The Statistical Sports Fan

Judging Figure Skating Judges

Robin Lock **Kari Frazer Lock**

Figure skating is a wonderful sport, combining athleticism with artistry. However, unlike most other sports, figure skating judging is subjective. It is not like hockey, soccer, or lacrosse in which the team that scores the most goals wins, or like swimming or running in which the fastest person wins. Rather, a panel of judges decides the winner of a figure skating competition. The winner is chosen based on the subjective opinions of human beings who rank the performances of each of the skaters in the competition. This subjectivity sometimes causes skaters, coaches, commentators or fans to question the fairness of the judging for a particular event. How can we determine *statistically* when a specific figure skating judge produces a ranking of the skaters that is significantly different from the rankings of the other judges? We describe a technique, using a bootstrap distribution, for identifying an inconsistent judge and apply the method to competitions from the 2002 Winter Olympic Games.

Measuring a Judge's Judgments

The system for determining the final placement of skates in a competition is somewhat complex (see sidebar) due to safeguards that are designed to prevent any one judge from exerting too much influence on the final result. See Basset & Persky (1994) or Russell (1997) for additional discussion of the merits of the "best of majority" method. The primary feature of this judging system that is important for our comparison of judges is that each judge produces a rank ordering of all of the skaters in a competition. For example, Table 1 gives the rankings for each of the nine judges (and the final placements) for the ladies free skate event at the 2002 Winter Olympics. Judges are human, they each have their own tastes and preferences, they may notice different elements of a particular performance and therefore we should not expect them to produce identical rankings for a particular set of skating performances. A certain degree of variability in the judge's rankings is inevitable, particularly in a close competition where the distinctions between the quality of the performances are small. But, occasionally one judge appears to stand out as being in noticeable disagreement with the other judges. Our task is determine when the deviations of one judge's rankings are significantly larger than one would expect to see, given the variability of the rankings for all the judges in that competition.

Since the methods for determining the final placement of skaters should not be unduly influenced by one inconsistent judge, we can determine that one judge is in significant disagreement with the other judges if that judge's rankings differ significantly from the final placement of the skaters. Although disagreeing with the other judges might not necessarily be bad, we will refer to the extent that a judge's rankings match the final placement of the skaters as the "success" of that judge. The rankings of a successful judge will closely match the final placement of the skaters, while the rankings of an unsuccessful judge will disagree with the final placements. To determine how much an individual judge agrees with the final placement, we will look at the Spearman rank correlation (just a correlation between the ranks of the data) between that judge's ranking and the final placement of the skaters. A high correlation will indicate a successful judge, while a lower correlation will indicate a less successful judge.

Table 2 gives the rank correlations of each judge with the final placements, from the ladies free skate event at the 2002 Winter Olympics.

The first thing to notice when looking at these correlations is that they are all extremely high. A perfect

Kari Frazer Lock is a junior majoring in mathematics and psychology at Williams College. She has earned United States Figure Skating Association gold medals in freestyle, moves in the field, and ice dancing. She skates professionally in shows across the U.S. and abroad.

Table 1: Ranks given by nine judges of the Ladies Free Skate at the 2002 Winter Olympics

Final Placement	Skater		J1	J2	J3	J4	J5	J6	J7	J8	J9
1	HUGHES Sarah	USA	1	4	3	4	1	2	1	1	1
2	SLUTSKAYA Irina	RUS	3	1	1	1	4	1	2	3	2
3	KWAN Michelle	USA	2	3	2	2	2	3	3	2	3
4	COHEN Sasha	USA	5	2	4	3	3	4	4	4	4
5	SUGURI Fumie	JPN	4	8	5	5	5	7	5	5	5
6	BUTYRSKAYA Maria	RUS	6	5	8	7	12	5	8	7	6
7	ROBINSON Jennifer	CAN	7	7	7	9	6	8	10	6	7
8	SEBESTYEN Julia	HUN	8	10	12	8	7	6	12	8	8
9	KETTUNEN Elina	FIN	9	9	13	6	12	10	7	11	14
10	VOLCHKOVA Viktoria	RUS	10	6	14	11	10	12	6	9	15
11	MANIACHENKO Galina	UKR	13	12	11	12	16	11	11	10	9
12	FONTANA Silvia	ITA	14	11	18	16	9	15	9	12	10
13	LIASHENKO Elena	UKR	15	13	6	10	8	14	13	14	16
14	ONDA Yoshie	JPN	11	14	10	15	15	13	15	13	11
15	HUBERT Laetitia	FRA	12	17	17	13	11	16	14	15	13
16	MEIER Sarah	SUI	16	16	9	14	14	9	16	16	12
17	GUSMEROLI Vanessa	FRA	17	15	15	17	17	18	17	17	17
18	SOLDATOVA Julia	BLR	19	18	22	20	21	17	18	18	19
19	HEGEL Idora	CRO	20	21	16	22	18	19	21	19	18
20	GIUNCHI Vanessa	ITA	18	19	20	21	19	20	20	20	20
21	BABIAKOVA Zuzana	SVK	22	20	19	19	20	21	19	22	22
22	KOPAC Mojca	SLO	21	22	23	18	22	22	22	21	21
23	LUCA Roxana	ROM	23	23	21	23	23	23	23	23	23

correlation (where the judge agrees exactly with the final placements) is 1.0, and only one of these correlations is even below 0.9. This tells us that each of the judges is in general agreement with the final placement of the skaters (and so in general agreement with each other). This is good news! It tells us that judges are basing the judging more on the skaters than on individual preferences (if they were judging based on individual preferences, the correlations would not be so high). However the skaters are being ranked (hopefully it is based on their skating performance, but it could also be based on their reputation, among other things), they are being ranked in a way in which the judges agree. The rankings are not arbitrary from judge to judge, and thus the final placements are more credible and meaningful.

Looking at the correlations for each judge, you will notice that the correlation between Judge #3's rankings and the final placement is lower than the correlations of the other judges. Obviously, one judge must have the lowest correlation. How do we know if the correlation for Judge #3 is simply the lowest correlation of these judges, or if it is *significantly* lower than the correlations of the other judges?

Ordinarily, to determine if a sample statistic (such as Judge #3's correlation) is statistically significant, we would compare the statistic to some underlying distribution, look at how far out on the distribution it lies, and calculate the probability of a sample statistic being that far out if all of the judges were consistent. If this probability is small, then the sample statistic is statistically significant and we would conclude that the judge is inconsistent with the others in the panel. But what is our underlying distribution in this case? We can't really compare the sample results to a population of all possible judges, or to all events, since each event is different and some events are harder to judge than others (sometimes there is a clear order in which the skaters should be ranked, which would give extremely high correlations, and sometimes all the skaters are about the same, which would give low correlations). Here we are only comparing the correlation of Judge #3 to the correla-

Table 2: Rank correlations for judges of the Ladies Free Skate

J1	J2	J3	J4	J5	J6	J7	J8	J9
.979	.970	.876	.953	.928	.960	.966	.993	.951

After watching an individual or pair skate, each judge awards two scores (on a 0-6.0 scale); one for the *technical merit* of the performance and the other score for its *artistic presentation*. These scores are then added together to give a combined score for each skater. After all the skaters have competed, they are ranked for each judge according to these combined scores, with the skater with the highest combined score receiving a first from that judge. In the case of a tie, the skater with the higher presentation mark is ranked higher. This results in each skater receiving an ordinal (rank) from each judge, with 1 being the top rank. The ordinals for the ladies free skate from the 2002 Winter Olympics are shown in Table 1. Final placements are determined by comparing the ordinals of the skaters.

The key factor in determining most final placements is a skater's *median ordinal*, the position where a majority of the judges (at least 5 of 9 in our case) place the skater at or better. A skater with a better median ordinal will always finish ahead of a skater with a worse (larger) median ordinal. When skaters finish with the same median ordinal, preference is given to the skater who has more judges giving that rank or better (called the *size of the majority*). If a tie still exists, the actual ordinals on the majority side (those at or better than the median) are added, with preference given to the smaller sum. If the tie is still unbroken, the ordinals for all of the judges are added. If that fails to determine a winner, the skaters are officially listed as tied.

Let's see how these rules apply to determine some of the final placements for the ladies free skate shown in Table 1. Five of the nine judges (a majority) placed Sarah Hughes in first, so she won the top spot (and the gold medal). Irina Slutskaya and Michelle Kwan each had a median ordinal of 2, but Kwan only got five judges at 2nd or better, while Slutskaya had six, so Slutskaya finishes second and Kwan drops to third. Sasha Cohen and Fumie Suguri had median ordinals of 4 and 5 respectively, so they are easily placed in those positions. Next come Maria Butyrskya and Jennifer Robinson, each with a median ordinal of 7 and exactly six of the nine judges placing them at 7th or better. Butyrskya's majority included ranks of 6, 5, 7, 5, 7, 6 while Robinson's were 7, 7, 7, 6, 6, 7, so Butyrskya's sum is smaller and she gets the sixth spot and Robinson goes to seventh in the final placements.

An important goal of this system is to prevent a lone judge from single-handedly helping or hurting a skater by giving them a mark much lower or higher than they deserve. For example, if you look at the results for Sasha Cohen in Table #1, all that mattered in her ranking was that 8 of her 9 ordinals were 4th or better. She would have received the same final placement if Judge #2 had not been so generous giving her 2nd or if Judge #1 had ranked her 17th instead of 5th. While the judges' opinions should certainly determine the final outcome of the event, no rogue judge should be able to unduly influence the results with rankings that are inconsistent with the others on the panel.

tions of the other eight judges of the event, so its difficult to determine a probability for how "unusual" the correlation of 0.876 is in this case.

Constructing a Bootstrap Distribution

A general technique for using the data in a sample to produce a reference distribution for a sample statistic is called the *bootstrap*. The basic idea is to randomly select elements of the sample itself to generate new samples and then to examine the distribution of the statistic in question for all of these new samples. Thus we don't need to make assumptions about the distribution of the underlying population itself; instead we let the bootstrap samples reveal relevant structure. In our case of figure skating judges, we use the rankings provided by the nine actual judges to produce a much larger "population" of judges with similar rankings. We can then create a whole distribution of rank correlations of the rankings of these simulated judges with the actual final placements of the skaters in the competition and see where the correlation of Judge #3 fits in this distribution. For more information on bootstrap techniques see Efron and Tibshirani (1993).

We start with the 9 ordinals (rankings) each skater received (one from each "real" judge). To create a simulated judge's score for a particular skater, randomly choose one of the 9 ordinals actually received by that skater. For example, since Michelle Kwan received five 2nd's and four 3rd's, each simulated judge has a 5/9 probability of giving Michelle 2nd, and a 4/9 probability of giving Michelle 3rd. (Thus, the simulated judge is judging the event as a real judge would) Thus we are simulating the behavior of a real judge, since the simulated ordinal given each skater was actually received by the skater from a real judge, and if more actual judges gave a skater one ordinal, more simulated judges would tend to give a skater that ordinal. Then repeat this random selection from the ordinals received by the other skaters. In actual competitions, a judge rarely gives two skaters the same ordinal (an exact tie); but this could easily occur for our simulated judges (for example, the random selection might choose ordinals of "3" for both Irina Slutskaya and Sasha Cohen). Ties are routinely handled in a rank correlation by averaging, so Slutskaya and Cohen would both be given a 3.5 rank and the next best ordinal would get rank 5.

If you'd like to generate your own new "judge" by hand, use 23 digits from a random number table (ignoring zeros) to determine which of the nine judge's ordinal you'll choose for each of the 23 skaters in Table 1, then determine the new judge's ranking by ranking those ordinals and averaging ties. For example, suppose that you enter a random number table and find the first five digits to be: 26885. Then your simulated judge would assign judge 2's ranking to the first skater (4 for Hughes), judge 6's ranking to the second skater (1 for

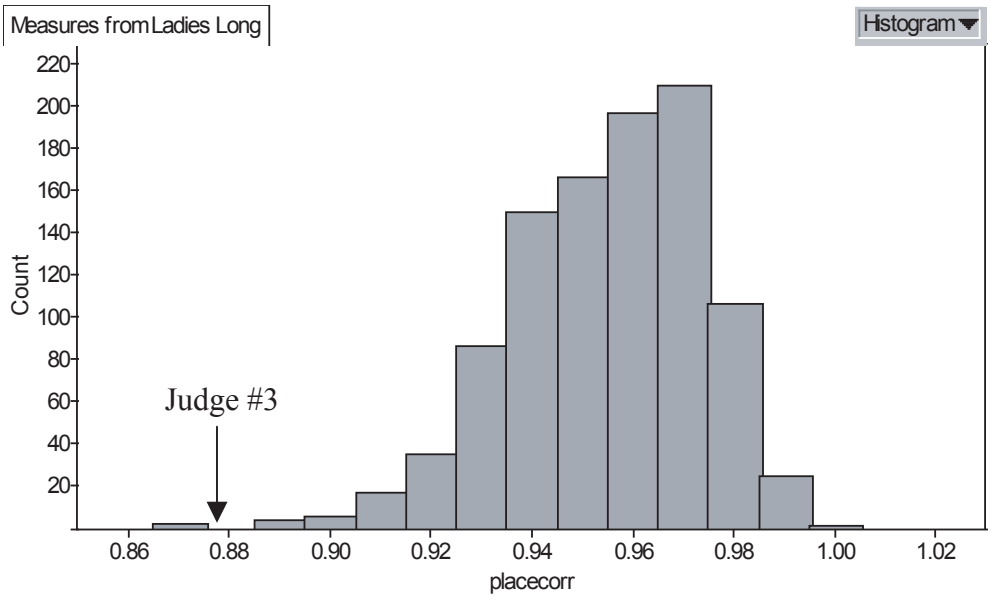

Figure 1. Rank correlations for 1,000 simulated judges of the Ladies Free Skate

Slutskaya), judge 8's ranking to skaters 3 and 4 (2 for Kwan, 4 for Cohen), and judge 5's ranking to the fifth skater (5 for Suguri) and so on. Accounting for the tie (between Hughes and Cohen) and assuming that no later skater is ranked in the top five, the "random" judge's rankings would start with (1) Slutskaya, (2) Kwan, (3.5) Hughes, (3.5) Cohen and (5) Suguri.

Using random selections (with the software package Fathom), we generated 1000 simulated judges, and found the correlation of the rankings of each simulated judge with the final placements. Since the correlation is a measure of the "success" of the judge in matching the actual final placements, this set of simulated correlations provides a yardstick for identifying where a typical judge's correlation should lie for this particular event, considering random variation. A histogram of these bootstrap correlations is shown in Figure 1.

We can see that Judge #3's correlation of 0.876 falls toward the extreme lower end of the distribution. The great majority (998 out of 1000 to be exact) of the generated judges have correlations higher than that of Judge #3. If Judge #3 really was judging in the same way the other judges were, it would be very rare for her to be that far out in the distribution. The approximate p-value from the bootstrap distribution is the proportion of simulated judges with a correlation as low or lower than the judge in question. Thus, the p-value for Judge #3 is about 0.002 and we can conclude that the correlation of Judge #3 is significantly low and, therefore, Judge #3's rankings are not consistent with the other judges. This could indicate a bias, poor quality judging, a mistake by the judge, or most likely just a difference in opinion. But, whatever the reason, Judge #3's assessment of the Ladies Free Skate event at the 2002 Winter Olympics was significantly different from the other 8 judges of that event.

The Notorious French Judge

What about the famous case of the French judge of the pairs long program at the 2002 Winter Olympics? She was accused of being biased to favor the Russian pair over the Canadians. But did she really judge the event significantly differently than the other judges? Below are the correlations for the 9 judges of the pair long program (Table 3).

The first thing you should notice about these correlations is that they are extremely high! Every one of them is above 0.98, which means that the judges agreed very much on the ranking of the skaters in that event. These strong correlations could be due to very accurate judging by the panel, clear differences between the performances of the skaters, or a tendency to pre-judge a competition and placed skaters according to their past reputations. Where was the notorious French judge among these correlations? Surprisingly, she was Judge #4, who had the highest correlation of the whole panel.

Table 3: Rank correlations for judges of the Pairs Long Program

J1	J2	J3	J4	J5	J6	J7	J8	J9
.994	.994	.988	.998	.997	.986	.992	.994	.983

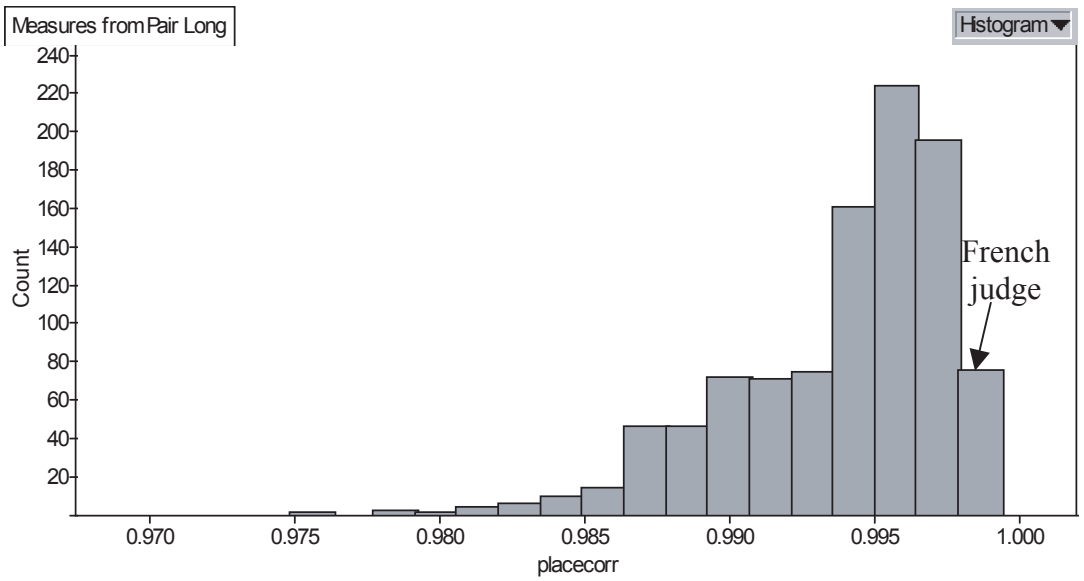

Figure 2. Rank correlations for 1000 simulated judges of the Pairs Long Program

In fact, when we did the bootstrap (see Figure 2), she was significantly *more* in agreement with the final placements than the other judges. Her rankings differed from the final placements by just a single permutation of the 8th and 9th place positions. By our definition of "successful" she judged the event extremely accurately, yet she was accused of bias. Why? Her rankings of the 1st and 2nd place skaters were the same as the original final placements, but were critical for determining those placements since the other 8 judges had split evenly, giving four firsts and four seconds each to the Russian and Canadian pairs. Many observers of the competition believed that the Canadian pair had skated a superior program, so the judging was scrutinized with extra attention. The controversy erupted with allegations that the French judge had been pressured or agreed to a deal to favor the Russian pair over the Canadians before the competition even started. In light of these allegations, the French judge's rankings were disregarded, producing an exact tie for the top spot and duplicate gold medals were awarded. While the distinction between first and second in a close competition is very important to the skaters, their fans and countries, a single permutation may have relatively little affect on the rank correlations between an individual judge and the final placements. So, our bootstrap procedure would not detect that sort of bias in judging.

Conclusion

The bootstrap technique provides a means to assess when the correlation between an individual judge's rankings of the skaters in a competition and the final placements of those skaters by the entire panel of judges is unusually low. While we have applied these ideas to two events from the 2002 Winter Olympics, they could also be applied to other figure skating competitions at various levels or to other "judged" events such as gymnastics, diving, or freestyle skiing. Interesting avenues for future work would be to try to characterize the distribution of rank correlations among judges for different events, levels of competition, numbers of judges, or types of sports. Is the skewed shape of the bootstrap distribution that we see in our two examples typical of most cases? Do pairs skating competitions tend to produce higher correlations than individual events? Can we follow the same judge over several competitions to determine a consistent pattern of disagreement? These methods can help determine whether a judge is really inconsistent with the rest of the judges or just exhibiting the sort of random variation in rankings that one would naturally expect for that particular competition.

References

Bassett, G.W. and Persky, J. (1994), "Rating Skating," *Journal of the American Statistical Association*, 89, 1075–1079.

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Russell, E. (1997), "Choosing a Rating System," *STATS: The Magazine for Students of Statistics*, 19, 13–17.

Web Resource

Judging results for figure skating at the 2002 Winter Olympic Games can be found at the United States Figure Skating Associations website *http://www.usfsa.org/olys02/results/*