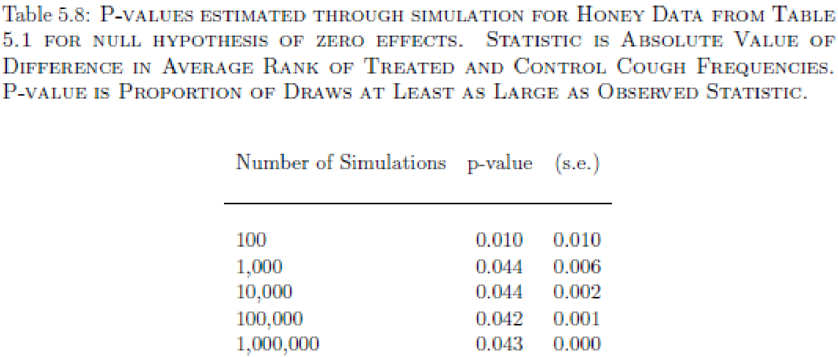


Table 5.8: P-values estimated through simulation for Honey Data from Table 5.1 for null hypothesis of zero effects. Statistic is Absolute Value of Difference in Average Rank of Treated and Control Cough Frequencies. P-value is Proportion of Draws at Least as Large as Observed Statistic.

Number of Simulations	p-value	(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

$$T^{\text{rank}} = \left|\overline{R}_t - \overline{R}_c\right|$$

Fisher Exact P-values with Covariates

First, one can use the pretreatment variables to transform the observed outcome.

For instance, if the pretreatment variable is analogous to the outcome, but measured prior to assignment to treatment or control.

Thus, define

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

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

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

Such gain scores are often used in educational research. One should resist the temptation, though, to interpret the gain $Y_i^{\prime,\text{obs}}$ 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.

Fisher Exact P-values with Covariates

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 $Y_i^{\prime,\text{obs}}$ 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

Fisher Exact P-values with Covariates

$$
\begin{aligned}
T^{\text{gain}} &= \frac{\sum_{i:W_i=1} Y_i^{\prime,\text{obs}}}{N_t} - \frac{\sum_{i:W_i=0} Y_i^{\prime,\text{obs}}}{N_c} \qquad (10)\\
&= \frac{\sum_{i:W_i=1}\left(Y_i^{\text{obs}} - X_i\right)}{N_t} - \frac{\sum_{i:W_i=0}\left(Y_i^{\text{obs}} - X_i\right)}{N_c}\\
&= \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} - (\overline{X}_t - \overline{X}_c),
\end{aligned}
$$

where $\overline{X}_c = \sum_{i:W_i=0} X_i/N_c$ and $\overline{X}_t = \sum_{i:W_i=1} X_i/N_t$ are the average value of the covariate in the control and treatment group respectively.

Fisher Exact P-values with Covariates

Compare this test statistis with the statistic based on the simple difference in average outcomes, $T^{\text{ave}} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$.

The difference between the two statistics is equal to the difference in pre-treatment averages by treatment group, $\overline{X}_t - \overline{X}_c$.

This difference is, on average (that is, 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 test statistic $T^{\text{gain}} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} - (\overline{X}_t - \overline{X}_c)$ will therefore generally differ from that of $T^{\text{ave}} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$, and thus so will be the associated p-value.

Fisher Exact P-values with Covariates

An alternative transformation involving the pre-test score is to use

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

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

$$T^{\text{prop}-\text{change}} = \overline{Y''}_t - \overline{Y''}_c = \frac{1}{N_t} \sum_{i:W_i=1} \frac{Y_i^{\text{obs}} - X_i}{X_i} - \frac{1}{N_c} \sum_{i:W_i=0} \frac{Y_i^{\text{obs}} - X_i}{X_i} \qquad (11)$$

Fisher Exact P-values with Covariates

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, e.g., a pre-test score in an educational testing example, or lagged earnings in a job training example.

Both T^{gain} and $T^{\mathrm{prop-change}}$ use the covariates in a very specific way: transforming the original outcome using a known, pre-specified function.

Most 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.

Fisher Exact P-values with 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

If the covariates are powerful predictors of the potential outcomes the test statistic

$$T^{\text{reg}-\text{coef}} = \hat{\beta}_W, \tag{12}$$

is likely to be more powerful than those based on simple differences in observed outcomes.

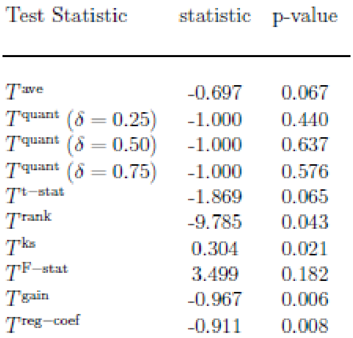
The validity of a test based on only one such statistic does not rely on the regression model being correctly specified.

Fisher Exact P-values for the Honey Data

Main outcome: cough frequency by treatment status

- $T^{\text{ave}} = |\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}|$
- T^{quant}, $\delta = 0.25$, $\delta = 0.5$, and $\delta = 0.75$
- $T^{\text{t}-\text{stat}}$
- T^{rank}
- Kolmogorov-Smirnov based test statistic. As can be seen from Table 5.2, this maximum difference occurs at $y = 2$, where $\hat{F}_{Y(1)}(2) = 0.63$ and $\hat{F}_{Y(0)}(2) = 0.32$
- $T^{\text{Hotelling}}$ based on cough frequency and cough severity.
- T^{gain} using pretreatment variable `cfp`
- $T^{\text{reg}-\text{coef}}$ pretreatment variable `cfp`

Fisher Exact P-values with Covariates



Test Statistic	statistic	p-value
T^{ave}	-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
$T^{\text{F-stat}}$	3.499	0.182
T^{gain}	-0.967	0.006
$T^{\text{reg-coef}}$	-0.911	0.008

P-values estimated using $1,000,000$ draws from the randomization distribution. Note the substantially lower p-values with the regression estimators. 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).

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 difference in average ranks by treatment status, and 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, in absolute value, the realized value of the test statistic.

In most situations, however, researchers are not solely interested in obtaining p-values for sharp null hypotheses.

Conclusion

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 shall discuss a framework for inference that does directly apply in such settings, at least asymptotically, while maintaining a randomization perspective; this was developed by Neyman (1923).

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."

Conclusion

Extensive work on exact inference using the randomization distribution, considerably extending Fisher's work in this area, has been done by Kempthorne and in the recent literature by Rosenbaum (e.g. Rosenbaum (2002, 2009)).

Randomization tests based on residuals from regression analyses are discussed in Gail, Tian, and Piantadosi (1988).

A Bayesian approach to the analysis of randomized experiments is developed in Rubin (1978). 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).