ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

Statistical Inference:

Part 4, Multiple Inference

A large portion of economic research consists of testing a single or small number of a priori hypotheses. For example, we might test whether class size affects student performance, whether air pollution increases infant mortality, or whether deregulation in a particular industry affects prices. In some cases, however, a research project tests multiple hypotheses. For example, we might test whether a government program has effects on an entire range of outcomes or we might test a large number of factors that we think could determine demand for a particular good. Exercises such as these are often referred to somewhat pejoratively as "data-mining," but there is nothing inherently wrong with them as long as you use testing procedures that account for the large number of tests being conducted.

1 The Multiplicity Problem

Consider a case in which you test 10 hypotheses using 10 independent test statistics. For example, you might test whether various subgroups display positive returns to getting a PhD (white males, black males, white females, black females, etc.). Suppose that all the null hypotheses are true, i.e., PhDs do not increase earnings for any subgroup. For any given subgroup, the probability of (falsely) rejecting the null hypothesis at the $\alpha = 0.05$ level is 5%. But the probability of (falsely) rejecting at least one of the 10 null hypotheses is $1 - P(\text{Reject None}) = 1 - .95^{10} \approx 1 - 0.60 = 0.40$. Thus there is a 40% probability of experiencing at least one Type I error despite the fact that we are controlling the probability of committing a Type I error at 5% for any individual test. Put simply, if you test enough true hypotheses, eventually you are bound to reject one of them.

There are several ways to quantify the multiple inference problem. One way is to focus

on the familywise error rate (FWER). The FWER quantifies the probability of rejecting at least one true null hypothesis from a group, or family, of hypotheses being tested. Formally, consider testing a family of M hypotheses, $H_1, H_2, ..., H_M$, of which J are true ($J \leq M$). FWER is the probability that at least one of the J true hypotheses in the family is rejected. As more hypotheses are added to a family, the probability of rejecting at least one of them at a given α-level increases, and hence FWER increases. The multiple inference problem, as quantified by FWER, therefore gets worse as the number of hypotheses tested increases, all other things being equal.

An alternative way to quantify the multiple inference problem (that is becoming more popular) is the false discovery rate (FDR). The false discovery rate is the expected proportion of rejections that are Type I errors. Formally, define V as the number of false rejections, U as the number of correct rejections, and $t = V + U$ as the total number of rejections. FWER is the probability that V is greater than 0 (i.e., the probability that we experience at least one false rejection). FDR, in contrast, is the expected proportion of all rejections that are Type I errors (i.e., false rejections), or $E[Q = V/t]$ (when $t = 0$, Q is defined to be 0). If all null hypotheses are true, then the total number of rejections equals the number of false rejections ($t = V$), and FWER and FDR are equivalent (Q equals 0 when there are no rejections and 1 when there are one or more rejections, so FDR $= E[Q] = P(t > 0) = P(V > 0) =$ FWER).

However, when some false hypotheses are correctly rejected, then FDR is less than FWER because the expected proportion of rejections that are Type I errors is less than the probability of making any Type I error. One way to see this is to note that the expected proportion of rejections that are Type I errors is equal to the probability of making at least one Type I error times the expected proportion of rejections that are Type 1 errors conditional on making at least one Type I error. Controlling FDR at a given level is therefore often less conservative than controlling FWER at the same level, resulting in increased statistical power.

2 Solution 1: Reduce the Number of Tests

There are two distinct ways to solve the multiple inference problem. One way is to simply reduce the number of tests being conducted. This method avoids adjusting the p-values on individual tests to control FWER or FDR – these adjustments generally reduce the power of any given test. Reducing the number of tests, however, limits the scope of hypothesis testing. The other approach maintains the number of tests but adjusts the p-values to reflect this fact. This approach allows for an arbitrarily large number of tests, but the power of each specific test can fall as the number of tests conducted grows. We first examine methods for limiting the number of tests being conducted.

The basic approach to limiting the number of tests is to jointly test a family of hypotheses instead of individually testing each of the hypotheses. The structure of the joint test will often be context specific, so we will discuss a few examples to give you a feeling for how you might proceed. In the preschool studies (Anderson 2008), the multiple inference issue centered around the large number of outcomes being studied – in each preschool experiment I observed a variety outcomes for each child (e.g., high school graduation status, standardized test scores, employment, earnings, arrests, etc.). The biostatistics literature (in which multiple inference is a well-recognized problem) refers to this problem as "multiple endpoints" (usually in the context of a clinical trial). Instead of testing whether preschool affected each outcome (i.e., endpoint) on an outcome by outcome basis, I tested whether preschool had an effect on a composite index of outcomes. To create this index, I coded all outcomes so that the positive direction indicated a "better" outcome and standardized outcomes by subtracting their respective means and dividing by their standard deviations (thus all outcomes were normalized to have a standard deviation of one). I then condensed multiple outcomes into a single index by taking the mean of all the outcomes in a given group (i.e., for each individual, I calculated the mean of his or her standardized outcomes).[1] By regressing this composite index on the treatment indicator, I effectively reduced the number of tests conducted.

[1]To improve efficiency, I also weighted by the inverse of the outcome variance-covariance matrix – see Anderson (2008) for details. This is generally not so important, however.

The test described above is useful for determining whether a treatment has a broad effect on a range of outcomes. It is inherently directional, however – if the treatment improves some outcomes but degrades others, the net effect on the composite index may be zero. If you have no a priori reason to believe that most outcomes should be affected in the same direction, an alternative is to apply a nondirectional test. For example, you could construct an SUR system in which there are M equations, one equation for each outcome. The explanatory variables would be the same for each equation, so the coefficient estimates would be no different than when running each equation separately. Estimating the equations as a system, however, allows you to test the hypothesis that the coefficients in all of the equations are equal to zero, i.e., the treatment has no effect on any outcome.

If the multiple inference problem occurs on the right-hand side of the regression instead of the left-hand side (i.e., you are testing whether a variety of factors demonstrate any effects on a particular outcome), the solution is even easier – simply apply a joint F-test to all the coefficients of interest.

The drawback to any joint testing procedure, however, is that it limits your ability to draw conclusions regarding specific hypotheses. If the joint test rejects, all you can conclude is that at least one of the hypotheses is false, but you don't really know which one. In many scenarios, then, you may wish to test individual hypotheses and adjust the p-values rather than consolidating tests.

3 Solution 2: Control FWER

There are two general approaches to adjusting the p-values of individual tests. The first approach controls the familywise error rate (FWER), or the probability of making *any* false rejection. A simple way to control FWER is to use the Bonferroni correction, which entails multiplying each p-value by the number of tests conducted. Although this procedure does control FWER, it is overly conservative, resulting in very poor power. It suffers in particular from three problems. First, it computes an upper bound rather than an exact probability (it

is common, for example, for Bonferroni p-values to exceed 1). Second, when a hypothesis is rejected, it can be removed from the family of null hypotheses being tested in order to increase the power of the remaining tests. Bonferroni makes no allowance for this. Finally, Bonferroni does not incorporate dependence between p-values. This can substantially increase power if p-values are highly correlated (in an extreme case, if all outcomes are perfectly correlated, FWER adjusted p-values and the unadjusted p-values should be equal).

A more powerful way to control FWER that addresses these three issues is the free step-down resampling method. This elaborately named procedure is implemented as follows:

1. Suppose that you are testing M hypotheses using M p-values. Sort the p-values, p_r, in order of decreasing significance (increasing p-value), i.e. such that $p_1 < p_2 < ... < p_M$.

2. Simulate the data set under the null hypothesis of no treatment effect by resampling $\mathbf{x_i}$ from its empirical distribution (and leaving $\mathbf{y_i}$ fixed). If you are testing the effect of one treatment on multiple outcomes, then $\mathbf{x_i}$ contains only one element.

3. Calculate a set of simulated p-values, $p_1^*, ..., p_M^*$, using the simulated treatment status variable(s). Note that they will not display the same montonicity as $p_1, ..., p_M$.

4. Enforce the original monotonicity: Compute $p_r^{**} = \min\{p_r^*, p_{r+1}^*, ..., p_M^*\}$. ($r$ denotes the original significance rank of the p-value, with $r = 1$ being the most significant and $r = M$ being the least significant)

5. Perform $L = 100,000$ replications of steps 2 through 4. For each p-value p_r, tabulate S_r, the number of times that $p_r^{**} < p_r$.

6. Compute $p_r^{fwer*} = S_r/L$.

7. Enforce monotonicity a final time: $p_r^{fwer} = \min\{p_r^{fwer*}, p_{r+1}^{fwer*}, ..., p_M^{fwer*}\}$. (This final monotonicity enforcement ensures that larger unadjusted p-values always correspond to larger adjusted p-values.)

The crucial steps of this algorithm are steps 2 through 4. Steps 2 and 3 are all about ensuring that the dependence structure between the different p-values is preserved. This structure is preserved because each observation is resampled with the correlation structure of its outcomes or treatments intact. We therefore expect $p_1^*, ..., p_M^*$ to be positively correlated (if the original outcomes or treatments were positively correlated), and the minimum p-value of a set of M positively correlated p-values is generally greater than the minimum p-value of a set of M independent p-values. Incorporating dependence thus increases the probability that $p_r < p_r^{**}$, reducing S_r and increasing the probability of rejection.

Step 4 performs the key multiplicity adjustment when the simulated p-value for p_r, p_r^*, is replaced with $\min\{p_r^*, p_{r+1}^*, ..., p_M^*\}$. The original p-value, p_r, is thus judged against the distribution of the minimum p-value of a set of $M - r + 1$ p-values. This makes the adjusted p-value more conservative than a standard p-value, which is implicitly judged against the distribution of the minimum p-value of a set of one p-value, but less conservative than the Bonferroni correction, which implicitly judges every p-value against the distribution of the minimum p-value of a set of M p-values.

Consider a concrete example of FWER adjusted p-values. Suppose that we are testing whether preschool affects $M = 10$ possible outcomes. Further suppose that the smallest p-value for the ten outcomes that we test is 0.020 and that the analogous FWER adjusted p-value, calculated via the free step-down resampling method, is $p^{fwer} = 0.110$. Suppose that we simulate the preschool data 100,000 times under the null hypothesis of no treatment effect (i.e., we randomly assign treatments to each of the observations). If we compute a set of 10 p-values for each simulation, the minimum p-value of that set will be less than or equal to the unadjusted p-value of 0.020 approximately 11 percent of the time. A minimum observed p-value of 0.020 is therefore not unlikely under the null given the number of tests conducted (10). For unadjusted p-values above the family's minimum p-value, the number of tests in the family effectively decreases, making the adjustment less severe.

FWER control limits the probability of making *any* false rejection. It is thus well suited to cases in which the cost of a false rejection is high. However, in many cases we are willing

to tolerate some false rejections in exchange for greater power (we are, after all, economists – our entire discipline is about making tradeoffs). Balancing false rejections versus greater power is particularly appealing when testing a large number of hypotheses, because FWER adjustments become increasingly severe as the number of tests grows – it is inherent in controlling the probability of making a single false rejection.[2] An alternative method of addressing the multiplicity problem that often affords better power is to control the false discovery rate, or the expected proportion of rejections that are Type I errors.

4 Solution 3: Control FDR

Controlling FDR formalizes the tradeoff between correct and false rejections and reduces the penalty to testing additional hypotheses. As discussed in Section 1, controlling FDR at a given level often requires less stringent p-value adjustments than controlling FWER at the same level, resulting in increased power.

Controlling FDR is actually quite simple (easier than controlling FWER). Suppose that we test M hypotheses $H_1, ..., H_M$, and again let the hypotheses be sorted in order of decreasing significance, such that $p_1 < p_2 < ... < p_M$. Suppose $q \in (0, 1)$. In general we will choose $q = 0.05$. Let c be the largest r for which $p_r < qr/M$. Rejecting all hypotheses $H_1, ..., H_c$ controls FDR at level q for independent or positively dependent p-values. In other words, beginning with p_M, check whether each p-value meets the condition $p_r < qr/M$. If it does, reject it and all smaller p-values. If it does not, check whether the next smallest p-value meets that condition. Continue until the condition is finally met or you run out of p-values. I refer to this procedure as the BH procedure because it was developed in Benjamini and Hochberg (1995).

This procedure is in fact conservative in that it controls FDR at level $q(m_0/M)$, where m_0 is the number of true null hypotheses. We do not observe m_0, but if we did we could

[2] In an extreme case, suppose we are testing 1,000 hypotheses. If we want to limit the probability of making just a single false rejection to only 5%, then we will have to apply very severe adjustments to the individual p-values.

"sharpen" the procedure by replacing qr/M with qr/m_0. Since $qr/m_0 \geq qr/M$, the sharpened procedure would provide greater power if at least one null hypothesis were false.

A two-stage procedure that estimates the number of true hypotheses to achieve sharpened FDR control is implemented as follows:

1. Apply the BH procedure at level $q' = q/(1+q)$. Let c be the number of hypotheses rejected. If $c = 0$, stop. Otherwise, continue to step 2.
2. Let $\hat{m}_0 = M - c$.
3. Apply the BH procedure at level $q^* = q'M/\hat{m}_0$.

By incorporating the number of hypotheses rejected in the first stage into the second stage, this procedure provides better power than the standard BH procedure while still controlling FDR at level q for independent p-values.

The BH and two-stage procedures both report whether a hypothesis was rejected at level q, but they do not report the smallest level q at which the hypothesis would be rejected. This value – which is the natural analog to the standard p-value – can easily be computed for all hypotheses by performing the procedure for all possible q levels (e.g., 1.000, 0.999, 0.998,...) and recording when each hypothesis ceases to be rejected. (I have Stata code available to easily calculate these FDR "q-values.").

To understand in practice why FDR control is less conservative than FWER control, consider how the BH and free step-down resampling procedures treat the median p-value, $p' = p_{M/2}$, in a set of M p-values. Roughly, the BH procedure rejects $H' = H_{M/2}$ if $p_{M/2} < \alpha(M/2)/M = \alpha/2$, while the free step-down resampling procedure rejects $H_{M/2}$ if $p_{M/2}$ exceeds the minimum of a family of $M/2$ simulated p-values at a rate less than α. The former equates to adjusting the p-value by a factor of 2, while the latter equates to adjusting the p-value by a factor of up to $M/2$. For large M, the difference becomes substantial. Note also that M does not appear on the right side of the expression $p_{M/2} < \alpha/2$. If additional

p-values – distributed similarly to the existing p-values – are added to the family of tests, the FDR adjustment to the existing p-values need not become more stringent in expectation.

For analyses testing many hypotheses, I would generally recommend controlling FDR rather than FWER, unless the cost of a false rejection is extraordinarily high. FWER control becomes exceedingly conservative for very large numbers of hypotheses, and it is unclear why we would assign such a high weight to avoiding any false rejections while assigning virtually no weight to avoiding "false acceptances."

5 Additional References

Benjamini, Y. and Y. Hochberg. "Controlling the False Discovery Rate." *Journal of the Royal Statistical Society: Series B*, 1995, 57, 289-300.