Chapter 6

Statistical Testing

6.1 Introduction

Two questions analysts might have about MaxDiff utilities have answers in statistical testing:

1. Do the utilities from one subgroup of respondents differ from those of some other subgroup(s) of respondents? For example, do males and females (or Belgians and Poles, or people from eight different income strata or three different political parties) have different preferences for the items included in a MaxDiff experiment? We call these *Tests for between-group differences.*
2. *Tests for between-item differences,* on the other hand, answer the question "Does the utility of one item differ from that of another (or which items have higher utilities than which other items)?"

In the sections below we describe omnibus statistical tests and then more detailed tests from both classical and Bayesian perspectives.

A significant result from a statistical test performed at 95% confidence means that such a result would occur no more than 5% of the time by chance alone. Say we have a small MaxDiff with 10 items. To compare males' and females' utilities we plan to run 10 independent groups t-tests. Because we have 10 such tests rather than just one, the chance of at least one of the tests returning a significant result, by chance alone, rises to $1.00 - 0.95^{10}$ or 40%. By running multiple tests our chance error rate inflated from 5% to 40%! We call that inflated

error *experimentwise error* (and the excess significant results we call instances of false discovery).

To combat experimentwise error, a cautious analyst can do one of two things. First, she can run an "omnibus" statistical test which tells her whether any differences at all are significant, and then only if so does she then follow-up with the various individual stat tests. Or, the analyst can forego the omnibus test and follow her various pairwise tests with a *post hoc* correction for experimentwise error. As these involve more advanced methods, we mostly relegate omnibus tests and *post hoc* corrections for experimentwise error to this chapter's Appendix.

6.2 Tests for Between-Group Differences

Using Aggregate MNL Utilities

In linear regression analysis, we test for between-group differences using a Chow test. Creating a dummy variable coded as 1 for one subgroup of respondents and 0 for a second subgroup of respondents, we run the regression model twice: once using the original predictor variables and once using both the original predictors and the interactions of the dummy variable for group membership with each predictor variable. A statistical test for improvement in R^2 serves as an omnibus test for the regression model, telling us whether or not *any* of the coefficients differ significantly between the subgroups. A significant omnibus test enables us check which particular regression coefficients differ between the groups, indicated by significant interaction terms.

Initially analysts ran a logit equivalent of the Chow test, until Louviere and Swait (1993) pointed out that the logit scale factor could confound the test and that possible differences in scale required a more complicated test, the Swait and Louviere test that appears in the Appendix.

Using Individual Respondent-Level Utilities

After learning from the Swait and Louviere omnibus test in the Appendix that significant between-group differences in utilities exist (or before confirming that we're not being victimized by experimentwise error using the Benjamini-Hochberg Procedure that also appears in the Appendix) we decide to run some between-group tests

on individual respondent-level utilities. Classical and Bayesian statistics have different ways of performing these tests.

Classical Between-Groups Test

To run the classical t-test or ANOVA, we first need to normalize the utilities to try to adjust for the fact that different respondents might have answered the MaxDiff questions with different amounts of consistency, for example using the Zero-Centered Diffs procedure described in Chapter 5. Now imagine that we've done so for a MaxDiff survey among dog owners about how much they like various breeds of dogs.

To learn whether men and women have different preferences for breeds we can use our favorite statistical software to compare males' and females' (normalized) utilities for any breeds of interest. When comparing exactly two groups we use the t-test and in cases where we want to test for differences across more than two groups we use the F statistic from a one-way analysis of variance (ANOVA). You can run these tests in your favorite statistical software or even in Excel.

Bayesian Between-Groups Test

The Bayesian test turns out to be pretty easy. HB estimation (Chapter 4) uses an iterative procedure that produces thousands of "draws" (or estimates) for both the upper- and lower-level models that make up the model hierarchy. The upper-level draws we call draws of alpha (where alpha is our current estimate of the population mean utility vector). The lower-level draws we call draws of beta (where beta is our current estimate of an individual's utility vector). We typically ignore the draws before the point where the model converges and becomes stable (typically the first 5K or 10K draws). We then use the remaining draws (which Sawtooth Software's HB estimation program can write to a convenient .CSV file) to conduct statistical tests in a way that only involves counting.

Statistical testing for HB requires a shift in mindset. You examine the distributions of draws of parameters to see if a strong majority of these draws falls on either one side or the other of the null hypothesis.

For the dog breeds example, let's imagine we conducted a MaxDiff experiment, estimated utilities via HB, and we were interested in developing a 95% credible interval (often referred to as a confidence interval) for the Beagle utility. We open the file containing the successive estimates of alpha for the population. Alpha is a vector containing all the parameters in the model, but we want to focus just on one element (column) of that vector: the Beagle utility. Imagine that you have 20,000 total draws (10,000 before convergence and then another 10,000 after) and that the first few draws of alpha (after convergence) look something like the following:

Draw#	Beagle
10001	-0.30285
10002	-0.29823
10003	-0.29801
10004	-0.30114
10005	-0.30159

One way to develop the 95% credible (confidence) interval is simply to sort the 10,000 "used" draws from lowest to highest and then pick out the 2.5th and 97.5th percentile values. Those mark the endpoints of the 95% credible interval (the interval that contains 95% of the estimates).

If you employ covariates in your HB model based on membership in the groups you want to test (e.g., males vs. females) you obtain estimates of alpha by different respondent groups. Imagine that gender is dummy-coded as a single covariate parameter in the design matrix, where one of the states (male) is assigned as the reference (0) level and the other state (female) is assigned to be 1. Now, when you open the file containing the successive estimates of alpha draws, you'll find the intercept alpha parameters (corresponding to the mean utilities among males for all the items) followed by the alpha parameters associated with the adjustments to the intercept values to result in mean female utilities for the items in the model. If we want to know if males prefer beagles more than females do (relative to the average breed of dog), we count for what percent of the used draws the males' utility for that level exceeds the females' utility (where the female's utility for beagles is equal to the intercept beagle parameter plus the adjustment for being female parameter for beagle). If 92% of the draws show the males' mean utility exceeds the females' utility,

then we are 92% confident that males like beagles more than females do (relative to the other breeds of dog included in the experiment). This procedure involves repeated runs of the HB analysis with as many covariates as stat tests one wants to run, so we're not aware of many practitioners who use it – much more often we see folks use the classical between groups test described above.

6.3 Tests for Between-Item Differences

We don't worry much about an omnibus test for differences among items. We have yet to come across a MaxDiff study with no significant between-item differences.

Using Aggregate MNL Utilities

Consider the utilities for our MaxDiff study of preferences for different dog breeds among dog owners:

Breed	Utility	Std Error
Toy Fox Terrier	0.474	0.133
Chihuahua	-0.127	0.129
Dachshund	0.040	0.131
Spanish Galgo	1.339	0.130
. . .	. . .	. . .
Indian Pariah Dog	0.347	0.128

To test if there is a statistically significant difference between items in a MaxDiff, realize that each respondent provides utilities for each breed, so that the observations of breeds are dependent. So, you compute a dependent groups t-test using the following formula:

$$t = \frac{U_1 - U_2}{\sqrt{SE_1{}^2 + SE_2{}^2 - 2COV_{12}}}$$

Where U_1 and U_2 are the utilities (effects) of the two levels and the expression in the denominator is the pooled standard error for the two levels (SE_1 is the standard error for U_1, SE_2 is the standard error

for U_2, and COV_{12} is the covariance[4] between the two levels). If the covariance of the two levels is 0.0012, the t-value for the statistical test of difference in utilities between how much dog owners like Chihuahuas and Indian Pariah Dogs is

$$(0.347\text{--}0.127)/\sqrt{0.128^2 + 0.129^2 - 2(0.0012)} = 2.71.$$

Since this t-value has an absolute magnitude greater than 2.58, we are greater than 99% confident that dog owners like Indian Pariah dogs more than they like Chihuahuas.

Classical Test for Between-Item Differences

There are again Classical and Bayesian statistical methods for running between-item tests.

Classical Test for Between-Item Differences

If you familiar with running *post hoc* tests for ANOVA models the easiest way to test for between item differences is using a repeated measures analysis of variance (or RM-ANOVA) followed by *post hoc* Tukey tests. An (unsurprising) significant result for the RM-ANOVA indicates that at least one pair of items have significantly different utilities and the Tukey *post hoc* tests tell you specifically which pairs of items are significantly different. For example, if we have 20 items in our MaxDiff, if we want to test for between-item differences, then we have a total of 190 paired comparisons to test. Testing this as a RM-ANOVA with a Tukey *post hoc* test will properly account for the experimentwise error from running so many tests. See Stevens (1996) for a useful discussion of the Tukey test for RM-ANOVA.

We've yet to come across anyone else using this approach. Most folks who use classical statistical tests simply run large numbers of dependent t-tests in their favorite statistical software packages. If you do this, however, you should at least correct for the number of t-tests you run by using the Benjamini-Hochberg Procedure described in the Appendix.

---

[4] If using Sawtooth Software's Lighthouse Studio to compute the aggregate logit scores, you can request the covariances from the Advanced Settings dialog.

Bayesian Test for Between-Item Differences

Let's imagine that we have successive estimates of alpha for Foxhounds and Miniature Poodles, and the first five draws (after convergence) look like this:

Draw#	Foxhound	Miniature Poodle
10001	-0.25136	-0.12476
10002	-0.24921	0.01873
10003	-0.22384	-0.08757
10004	-0.21998	-0.08897
10005	-0.23179	0.00622

If we want to compute our degree of confidence that respondents prefer Miniature Poodles more than Foxhounds, we simply count for how many draws of alpha the utility estimate for the Poodles is greater than for Foxhounds. If 99.17% of the alpha draws had Miniature Poodles preferred to Foxhounds, then we are 99.17% confident that our respondents prefer Miniature Poodles to Foxhounds. To correct for multiple comparison error when we run more than one of these tests, see the Benjamini-Hochberg procedure in the Appendix.

6.4 Appendix

Omnibus Tests and Post Hoc Corrections for Experimentwise Error

We show these tests in the Appendix not because we find them unimportant, but because few analysts use them. Certainly these methods add time and effort to the testing process that can pinch the schedules of many commercial studies.

The Swait and Louviere Test for Between-Group Differences

We know that the MNL model has a scale factor as described in Chapter 4. The scale factor affects the size of the utilities such that respondents who answer more consistently, i.e. with less response error, tend to have utilities with larger absolute values across the board, while respondents who answer less consistently, i.e. with more response error, tend to have smaller (closer to zero) utilities. Imagine a situation where Poles and Belgians have the same preferences for the items but that the Poles make their MaxDiff choices more consistently. In this case the utilities might look different, because the Poles

would have larger positive and larger negative utilities while the Belgians' utilities would be smaller (i.e., closer to zero). In order to test properly for differences in logit models we need to take into account differences in respondent consistency, that is, differences in the logit scale factor.

Swait and Louviere (1993) supply just such a test. Their test procedure is sequential and it involves running four models for any comparison between two groups.

1. We use the data matrix from the first group to be compared, X_1, to run an MNL model which gives us a set of utilities β_1 and a log likelihood fit of LL_1.
2. Similarly, we use the X_2, the data matrix from the second group of respondents, to power a model that yields a second set of utilities, β_2, and a corresponding log likelihood LL_2.
3. For the third model we just concatenate the two data matrices $X_1 + X_2$ to get a set of utilities β_P and a fit statistic LL_P for the pooled model.
4. Finally, we run a scale-constrained model where we use the pooled data matrix, only with the scale parameter μ for the second data matrix that maximizes the log likelihood fit of the model: $X_1 + \mu_2 X_2$. One finds the value of the scale parameter μ_2 by running the model time and again, moving μ_2 in the direction that improves log likelihood, until the log likelihood stops improving. The log likelihood function's concavity guarantees a unique maximum discoverable through this simple search procedure. What results is a set of coefficients b_μ and LL_μ.

The test procedure evaluates three related hypotheses:

H_1: $\beta_1=\beta_2$ and $\mu_1=\mu_2$
H_{1A}: $\beta_1=\beta_2$
H_{1B}: $\mu_1=\mu_2$ (because we cannot identify both scale parameters, we can set $\mu_1=1$ and compute μ_2 relative to it)

The first test in the sequence uses the fit statistics from the two groups' separate models and from the scale-constrained model, as follows:

$$\lambda_A = -2 [LL_\mu - (LL_1 + LL_2)]$$

If λ_A exceeds the critical chi-squared value with K+1 degrees of freedom (K is the number of parameters in each of the three logit models) then we reject H_{1A} and H_1. We conclude that the coefficients (utilities) differ for the two groups and we stop the testing sequence. Only at this point would it be appropriate to proceed to test individual coefficients to see which ones differ between the groups using t-tests similar to those described in the body of the chapter.

If we fail to reject H_{1A} we can go on to test H_{1B} as follows using the fit statistics from the pooled and the scale-constrained models:

$$\lambda_B = -2 (LL_P - LL_\mu)$$

If λ_B exceeds the critical chi-squared value for a test with 1 degree of freedom, then we reject H_{1B} and conclude that the two models have different scales.

Note that with this procedure we may find that the models have different utilities or that they have different scales, or we may find that their scales and utilities do not differ significantly. We will never be in a position to conclude that two models have both different scales and different utilities, however.

Continuing the example above, we might reject the first hypothesis and conclude that the utilities of Belgians and Poles differed significantly. Alternatively, we might reject the second hypothesis and conclude that Belgians and Poles had significantly different scale factors (i.e., amounts of response error) but underlying utilities that did not differ significantly. Finally, we might reject neither hypothesis and conclude that Poles and Belgians differ neither in their utilities nor in the consistency with which they answer their choice questions.

Tests involving more than two groups would be analogous.

The Benjamini-Hochberg Procedure for Experimentwise Error

If we opt not to perform the omnibus test above, we can go ahead and run all of our various classical t-tests or their Bayesian equivalents and then afterwards account for the fact we've run lots of tests by using the Benjamini–Hochberg (BH) Procedure (Benjamini and

Hochberg 1995). Unlike other corrections for experimentwise error, BH controls false positives without the danger of inflating the number of false negatives.

The procedure works as follows:

1. Rank the p-values of all your tests from lowest to highest
2. Calculate a critical p-value for each rank by multiplying the rank by the error rate you want to allow across all tests (e.g., 0.05) and divide by the number of tests (so if we run 45 statistical tests and we want a 0.05 experimentwise error rate the critical value for the first rank would be 1 * 0.05 /45 = 0.00111. For the second ranked test it would be 0.00222, etc.)
3. Identify the highest rank for which the observed p-value is less than the critical p-value
4. For that rank and all lower ranks reject the hypothesis that the interaction effect is 0

For example, say we've run 45 statistical tests and five of them appear to be significant. We now table the observed p-values and the BH critical p-values for the first 5 ranks:

Rank	Observed p	Critical p
1	0.0021	0.00111
2	0.0023	0.00222
3	0.0025	0.00333
4	0.0158	0.00444
5	0.0362	0.00555

The third rank is the highest wherein the observed p is lower than the critical p, so in this study we accept the first, second and third ranked tests as significant.

So Benjamini-Hochberg is a tool we can use to decide which tests are "really" significant when we run a number of tests and inflate our experimentwise error rate.