Confounding

A Case Study in developing and Building Data Wisdom

Background: Mexican Diver **Fernando Platas** lost the 2000 Olympic gold medal in the 3-meter springboard diving competition by an extremely narrow margin to **Ni Xiong** of China.

https://youtu.be/KjBsU7LIJ74?t=10s

Did Xiong's high scores from the Chinese judge, Facheng Wang, cost Platas the gold medal?

Scores for 3 of Xiong's Dives

NZL	CHN	GER	NOR	FRA	USA	PUR	Trimmed Mean
8.0	8.5	7.5	7.5	8.0	7.5	8.0	7.8
8.0	8.0	7.0	8.5	8.0	7.5	7.0	7.7
8.5	8.5	8.0	9.0	8.0	8.5	8.5	8.4

In the 2000 Olympic 3-meter springboard and 10-meter platform events for men and women, there were:

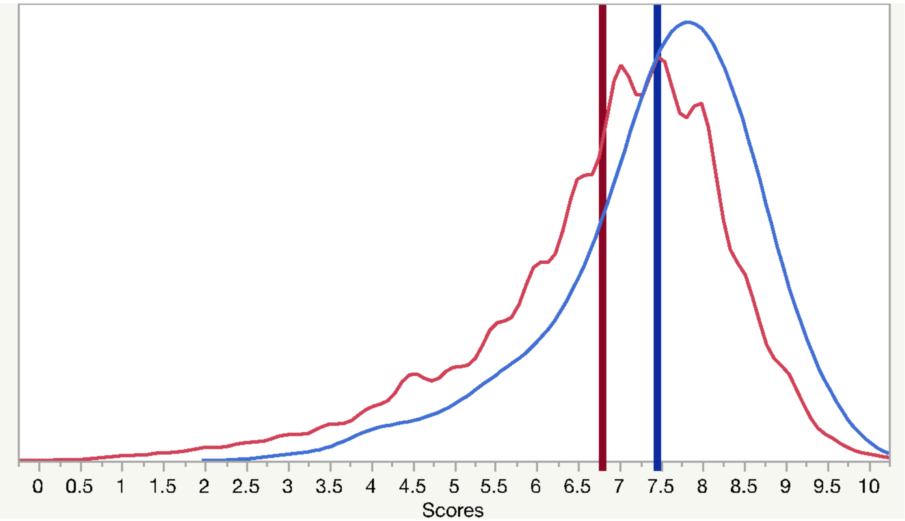
- 156 divers
- 25 judges
- 1541 dives
- Scores can (and did) range from 0 to 10 in steps of 0.5,
- average score = 6.83 (median of 7)
- SD= 1.47 (across all dives)
- 314 dives were scored by a judge whose nationality matched that of the particular diver.
- SD= .41 (within a single dive)

Are judges biased in favor of divers from their own nationality? Can we use the data to measure bias? To answer this question, we need to define “bias”

Bias

"A biased judge is one who tends to award higher scores to her own countrymen than to divers from other countries."

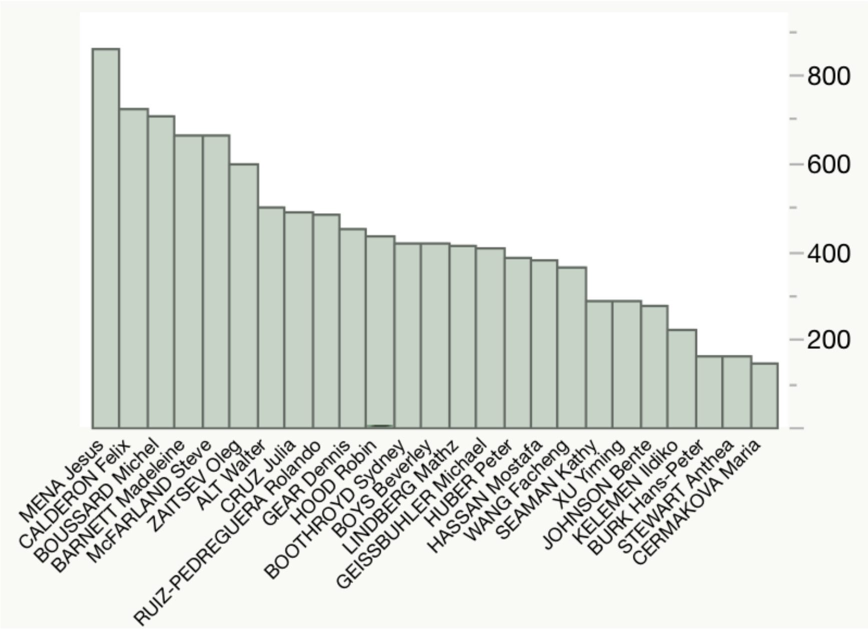
Are Judges biases when their Nationality "Matches" the diver's Nationality?



Blue Curve is histogram of Matches Dives. Red curve is histogram of non-matched dives
The lines are the means for each group

Difference in Means:

Estimate of Nationality Bias = 7.46 – 6.81 = .65 points



Number of scored dives for each judge

- Some divers are more likely to have their dive scored by a judge whose nationality matches their own and others are less likely.
- Indeed, there are many divers that are never scored, not even once, by a judge whose nationality matches their own.
- Possible confounding: divers that are judged less frequently (or not at all) by judges who match their nationalities, are much weaker divers on average.

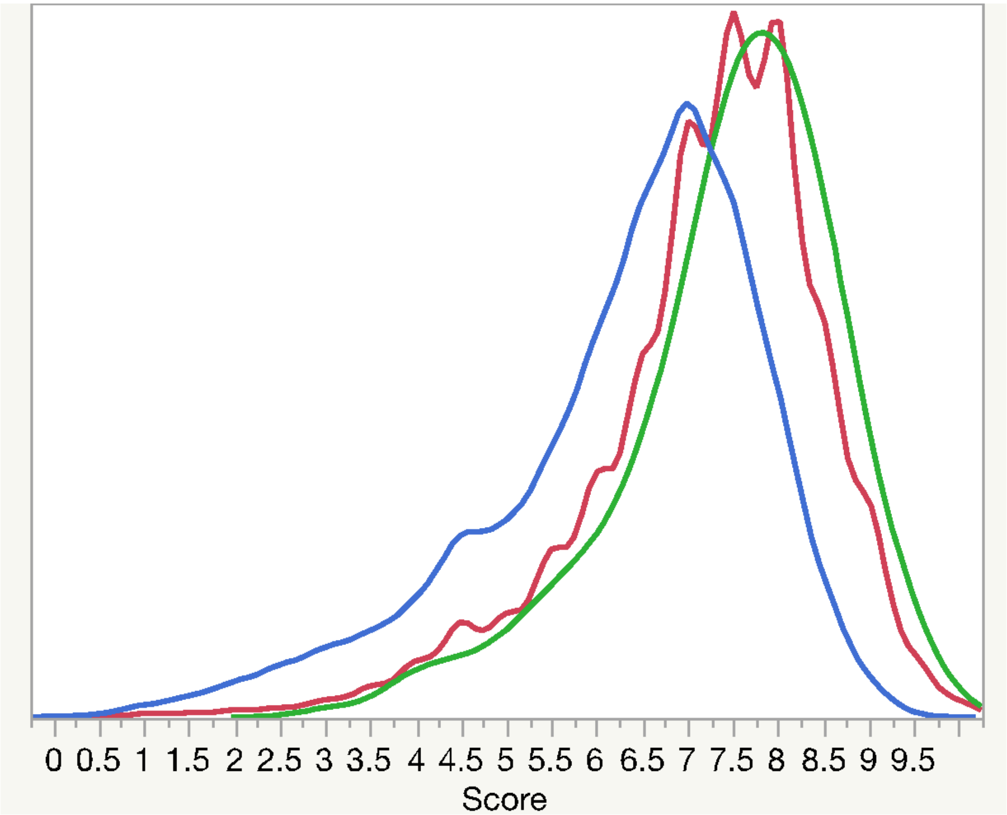
To check if this is so, divide all the scores into three mutually exclusive groups:

1. All Scores for dives where there **is a nationality match** between judge and diver.

2. Scores where diver and judge **do not match nationality**, where the diver has *never been scored* by a judge whose nationality matches their own.

3. Scores where diver and judge **do not match nationality**, where the diver has been scored by a judge whose nationality matches in the entire competition (at least once.)

Then we create a histogram of the scores for each group and compare them by graphing them jointly:



Histograms for the three groups

Group 1 (never matched divers- Blue)
Group 2 (Red, not-matched dives, made by divers that sometimes match)
Group 3 (Green, matched dives)

Group	Number of Dives	Mean	Std Dev
Never-Matched Divers	4333	6.285	1.53
Not Matched Dives- but sometimes matched divers	6140	7.186	1.31
Matched Dives	314	7.461	1.21

The mean for the blue group (Group 1: divers whose nationality never matches that of a judge) is a lowly 6.285, which is way less than the overall average of 6.83.
In contrast, scores for dives that are not matched by divers that do have matches, the mean (Group 2: red histogram) is 7.18.

This is nearly a one-point difference in average score. Thus, the group of divers that match a judge in nationality (at least once) are much better divers!

Conclusion: As a group their appears to be nationality bias.

But there is still possible confounding. The divers who are matched more often may be better divers.

So now let’s look at an individual Judge: Chinese Judge Wang.

- Wang average score for Chinese divers = 8.45
- Wang average score for non-Chinese divers = 6.97

Wang *did* give substantially higher scores to Chinese divers than to non- Chinese divers.

Conclusion: Wang is biased towards Chinese divers!

WRONG!

A possible confounding factor:

- The average score by Chinese divers from the whole competition, including all judges, was 8.16.
- The average score is 6.72 for all other divers.

This is about 1.5 points difference, which matches the Wang “bias.”

We have failed to control for the fact that Chinese divers are better on average, and so when Judge Wang “matches” the divers are Chinese and get better scores, on average.

Let’s Revised Definition of Bias

“a judge who tends to award scores higher than the panel average to his own countrymen.”

So, on each dive, measure the “**discrepancy**” for a given judge by comparing that Judge’s score to the to the average panel score for the dive.

Negative discrepancy = judge’s score is BELOW average (compared to panel average)

Positive discrepancy = judge’s score is ABOVE average (compared to panel average)

Chinese Diver Li Xiong's scores and discrepancies for an example dive

	NZL	CHN	GER	NOR	FRA	USA	PUR	AVERAGE
Score	8.0	8.5	7.5	7.5	8.0	7.5	8.0	7.86
Discrepancy	+0.14	+0.64	-0.36	-0.36	+0.14	-0.36	+0.14	0

So, on this dive, the Chinese judge has the largest discrepancy.

Overall his average discrepancy for Chinese divers is +.17.

So, he must be guilty of bias according to even this revised definition!

WRONG AGAIN!

If you agree with this conclusion, you are again guilty of the FALSE CAUSE fallacy.

You assume that since Judge Wang's discrepancy for Chinese divers is greater than 0, then he must be biased.

What if Judge Wang just tends to give out higher score to ALL divers?

A REVISED REVISION of the DEFINITION OF BIAS:

“a biased judge is one who awards higher scores than other judges to his own countrymen, *but fails to award higher scores to non-countrymen*.”

So, what is Judge Wang’s average discrepancy for all divers? Also +.17
Clearly unbiased, as the difference between discrepancies is 0.

<u>In contrast-</u>

American Judge McFarland discrepancy for American divers = +.234

American Judge McFarland discrepancy for non-American divers = +.012. This is a difference of +.22 points for American divers.

That’s not nothing: how about the Russian Judge? He was +.27 for Russians and -.02 for everyone else (DoD = .29) which is even more biased that McFarland. The German and Austrian Judge were even more biased.

MORAL OF THE STORY: CETERIS PARIBUS ALL OTHER THINGS BEING EQUAL

When we use the method of comparisons, it important to add the caveat "all other things being equal."

Consider the original definition of bias with this added caveat

"A biased judge is one who tends to award higher scores to her own countrymen than to divers from other countries."

Assuming ALL OTHER THINGS ARE EQUAL

The reason why this definition failed, is that other things were decidedly not equal. Chinese divers are on average much better divers and the Chinese judge was a generous judge that tended to give higher scores.

Can you think of another explanation for these biases other than favoritism?

Swiss judge Michael Geissguhler- difference of discrepancies = +.68 in favor of his countrymen.

But Switzerland is a small country, so Judge Geissguhler's assessment of bias is based on just 3 Swiss dives.

The critical statistical question here is:

"Could the observed effect be due to chance?"

Is this differentce important or significant enough to make a difference?

Recall:
Judge McFarland average discrepancy: .012
Judge McFarland average American discrepancy: .234
DoD: .234-.012= .22
The sample size = 42 American divers

Championships get decided by .22 point discrepancies. So yes, it is "significant".

Review:

If you first compare matched scores to unmatched scores you observe a big difference. This is because matched scores more often involve better divers because judges are more often from bigger countries with better divers.

As a first step, control for the quality of the divers.

This isn’t enough. You also have to control for the judge.

After controlling for the quality of the divers and the judges you can compare the scores for nationality bias.