**Biostatistics Lab using SAS Software**

Graduate Education in Public Health

Fridays (10:00–11:30 am ET)

Location: Room NCPC 206

**Course Director/Instructor:**

Yvonne Phillips

Office Hours: By Appointment

Email: [yphillips@msm.edu](mailto:yphillips@msm.edu)

# COURSE DESCRIPTION

This 6-week syllabus is designed to provide beginners with a comprehensive introduction to data analysis using SAS. Throughout the course, participants will gain hands-on experience, learn essential data manipulation and analysis techniques, and complete a final project to apply their newfound knowledge.

# COURSE OBJECTIVES

Students who successfully complete this course will be able to:

* Develop students' proficiency in analyzing and manipulating datasets. Through this course, students will enhance their abilities in several aspects of data handling, including preprocessing, exploration, and visualization.
* Provide students with practical expertise in the areas of data mining, visualization, and communication. This will be achieved by engaging students in hands-on projects that utilize spreadsheet-based technologies and the SAS programming language.

Tools used for data analysis include Excel, and SAS software. SAS software includes very rich learning resources, the ability to perform modeling, visualization, reports, dynamic charts, etc.

**TECHNOLOGY REQUIREMENTS**

**Computer/Internet Access/Virtual Computer Lab account**

* PC or laptop with Windows XP or above, a screen 10 in or greater and an Internet connection, preferably DSL or cable.
* Morehouse School of Medicine email account, Word processing software e.g., Microsoft Word or Open Office,
* A wired Internet connection (DSL, LAN, or cable connection is desirable with a Bandwidth of 512Kbps for connecting to Blackboard Learn course management system is recommended).
* Personal computer and internet connection problems do not excuse the requirement to complete all coursework in a timely and satisfactory manner. Each student needs to have a backup method to deal with these inevitable problems. These methods might include the availability of a backup PC at home or work, the temporary use of a computer at a friend's home, the local library, office service companies, Starbucks, a TAMUC campus open computer lab, etc.

# COURSE GRADING

Grades will be determined as follows:

* + **Project 4 10%**

# TOTAL 10%

**Course Components Lectures**

The class is a synchronous class that meets twice a week for lectures and joint class activities. The class activities are designed to help you master the relevant materials, to work on your homework, and get you started on your project.

**Class participation (10%)**: Students will be evaluated based on their participation in activities and discussions during class hours. All students must be punctual and attend all scheduled lectures unless excused. The Course Director must excuse any absences in advance, except for unavoidable emergencies. However, such emergencies are rare events. Attendance and punctuality are considered part of the participation grade. From my previous experience, attendance and final grades tend to be correlated.

**Homework:** The homework is going to provide an opportunity to learn data science skills and to test your understanding of the material. See the homework as an opportunity to learn, and not to “earn points”. The homework will also be graded to reflect this objective.

**Project:** The project is an individual contribution, in a few weeks, you will start to work on a data science project. The goal of the project is to go through the complete data analytics to answer questions you have about some topic of your own choosing. You will be provided with the data. You will design your visualizations, run statistical analyses, and communicate the results.

**NOTE:** This syllabus is subject to change at the discretion of the instructor(s). Any revisions will be communicated to students and reflected on the syllabus document posted in Canvas.

# GRADING SCALE:

|  |  |
| --- | --- |
| **GRADE** | **GRADE POINT** |
| **A** | 96-100 |
| **A-** | 90-95 |
| **B+** | 85-89 |
| **B** | 80-84 |
| **B-** | 75-79 |
| **C** | 70-74 |

|  |
| --- |
| **COURSE SCHEDULE AND IMPORTANT DATES** |

**The tentative weekly course of instruction**

|  |
| --- |
| Week 1  Introduction to SAS |
| Introduction to the course, its objectives, and the importance of data analysis in various fields. Overview of SAS as a powerful tool for data analysis.   * Course Overview and Introduction to Data Analysis * Installing and Setting Up SAS   + Detailed instructions on how to install SAS software (if not already installed) and set up the environment.   + Explanation of the SAS interface and different components. * Exploring the SAS Interface   + A hands-on tour of the SAS interface, including the SAS Editor, Log, and Output windows.   + Introduction to the SAS code editor and how to execute code. * Reading Data that is Aligned in Columns: The data is aligned in columns and SAS reads the data with column input. (see page 59 of the manual)   + Working with the INPUT statement   SAS Manual (PDF): [Step-By-Step Base SAS Programming Manual](https://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_913/base_step_10071.pdf) |
| Week 2  Introduction to Data Exploration |
| * Explanation of data exploration's significance in the data analysis process. * The LIBNAME function * Working with SAS Data Sets   + Understanding the structure of SAS data sets.   + Importing data from a CSV file into a SAS data set.   + Basic data manipulation techniques like sorting and filtering data. * Introduction to the SAS DATA step for data manipulation. * Overview of the key steps involved in data exploration. |
| Week 3  Introduction to Data Exploration |
| In this section, you will learn how to do the following: display information about a SAS data set create a new SAS data set from an existing SAS data set rather than creating it from raw data records.   * Using DATA step for tasks such as creating new variables, subsetting data, and summarizing data. * Descriptive Statistics with SAS   + Introduction to basic descriptive statistics: mean, median, mode, standard deviation, and variance.   + Using SAS functions like MEAN, MEDIAN, and STD to calculate these statistics. |
| Week 4  Data Visualization with SAS |
| * Introduction to data visualization principles.   + Creating basic plots like bar charts, scatter plots, and histograms using SAS procedures. * Customizing Graphs and Plots   + Exploring customization options in SAS for modifying plot appearance, labels, and titles.   + Adding features like legends and annotations to enhance visualizations. |
| Week 5  Data Visualization with SAS |
| TBD |
| Week 6  Statistical Tests with SAS |
| TBD |

Students are introduced to several fundamental SAS functions and procedures that form the basis for data analysis and data exploration.

1. Introduction to Data Analysis and SAS
   1. DATA Step: The DATA step is fundamental in SAS. It's used for data manipulation, creating new variables, and subsetting data.
   2. PROC IMPORT: This procedure is used to import data from external sources like CSV, Excel, or databases into SAS datasets.
   3. PROC SORT: Sorts data observations based on one or more variables.
   4. SET Statement: Used to read data from an existing dataset into a new dataset.
   5. MERGE Statement: Merges two or more datasets based on common variables.
   6. FORMAT Statement: Allows you to apply user-defined formats to variables for better data representation.
2. Data Exploration and Visualization: Students will learn how to manipulate, summarize, and visualize data, which are essential skills for understanding and interpreting data effectively.
   1. MEAN, MEDIAN, and STD functions: Used for calculating the mean, median, and standard deviation of numerical variables.
   2. FREQ Procedure: Helps create frequency tables for categorical variables, showing the distribution of categories.
   3. UNIVARIATE Procedure: Generates various descriptive statistics like histograms, box plots, and percentiles for exploring data distributions.
   4. SGPLOT Procedure: Used for creating customizable and informative graphs, including bar charts, scatter plots, and histograms.
   5. CORR Procedure: Calculates the correlation matrix to explore relationships between numerical variables.
   6. REG Procedure: Conducts simple linear regression analysis to explore relationships between dependent and independent variables.
3. For Numeric Variables (Continuous Data):
   1. Graphical Analysis:
      1. PROC UNIVARIATE: This procedure provides comprehensive graphical and non-graphical analysis of numeric variables. It can produce histograms, box plots, quantile-quantile (Q-Q) plots, and various summary statistics.
      2. PROC GPLOT: This procedure is useful for creating scatter plots, line plots, and other graphical representations of numeric variables.
      3. PROC SGPLOT: Offers advanced capabilities for creating customizable and informative graphs, including histograms, density plots, and box plots for numeric data.
   2. Non-Graphical Analysis:
      1. MEAN, MEDIAN, and STD functions: These functions are fundamental for calculating the mean, median, and standard deviation of numeric variables.
      2. PROC MEANS: Provides various statistics such as mean, median, minimum, maximum, and percentiles for numeric variables.
      3. PROC SUMMARY: Offers a way to compute summary statistics for numeric variables, including count, sum, and various percentiles.
4. For Categorical (Character) Variables:
   1. Graphical Analysis:
      1. PROC FREQ: This procedure is used for graphical and non-graphical analysis of categorical variables. It generates frequency tables and bar charts for categorical data.
      2. PROC SGPLOT: While often used for numeric data, SGPLOT can also create bar charts and other graphical representations of categorical variables.
   2. Non-Graphical Analysis:
      1. COUNTC and FREQ functions: These functions can be used to count occurrences of values in categorical variables.
      2. PROC FREQ: Besides its graphical capabilities, PROC FREQ provides valuable non-graphical information, including counts and percentages of categories.
      3. PROC TABULATE: This procedure allows you to create tables of frequencies and percentages for categorical data.
      4. PROC ANOVA: PROC ANOVA is a SAS procedure used for performing analysis of variance (ANOVA). It is used to assess the differences between group means and determine whether there are statistically significant differences among the means of multiple groups.
      5. TTEST: function is a statistical test used to compare the means of two independent groups to determine if there is a statistically significant difference between them. It is often used to assess whether the means of two samples are significantly different from each other, with the t-statistic indicating the magnitude of the difference and the p-value determining whether the difference is statistically significant.
      6. TUKEY: Tukey is often used in the context of Tukey's Honestly Significant Difference (HSD) test, which is a post hoc test used in ANOVA to determine which specific groups differ significantly from each other after a significant main effect has been found. The Tukey test helps identify where these differences exist while controlling for familywise error rate.