New physics in the garden of forking paths

Andrew Fowlie

Mini-Workshop of LHC anomalies 2023

Xi’an Jiaotong Liverpool University

西交利物浦大学

The truth will set you free.

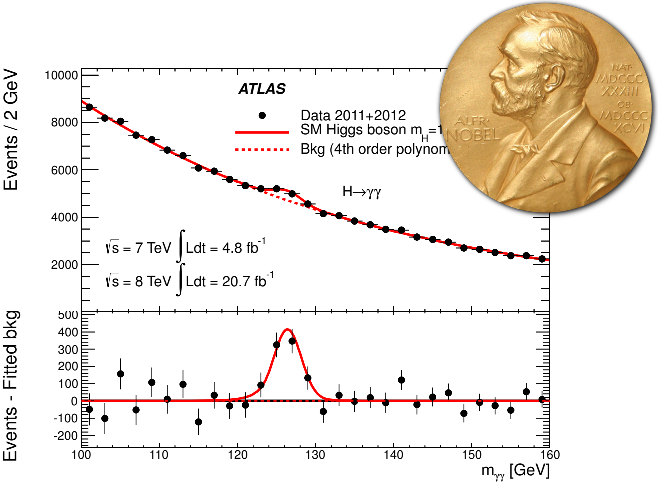
But first it’s gonna piss you off.

Outline

1. Distinguishing signals from noise
2. Dangerous research practices — HARKing and dredging
3. New physics in the garden of forking paths

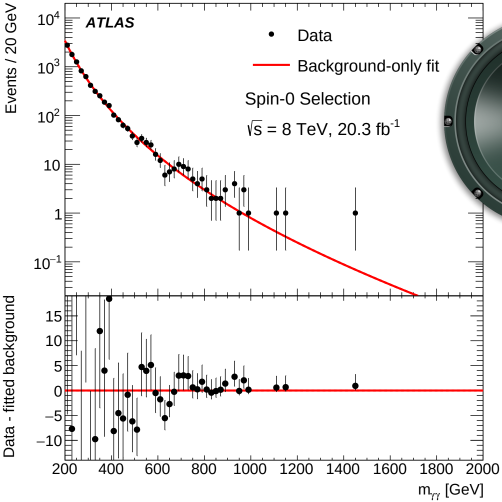
Distinguishing signals from noise

Nobel prize



Signal — Higgs discovery (Aad et al., 2013)

Noise



Noise — 750 GeV bonanza (Aad et al., 2015)!

*The first principle is that you must not fool yourself —
and you are the easiest person to fool*

R. Feynman (1974). “Cargo Cult Science”.

Distinguishing signals from noise

- We need a statistical methodology for distinguishing signals from noise
- In high-energy physics, that methodology typically involves a *p*-value
- We'll still make mistakes
- But perhaps we can reduce or control chances of mistakenly interpreting noise as a signal

P-values

P-value (Wasserstein and Lazar, 2016)

The p-value, p, is the probability of observing data as or more extreme than that observed, given the null hypothesis, H_0, i.e.,

$$p = P\left(\lambda \geq \lambda_{\text{Observed}} \,\middle|\, H_0\right)$$

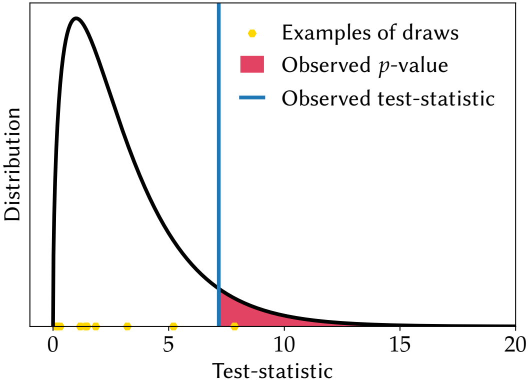
where λ is a test-statistic that summarises the data and defines extremeness, and H_0 specifies the distribution of λ

See Demortier, 2008 for discussion about composite null hypotheses that don't uniquely specify the distribution of λ.

Test-statistic often based on (profiled) likelihood ratio (Neyman and Pearson, 1933)

P-values

Thus *p* is a tail probability.



Type-1 errors

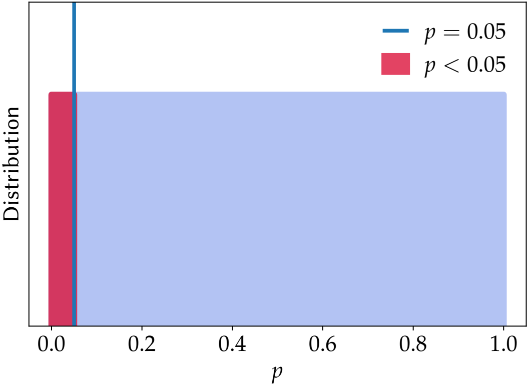
Type-1 error

Reject the null hypothesis when it is true — interpret noise as a signal

- We construct a rule so that we'd make type-1 errors at a pre-specified rate in the long-run in an ensemble of experiments
- That is, control type-1 error rate, α
- We can't only consider only the observed data or the analysis we performed on the observed data
- We must consider the whole ensemble of repeats — all the analyses we would have done were the data different

Controlling the type-1 error rate

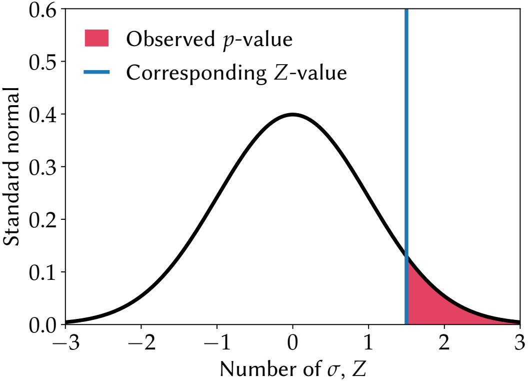
The p-value enables us to control type-1 error rate because, as it is tail probability, it is uniformly distributed under the null



Placing a threshold $p < \alpha$ controls the type-one error rate to be α

5σ discovery threshold

In particle physics, it's conventional to translate p-values into Z-values. 5σ corresponds to about $p = 10^{-7}$



Anything between $2\sigma \lesssim Z < 5\sigma$ is called an anomaly

Interpreting *p*-values

Very popular in particle physics and elsewhere. Two possibly contradictory interpretations:

- P is a measure of evidence against H_0: small $p \Rightarrow H_0$ implausible
- P is a means to control error rate: if we reject null when p-value ≤ 0.05, for example, becomes error theoretic approach with type 1 error rate 0.05

Interpreting 5σ

Consequently, there are two possibly contradictory interpretations of a 5σ rule:

- 5σ is a threshold on strength of evidence — extraordinary claims require extraordinary evidence
- 5σ is a desired type-1 long-run error rate

Dangerous research practices — HARKing and dredging

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment

However, this should not be surprising. It can be proven that most claimed research findings are false

J. P. A. Ioannidis (2005). “Why most published research findings are false”. *PLoS medicine* 2.8, e124.

A NO MALIES
NON-POISSON DISTRIBUTED
UN-CON-FIR-MED
SYSTEMATIC
RIP
† 3σ
4.5 σ
TAILS OF DISTRI-BUTIONS

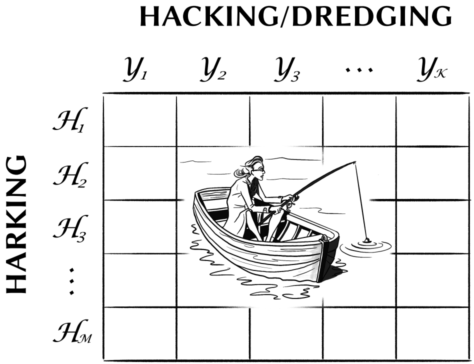
G. D’Agostini (2016). “The Waves and the Sigmas (To Say Nothing of the 750 GeV Mirage)”. arXiv: `1609.01668 [physics.data-an]`.

Causes of reproducibility crisis

- Hacking
- HARKing
- Underpowered studies
- File-drawer problem and dangerous incentives

D. Bishop (2019). “Rein in the four horsemen of irreproducibility”. *Nature* 568.7753, pp. 435–435.

Hacking and HARKing — two axis of wonky statistics



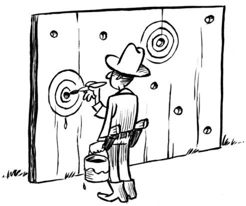
HARKing = Hypothesising After Results Are Known

Original artwork distributed under CC-BY license. Artwork Viktor Beekman & concepts Eric-Jan Wagenmakers.

Good scientists are skilled at looking hard enough and subsequently coming up with good stories (plausible even to themselves, as well as to their colleagues and peer reviewers) to back up any statistically-significant comparisons they happen to come up with

A. Gelman and E. Loken (2013). "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time".

HARKing

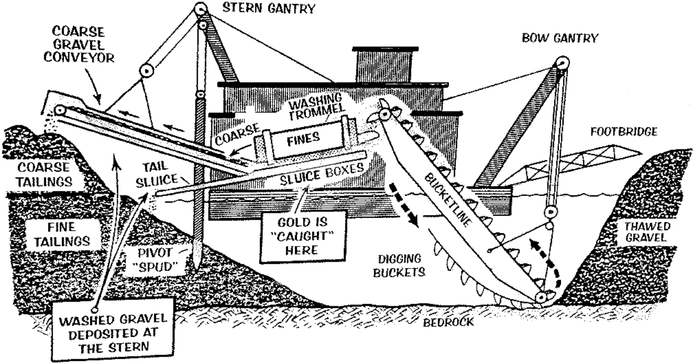


Exploratory research	vs.	Confirmatory research
T1 errors out of control	vs.	T1 errors under control
No evidence	vs.	Compelling evidence

Original artwork distributed under CC-BY license. Artwork Viktor Beekman & concepts Eric-Jan Wagenmakers.

Dredging

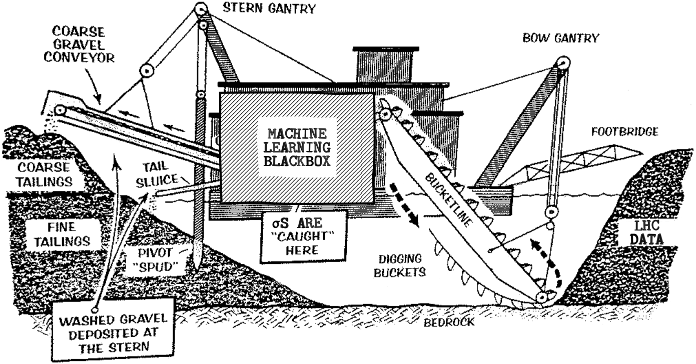
Dredge one million tonnes of river gravel, discover gold



Original artwork Harry Matthews (2009).

Dredging

Dredge one million terabytes of LHC data, discover anomalies



Original artwork Harry Matthews (2009).

Over-interpretation of noise is facilitated by the extent to which data analysis is rapid, flexible and automated

In a high-dimensional dataset, there may be hundreds or thousands of reasonable alternative approaches to analysing the same data

M. R. Munafò et al. (2017). “A manifesto for reproducible science”. *Nature Human Behaviour* 1.1.

During data analysis it can be difficult for researchers to recognise P-hacking or data dredging because confirmation and hindsight biases can encourage the acceptance of outcomes that fit expectations or desires

This, unfortunately, is not scientific discovery, but self-deception

We need measures to counter the natural tendency of enthusiastic scientists who are motivated by discovery to see patterns in noise

M. R. Munafò et al. (2017). "A manifesto for reproducible science". *Nature Human Behaviour* 1.1.

Power failure

- Power = probability reject null hypothesis when it is false
- Suppose I search for a signal, $s = 1$, on top of a background, $b = 100$ and I see an anomaly $o = 130$
- Despite apparent 3σ, this isn't much evidence for signal of anticipated size $s = 1$, as $o \geq 130$ approximately just as likely under each hypothesis
- This experiment was underpowered at anticipated effect size
- Underpowered studies have limited ability to discriminate noise from signal and may be worse than no study at all

K. S. Button et al. (2013). "Power failure: why small sample size undermines the reliability of neuroscience". *Nature Reviews Neuroscience* 14.5, pp. 365–376.

File-drawer problem and dangerous incentives

- If you find a significant result, publish it
- If you don't, stick it in the file-drawer
- Leads to publication bias and inflated error rates in published papers
- Connected to incentives to produce significant results, rather than quality research

R. Rosenthal (1979). "The file drawer problem and tolerance for null results". *Psychological Bulletin* 86.3, pp. 638–641.

New physics in the garden of forking paths

The garden of forking paths

Can we find a more nuanced perspective suitable for high-energy physics?

We think the real story is that researchers can perform a reasonable analysis given their assumptions and their data, but had the data turned out differently, they could have done other analyses that were just as reasonable in those circumstances

A. Gelman and E. Loken (2013). "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time".

In this garden of forking paths, whatever route you take seems predetermined, but that's because the choices are done implicitly

The researchers are not trying multiple tests to see which has the best p-value; rather, they are using their scientific common sense to formulate their hypotheses in reasonable way, given the data they have

The mistake is in thinking that, if the particular path that was chosen yields statistical significance, that this is strong evidence in favour of the hypothesis

A. Gelman and E. Loken (2013). "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time".

Garden of forking paths

- Many possible ways to analyse observed data
- The choices may be reasonable — i.e., no concerted effort at hacking, HARKing or dredging — for the given data
- Would have made different choices, though, for different data
- Without a clear analysis plan, hard to compute significance and account for all possible choices
- Statistical significance needn’t imply that error rates are under control or strong evidence

Are HEP error rates under control?

- Whose error rates? Mine? Yours? HEP-EX? HEP-PH?
- In LHC run-1, there were arguably fewer HEP-EX anomalies than there should have been (Nachman and Rudelius, 2012)
- On the other hand, there are regularly anomalies that generate excitement and activity in HEP-PH and disappear
- These aren’t formally type-1 errors, but are these a kind of error? They are damaging to credibility of HEP-PH (Fowlie, 2021)
- There are even claims of 5σ here and there in HEP-PH

Precautions in HEP-EX

- Coordination from statistics working groups
- Blinding — no hacking or HARKing
- Commitment to publish regardless of significance — no file-drawer
- Usually testing well-motivated theories that we are sensitive to — fewer underpowered tests
- Community effort — reduced dangerous incentives
- Redefined significance — 5σ threshold versus 0.05
- LHC results are already replicated — ATLAS and CMS
- Typically report a global p-value — corrected for multiple comparisons

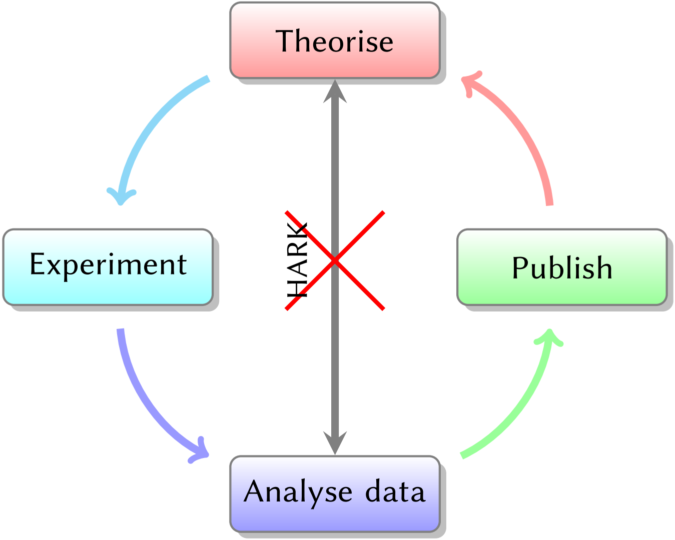
How many of these precautions apply to claims made in HEP-PH?

Risks in HEP-EX

- P-value known to overstate evidence. E.g., a famous result that $p = 0.05$ means that the plausibility of the null hypothesis is at least about 30%
- So many independent analyses and searches: even if the rate of errors under control, anomalies are inevitable even if no new physics
- Tests and measurements that aren't motivated by any specific alternative theory
- Underestimated systematic errors in "dirty" observables

How many apply to b-physics? Where are heaps of anomalies at?

Is this cycle possible in HEP-PH?



HARKing inevitable and vital in HEP-PH

- We are not HEP-EX
- How could we theorise, design an experiment, collect data and test our theory?
- Experiments are so expensive that we can afford one in the whole world
- Perhaps HARKing inevitable and vital
- HARKing problematic to the extent that it misleads about strength of evidence
- Closer coordination between explanations from HEP-PH and future analyses in HEP-EX may help

Draw a distinction between SHARKing (Secretly HARKing) and THARKing (Transparently HARKing)

THARKing can promote the effectiveness and efficiency of both scientific inquiry and cumulative knowledge creation

Failure to THARK in high-stakes contexts where data is scarce and costly may also be unethical

J. R. Hollenbeck and P. M. Wright (2017). "Harking, Sharking, and Tharking: Making the Case for Post Hoc Analysis of Scientific Data". *Journal of Management* 43.1, pp. 5–18.

The question of whether HARKing's costs exceed its benefits is a complex one that ought to be addressed through research, open discussion, and debate

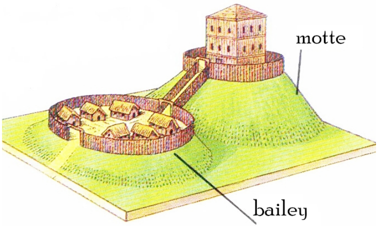
N. L. Kerr (1998). "HARKing: Hypothesizing After the Results are Known". *Personality and Social Psychology Review* 2.3, pp. 196–217.

HEP-PH file-drawer

- Undoubtedly a publication bias/file-drawer effect in HEP-PH — no one is ambulance chasing null results
- Not clear that it is always a problem
- Do we really expect HEP-PH papers on null results of searches? Don't they belong in a file drawer, not arXiv?
- Anomaly detection — perform "model-independent" search for anomalies in massive dataset
- This one activity in HEP-PH in which precautions to avoid file-drawer, dredging and HARKing could be beneficial
- E.g., encouraging pre-registration of planned anomaly hunts prior to public data releases

Conclusions

- Hard to draw reliable conclusions from noisy data
- Strict precautions in HEP-EX, though risks remain
- Best practices advocated aren't easily applied to HEP-PH
- Imponderable multiple corrections in HEP-PH due to garden of forking paths
- That said, post-hoc analyses and reinterpretations are valuable
- Though hard to argue that they could provide compelling evidence for new physics



Evidence interpretation of P hard to justify, but aligns with how we think; error theoretic on firmer footing but requires mental gymnastics

- Bailey — provides evidence
- Motte — controls type-1 error rate

You don’t know *p*

- *P* isn’t formally a measure of evidence
- *P* isn’t an error rate. It’s a means to controlling an error rate. The error rate itself was specified prior to even collecting data (e.g., 5σ)

Yet it’s widely misinterpreted as both of those things

> *The fact that academics don’t know what p means is a symptom of the fact that p doesn’t tell anything worth knowing*

E. Wagenmakers (2020).

Some proposed solutions in social sciences

- Training in statistical methods and problems
- Blinding — no hacking or HARKing
- Pre-registration of experimental design and analysis plan — no hacking or HARKing
- Encourage collaboration; collect more data and increase power
- Commitment to publish pre-registered work, regardless of significance — no file-drawer
- Abandon thresholds for significance — no dangerous incentives
- Redefine significance — lower 0.05 threshold
- Abandon attempts to control rates of error

References I

- Original artwork distributed under CC-BY license. Artwork Viktor Beekman & concepts Eric-Jan Wagenmakers.
- Original artwork Harry Matthews (2009).
- Aad, G. et al. (2013). “Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC”. *Phys. Lett. B* 726. [Erratum: Phys.Lett.B 734, 406–406 (2014)], pp. 88–119. arXiv: `1307.1427 [hep-ex]`.
- — (2015). “Search for resonances decaying to photon pairs in 3.2 fb^{-1} of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector”.
- Bishop, D. (2019). “Rein in the four horsemen of irreproducibility”. *Nature* 568.7753, pp. 435–435.
- Button, K. S. et al. (2013). “Power failure: why small sample size undermines the reliability of neuroscience”. *Nature Reviews Neuroscience* 14.5, pp. 365–376.

References II

- D’Agostini, G. (2016). “The Waves and the Sigmas (To Say Nothing of the 750 GeV Mirage)”. arXiv: `1609.01668 [physics.data-an]`.
- Demortier, L. (2008). “P Values and Nuisance Parameters”.
- Feynman, R. (1974). “Cargo Cult Science”.
- Fowlie, A. (2021). “Comment on ”Reproducibility and Replication of Experimental Particle Physics Results””. arXiv: `2105.03082 [physics.data-an]`.
- Gelman, A. and E. Loken (2013). “The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time”.
- Hollenbeck, J. R. and P. M. Wright (2017). “Harking, Sharking, and Tharking: Making the Case for Post Hoc Analysis of Scientific Data”. *Journal of Management* 43.1, pp. 5–18.

References III

- Ioannidis, J. P. A. (2005). “Why most published research findings are false”. *PLoS medicine* 2.8, e124.
- Kerr, N. L. (1998). “HARKing: Hypothesizing After the Results are Known”. *Personality and Social Psychology Review* 2.3, pp. 196–217.
- Munafò, M. R. et al. (2017). “A manifesto for reproducible science”. *Nature Human Behaviour* 1.1.
- Nachman, B. and T. Rudelius (2012). “Evidence for conservatism in LHC SUSY searches”. *Eur. Phys. J. Plus* 127, p. 157. arXiv: `1209.3522 [stat.AP]`.
- Neyman, J. and E. S. Pearson (1933). “On the Problem of the Most Efficient Tests of Statistical Hypotheses”. *Philos. Trans. Roy. Soc. London Ser. A* 231, pp. 289–337. ISSN: 02643952.
- Rosenthal, R. (1979). “The file drawer problem and tolerance for null results”. *Psychological Bulletin* 86.3, pp. 638–641.
- Wagenmakers, E. (2020).

References IV

Wasserstein, R. L. and N. A. Lazar (2016). “The ASA’s Statement on p-values: Context, Process, and Purpose”. *The American Statistician* 70.2, pp. 129–133.