Significance and p-values

How can we study the effect of Chance Variation?

We know that if the judges are biased in favor of their countrymen then their average discrepancy will be large. But judges are human and their scores will vary.. just because.

This unexplainable (irreducible) variation is called "Chance Variation".

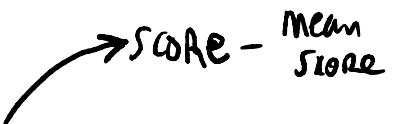*difference of discrepancies between nationality matches and non matches*

PROBLEM

How large does the $\boxed{DoD}$ have to be in order to be convinced that the DoD is **not** caused by chance variation?

Is it signal or noise?

This is the FUNDAMENTAL question in modern science.

A Study of Variation in Scores

Diver: Jesus-Iory Aballi, Cuba.
Event: 10 Meter.
Round: Prelim
Mean: 7.429.

→ score − mean score

Judge	Score	Deviation
NZL	7	-0.429
GER	7.5	0.0714
SWE	7.5	0.0714
USA	8	0.5714
MEX	7.5	0.0714
ZIM	7	-0.429
ESP	7.5	0.0714

Repeat for all dives.
The root mean square
of these deviations is .409. The
distribution has an almost perfect Bell
shape.
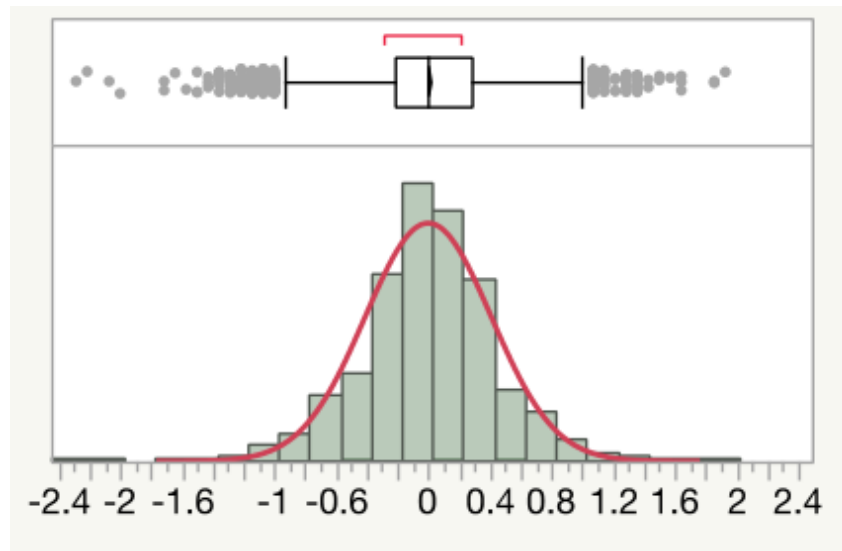
McFarland gave the American divers an average score of 7.79 and the non-Americans an average score of 6.70. This is not evidence of bias, because the Americans are very good divers.

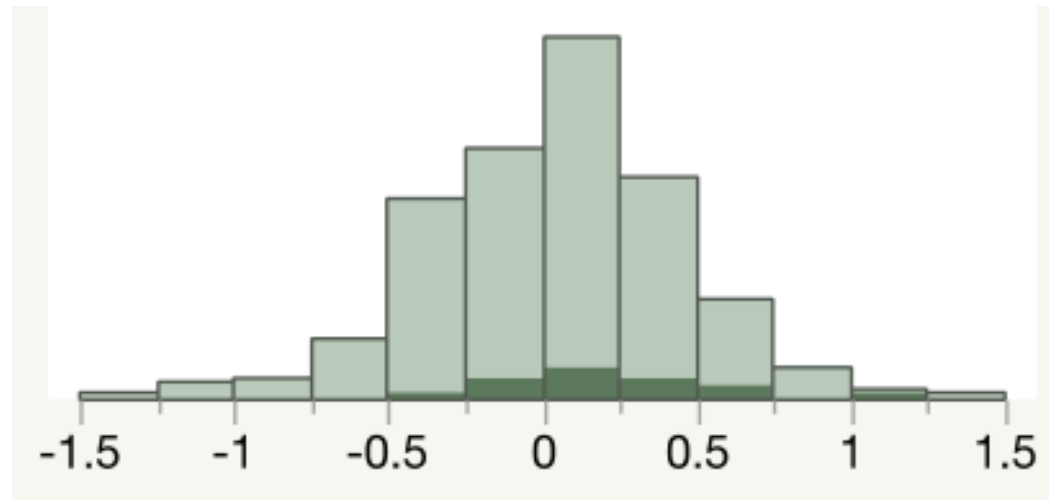American divers are much better as the chart below clearly shows:
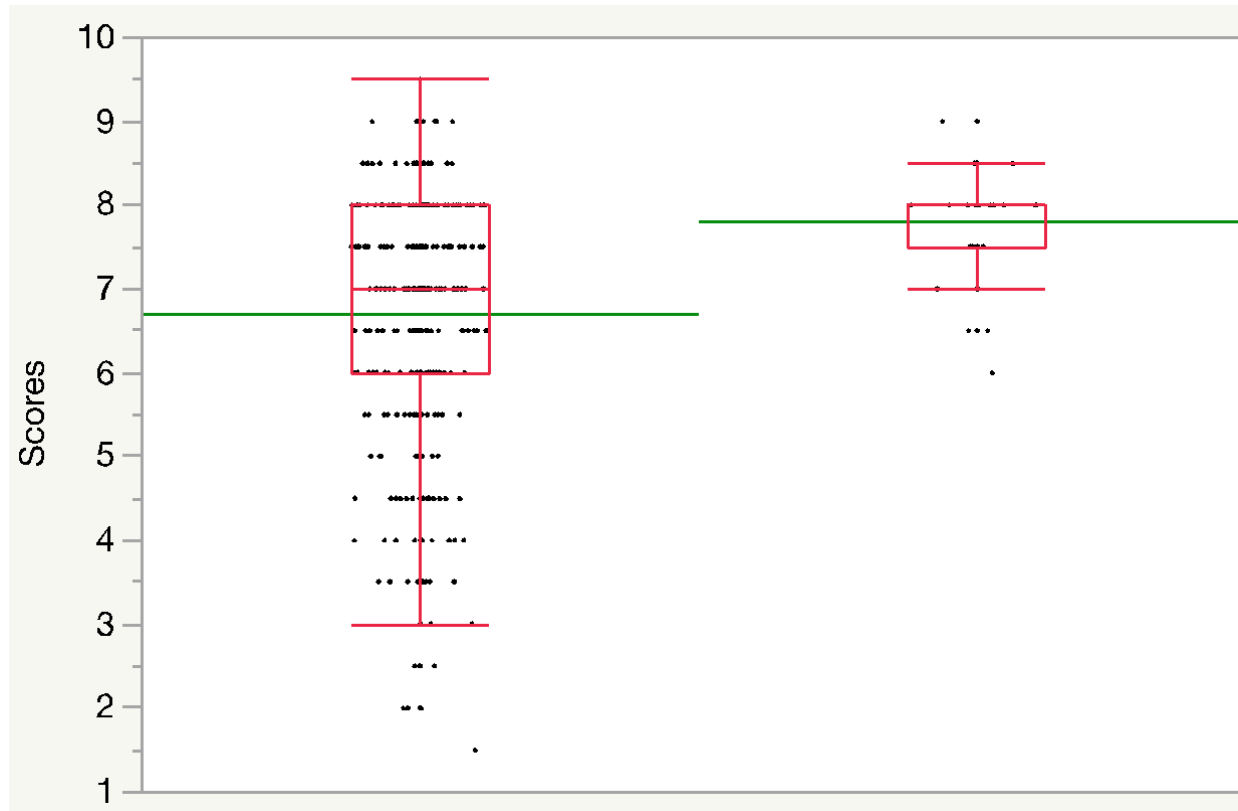
Right: Box Plot of American Diver's scores (mean = 7.79)
Left: Box Plot of Non-American Diver's scores (mean = 6.70)

Judge McFarland judged 657 dives

Distribution of discrepancies

- The mean discrepancy is .02.

- The mean discrepancy for the 42 dives by Americans was .20. The mean discrepancy for everyone else is .01. The DoD is .20-.01 = .19

- You can see that visually on the histogram: the 42 Americans are shaded.
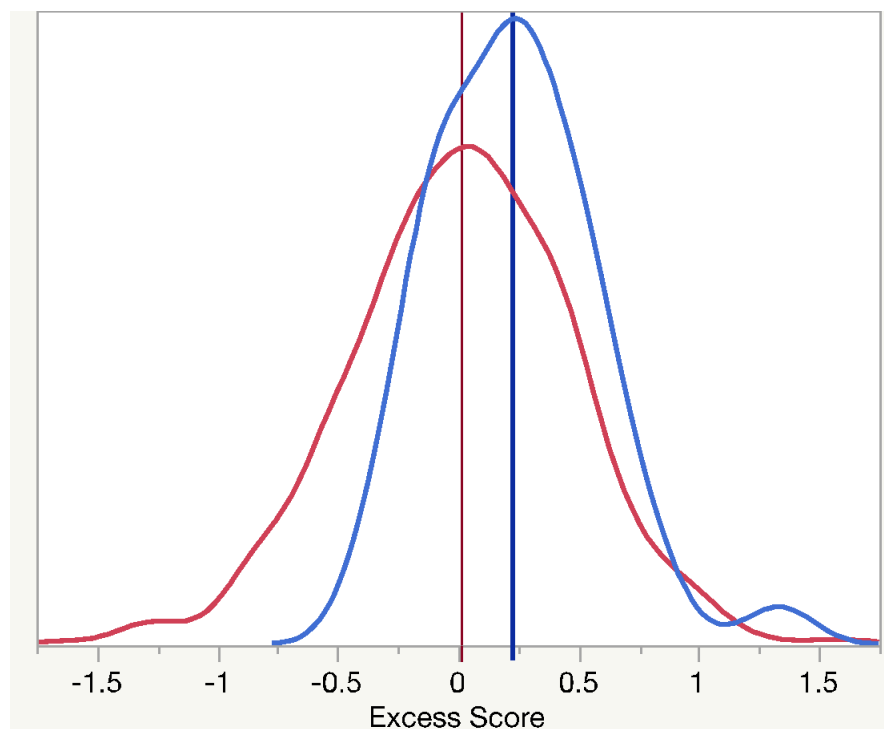
Judge McFarland excess scores for American Divers (Blue) compared to Non-American Divers (Red)

But is this DoD due to noise / random chance?

If American judge McFarland had been unbiased and his discrepancies had been randomly distributed across all divers, *how likely* is it that his DoD (difference of discrepancies) would have been +0.19 or higher?

To answer this question, we do a **Permutation Study**.

Take all of McFarland's dives and *randomly* permute the nationality labels of the divers so that the 42 matches are now on a completely different set of divers (who aren't really American, they are just labeled as such).

If he were unbiased, then permuting the nationality labels shouldn't lead to a significantly different DoD, i.e. the DoD = +0.19 should be plausible.

Example:

randomly switch the order of the Country Column → (annotation pointing to "Permuted Country" column)

Event	Round	Diver	Country	Permuted Country	Rank	DiveNo	Difficulty	JScore	Judge	JCountry	Nationality Match of Judge and Diver	Permutation Match of Judge and Driver
M10mPF	Prelim	TIAN Liang	CHN	USA	1	2	3	8	McFARLAND …	USA	No	Yes
M10mPF	Prelim	MEYER Heiko	GER	CHN	12	2	3.6	3.5	McFARLAND …	USA	No	No
W10mPF	Prelim	SANTOS Leire	ESP	ROM	27	2	3	5	McFARLAND …	USA	No	No
M10mPF	Prelim	WATERFIELD …	GBR	AUT	33	2	3.8	3.5	McFARLAND …	USA	No	No
W3mSB	Prelim	FU Mingxia	CHN	RUS	1	4	3	6.5	McFARLAND …	USA	No	No
W3mSB	Semi	HARTLEY Blythe	CAN	UKR	7	1	1.6	6.5	McFARLAND …	USA	No	No
W10mPF	Prelim	KONSTANTAT…	GRE	KAZ	38	2	2.8	4	McFARLAND …	USA	No	No
W10mPF	Prelim	SAEZ-de-…	ESP	USA	15	3	2.9	5	McFARLAND …	USA	No	Yes
W10mPF	Prelim	KONSTANTAT…	GRE	BRA	38	1	3.1	2	McFARLAND …	USA	No	No
M3mSB	Prelim	FRECE Richard	AUT	INA	31	1	3.1	6.5	McFARLAND …	USA	No	No
W10mPF	Semi	SAEZ-de-…	ESP	BLR	14	2	2	6	McFARLAND …	USA	No	No
M10mPF	Prelim	SKRYPNIK …	UKR	MAS	23	1	2.7	6	McFARLAND …	USA	No	No
W3mSB	Semi	GUO Jingjing	CHN	AUS	2	3	1.9	8	McFARLAND …	USA	No	No
M3mSB	Prelim	ALVAREZ Rafael	ESP	USA	16	2	3	6.5	McFARLAND …	USA	No	Yes
M3mSB	Prelim	BIMIS Thomas	GRE	KAZ	32	2	3.1	5.5	McFARLAND …	USA	No	No
M3mSB	Prelim	DOBROSKOK …	RUS	PUR	17	3	3.5	5	McFARLAND …	USA	No	No
M3mSB	Prelim	SALAZAR …	CUB	UKR	15	3	3	6.5	McFARLAND …	USA	No	No
M3mSB	Prelim	URAN Juan-…	COL	CHN	41	2	3.1	4.5	McFARLAND …	USA	No	No
W10mPF	Prelim	ALCALA …	MEX	USA	30	1	2.8	6.5	McFARLAND …	USA	No	Yes
M10mPF	Prelim	AVTANDILYAN …	ARM	UKR	38	2	3	4	McFARLAND …	USA	No	No
M10mPF	Prelim	HAJNAL Andras	HUN	BLR	34	1	3	3.5	McFARLAND …	USA	No	No
W10mPF	Prelim	KONSTANTAT…	GRE	CUB	38	3	3	3.5	McFARLAND …	USA	No	No
W10mPF	Prelim	REIFF Marion	AUT	GBR	37	3	3.1	3.5	McFARLAND …	USA	No	No
M3mSB	Prelim	RODRIGUEZ …	MEX	MEX	29	1	3.1	6.5	McFARLAND …	USA	No	No

The key idea here is that McFarland's **overall average discrepancy will be the same**, but the difference in discrepancies (DoD) **will be different** because the dives will be divided into different groups under permutation.

The DoD is computed by comparing two groups:

1. the discrepancies for the randomly selected "American" divers
2. the discrepancies of the randomly selected "non-American" divers.

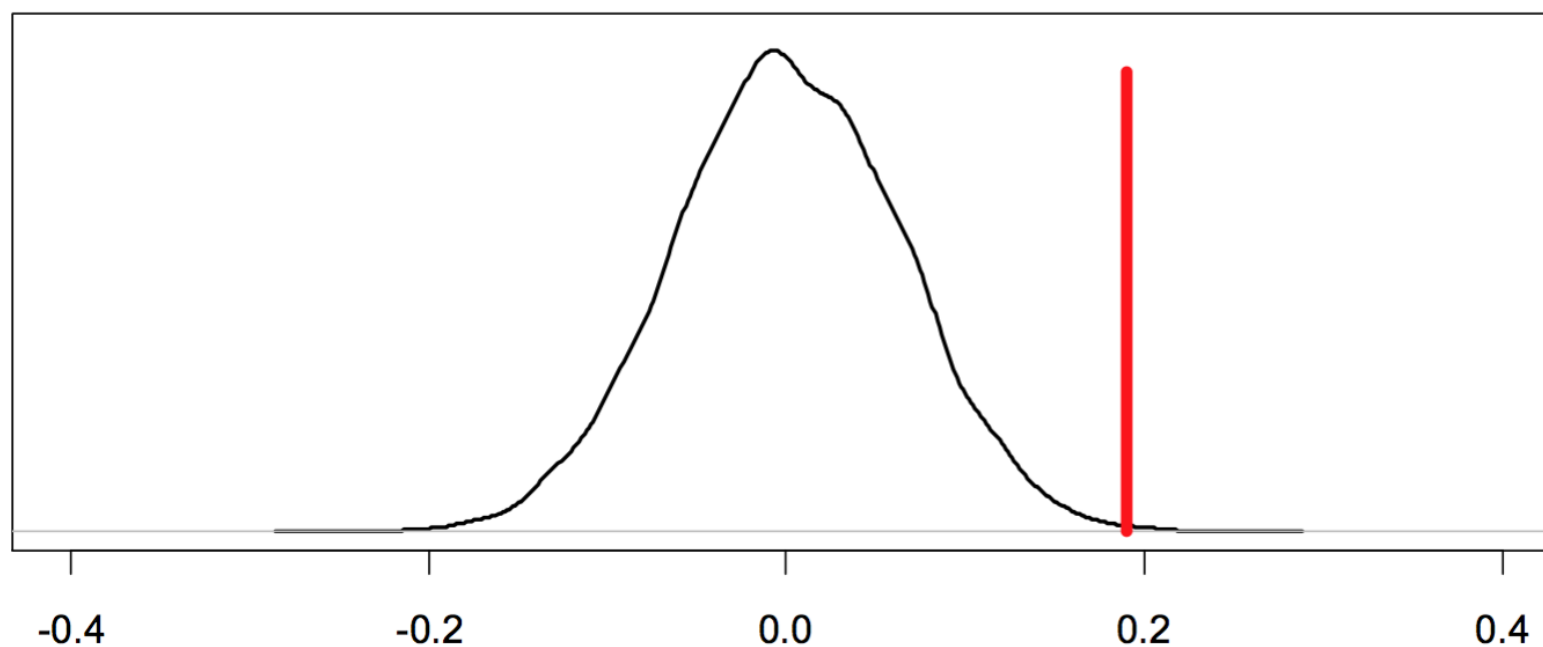**Permutation Distribution:**

Now divide the 657 scored dives into 2 groups **at random**; one with 615 and the other with 42.

For each random selected you calculate a DoD.

Repeat as many times as you want.

Then see where the actual assignment by nationality compares to the randomly created divisions. Make the histogram of DoD created from the permutations and then see where the actual DoD lands….

Have we found a cheater?

This is the distribution of the DoD computed for every permuted assignment. The red-line is Judge McFarland's actual DoD.

The proportion of the area to the right of the red line is the proportion of times the permuted assignments produced a DoD as large as McFarland's actual value of .22

It's not very often: 1 out of 1000.

This is called a p-value.

This p-value means:

assuming that the judge was unbiased, the probability of observing a DoD of $\geq .22$ is $\frac{1}{1000}$.

This is very small!

Tests of Significance

- The scientific method begins with a **research hypothesis**.

- This is what he or she wishes to establish.

- It is natural to try to collect evidence that confirms the research hypothesis.

Our Research Hypothesis:

- Judge McFarland is biased towards Americans.

- He gives even higher scores to American divers than he gives to non-Americans even after controlling or adjusting for the quality of the dives and his natural tendencies to be a slightly easier grader than other judges.

Confirmatory Evidence:

His 42 scores for American divers, average .22 more points than the 672 scores for non-American dives.

The idea that will change your life

The Scientific Method

The scientific method rejects confirmatory reasoning. It is too subject to what is now called "*confirmation bias*"- the tendency to cherry pick evidence that supports our ideas while ignoring or explaining away evidence that contradicts.

The **scientific method** reverses the approach:

- Begins with a Null Hypothesis which is the opposite of the research hypothesis.

- **Goal:** Assemble evidence that cannot possibly have happened if the Null were true.

- A Hypothesis is **testable** if evidence can be brought that disprove or *falsify* the opposite.

- Only hypotheses whose nulls (opposites) are ***falsifiable*** are scientific.

The research hypothesis is that the American Judge exhibits nationality bias.

The Null Hypothesis is that he not biased.

The statistical representations of these hypotheses:

- **Research Hypothesis:** McFarland's bias of .22 points on average is real and not caused by chance variation. He is biased towards Americans.

- **Null Hypothesis:** McFarland's bias of .22 points is just variability at work and not bias.

Process:

Measure the evidence from the perspective of the null hypothesis. Assume it is true and then study the data. If the data is implausible or highly contradictory of the null then reject it, thereby proving your point.

To prove there is nationality bias show that the scores for matched divers are sufficiently high that that chance variation cannot explain it.

This is called **Statistical Significance.**

The p-value: how likely is it to get a result as extreme as the one that was observed, under the Null

The p-value measures the strength of the evidence against the null.

Statistical Significance is usually determined by p-values.
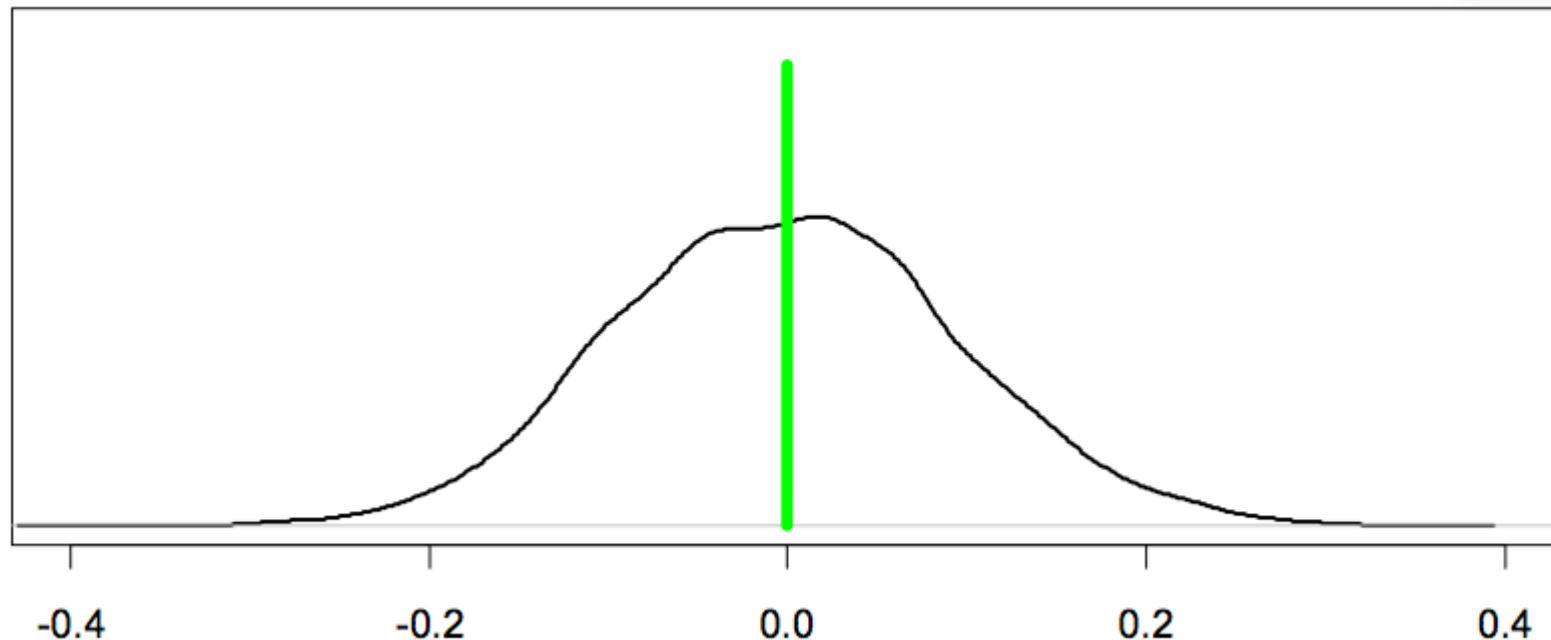
Statistical Question:

Is **chance variation** a reasonable explanation for any observed differences in data or must there be another explanation?

The permutation test allowed us to understand the reasonable effect of chance variation on the results.

It controls for chance.

Judge Wang's DoD for his 22 matched dives

What does this mean?

The DoD for the 22 Chinese divers, is exactly equal to the DoD for a random choice of 22 dives, on average.

Judge Wang tends to give slightly higher scores than other judges, but not any higher for Chinese divers. The p-value?

Judge	Number of Matched Dives	Average Discrepancy for Matched Dives	Number of Non-Matched Dives	Average Discrepancy for Non- Matched Dives	Difference of Discrepancies (DoD)	Permutation p-value
Alt, Walter (GER)	25	+0.31	473	-0.08	0.39	**<0.0001**
Barnett, Madeleine (AUS)	38	+0.18	623	-0.11	0.29	**<0.0001**
Boothroyd, Sydney (GBR)	16	+0.32	395	+0.04	0.28	**0.0042**
Boussard, Michel (FRA)	10	0.00	692	-0.11	0.11	**0.1918**
Boys, Beverley (CAN)	13	+0.27	398	+0.06	0.21	**0.0202**
Burk, Hans-Peter (GER)	10	+0.37	149	-0.09	0.46	**0.004**
Calderon, Felix (PUR)	5	+0.23	712	-0.07	0.30	**0.0633**
Cruz, Julia (ESP)	11	+0.29	475	-0.02	0.30	**0.003**
Geissguhler, Michael (SUI)	3	+0.67	398	-0.01	0.68	**0.0015**
Huber, Peter (AUT)	8	+0.31	374	0.00	0.31	**0.0162**
McFarland, Steve (USA)	42	+0.20	615	+0.01	0.19	**0.0013**
Mena, Jesus (MEX)	28	+0.25	828	-0.06	0.30	**<0.0001**
Ruiz-Pedreguera, Rolando (CUB)	11	+0.29	470	+0.01	0.28	**0.0033**
Seamen, Kathy (CAN)	16	+0.15	265	-0.00	0.16	**0.0730**

Notice that the p-value for the Swiss Judge is very small:

Even though he only has 3 matched dives in total (out or 401 dives in total)
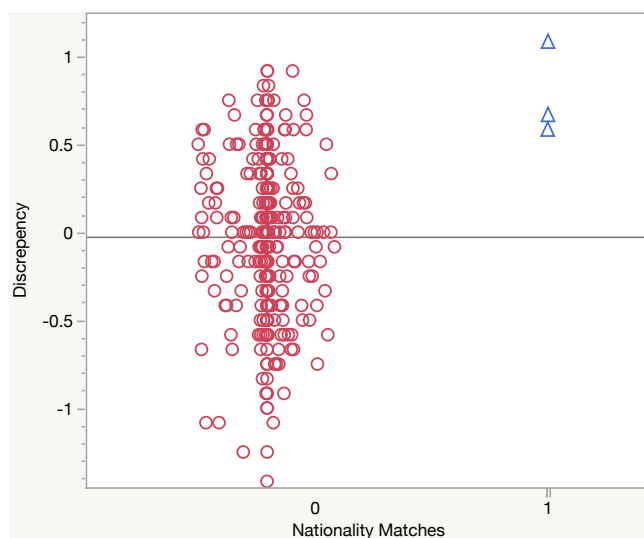
Sometimes even a small number of samples can be very informative.