Chapter 12

Statistical Testing

12.1 Introduction

We (or our clients) often want to know how confident we are in our estimates. What's the 95% confidence interval? Is a price coefficient different from zero? Is one attribute level preferred to another? Do groups of respondents differ in their preferences? How about confidence surrounding shares of preference from a choice simulator? This chapter deals with these topics. We start by illustrating statistical hypothesis tests with aggregate logit, then change our focus to statistical tests involving the commonly used HB (hierarchical Bayes) approach.

This chapter contains many sections, so it may be helpful to give you an outline for quick reference:

12.5 Bayesian Tests for HB Estimation
- A. Confidence intervals and HB
- B. Significance of attributes and levels with HB
- C. Tests for differences between levels within attributes with HB
- D. Tests for differences between product concepts with HB
- E. Tests for differences between groups of respondents on attribute levels with HB
- F. Tests for interaction effects with HB

12.6 Frequentist Tests for Shares of Preference
- A. Confidence intervals
- B. Testing differences in shares of preference

12.7 HB Testing for Product Shares of Preference

12.8 Appendix: The Swait and Louviere Test for Between-Group Differences in MNL Utilities

Generally, we recommend the Bayesian tests. The frequentist tests based on individual-level utility estimation (point estimates) also tend to work quite well, though in our experience they tend to be a bit less conservative than the Bayesian tests (they are more apt to find significant differences and to understate the degree of uncertainty).

12.2 Statistical Testing with Aggregate Logit

Pooled (aggregate) logit is often used as a quick estimation approach for estimating summary part-worth utility weights for a sample of respondents. A typical report for an aggregate logit estimation is shown below (Exhibit 12.1).

Log-likelihood for this model	-4606.05762		
Log-likelihood for null model	-5198.60385		
Difference	592.54623		
Chi-Square	1185.09247		
Degrees of Freedom	7		
P-Value	0.00000		
Variable	**Effect**	**Std Error**	**T Ratio**
High-Flyer Pro, by Smith and Forester	0.54582	0.03520	15.50595
Magnum Force, by Durango	0.36009	0.03518	10.23618
Eclipse+, by Golfers, Inc.	-0.37493	0.04087	-9.17325
Long Shot, by Performance Plus	-0.53098	0.04271	-12.43228
Drives 5 yards farther than the average ball	-0.47410	0.03281	-14.44912
Drives 10 yards farther than the average ball	0.12715	0.02916	4.36104
Drives 15 yards farther than the average ball	0.34694	0.02845	12.19481
Price (Linear Coefficient):	-0.21816	0.01010	-21.59281
NONE	0.00504	0.04135	0.12185

Example Aggregate Logit Report
Exhibit 12.1

A. Overall significance of the model

With aggregate logit, a single vector of part-worths is fit to all respondents' choices (see Chapter 10 for more details). A statistical test of the overall significance of the model is shown at the top of the report in the form of the fit (log-likelihood) of the null model compared to the fit that the estimated part-worths provide (log-likelihood for this model). With frequentist statistical testing, the null hypothesis is that the part-worth utilities are all zero and the model fit is not statistically significant. The null model is uninformative and predicts equal likelihood for each concept in each task. This data set has 250 respondents, 15 tasks per respondent, where each task has 4 concepts. Thus, the null log-likelihood is equal to $\ln(0.25)(15)(250) = -5198.60$. However, the model we've fit has part-worth utilities different from zero and the log-likelihood fit of this model is -4606.06.

Twice the difference between the model fit (-4606.06) and the null fit (-5198.60) is distributed as chi-square. This is called the *2 log-likelihood test.*

$$(1) \qquad \text{chi-square} = 2(LL_m - LL_n)$$

where:

LL_m is the log-likelihood of this model

LL_n is the log-likelihood of the null model

For the model shown in Exhibit 12.1, the chi-square is equal to 2(-4606.05762 - -5198.60385) = 2(592.54623) = 1185.09247, with degrees of freedom equal to the number of parameters estimated in the model, or 7. The reported p-value is the likelihood of observing a chi-square statistic this large due to chance, given 7 degrees of freedom[25]. That reported p-value is extremely small (0.00000) after rounding to the nearest five decimal places of precision. We are therefore better than 99.999% confident (1 - p) that the model provides significant fit to the data.

How does one count degrees of freedom? In Chapter 10, we demonstrated how to dummy-code a CBC model to avoid linear dependency in the design matrix. The first two attributes in this model are dummy-coded with k-1 levels, where k is the number of levels within the attribute. The Price attribute in this model is estimated as a single linear coefficient (1 parameter). An additional parameter is used to estimate the None (another 1 parameter). Therefore, the total number of parameters in this model is (4-1)+(3-1)+1+1=7.

B. Comparing two (nested) aggregate logit models

We might consider running a second logit model (using the same respondents and choice tasks) where Price is dummy-coded rather than coded as a single coefficient (Price has 4 levels in this study), leading to a new model with 9 total parameters to estimate. We might want to test whether the new larger model provides better fit than the previous model—a *statistically significant* better fit. The new log-likelihood for the model (not shown here) with dummy-coded price is

[25] You can always turn to a chi-square table in your favorite statistics text book to look up p-values, but a more convenient way to compute directly the exact p value is to use the =CHIDIST(CHI_Value, DF) formula provided in Excel. Specifying =CHIDIST(1185,7) in a cell returns a p-value of 1.2E-251.

-4600.39426. The chi-square test for the difference in model fit is equal to:

$$(2) \qquad \text{chi-square} = 2(LL_n - LL_o)$$

where:

LL_n is the log-likelihood of the new model
LL_o is the log-likelihood of the old model

The chi-square is equal to 2(-4600.39 - -4606.06) = 11.34 with degrees of freedom equal to the difference in number of parameters estimated for the two models, or 9 - 7 = 2. Again, referring to a chi-square table, we find that the p-value associated with this chi-square critical value (2 degrees of freedom) is 0.00345. We are 99.655% confident that the new model (with dummy-coded price) provides a better fit than the old model (with linear price).

Tests for non-nested models are also available which allow you to determine if you have included the right set of variables in the model. This situation usually occurs when you build a model that combines conjoint data with other predictors (e.g., demographics), a discussion that takes us beyond the scope of this book (see Ben-Akiva and Swait 1986).

In methodological studies we often see tests for differences in models with the same structure but collected from different sets of respondents (e.g., do the utilities of males differ from those of females; do the utilities differ for people from different regions or countries; do they differ between this wave of the study and the last wave, etc.). This test, the Swait and Louviere test, is a technical enough topic that we have included it in the Appendix to this chapter.

C. Significance of attributes and levels

In the model shown in Exhibit 12.1, we estimated a single coefficient for price. The relevant portion of that aggregate logit report is shown below:

	Effect	Std Err	T Ratio
Price (Linear Coefficient):	-0.21816	0.01010	-21.59281

It's common to express our confidence about the precision of the coefficient (or a part-worth utility weight, in the case of dummy-coding) in terms of a 95% confidence interval. The 95% confidence interval is calculated by taking the utility weight estimate plus or minus 1.96 standard errors. The utility weight estimate is -0.21816 and the standard error is 0.01010, leading to a 95% confidence interval of [-0.23796, -0.19836]. This interval is commonly interpreted as follows: if we were to draw new samples and repeated the experiment hundreds or thousands of times and computed a new confidence interval each time based on our model, the confidence intervals would contain the true population mean 95% of the times.

Often, researchers are interested in testing whether a coefficient such as the price weight is different from zero. We do this via hypothesis testing and computing a critical value called the t-ratio (the t-value). We calculate the t-ratio by dividing the coefficient (the effect) by the standard error, or in this instance -0.21816 / 0.01010 = -21.59281. T-values with absolute magnitudes greater than 1.96 give us at least 95% confidence that the coefficient is different from zero[26]. T-values with absolute values greater than 2.58 provide at least 99% confidence. Thus, given this t-value of -21.59281 we can reject the null hypothesis (that the coefficient for price is zero) with greater than 99% confidence.

When the part-worth utilities are estimated via effects-coding (a form of dummy-coding that constrains the sum of utilities within each attribute to be zero), the t-values associated with each part-worth let us compute the confidence that each part-worth utility is different from their average (zero). The results for the brand attribute from the model shown in Exhibit 12.1 are given below.

[26] For practical purposes, we assume that the number of observations is large enough such that the t-distribution has converged to the z-distribution. Thus, we use the critical values for 95% and 99% confidence from the z-distribution, assuming a two-tailed test.

Variable	Effect	Std Error	T-Ratio
High-Flyer Pro, by Smith and Forester	0.54582	0.03520	15.50595
Magnum Force, by Durango	0.36009	0.03518	10.23618
Eclipse+, by Golfers, Inc.	-0.37493	0.04087	-9.17325
Long Shot, by Performance Plus	-0.53098	0.04271	-12.43228

Given that we are using effects-coding (that constrains part-worth utilities to be zero-centered), you would not be surprised to find one or more of the part-worth utilities for a multi-level attribute to be close to zero. Such a finding would not necessarily mean that this particular level of middling preference was being ignored by respondents.

Another potential oddity with conducting statistical tests on aggregate part-worth utilities for unordered attributes is that differences of opinion across people regarding fairly important attributes can nearly cancel out and lead to population part-worth utility weights near zero. Heterogeneity (differences in tastes) may make it look like the brand or color attribute is relatively less important or completely unimportant for the population, though the truth may be that the attribute is actually very important to most individual respondents. Imagine the case in which an equal number of respondents have different favorite brands, leading to their preferences cancelling each other out when viewed in the aggregate. This issue is avoided when conducting statistical tests on the part-worth utilities for ordered attributes that have expected logical preference order (such as for the performance attribute in Exhibit 12.1).

D. Tests for differences between levels within attributes

Consider the part-worth utilities for the performance attribute:

	Effect	Std Error	T-Ratio
Drives 5 yards farther than the average ball	-0.47410	0.03281	-14.44912
Drives 10 yards farther than the average ball	0.12715	0.02916	4.36104
Drives 15 yards farther than the average ball	0.34694	0.02845	12.19481

If none of the part-worths for the levels within this ordered attribute (where we expect everybody to prefer higher performance to lower performance) has a significant t-value, then we might conclude that this attribute was not important. (More formally, the researcher could code this as a linear term such that there was a single coefficient to summarize the effect of performance with a single standard error for computing a more proper t-test of significance.)

Often researchers want to know if there is a statistically significant difference between levels of the same attribute. For example, the researcher may wonder if respondents on average prefer a golf ball that drives 15 yards further than the average ball (utility = 0.34694) compared to one that drives 10 yards further (utility = 0.12715). In other words, is the difference we observe in the model statistically significant? We can compute a t-value using the following formula:

$$t = \frac{U_1 - U_2}{\sqrt{SE1^2 + SE2^2}} \quad (3)$$

where U_1 and U_2 are the utilities of the two levels and the denominator computes the pooled standard error for the two levels, where SE1 and SE2 are the standard errors for U_1 and U_2. (Note, this is a shortcut formula that ignores covariances.) Thus, the t-value for the statistical test of difference in utilities between the levels is

$$(0.34694 - 0.12715) / \sqrt{0.02845^2 + 0.02916^2} = 5.39501$$

Since this t-value has an absolute magnitude greater than 2.58, we are greater than 99% confident that driving 15 yards is preferred to driving 10 yards further than the average ball for the population.

E. Testing for significant interaction effects

Similar to the test described in the previous section for testing dummy-coded part-worth terms for price vs. a linear term, you can test whether adding interaction effects to the model improves the model fit. The model shown in Exhibit 12.1 estimates the independent effects of three separate attributes: brand, performance, and price. We compare the log-likelihood of the model with main effects only to the log-likelihood of the model that includes these same main effects plus an interaction effect between, say, two attributes.

Prior to investigating potential interaction effects, we need to be confident that the experimental design supports the precise estimation of interaction terms. This particular data set uses a randomized design, where respondents are randomly assigned to one of hundreds of different blocks (versions) of CBC tasks, where each block features near-perfect one-way and two-way level balance. Such a design usually supports the efficient estimation of all potential first-order interaction effects (between attributes taken two at a time).

As shown in Exhibit 12.1, the log-likelihood fit for the main-effect only model is -4606.06 and the number of parameters in the model is 7. Let's imagine we hypothesize that the price coefficient may differ depending on the brand. To test the interaction effect between brand and price, we add additional interaction terms to the model (as described in Chapter 9). The brand attribute is effects-coded, with 4 - 1 = 3 parameters. Price is coded in this model (Exhibit 12.1) as a linear term. Thus, the interaction between brand and price adds 3 x 1 = 3 new parameters to the model. When we run the new model, the new model fit is a log-likelihood of -4593.39, representing an improvement versus the main-effects model of -4593.39 - -4606.05 = 12.66. Two times that amount (25.32) is distributed as chi-square, with degrees of freedom equal to the difference in the number of parameters estimated between the two models, or 3, leading to a p-value of 0.00001. Thus, we are 99.999% confident that the interaction term between brand and price adds significant fit to the model. If the previous test suggests a significant interaction, we should examine the relevant price and brand x price interaction effects for the augmented and improved model as shown below in Exhibit 12.2:

	Effect	**Std Err**	**T-Ratio**
Price (Linear Coefficient):	-0.22930	0.01058	-21.67339
High-Flyer Pro, by Smith and Forester x Price:	0.01431	0.01805	0.79264
Magnum Force, by Durango x Price:	0.07772	0.01785	4.35464
Eclipse+, by Golfers, Inc. x Price:	-0.00824	0.02034	-0.40530
Long Shot, by Performance Plus x Price:	-0.08379	0.02214	-3.78408

Interaction Effects between Brand and Price

Exhibit 12.2

Because we used effects-coding for the brand attribute, the interaction effects are zero-centered. The overall fit to the model improved with the addition of the three interaction terms, so we would expect at least one of the individual t-tests associated with the interaction terms to be statistically significant (actually, two of the t-ratios for interaction terms are significant beyond the 99% confidence level).

The interaction effects represent the adjustment to the price slope per brand after accounting for the main effect of price. The total price effect (slope) for Magnum Force is equal to -0.22930 + 0.07772 = -0.15158. Magnum Force is less sensitive to price than the average of the brands included in the experiment. Furthermore, the 0.07772 weight for the interaction term between Magnum Force and price is statically significant, with a t-ratio of 4.35464, or better than 99% confidence.

12.3 Multiple Independent Tests and the Alpha Inflation Problem

For statistical testing, it is typical to set a threshold for confidence such as 95%. Setting a confidence threshold of 95% means that you are setting the alpha (the likelihood of falsely declaring that an effect is significant when it actually is not) at 5% (called Type I error, or false positives). As you conduct more and more independent tests within the same "family" of hypotheses at a given alpha, the likelihood of falling prey to at least one false positive increases (called *experiment-wise error* or *family-wise error*). What exactly constitutes the same "family" is a bit elusive. It could mean the same kind of statistical test or multiple tests involving the same (or very similar) statistical constructs.

A. The Bonferroni Correction

Imagine you've conducted a CBC study that supports the estimation of all potential first-order interaction effects. Further imagine that in truth the only effects for the total population that are significant are the main effects (the independent effects of attributes). But, you've interviewed a sample and each respondent answers with error. Being the thorough researcher you are, you decide to conduct 2 log-likelihood tests to investigate *all* potential interaction effects taken one at a time. For example, with six attributes there are (6)(5)/2 = 15 pos-

sible interactions between attributes taken two at a time. With 10 attributes, there are (10)(9)/2 = 45 possible interaction effects.

For the first interaction effect you test, the likelihood of false positive is just 5%. But, after you have conducted 10 independent tests to investigate 10 potential first-order interaction effects (and this, we think, certainly involves tests within the same family), the likelihood of declaring a false positive has increased to $1 - 0.95^{10} = 40.1\%$! After 30 independent tests where in truth there are no significant effects for the population, the likelihood of finding at least one significant interaction at the 95% confidence level has increased to $1 - 0.95^{30} = 78.5\%$. The alpha inflation curve is illustrated in Exhibit 12.3.

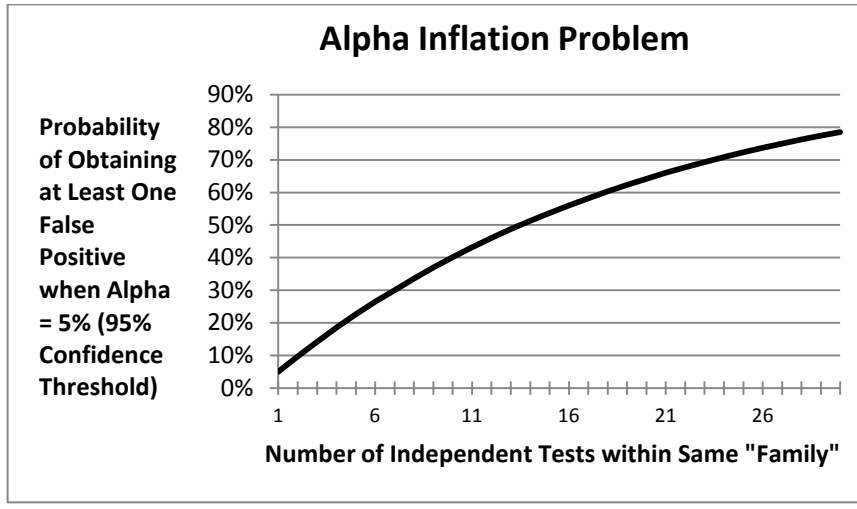

Exhibit 12.3

The well-known Bonferroni correction adjusts your alpha values when conducting multiple independent tests such that the likelihood of finding *any* false positives remains at your target alpha rate. It is performed by simply dividing the alpha rate by the number of independent tests within the same family. For example, if you want the likelihood of any false positives to remain at 5% over the course of 10 independent tests within the same family, you divide the alpha rate by 10, or 0.05/10 = 0.005. The new critical value for each of your tests is the value (e.g., from the f-test, t-test, or chi-square test) associated with 1 - 0.005 = 99.5% confidence.

Unfortunately, according to McDonald (McDonald 2014), "if you have a large number of multiple comparisons and you're looking for many that might be significant, the Bonferroni correction may lead to a very high rate of false negatives."

B. Benjamini-Hochberg (BH) Procedure

A more robust approach is the Benjamini-Hochberg (BH) procedure for controlling family-wise error, as it controls false positives without the danger of inflating the number of false negatives. The procedure works as follows:

1. Rank the p-values 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 for a 45 interaction study if we wanted a 0.05 experiment-wise 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.

In a 45 interaction example, assume these observed and critical p-values for the first five ranks (and that for all higher ranks the observed p-value exceeds the critical p-value):

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.

Note that we can make the process less mechanistic. For example, we might suspect that two of our tests in particular might reveal significant interactions. In this case, we might run those two tests each with p-value cutoffs of 0.05 but then use BH on the remaining 43

tests (alternatively, a theory-driven analyst might conduct just those two tests and not run the remaining 43 at all).

So Benjamini-Hochberg is a tool we can use to decide which interactions to include in our models when we want to avoid family-wise error.

Conjoint data tend to be robust, and the inclusion of a false-positive interaction effect in an aggregate logit CBC model is typically not very damaging to prediction. In other research situations such as in medical research, avoiding false positives is much more important, since it avoids money spent on further research to confirm the false positive or the potential embarrassment if new research cannot confirm your finding.

12.4 Frequentist Tests for Individual-Level Part-Worth Utilities

Individual-level modeling has become the norm, at least within the marketing research field for CBC experiments. HB is the most popular approach[27], but other methods are available as well, such as random parameters logit. Many conjoint analysts have become familiar with t-tests and f-tests and regularly use cross-tabulation software that conveniently produce these tests. It isn't surprising, then, that researchers lean on their familiar tools to conduct statistical testing for conjoint analysis when individual-level part-worth utility scores are available to them.

We should note that applying the traditional frequentist tests to individual-level part-worth utilities from HB (the point estimates, representing the average of many draws) is not quite proper, but we'll describe them anyway in this section before discussing the more proper Bayesian tests.

A. Normalization of individual-level utilities (Diffs)

Prior to using individual-level part-worth utilities (the point estimates from HB or OLS utilities for a ratings-based conjoint) in t-test or f-tests, we should normalize them so that each respondent's part-worth

[27] Prior to the popularity of HB analysis for CBC, early conjoint analysis researchers often used ratings-based conjoint methods with OLS regression estimation to produce individual-level part-worths. It was common to use these part-worth utility scores in cross-tabulation packages for tabulating the data by different segments.

utilities are roughly on the same scale. A common way to do this is a method that Sawtooth Software calls *Diffs*, but there are other procedures (such as equalizing the standard deviation of the scores within respondents) that also could be done. For Diffs we find the multiplier for each respondent to apply to this respondent's part-worths such that the resulting sum of the differences between best and worst levels across attributes is equal to some constant such as the number of attributes times 100 (or you could choose any other target constant such as 10 or 1).

B. F-tests and matched sample t-tests

Now that you have normalized the utilities (the point estimates from HB or OLS utilities for a ratings-based conjoint), you can conduct the standard t-tests and f-tests. For comparing groups of people on the part-worth utilities, the f-test is commonly used.

For comparing levels within attributes, we recommend matched sample t-tests. Let's assume the normalized average part-worth utility for level 2 is higher than level 1 for the sample (within the same attribute). We want to know if that difference is statistically significant. For each respondent, we create a new variable called DiffUtil21 equal to Util2 - Util1, where Util2 is the utility for level 2 and Util1 is the utility for level 1. The t-test is computed by taking the mean of DiffUtil21 and dividing it by the standard error of DiffUtil21. If the resulting t has an absolute magnitude of 1.96 or greater, we are at least 95% confident that there is a statistically significant difference between the utilities of these two levels.

Both f-tests and matched sample t-tests can be performed on attribute importance scores as well. Attribute importance scores are computed by percentaging the ranges of attributes within each individual.

12.5 Bayesian Tests for HB Estimation

Upon reading this title for Bayesian tests for HB, you might have assumed that Bayesian tests are going to be harder to understand than the frequentist tests we've described to this point. Not true! Bayesian statistical tests are actually easier to understand and implement. Don't believe us? Keep on reading.

HB estimation (described in Chapter 10) leads to 1000s of posterior draws 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 prior to assuming convergence (typically the first 5K or 10K draws). We can use the remaining draws (which our HB estimation program may have written to convenient .CSV files) to conduct statistical tests and it just involves counting!

Statistical testing for HB requires a shift in mindset. For example, you do not test to see if a new parameter adds significant fit to the model. There isn't anything akin to the 2 log-likelihood test for aggregate logit that we apply to HB models. Rather, you examine the distributions of posterior draws of parameters to see if a strong majority of these draws (a preponderance of the evidence) falls on either one side or the other of the null hypothesis. With HB, you actually do not want to maximize the fit of the lower-level model (the individual-level models). You strike a compromise between fitting the lower- and upper-level models that you believe based on your priors will near-maximize the fit to new observations not included in the model.

A. Confidence intervals and HB

Let's imagine we conducted a CBC experiment, estimated the part-worth utilities via HB, and we were interested in developing a 95% confidence interval for the price coefficient. 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 price coefficient. Imagine that you have 20,000 total draws, you throw away the first 10,000 and the next few draws of alpha look something like the following:

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

To develop the 95% confidence interval, you simply sort the 10,000 used draws from lowest to highest and then pick out the 2.5th

and 97.5th percentile values. That's the 95% confidence interval (it contains 95% of the estimates)! (Note: the same approach works for individual-level draws for developing confidence intervals within individuals on a parameter of interest.)

B. Significance of attributes and levels with HB

Following the previous example, we can compute the degree of confidence that the price coefficient is less than 0. We simply count for how many of the used draws of alpha (those after convergence) the price coefficient is less than 0. If 99.925% of the draws of the price coefficient are negative, then we are 99.925% confident that the price slope is significantly less than zero.

C. Tests for differences between levels within attributes with HB

Let's imagine that we have successive estimates of alpha for the performance attribute and the first five (after convergence) look like:

Draw#	5 yards further	10 yards further	15 yards further
10001	-0.25136	-0.12476	0.37612
10002	-0.24921	0.01873	0.23048
10003	-0.22384	-0.08757	0.31141
10004	-0.21998	-0.08897	0.30895
10005	-0.23179	0.00622	0.22557

If we want to compute the degree of confidence that driving 15 yards further is better than driving 10 yards further for a golf ball, we simply count for how many draws of alpha the part-worth utility estimate for 15 yards further is greater than 10 yards further. If 99.17% of the alpha draws had 15 yards preferred to 10 yards, then we are 99.17% confident that driving 15 yards is preferred to 10 yards.

D. Tests for differences between product concepts with HB

We can sum part-worth utilities using estimates of alpha (after convergence) to compare whether the preference for one product concept exceeds that of a different product concept. For each alpha draw after convergence is assumed, sum the part-worth utilities associated with the two product concepts you are testing. Count for how many

draws of alpha one product concept's total utility exceeds the other concept's total utility. If 99.17% of the alpha draws show that the total utility for concept 1 exceeds concept 2, then we are 99.17% confident that concept 1 is preferred to concept 2.

E. Tests for differences between groups of respondents on attribute levels with HB

If you employ covariates in your HB model based on group membership (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 for males for all the parameters in your model) followed by the alpha parameters associated with the adjustment from the intercept parameters for females. In other words, the average female utility for a parameter is equal to the intercept plus the female covariate value for that parameter. If we wanted to know if males preferred the attribute level "drives 15 yards further than the average golf ball" more than females (relative to the other two levels of this attribute), we count for what percent of the used draws the males' utility for that level exceeds the females' utility. If 92.53% of the draws show the males' mean utility exceeds the females' utility, then we are 92.53% confident that males value this level more than females (relative to the other two levels of this attribute).

F. Tests for interaction effects with HB

Most of the interaction effects we observe in aggregate models like aggregate logit or (low-dimension) latent class solutions for CBC are due to unrecognized heterogeneity. Under individual-level estimation via an approach like HB, you will only occasionally find that interaction effects help a CBC model make significantly better predictions of choice. Still, it is helpful to select experimental designs that permit the precise estimation of all first-order interaction effects and then test whether the inclusion of the most promising ones can improve the model's predictions. Even then, it is rare to add more than just one or a *very* few interaction effects to a CBC HB model. HB models are

much slower to run than aggregate logit and including more than just a few interaction effects between attributes taken two at a time might lead to extremely long run times.

When using effects-coding of categorical attributes, both the main effects and interaction effects are zero-centered. HB estimation provides a straightforward way for you to test interaction effects by enumerating the used alpha draws. For each interaction term, you can count what percent of the used draws fall on either side of 0. For example, if the interaction effect between Brand A and Price 1 leads to 99% of its draws of alpha being greater than zero, you are 99% confident this interaction is statistically significant.

The problem with this Bayesian test of significance for interaction effects is that it is extremely sensitive and will tend to identify interaction effects well beyond the 99% confidence level that, while statistically significant, make little change to the model predictions and slow HB estimation down quite a bit. One approach we've implemented at Sawtooth Software is to jack-knife repeatedly across choice tasks, each time holding out one or a few (randomly selected) of the calibration CBC tasks for validation. For each replication, we estimate HB models using the remaining tasks and then predict the held out tasks. We compare the hit rate (averaged across all the replicates) for main-effects only models to the hit rate when an interaction effect is also included. This jack-knife procedure must be repeated many times to stabilize the hit rates, so it is extremely helpful if this can be run in parallel (across multiples machines or multiple cores on the same machine). In our experience, we look for interaction effects that can increase the hit rate for holdout tasks (relative to the main-effects model) by perhaps 1%, 2%, or more.

12.6 Frequentist Tests for Shares of Preference

In Chapter 14, we describe how to specify simulation scenarios (typically an array of competitive product offerings) and use the part-worth utilities within a choice simulator to predict what percent of the respondents would choose each of the product alternatives. Practitioners tend to use individual-level models (such as estimated via HB) when constructing such choice simulators.

A. Confidence intervals

Given the types of choice simulators (often built in Excel) commonly used in practice that usually employ point estimates (a single vector of part-worths for each respondent) to predict choices, most practitioners default to the frequentist tests and traditional ways of developing confidence intervals. For each individual, we obtain shares of preference that sum to 100%. This leads to quite easy calculations of the standard deviation across respondents for those shares and the standard error estimates which are given by dividing the standard deviation by the square root of the sample size. The 95% confidence interval for any share of preference can be obtained by taking the mean share of preference +/- 1.96 times its standard error.

B. Testing differences in shares of preference

We recommend using a matched sample t-test, as follows. Let's assume the shares of preference for each individual are available and we see that the share of preference for product 2 is preferred to product 1 for the sample. We want to know if that difference is statistically significant. For each respondent, we create a new variable called DiffSOP21 equal to SOP2-SOP1, where SOP2 is the share of preference for product 2 and SOP1 is the share of preference for product 1. The t-test is computed by taking the mean of DiffSOP21 and dividing it by the standard error of DiffSOP21. If the resulting t has an absolute magnitude of 1.96 or greater, we are at least 95% confident that there is a statistically significant difference between these two shares of preference.

12.7 HB Testing for Product Shares of Preference

Applying frequentist approaches to develop confidence intervals or t-tests on shares of preference that were simulated from the point estimates of beta (within-respondent averages of the beta draws) is not proper and overstates our confidence, while understating the width of the confidence intervals. It ignores the uncertainty in the individual-level draws. Leading academics and practitioners recommend that if it is very important to obtain proper estimates of precision (such as in litigation), then one should simulate shares of preference from the upper-level model (one that includes useful and significant covari-

ates). The draws for shares of preference from the upper-level model simulations then can be enumerated to obtain confidence intervals and for statistical testing (such as for testing whether a product that contains a superior feature is better than another product without that feature). It is well beyond the scope of this chapter to delve into this topic and we refer the reader to a tutorial on this subject presented at the American Marketing Association's ART/Forum by Greg Allenby (Ohio State) and Tom Eagle (Eagle Analytics of California). We recommend you contact these authors directly to request the materials (Allenby and Eagle 2014).

12.8 Appendix: The Swait and Louviere Test for Between-Group Differences in MNL Utilities

A common question arises when we wonder whether a statistical model has the same coefficients for different groups of respondents. In linear regression models we can use something called a Chow test (Chow 1960). Imagine we have a regression model we've built and we want to know whether respondents in Poland have the same coefficients as respondents from Belgium. To conduct a Chow test we run two models. In the first, run a regression analysis with the whole sample, including respondents from both Poland and Belgium. In the second model, we add a dummy variable for country (1 = Poland, 0 = Belgium) and we build a model that has (a) the main effects for all of our independent variables, plus (b) the main effect for country, and (c) the interactions of the independent variables with the country variable. We want to see if the expanded model, the one with the dummy variable and the interactions, fits the data significantly better than does the other model. We can test this using a standard test for improvement in R^2 for nested regression models.

The aggregate multinomial logit models we use for choice experiments, however, have a complication that prevents us from using a test as simple as the Chow test. We know that the logit model has a scale factor as described in Chapter 10. The scale factor affects the size of the utilities such that respondents who answer more consistently (that is 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. Following our example above, we can imagine a situation where Poles and Belgians have the same preferences

for the attributes and levels but that the Poles answer their choice questions more consistently. In this case, the models might look different, because the Poles would have larger positive and larger negative utilities while the Belgians' utilities would be smaller. 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. First, run the four models:

1. We use the data matrix from the first group to be compared, X_1, to run a 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, multiplying the values of X for second data matrix by a positive constant, thus 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:

$$\lambda_A = -2 [LL_\mu - (LL_1 + LL_2) \quad (4)$$

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:

$$\lambda_B = -2 (LL_P - LL_\mu) \quad (5)$$

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 they 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.