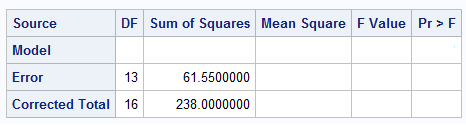
MSDS 6371: Midterm Exam 1 (FALL 2015)

1. (2pts each) Circle all that are true:
   1. For a left tailed test, if the test statistic is less than the critical value, the correct conclusion is to fail to reject the Ho**.**
   2. **If you increase alpha, the power will increase.**
   3. **All other things equal, if you increase the confidence, the interval will get wider.**
   4. The 2 Sample T-Test is resistant to outliers.
   5. **A Rank-Sum test is a good test to use with censored data**.
   6. For a fixed alpha, will never be bigger than , no matter what the sample size.
   7. If the effect size increases (the distance between the null mean and assumed actual mean) the power will decrease.
   8. **For a two sample confidence interval for the difference in means (µ1-µ2), if both limits are positive, that is evidence that the two population means are not equal and that µ1 is bigger.**
   9. The F-Test for Equality of variance is assumption free.

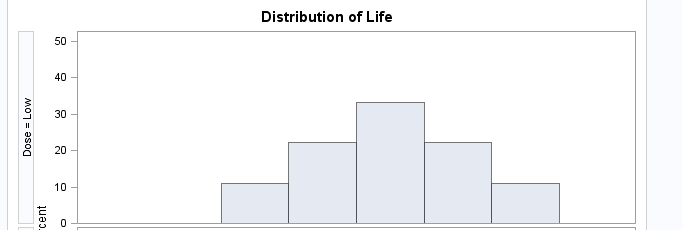


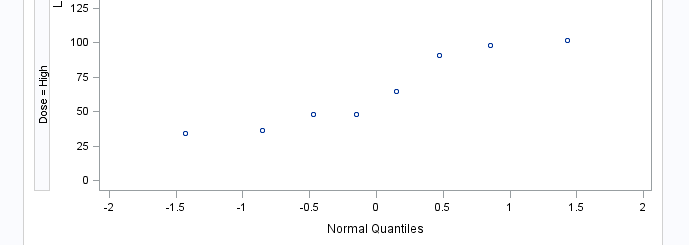
1. ( 2 pts each) Given the ANOVA table above, find:
2. The Mean Square Error: **58.817**
3. The F Value **12.423**
4. R2 **0.741**
5. How many degrees of freedom were used in estimating the MSE? **3 degrees of freedom**
6. **(2 pts each) Circle all** that are true about the Central Limit Theorem (CLT) assuming n > 30.
   1. The sampling distribution of the sample standard deviation will be normally distributed.
   2. **The mean of the sampling distribution of the mean will be the same as the mean of the original distribution the sample was pulled from.**
   3. **The sampling distribution of the sample mean will be normally distributed**.
   4. The sampling distribution of the sample mean will be the same as the original distribution.
   5. **The standard deviation of the sampling distribution of the sample mean (the standard error) will be the same as the standard deviation of the original distribution.**
7. S.W. Laagakos and F. Mosteller of Harvard University fed 38 randomly chosen female mice different doses of red dye number 40 and recorded the time of death in weeks. To decide which mouse receives which treatment, a random number from 1 to 4 is drawn for each mouse. The mouse receives the treatment based on this number:
8. Control Group: No Dosage
9. Low Dosage
10. Medium Dosage
11. High Dosage

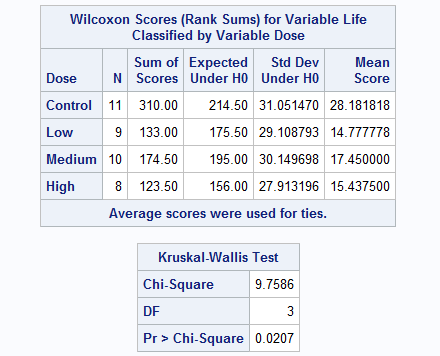
Dosage and time of death are shown in the data file red40.csv.

* 1. (9pts) Test the claim that there are any differences in location (mean or median) between the 4 groups. Perform a complete analysis.

**We are testing to find out if there are any differences in the lifespan of mice given different doses of the dye red 40. While there is no evidence against normality for some of the test groups (such as the low dose, see histogram below) others do not seem normally distributed (most especially the high dosage group, see QQ plot below). For that reason we will use the Kruskal Wallis test to test for a difference in medians. This is ideal as the Kruskal Wallis has no requirement for normality and our sample sizes are all greater than five.**



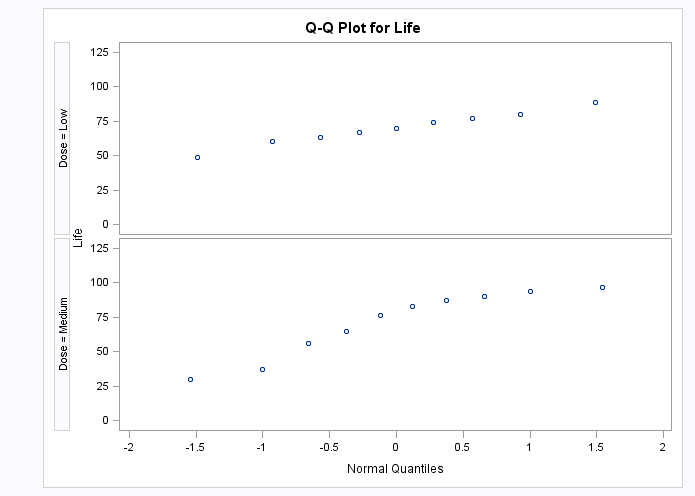


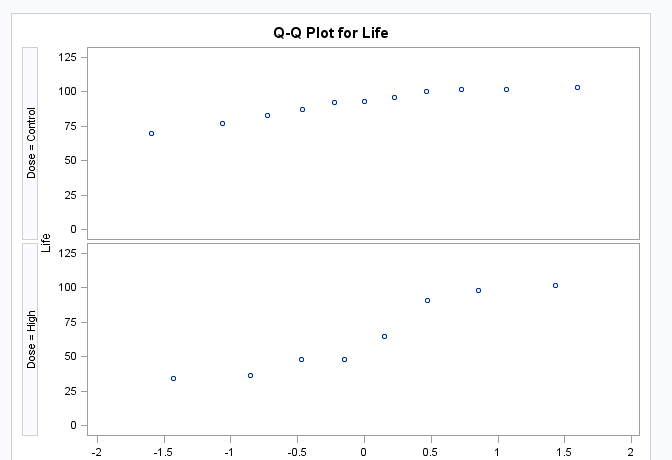
**Our null hypothesis is that all groups have an equal median while the alternative is that at least one median is different.** 

**Our Chi-Square value of 9.7486 gives us a p-value of 0.0207. Given our alpha of 0.05 we reject the null hypothesis that the median life spans of the mice is equal regardless of dose given. Instead there is strong evidence that there is a difference is the median life span though we cannot say at this time which one(s) differ.**

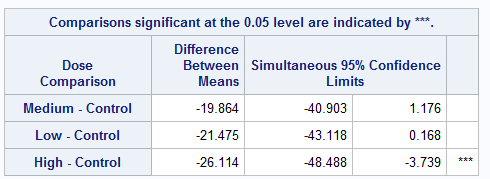
* 1. (9 pts) If the results of the above analysis provide evidence of a difference in mean or median time to death, which of the dosage treatments have significantly different mean or median times to death than the control (no dosage)? For this problem, make sure and state the problem, address any assumptions, provide the appropriate visual aids (tables, plots, etc), provide a clear and concise conclusion as well as the scope. Individual analysis may vary but should at least contain the above.

**Since there is no option to use Dunnett’s procedure with the Kruskal Wallis test in SAS we will be using it to test the difference in means between the dosage groups and the control groups. While there was some evidence against normality in the high dosage group it was not very strong and the other groups have little evidence against normality (see QQ plots below)**

****



**After running Dunnett’s procedure we can conclude only one of the dosages has a significant difference from the control at alpha = 0.05. The differing one is the high dosage. On average the mice receiving a high dose of red40 died younger than the control group. The 95% confidence interval for the control’s life span difference is 3.7 to 48.5 weeks.**

****

**Since the mice were randomly selected and the treatments randomly assigned we can state that red40 dye in high doses caused the shorter life span. We can apply these results to all female mice of the type chosen for the trial.**

* 1. (9pts) Would you feel comfortable using a contrast to test for a significant difference between the mean times to failure of the control and low dosage groups? If so perform this contrast. If not explain why?

**Since the low and control groups are both normally distributed and have similar standard deviations I will perform a contrast. Weights of 1 0 -1 0 were used as per the order SAS puts the data in. Our null hypothesis is that there is no difference in the means between control and low and the alternative is that there is a difference. After running the test the following table was generated.**

****

**Given the p-value of 0.0196 we reject the null hypothesis at alpha = .05. This test would indicate that there is a difference in the mean life spans between these two groups. This somewhat contradicts the test run above which only found significant difference between the high and control groups. However the confidence interval generated by that test (-43.118 to 0.168 weeks) would seem to agree that there is a difference since zero is on the very edge of that interval.**

1. (2 pts) Only two more to go! You’re half way done! For this question simply identify the first Department Chair of Statistics at SMU. Google search SMU STATISTICS.

**In 1961 Paul Minton was the first chair of the new Statistics Department.**

1. (8 pts) Assume Merck Pharmaceuticals wanted to test a new cancer drug. In order to do so they tested it first on rats. 40 rats with the same cancer were selected (not at random from any bigger population) for the study and 20 were randomly selected to be administered the drug and 20 were not given any drug at all. The response variable is a cancer score in which the higher the number the more cancers cells are present in the rat. Histograms and QQ plots revealed evidence against the data coming from a normal distribution as well as visual evidence against equal standard deviations. Therefore, since only 20 samples were taken, the researchers were skeptical that they could use a pooled t test to test for difference in mean cancer score between the two groups. They also noticed that the sample with the bigger sample mean had the bigger sample standard deviation. For this reason a log transformation was performed. The researchers were satisfied that the log transformed data met the assumptions to perform a pooled t test. They performed the test on the logged data and the results are below. Assume all the assumptions of the pooled ttest are met with the logged data. Use the test results to interpret the test results for the researchers. Your answer should consist of a step 6 conclusion and scope discussion; all inference should be on the original scale (cancer\_score units, not logged units.). Just provide Step 6 (Conclusion) and the Scope. You **do not** need to do State the problem, address any assumptions or do Steps 1 – 5.

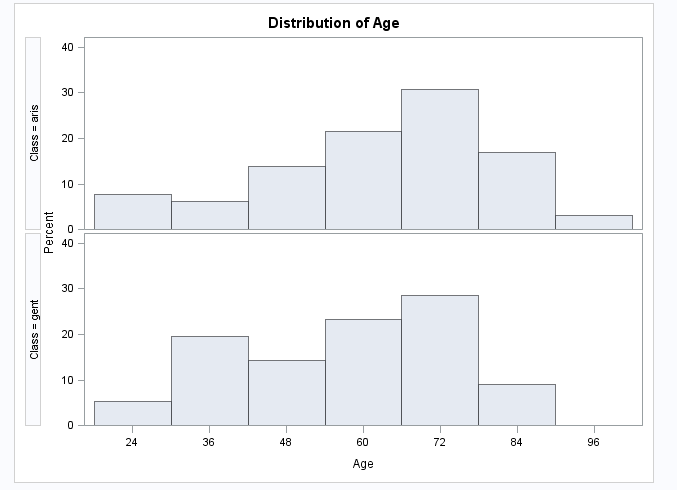
|  |
| --- |
|  |

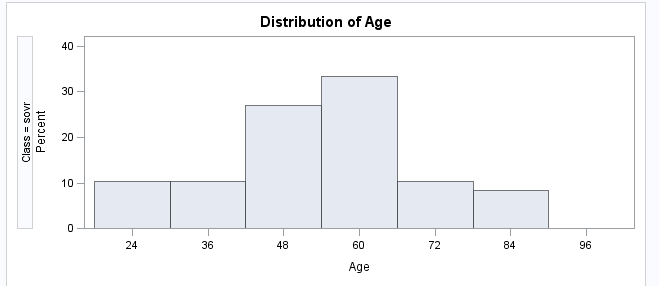
**The control group had an average cancer cell count of 25,304 while the rats who received the drug had an average cancer cell count of 4,627. This is an average difference of 20,677.6 cells. Given the very low p-value (<.0001) we can reject the null hypothesis that there is no difference in cell counts between the two groups. Instead we conclude that the drug had both a statistical and practical effect on the number of cancer cells in the rats. More specifically we can say that the drug was the cause and had the effect of reducing the number of cancer cells present in the rats.**

**Since these rats were not randomly selected from a larger population these results can only be applied to this particular group of rats.**

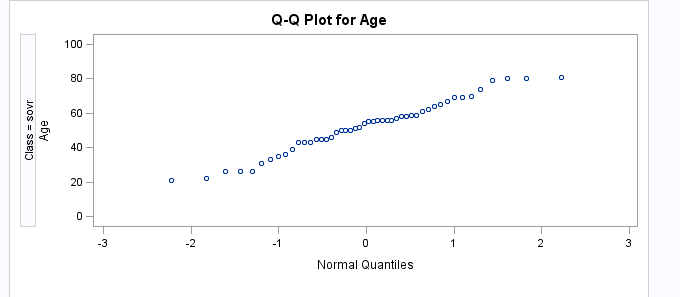
1. A researcher wants to test the hypothesis that life expectancy between the classes (sovereignty, gentry, aristocracy) in mid 1800’s Europe was significantly different. To do this the researcher obtained a sample of the ages at death of 48 sovereignty, 56 gentry and 65 aristocracy members. These were obtained by looking at the first few pages of extensive records kept for families of each class. The records were in alphabetical order by first name. These data are in EuroDeathAge.csv.
   1. (9 pts) Test the claim that the mean or median life expectancy for the sovereignty is less than that of the gentry. Conduct a complete analysis.

**All three of the sample groups have no evidence against normality and appear to have similar distributions. There seems to be a slight left skew in the gentry and aristocracy groups but it isn’t very extreme (see histogram and QQ plots below). As such we will use a standard t-test.**

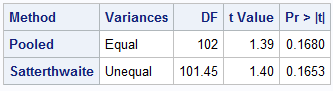
****

****

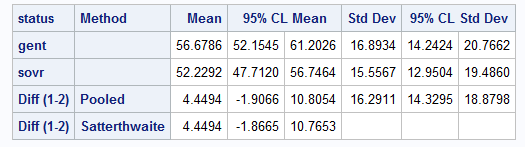




**Our null hypothesis is that there is no difference in means and the alternative is that the mean life span of the gentry is higher than that of the sovereignty. Our critical t is 1.6599 since the degrees of freedom are 102 and our alpha is 0.05. The t statistic is 1.39 and the p-value is 0.084 (the p-value from the table is divided by 2 since it is a one tailed test).**

****

**Since the p-value is greater than 0.05 we fail to reject the null hypothesis that the gentry have a longer mean lifespan than the sovereignty.**

****

**Our 95% confidence interval for the difference between gentry and sovereignty mean life spans is -1.9066 to 10.8054 which does contain zero and is consistent with failing to reject the null hypothesis.**

* 1. (9 pts) Next the researcher wants to test if the aristocracy has a greater mean life expectancy than the average of the sovereignty and gentry classes. Would a contrast be appropriate here? Why would you want to use a contrast if possible?

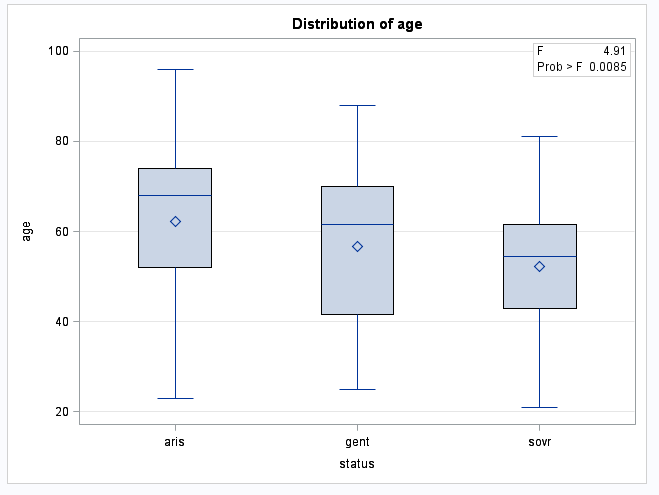
**Given the similar standard deviations and lack of evidence against normality it would be appropriate to use a contrast. A contrast is good to use since it is more precise than a standard ANOVA since it gives actual comparisons between the groups and not just telling if one is different from any of the others. Using the contrast coefficients we can set how these comparisons are made by giving the means tested either positive or negative coefficients and by their magnitude.**

* 1. (9 pts) If appropriate, use a contrast to perform the analysis. Conduct a complete analysis.

**As stated above we see similar standard deviations and have no evidence against normality. Our null hypothesis is that the aristocracy has the same mean life expectancy as the gentry and sovereignty. The alternative is that the aristocracy has a greater mean life expectancy than the other two. We are using an alpha of 0.05. Coefficients of 2 -1 -1 were used and the following table was generated.**

****

**Given the very low p-value (0.0042) we reject the null hypothesis and conclude that the aristocracy did have a longer mean life expectancy than the gentry and sovereignty. The below box plot is consistent with this conclusion as the median, mean, and max life expectancy of the aristocracy are higher than the other two.**

****

1. BONUS TO COME!!!!