Summary of Analyzing and Summarizing Data

Introduction

Typically, you want to do all your analyzing, summarizing, and modeling of large data in CAS. Remember, CAS stores entire tables in memory for multi-threaded parallel processing, resulting in extremely fast results. When the processing is complete, CAS can return a subset of the data or summarized results to the compute server for additional processing, visualization, and reporting.

Analyzing Data with CAS Actions

- Generate simple descriptive statistics of numeric variables including the sample mean, sample variance, sample size, and sum of squares:

```
simple.summary /
      table={castable},
      inputs={column-names},
      subset={summary-statistics},
      casout={casouttable},
      <, additional parameters>;
```

- Perform aggregation on selected variables with additional flexibility and control:

```
aggregation.aggregate /
      table={castable},
      varSpecs={
                  {name="column-name", agg="summary-statistic" | subset={summary-
statistics},
                  {colNames={column-names},
                  <,{additional columns}>
      };
      saveVariableSpecification=TRUE | FALSE,
      casout={casouttable}
      <,{additional columns}>;
```

- Generate a frequency distribution for one or more variables:

```
simple.freq /
      table={castable},
      inputs={column-names},
      casout={casouttable}
      <,additional parameters>;
```

- Construct the frequency and crosstabulation tables:

```
freqTab.freqTab /
      table={castable},
      tabulate={
                  {vars={column-names},<, cross={column-names}>}
```

```
                <, {freqTab_tabulate-n}, ...>
    }
    tabdisplay="CROSSLIST" | "LIST",
    includeMissing=TRUE | FALSE,
    order="FORMATTED" | "FREQ" | "INTERNAL"
    <,additional parameters>;
```

- Performs one-way or two-way tabulations:

```
simple.crossTab /
    table={castable},
    row="column-name",
    col="column-name",
    weight="column-name",
    aggregator="aggregator-value",
    <, additional parameters>;
```

Visualizing and Reporting

- There are two programming options to create visualizations.
- The first and easiest option is to reference a CAS table in a SAS ODS graphics procedure. This works well on data that is not extremely large.
- If a CAS table is referenced in an ODS Graphics procedure, CAS will send the ENTIRE table to the compute server for processing.
    - If the CAS table is a manageable size, the compute server will process the data and create a data visualization
    - If the table exceeds the transfer size limit, an error is returned. This protects you from transferring extremely large tables.
    - The table size limit that CAS can send to the compute server is typically set by the administrator