Mert Dogan
Z1910844

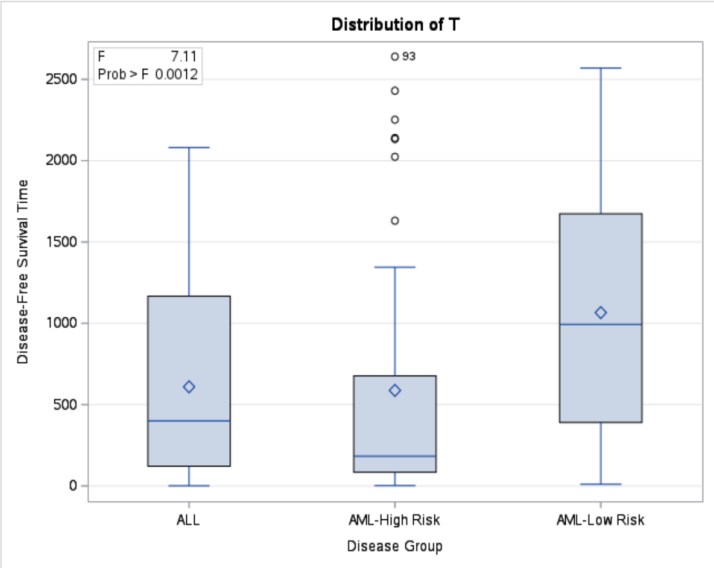
Assignment 3
10/02/2023

1. Open the BMT file available in the SASHELP directory. The file contains data from 137 bone marrow transplant patients. The Group variable contains data on patient classification into one of three risk categories: ALL (acute lymphoblastic leukemia), AML-Low Risk (acute myelocytic leukemia, low risk), and AML-High Risk. The dependent variable of interest is the disease-free survival time in days, T. Conduct the test to infer if the mean days of survival is different for the three disease groups and state your results.

Dependent Variable: T Disease-Free Survival Time

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7186442.04	3593221.02	7.11	0.0012
Error	134	67675769.84	505043.06		
Corrected Total	136	74862211.88			

In our analysis, 137 bone marrow transplant patients are categorized and examined based on three risk groups: ALL (acute lymphoblastic leukemia), AML-Low Risk (acute myelocytic leukemia, low risk), and AML-High Risk. We would like to examine if the mean disease-free survival time in days, T, is different for the three groups. The ANOVA test yielded a p-value of (0.0012) from the F statistic. P-value (0.0012) being less than the nominal alpha (0.05), we reject null and conclude that the average disease-free survival time in days differs from the others for at least one group.



By checking Tukey's multiple comparison test, we would like to infer which group means differ significantly. According to the p-values of conducting a test for the difference between the group means, we can conclude that Group 1 (ALL) is different from Group 3 (AML-Low Risk) with a p-value of 0.0081 and Group 2 (AML-High Risk) is also different from Group 3 with a p-value of 0.0031. Group 1 (ALL) and Group 2 (AML-High Risk) are not significantly different from each other with a p-value of 0.9890.

Overall, we have conducted an ANOVA test that explains the relationship between risk categories and disease-free survival time. We can conclude that the survival time of the patients differs based on their corresponding risk groups. We can also conclude that the survival time of AML-Low Risk is significantly different from ALL and AML-High risk, but the survival time of ALL and AML-High Risk groups are not significantly different from each other.

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Group	T LSMEAN	LSMEAN Number
ALL	609.42105	1
AML-High Risk	587.28889	2
AML-Low Risk	1065.77778	3

Least Squares Means for effect Group
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: T

i/j	1	2	3
1		0.9890	0.0081
2	0.9890		0.0031
3	0.0081	0.0031	

2. The heart dataset in the SASHELP directory contains data from the Framingham heart study. We would like to explore the relationship between Smoking_status (5 categories) and the AgeatDeath variables. (Note that although the study contains data from 5209 individuals, only a subset of them are dead). Conduct the appropriate analysis and state your conclusions.

Class Level Information

Class	Levels	Values
Smoking_Status	5	Heavy (16-25) Light (1-5) Moderate (6-15) Non-smoker Very Heavy (> 25)

Number of Observations Read	5209
Number of Observations Used	1971

Dependent Variable: AgeAtDeath Age at Death

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	19086.1404	4771.5351	47.21	<.0001
Error	1966	198696.2137	101.0662		
Corrected Total	1970	217782.3541			

In our analysis, out of the 5209 individuals, only 1971 of them will be included due to having information available on the AgeatDeath variable, as they are the individuals who have passed away. We will examine the relationship between Smoking_status (5 categories listed as): Non-smoker, Light (1-5), Moderate (6-15), Heavy (16-25), Very Heavy (>25), and the age at death for these individuals. We would like to examine if the mean age at death is different for these five groups. The ANOVA test yielded a p-value of (<0.0001) from the F statistic. P-value (<0.0001) being less than the nominal alpha (0.05), we reject null and conclude that the average age at death differs from the others for at least one group.

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Smoking_Status	AgeAtDeath LSMEAN	LSMEAN Number
Heavy (16-25)	68.0248307	1
Light (1-5)	70.5240642	2
Moderate (6-15)	68.5915493	3
Non-smoker	73.7609428	4
Very Heavy (> 25)	65.4092827	5

Least Squares Means for effect Smoking_Status
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: AgeAtDeath

i/j	1	2	3	4	5
1		0.0357	0.9616	<.0001	0.0109
2	0.0357		0.3078	0.0006	<.0001
3	0.9616	0.3078		<.0001	0.0073
4	<.0001	0.0006	<.0001		<.0001
5	0.0109	<.0001	0.0073	<.0001	

By checking Tukey's multiple comparison test, we would like to infer which group means differ significantly. According to the p-values of conducting a test for the difference between the group means, we can conclude that Group 1 (Heavy smokers) is not significantly different from Group 3 (Moderate smokers) with a p-value of 0.9616 and Group 2 (Light smokers) is also not significantly different from Group 3 (Moderate smokers) with a p-value of 0.3078. All other group mean comparisons yield a significant difference with p values less than .05.

Overall, we have conducted an ANOVA test that explains the relationship between smoking status and age at death. We can conclude that the age at death of individuals differs based on their corresponding smoking status. We can also conclude that other than 'heavy and moderate smokers' and 'light and moderate smokers', average age at death of individuals for all the other smoking status comparisons are significantly different from each other.

3. Use the HBAT_200 dataset for this question. We would like to examine the relationship between (X1) Customer Type (1= customer for less than a year, 2 = between 1 to 5 years, 3= customer for more than 5 years) and whether they would be likely to recommend (X20). Conduct the analysis and state your conclusions.

Class Level Information		
Class	Levels	Values
X1___Customer_Type	3	1 2 3

Number of Observations Read	200
Number of Observations Used	200

Dependent Variable: X20___Likely_to_Recommend X20 - Likely to Recommend

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	71.0427574	35.5213787	43.11	<.0001
Error	197	162.3159926	0.8239390		
Corrected Total	199	233.3587500			

In our analysis, we use the HBAT_200 dataset to analyze the relationship between the customer type (based on the number of years they have been a customer) and their likelihood to recommend (X20). Customer type is categorized into 3 groups: Type (1= customer for less than a year, 2 = between 1 to 5 years, 3= customer for more than 5 years) and we would like to examine whether the mean likelihood of recommending differs among the three groups. The ANOVA test yielded a p-value of (<0.0001) from the F statistic. P-value (0.0001) being less than the nominal alpha (0.05), we reject null and conclude that the average likelihood of recommending differs from the others for at least one group.

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

X1___Customer_Type	X20___Likely_to_Recommend LSMEAN	LSMEAN Number
1	6.14117647	1
2	7.20937500	2
3	7.52205882	3

Least Squares Means for effect X1___Customer_Type
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: X20___Likely_to_Recommend

i/j	1	2	3
1		<.0001	<.0001
2	<.0001		0.1204
3	<.0001	0.1204	

By checking Tukey's multiple comparison test, we would like to infer which group means differ significantly. According to the p-values of conducting a test for the difference between the group means, we can conclude that Customer type 1 (less than a year customer) is significantly different from Customer type 2 (customer for 1-5 years) and Customer type 3 (customer for more than 5 years with a p-value of <0.0001 for both comparisons. We can also conclude that the average likelihood of recommending between Customer types 2 and 3 are not significantly different with a p-value of 0.1204 greater than .05.

Overall, we have conducted an ANOVA test that explains the relationship between customer type and their likelihood of recommendation. We can conclude likelihood of recommending differs based on the customer type. We can also conclude that the average likelihood of recommendation for customer type 1 is significantly different from customer types 2 and 3, but the average likelihood of recommendation is not significantly different between customer types 2 and 3.