MA409: Statistical Data Analysis (SAS)

Assignment 1

Note: For problems 4-7, please work with SAS and provide the SAS code in a separate ***.sas file*** and the outputs from SAS (using screenshots) together with problems 1-3 in a ***PDF file***. The SAS datasets are under *~/my_shared_file_links/u44964992/Assignments/* on SAS OnDemand for Academics.

1. Compute the population kurtosis for the following two distributions with probability density functions (keep two decimal points in your results):

$$f_1(x) = \begin{cases} 0.3334, & \text{if } |x| < 0.9399 \\ \dfrac{0.2945}{x^2}, & \text{if } 0.9399 \le |x| < 2.3242 \\ 0, & \text{if } |x| \ge 2.3242 \end{cases}, f_2(x) = \begin{cases} 0.4082 - 0.1667|x|, & \text{if } |x| < 2.4495 \\ 0, & \text{if } |x| \ge 2.4495 \end{cases}.$$

   State your findings about what kurtosis measures. (10 points)

2. Suppose that two teaching assistants (TA) rank the final projects of 10 groups from best to worst (1 to 10) for our course. The following table shows the rankings that each TA assigned to the groups:

Group	TA 1	TA 2
Group 1	1	2
Group 2	2	1
Group 3	3	4
Group 4	4	3
Group 5	5	6
Group 6	6	5
Group 7	7	8
Group 8	8	7
Group 9	9	10
Group 10	10	9

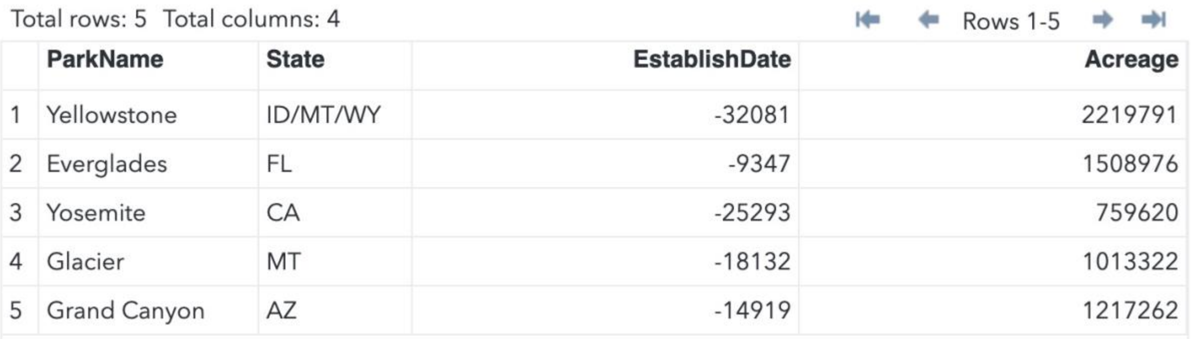
   Compute the Kendall's tau between the rankings of the two TAs and briefly state your conclusion. (10 points)

3. Read the first four pages of an article on missing data (click here). Let Y denote a single variable with missing data, $\boldsymbol{X}$ denote a set of variables that are always observed, and M denote an indicator variable defined based on whether Y is missing:

$$M = \begin{cases} 0, & \text{if } Y \text{ is missing} \\ 1, & \text{if } Y \text{ is observed} \end{cases}.$$

Write down the equations expressing the assumptions of missing completely at random (MCAR) and missing at random (MAR) with $\boldsymbol{X}, Y, M$ and briefly explain the equations. (10 points)

4. Sometimes raw data are not straightforward numeric or character. For example, we humans easily read the number 1,000,001 as one million and one, but SAS may see it as a character string. The data "*NationalPark.txt*" contain information about U.S. national parks: name, state (or states), date established, and size in acres. Provide the SAS code to read "*NationalPark.txt*" that produce a SAS dataset the same as follows:



Total rows: 5 Total columns: 4

	ParkName	State	EstablishDate	Acreage
1	Yellowstone	ID/MT/WY	-32081	2219791
2	Everglades	FL	-9347	1508976
3	Yosemite	CA	-25293	759620
4	Glacier	MT	-18132	1013322
5	Grand Canyon	AZ	-14919	1217262

Hint: please click here. (10 points)

5. The World Health Organization (WHO) collected data in countries across the world regarding the outbreak of swine flu cases and deaths in 2009. The data "*sff.sas7bdat*" includes information on cases and deaths per country by month during the epidemic.

   (1) Use *PROC CONTENTS* to get the names, types, lengths, and labels of the variables in "*sff.sas7bdat*". Then count the number of countries within each continent and put the screenshot of the results here. (5 points)

   (2) Count the number of countries per continent that reported no cases during the first month of the outbreak (April), the number of countries per continent that had at least one case during April, as well as the number of countries per continent that we cannot tell whether there were cases during April. Put the screenshot of the results here. (10 points)

   (3) To find potential errors in the data, output countries that reported a first death date, but no first case date. This output should include only the variables continent, country, first case date, last reported number of cases, and first death date. Make sure that countries on the same continent are listed together and dates are human readable. (5 points)

6. Researchers at a local medical center have just completed enrollment for a clinical trial of a new cholesterol-lowering medication for use in subjects with borderline high total cholesterol. They

keep their enrollment data in two SAS datasets so as to not bias the clinicians. The "*visits.sas7bdat*" dataset contains basic information about each subject at their baseline visit (*Visit* = *0*). The "*txgroup.sas7bdat*" dataset contains information about whether each subject received the treatment or a placebo.

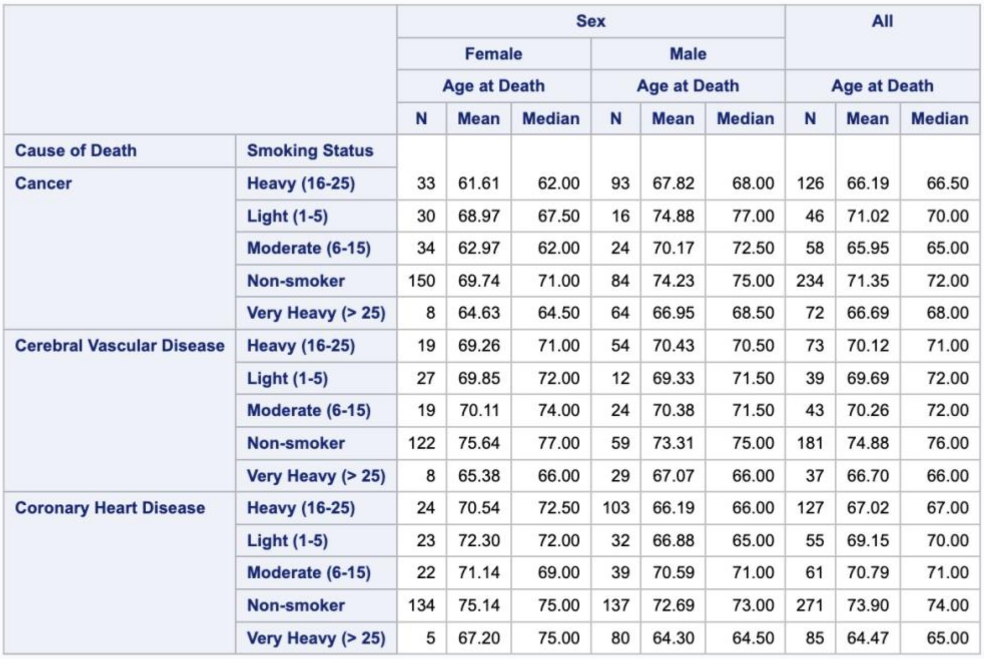
(1) Check if each of the two datasets has multiple records for the same subject. (5 points)

(2) Combine the *visits* and *txgroup* datasets into one that identifies the treatment group for each subject **using a DATA step**. Make sure that the resulting dataset has only one record per subject. (5 points)

(3) Using the combined dataset from (2), calculate the median baseline cholesterol measurement by treatment group, and use the information to create a variable (with name *Abovemedian*) that classifies subjects as less than or equal to the corresponding median (*Abovemedian=0*), or more than the median (*Abovemedian=1*). Do this without typing the calculated median values by hand into your code. Hint: compute the medians and store them in a new dataset, then combine it with the dataset from (2) and define the required variable. (5 points)

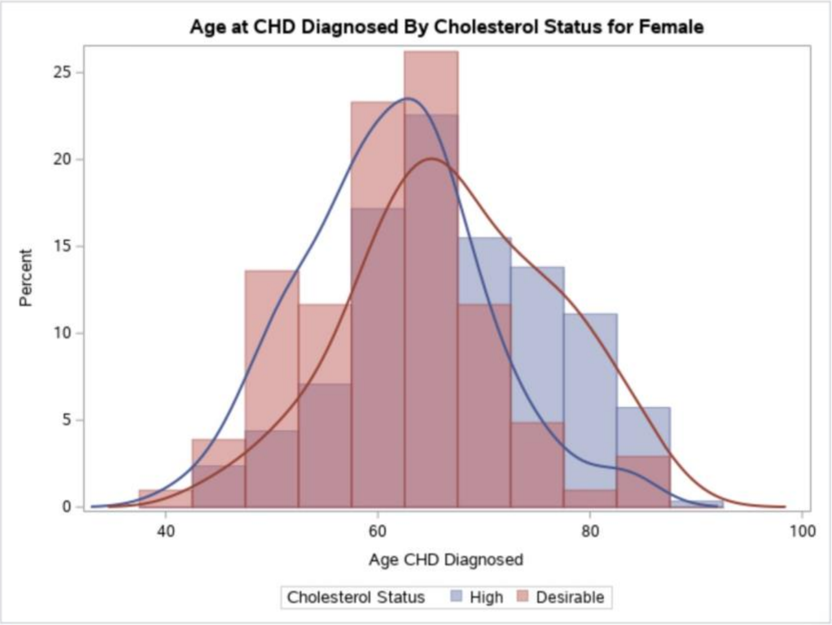
7. The *SASHELP.Heart* dataset we used in Example 2.14 of the lecture notes provides the results from the Framingham Heart Study which contains 5209 observations.

(1) Use *PROC TABULATE* to generate the same output table as below. (10 points)



		Sex						All		
		Female			Male					
		Age at Death			Age at Death			Age at Death		
		N	Mean	Median	N	Mean	Median	N	Mean	Median
Cause of Death	Smoking Status									
Cancer	Heavy (16-25)	33	61.61	62.00	93	67.82	68.00	126	66.19	66.50
	Light (1-5)	30	68.97	67.50	16	74.88	77.00	46	71.02	70.00
	Moderate (6-15)	34	62.97	62.00	24	70.17	72.50	58	65.95	65.00
	Non-smoker	150	69.74	71.00	84	74.23	75.00	234	71.35	72.00
	Very Heavy (> 25)	8	64.63	64.50	64	66.95	68.50	72	66.69	68.00
Cerebral Vascular Disease	Heavy (16-25)	19	69.26	71.00	54	70.43	70.50	73	70.12	71.00
	Light (1-5)	27	69.85	72.00	12	69.33	71.50	39	69.69	72.00
	Moderate (6-15)	19	70.11	74.00	24	70.38	71.50	43	70.26	72.00
	Non-smoker	122	75.64	77.00	59	73.31	75.00	181	74.88	76.00
	Very Heavy (> 25)	8	65.38	66.00	29	67.07	66.00	37	66.70	66.00
Coronary Heart Disease	Heavy (16-25)	24	70.54	72.50	103	66.19	66.00	127	67.02	67.00
	Light (1-5)	23	72.30	72.00	32	66.88	65.00	55	69.15	70.00
	Moderate (6-15)	22	71.14	69.00	39	70.59	71.00	61	70.79	71.00
	Non-smoker	134	75.14	75.00	137	72.69	73.00	271	73.90	74.00
	Very Heavy (> 25)	5	67.20	75.00	80	64.30	64.50	85	64.47	65.00

(2) Plot the histograms with kernel density curves of age at CHD (coronary heart disease) **for female by cholesterol status who are of desirable or high status**. The plot should look like the following plot (don't have to be exactly the same but should clearly show the two histograms). (5 points)



(3) Create a macro that implement this functionality: specify a categorical variable, if today is Monday or Wednesday, draw a histogram of age at death by the categorical variable; else if today is Tuesday or Thursday, draw a barplot of age at death by the categorical variable. Apply the macro with *Sex* being the specified categorical variable. Hint: you may want to use the *&SYSDAY* automatic macro variable. (10 points)