

Gymnastics Judging Final Report

by

Bryan Hartman

Scott Lynch

Ben Pope

Problem Statement

The Olympic Games are a big business. NBC recently paid over 4 billion dollars for the rights to the 2014, 2016, 2018, and 2020 Olympic games.¹ For context that is roughly the same amount of money per day of broadcasting that CBS and FOX pay for the rights to broadcast NFL games², the crown jewel of American Sports. Of all the Olympic sports the primary driver is Gymnastics. During the London 2012 Olympics, the primetime broadcast that included some of the marquee Olympic Gymnastic events drew over 34 million viewers³. That is more than last year's College Football Playoff Championship⁴ and NCAA Men's Basketball Championship⁵ *Combined*. With so much money on the line any scandal can be a significant public relations issue and could have the potential to harm ratings. If gymnastics is known for being one of the most popular Olympic sports, it is also known for having a history of judging scandals and questionable scores. In 2004 Fans booed for 10 minutes after Alexei Nemov performed on the high bar and eventually two judges raised their scores for the athlete. In the same Olympics the Korean team threatened to sue the IOC over perceived low scoring of an athlete's performance⁶. Using a new scoring system in 2012 did not stop the scandals and subjectivity. Perceived low scores prevented American Jordyn Weiber from competing in the all-around final and the Japanese were elevated from 4th to 2nd after lodging a protest of their scores⁷.

Modern statistics can address the problem of judging gymnastics in multiple ways. Constant statistical analysis of individual judges scoring can help bring to light hidden bias in judges to help them improve their performance, and help governing bodies select the most fair judges. There are three types of hidden bias that we will examine in this report. First is general bias, judges who consistently score high or low relative to their peers. Second is country of origin bias, judges who have an inherent bias toward countries with a history of success. Lastly is 'last will be first' bias, judges who reserve the highest scores for those who perform later in the competition. Identifying these types of bias in judges will help improve scoring and transparency in the sport. The public trust in the judging process is paramount.

Data Source

For our data analysis, we focused our attention on the Artistic Women's Gymnastics events from the 2012 London Olympics. This was the most recent Summer Olympics and the first to adopt a new scoring system for women's gymnastics. We gathered our data from the website www.sports-reference.com. This website compiles and shares data from many different sources across the major American sports as well as the Olympics. The origin of the data itself comes from the group

¹ <http://tvbythenumbers.zap2it.com/2011/06/07/report-nbc-wins-latest-olympics-tv-rights-bid/94902/>

² <http://www.adweek.com/news/television/nfl-hammers-out-nine-year-rights-renewals-nbc-cbs-fox-137128>

³ <http://www.adweek.com/news/television/london-2012-nbc-record-ratings-pace-142297>

⁴ http://www.nytimes.com/2016/01/13/sports/ncaaf/college-football-championship-game-tv-ratings-drop-23-percent.html?_r=0

⁵ <http://www.usatoday.com/story/sports/ncaab/2013/04/09/ncaa-mens-basketball-title-game-cbs-overnight-tv-ratings/2067107/>

⁶ http://usatoday30.usatoday.com/sports/olympics/athens/gymnastics/2004-08-24-judging-cover_x.htm

⁷ <http://www.buddytv.com/articles/olympics/olympics-2012-scandal-takes-ce-46841.aspx>

OlyMADMen, who are an international consortium of Olympic historians and statisticians that have collected data on all Olympians since the 1980s. The leader of the group, Bill Mallon, is the founding member and served as the past-president of the International Society of Olympic Historians (ISOH) and former editor of the ISOH Journal that covered Olympic history. The OlyMADMen database contains over 50 million records of data, representing an estimated 100,000 man-hours of work and is widely considered the most reputable source for historic Olympic data.

Our focus for the analysis is on the Artistic Women's Gymnastics Balance Beam event during the Women's Individual All-Around Qualification Phase from the 2012 London Olympics. The data is organized by event in rank order of the athletes from the competition. For each athlete, biographical data, the overall score for the event, the difficulty score for the event, and each judges' execution score for the event.

Method

General: We believed that the following model describes the scores in this event. True scores of each performance are distributed normally $y \sim N(\mu_p, \sigma_p^2)$, the bias of each judge is distributed normally $z \sim N(\nu_j, \tau_j^2)$ and the actual scores for each performance are distributed $\sim N(y + z, \sigma^2)$. We will use two different statistical methods, Bootstrapping and Expectation Maximization, to estimate each of these parameters and based on these parameters we will be able to answer the questions posed above. In general we developed estimates for $z \sim N(\nu_j, \tau_j^2)$ to determine if one or more of the judges tend to score higher or lower to answer the question of whether or not judges have a general bias. Secondly, we will bin the data into 'power countries' and non-'power countries'. We then estimated $z \sim N(\nu_j, \tau_j^2) | z \in \text{Power Countries}$ and compared them to the $z \sim N(\nu_j, \tau_j^2) | z \in \text{Non - Power Countries}$ to determine if there was a difference. For determining 'power countries', we tallied the number of overall medals from the previous 3 world championships. The leading countries were China, US, Russia, and Romania. Lastly we categorized the performances by 'beginning' 'middle' and 'end' of the competition to determine if $z \sim N(\nu_j, \tau_j^2)$ vary over time during the competition.

Expectation Maximization (EM): Our approach to implement the EM algorithm was to estimate the level of bias from each of the judges as well as the true scores for each of the gymnasts. We estimated these two below:

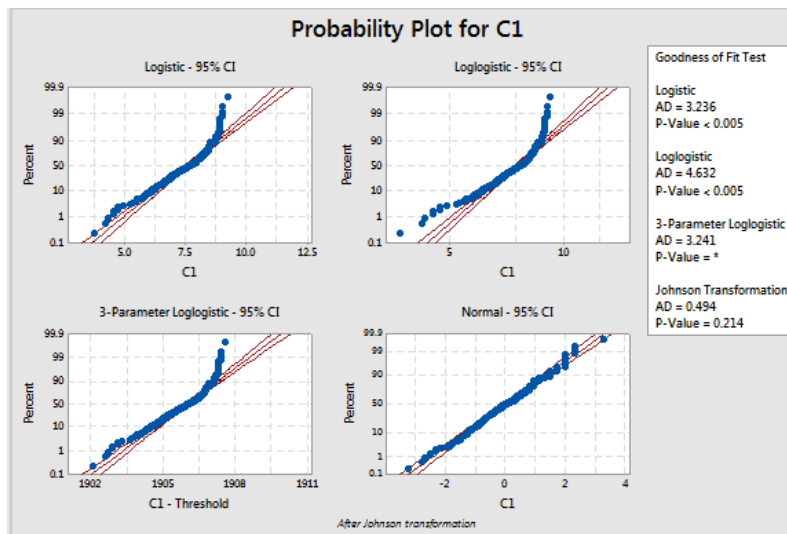
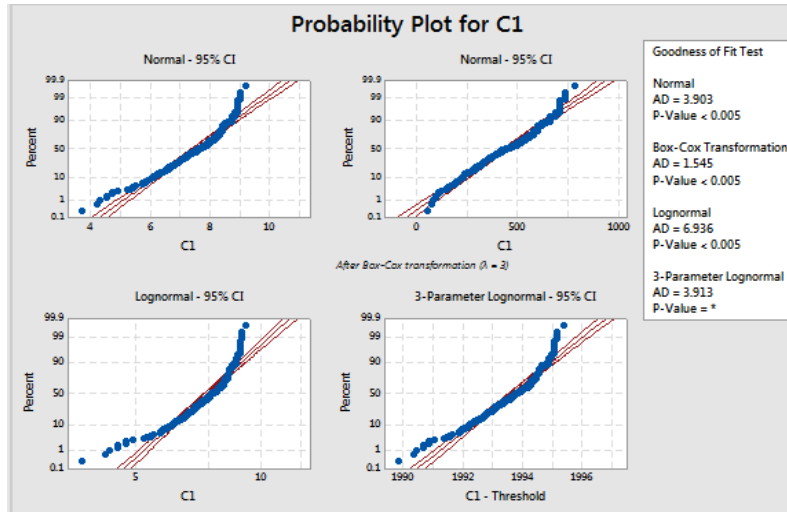
$$\text{True Score: } y \sim N(\mu_p, \sigma_p^2)$$

$$\text{Judge Bias: } z \sim N(\nu_j, \tau_j^2)$$

We derived the estimates for y (true score) and z (bias) from the conditional distribution of x (actual score).

$$x|y + z \sim N(y + z, \sigma^2)$$

Before we were able to implement the EM procedures we needed to ensure that x (the actual scores) were distributed normally. We ran a distribution fit test using Minitab and we were only able to achieve a strong fit through a Johnson SB transformation. Below are several of the results from our goodness of fit tests.



Clearly, the actual scores were not distributed normally. One thing we noticed however was that we could transform our data via the Johnson Transformation to achieve the normality requirement for our EM procedures. We performed this transformation via the Johnson SB Distribution and associated parameters:

$$x' = \gamma + \eta * \ln\left(\frac{x - \varepsilon}{\varepsilon + \lambda - x}\right)$$

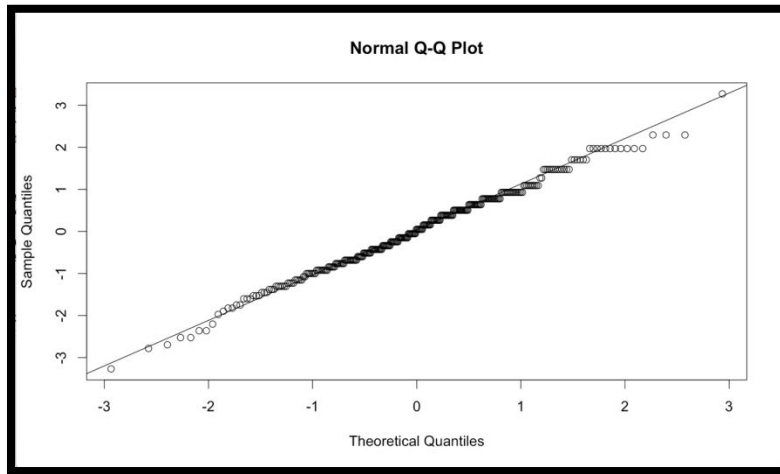
where, $\gamma(\text{shape parameter } 1) = -1.58052$, $\eta(\text{shape parameter } 2) = 1.35706$,
 $\varepsilon(\text{location parameter}) = 2.05860$, $\lambda(\text{scale parameter}) = 7.34161$, x = original data, x' = Johnson
transformed data.

Following this transformation, we performed a Shapiro-Wilk test as well as a qq plot to ensure our data
was normal and arrived at the following results.

Shapiro-Wilk normality test

data: x_ij

W = 0.9946, p-value = 0.3692



The results indicate that our transformation was successful and that we achieved our normality
requirement necessary to perform the following EM procedures. First, we had to calculate the
conditional bivariate normal distribution for y (true score) and z (bias) given x (actual score). We did this
by taking the joint pdf of all three and applying Bayes rule to arrive at the following conditional expected
value for y and z for our E-Step:

$$\begin{bmatrix} u_1^{pj} \\ u_2^{pj} \end{bmatrix} = E \begin{bmatrix} y^{pj} \\ z^{pj} \end{bmatrix} | X^{pj} = x^{pj} = \begin{bmatrix} u_p \\ v_j \end{bmatrix} + \begin{bmatrix} \sigma_p^2 \\ \tau_j^2 \end{bmatrix} * \left(\frac{x^{pj} - (u_p + v_j)}{\sigma_p^2 + \tau_j^2 + \sigma^2} \right)$$

$$\begin{bmatrix} \sigma_{11}^{pj} & \sigma_{12}^{pj} \\ \sigma_{21}^{pj} & \sigma_{22}^{pj} \end{bmatrix} = \begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \tau_j^2 \end{bmatrix} - \begin{bmatrix} \sigma_p^2 \\ \tau_j^2 \end{bmatrix} * \begin{bmatrix} \sigma_p^2 & \tau_j^2 \end{bmatrix} * \left(\frac{1}{\sigma_p^2 + \tau_j^2 + \sigma^2} \right)$$

*Note: We'll only need to update u_1^{pj} , u_2^{pj} , σ_{11}^{pj} , and σ_{22}^{pj} for each iteration since these are the only
parameters we'll use for the following M-Step. We computed the maximum likelihood for our
parameters used in our M-Step as follows:

$$u'_p = \frac{1}{5} * \sum_{j=1}^5 u_1^{pj}, p = 1, \dots, 60$$

$$v'_j = \frac{1}{60} * \sum_{p=1}^{60} u_2^{pj}, j = 1, \dots, 5$$

$$\sigma'_p = \frac{1}{5} * \sum_{j=1}^5 (\sigma_{11}^2 + (u_1^{pj} - u'_p)^2), p = 1, \dots, 60$$

$$\tau'_p = \frac{1}{60} * \sum_{p=1}^{60} (\sigma_{22}^2 + (u_2^{pj} - v'_j)^2), j = 1, \dots, 5$$

We ran our algorithm 1,000 times, resulting in updated true scores and judge bias. However, in order to scale our data back to its original levels and to compare our updated scores with the original scores, we had to inverse transform our data back into its original state using an inverse Johnson function as follows.

$$x = \frac{\exp\left(\frac{x' - \gamma}{\eta}\right) * (\varepsilon + \lambda) + \varepsilon}{\exp\left(\frac{x' - \gamma}{\eta}\right) + 1}$$

with the same original parameters described earlier.

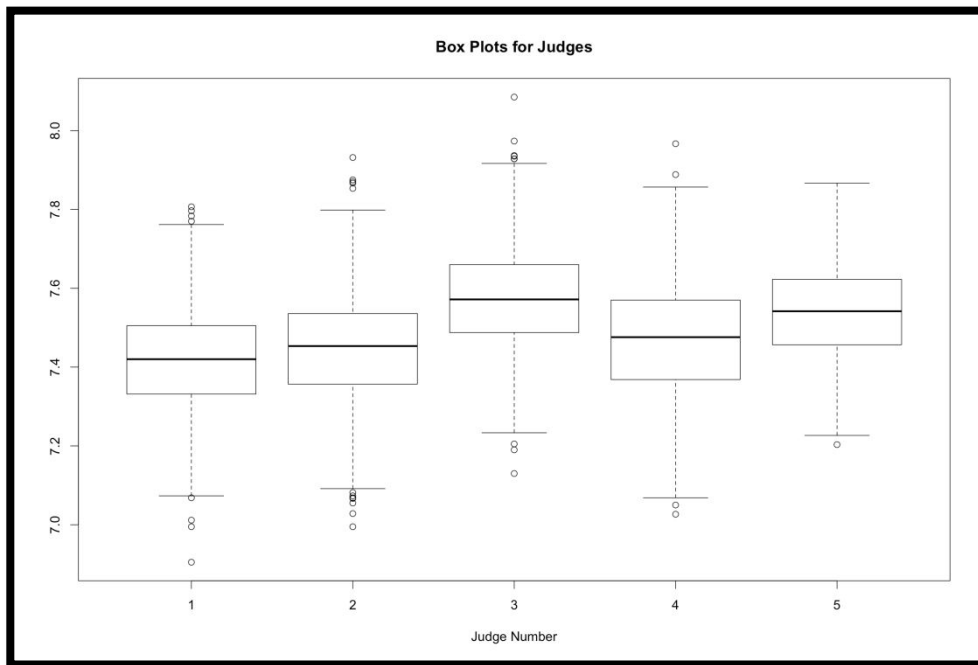
Bootstrapping: For the bootstrap procedure, we used a non-parametric resampling method to resample the judges' scores over 1,000 replications. We first examined all 60 of each judges' scores and resampled them 1000 times to generate a list of 1000 resampled averages. This list of 1000 averages is known to be normally distributed. We can compare the parameters of these distributions to determine if one judge gave higher or lower scores than the other judges. We also examined if judges are consistent in scoring gymnasts from the 'power countries' and the 'non-power countries'. There were 11 gymnasts from power countries. The difference between each judge's score and the average score for that gymnast was computed and bootstrapped. This gave a normal distribution for that judge's bias toward power countries. This method was repeated with the 49 gymnasts from the non-power countries to give a normal distribution for the bias toward non-power countries. Consistent judges should have very similar parameters between their bias for power countries and non-power countries. Finally, we examined if judges are consistent in scoring gymnasts based on performance order. All 60 gymnasts were divided into 5 subgroups, with 14 in Sub-group 1, 13 in Sub-group 2, 11 in Sub-group 3, 10 in Sub-group 4, and 12 in Sub-group 5. Gymnasts were placed in a subgroup based on a random drawing ensuring that complete teams were placed in the same subgroup. This randomization means that the level of competition in each subgroup should be equal. In terms of competition order S1 went first followed by S2, S3, S4, and S5. The data was re-sorted in order of competition. Each judge's scores

were bootstrapped for every subgroup and for all 60 gymnasts, and bias for each subgroup and judge was calculated by taking the difference between the mean of a judge in that subgroup and the mean of the judge for all gymnasts. Consistent judges should have a very small difference in means, as it should not matter when a gymnast competes compared to her score.

Evaluation and Final Results

Our final results are organized into three subsections and a final assessment. The subsections include the overall bias, the power country bias, and the subsection bias. See below for our evaluation of the methods and the final results.

Overall Bias: When using both the EM method and the Bootstrapping method we discovered that Judge 3 gives the highest scores and judge 1 gives the lowest scores. Below is the box plot for all judges' scores generated by the bootstrapping method.



Based on the results of our box plot, we can see that judge 3 appears to have a higher average than the other judges and judge 1 is slightly lower. For significance testing, a t test for the difference between means will not be useful in this case because each sample has $n=1000$. However, we can qualitatively see that the judges' middle 50% intervals overlap indicating that the judge's scores are still relatively close to one another. We calculated our judge bias as the difference between the individual judge's bootstrap average and the average between the all five judges. We then used this information to recalculate the rankings and determine what differences existed between the original scores and the bias corrected scores. Based on this methodology, we arrived at the following judge bias averages.

Judge	Overall Bias
Judge 1	-.0727
Judge 2	-.0424
Judge 3	.0857
Judge 4	-.0195
Judge 5	.0489

When we used the EM methodology we observed the overall bias to be as follows.

Judge	Overall Bias
Judge 1	-.0815
Judge 2	-.0548
Judge 3	.0818
Judge 4	-.0148
Judge 5	.0351

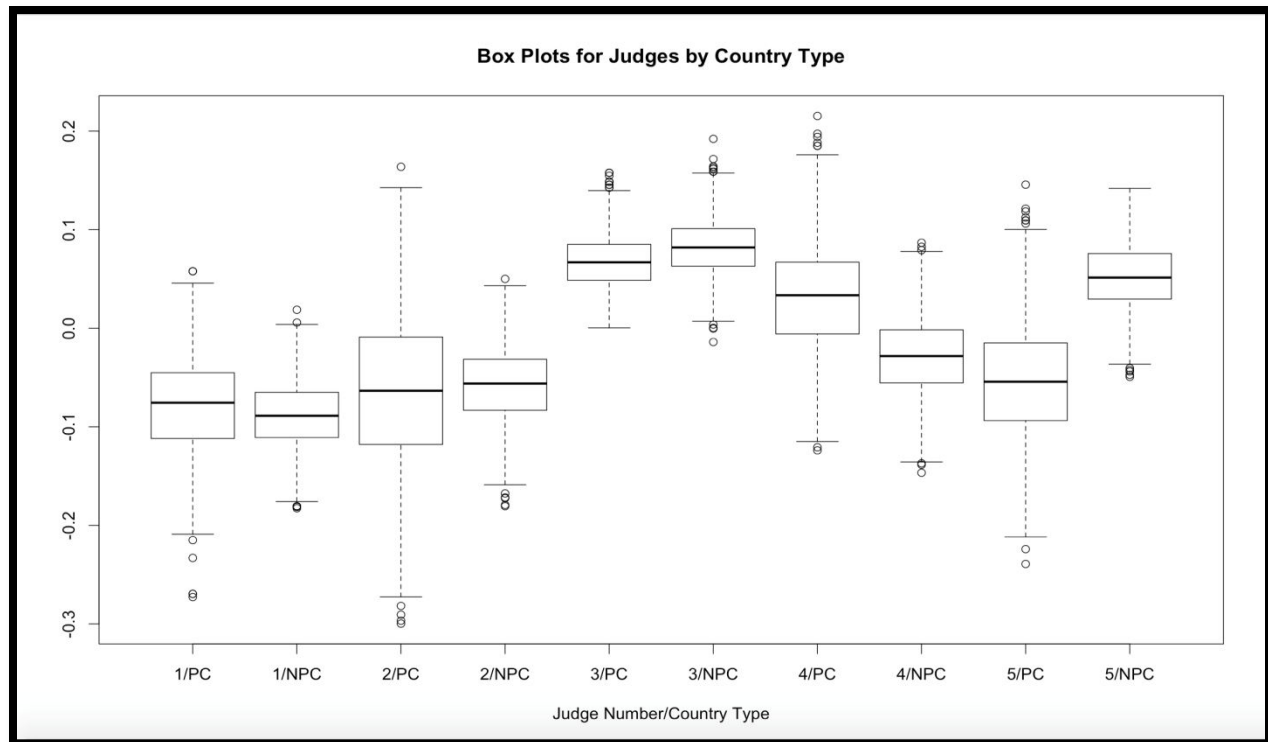
The bias scores in the two tables were very similar in exact value and exactly the same in rank order. This gives credence to both methods since they are independently validated by each other. It also enables us to say with confidence that judge 3 gives higher scores than judge 1.

After this analysis we removed judge bias and re-ranked the athletes, there were very few significant changes to the overall standings. In general the Bootstrap method was more conservative resulting in fewer and less dramatic position changes. The EM method was more aggressive and resulted in more and larger position changes. In both models about 1:3 of the 24 athletes in qualifying positions moved rank by one place. The largest jump was 6 places (41st to 35th place). The largest drop was 11 places (40th to 51st place). While judge bias was relatively minor, this analysis demonstrates that even small changes to the data can have a significant impact on the outcome of a competition.

Power Country Bias: The data for power country bias did not support the EM method. In order for the EM algorithm to work the data must be normally distributed. In the above case we were able to make it normally distributed using a Johnson transformation, but none of the known transformations were able to make the data subsets for power country and non-power country normal. Therefore, only the bootstrapping method was used for this portion.

Each judge's bias toward power countries and non-power countries is shown below as well as the two box plots for each judge.

Judge	Power Country Bias Mean	Non-Power Country Bias Mean	Power Country Bias Var	Non-Power Country Bias Var
Judge 1	-.079	-.086	.00022	.00015
Judge 2	-.058	-.060	.0003	.0003
Judge 3	.069	.081	.00002	.0001
Judge 4	.030	-.030	.0001	.0003
Judge 5	-.050	.051	.0001	.0001



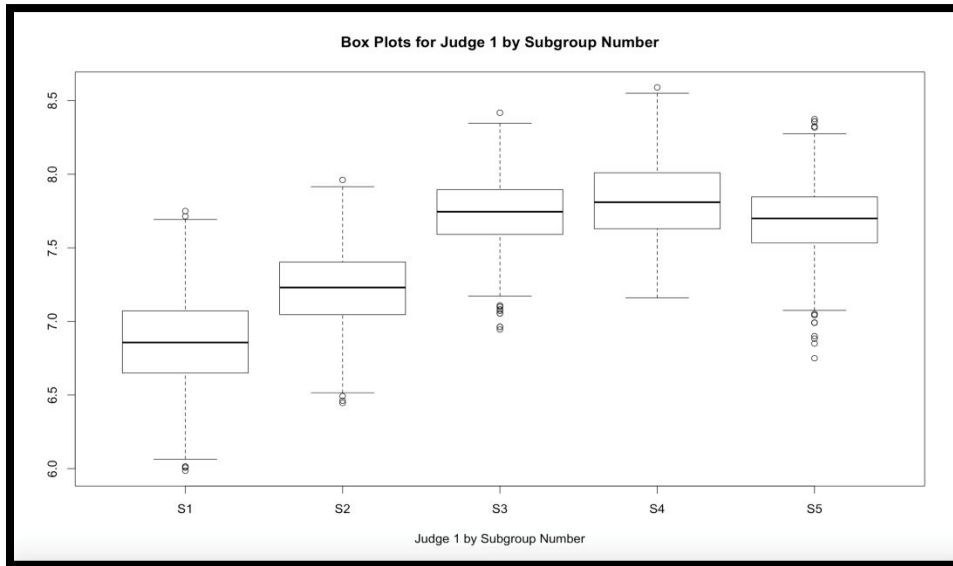
We concluded from this data that Judge 5 had a bias toward non-power countries and judge 4 had a bias toward power countries. Judge 5's bias was severe enough that he generally scored athletes from non-power countries .1 points higher than he would have scored the routine if the gymnast had been from a power country. That is a big difference. Although not all judges have a bias based on a country's gymnastic history having two of five judges display this bias is enough to warrant using it as evidence for further training and a criteria in judge selection.

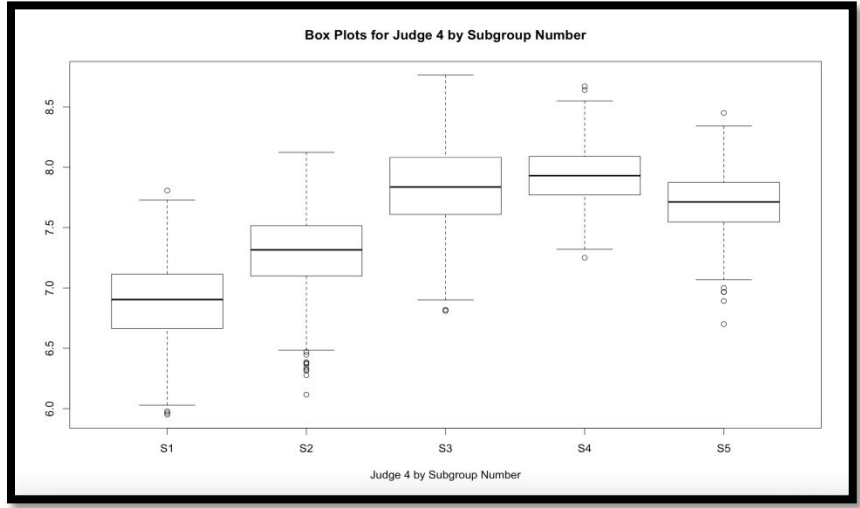
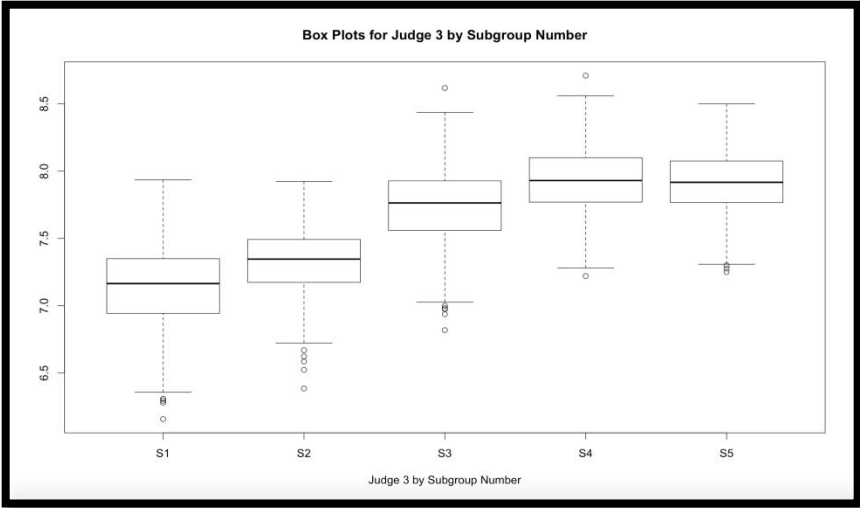
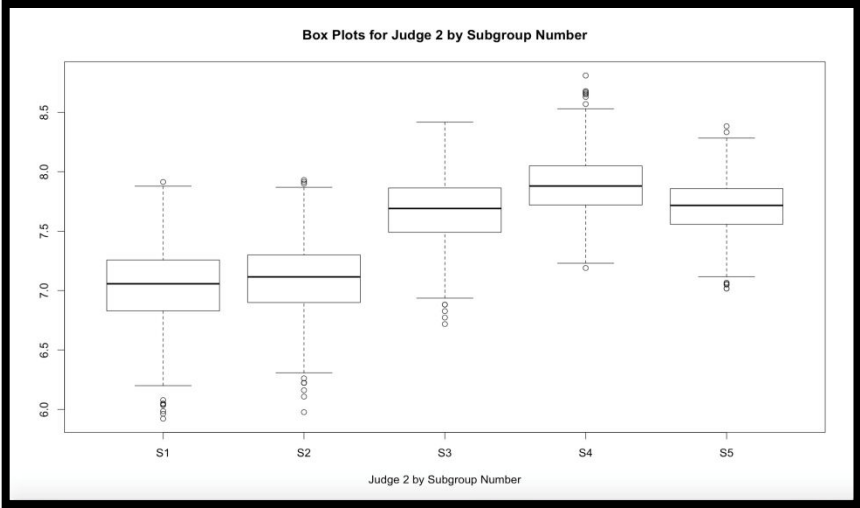
For significance testing, a t test for the difference between means was not useful in this case because each sample has $n=1000$. When the sample size is so high the t statistic is inordinately high and therefore the null hypothesis that the means were the same will always be rejected. However the objective is not to see if the population means were the same, only if they were similar. Qualitatively we saw that judge 1, 2, and 3 all had very similar and consistent biases between power countries and non-power countries and seemed to not favor power countries or non-power countries. However judge 4 had a .06 difference between his/her bias for power countries and non-power countries. Judge 4 seemed to favor power countries. Judge 5 had an even larger difference of .1, but judge 5 seemed to favor non-power countries. Having 2 of 5 judges show bias based on if the gymnast is in a power country or not showed a flaw in the judging system that can be addressed through awareness, training and judge selection.

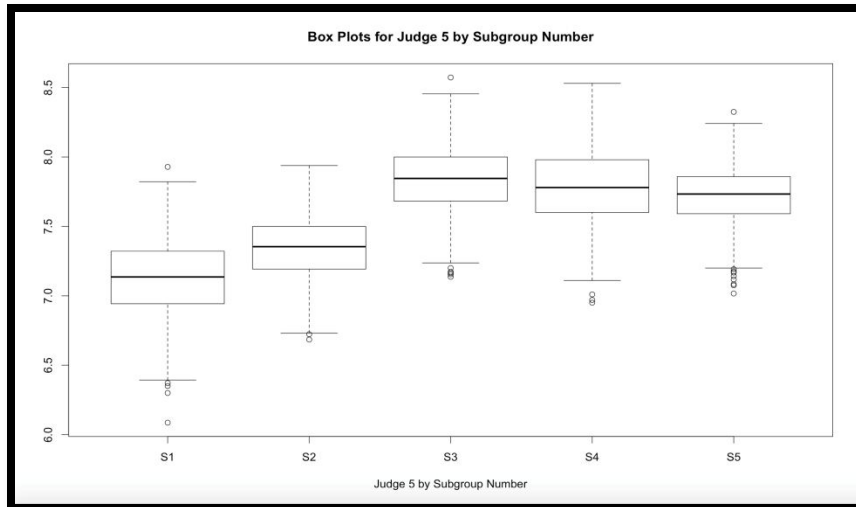
Subgroup Bias:

BIAS	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
Subgroup 1	-0.54582	-0.41416	-0.43498	-0.57848	-0.40321
Subgroup 2	-0.22162	-0.34811	-0.25288	-0.17911	-0.18259
Subgroup 3	0.327533	0.24333	0.154518	0.32135	0.301535
Subgroup 4	0.377137	0.463633	0.368475	0.455982	0.245213
Subgroup 5	0.23893	0.266968	0.311045	0.230323	0.19927

The results of the subgroup bias are shown in the above table. Subgroup 1 was scored in the most negative sense, followed by Subgroup 2, Subgroup 5, Subgroup 3, and Subgroup 4 being most favorably scored. This follows the trend that earlier competitors were judged more harshly than later competitors. The bias scores do not exclusively reflect only time bias, as certain power countries (such as China in Subgroup 4 and the United States in Subgroup 3) further inflated the bias. The box plots for each judge are shown below:







Consistent with all judges, the scores in Subgroups 3, 4, and 5 were all higher and had no overlap with the scores in Subgroups 1 and 2. This further shows the trend that as groups go later in the day the scores go higher. Once again Subgroup 4 had the highest mean out of all the groups, which points to a combination of time of competition and the teams involved in each Subgroup. The judge's scores were recalculated by subtracting the bias to get the true score. Averages were determined, and gymnasts were re-ranked according to their new true scores. In terms of the top 24 qualifying the originally ranked 23rd and 24th gymnasts were dropped out and the 33rd and 36th gymnasts jumped in. The remaining 22 gymnasts were reshuffled, which had no effect on who advanced and who did not.

EM

The EM method proved to be more extreme in its measurement of bias. The table of bias for each judge is below:

EM BIAS	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
Subgroup 1	-0.39286	-0.39286	-0.41429	-0.42143	-0.39286
Subgroup 2	-0.11538	-0.27692	-0.06923	-0.00769	0.007692
Subgroup 3	0.527273	0.436364	0.454545	0.445455	0.5
Subgroup 4	-0.11	0.12	-0.04	0.02	-0.14
Subgroup 5	0.191667	0.258333	0.175	0.075	0.108333

As with the bootstrapping method, Subgroups 1 and 2 had negative bias, with the exception of Judge 5 for Subgroup 2. Subgroup 3 was determined to have the highest bias. The surprise result was Subgroup 4, which showed to have a negative bias in three out of the five judges while in the bootstrapping method it had the highest positive bias. Looking into the actual data the 42nd and 48th ranked gymnasts in Subgroup 4 drove the average bias down with a bias of -1.5 or greater for most judges, which is improbably large for gymnastics. Further in the re-ranking of gymnasts over half of the qualifiers were dropped out of the top 24, while the two gymnasts tied for 29th respectively placed 1st and 4th. Though there is certain expectation that some gymnasts would qualify unfairly due to the bias,

this points to the shortcomings of the EM method as it is less robust than the bootstrapping method. Overall the trend still holds that the later subgroups had more positive bias than the earlier subgroups.

Overall, these results must be taken with reservations. We do believe there is a trend in earlier competitions being judged more harshly than later competitions. This belief and the statistical results supporting this are dependent on the notion that the gymnasts are placed randomly in a subgroup. While the intention of randomness was true, the actual placement showed skewness towards the later subgroups, as most of the power countries were placed in subgroups 3, 4, and 5.

Final Assessment

In our project proposal, we formed three hypotheses for the types of bias we expected to find during our analysis. First, we expected to see that certain judges gave higher or lower scores to everyone relative to other judges. Our analysis confirmed this expectation since judge 1 and judge 3 appeared to have differences from one another. Second, we expected to see confirmation bias from judges toward 'power countries'. Judges 4 and 5 appeared to have some of bias toward these countries. Lastly, we expected to see the gymnasts performing in later subgroups receiving higher scores. This was identified, but the results were inconclusive given the nature of the countries assigned to these later subgroups. Overall, we believe our analysis results were close to what we expected.

We encountered a few problems during the data analysis process. For one, we were not able to effectively perform formal statistical significance testing procedures. This was due to the high number of samples generated from our bootstrap method as well as our EM procedure only generating parameters for our distributions. Instead, we used a qualitative approach to compare judge bias. Another issue was assessing the subgroup bias. Since the subgroups appeared to have much stronger competitors in later subgroups, it was difficult to identify judge bias.

In order to improve our results, we could examine several follow-up options. For one, we could assess different gymnastics events as well as other Olympics. While our study was representative of the population we set out to examine (all competitors and judges scores for the Beam Qualifying event at the London Olympics), we could look at additional events and Olympics. This would help us identify trends in judge bias.

Overall, we believe our approach met our study objectives. We were able to identify where bias existed and the effects that it can have on overall scoring. If judges received this information during the early stages of the competition, they would be able to root out unintentional bias which would lead to better assessments during the main events. We strongly recommend that judge governing bodies use similar types of procedures to educate and inform their judges.