MSDS 6371 Analysis Question



This year, the Winter Olympics are in PyeongChang, South Korea, and the world is watching! Surprisingly, the sport of curling has been around since the first Winter Olympics in 1924, but it wasn’t recognized as a medal sport until 2006. This year, controversy has struck curling as the bronze medal winner [Aleksandr Krushelnitckii](https://www.olympic.org/pyeongchang-2018/results/en/curling/athlete-profile-n3043371-aleksandr-krushelnitckii.htm) has tested positive for a banned substance known as meldonium. This is the same substance for which former world tennis number one Maria Sharapova tested positive in 2016.

There has been great debate about any advantage athletes may gain when taking meldonium. World doping expert Don Catlin concluded, “There’s really no evidence that there’s any performance enhancement from meldonium – zero percent.”

Use the Curling.csv data set to answer the questions below.

**FACTS** to use in the analysis unless otherwise specified (READ THESE!):

1. Assume that his curling scores for each tournament come from a **normal** distribution.
2. Assume that the **standard deviations** of his scores for tournaments are consistent (not unequal).
3. Although this may be a questionable assumption, assume that the scores within and between tournaments are **independent** of one another.
4. Assume that it was known that he did NOT test positive for meldonium in ANY of the non-Olympic tournaments.
5. Besides the Olympics, the six other tournaments under study were randomly selected from all of his tournaments.
6. In curling, a higher score is better.
7. (20 points) Conduct an analysis to test for any evidence of different mean or median curling scores between tournaments. In other words, is there evidence that at least one mean or median is different from the others or that at least one pair of means / medians is different from each other?
   * Perform a **COMPLETE ANALYSIS**, including, but not limited to, addressing the assumptions of the test (using the FACTS (1)-(6) found prior to part (a), no need to analyze graphs) and a 6-step hypothesis test (including a scope of inference).
   * Provide your SAS or R code and screen shots of relevant output (not a dump of the output!).
   * Use an alpha = .05 level of significance.
   * Once you are finished with the analysis above, address the assumptions of the data (disregarding facts (1), (2), and (3) found prior to part (a)) like you normally would, using relevant graphs, etc. Write a sentence or two describing the analysis you would perform given your analysis of the assumptions via graphs, etc. It may be exactly what you already performed for this question or a different analysis. In the interest of time, do **NOT** perform this analysis.

Problem: Test if there is any evidence that there is a difference in the mean scores between tournaments.

Assumptions:

1. Normality - Given above, we will assume that the data is normal.
2. Equal Standard Deviations – Given above, we will assume equal variances
3. Independence – Given above, we will also assume independence.

Given the following assumptions were met above, we can run a one way anova test since we have more than 2 samples ( we have 7 samples of tournaments to compare)

Code Used:

PROC IMPORT OUT= WORK.curling

DATAFILE= "/home/marinfamily1010/sasuser.v94/Data/Curling.csv"

DBMS=CSV REPLACE;

GETNAMES=YES;

DATAROW=2;

RUN;

proc sql;

create table work.curling2 as

select \* from WORK.curling

order by Tournament asc;

run;

proc univariate data = WORK.curling2;

by Tournament;

histogram score;

qqplot score;

run;

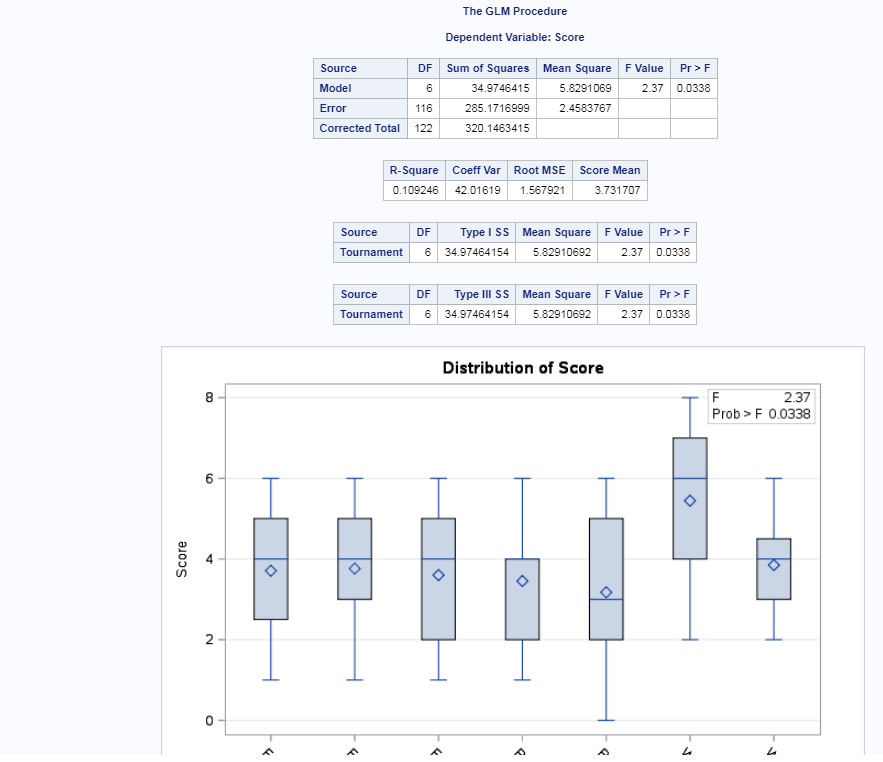
proc glm data = WORK.curling2 ;

class Tournament;

model score = Tournament;

run;

Output:



Step 1:

H0: u1 = u2 = u3 = u3 = u4 = u5 = u6 = u7

Ha: ui <> uj for some i,j

Step 2: Critical Value

We can skip this for anova

Step 3 Test Statistic: F = 2.37

Step 4: P-value: .0338

Step 5: We reject the null with an alpha of .05

Step 6:

There is strong evidence to suggest at the alpha = .05 level of significance (p-value <.0338) to reject the null hypothesis that the means between the tournaments are the same. Note that I didn’t provide a confidence interval for one way anova. Seems like all examples of this within powerpoints and hw solutions, none was provided. Confused about this.

Scope of Inference: This is an observational study as there was no randomized effect nor experimentation. The games were observed and recorded for one person and no causal inference can be made to a bigger population other than the person observed.

Now let’s address the assumptions without the assumptions made previously.

Using the following code to look at charts:

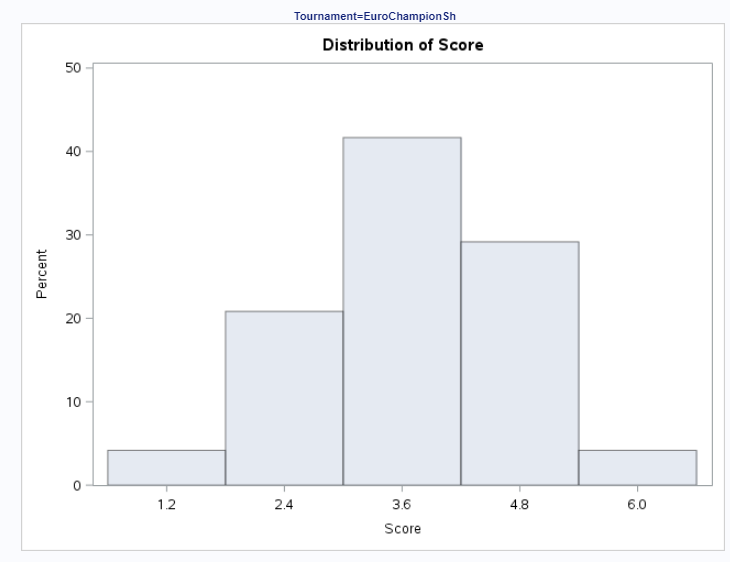
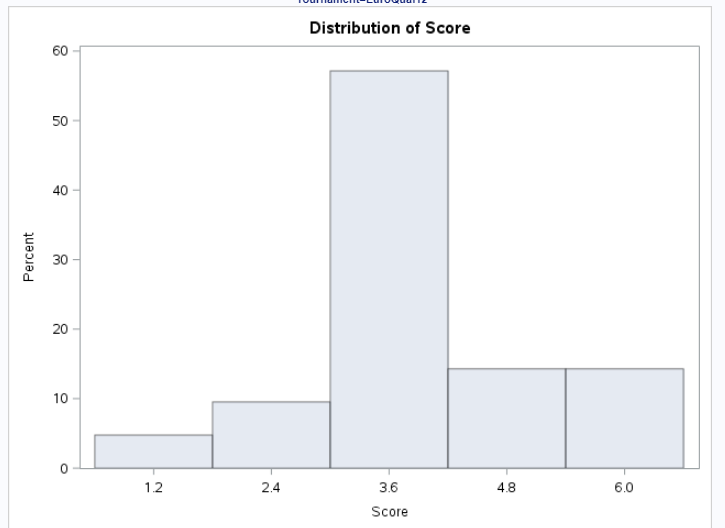
proc univariate data = WORK.curling2;

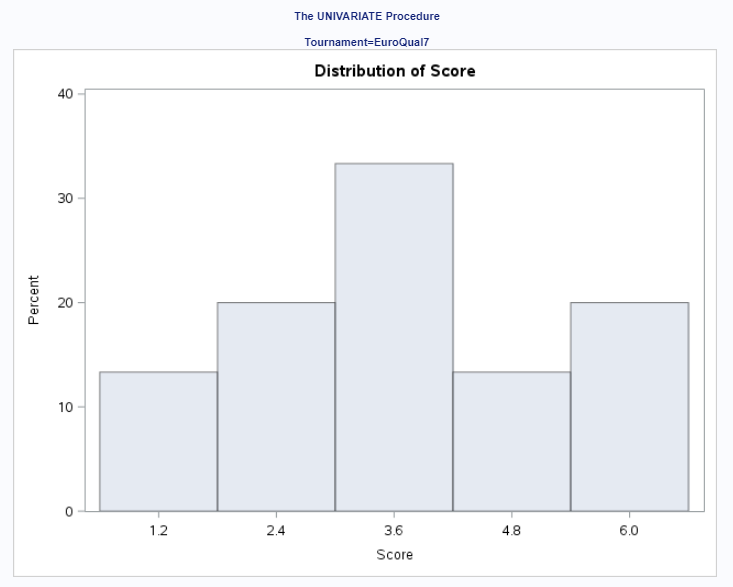
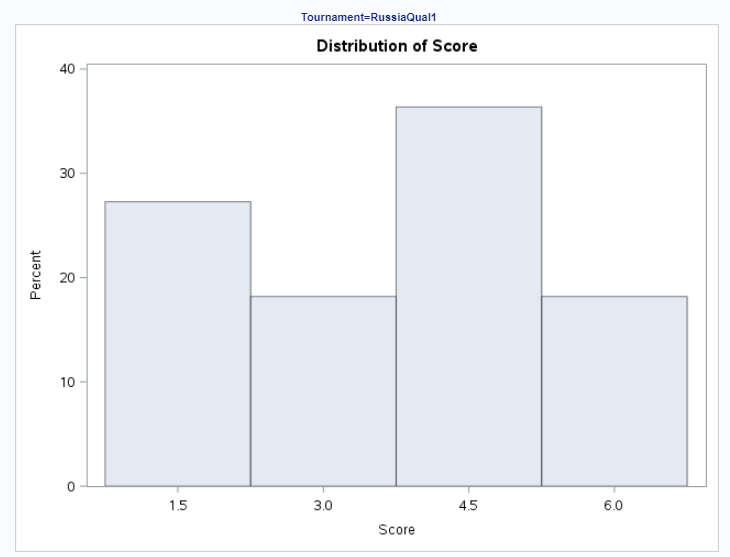
by Tournament;

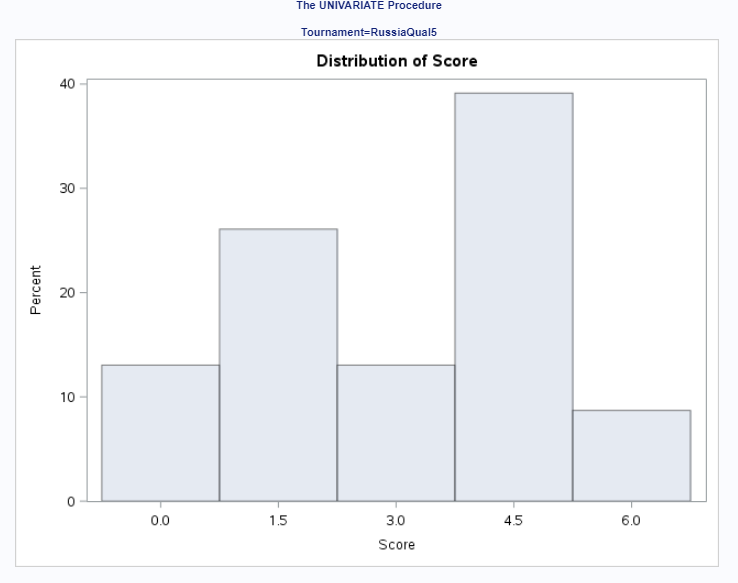
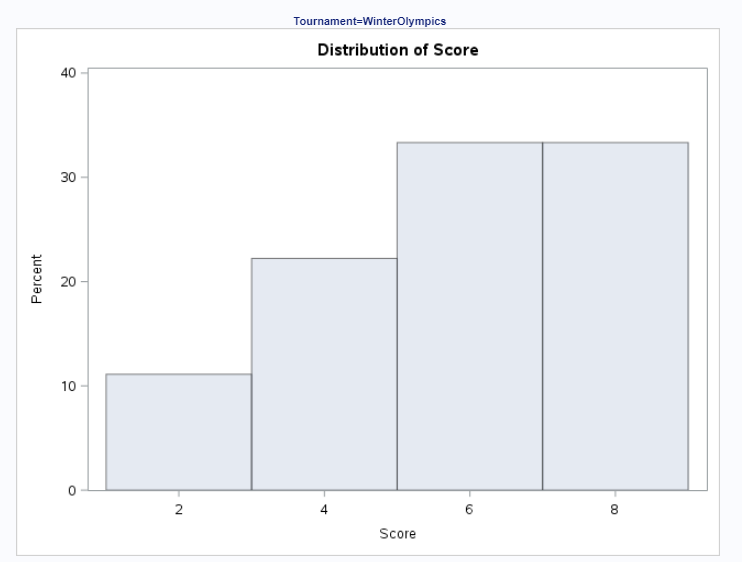
histogram score;

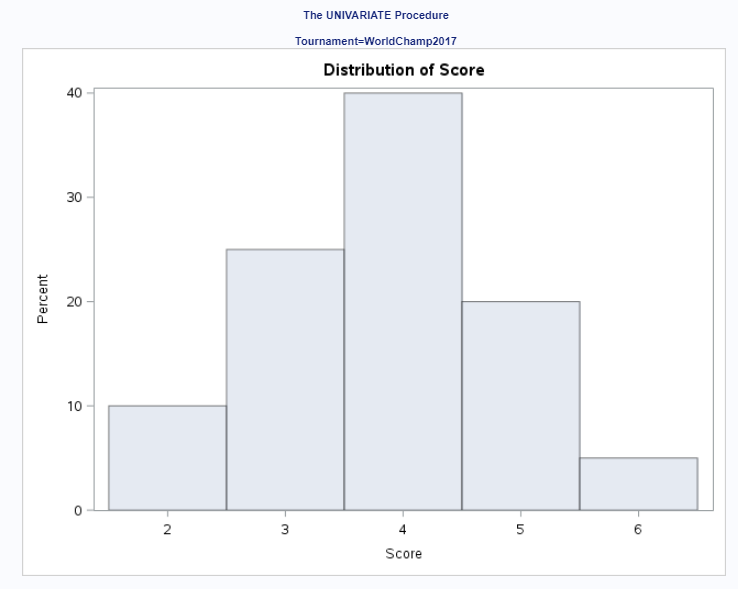
qqplot score;

run;

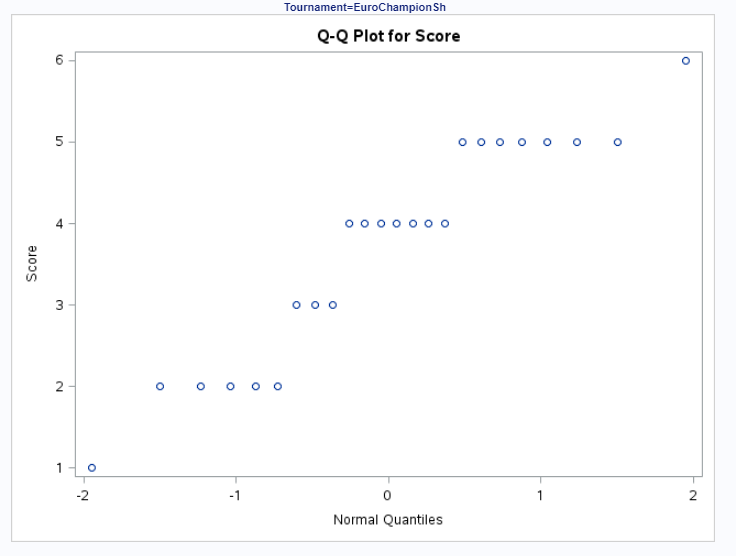
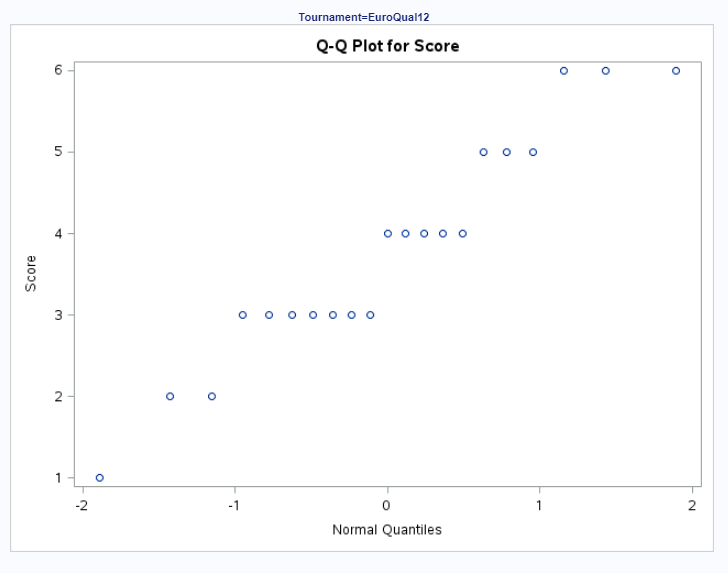
 

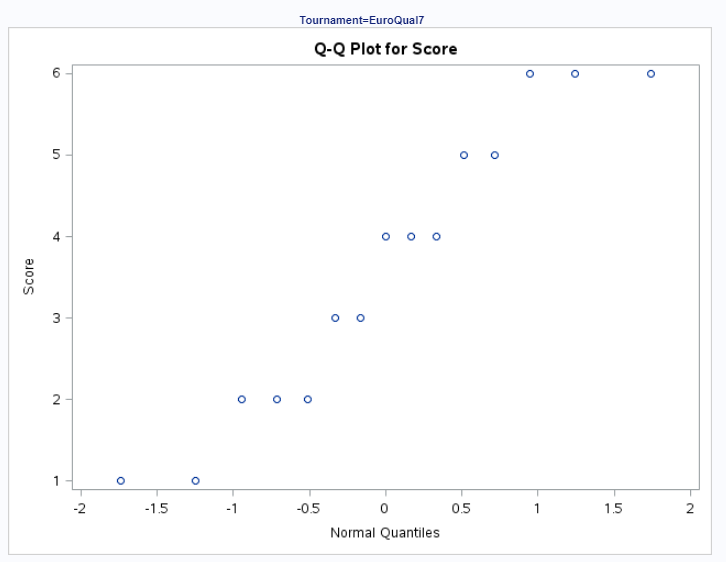
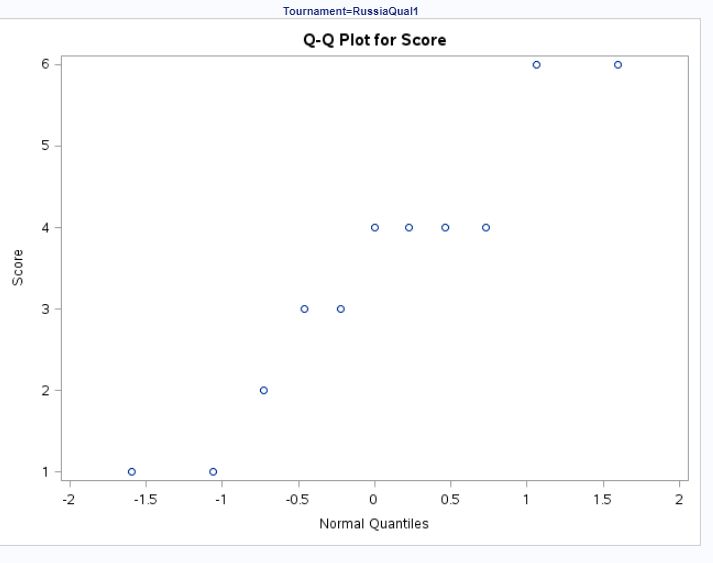
 

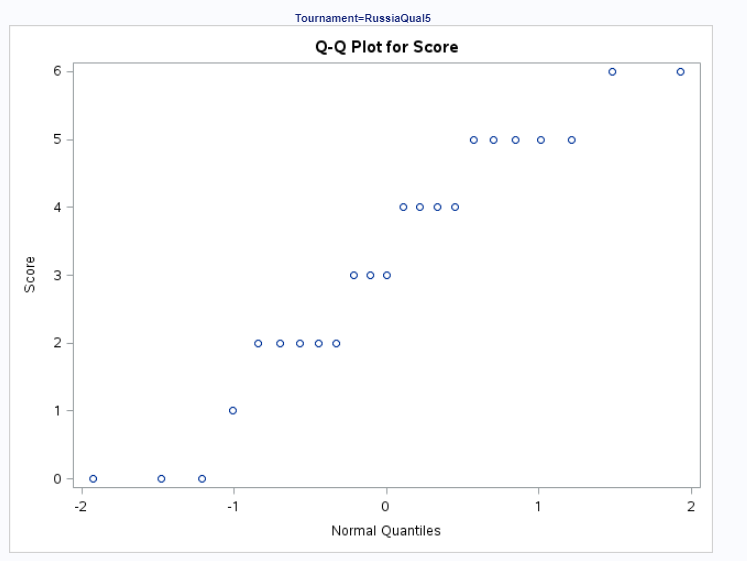
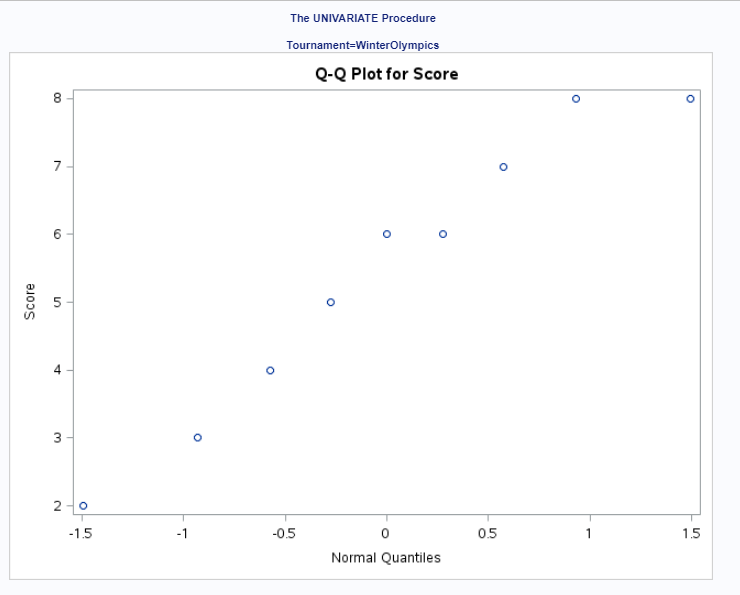
 

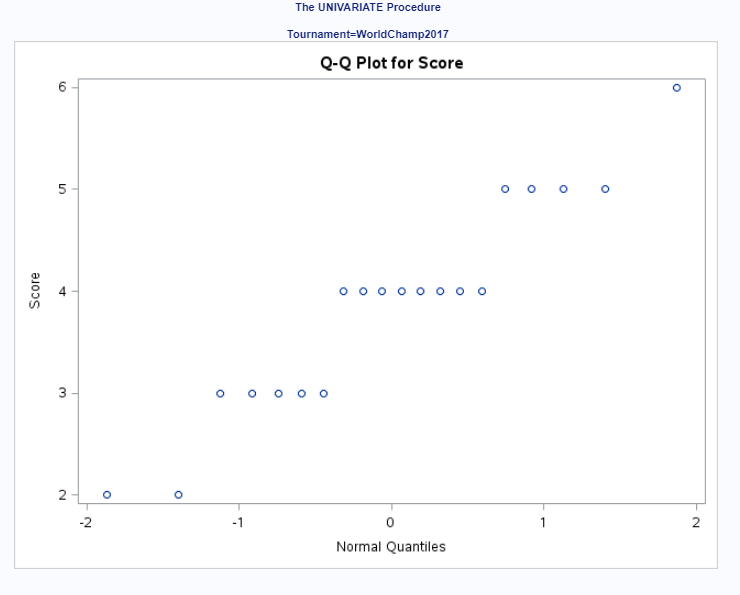


Given the histograms for the 7 tournaments above, the data appear to be normal. Russiaqual looks a little strange, but I think it can still pass as being normal. Here I would say that the noramality assumption passes by looking at the histograms alone.

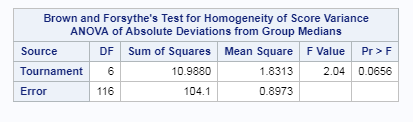
 



When looking at the qqplots, the normality is a little questionable, but one can argue given the histograms and being on the fence with the qqplots that the data is somewhat normal.

Equal SDs



Given that the p-value of .0656 is greater than .05, we can fail to reject the null that the variances are not equal and we can agree with the earlier assumptions that the variances are equal.

Given that I agree that the data is normal and that the variances are equal and I am assuming independence, I would have chosen to do a 1 way anova to determine if there was at least 1 mean between the 7 tournaments that were different as above.

1. (15 points) Is there any evidence that any of the mean or median curling scores from any of the six randomly selected tournaments where Aleksandr tested negative for meldonium are **less than** his mean or median curling score in PyeongChang (where he tested positive for meldonium)? In other words, which “negative for meldonium” tournament scores, if any, are significantly lower than scores in PyeongChang (“positive for meldonium”)?
   * Address the assumptions that are necessary for the test you want to use, but you may assume facts (1), (2), and (3), listed prior to part (a), are true. (No need to analyze graphs.)
   * While a complete analysis is not necessary, your analysis should include a **clear conclusion** that explains your results and supports them with relevant statistics (appropriate **confidence intervals** to quantify significant differences, **p-values**, etc.).
   * Provide your SAS or R code and screen shots of relevant output.
   * Use the alpha = .05 level of significance.
   * You may use two-sided tests for full credit, but it would be even better to ensure that your “sidedness” matches the question of interest.
   * Once you are finished with the analysis above, address the assumptions of the data (without assuming the first three facts prior to part (a)) like you normally would, using relevant graphs, etc. You may summarize the salient points of the assumptions discussion from part (a) rather than recopying the work. Then, write a sentence or two describing the analysis you would perform given your analysis of the assumptions via graphs, etc. It may be exactly what you already performed for this question or a different analysis. In the interest of time, do **NOT** perform this analysis.

Problem: Determine that the means of all the other games are different than the means of the Winter Olympics.

Assumptions:

1. Normality - Given above, we will assume that the data is normal.
2. Equal Standard Deviations – Given above, we will assume equal variances
3. Independence – Given above, we will also assume independence.

Code:

proc sql;

create table work.curling3 as

select case when tournament = 'WinterOlympics' then 'WinterOlympics' else 'Other' end as tournament, score

from work.curling2 order by tournament asc;

run;

data quantile;

myquant = quantile('t',.975, 122); /\*199 is N - 1 for ttest AND .05 FOR TWO TAIL\*/

run;

proc print data = quantile;

run;

proc ttest data = work.curling3 alpha= .05 h0=0 sides=2;

class tournament;

var score;

run;

I’m going to group all other tournaments together as Other and leave the Winter Olympics alone and do a two sample t-test given that the assumptions are met as stated above.

### 6 Step Process:

Step 1: Hypothesis

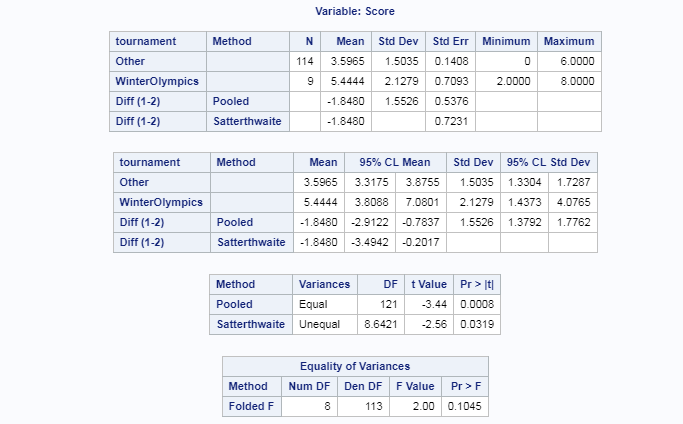
Ho = u1 = u2

Ha = u1 <> u2

#### Step 2: Critical Value

1.97960

#### Step 3: Test Statistic -3.44



Step 4: P-value

.0008

#### Step 5:

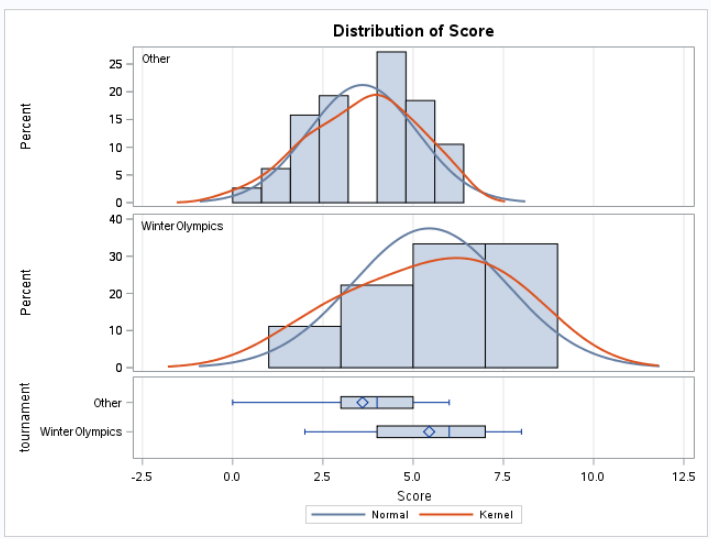
Reject the null

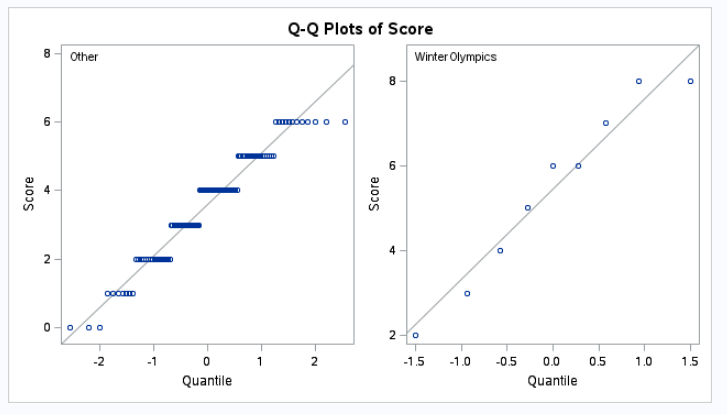
#### Step 6:

There is strong evidence to suggest at the alpha = .05 level of significance (p-value <.0008) that the mean score for all other games is not the same as the mean score at the Winter Olympics with a confidence interval of (.7837,2.9122).

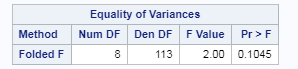
Scope of Inference: This is an observational study as there was no randomized effect nor experimentation. The games were observed and recorded for one person and no causal inference can be made to a bigger population other than the person observed.

Now let’s address the assumptions without the assumptions made previously.





Given the histogram and qqplots above, I would agree with the initial assumptions at the data appear to be normal.



I would also fail to reject the null that the variances are different and assume indepdence. Given all of these assumptions, I would not change the analysis I did above. I grouped the data into Other and did a 2 sample t-test comparing to the Winter Olympics.

1. (12 points) The data set also includes matches that were designated “Home,” “Away,” and “Olympic” matches. It is thought that his “Home” matches have a higher mean score than his “Away” matches. (This is a common sporting theory, whether real or perceived.) Therefore, we would like to test the scores of his “Home” matches against the scores of his “Away” matches. Is there any evidence that his “Away” matches have **lower** scores than the “Home” matches? In other words, is the mean of the away tournament mean scores significantly less than the mean of the home tournament mean scores?
   * Use a contrast to test for this difference. (Do **NOT** relabel or recode the data.)
   * There is no need to perform preliminary tests to see if “Home” tournaments have the same mean/median scores or if “Away” tournaments have the same mean /median scores, as being grouped under the same value (“Home” or “Away”) is sufficient reasoning to perform a contrast that the question of interest addresses.
   * Assumptions do NOT need to be analyzed from charts and graphs, only from the first three FACTS prior to part (a).
   * It is NOT necessary to show the six steps of a hypothesis test, but provide a **clear conclusion** supported by relevant **confidence interval(s)** and **p-value(s)**.
   * Include your SAS or R code, including your contrast weights, and screen shots of relevant output.
   * Use alpha = 0.1 significance level.
   * You may use a two-sided test for full credit, although it would be even better to match the “sidedness” to the question of interest.

We are assuming assumptions have been met:

1. Assume that his curling scores for each tournament come from a **normal** distribution.
2. Assume that the **standard deviations** of his scores for tournaments are consistent (not unequal).
3. Although this may be a questionable assumption, assume that the scores within and between tournaments are **independent** of one another.

Therefore, we can run a contrast model to test between Away and Home. There is a 3rd place which is the Olympics.

Code:

proc glm data = work.curling2;

class hao;

model score = hao;

contrast 'Contrast Away v Home' hao 1 -1 0;

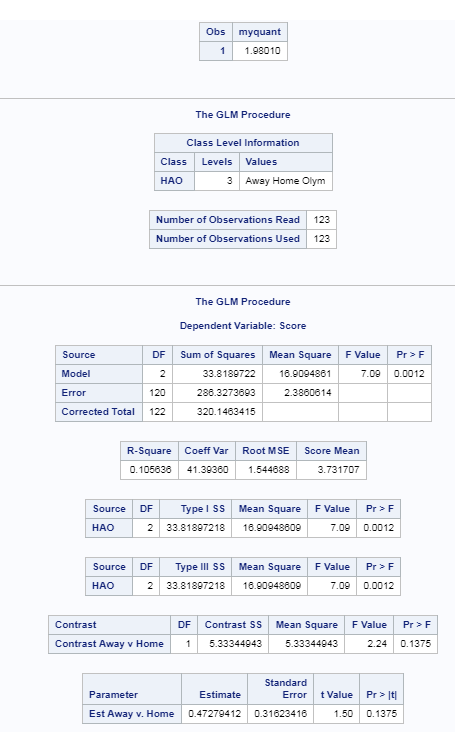
estimate 'Est Away v. Home' hao 1 -1 0;

run;

data quantile;

myquant = quantile('t',.975, 122-3);

run;



We reject the null hypothesis that the Home scores are equal to the away scores with a p-value of .0012. A confidence interval for the difference in means is: .4727 +/- .626 = (.1533, 1.0987)

1. (5 points) What are the advantages/differences when using a contrast for part (c) rather than simply running a two-sample t-test?
   * There are several. Two **distinct** advantages/differences will earn full credit, though there are more.
   * If you can think of more than two, excellent! List them.

By keeping the data together by using a contrast rather than simply running a 2 sample t-test for 2 of the groups vs the rest we benefit from the following:

1. Larger Degrees of Freedom
2. More Power
3. As power increases, the standard error decreases
4. As power increases, the effect size increases