MA409: Statistical Data Analysis (SAS)

Assignment 1 (Feb 27 – Mar 19)

1.1 The sample standard deviation is defined as $S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}}$. Why are we dividing $\sum_{i=1}^{n}(X_i - \bar{X})^2$ by $n-1$ instead of n? Please provide the corresponding mathematical justification. (10 points)

1.2 Show that the population kurtosis of a normal distribution is 3. (10 points)

1.3 Let X and Y be two continuous variables. If the Pearson's correlation coefficient of X and Y is 0, then X and Y are "uncorrelated"; if the joint probability density of X and Y equals the product of the densities of X and Y, i.e., $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, then X and Y are "independent". Show that if X and Y are independent, they must be uncorrelated. Then use a counter-example to show that if X and Y are uncorrelated, they are not necessarily independent. (10 points)

1.4 Consider a hypothetical clinical trial involving liver cirrhosis patients, the prothrombin index (a measure of liver function, higher value suggests better liver function) is recorded at study entry and 10 follow-up visits (one every 3 months). Note all the patients have 11 prothrombin index records, as some patients have missing records due to death of liver cirrhosis. Which type of missing is in the data? State your thoughts and rationale. (10 points)

1.5 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.

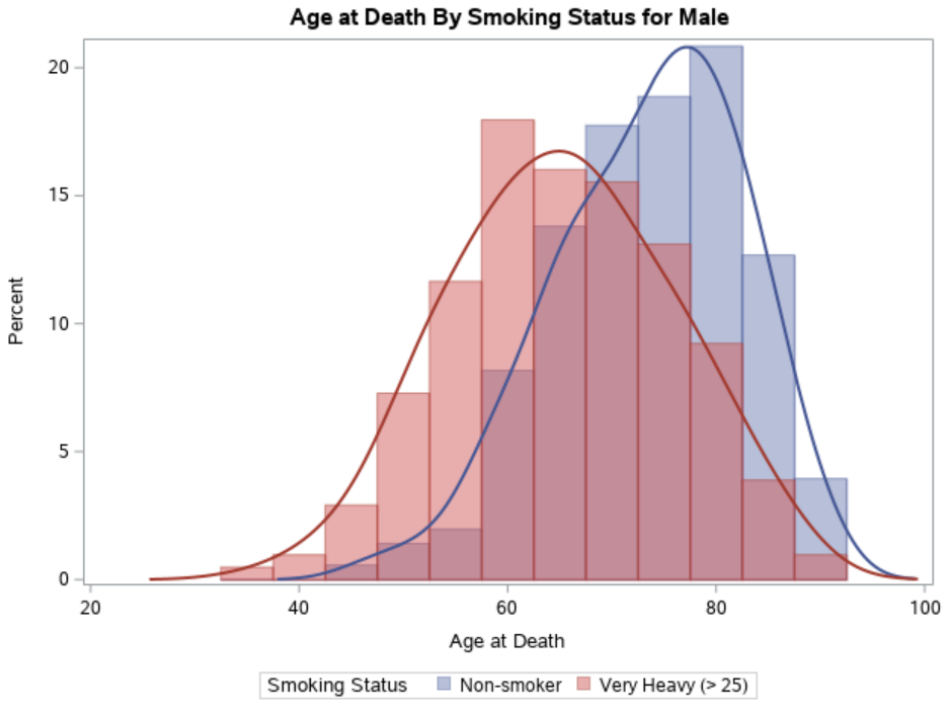
(1) Use *PROC MEANS* to compute the age at death by sex and smoking status, report the number of missing values, mean, median, standard deviation, skewness, kurtosis, limit the decimal places to 4. (10 points)

(2) 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
Smoking Status	Weight Status									
Heavy (16-25)	Normal	31	66.29	69.00	104	69.06	70.00	135	68.42	70.00
	Overweight	72	67.29	67.00	221	68.08	67.00	293	67.88	67.00
	Underweight	4	68.25	70.50	11	66.82	67.00	15	67.20	67.00
Light (1-5)	Normal	26	68.85	67.50	18	71.89	70.00	44	70.09	68.00
	Overweight	78	71.15	72.00	50	70.60	71.00	128	70.94	72.00
	Underweight	7	67.43	70.00	8	69.00	69.00	15	68.27	70.00
Moderate (6-15)	Normal	44	66.82	66.00	25	68.48	70.00	69	67.42	67.00
	Overweight	51	67.61	68.00	80	70.81	71.00	131	69.56	70.00
	Underweight	9	65.11	65.00	4	64.75	63.50	13	65.00	65.00
Non-smoker	Normal	91	71.48	73.00	58	73.10	74.50	149	72.11	74.00
	Overweight	434	74.53	76.00	291	73.48	75.00	725	74.11	75.00
	Underweight	10	74.00	74.50	5	79.40	77.00	15	75.80	76.00
Very Heavy (> 25)	Normal	8	62.50	65.50	51	65.67	66.00	59	65.24	66.00
	Overweight	20	68.65	71.00	149	64.87	65.00	169	65.32	65.00
	Underweight	3	69.67	68.00	5	65.00	58.00	8	66.75	67.00

(3) Use PROC SQL to compute the number of observations with non-missing age at death and mean age at death by sex, smoking status and weight status, only output the groups with no missing sex, smoking status and weight status and number of observations with non-missing age at death greater than 20, order the output by mean age at death in descending order. (10 points)

(4) Plot the histograms with kernel density curves of age at death for males by smoking status who are non-smokers or very heavy smokers. The plot should look like the following plot (don't have to be exactly the same, but should clearly show the two histograms). (15 points)



(5) Create a macro that implement this functionality: by specifying a dataset name and a variable name in the dataset, draw a histogram of the variable if the variable is a numerical variable, and draw a bar chart of the variable if the variable is a categorical variable. Then use the macro to draw a histogram of age at death as well as a bar chart of cause of death for the *SASHELP.Heart* dataset. You may want to use the *VARTYPE* function, click here for the documentation. (15 points)

For the (1)-(5) above, please provide the SAS code you use and the outputs from SAS.