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.
8. (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.

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.
2. (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.