**File Descriptions**

SpecSheet - AttributeAppraisal.xlsx: A description of the required inputs and outputs for the RAT-STATS unrestricted attribute appraisal module.

Examples - AttributeAppraisal.xlsx: Two example RAT-STAT inputs.

VBCode - AttributeAppraisal.xlsm: Macro that replicates the RAT-STATS implementation of the unrestricted attribute appraisal module.

Also refer to pages 36 through 39 (sections 2-2 through 2-5) in the RAT-STATS 2010 Companion Manual and pages 106 through 113 (sections 3-4 through 3-11) of the RAT-STATS 2010 User Manual. Both documents can be found on the following page <https://oig.hhs.gov/compliance/rat-stats/>. The data limitations associated with RAT-STATS can be found on pages 386 through 393 (A-1 through B-5) of the RAT-STATS 2010 User Guide.

**Exception Checking**

Your program does not need to provide the exact same error messages as RAT-STATS, but exceptions handled by RAT-STATS must be handled by your submissions as well. Exceptions will be checked through the use of test datasets.

**General Overview**

If a user pulls a simple random sample and counts the number of characteristics of interest within the sample, then the unrestricted attributed appraisal module can be used to calculate a valid statistical estimate based on the sample results. The unrestricted attribute appraisal module uses what is known as an “exact” interval based on a hyper geometric distribution. The interval is referred to as being “exact” because it is guaranteed to provide at least the target confidence level regardless of the size of the sample or composition of the frame.

In practice, RAT-STATS calculates the lower limit for a unrestricted attribute appraisal by finding the smallest number of errors in the universe where the probability of observing the sample error count or greater is less than the alpha level (i.e., [1-confidence level]/2). The upper limit is then calculated by finding the largest number of errors in the universe where the probability of observing the sample error count fewer is less than the alpha level. In simpler terms, the approach is based on finding the largest and smallest frame for which chance of observing the sample is just unlikely enough to match the target confidence level.

For example suppose an individual selected 10 transactions from a universe of 100, observes 4 errors, and wants to calculate a two-sided 90 percent confidence interval. To find the lower limit of a 90 percent confidence interval we need to identify the smallest number of errors that there could be in the 100 items such that there is at least a 5 percent chance of observing 4 or more errors in our sample of 10 items. A simple case would be if there were 0 errors in the universe. In this case the chance of observing 4 or more errors in our 10 samples would be 0 percent, because we cannot have 4 errors in the sample if there are 0 errors in the universe!

What about when the frame has 16 errors? In this case the chance of selecting 4 or more errors in a sample of 10 is 5.149 percent (we will talk about how this is calculated later). If the frame only had 15 errors the probability drops to 4.08 percent. Recall that for the lower limit, the goal is to find the number of errors in the 100 samples such that the chance of observing results greater than or equal to the sample results is just unlikely enough. “Unlikely enough” being defined as [1-confidence leve]/2, or 5 percent for a 90 percent confidence level. It follows directly that the lower limit for this case is 16 errors (or 16 percent since the universe contains 100 items).

The upper limit is identified in a similar fashion. First we note that the probability of observing 4 or fewer errors in the sample given 68 errors in the universe is 5.39 percent. If the universe contained 69 errors then the probability of observing 4 or fewer errors drops to 4.59 percent. The upper limit is defined based on the largest number of errors in the universe such that the chance of observing 4 or fewer errors in the sample is greater than 5 percent. It follows directly that the upper limit for this example is 68 errors (or 68 percent since the universe contains 100 items).

Consider what happens when the numbers from this example are entered into RAT-STATS. In particular, we can enter 4 for the items with the characteristic of interest, 100 for the universe, and 10 for the sample size. With these results the resulting 90 percent confidence interval ranges from 16 to 68. Notice how these align with the numbers identified in the previous paragraph.

The gap in the above explanation is that it does not explain how to calculate the probability of observing a number of errors less than or equal[[1]](#footnote-1) to our sample error count given a particular number of errors in the universe. This probability is known as the cumulative distribution for a hypergeometric distribution. There is no direct approach that I am aware of for calculating this function. A simpler, but still complex calculation is for the probability of observing a specific number of errors in the sample for a particular number of errors in the universe (i.e. the hypergeometric function rather than the cumulative hypergeometric function).

For example, the cumulative hypergeometric distribution function outputs the probability of observing 4 or fewer errors in our sample given, for example, 50 errors in the universe. The hypergeometric distribution can only provide the probability of a single number of errors. Consequently, it could be used to calculate the probability of observing exactly 4 errors in our sample given 50 errors in the universe. The hypergeometric distribution can be used to calculate the cumulative hypergeometric function. In the present case, you could calculate the probability of exactly 4 errors, exactly 3 errors, exactly 2 errors, etc. and sum them together. This latter calculation is, in fact, the most common approach for calculating the probabilities for the cumulative hypergeometric distribution.

The function for the hypergeometric distribution cannot be programmed directly since the numbers involved will lead to an overflow for large sample sizes and/or universe counts. The provided Excel macro include one efficient calculation of the hypergeometric distribution; however, others exist as well.

A final hurdle involves searching the possible error counts in the universe to find the upper and lower limits. It would be unnecessarily time consuming to search through all possible universe counts. An easy way to cut-down on the search time is to choose a good starting point. One approach along these lines is to first calculate an exact interval based on a binomial distribution (i.e., the Clopper Pearson interval). The Clopper Pearson interval has a closed solution and is guaranteed to be wider than an interval based on the hypergeometric distribution. Consequently, one can use the Clopper Pearson interval to set the starting points for the number of errors in the universe and then adjust the universe error counts until the upper and lower limits are found. Any search method can be used as long as it leads to the correct result and can be completed within a reasonable time.

1. The logic of this paragraph applies equally to calculating the probability of observing results greater than or equal to our sample results. [↑](#footnote-ref-1)