What's the question?

David J. Hand
Imperial College, London

The Statistics Wars and Their Casualties Online Workshop
1 December 2022

- Science is an iterative process, cycling through theory generation, data collection, and comparison of theory with data

- Different scientific schools favour different theories, until accumulating evidence/data renders one or more untenable

- Tensions can arise between proponents of different theories

- The discipline of statistics is susceptible to similar tensions

- Not between theories of “what is”, but between strategies for shedding light on the real world from limited empirical data

The classic case of a tension in statistics is the debate between different schools of inference

This has often been heated as is illustrated by the recent Wasserstein et al (2019) call for a ban on the use of certain terms, saying:

'it is time to stop using the term "statistically significant“ …
'a declaration of "statistical significance" has today become meaningless’

In response to Wasserstein and Lazar's call to stop using the term statistically significant, I wrote:

> *"…banning any potentially useful method … is short-sighted, unscientific and Procrustean, …[it] damages the capability of science to advance, and, worse still, feeds into public mistrust of the discipline of statistics"*
>
> Hand D.J. (2021) Trustworthiness of statistical inference. *Journal of the Royal Statistical Society, Series A,* **185**, 329-347

"…the fact that a tool can be misunderstood and misused is not a sufficient justification for discarding that tool. Rather, methods for calling out and avoiding such mistakes are required…"

Mayo D.G. and Hand D.J. (2022) Statistical significance and its critics: practicing damaging science, or damaging scientific practice. *Synthese*. https://link.springer.com/content/pdf/10.1007/s11229-022-03692-0.pdf

Wasserstein *et al* are concerned with statistical ***inference***

But I contend that this is merely an illustration of a bigger problem:

- The failure to recognise that different tools are suited to answer different questions
- The failure to articulate the scientific questions sufficiently clearly to permit informed choice of tool
- Poor mapping of the scientific question to the statistical question
- Misuse of poorly understood statistical tools

I will illustrate this by looking at some familiar statistical tools

There are two steps in formulating testable hypotheses

Step 1:

Formulate the ***scientific*** question

Necessarily involves simplifications

Step 2:

Translate the scientific question into a statistical question

Risk of mismatch, oversimplification, distortion

Both steps need to be precise for accurate conclusions to be drawn

From the earlier talks on 22nd and 23rd September:

Stephen Senn: "*In my opinion if there are problems, they have a lot less to do with choice of inferential system or statistic and a lot more to do with appropriate calculation*"
https://phil-stat-wars.com/2022/10/10/the-statistics-wars-and-their-casualties-videos-slides-from-sessions-1-2/

Christian Hennig: "*mu1, mu2 are thought constructs defined within the normal model…The real hypothesis of interest is about whether one of the teachers gives systematically higher marks. …*" https://phil-stat-wars.com/2022/10/10/the-statistics-wars-and-their-casualties-videos-slides-from-sessions-1-2/

"*Note that the Central Limit Theorem is about estimating E(X), which may not be in line with the interpretative hypothesis, so what the t-test does based on CLT may be misleading even though CLT applies*"
https://phil-stat-wars.com/2022/10/10/the-statistics-wars-and-their-casualties-videos-slides-from-sessions-1-2/

Example 1:

We all know that the mean is sensitive to values of a handful of extreme points

- Can make the mean as large as you like by increasing the largest value, while the median stays unchanged

So you sometimes see statements like:

"The choice between mean and median should depend on how the data are distributed …

The mean is a better measure of "location" than the median for symmetric distributions, but otherwise the median should be used"

But this is ***wrong***

It makes no reference to the question being asked

That is, to what you want to know

Suppose I randomly choose the remuneration of a new recruit from a very skewed distribution of salaries

What interests ***me*** is the ***mean*** of the distribution:

my total wage bill is the product of the mean and the number of employees

What interests ***a potential new recruit*** is the ***median***:

to the potential new recruit the mean is of no interest, since she's almost certain to receive substantially less than the mean

The choice between mean and median depends on what you want to know

1994 Baseball players strike in US

Median $ 337,500 vs Mean $ 1,049,589

Arithmetic means can be misused -> ban arithmetic means?

Example 2:

- I want to buy a new car
- I'm weighing up the merits of two alternative types, A and B
- In terms of which of type A or B is more cost (fuel) efficient

To avoid confusion about the units I will work in kilometres and Litres

Miles -> kilometres (km)

Gallons -> litres (L)

To test fuel efficiency I take samples of type A and B cars

For simplicity of exposition:

- ***samples of size 3***
- ***and unrealistic values so I can illustrate the ideas easily***

I then collect the data

	Type A			Type B		
	Car 1	Car 2	Car 3	Car 1	Car 2	Car 3
B(km/L)	1	1	13	3	3	3

	Type A				Type B			
	Car 1	Car 2	Car 3	Av.	Car 1	Car 2	Car 3	Av.
B(km/L)	1	1	13	5	3	3	3	3

→ ***type A is better – more km/L***

I'm on the verge of buying a Type A car, when I mention it to my neighbour, who happens to be French

She asks to see my data and she says I have made a mistake

She says: ***car fuel efficiency should be calculated in L/km, not km/L***

	Type A				Type B			
	Car 1	Car 2	Car 3	Av.	Car 1	Car 2	Car 3	Av.
F(L/km)	1	1	1/13	0.692	1/3	1/3	1/3	0.333

→ ***type B is better – fewer L/km***

Summary:

	Type A				Type B			
	Car 1	Car 2	Car 3	Av.	Car 1	Car 2	Car 3	Av.
B(km/L)	1	1	13	5	3	3	3	3

→ ***type A is better – more km/L***

F(L/km)	1	1	1/13	0.692	1/3	1/3	1/3	0.333

→ ***type B is better – fewer L/km***

Convert to km/L for comparison

Summary:

	Type A				Type B			
	Car 1	Car 2	Car 3	Av.	Car 1	Car 2	Car 3	Av.
B(km/L)	1	1	13	5	3	3	3	3

→ ***type A is better – more km/L***

F(L/km)	1	1	1/13	0.692	1/3	1/3	1/3	0.333

→ ***type B is better – fewer L/km***

Convert to km/L for comparison

F(km/L) 1/0.692 = 1.444 1/0.333 = 3

→ ***type B is better – more km/L***

One of the two researchers is drawing the wrong conclusion
But which?

It's obvious ***why*** B and F are drawing different conclusions:
Reciprocal is a nonlinear transformation, so

$$AM\big(f(x)\big) \neq f\big(AM(x)\big)$$

where $f(x) = 1/x$ and AM is the arithmetic mean

(At least) one of the two researchers is ***misusing*** the arithmetic mean

So arithmetic means can be misused -> ban arithmetic means?

And since the harmonic mean is the arithmetic mean on the reciprocal scale
Harmonic means can be misused -> ban harmonic means?

If we really consider the alternatives *km/L* and *L/km* to be ***equally legitimate*** then something is wrong with our analysis

Something in the ***mapping*** from the scientific question to the statistical question is misleading us

If the scientific question says analyses based on x and 1/x should reach the same conclusion
the statistical analysis must be using information in the x and 1/x scales which is irrelevant

The original (scientific) question asked:

"which of type A or B is more cost (fuel) efficient"

"More" is a purely ordinal relationship!

But the arithmetic mean also uses **size**

We should be estimating

Prob(a randomly chosen type A car will be more fuel efficient than a randomly chosen type B car)

A {1, 1, 13} B {3, 3, 3}

Prob(A > B) = 1/3

[Which is equivalent to Prob(1/A < 1/B) = 1/3]

That is, Prob(A > B) < 0.5

So type B is better

Incidentally and parenthetically ….

In general, to answer the "more than" question
we must have invariance to any monotonic transformation g

$$Av(x_1, \ldots, x_n) \geq Av(y_1, \ldots, y_m)$$
$$\Leftrightarrow$$
$$g^{-1}\left[Av\left(g(x_1), \ldots, g(x_n)\right)\right] \geq g^{-1}\left[Av\left(g(y_1), \ldots, g(y_m)\right)\right]$$

The median is invariant to such transformations
And you might have been worried by the skew {1, 1, 13} distribution
(despite what I said earlier)

But Median A > median B $\not\Rightarrow$ Prob(A > B) > 0.5
So invariance to monotonic transformations is a necessary but not sufficient condition to answer our question

In the above, I have supposed I am buying a single car

In that case, wanting to compare the performance of single randomly chosen cars of each type is a sensible question

But suppose I am operating a fleet of taxis

Then I would not be interested in individual cars but in the performance of my fleet as a whole

$$\frac{Total\ distance\ travelled}{Total\ fuel\ used}$$

This is equivalent to

$$\frac{\sum_i u_i/n}{\sum_i v_i/n}$$

where u_i and v_i are the distance travelled and the fuel used for the *i*th car in a given time

This is the ***ratio of the arithmetic means*** of the distance travelled and fuel consumed, not the ***arithmetic mean of the ratios***

Example 3: Price inflation

The Retail Price index, RPI, first calculated in 1947,
uses ***arithmetic means*** to summarise individual price quotes

But in 2030 the UK government will switch to the CPIH
Because many see the arithmetic mean as *inappropriate*
CPIH uses ***geometric means*** (used in the EU and USA)

There are consequences. For example:

- Defined benefit pensions will have annual uprating reduced, losing 4-9% of their lifetime value
- Index-linked gilts, government bonds with payments that change with inflation, will have interest payments reduced by £90-£100 billion

The arithmetic mean has been used in error since 1947 ?

So arithmetic means can be misused -> ban arithmetic means?

Example 4: Which of treatments A or B is better?

This is an informal question: *what do you mean by "better"?*

We need to make the ***scientific question*** more precise
Even before we can try to translate it into a statistical question

Example 4a: Which of treatments A or B is better?

Do we mean an explanatory or pragmatic question?

Explanatory: Which of A or B is better, all other things being equal (lab. conditions, physiological, biological)?

The scientific question, the understanding question

Pragmatic: Which of A or B is better, as used in practice?

The operational question, the use question

e.g. Treatment of cancer by radiotherapy

Wish to compare

A: Radiotherapy, preceded for 30 days by sensitising drugs

B: Radiotherapy alone

Explanatory comparison of A with B:

To make the only difference between A and B whether or not the sensitising drug has been taken

in B we must precede the radiotherapy with 30 days placebo

Pragmatic comparison of A with B:

We would not make people wait 30 days in practice

so in B, we must start radiotherapy immediately

So the different scientific questions have ***different designs***

A: Radiotherapy, preceded for 30 days by sensitising drugs
B: Radiotherapy alone

Explanatory comparison of A with B:

Want to avoid concluding that there is a difference when there is none

Or vice versa

Pragmatic comparison of A with B:

Want to guard against identifying poorer as better

Concluding there is a difference when there is none does not matter

These have ***different error structures***

And so will likely require different sample sizes

A: Radiotherapy, preceded for 30 days by sensitising drugs
B: Radiotherapy alone

Explanatory comparison of A with B:

Analyse only those who stick to protocol

Pragmatic comparison of A with B:

Include side-effects withdrawals as treatment failures

So ***different samples are being analysed***

A: Radiotherapy, preceded for 30 days by sensitising drugs
B: Radiotherapy alone

Explanatory comparison of A with B:

Eliminate superfluous variability

Study a homogeneous sample of people

Pragmatic comparison of A with B:

Subjects should be a sample of those to be encountered in practice – i.e. heterogeneous

Different populations are being studied

A: Radiotherapy alone
B: Radiotherapy, preceded for 30 days by sensitising drugs

So to clarify the scientific question

Which of treatments A or B is better?

We need to know if the study is pragmatic or explanatory

- ***different designs***
- ***different error structures***
- ***different sample sizes***
- ***different samples analysed***
- ***different populations***

Example 4b: Which of treatments A or B is better?

Really mean: *Which of treatments A or B is better [for the next patient]?*

To find out, I will give A and B to patients and study the outcome

Patients differ, so cannot make definitive statement about "the next patient"

So I change my question to

Which of A and B is better on average?

And here I will make the common assumption that by "average" I mean the *Arithmetic Mean* $\mathcal{AM}$

But what if I cannot give *both* A and B to each patient

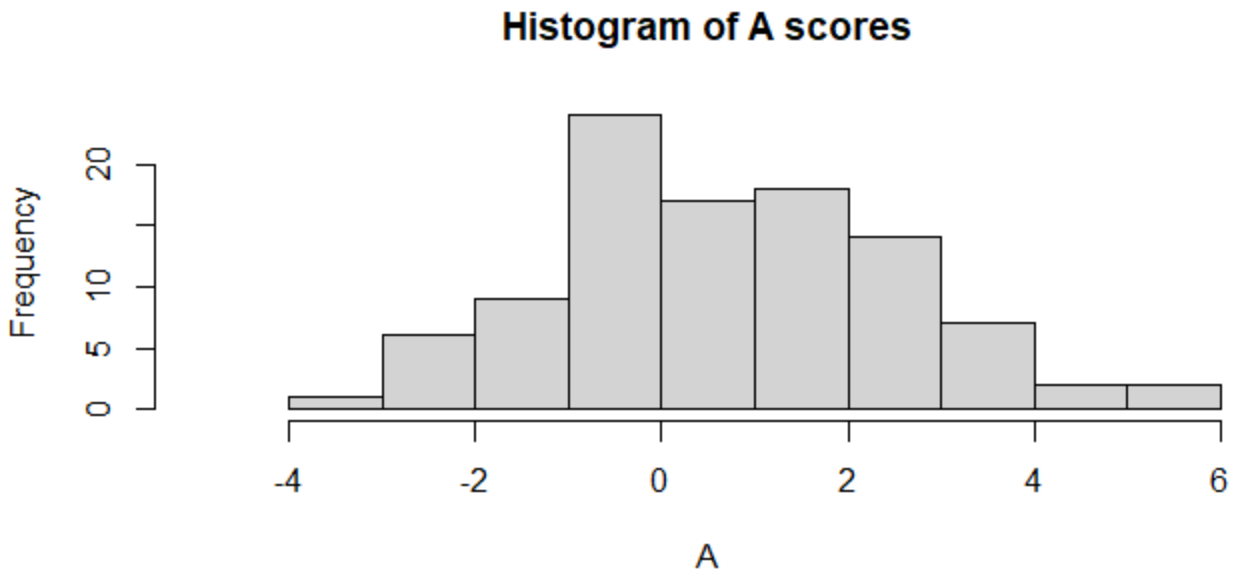
- cure, survival time, …

Then I randomly assign patients to A and B

And that's fine because $\mathcal{AM}$ is a linear function

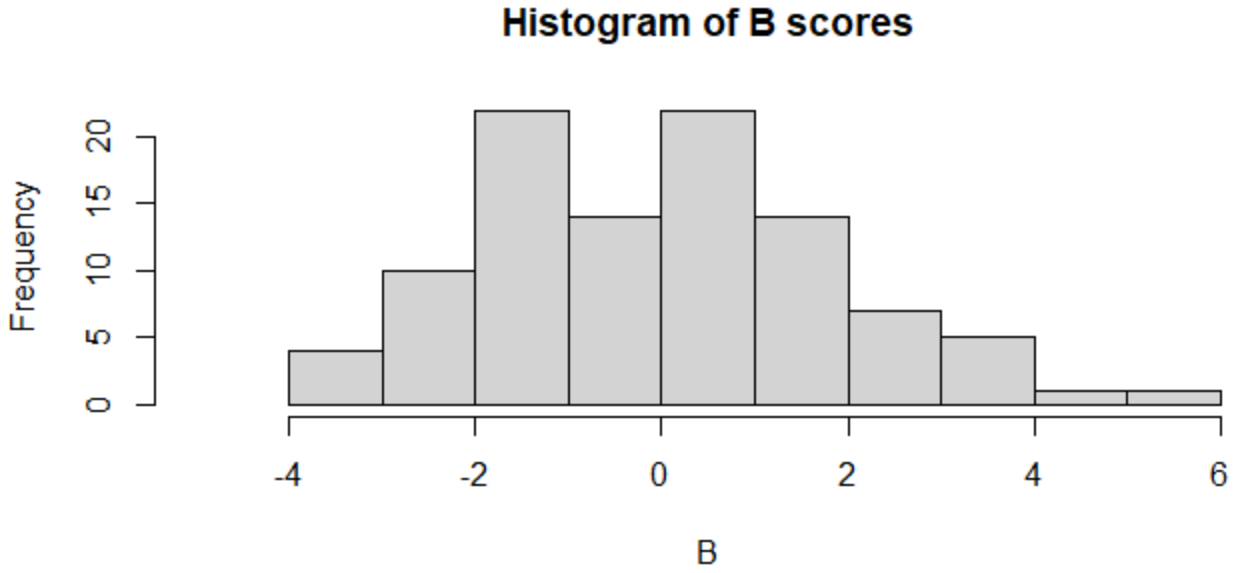
$$\mathcal{AM}(\text{A-B}) = \mathcal{AM}(\text{A}) - \mathcal{AM}(\text{B})$$

And we can use the two sample t-test
(for example, if the score distributions are Normal with equal variance)



$\mathcal{AM}$(A) = 0.75

sd(A) = 1.83



$\mathcal{AM}$(B) = -0.01

sd(B) = 1.83

Except it's not fine (as you all know):

The standard deviation of A-B scores
when A and B are *from the same person*

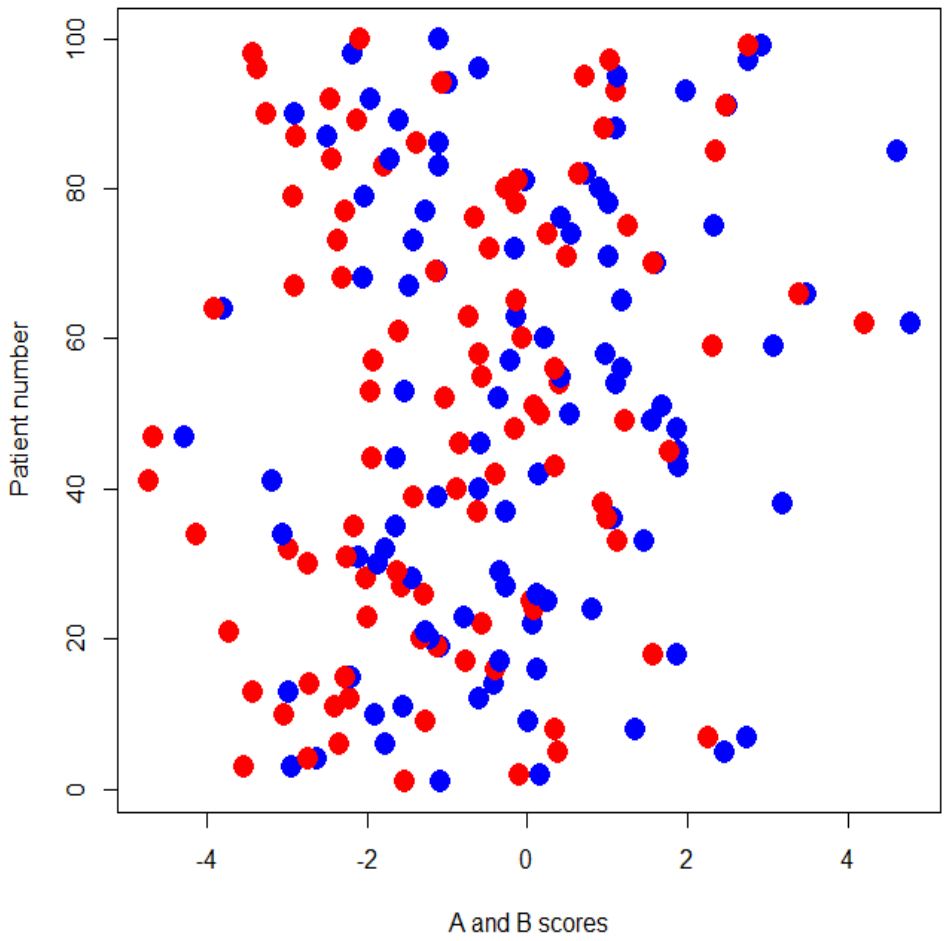
may be very different from the standard deviation
when the A and B scores are *from different (independently chosen)* people

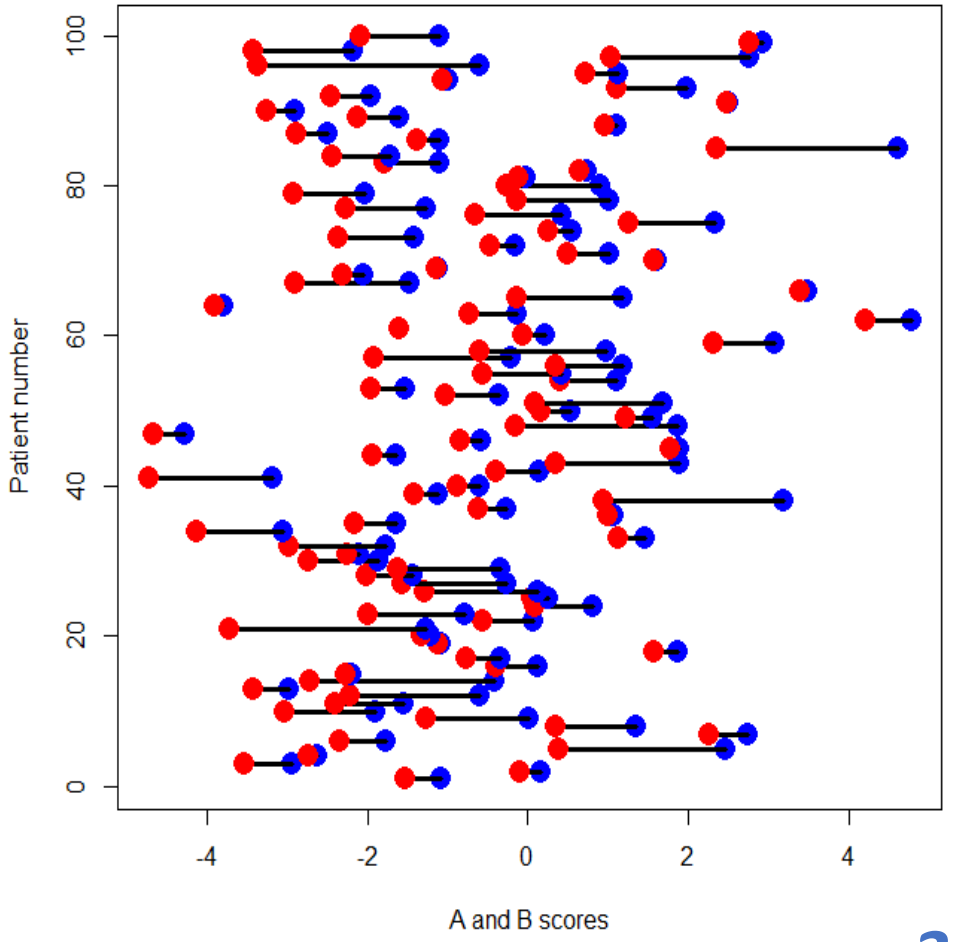
$\mathcal{AM}$(A) = 0.75 blue

$\mathcal{AM}$(B) = -0.01 red

Sample-sd(A-B|indep) = 2.24

Sample-sd(A-B|matched) = 0.63



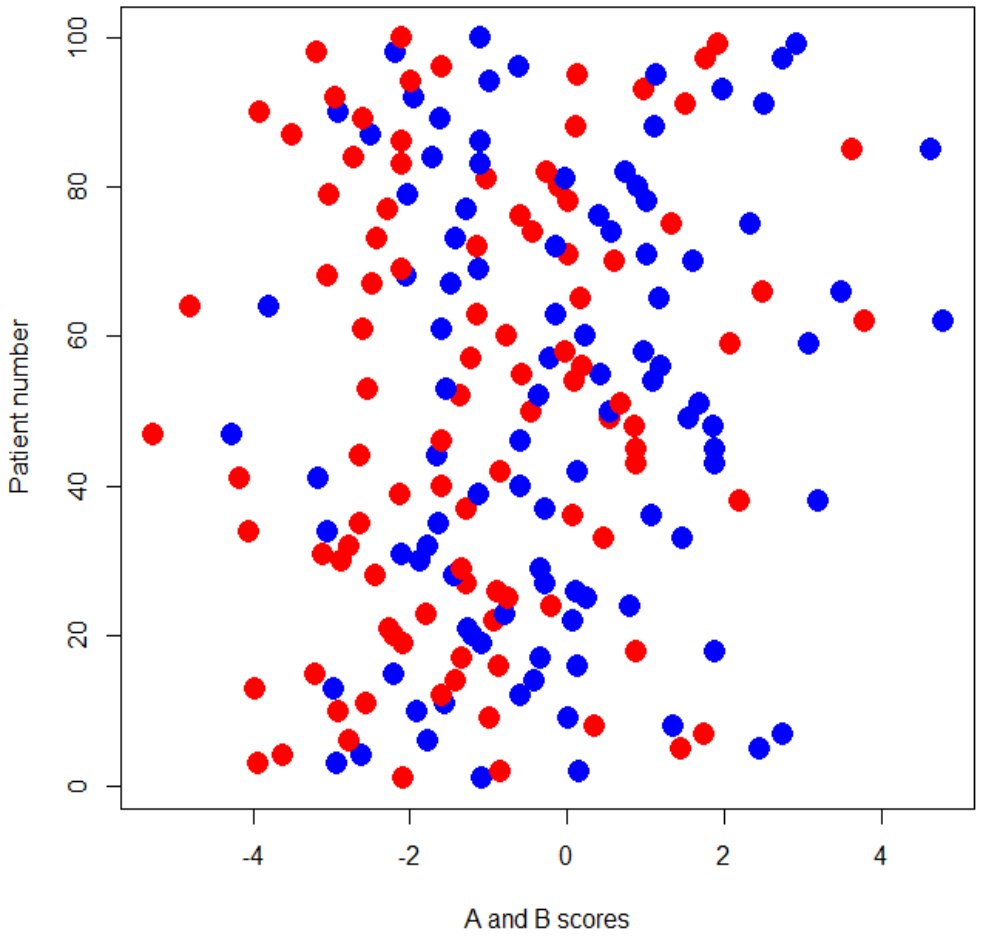


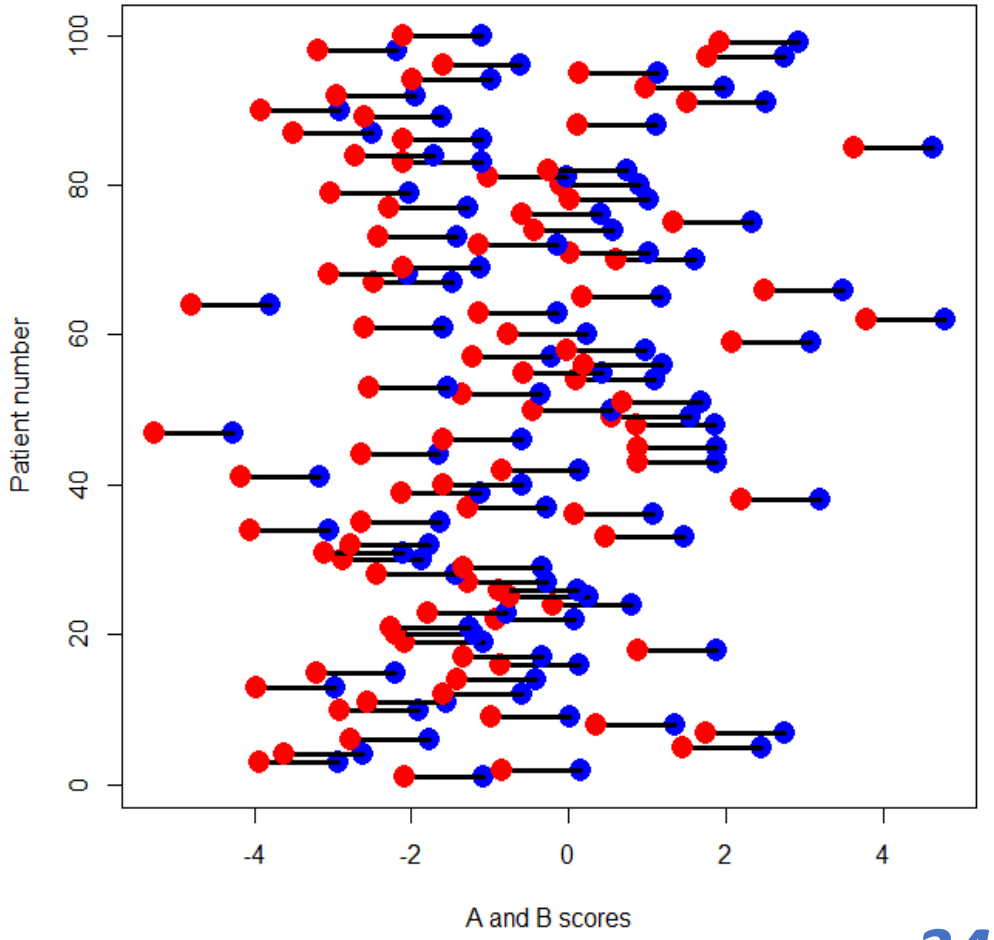
$\mathcal{AM}$(A) = 0.75

$\mathcal{AM}$(B) = -0.25

Sample=sd(A-B|independent) = 2.58

Sample-sd(A-B|matched) = 0





Ideally we would use matched-pairs t-test

but we can't because we are unable to measure both A and B for each person

so we need a more sophisticated approach than simply comparing the means

So arithmetic means can be misused -> ban arithmetic means?

But is any of this right?

My original question was: ***Which of A and B is better?***

This question makes no reference to ***how much*** better A is than B
Again it makes no reference to the ***size*** of the difference

My real question simply concerns **sign**(A-B) not **size**(A-B)

And since we have to look at these for a body of previous people
My real question becomes

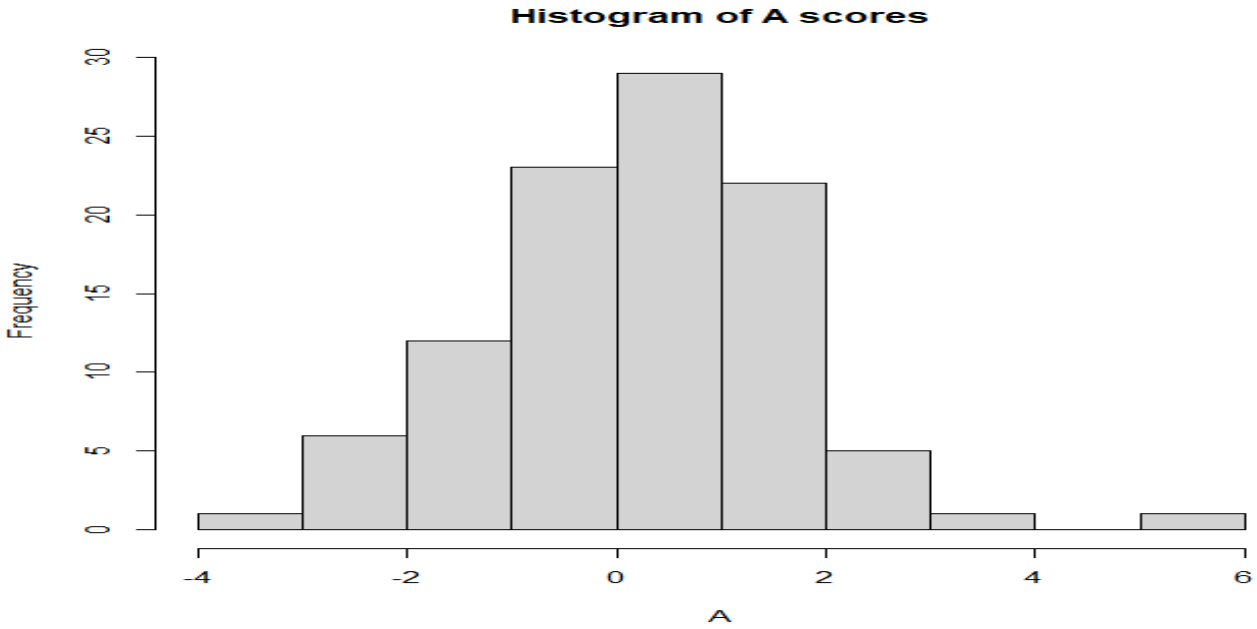
Is Prob(A-B>0)>1/2? *NOT* ***Is AM(A-B)>0?***

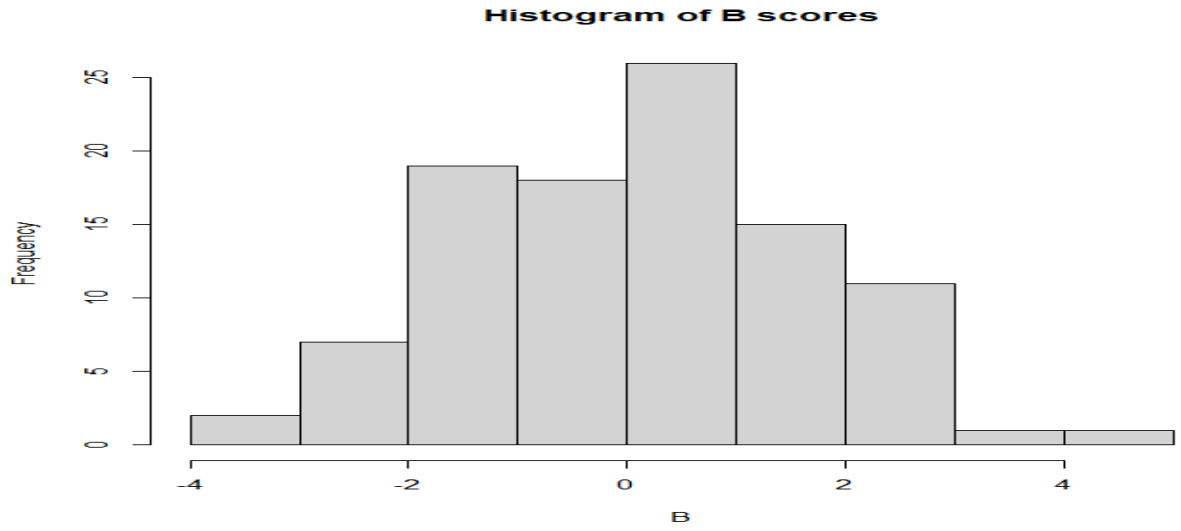
Our question should not involve the arithmetic mean at all

So arithmetic means can be misused -> ban arithmetic means?

This matters because we can have

$$\mathbb{E}[A-B] > 0 \quad \text{while} \quad P(A-B>0) < 1/2$$





Here:

$\mathbb{AM}(A-B) = 0.13 > 0$

while

$P(A-B>0) < 0.45 < 1/2$

In short:

My examples used the arithmetic mean

The arithmetic mean can be misused

And used in poorly understood ways

Should we therefore ban its use?

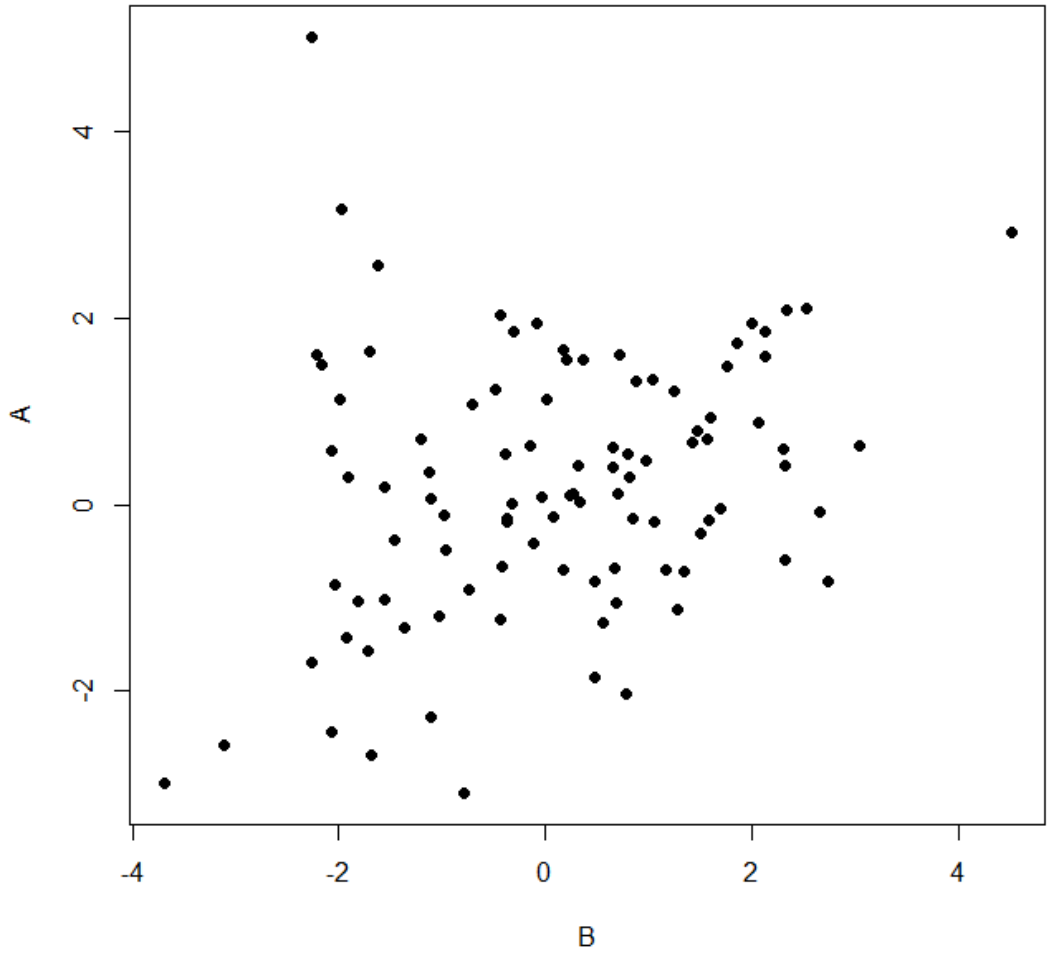
Or should we think carefully about what we want to know
And use the right tool for the question?

The arithmetic mean: don’t say it and don’t use it

But it gets worse!

Allowing for the correlation (which I know because it's synthetic data)

P(A-B>0) = 0.45 < 1/2

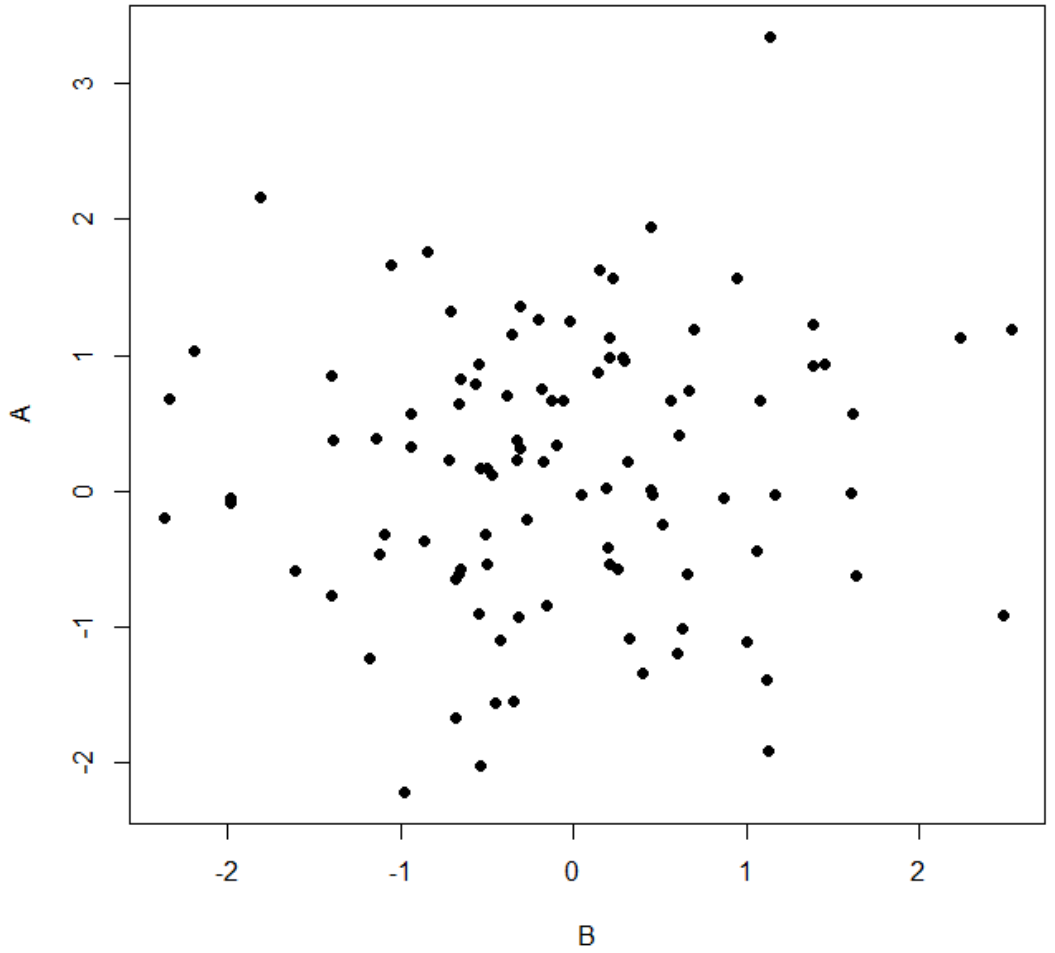


But if we assume independence

As does the standard nonparametric Mann-Whitney-Wilcoxon (MWW) test

P(A-B>0) = 0.58 > ½

So the standard two group approach would conclude that A is better than B when the reverse is the case



So the MWW test can be misused -> ban the MWW test?

Example 5: choice of criterion for supervised classification methods

Sensitivity, specificity, positive predictive value, negative predictive value, recall, precision, error rate, kappa statistic, Youden statistic, F-measure, KS-statistic, maximum proportion correctly classified, Area Under the ROC Curve, Gini coefficient, H-measure, ….

ER – error rate

Proportion of test cases misclassified

AUC – area under ROC curve

Probability that a randomly chosen class 0 object will have a lower score than a randomly chosen class 1 object

$$AUC = \int F_0(t) f_1(t) dt$$

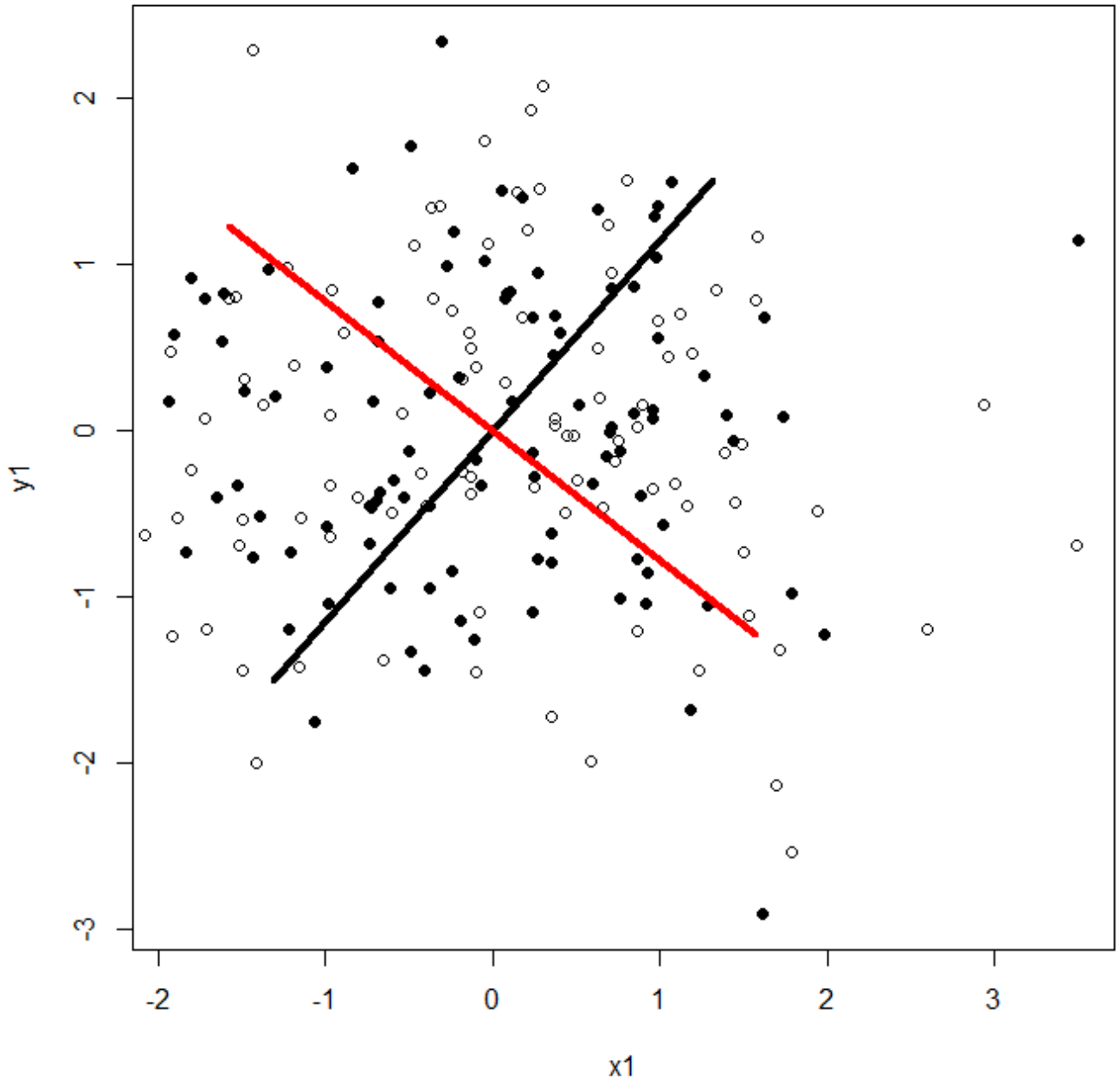
The MWW statistic

Best decision surfaces

AUC in Black

Error rate in Red

Angle between them is 93°



Error rate is used in over 95% of papers assessing the performance of machine learning algorithms

Error rate assumes two kinds of misclassification are equally serious, which is seldom true

Error rate can be misused -> ban the use of error rate?

AUC is widely used in medicine, credit scoring, machine learning, signal detection, etc

AUC is a linear transformation of the expected misclassification loss

But this expectation is calculated over ***different*** distributions of the cost ratio for ***different*** classifiers

This is irrational: the cost ratio distribution should depend on the problem and researcher, not the classifier

AUC can be misused

Should AUC be banned?

Summary

Basic statistical tools can be and are misused

- Even something so basic as the arithmetic mean

But we do not therefore call for a ban on the use of such tools

Rather we call for education

- In their properties
- In how to use them
- In how to interpret them
- In any necessary conditions or assumption in using them

Exactly the same should apply to significance tests

Further reading

Hand D.J. (1992) On comparing two treatments, *The American Statistician*, **46**, 190–192.

Hand D.J. (1994) Deconstructing statistical questions (with discussion), *Journal of the Royal Statistical Society*, *Series A*, **157**, 317–356.

Hand D.J. (2022) Trustworthiness of statistical inference. *Journal of the Royal Statistical Society*, *Series A,* **185**, 329-347.

Hand D.J. and Anagnostopoulos C. (2022) Notes on the H-measure of classifier performance. *Advances in Data Analysis and Classification*, DOI: 10.1007/s11634-021-00490-3.

Mayo D.G. and Hand D.J. (2022) Statistical significance and its critics: practicing damaging science, or damaging scientific practice. *Synthese*. https://link.springer.com/content/pdf/10.1007/s11229-022-03692-0.pdf

Senn S. (2014) https://errorstatistics.com/2014/07/26/s-senn-responder-despondency-myths-of-personalized-medicine-guest-post/#more-15521