SESUG Paper 85-2025

AP Statistics and SAS

Immanuel Pulavarti, Central Bucks West High School

ABSTRACT

This paper explores the integration of SAS software into the AP Statistics curriculum to promote a deeper understanding of statistical concepts and data analysis. By leveraging the power of SAS, students gain practical experience with real-world data, enhancing their ability to perform descriptive and inferential statistical procedures such as hypothesis testing, confidence intervals, regression analysis, and probability modeling. The use of SAS encourages active learning by allowing students to visualize data distributions, automate calculations, and interpret complex outputs. Through hands-on coding and data manipulation, students develop computational thinking and analytical skills that are increasingly vital in academic and professional environments. Incorporating SAS into AP Statistics bridges the gap between theoretical knowledge and applied statistics, better preparing students for college-level coursework and future careers in data science and related fields.

INTRODUCTION

AP Statistics focuses on how to collect, analyze, and interpret data. Students learn to summarize patterns, make predictions, and test claims using statistical methods. SAS is a software platform that supports these skills by providing tools for organizing data, running calculations, and creating graphs. Using SAS alongside AP Statistics helps connect theory to practice, giving students a way to explore concepts with real datasets. This paper describes how SAS can be applied to key AP Statistics topics. Some topics not included - probability, sampling methods, normal distributions, confidence intervals, hypothesis testing, correlation and regression, and simulation-based inference can also be better understood using SAS.

DESCRIPTIVE AND EXPLORATORY ANALYSIS

Descriptive and exploratory analysis is the first step in understanding a dataset. In AP Statistics, this includes summarizing numerical data, examining distributions, and identifying patterns before performing formal inference. These methods help reveal the shape, center, and spread of the data. SAS offers several procedures for this purpose, such as PROC MEANS and PROC FREQ for numerical and categorical summaries, and PROC SGPLOT for visualizing relationships and distributions. The following subtopics show how these tools can be applied to explore data in ways that align with the AP Statistics curriculum.

SUMMARIZING NUMERICAL DATA

A key part of descriptive analysis is calculating summary statistics for numerical variables. SAS's PROC MEANS makes this process simple and efficient. For example, using the built-in sashelp.cars dataset, we can summarize the MPG_City variable:

```
/* Summarize city miles per gallon for cars */

proc means data=sashelp.cars mean median min max std;

    var MPG_City;

run;
```

This code produces the mean, median, minimum, maximum, and standard deviation for city miles per gallon across all cars in the dataset. Using real SAS datasets like sashelp.cars ensures the code runs without errors while illustrating how to perform descriptive statistics in practice.

DATA MANAGEMENT AND PREPARATION USING SAS

Once initial descriptive and exploratory analyses are complete, the next step in the analytical process is preparing the data for deeper investigation. Data rarely arrives in a perfectly clean and analysis-ready form; instead, it often requires filtering, sorting, recoding, merging, or creating new variables. In AP Statistics, these preparation steps ensure that datasets match the requirements of the statistical methods being applied, such as randomization conditions, grouping structures, or transformation needs. SAS offers a comprehensive set of tools for managing and preparing data. The DATA step allows for precise observation-level manipulation, while procedures such as PROC SORT, PROC TRANSPOSE, and PROC SQL enable efficient reorganization and integration of datasets. Built-in functions further support recoding variables, creating computed values, and handling missing data. The following subtopics will demonstrate how these features can be applied in practical scenarios, from filtering records based on specific conditions, to creating new variables for calculated measures, to combining multiple datasets for a unified analysis. These techniques not only support the AP Statistics curriculum but also mirror the processes used in real-world data analysis projects.

FILTERING AND SELECTING DATA

Filtering and selecting data is a critical step in preparing a dataset for analysis. By removing irrelevant observations, analysts can focus on the cases that meet the requirements of the study. In AP Statistics, this might involve filtering out extreme outliers, selecting data from a specific category, or narrowing the sample to meet randomization conditions. In SAS, filtering can be accomplished using the WHERE statement in many procedures or within a DATA step. The example below uses the built-in sashelp.cars dataset to display only vehicles with more than 30 miles per gallon in the city:

```
proc print data=sashelp.cars;
   where mpg_city > 30;
run;
```

Output 1 shows the resulting subset of cars that meet the MPG condition. Filtering in this way ensures that subsequent analyses are relevant to the research question and do not include extraneous data.

Using WHERE Statements

The WHERE statement in SAS is a flexible way to filter observations based on specific conditions. It can be used within many procedures, including PROC PRINT, PROC MEANS, and PROC FREQ. Using WHERE allows analysts to focus on a subset of the data without creating a new dataset, which is efficient for exploratory analysis. The following example filters the sashelp.cars dataset to display only vehicles with more than 30 miles per gallon in the city that are made by Toyota:

```
proc print data=sashelp.cars;
   where mpg_city > 30 and make = "Toyota";
run;
```

Output 2 shows the resulting subset of cars that meet both conditions. This method ensures that the analysis is restricted to relevant observations while keeping the original dataset intact.

Using PROC SQL Filters

PROC SQL provides an alternative method for filtering data in SAS. Unlike the WHERE statement, which is used within many procedures, PROC SQL allows you to both filter and manipulate data in a single step, using standard SQL syntax. This approach is especially useful when combining multiple datasets or performing more complex queries. The following example selects all vehicles from the sashelp.cars dataset with city miles per gallon greater than 30 and made by Toyota:

```
proc sql;
   from sashelp.cars
   where mpg_city > 30 and make = "Toyota";
quit;
```

This displays the filtered results. Using PROC SQL in this way ensures that analysts can efficiently subset data while retaining full control over selection conditions.

CREATING AND MODIFYING VARIABLES

Creating and modifying variables is a key step in preparing data for analysis. New variables can be derived from existing data to capture additional insights, calculate rates or ratios, or recode categorical values for easier analysis. In AP Statistics, this aligns with tasks like creating a new variable for residuals, differences, or standardized scores. In SAS, the DATA step allows analysts to create new variables or modify existing ones efficiently. The following example demonstrates how to calculate a new variable, MPG_Difference, representing the difference between highway and city miles per gallon in the sashelp.cars dataset:

```
data cars_modified;
   set sashelp.cars;
   MPG_Difference = MPG_Highway - MPG_City;
run;
```

Output 3 shows the first few rows of the dataset with the new MPG_Difference variable added. This process enables further analysis on derived measures, supporting more in-depth statistical exploration.

Creating Computed Variables

Computed variables allow analysts to derive new information from existing data. These variables can represent differences, ratios, or other mathematical transformations that provide additional insight for statistical analysis. In AP Statistics, computed variables are often used to calculate differences between groups, percentages, or standardized scores. The following SAS example creates a new variable, MPG_Ratio, representing the ratio of highway to city miles per gallon in the sashelp.cars dataset:

```
data cars_computed;
    set cars_modified;
    MPG_Ratio = MPG_Highway / MPG_City;
run;
```

Output 4 displays the first few rows of the dataset with the newly created MPG_Ratio variable. Using computed variables in this way allows for more nuanced analysis and prepares the dataset for further statistical procedures.

Recording Categorical Variables

Recoding categorical variables is often necessary to simplify analysis, combine levels, or create meaningful groupings. In AP Statistics, this might involve grouping similar categories together to increase sample sizes or to make comparisons more interpretable. The following example recodes the Type variable in the sashelp.cars dataset to create a new variable, Car_Type_Group, which groups “SUV” and “Truck” as “Large” and all other types as “Small/Medium”:

```
data cars_recoded;
    set work.cars_modified;
    if Type in ("SUV","Truck") then Car_Type_Group = "Large";
    else Car_Type_Group = "Small/Medium";
run;
```

Output 5 shows the dataset with the new Car_Type_Group variable, demonstrating how recoding can simplify categorical data for analysis. This approach ensures clearer interpretation and more effective comparison of groups.

SORTING AND ORGANIZING DATA

Sorting and organizing data is an essential step in preparing datasets for analysis. Properly sorted data makes it easier to identify trends, perform group comparisons, and generate accurate reports. In AP Statistics, sorting might be used to arrange data by a categorical variable for group summaries or by a numerical variable to identify extreme values. SAS provides the PROC SORT procedure to arrange datasets efficiently. The following example sorts the sashelp.cars dataset by the Make variable in ascending order and by MPG_City in descending order:

```
proc sort data=sashelp.cars out=work.cars_sorted;
   by Make descending MPG_City;
run;

proc print data=work.cars_sorted(obs=10);
run;
```

This displays the first ten rows of the sorted dataset, showing cars arranged alphabetically by make and then by city miles per gallon in descending order. Sorting data in this manner simplifies subsequent analysis and ensures results are presented clearly and logically.

CONCLUSION

Effective data analysis begins with a solid understanding of the dataset, followed by careful preparation to ensure accuracy and clarity. Descriptive and exploratory methods were applied to summarize numerical and categorical data, revealing patterns and distributions that inform subsequent statistical procedures. SAS procedures such as PROC MEANS, PROC FREQ, and PROC SGPLOT provide powerful tools for these initial analyses, allowing students to visualize and quantify key characteristics of the data in a way that reinforces AP Statistics concepts. Beyond exploration, data management and preparation are essential to producing reliable results. Techniques such as filtering with WHERE statements or PROC SQL, creating computed variables, recoding categorical data, and sorting datasets ensure that the data is structured appropriately for analysis. By applying these methods on the sashelp.cars dataset, students can see how SAS not only streamlines data handling but also deepens understanding of statistical principles, including group comparisons, derived measures, and focused inference. By integrating exploratory analysis with systematic data

preparation, SAS enables students to approach statistical questions with confidence, accuracy, and efficiency. Mastery of these tools strengthens comprehension of both the data and the underlying statistical concepts, making learning interactive, practical, and reproducible. Ultimately, these skills provide a clear path from raw data to meaningful conclusions, demonstrating how SAS facilitates both statistical analysis and the learning process in AP Statistics.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Immanuel Pulavarti
267-895-0288
immanuelap@yahoo.com