What's wrong with how we teach frequentist estimation and inference? And what should we do about it?

(A: Teach coverage and interval estimation)

Mark E Schaffer (Heriot-Watt University, Edinburgh)

ENBIS-22 Conference, Trondheim, Norway, 26-30 June 2022.

P-values and "Statistical Significance"

P-values, "statistical significance", "null hypothesis significance testing" (NHST)

- Much attention in the applied statistics literature in recent years, most of it critical.
- Economics profession just starting to pick up on this (e.g., JEP Summer 2021 symposium on statistical significance, with contributions from Imbens (2021), Kasy (2021), Miguel (2021).)
- American Statistical Association 2016 "Statement on Statistical Significance and P-Values": "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold." (Wasserstein and Lazar, 2016)
- 2019 *Nature* paper by Amrhein et al. (2019), cosigned by over 800 researchers (including me): researchers should "retire statistical significance" in favor of more nuanced interpretation.

A typical (economics) NHST example

A researcher estimates

$$y_i = x_i\beta + \varepsilon_i \tag{1}$$

usually with some "controls", and then tests the null hypothesis

$$H_0 : \beta = 0 \tag{2}$$

based on the estimated $\hat{\beta}$ and its standard error. If the p-value is less than 5%, the researcher declares victory: β is "statistically significant" and it's time to write it up and send it off to a journal.

There is a long list of reasons why this is Bad Practice.

First on my list: by itself, testing $H_0 : \beta = 0$ tells us almost nothing.

A typical (economics) NHST example (continued)

As economists, we almost always want to know the answers to "How big is the effect?" and "How precisely is it estimated?" (Economics very mainstream here.)

By itself, testing $H_0 : \beta = 0$ helps answer neither of these questions.

Say we reject $H_0 : \beta = 0$.

- What if $\hat{\beta}$ is extremely small but extremely precisely estimated?
- What if $\hat{\beta}$ is very large but the standard error is also huge?

Our conclusions should be very different!

(It is amazing that so many papers with this mistake still get circulated.)

All pretty obvious (to this audience, anyway).

NHST and teaching econometrics

Most of the wider debate has been about the problems of misuse of statistical significance, p-values, NHST etc. in the practice of research.

But the problem is rife in teaching, as a casual skim of u/g econometrics textbooks (and statistics textbooks more generally) will reveal.

(NB: Nostra culpa! Looking at my old teaching materials makes for uncomfortable reading in places.)

The pervasiveness of the problem in research over many decades is hard to explain without a causal role for how statistics and econometrics is taught.

Interval estimation and coverage

What should be done instead?

$$y_i = x_i\beta + \varepsilon_i$$

My recommendation:

Interval estimation and coverage probability should be the key teaching outcomes.

- Report $[\hat{\beta}_{LL}, \hat{\beta}_{UL}]$ as the key estimand – **not** $\hat{\beta}_{OLS}$.
- Teaching interval estimation rather than point estimation automatically emphasises uncertainty.

NB: By "interval estimation" I mean frequentist confidence intervals (unless it's a Bayesian course).

An u/g textbook example: the gender wage gap

Empirical exercise results summarised:

- "Based on this sample, we estimate the firm's gender wage gap to be [15%, 21%] based on a 95% confidence interval."
- "Based on this sample, we estimate the firm's gender wage gap to be [1%, 35%] based on a 95% confidence interval."

It's easy to see, and to teach, the difference between these two results: the first estimate is obviously more precise, and the metric is easy to understand.

Compare traditional NHST approach:

- "At the 5% significance level, we can reject the null hypothesis that there is no discrimination. By the way, the p-value is 0.00004%."
- "At the 5% significance level, we can reject the null hypothesis that there is no discrimination. By the way, the p-value is 4%."

What are students to make of this? Thoroughly opaque.

Teaching coverage

To interpret these intervals, students need to understand what "coverage" means. Teaching this concept is easier than it sounds (and **much** easier than teaching p-values).

Definition of coverage: "The coverage probability [or just coverage] of an estimation procedure for a parameter β is the probability that the estimated interval will contain the true β."

Definition of a 95% confidence interval: An interval estimation method with 95% coverage. In repeated samples, 95% of the estimated intervals will contain the true β.

Need to emphasise to students: (1) **The interval (not the β) is random** – it's based on a sample dataset. (2) 95% coverage applies to the **method**.

Teaching this is easier than it sounds, because there are good analogies available.

Teaching coverage

Time permitting, the following slides will be replaced by a live demonstration.

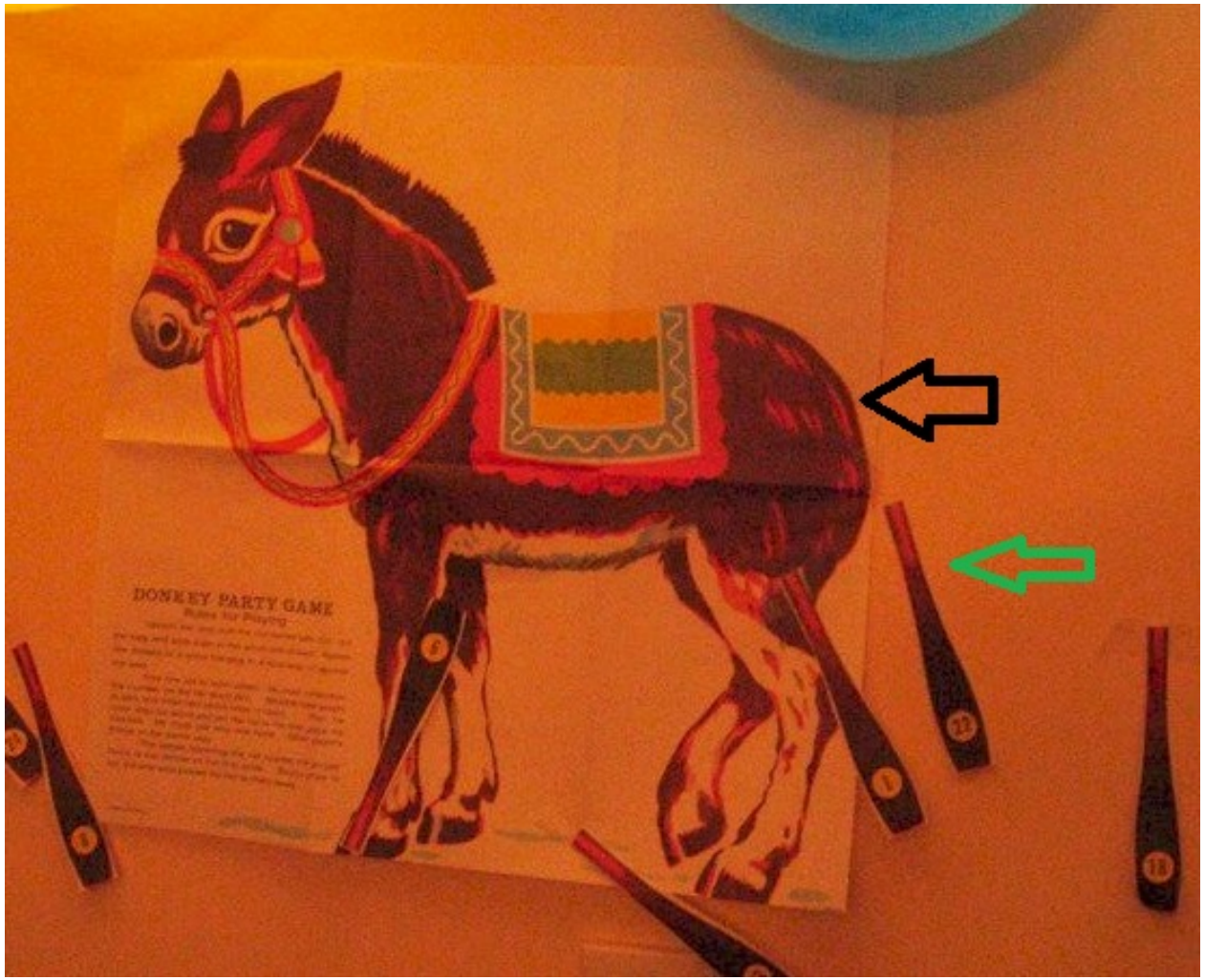
Pin-the-Tail-on-the-Donkey:

Pin-the-Tail-on-the-Donkey is a well-known children's party game.

- A large poster of a donkey is put on a wall. The donkey is missing its tail.
- The child playing is given a tail with a pin or something sticky so that the tail can be attached to the donkey in the appropriate place [sic].
- The catch is that the child is blindfolded and then spun around so that they are disoriented.
- The child is then pointed towards to donkey poster and told to try to put the tail as close as possible to where the tail belongs.
- The other children at the party can yell clues and suggestions to the blindfolded child: "Higher!" "To your right!" "Lower!" And so on.

Pin-the-Tail-on-the-Donkey

The picture here shows the aftermath of a play of the game. The target point – where the tail belongs – is indicated by the black arrow. The winning player's tail is indicated by the green arrow.



Interlude: Confidence intervals vs Prediction intervals?

$$y_i = x_i\beta + \varepsilon_i$$

Confidence interval: an interval for the (fixed) parameter β.

Prediction interval: an interval for the (random) outcome y_i.

Typically we teach confidence intervals first.

But for teaching interval estimation and coverage probability, there's a case for starting with prediction intervals, as we'll see.

Teaching coverage

Pin-the-Ring-on-the-Donkey differs from conventional Pin-the-Tail-on-the-Donkey in two key respects:

- In Pin-the-**Ring**-on-the-Donkey you're blindfolded as usual, but instead of a tail with a pin, you're trying to place a ring on the poster where the donkey's tail goes. If the ring contains the point where the tail goes, you get a point; if it doesn't, you get nothing.
- The exact placement of the donkey is determined **after** you put on the blindfold. Each time you play, the donkey is takes up a new (random) place.

The features of this game share a near-complete range of characteristics with frequentist prediction intervals.

And because it's based on a children's game – indeed, one often familiar to many students already (depending on the country...) – it's easy to teach.

Pin-the-Ring-on-the-Donkey and Frequentist PIs

- The (random) placement of the donkey's rear is our out-of-sample random outcome y_{oos}.
- The ring is our prediction interval.
- At the end of the play, we take off our blindfold, and we see where the (random) donkey is (y_{oos}).
- We also see where our ring ended up.
- The placement of the ring is random; each time we play the game, the ring will end up in a different place.
- But any one time we play the game, it will either contain, or not contain, the realised out-of-sample point where the tail belongs.
- The ring is no longer random after the end of play.

Pin-the-Ring-on-the-Donkey and Frequentist PIs

Still more:

- The size of the ring corresponds to the confidence level. Want to win more often? Use a larger ring. (Set the level higher and the prediction interval will be wider.)
- The clues yelled out by our friends are datapoints. Very few clues (a small dataset) and we are likely to do poorly; more clues and we will do better.
- The frequentist properties of interval construction are analogous to playing the game repeatedly. We know that if we play the game over and over, 95% of the time we will score a point, just as 95% of the time our prediction interval will contain the actual out-of-sample outcome y_{oos}.
- But in any one play of the game we might or might not score.

Mystery Pin-the-Ring-on-the-Donkey and Frequentist CIs

What about **confidence intervals for the parameter** β? Tweak the game.

Mystery Pin-the-Ring-on-the-Donkey differs from Pin-the-Ring-on-the-Donkey in two respects:

- The donkey is placed once, before the game is ever played, and never moves. (Fixed location, not random.)
- In **Mystery** Pin-the-Ring-on-the-Donkey, you never find out the result of any particular game. The blindfold never comes off.

The features of this game share a near-complete range of characteristics with frequentist confidence intervals.

No need to go through all the details – the list is similar to that for prediction intervals.

Mystery Pin-the-Ring-on-the-Donkey and Frequentist CIs

- Key difference between the standard and "Mystery" versions: with prediction intervals, we can check whether our prediction intervals have the expected coverage probability by looking at multiple out-of-sample observations.
- In fact, we can use out-of-sample observations to construct our prediction intervals.
- With confidence intervals, the magic of mathematical statistics tells us what the coverage will be **under certain assumptions**. But we have to believe these assumptions! ("Have confidence in them" ... sorry.) Important to emphasise this to students.
- In "Mystery Pin-the-Ring-on-the-Donkey" we **never find out** if we won, and in real-word applied statistics, we never find out if our interval **really did** contain the true β.
- ... But sometimes we find out if it *didn't*.

Confidence Intervals and Realized Confidence Intervals

We extend "Mystery Pin-the-Ring-on-the-Donkey" by introducing a formal set of rules: "*Olympic* Mystery Pin-the-Ring-on-the-Donkey".

In the modern javelin competition, the javelin has to land in a well-defined "sector" that extends outwards from the point where the javelin is launched. If the thrown javelin lands outside this sector, it is an illegal throw and does not count.

In "Olympic Mystery Pin-the-Ring-on-the-Donkey", the ring, when placed, has to contain part of the poster on which the donkey appears. If the ring is placed outside the poster, that play is disqualified.

The player is told **after** the play *whether or not the ring was placed on the poster*, i.e.,whether or not it was a legal play.

Equivalently, the donkey is removed from the poster and then the blindfold comes off, so the player can see whether the ring was placed legally or not.

Confidence Intervals and Realized Confidence Intervals

This is a simple but effective way to convey that the coverage property of confidence intervals applies **only to the procedure**.

- Prior to any play of "Olympic Mystery Pin-the-Ring-on-the-Donkey", a player may have a coverage probability of 95%, i.e., a 95% change of scoring a point.
- But after the donkey is removed from the poster and the player removes their blindfold, they can see whether it was a legal play.
- If it was not a legal play, then the ring covers an area where it is literally impossible for the donkey's tail to belong.
- **Yet the ex ante coverage probability of the procedure was (and remains, for future plays) 95%.**
- Similarly, a realized confidence interval can contain values that are literally impossible for β to take, and yet the procedure can have nominal coverage equal to actual coverage.

Conclusions

- The problems with NHST and the misuse of p-values is still widespread in applied statistics.
- We should be optimistic: the very fact that there is widespread recognition in the research and academic community that there is indeed a problem is good news indeed.
- But progress will be slow until we change how we teach our students.
- We should teach our students interval estimation as the key learning outcome, and – in the frequentist setting – coverage as the key concept.
- I have suggested some tools for how to do this in an accessible and easy-to-understand way.

Thank you!

What's wrong with how we teach frequentist estimation and inference?
And what should we do about it?

(A: Teach coverage and interval estimation)

Mark E Schaffer (Heriot-Watt University, Edinburgh)

ENBIS-22 Conference, Trondheim, Norway, 26-30 June 2022.

References I

G.W. Imbens. Statistical significance, p-values, and the reporting of uncertainty. *Journal of Economic Perspectives*, (3):157–174, 2021.

G.W. Kasy. Of forking paths and tied hands: Selective publication of findings, and what economists should do about it. *Journal of Economic Perspectives*, (3):175–192, 2021.

E. Miguel. Evidence on research transparency in economics. *Journal of Economic Perspectives*, (3):193–214, 2021.

Ronald L. Wasserstein and Nicole A. Lazar. The asa statement on p-values: Context, process, and purpose. *The American Statistician*, 70 (2):129–133, 2016. doi: 10.1080/00031305.2016.1154108. URL `https://doi.org/10.1080/00031305.2016.1154108`.

Valentin Amrhein, Sander Greenland, and Blake McShane. Scientists rise up against statistical significance. *Nature*, (567):305–307, 2019.