Statistical Hypothesis Tests

Kosuke Imai
Department of Politics, Princeton University

March 24, 2013

In this lecture note, we discuss the fundamentals of statistical hypothesis tests. Any statistical hypothesis test, no matter how complex it is, is based on the following logic of *stochastic proof by contradiction*. In mathematics, proof by contradiction is a proof technique where we begin by assuming the validity of a hypothesis we would like to disprove and then derive a contradiction under the same hypothesis. For example, here is a well-know proof that $\sqrt{2}$ is an irrational number, i.e., a number that cannot be expressed as a fraction.

> We begin by assuming that $\sqrt{2}$ is a rational number and therefore can be expressed as $\sqrt{2} = a/b$ where a and b are integers and their greatest common divisor is 1. Squaring both sides, we have $2 = a^2/b^2$, which implies $2b^2 = a^2$. This means that a^2 is an even number and therefore a is also even (since if a is odd, so is a^2). Since a is even, we can write $a = 2c$ for a constant c. Substituting this into $2b^2 = a^2$, we have $b^2 = 2c^2$, which implies b is also even. Therefore, a and b share a common divisor of two. However, this contradicts the assumption that the greatest common divisor of a and b is 1.

That is, we begin by assuming the hypothesis that $\sqrt{2}$ is a rational number. We then show that under this hypothesis we are led to a contradiction and therefore conclude that the hypothesis must be wrong.

Statistical hypothesis testing uses the same logic of proof by contradiction and yet there is uncertainty, hence the word "stochastic." Specifically, we can never conclude, with certainty, that a hypothesis is incorrect. Instead, we argue that the hypothesis is likely to be incorrect. Theory of statistical hypothesis testing allows us to quantify the exact level of confidence we have in this uncertain conclusion.

1 Hypothesis Tests for Randomized Experiments

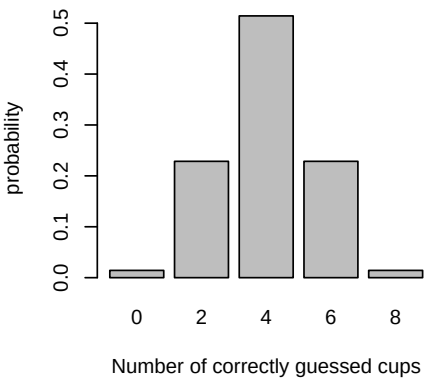
Ronald Fisher invented the idea of statistical hypothesis testing. He showed, for the first time in the human history, how one can randomize the treatment assignment and conduct a hypothesis test. Following Fisher, Neyman also developed another hypothesis testing procedure for randomized experiments. Both procedures are called *randomization tests* because they directly exploit the randomization of the treatment assignment, using it as the sole basis for inference. These tests are also design-based in that they exclusively rely on the design features of randomized experiments without necessitating the specification of a probability model.

1.1 Lady Tasting Tea

In his book, Fisher (1935) illustrated his idea with the following famous example. During an afternoon tea party at Cambridge University, England, a lady declared, "Tea tastes different depending on whether the tea was poured into the milk or whether the milk was poured into the tea." Fisher designed a randomized experiment to answer the question of whether the lady had real ability to distinguish tastes or simply was bluffing. Fisher's experimental design was as follows. Eight identical cups were prepared and four were randomly selected where the milk was poured into the tea. For the remaining four cups, the milk was poured first. The lady was then asked to identify, for each cup, whether the tea or the milk had been poured first. What happened? The lady correctly classified all the cups! Was the lady lucky? Or, does she actually possess the ability to detect the difference in tastes as she claims?

Fisher proposed to calculate the probability of observing the outcome that is at least as extreme as the outcome you actually observed under the null hypothesis and called it *p-value*. If this probability is small, then you conclude that the null hypothesis is likely to be false. The idea is that under this null hypothesis a small p-value implies a possible inconsistency between the observed data and the hypothesis formulated. In the current example, we can calculate the p-value exactly using the randomization of treatment assignment without an additional assumption. Recall our null hypothesis, which is that the lady has no ability to detect the order. Under this hypothesis, the lady would classify each cup in the exactly same way as she did in the experiment regardless of the actual order. This is illustrated in the table below where each row represents a cup and the actual order is given in the second column: "T" and "M" signify whether the tea or the milk was poured first. The third column presents the lady's guess which is the same as the actual realization, showing that she classified all cups correctly.

cups	actual	lady's guess	other scenarios				$\cdots$
1	M	M	T	T	T	T	
2	T	T	T	T	M	M	
3	T	T	T	T	M	M	
4	M	M	T	M	T	M	
5	M	M	M	M	M	M	
6	T	T	M	M	T	T	
7	T	T	M	T	M	T	
8	M	M	M	M	T	T	
correctly guessed		8	4	6	2	4	$\cdots$



Finally, the rest of columns give hypothetical scenarios showing the other possible permutations that could have resulted from randomization. For example, it could have been that the tea was poured before milk for the first four cups as shown in the fourth column. The bottom row presents the number of correctly classified cups for each hypothetical scenario under the null hypothesis that the lady cannot tell the difference and therefore would have guessed the same way for any permutation. For example, under the null hypothesis, the lady would have given the same answer regardless of the actual realization and so the number of correctly guessed cups equals four if the realization corresponds to what is given in the fourth column. The distribution of this test statistic is given above. In this case, there are $\binom{8}{4}$ or 70 ways of such hypothetical scenarios including the actual realization. Therefore, the p-value, i.e., the probability of observing the value that is at

least as extreme as what is actually observed, is 1/70. From this calculation, Fisher concluded the observed data are inconsistent with the hypothesis that the lady had no ability.

This procedure is called *Fisher's exact test* because regardless of sample size, the test computes the exact p-value for testing the null hypothesis. The test is also called a permutation test because it computes all the permutations of treatment assignments.

1.2 Statistical Hypothesis Testing Procedure

The lady tasting tea example contains all necessary elements of any statistical hypothesis testing. Specifically, the statistical hypothesis testing procedure can be summarized as the following six steps,

1. Choose a null hypothesis H_0 and its alternative H_1.
2. Choose a threshold α, the maximal probability of *Type I error* one is willing to tolerate. Type I error represents false rejection of null hypothesis when it is true.
3. Choose a test statistic, which is a function of observed data.
4. Derive a distribution of the test statistic under the null hypothesis. This distribution is often called reference distribution.
5. Compute the p-value by comparing the observed value of the test statistic against its reference distribution.
6. Reject the null hypothesis if the p-value is less than the pre-specified threshold α and retain the null hypothesis otherwise.

Let's revisit Fisher's example in terms of these six steps. In Fisher's example, the null hypothesis is given by $H_0 : Y_i(1) - Y_i(0) = 0$ for all i and an alternative is $H_1 : Y_i(1) - Y_i(0) \neq 0$ for at least some i. This null hypothesis is said to be *sharp* because the hypothesis is specified for each unit. A sharp null hypothesis is strong in that it assumes zero effect for every unit i. Contrast this with a weaker null hypothesis that the average treatment effect is zero where each unit may have non-zero treatment effect and yet the effect is zero on average across all units. For Step 2, we choose $\alpha = 0.05$ as the threshold, allowing for up to 5 percent Type I error rate. The test statistic is the number of correctly classified cups and the reference distribution was derived above. The p-value is computed as $\Pr(S \geq 8) = 1/70$ under the null hypothesis. Since the pre-specified threshold is 0.05 and $1/70 < 0.05$, we reject the null hypothesis.

Can we generalize Fisher's exact test to any number of observations? Suppose that we have a completely randomized experiment with a binary treatment (say, canvassing a voter). The total sample is n units and randomly selected n_1 units are assigned to the treatment condition and the rest is assigned to the control group. The outcome is binary (say, turnout) and the test statistic is written as $S = \sum_{i=1}^{n} T_i Y_i$ (the number of treated units who voted). Under the sharp null hypothesis that the treatment has no effect on the outcome for all units $H_0 : Y_i(1) - Y_i(0) = 0$, Fisher shows that the distribution of the statistic is the *hyper-geometric distribution* whose probability mass function is given by,

$$\Pr(S = s) \quad = \quad \frac{\binom{m}{s}\binom{n-m}{n_1-s}}{\binom{n}{n_1}}. \tag{1}$$

where $m = \sum_{i=1}^{n} Y_i$. (From this expressions, you might guess that the distribution can be approximated by the Binomial distribution $(n_1, m/n)$ when n is large and n_1/n is small, and in fact this is the case.) In Fisher's original example, we had $n = 8$, $n_1 = 4$, and $m = 4$. The number of permutation grows exponentially as n increases, making it difficult to compute the exact distribution. However, it turns out that we can analytically obtain the following exact mean and variance of this random variable,

$$\mathbb{E}(S) \;=\; \frac{n_1 m}{n}, \quad \text{and} \quad \mathbb{V}(S) = \frac{m n_1 (n - n_1)}{n(n-1)} \left(1 - \frac{m}{n}\right), \tag{2}$$

These moments can be then used to obtain an asymptotic approximation for a large n via the central limit theorem, i.e., $\{S - \mathbb{E}(S)\}/\sqrt{\mathbb{V}(S)} \xrightarrow{d} \mathcal{N}(0, 1)$. In the statistical software **R**, the exact calculation is done for a small sample size while the simulation method is used for a large sample (see the function called `fisher.test()`). Fisher's exact test was later generalized by McNemar (1947) (in psychometrics) and Mantel and Haenszel (1959) (in epidemiology) to matched-pair and stratified designs, respectively.

1.3 Testing the Population Average Treatment Effect

Now, let's consider a statistical hypothesis test about the average treatment effect under Neyman's framework we discussed earlier in the course. Suppose that our null hypothesis assumes zero *average* treatment effect, $H_0 : \mathbb{E}(Y_i(1) - Y_i(0)) = 0$. This setup contrasts with Fisher's sharp null hypothesis where each unit is assumed to have zero treatment effect. As a little digression, we note that Neyman and Fisher disagreed with each other about how the statistical hypothesis test should be conducted. In discussing Neyman *et al.* (1935), Fisher and Neyman argued against each other (see page 173),

> Dr. Neyman: [...] So long as the *average* yields of any treatments are identical, the question as to whether these treatments affect *separate* yields on *single* plots seems to be uninteresting and academic, and certainly I did not consider methods for its solution.
>
> Professor Fisher: It may be foolish, but that is what the z test was designed for, and the only purpose for which it has been used.
>
> Dr. Neyman: I am considering problems which are important from the point of view of agriculture. And from this viewpoint it is immaterial whether any two varieties react a little differently to the local differences in the soil. What is important is whether on a larger field they are able to give equal or different yields. [...]
>
> Professor Fisher: I think it is clear to everyone present that Dr. Neyman has misunderstood the intention – clearly and frequently stated – of the z test [...] It may be that the question which Dr. Neyman thinks should be answered is more important than the one I have proposed and attempted to answer. I suggest that before criticizing previous work it is always wise to give enough study to the subject to understand its purpose. Failing that it is surely quite unusual to claim to understand the purpose of previous work better than its author. [...]

Despite this disagreement, we will see that the methods proposed by Fisher and Neyman share a common feature that they take advantage of experimental design without making modeling assumptions.

Under Neyman's framework, we begin by considering the difference in means estimator,

$$\hat{\tau} \;=\; \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)Y_i \tag{3}$$

where n is the sample size and $n_1 = \sum_{i=1}^{n} T_i$ ($n_0 = n - n_1$) is the size of treatment (control) group. We assume the complete randomization of treatment assignment, which means that n_1 and n_0 are pre-specified. Previously, we showed that the asymptotic sampling distribution of this statistic is given by,

$$\sqrt{n}(\hat{\tau} - \tau) \;\xrightarrow{d}\; \mathcal{N}\left(0,\; \frac{\sigma_1^2}{k} + \frac{\sigma_0^2}{1-k}\right) \tag{4}$$

where $k = n_1/n$, $\tau = \mathbb{E}(Y_i(1) - Y_i(0))$, and $\sigma_t^2 = \mathbb{V}(Y_i(t))$ for $t = 0, 1$. Using this fact, we can easily derive the reference distribution by substituting $\tau = 0$. Furthermore, using a consistent estimate of σ_t^2 and applying the Slutsky Theorem, we have the following approximate reference distribution,

$$Z \;=\; \frac{\hat{\tau}}{\text{s.e.}} \;\xrightarrow{d}\; \mathcal{N}(0,\; 1) \quad \text{where} \quad \text{s.e.} \;=\; \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}} \tag{5}$$

Thus, we can use this z-score as a test statistic and compute the p-value based on the asymptotic reference distribution, which is the standard normal.

If the alternative hypothesis is given by $H_1 : \mathbb{E}(Y_i(1) - Y_i(0)) \neq 0$, it is called *two-sided* because it allows for the possibility that the population average treatment effect (PATE) is either negative or positive larger or smaller relative to the true value specified under the null hypothesis. In this case, we can calculate the p-value as the probability that the z-score takes the value more extreme than the observed value in terms of its absolute magnitude, i.e., $p = 2 \times \Phi(Z \geq |Z_{\text{obs}}|)$ where Z_{obs} is the observed value of the test statistic Z and $\Phi(\cdot)$ is the distribution function of the standard normal random variable. On the other hand, if we assume that the PATE never takes a negative value, then we can use the one-sided alternative hypothesis, $H_1 : \mathbb{E}(Y_i(1) - Y_i(0)) > 0$. In this case, when computing the p-value, we do not consider a large negative test statistic as an extreme value because we assume the true PATE is never negative. This means that we only need to consider a large positive value and thus the p-value is given as $p = \Phi(Z \geq Z_{\text{obs}})$. When Z_{obs} is positive, the one-sided p-value is smaller than the two-sided p-value. Thus, given the same threshold, the one-sided test is more likely to reject the null than the two-sided test.

1.4 Inverting Statistical Hypothesis Tests

As one might expect, there is a clear relationship between hypothesis tests and confidence intervals. To see this, consider the following general null hypothesis about PATE, $H_0 : \mathbb{E}(Y_i(1) - Y_i(0)) = \tau_0$ where we used $\tau_0 = 0$ in the above discussion. In this general case, the z-score statistic and its asymptotic distribution are given by,

$$Z \;=\; \frac{\hat{\tau} - \tau_0}{\text{s.e.}} \;\xrightarrow{d}\; \mathcal{N}(0,\; 1) \tag{6}$$

Now, consider the $(1-\alpha)\times 100\%$ level two-sided test where we reject the null hypothesis if the observed value of Z is greater in its absolute magnitude than the critical value $z_{1-\alpha/2}$ is defined as $\Phi(z_{1-\alpha/2}) = 1-\alpha/2$. For example, if $\alpha = 0.05$, then $z_{1-\alpha/2} \approx 1.96$. Thus, our decision rule is that we reject the null hypothesis $H_0 : \tau = \tau_0$ if and only if $|Z_{obs}| > z_{1-\alpha/2}$.

Suppose that we try different values of τ_0 and define a set of these values for which we fail to reject the null hypothesis. This set can be formally written as,

$$\begin{aligned} \mathcal{A} &= \left\{ \tau_0 : \left| \frac{\hat{\tau} - \tau_0}{\text{s.e.}} \right| \leq z_{1-\alpha/2} \right\} && (7) \\ &= \left\{ \tau_0 : \hat{\tau} - z_{1-\alpha/2} \times \text{s.e} \leq \tau_0 \leq \hat{\tau} + z_{1-\alpha/2} \times \text{s.e} \right\} && (8) \end{aligned}$$

The second equality shows that this set is equivalent to the confidence interval, establishing the one-to-one correspondence between statistical tests and confidence intervals. That is, we can *invert* a statistical test to obtain a confidence interval: if you collect all null values for which a $(1-\alpha)\times 100\%$ level test fails to reject the null hypothesis, then these values form the confidence interval of the same level.

1.5 More on Permutation Tests

Here, we discuss additional materials related to permutation tests.

Continuous outcome. Fisher's original formulation assumed that the outcome is binary, but his test can be easily extended to other types of outcome variables. For continuous outcomes, the *Wilcoxon rank-sum statistic* (Wilcoxon, 1945) is often used. Let R_i be the rank of the observed outcome variable for unit i where n observed values are ordered from the smallest to the largest and numbered from 1 to n (ties will be considered below). The rank sum statistic is given by $S = \sum_{i=1}^{n} T_i R_i$. This statistic is closely related to the *Mann and Whitney statistic* (Mann and Whitney, 1947), defined as $S - n_1(n_1+1)/2$, which, unlike the Wilcoxon rank-sum statistic, has in general an identical distribution when the treatment and control groups are switched. The exact distribution of the Wilcoxon rank-sum statistic can be obtained by enumerating all possible permutations of the treatment assignments as before. When a sample is too large to conduct this enumeration, one can use an asymptotic approximation based on the following mean and variance,

$$\mathbb{E}(S) = \frac{1}{2} n_1 (n+1), \quad \text{and} \quad \mathbb{V}(S) = \frac{1}{12} n_1 n_0 (n+1). \tag{9}$$

In **R**, the function `wilcox.test()` can be used to conduct the Wilcoxon and Mann-Whitney rank-sum tests.

When outcomes are ordered but discrete, ties may exist. In this case, it is customary to assign the average of tied ranks. For example, if four tied observations can have rank 21, rank 22, rank 23, or rank 24, then the average of these ranks, i.e., 22.5, will be assigned to all of them. The expectation is unaffected, but the variance needs to be adjusted. In particular, if we assume there exist d distinct values and m_i to represent the number of observations that take the ith smallest distinct value, then the variance is given by,

$$\mathbb{V}(S) = \frac{1}{12} n_1 n_0 (n+1) - \frac{n_0 n_1 \sum_{i=1}^{d} m_i (m_i^2 - 1)}{12 n (n-1)}. \tag{10}$$

When $m_i = 1$ for all i, this reduces to the usual variance. Also, the correction term is typically small when n is large unless the number of ties increases together with the sample size.

Inverting Fisher's exact test. We showed earlier that there is a one-to-one relationship between hypothesis tests and confidence intervals. We may wonder how to invert Fisher's exact test and obtain an exact confidence interval. To do this, consider a case where the outcome variable is not binary but rather is a continuous variable. A general null hypothesis is $H_0 : Y_i(1) - Y_i(0) = \tau_0$. We note that this sharp null hypothesis assumes the *constant additive treatment effect* model where every unit has the same additive treatment effect. Under this null hypothesis, we know the values of both potential outcomes for each unit. For example, we observe $Y_i(0)$ for a control unit and $Y_i(1)$ is unobserved. And yet under this null hypothesis, we have an imputed value of the missing potential outcome $Y_i(1) = Y_i(0) + \tau_0$. Thus, under the null hypothesis, we observe all potential outcomes and it is easy to derive the reference distribution of a test statistic. Suppose that our test statistic is the difference in sample means between the treatment and control group outcomes. Then, for every permutation of treatment assignment, we compute the value of this test statistic, yielding the reference distribution. We can compare the observed value of the test statistic against this reference distribution in order to compute the p-value and decide whether or not to reject the null hypothesis. Collecting the null values for which we fail to reject the null hypothesis at the α level, we obtain the $(1 - \alpha) \times 100\%$ confidence interval.

The problem of this procedure, however, is that the constant additive treatment effect model is too restrictive. This relates to the limitation of a sharp null hypothesis raised by Neyman in his exchange with Fisher quoted above. It is difficult to imagine that the treatment effect is constant across units in any social science experiment. Despite the fact that it allows one to compute the exact p-value regardless of sample size, this makes Fisher's exact test less attractive to practitioners where treatment effect heterogeneity is a rule rather than an exception. Is it possible to address this limitation? One possible approach is described below.

Population Inference. Lehmann (2006) shows that in certain settings, population inference can be made using the same permutation tests that were originally designed for sample inference. He shows that identical permutation methods can also lead to population inference under a particular model. Consider the following null hypothesis of no treatment effect where the outcome variable is assumed to be continuous,

$$H_0 : F_{Y(1)} \quad = \quad F_{Y(0)}, \tag{11}$$

where $F_{Y(t)}$ represents the distribution function for a potential outcome $Y_i(t)$ (independently identically distributed or i.i.d.), i.e., $Y_i(t) \overset{\text{i.i.d.}}{\sim} F_{Y(t)}$ for $t = 0, 1$. Note that unlike the sharp null hypothesis this only requires the marginal distributions of the two potential outcomes (but not necessarily their values) are identical. That is, some units can have non-zero treatment effects so long as the distributions of two potential outcomes remain the same.

Now, suppose we use the Wilcoxon's rank sum test. Then, its reference distribution is completely characterized by the probability that n_1 units in the treatment group take a certain set of ranks, $r = (r_1, r_2, \ldots, r_n)$, regardless of how each of these units is assigned to an element of vector r. Under the null hypothesis of equation (11), this probability is equal to $1/\binom{n}{n_1}$ because each subset of n_1 observations from a total of n observations is equally likely to take this particular set of ranks (and there are $\binom{n}{n_1}$ such subsets). This probability equals exactly the one we obtain for sample inference! Thus, in a completely randomized experiment, the Wilcoxon's rank sum statistic has the same exact reference distribution regardless of whether one is using a sharp null hypothesis or a population-level null hypothesis of equation (11). This means that the inference

based on the Wilcoxon's rank sum statistic is valid under either null hypothesis. However, one limitation of this approach is that it relies on the complete randomization of the treatment and the use of a specific test statistic. Also, it is not easy to deal with ties in this population inference framework, which has implications for discrete ordered outcomes.

To obtain confidence sets and point estimates in this population inference framework, the following *shift model* is often considered.

$$F_{Y(1)}(y) \;=\; F_{Y(0)}(y + \Delta), \tag{12}$$

for all y and Δ is a constant. The model says that the marginal distributions of the two potential outcomes look identical except that their locations differ by an unknown amount Δ. This difference then represents the causal quantity of interest which we want to make inferences about. Unlike quantities like $Y_i(1) - Y_i(0)$, Δ does not permit an easy interpretation. However, this setting allows for the potential existence of treatment effect heterogeneity. To construct a confidence set based on permutation tests, we invert the permutation tests by specifying the null hypothesis: $H_0 : F_{Y(1)}(y) = F_{Y(0)}(y + \Delta_0)$ for various values of Δ_0. Under this null hypothesis, adding Δ_0 to $Y_i(0)$ will give a random variable that has the same distribution as $Y_i(1)$. Thus, we can apply the above argument and use the Wilcoxon's rank sum statistic to construct a confidence set for population inference in the same manner as done in sample inference.

Finally, one common method used to obtain an point estimate of Δ in equation (12) is due to Hodges and Lehmann (1963). An intuition is that since under the shift model $Y_i(0) + \Delta$ and $Y_i(1)$ are i.i.d., we will try to find the value of Δ such that the sample distributions of these two variables are as similar to each other as possible. If there is a unique such value, that is the Hodges-Lehmann estimate. Otherwise, the interval can be identified and its mid-point will be taken as the estimate. Formally, we assume that the test statistic, $f(Y, T, \Delta)$, is a nondecreasing function of Δ and when $\Delta = 0$ the distribution of this statistic is symmetric about a fixed point μ. Then, the Hodges-Lehmann estimator is defined as,

$$\widehat{\Delta} \;=\; \frac{1}{2}\left[\sup\{\Delta : f(Y, t^{obs}, \Delta) > \mu\} + \inf\{\Delta : f(Y, t^{obs}, \Delta) < \mu\}\right]. \tag{13}$$

Now, it can be shown that the reference distribution of the Wilcoxon's rank sum statistic is symmetric about $n(n+1)/2$ (which implies that the Mann-Whitney statistics are symmetric about 0). The statistic is also nondecreasing in Δ and thus, one can immediately apply the Hodges-Lehmann estimator. Hodges and Lehmann (1963) shows that although this estimator is biased in general it is consistent and asymptotically normal.

2 Power of Statistical Hypothesis Tests

When conducting a statistical hypothesis test, we potentially make two kinds of error. First, it is possible that the null hypothesis is true but we incorrectly reject it. This error is called *Type I error*. Another possibility is that the null hypothesis is false but we fail to reject it. This is known as *Type II error*. The following table illustrates these two types of error associated with hypothesis testing.

	Reject H_0	Retain H_0
H_0 is true	**Type I error**	Correct
H_0 is false	Correct	**Type II error**

We have already seen how statistical hypothesis testing procedures control for Type I error. If the null hypothesis is true, then the probability that we falsely reject it is given by the pre-determined *level* or *size* of one's hypothesis test. The level of test corresponds to the threshold where if the p-value is less than this threshold the null hypothesis is rejected. Thus, the Type I error probability is under complete control of researchers. We can pick the level of test in order to specify the degree to which we are willing to tolerate false decisions if the null hypothesis is actually correct.

However, there is an explicit trade-off between the two kinds of error. To see this, we do the following thought exercise. What is a simple decision rule that completely eliminates Type I error? Clearly, if we never reject the null, then we never make Type I error regardless of whether or not the null hypothesis is true. However, this increases the chance that we make Type II error. The same argument applies in the opposite direction: always rejecting the null hypothesis eliminates Type II error but increases the Type I error probability. This dilemma arises because we do not know whether the null hypothesis is actually true. This illustrates the importance of considering both types of error when conducting statistical hypothesis tests. Applied researchers have a tendency to focus on Type I error while being unaware of the possibility that they may be committing Type II error. In fact, one could even argue that Type II error is more important than Type I error because we rarely believe that the null hypothesis is true. For example, the null hypothesis under Neyman's framework that the PATE is zero is unrealistic: PATE might be small but probably is not exactly zero.

2.1 Interpretation of Statistical Hypothesis Tests

Before we begin our formal analysis, the discussion above highlights the issue of how one should interpret the results of statistical hypothesis tests. In particular, the failure to reject a null hypothesis should not be interpreted as evidence indicating that the null hypothesis is true. In other words, it is possible that we may not be able to reject the null hypothesis because we are committing Type II error, not because the null hypothesis is actually true. In the context of analyzing randomized experiments under Neyman's framework, even if we fail to reject the null hypothesis that the PATE is zero, this does not necessarily mean the PATE is exactly zero. It is also possible, for example, that the PATE is not zero but we do not have a sufficient sample size to detect it. What the failure to reject the null hypothesis implies is that the observed data are consistent with the null hypothesis but the null hypothesis may not hold under the true data generating process.

A similar caution of interpretation applies to the p-value. Recall that the p-value is defined as the probability, computed under the null hypothesis, of observing a value of test statistic at least as extreme as its observed value. While a smaller p-value indicates stronger evidence against the null hypothesis, a large p-value can occur either because the null hypothesis is true or because the test is unable to reject the null even though the null is false. The p-value is also neither the probability that the null hypothesis is true nor the probability that the alternative hypothesis is false. In fact, the probability that the null hypothesis is actually true is either exactly one or zero (and we do not know which)! Finally, even if we observe the p-value that is less than the pre-specified threshold, this does not mean that the finding is scientifically significant as the result only implies the *statistical significance*. In order to assess the scientific importance of findings, one must estimate the effect size going beyond the question of whether the effect is zero or not.

2.2 Power Analysis

We now conduct a formal analysis of Type II error. Such an analysis is called *power analysis*. Formally, the power of hypothesis is defined as the probability that we reject the null hypothesis. If the null hypothesis is false, this equals one minus the probability of Type II error. Therefore, we wish to maximize the power of statistical hypothesis test while keeping its size at a pre-specified level. To calculate the power of statistical hypothesis test, we implement the following procedure:

1. Specify all elements of a statistical hypothesis test for which you wish to calculate the power. This includes a test statistic, the null hypothesis to be tested, and the level of statistical hypothesis test.

2. Specify the true data generating process, which differs from the null hypothesis (if it is the same as the true data generating process, then the power will be identical to the size of test). This is the scenario under which you wish to calculate the power.

3. Calculate the probability that you will reject the null hypothesis under the specified scenario.

The power analysis is useful because it allows one to examine the necessary sample size under a variety of data generating processes. Researchers wish to have a large enough sample size so that they can reject the null hypothesis when it is actually false. When the departure from the null hypothesis is small, a large sample size is needed to reject the null hypothesis. Thus, the power analysis asks researchers to specify the minimum degree of departure from the null hypothesis they wish to detect with a certain probability. Given this specification, researchers can calculate the minimum sample size.

The following example will clearly illustrate how power calculation is done. Suppose that we have a simple random sample of size n. For each unit, we observe a binary variable X_i (say, whether a voter supports Obama). We wish to test the null hypothesis $H_0 : p = 0.5$ where $p = \Pr(X_i = 1)$ representing the population proportion of Obama supporters. Our test statistic is the following z-score, which is asymptotically distributed as normal,

$$Z \;=\; \frac{\hat{p} - 0.5}{\sqrt{0.5 \times (1 - 0.5)}} \;=\; 2\hat{p} - 1 \;\overset{\text{approx.}}{\sim}\; \mathcal{N}(0,\; 1) \tag{14}$$

Assume that we use an $\alpha = 0.05$ two-sided test. Then, we reject the null hypothesis if and only if the observed value of Z is either greater than 1.96 or less than -1.96.

Now, consider the following scenario. We would like to calculate the probability that we successfully reject the null hypothesis when the true data generating process is based on $p = 0.6$ (i.e., the population proportion of Obama supporters is 0.6). To do this, note that the sampling distribution of Z statistic under this data generating process is,

$$\hat{p} \;\overset{\text{approx.}}{\sim}\; \mathcal{N}\left(p,\; \frac{p(1-p)}{n}\right) \tag{15}$$

where $p = 0.6$. Therefore, the sampling distribution of the test score under this data generating process is,

$$Z \;=\; 2\hat{p} - 1 \;\overset{\text{approx.}}{\sim}\; \mathcal{N}\left(2p - 1,\; \frac{4p(1-p)}{n}\right) \tag{16}$$

The power is calculated as the probability that Z takes the value greater than 1.96 or less than -1.96 under this approximate distribution. The former can be calculated as,

$$\Pr(Z > 1.96) \approx \Pr\left(\frac{Z-(2p-1)}{\sqrt{4p(1-p)/n}} > \frac{1.96-(2p-1)}{\sqrt{4p(1-p)/n}}\right) = \Phi\left(\frac{2p-0.96}{\sqrt{4p(1-p)/n}}\right) \quad (17)$$

where $\{Z-(2p-1)\}/\sqrt{4p(1-p)/n}$ is asymptotically distributed as the standard normal distribution. A similar calculation yields that the probability that Z is less than -1.96 equals $\Phi\{(-2.96-2p)/\sqrt{4p(1-p)/n}\}$.

Given this calculation of power function, we can change the sample size n and the true probability p to compute the power of hypothesis test under various settings. This will allow us to examine the sample size required for a hypothesis test to have enough power so that the departure from the null hypothesis can be detected.

2.3 Multiple Testing

The setup so far assumes that we conduct a single hypothesis test. In practice, however, researchers conduct more than one hypothesis test in a single research paper. For example, after conducting a randomized experiment, we may wish to test whether the treatment affects each of several outcome variables of interest. However, the problem of such multiple testing is the issue of false discovery. Suppose that we conduct 10 independent statistical tests with $\alpha = 0.05$ level and yet (unknown to us) all of these 10 null hypotheses are false. Then, the probability that you will find at least one statistically significant result is $1 - 0.95^{10} \approx 0.4$. That is, even when the treatment has no effect on any of the 10 outcomes, we are likely to find some statistically significant result.

There exist a large body of statistical literature on the multiple testing problem. First, we can come up with a method that controls for the *family-wise error rate* (FWER), which is the probability of making at least one Type I error. In the above example, FWER is 0.4 and we would like to lower it. The *Bonferroni correction* is a method is based on the following Bonferroni inequality,

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \Pr(A_i) \quad (18)$$

where $\{A_i\}_{i=1}^n$ is a sequence of events. We apply this inequality to multiple testing by letting n be the number of hypothesis tests and A_i be the event that we do not reject the ith null hypothesis when it is true.

$$\text{FWER} = \Pr\left(\bigcup_{i=1}^{n} \text{falsely reject } i\text{th null}\right) \leq \sum_{i=1}^{n} \Pr(\text{falsely reject } i\text{th null}) \quad (19)$$

Thus, if we wish to make sure that the FWER is at most α, then one can reject each null hypothesis when the p-value is less than α/n rather than α.

The problem of this Bonferroni correction is that it is often too conservative to be useful especially when n is large. While there are some improvements of this procedure (e.g., Holm, 1979), in a seminal article, Benjamini and Hochberg (1995) propose to consider the *false discovery rate* (FDR),

$$\text{FDR} = \mathbb{E}\left\{\frac{\text{number of false rejections}}{\max(\text{total number of rejections}, 1)}\right\} \quad (20)$$

which equals zero when there is no rejection of null hypothesis. For independent tests, Benjamini and Hochberg (1995) develop the following testing procedure that controls the FDR.

1. Order n p-values as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$.
2. Find the largest i such that $p_{(i)} \leq \alpha i/n$ and call it k.
3. Reject the null hypotheses for all the k tests associated with p-values $p_{(1)}, \ldots, p_{(k)}$.

This procedure is less conservative because $\alpha/n \leq \alpha i/n$ and yet still controls the FDR, which is easy to interpret. Benjamini and Yekutieli (2001) further prove that this procedure is valid even when tests are positively correlated and propose another procedure which works in a setting of arbitrary dependence among tests.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 1, 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**, 4, 1165–1188.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, London.

Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics* **34**, 2, 598–611.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 2, 65–70.

Lehmann, E. L. (2006). *Nonparametrics: Statistical Methods based on Ranks*. Springer, New York, Revised 1st edn.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* **18**, 1, 50–60.

Mantel, N. and Haenszel, W. (1959). Statical aspects of the analysis of data from retrospective studies of disease. *Journal of the National cancer Institute* **22**, 4, 719–748.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 2, 153–157.

Neyman, J., Iwaszkiewicz, K., and Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation (with discussion). *Supplement of Journal of the Royal Statistical Society* **2**, 107–180.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* **1**, 80–83.