Problem Statement (Scott Lynch)

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 games[2], 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 viewers[3]. That is more than last year's College Football Playoff Championship[4] and NCAA Men's Basketball Championship[5] *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. 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 performance[6]. Using a new scoring system in 2012 did not stop the scandals. 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 scores[7].

Modern statistics can address the problem of judging gymnastics in multiple ways. Constant statistical analysis of individual judges scoring can help the sport in three ways. First, it can identify judges who consistently score high or low relative to their peers, this is an issue that can be then addressed in judge selection, or judge training. Second, it can identify judges with patterns of abnormal scoring. This will help bring to light and weed out potentially corrupt or biased judges. Thirdly, it can help make a case to the public that most gymnastics judge are doing it right and giving an accurate representation of the scores. The public trust in the judging process is paramount. We will analyze a sample of Gymnastics scores to create a system to assess if 1) judges have a general bias, 2) judges have a targeted bias toward 'power countries', and 3) if there is an inflationary bias that occurs as the event progresses.

Data Source (Bryan Hartman)

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

[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/ncaafootball/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

American sports as well as the Olympics. The origin of the data itself comes from the group 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. The data for each athlete includes biographical data, the overall score for the event, the difficulty score for the event, and each judges' execution score for the event.

<u>Method</u> (Group Collective)

We believe 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 to estimate each of these parameters and based on these parameters we will be able to answer the questions posed above. In general we will develop estimates for $z\ (\nu_j, \tau_j^2)$ 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 will then estimate $z\ (\nu_j, \tau_j^2)|z\ \epsilon\ Power\ Countries$ and compare them to the $z\ (\nu_j, \tau_j^2)|z\ \epsilon\ Non - Power\ Countries$ to determine if there is 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 will bin the performances by 'beginning' 'middle' and 'end' of the competition to determine if $z\ (\nu_j, \tau_j^2)$ vary over time during the competition.

We will estimate the above parameters using two approaches. The first approach will be to perform the resampling technique of the bootstrap. For this approach, we will bootstrap each performers score by resampling the 5 judges' scores 10,000 times (with replacement). These 10,000 means will be normally distributed with $y{\sim}N\ (\mu_p, \sigma_p^2)$. We will then bootstrap the judges' scores by using each individual judge's scores for all 60 performers. This will give us an estimate for □ + □ and we will be able to use the given data for X to solve for the parameters in $z\ (\nu_j, \tau_j^2)$.

Our second analytical approach involves the use of the EM algorithm to estimate the parameters listed above to answer the same questions. We will then be able to compare our conclusions for each approach to see if both are consistent. We can also do quantitative analysis to determine if one method is a more efficient way of answering these questions.

Expected Results (Ben Pope)

While we may only see miniscule indications of judge bias, gymnastics event scores are separated by a razor thin margin. Therefore, it is paramount to identify these indications as it could mean the difference between a gold and silver. We have formed three hypotheses on what biases we may find in the data set. First, judges consistently give higher or lower scores to everyone compared to his or her peers. The degree to which a person is impressed by a performance is an innate personality trait, one that training or correction cannot eradicate. This manifests itself in judges being more skeptical or optimistic in their scoring. Second, we expect to see a confirmation bias in judges' scores. Countries such as the USA, Russia, and China are routinely at the top of the rankings for gymnastics and therefore we expect to see bias toward performers from 'power countries'. Thirdly, we expect that gymnasts that perform later in the event score higher than gymnasts that perform earlier. This so called "last will be first" phenomenon is a psychological effect in all sports that require judging.[8] To reduce this effect gymnastics judging has switched from end-of-sequence scoring to step-by-step scoring. In addition to these three biases that we predict, we will identify other potential biases that arise from our statistical analysis.

[8] http://www.slate.com/articles/sports/explainer/2012/07/olympic_gymnast_jordyn_wieber_upset_scoring_bias_favors_late_performers_.html