IEOR 151 – Lecture 6
Multiple Testing

1 Example: Comparing Restaurant Quality

Consider the following hypothetical situation: There is a chain of fast food restaurants that is facing decreased customer satisfaction and revenues, and management believes the problem is that consumer demand has caused the menu to become so large that the employees cannot keep service times low while being able to prepare the many different items on the menu. As an experiment, 25 of the restaurants are changed to a service model in which half of the menu are pre-made packaged items and the other half are items that require preparation. These are compared to another 50 restaurants that keep the old service model of a large menu, and a decision will be made on which approach to use in all the restaurants based on the the results of the comparison.

For these 75 restaurants, patrons were randomly selected to fill out a survey in which they were asked to numerically rate their satisfaction (from 1=very unsatisfied to 5=very satisfied) in the categories of service time, menu options, food quality, and food taste. These patrons were also asked to rate their likelihood of returning to the restaurant and likelihood of recommending the restaurant (from 1=very unlikely to 5=very likely). In addition, management is interested in how the different service models affect revenue per order, profit per order, service time per order, and customer count. Hypothesis tests were conducted comparing the two groups of restaurants on these qualities, and the average values for each groups and p-values are summarized below:

	New Menu	Old Menu	p-value
Service Time Ratings	4.2	3.1	$p = 0.002$
Menu Options Ratings	2.8	4.2	$p = 0.009$
Food Quality Ratings	4.3	4.5	$p = 0.010$
Food Taste Ratings	2.9	3.7	$p = 0.030$
Returning Likelihood	4.0	3.9	$p = 0.042$
Recommendation Likelihood	2.8	2.3	$p = 0.053$
Revenue per Order	18.42	17.23	$p = 0.005$
Profit per Order	2.79	1.32	$p = 0.006$
Service Time per Order	3:19	5:83	$p = 0.004$
Customer Count	1104	1152	$p = 0.012$

2 Risk in Hypothesis Testing

The first challenge with this scenario is that there is always some quantity of risk inherent in making decisions, and the p-value approach to hypothesis testing does not take this risk into account. Recall that the exact description of a p-value is the smallest significance level for which the hypothesis test would be rejected. Furthermore, recall that the significance level denotes the probability of rejecting the null hypothesis under the assumption that the null hypothesis is true. This framework does not require rigorously defining an alternative hypothesis, and so we have zero guarantees about other types of errors of decision making (e.g., accepting the null hypothesis when it is false).

Because of this mismatch, one approach to dealing with risk in hypothesis testing is to choose the significance level to reflect the level of risk of accepting or rejecting the null hypothesis. In situations where there is low risk for accepting the null hypothesis (and low risk for "rejecting" the alternative hypothesis), we can set the significance level to a lower value. In situations where there is low risk for rejecting the null hypothesis (and low risk for "accepting" the alternative hypothesis), we can set the significance level to a higher value. These significance levels are always subjective, but in this way we can qualitatively adjust for the level of risk we would like to take.

3 Difficulties with Multiple Testing

The second challenge with this scenario is that we are simultaneously performing multiple hypothesis tests, and this is tricky because the fact that multiple tests are being conducted must be taken into account in order to ensure that the false positive rate is below the desired significance level α. It particularly difficult because in many situations multiple tests are implicitly conducted.

As another hypothetical example, imagine that a scientific researcher performs a study to determine whether having a Bear Patrol keeps bears out of town. The researcher compares the average number of bear sightings in towns with and without a Bear Patrol, and determines that the difference is not statistically significant since they get $p = 0.75 > \alpha = 0.05$. But because the result does not reject the null hypothesis, the researcher does not publish the result. Now suppose that another k researchers perform the same study, but use data from different towns. Interestingly, some of the k researchers get $p < \alpha$, and conclude that Bear Patrols do keep bears out of town. These results are published in journals and widely advertised by the press.

Returning to this Bear Patrol example, an interesting question to ask is: What is the probability of a false positive (that at least one researcher concludes Bear Patrols do keep bears out of town) assuming that the null hypothesis is true. This is the same as the event that

all k researchers do not make a false positive error, and this probability is given by

$$P := \mathbb{P}(\text{false positive with } k \text{ tests}) = 1 - 0.95^k,$$

where we have assumed that each test is independent. A table of this probability for different values of k is given below. What this example shows is that the probability of making a false

k	1	2	5	10	100
P	0.05	0.10	0.23	0.40	0.99

positive error increases as more tests are conducted. This shows the need for a method to compensate for this.

4 Bonferroni Correction

Suppose there are k null hypotheses $(H_0^1, H_0^2, \ldots, H_0^k)$ with k corresponding p-values

$$(p_1, p_2, \ldots, p_k),$$

and assume that the desired maximum probability of having one or more false positives is given by the significance level α. The probability of having one or more false positives is also known as the familywise error rate.

The idea of the Bonferroni correction is to modify the testing procedure to reject hypothesis only when $p_i < \alpha/k$. The intuition of this can be given by using Boole's inequality (also known as the union bound) to bound the familywise error rate. Let I_0 be the indices of the null hypotheses that are true, then we have

$$FWER = \mathbb{P}(\cup_{i \in I_0}(p_i \leq \alpha/k)) \leq \sum_{i \in I_0} \mathbb{P}(p_i \leq \alpha/k) \leq k\alpha/k = \alpha.$$

Note that there is no assumption of independence.

4.1 Alternative Form

The Bonferroni correction can be defined in an alternative form. We define the *adjusted* p-value as

$$\tilde{p}_i = \min\{k \cdot p_i, 1\}.$$

If the adjusted p-value is below the original significance level $\tilde{p}_i < \alpha$, then the i-th null hypothesis is rejected.

4.2 Restaurant Example

Suppose that we wish to make conclusions in the restaurant example at the level $\alpha = 0.05$. Applying this method to the example data, we should reject any null hypothesis for which the p-value is below $\alpha/10 = 0.005$. This means that we should reject the following null hypothesis: No differences in service time ratings and service time per order. We should accept the remaining null hypothesis. The results are quite different than if we had rejected all tests with p-values below $\alpha = 0.05$.

5 Holm-Bonferroni Method

It turns out that the Bonferroni correction can be overly conservative, and more careful testing procedures can have greater power than the Bonferroni correction. An example is the Holm-Bonferroni method, which provides improvement over the Bonferroni correction. This method can be summarized as an algorithm:

1. Sort the p-values into increasing order, label these $p_{(1)}, p_{(2)}, \ldots, p_{(k)}$, and denote the corresponding null hypotheses as $H_0^{(1)}, H_0^{(2)}, \ldots, H_0^{(k)}$;
2. Let m be the smallest index such that $p_{(m)} > \alpha/(k-m+1)$; if no such m exists, then reject all hypotheses; if $m = 1$, then do not reject any hypothesis;
3. Otherwise, reject the null hypotheses $H_0^{(1)}, H_0^{(2)}, \ldots, H_0^{(m-1)}$ and accept the null hypotheses $H_0^{(m)}, H_0^{(m+1)}, \ldots, H_0^{(k)}$;

The idea is behind this test is a little more complicated than that of the Bonferroni correction. Let I_0 be the null hypotheses that are true, and define $n_0 = \#I_0$ to be the number of hypotheses in I_0. Let j be the smallest index satisfying $p_{(j)} = \min_{i \in I_0} p_i$, and note that $j \leq k - n_0 + 1$ because there are only $k - n_0$ remaining hypotheses. Next, observe that $\alpha/(k-j+1) \leq \alpha/n_0$, and the familywise error rate is given by $\mathbb{P}(p_{(j)} \leq \alpha/(k-j+1))$. And so it must be that

$$FWER = \mathbb{P}(p_{(j)} \leq \alpha/(k-j+1)) \leq \mathbb{P}(p_{(j)} \leq \alpha/n_0) \leq \sum_{i \in I_0} \mathbb{P}(p_i \leq \alpha/n_0) = n_0 \alpha/n_0 = \alpha.$$

5.1 Alternative Form

The Holm-Bonferroni method can be defined in an alternative form. We define the *adjusted* p-value as

$$\tilde{p}_{(i)} = \max_{j \leq i} \left(\min\{(k-j+1) \cdot p_{(j)}, 1\} \right).$$

If the adjusted p-value is below the original significance level $\tilde{p}_{(i)} < \alpha$, then the null hypothesis $H_0^{(i)}$ is rejected.

5.2 Restaurant Example

Suppose that we wish to make conclusions in the restaurant example at the level $\alpha = 0.05$. Applying this method to the example data, observe that the sorted p-values are given by: 0.002, 0.004, 0.005, 0.006, 0.009, 0.010, 0.012, 0.030, 0.042, 0.053. The "thresholds" are given by: 0.005, 0.006, 0.006, 0.007, 0.008, 0.010, 0.013, 0.017, 0.025, 0.050. The smallest index is $m = 5$ because $p_{(5)} = 0.009 > \alpha/(10 + 1 - 5) = 0.008$. As a result, we should reject the following null hypothesis: No differences in service time ratings, revenue per order, profit per order, and service time per order. We should accept the remaining null hypothesis. Compare the conclusions to those made when using the Bonferroni correction.