Finders, Keepers

Statistical fluctuations, causality, and whatnot in particle physics

Pietro Vischia[1]

[1] CP3 — IRMP, Université catholique de Louvain

UCLouvain
Institut de recherche
en mathématique et physique

LIP-Lisboa, LIP Seminars Cycle 2020

Contents

UCLouvain
Institut de recherche
en mathématique et physique

2016: breaking announcement by the American Statistical Association

ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

doi:10.1080/00031305.2016.1154108

Unprecedented policy statement

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

> Q: Why do so many colleges and grad schools teach $p = 0.05$?
> A: Because that's still what the scientific community and journal editors use.
> Q: Why do so many people still use $p = 0.05$?
> A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: "We teach it because it's what we do; we do it because it's what we teach." This concern was brought to the attention of the ASA Board.

Of course, it was not simply a matter of responding to some articles in print. The statistical community has been deeply concerned about issues of *reproducibility* and *replicability* of scientific conclusions. Without getting into definitions and distinctions of these terms, we observe that much confusion and even doubt about the validity of science is arising. Such doubt can lead to radical choices, such as the one taken by the editors of *Basic and Applied Social Psychology*, who decided to ban *p*-values (null hypothesis significance testing) (Trafimow and Marks 2015). Misunderstanding or misuse of statistical inference is only one cause of the "reproducibility crisis" (Peng 2015), but to our community, it is an important one.

When the ASA Board decided to take up the challenge of developing a policy statement on *p*-values and statistical significance, it did so recognizing this was not a lightly taken step. The ASA has not previously taken positions on specific matters of statistical practice. The closest the association has come

Frequentist and Bayesian stat in one slide

- Frequentist probability: the most familiar?
  - Based on the possibility of repeating—under similar conditions—an experiment many times
  - Repeat an experiment N times, observe n events of type A
  - Probability for any event to be of type A: empirical limit of the frequency ratio $P(X) = \lim_{N\to\infty} \frac{n}{N}$
  - Defined only for sets of data
- Bayesian probability: the most intuitive?
  - Based on the concept of degree of belief
  - A subjective definition by De Finetti based coherent bet: win given amount if X, win nothing if not X
    $P(X) := \frac{\text{The largest amount you are willing to bet}}{\text{The amount you stand to win}}$
  - Bet must be *coherent*: no guaranteed expectated profits (no Dutch book)
  - Depends on the knowledge of the observer *prior* to the experiment
  - Supposed to change when the observer gains more knowledge (e.g. after an experiment)

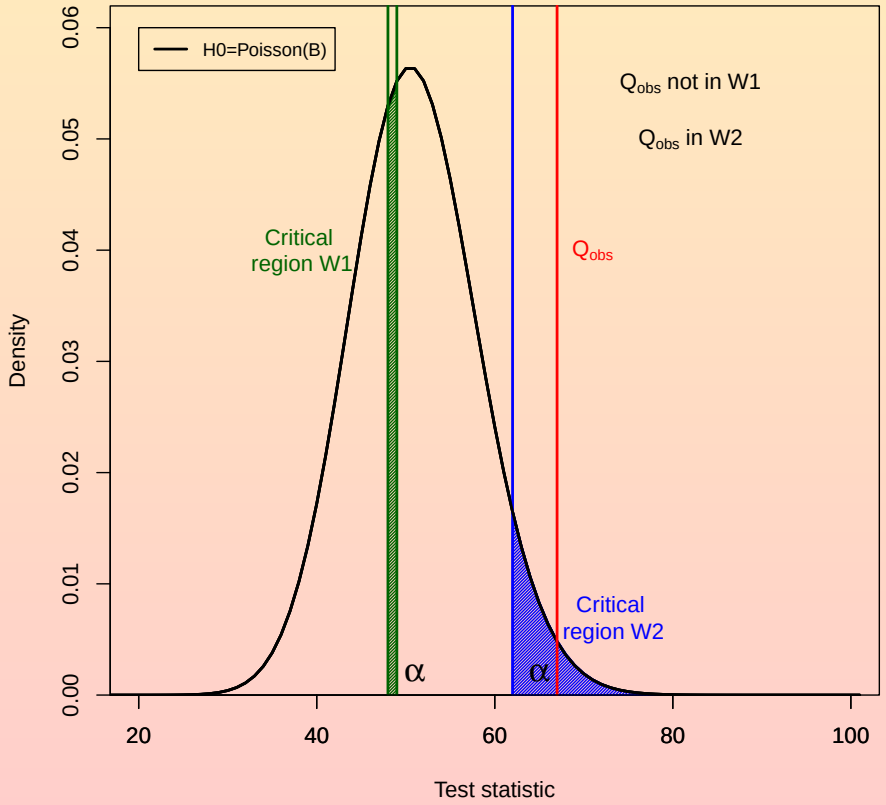
Book	Odds	Probability	Bet	Payout
Trump elected	Even (1 to 1)	$1/(1+1) = 0.5$	20	$20 + 20 = 40$
Clinton elected	3 to 1	$1/(1+3) = 0.25$	10	$10 + 30 = 40$
		$0.5 + 0.25 = 0.75$	30	40

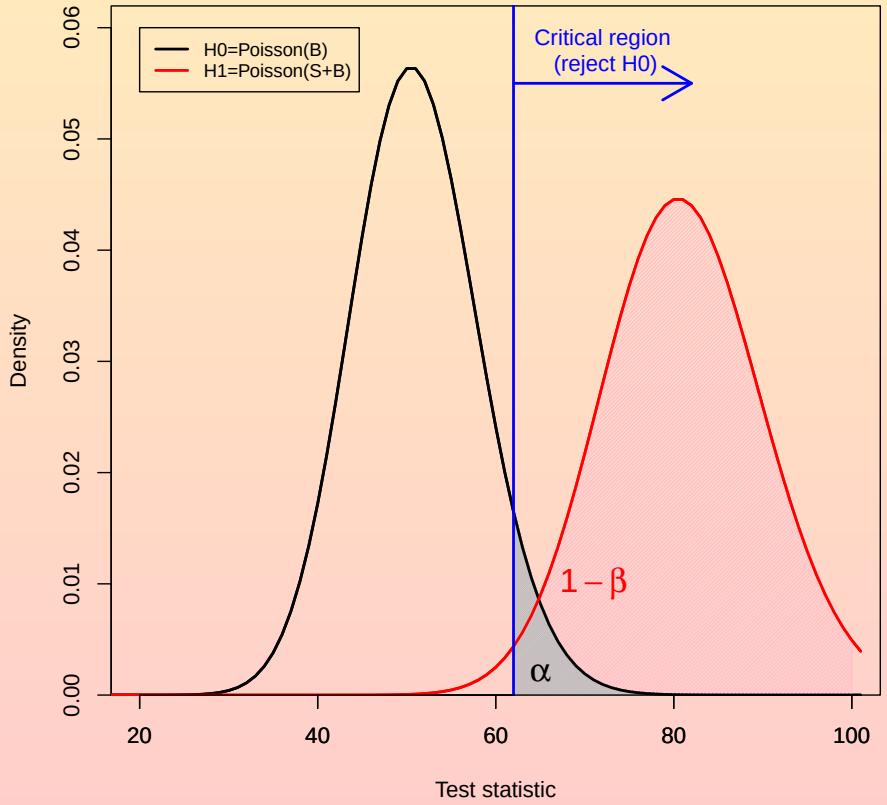
Hypotheses

- Hypothesis: a complete rule that defines probabilities (density functions) for a function of the data (*test statistic*)
  - Simple: completely specified (or each of its parameters is fixed to a single value)
  - Complex family of hypotheses parameterized by one or more parameters $P(\vec{x}; \theta) := P(\vec{x}; H(\theta)$.
- Statistical test
  - A statistical test is a proposition concerning the compatibility of H with the available data.
  - A binary test has only two possible outcomes: either accept or reject the hypothesis
- No reference to a ground truth

Frequentist test

- H_0: the hypothesis to test, which we assume true in absence of further evidence
- $\mathbf{Q}$: a function of the observations (called "*test statistic*"), defined in a space W
  - Critical region w: observations X falling into w are regarded as suggesting that H_0 is **not** true
- $\alpha := P(X \in w | H_0)$: level of significance: when small, a-priori preference to H_0
- Perform experiment, check where $\vec{x}_{obs}$ lies, *reject* H_0 if $\vec{x}_{obs} \in w$, *accept* H_0 if $\vec{x}_{obs} \notin w$
- Need alternative hypothesis H_1 to solve ambiguity in critical region choice
  - We can use our expectations about reasonable alternative hypotheses to design our test to exlude H_0
- If H_0 rejected, often H_1 is the new H_0 (explains better the data)
  - E.g. from (H_0:noHiggs, H_1:Higgs) to (H_1:Higgs , H_1:otherNewPhysics)



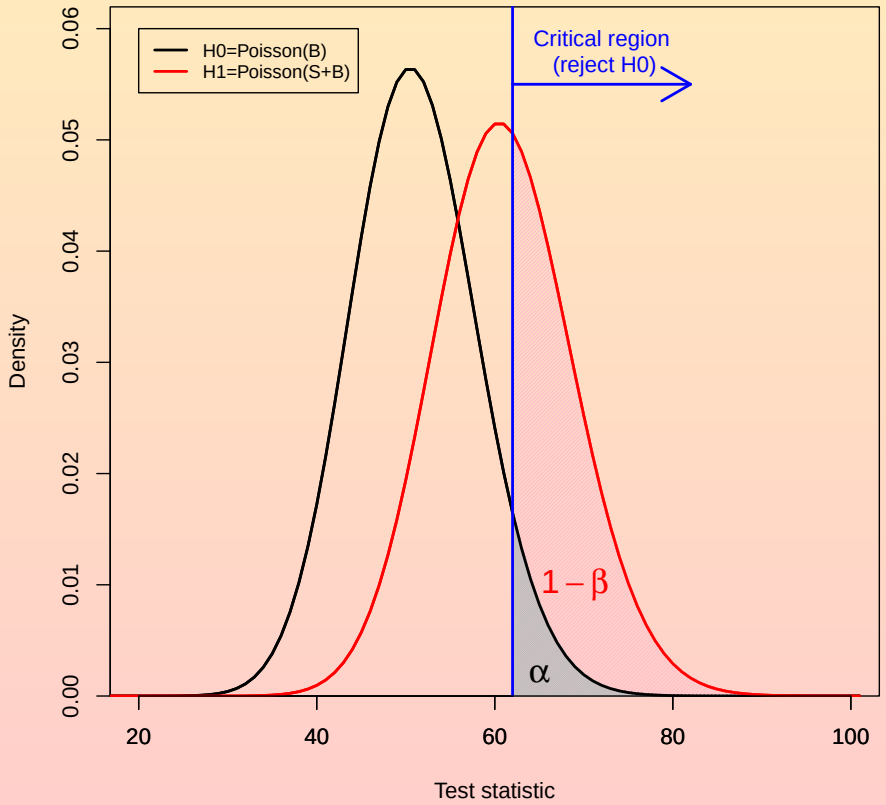


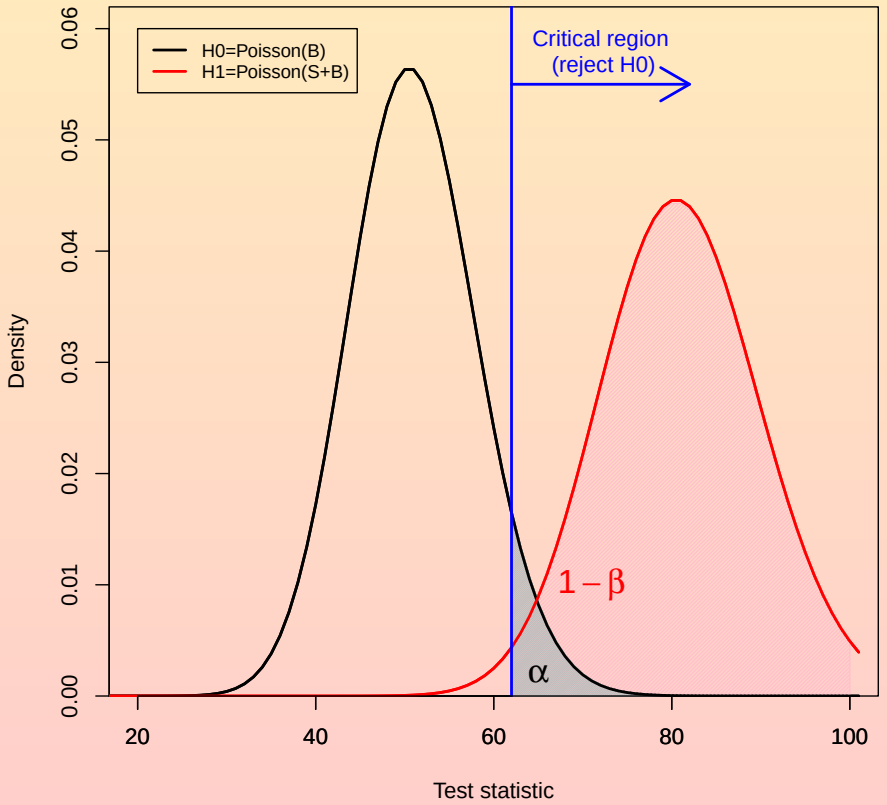
Plots from Vischia—in preparation with Springer

How useful is a test ? Type I and II errors

- *Power of the test*: how well it discriminates against the alternative hypothesis
  - $P(X \in w|H_1) = 1 - \beta$
  - Power $(1 - \beta)$ is the probabiliity of X falling into the critical region if H_1 is true
  - $P(X \in W - w|H_1) = \beta$
  - β is the probability that X will fall into the acceptance region if H_1 is true

	Choose H_0	Choose H_1
H_0 is true	$1 - \alpha$	α (Type I error)
H_1 is true	β (Type II error)	$1 - \beta$ (power)

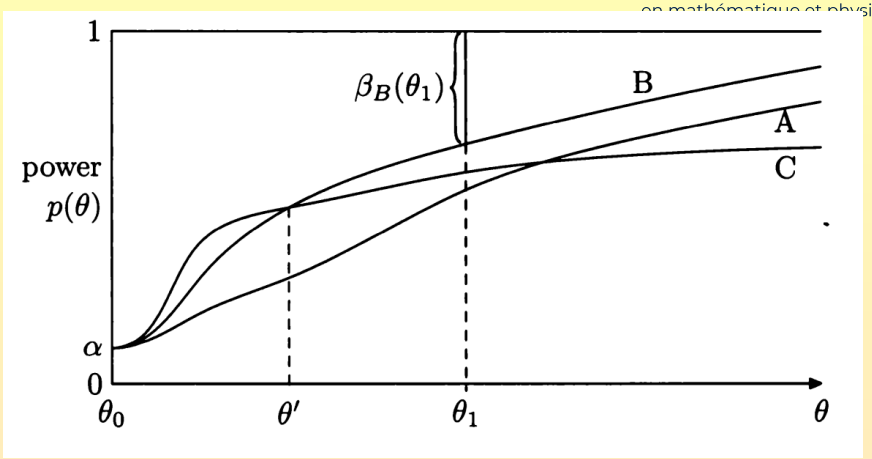


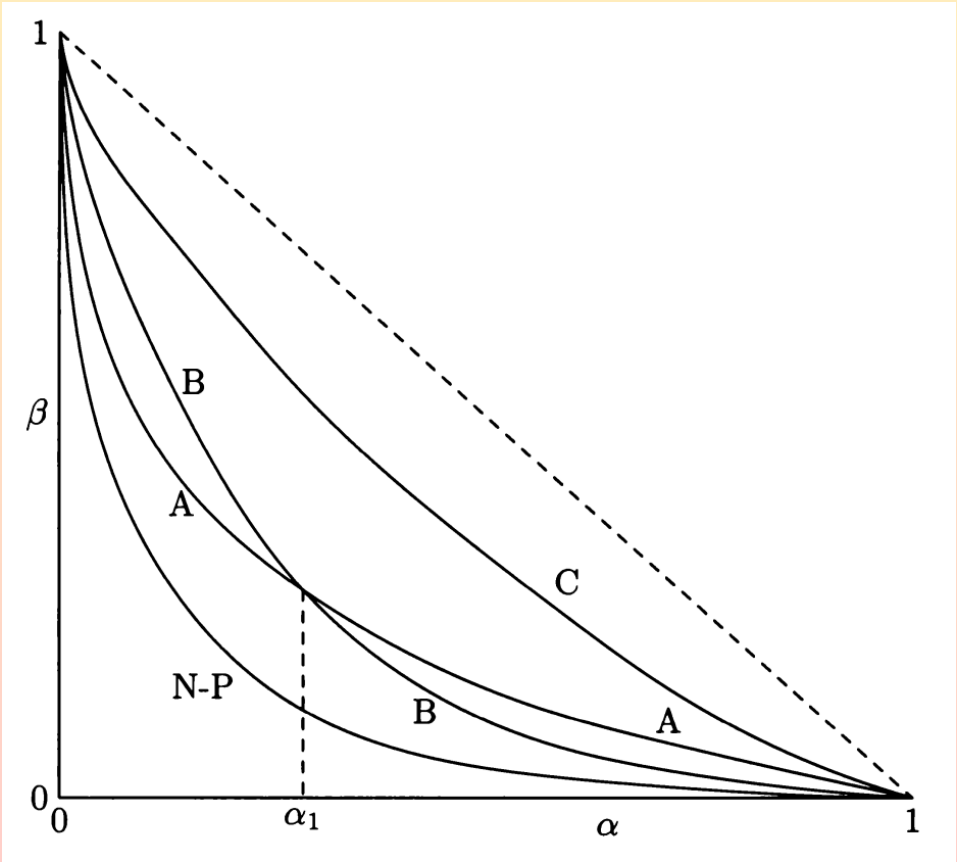


Plots from Vischia—In preparation with Springer

Choose a suitable test

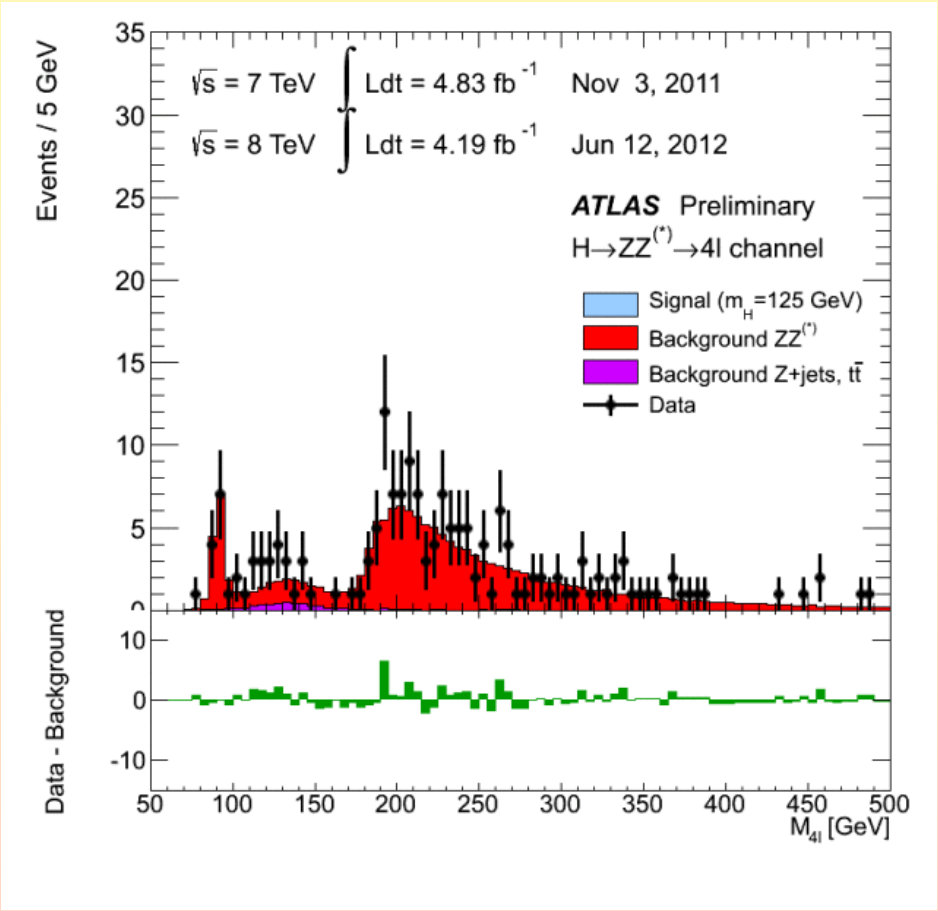
- For parametric (families of) hypotheses
  - $H_0 : \theta = \theta_0$
  - $H_1 : \theta = \theta_1$
  - Power: $p(\theta) = 1 - \beta(\theta)$
- For the null, $p(\theta_0) = 1 - \beta(\theta_0) = \alpha$
- Can choose a *more powerful test*
- For each value of $\alpha = p(\theta_0)$, compute $\beta = p(\theta_1)$, and draw the curve
- Curves closer to the axes are better tests
- Ultimately, though, choose based on the cost function of a wrong decision
  - Bayesian decision theory
- Neyman-Pearson test as the most powerful test
  - Simple (H_0:θ_0 vs H_1:θ_1) hypotheses
  - Choose critical region based on likelihood ratio $\ell(\mathbf{X}, \theta_0, \theta_1) := \frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)} \geq c_\alpha$
  - Valid for simple (non-parametric) hypotheses (when likelihood is computable)
  - Not necessary optimal for complex hypotheses

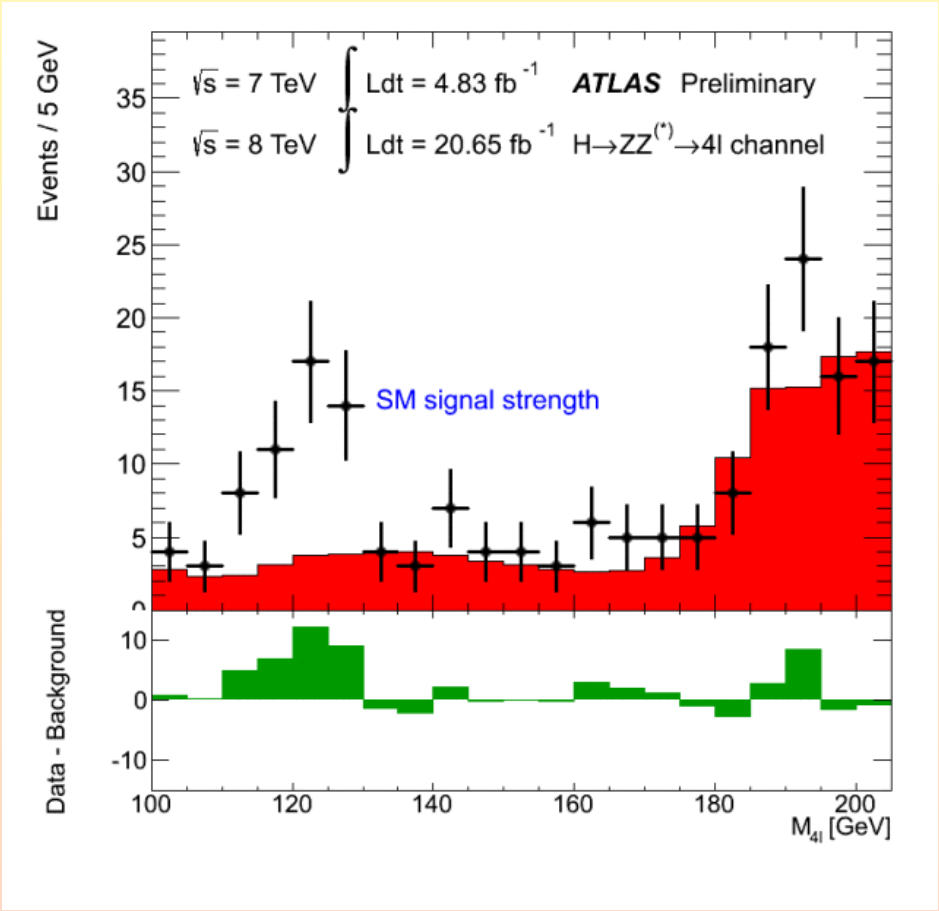




Plots from James, 2nd ed.

I have an excess, do I?

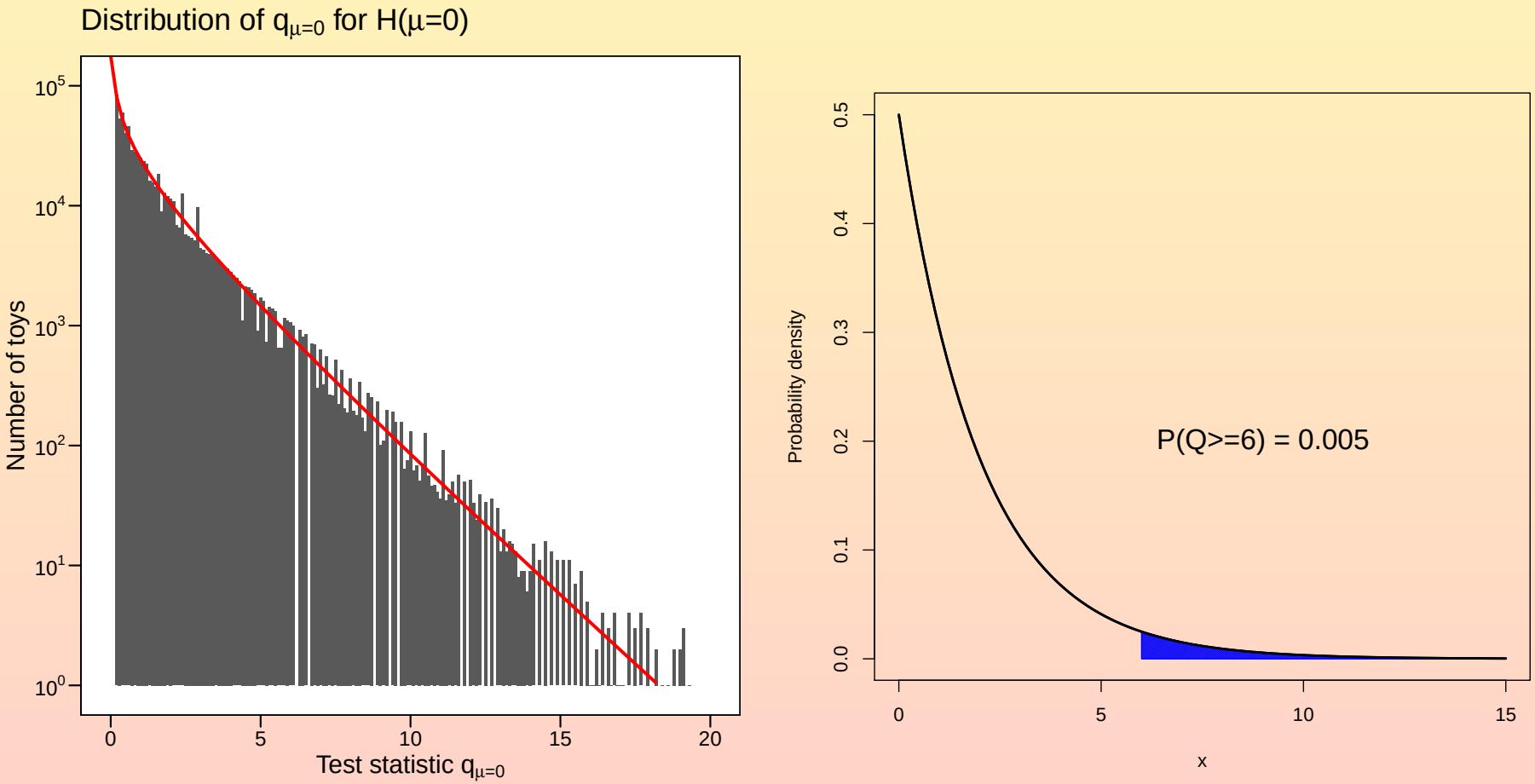




Plot from https://cds.cern.ch/record/2230893

Look only at the null hypothesis!

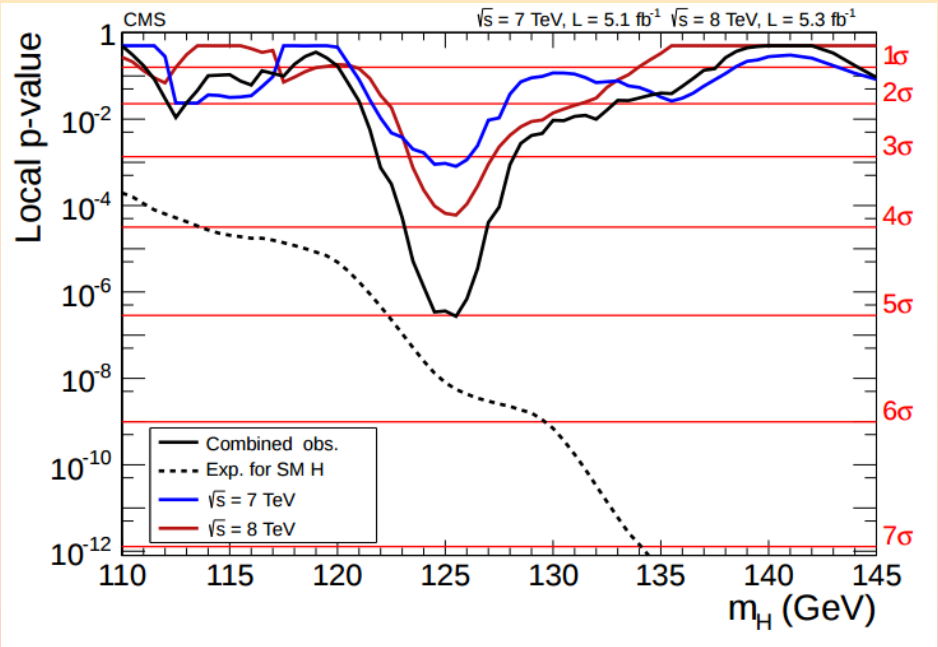
- Probability of obtaining a fluctuation with test statistic q_{obs} or larger, under the null hypothesis H_0
  - Distribution of test statistic under H_0 either with toys or asymptotic approximation (if N_{obs} is large, then $q \sim \chi^2(1)$)

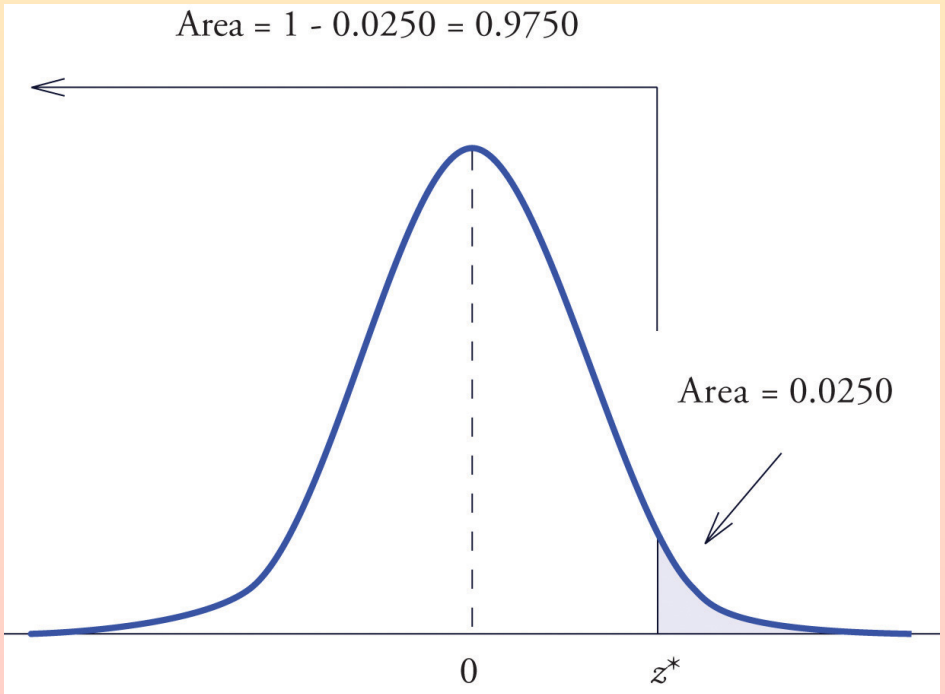


Plots from Vischia—in preparation with Springer

And the sigmas?

- Just an artifact to convert p-values to easy-to-remember $\mathcal{O}(1)$ numbers
  - 1σ: $p = 0.159$
  - 3σ: $p = 0.00135$
  - 5σ: $p = 0.000000285$
- No approximation involved, just a change of units to gaussian variances: one-sided tail area
  $\frac{1}{2\pi}\int_x^\infty e^{-\frac{t^2}{2}}dt = p$
  - p-value must be **flat** under the null, or interpretation is invalidated
- HEP: usually interested in one-sided deviations (upper fluctuations)
  - Most other disciplines interested in two-sided effects (e.g. 2σ: $p_{2sided} = 0.05$)





Left: ATLAS Collaboration, Right: https://saylordotorg.github.io/

Back to ASA: the six statements

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
   - The widespread use of "statistical significance" (generally interpreted as $p \leq 0.05$) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.
4. Proper inference requires full reporting and transparency
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.
   - ...supplement or even replace p-values with other approaches. These include methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates.

doi:10.1080/00031305.2016.1154108

Responses to ASA statement: redefine pvalue threshold or not use it at all

- Benjamin *et al.* (doi:/10.31234/osf.io/mky9j) proposed to switch to lower threshold ($p < 0.005$) and not use it as criterion for publication

> **One Sentence Summary:** We propose to change the default *P*-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005.

- Wagenmakers (doi:/10.3758/BF03194105) proposed to switch to Bayesian criteria

A practical solution to the pervasive problems of *p* values

ERIC-JAN WAGENMAKERS
University of Amsterdam, Amsterdam, The Netherlands

In the field of psychology, the practice of *p* value null-hypothesis testing is as widespread as ever. Despite this popularity, or perhaps because of it, most psychologists are not aware of the statistical peculiarities of the *p* value procedure. In particular, *p* values are based on data that were never observed, and these hypothetical data are themselves influenced by subjective intentions. Moreover, *p* values do not quantify statistical evidence. This article reviews these *p* value problems and illustrates each problem with concrete examples. The three problems are familiar to statisticians but may be new to psychologists. A practical solution to these *p* value problems is to adopt a model selection perspective and use the Bayesian information criterion (BIC) for statistical inference (Raftery, 1995). The BIC provides an approximation to a Bayesian hypothesis test, does not require the specification of priors, and can be easily calculated from SPSS output.

- Gelman (statmodeling.stat.columbia.edu) proposes to not limit ourselves to a single summary statistic or threshold
  - "I put much of the blame on statistical education, for two reasons"
  - "First [...] we typically focus on the choice of sample size, not on the importance of valid and reliable measurements."
  - "Second, it seems to me that statistics is often sold as a sort of alchemy that transmutes randomness into certainty, an *uncertainty laundering* [...] Just try publishing a result with p = 0.20"
  - "In summary, I agree with most of the ASA's statement on p-values but I feel that the problems are deeper, and that the solution is not to reform p-values or to replace them with some other statistical summary or threshold, but rather to move toward a greater acceptance of uncertainty and embracing of variation."

How to go Bayesian in model selection: test two models...

- The parameter θ might be predicted by two models M_0 and M_1: $P(\theta|\vec{x},M) = \frac{P(\vec{x}|\theta,M)P(\theta|M)}{P(\vec{x}|M)}$
  - A step further than yesterday in writing down the Bayes theorem: now multiple conditioning
  - $P(\vec{x}|M) = \int P(\vec{x}|\theta,M)P(\theta|M)d\theta$: *Bayesian evidence* or *model likelihood*
- Posterior for M_0: $P(M_0|\vec{x}) = \frac{P(\vec{x}|M_0)\pi(M_0)}{P(\vec{x})}$
- Posterior for M_1: $P(M_1|\vec{x}) = \frac{P(\vec{x}|M_1)\pi(M_1)}{P(\vec{x})}$
- The *odds* indicate relative preference of one model over the other
- Posterior odds: $\frac{P(M_0|\vec{x})}{P(M_1|\vec{x})} = \frac{P(\vec{x}|M_0)\pi(M_0)}{P(\vec{x}|M_1)\pi(M_1)}$
  - Posterior odds = Bayes Factor $\times$ prior odds
- $B_{01} := \frac{P(\vec{x}|M_0)}{P(\vec{x}|M_1)}$
- Various slightly different scales for the Bayes Factor
  - Interesting: deciban, unit supposedly theorized by Turing (according to IJ Good) as *the smallest change of evidence human mind can discern*

Jeffreys

K	dHart	bits	Strength of evidence
$< 10^0$	0	—	Negative (supports M_2)
10^0 to $10^{1/2}$	0 to 5	0 to 1.6	Barely worth mentioning
$10^{1/2}$ to 10^1	5 to 10	1.6 to 3.3	Substantial
10^1 to $10^{3/2}$	10 to 15	3.3 to 5.0	Strong
$10^{3/2}$ to 10^2	15 to 20	5.0 to 6.6	Very strong
$> 10^2$	> 20	> 6.6	Decisive

Kass and Raftery

$\log_{10} K$	K	Strength of evidence
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

Trotta

|lnB|	relative odds	favoured model's probability	Interpretation
< 1.0	< 3:1	< 0.750	not worth mentioning
< 2.5	< 12:1	0.923	weak
< 5.0	< 150:1	0.993	moderate
> 5.0	> 150:1	> 0.993	strong

Images from Wikipedia and from Roberto Trotta, Chair Lemaitre Lectures 2018

UCLouvain
Institut de recherche
en mathématique et physique

Bayesian model comparison of 193 models
Higgs inflation as reference model

$\ln(\mathcal{E}/\mathcal{E}_{\rm HI})$

Martin,RT+14

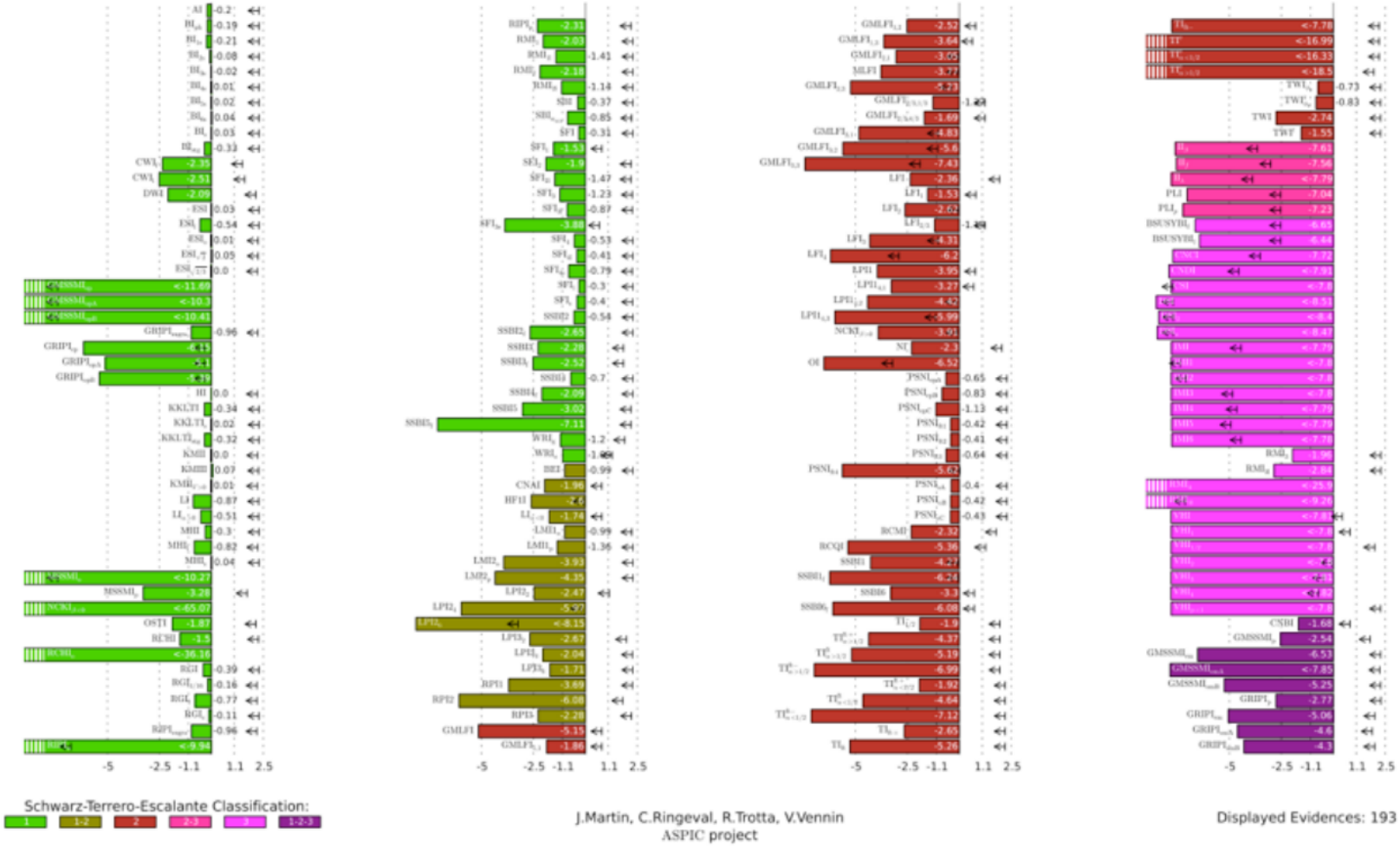
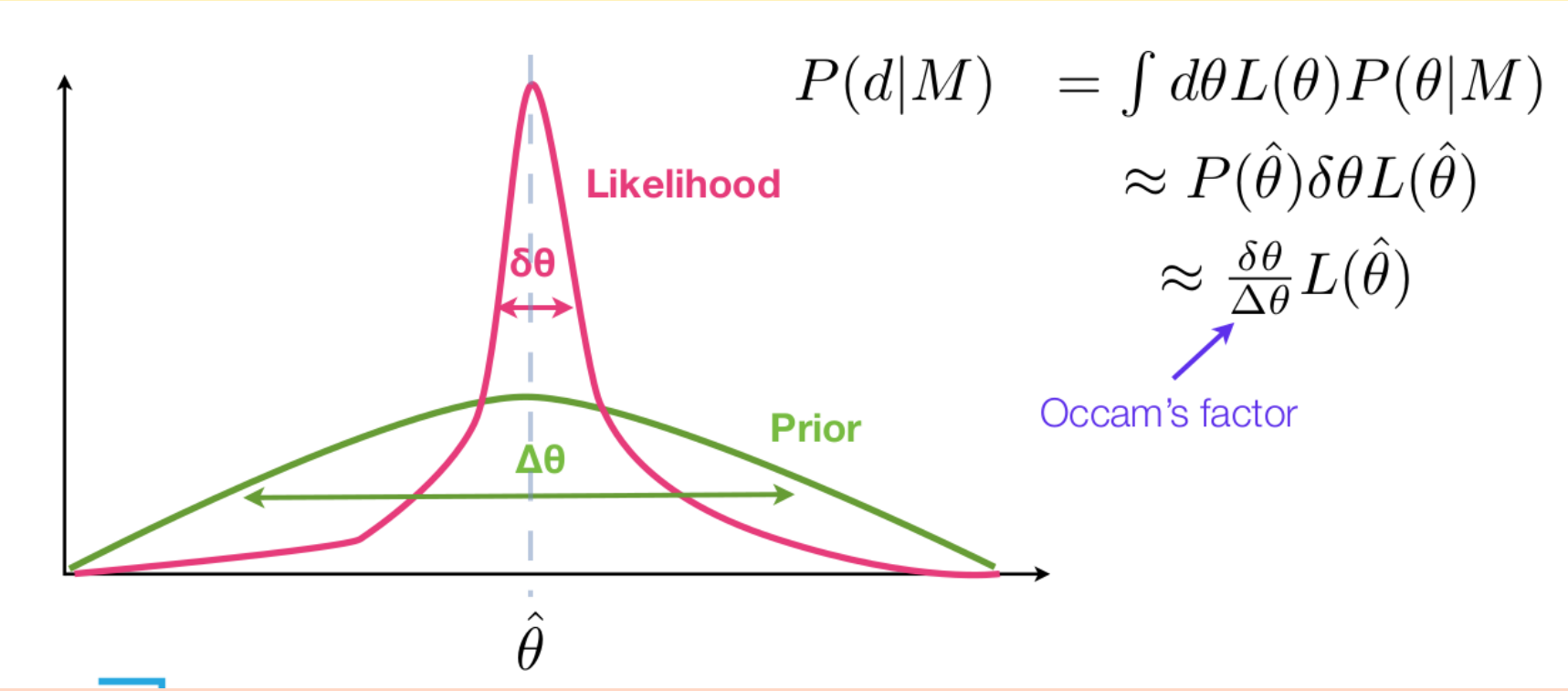


Image from Roberto Trotta, Chair Lemaitre Lectures 2018

Bayesian model selection — Discourage nonpredictive models

- The Bayes Factor also takes care of penalizing excessive model complexity
- Highly predictive models are rewarded, broadly-non-null priors are penalized



From Roberto Trotta, Chair Lemaitre Lectures 2018

Bayes vs p-values: the Jeffreys-Lindley paradox

- Data X (N data sampled from $f(x|\theta)$)<
  - H_0:$\theta = \theta_0$. Prior: π_0 (non-zero for point mass, Dirac's δ, counting measure)
  - H_1: $\theta ! = \theta_0$. Prior: $\pi_1 = 1 - \pi_0$ (usual Lebesgue measure)
- Conditional on H_1 being true:
  - Prior probability density $g(\theta)$
  - If $f(x|\theta) \sim Gaus(\theta, \sigma^2)$, then the sample mean $\bar{X} \sim Gaus(\theta, \sigma_{tot} = \sigma/N)$
- Likelihood ratio of H_0 to best fit for H_1: $\lambda = \frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\hat{\theta})} = exp(-Z^2/2) \propto \frac{\sigma_{tot}}{\tau} B_{01}$; $Z := \frac{\hat{\theta}-\theta_0}{\sigma_{tot}}$
  - λ disfavours the null hypothesis for large significances (small p-values), independent of sample size
  - B_{01} includes σ_{tot}/τ (Ockham Factor, penalizing H_1 for imprecise determination of θ), sample dependent!
- For arbitrarily large Z (small p-values), λ disfavours H_0, while there is always a N for which B_{01} favours H_0 over H_1

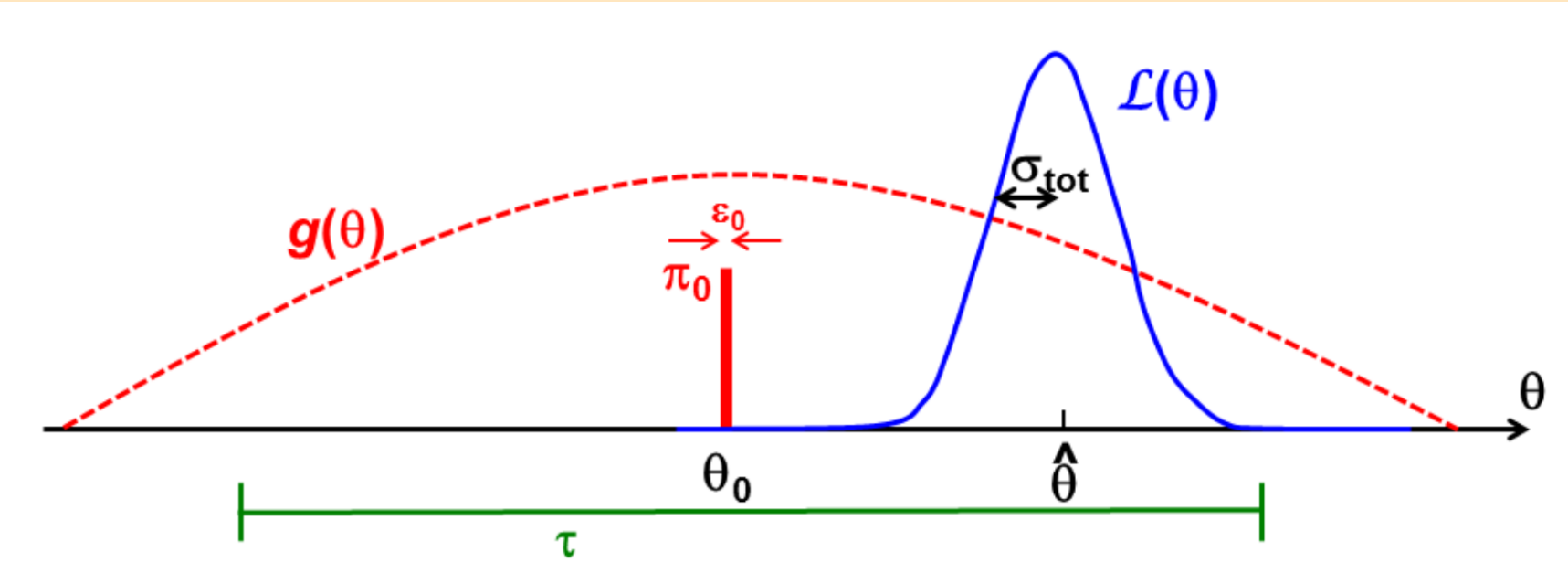


Image from Cousins, doi:10.1007/s11229-014-0525-z

Reproducibility crysis: is it a thing?

- It seems so: The Bayer Study (https://www.nature.com/articles/nrd3545)

Published: 31 August 2011

Reliability of 'new drug target' claims called into question

Asher Mullard

Nature Reviews Drug Discovery **10**, 643–644(2011) | Cite this article

Bayer halts nearly two-thirds of its target-validation projects because in-house experimental findings fail to match up with published literature claims, finds a first-of-a-kind analysis on data irreproducibility.

- “Irreproducibility was high both when Bayer scientists applied the same experimental procedures as the original researchers and when they adapted their approaches to internal needs (for example, by using different cell lines).”
- “High-impact journals did not seem to publish more robust claims, and, surprisingly, the confirmation of any given finding by another academic group did not improve data reliability.”

The funny bit

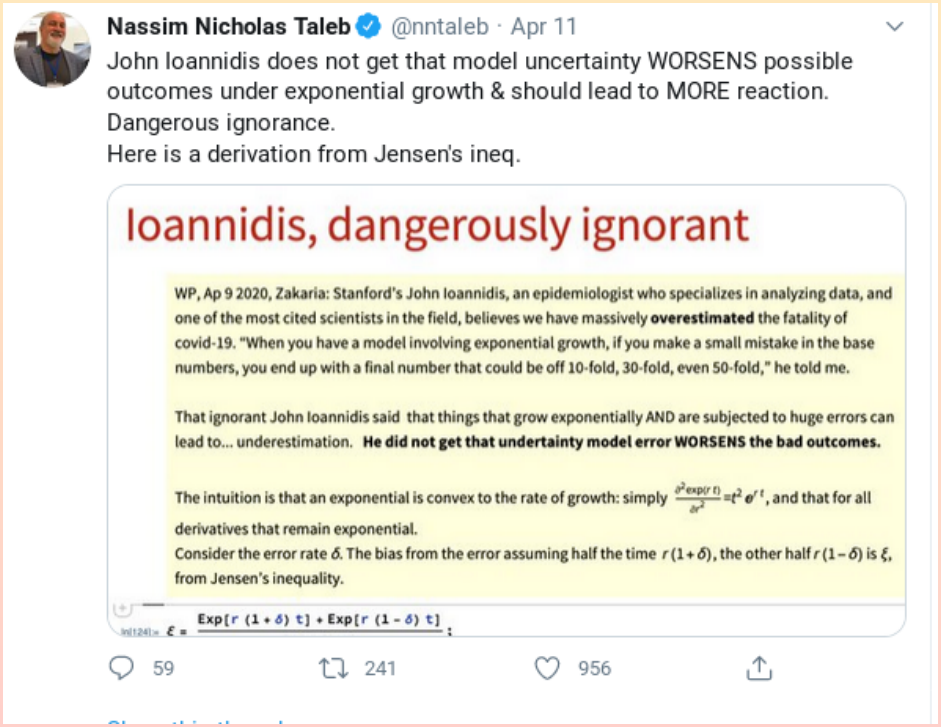
- Ioannidis (doi:/10.1371/journal.pmed.0020124) identifies several causes mostly linked to scientists' own biases
  - Investigator prejudice, incorrect statistical methods, competition in hot fields, publishing bias

Population-level COVID-19 mortality risk for non-elderly individuals overall and for non-elderly individuals without underlying diseases in pandemic epicenters

John P. A. Ioannidis, Cathrine Axfors, Despina G. Contopoulos-Ioannidis

doi: https://doi.org/10.1101/2020.04.05.20054361

- Then Ioannidis got accused of the same issues, just last month

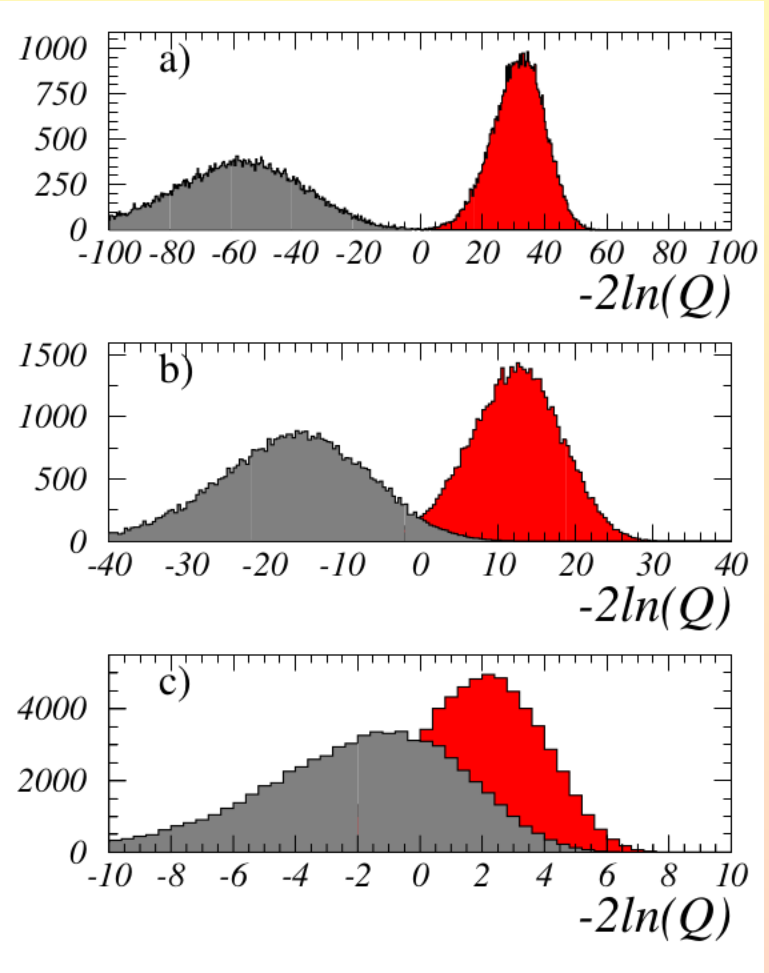


How HEP protects itself

- Goal: seamless transition between exclusion, observation, discovery (historically for the Higgs)
  - Exclude Higgs as strongly as possible in its absence (in a region where we would be sensitive to its presence)
  - Confirm its existence as strongly as possible in its presence (in a region where we are sensitive to its presence)
  - Maintain Type I and Type II errors below specified (small) levels
- Identify observables, and a suitable test statistic Q
- Define rules for exclusion/discovery, i.e. ranges of values of Q leading to various conclusions
  - Specify the significance of the statement, in form of confidence level (CL)
- Confidence limit: value of a parameter (mass, xsec) excluded at a given confidence level CL
  - A confidence limit is an upper(lower) limit if the exclusion confidence is greater(less) than the specified CL for all values of the parameter below(above) the confidence limit
- The resulting intervals are neither frequentist nor bayesian!

Get your confidence levels right

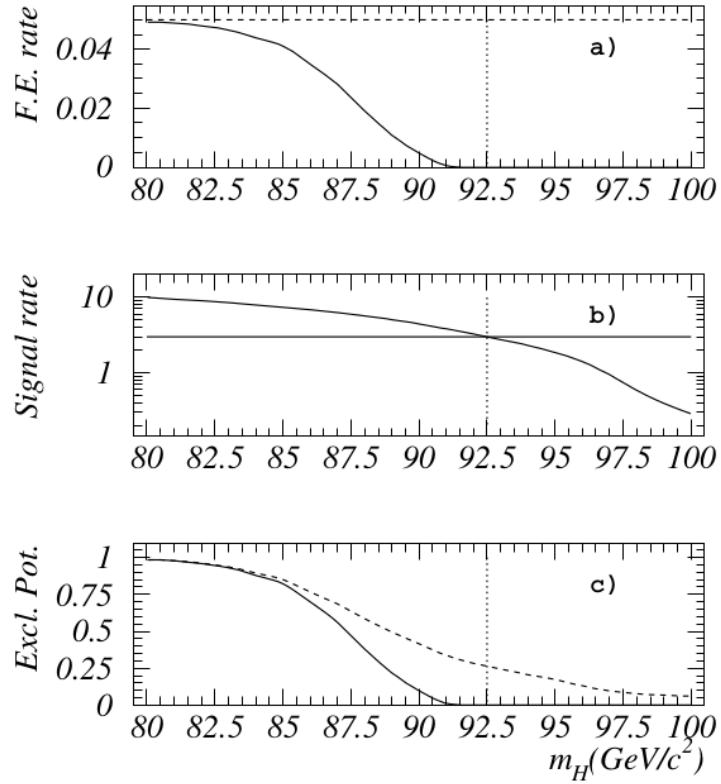
- Find a monotonic Q for increasing signal-like experiments (e.g. likelihood ratio)
- $CL_{s+b} = P_{s+b}(Q \leq Q_{obs})$
  - Small values imply poor compatibility with $S + B$ hypothesis, favouring B-only
- $CL_b = P_b(Q \leq Q_{obs})$
  - Large (close to 1) values imply poor compatibility with B-only, favouring $S + B$
- What to do when the estimated parameter is unphysical?
  - The same issue solved by Feldman-Cousins
  - If there is also underfluctuation of backgrounds, it's possible to exclude even zero events at 95%CL!
  - It would be a statement about future experiments
  - Not enough information to make statements about the signal
- Normalize the $S + B$ confidence level to the B-only confidence level!



Plot from Read, CERN-open-2000-205

Avoid issues at low signal rates

- $CL_s := \frac{CL_{s+b}}{CL_b}$
- Exclude the signal hypothesis at confidence level CL if $1 - CL_s \leq CL$
- Ratio of confidences is not a confidence
  - The hypotetical false exclusion rate is generally less than the nominal $1 - CL$ rate
  - CL_s and the actual false exclusion rate grow more different the more $S + B$ and B p.d.f. become similar
- CL_s increases coverage, i.e. the range of parameters that can be exclude is reduced
  - It is more conservative
  - Approximation of the confidence in the signal hypothesis that might be obtained if there was no background
- Avoids the issue of CL_{s+b} with experiments with the same small expected signal
  - With different backgrounds, the experiment with the larger background might have a better expected performance
- Formally corresponds to have $H_0 = H(\theta! = 0)$ and test it against $H_1 = H(\theta = 0)$
  - Test inversion!

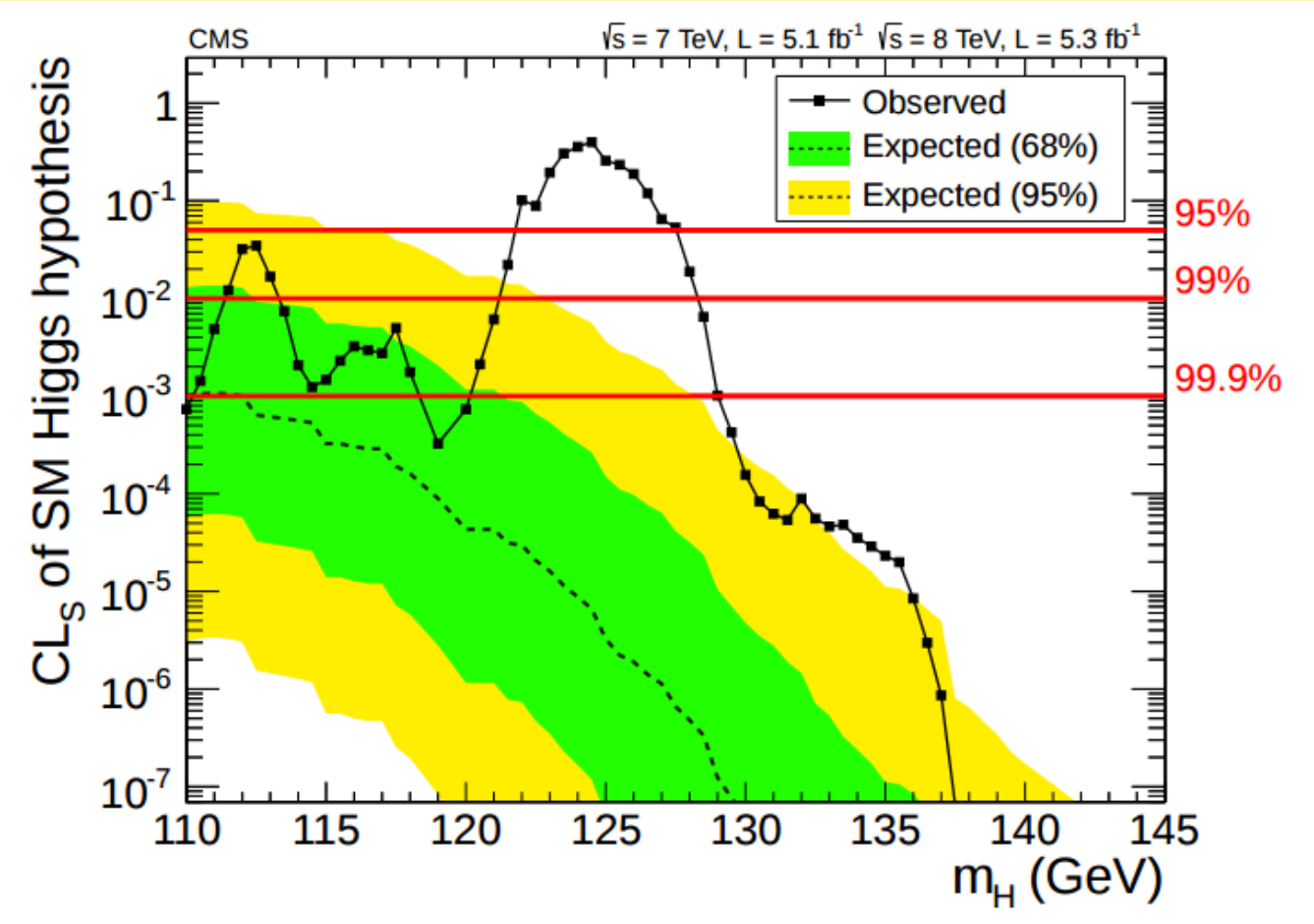


Dashed: CL_{s+b}
Solid: CL_s
$S < 3$: exclusion for a B-free search $\equiv 0$

Plot from Read, CERN-open-2000-205

That's what we used for the Higgs discovery!

- Apply the CL_s method to each Higgs mass point
- Green/yellow bands indicate the $\pm 1\sigma$ and $\pm 2\sigma$ intervals for the expected values under B-only hypothesis
  - Obtained by taking the quantiles of the B-only hypothesis



Plot from Higgs discovery paper

Fluctuations in HEP? The proposal of a 5σ criterion

UCLouvain
Institut de recherche en mathématique et physique

- Rosenfeld, 1968 (https://escholarship.org/uc/item/6zm2636q) *Are there any Far-out Mesons or Baryons?*
  - "In summary of all the discussion abouve, I conclude that each of our 150,000 annual histograms is capable of generating somewhere between 10 and 100 deceptive upward fluctuations [...] (we) should expect several 4σ and hundreds of 3σ fluctuations"

of 3σ fluctuations. What are the implications? To the theoretician or phenomenologist the moral is simple; wait for nearly 5σ effects. For the experimental group who have just spent a year of their time and perhaps a million dollars, the problem is harder. I suggest that they should go ahead and publish their tantalizing bump (or at least circulate it as a report.) But they should realize that any bump less than about 5σ constitutes only a call for a repeat of the experiment. If they, or somebody else, can double the number of counts, the number of standard deviations should increase by $\sqrt{2}$, and that will confirm the original effect.

My colleague Gerry Lynch has instead tried to study this problem "experimentally" using a "Las Vegas" computer program called Game. Game is played as follows. You wait until an unsuspecting "friend" comes to show you his latest 4σ peak. You draw a smooth curve through his data (based on the hypothesis that the peak is just a fluctuation), and punch this smooth curve as one of the inputs for game. The other input is his actual data. If you then call for 100 Las Vegas histograms, Game will generate them, with the actual data reproduced for comparison at some random page. You and your friend then go around the halls, asking physicists to pick out the most surprising histogram in the printout. Often it is one of the 100 phoneys, rather than the real "4σ" peak. Figure 3 shows two Game histograms, each one being one of the more interesting ones in a run of 100. The smooth curves drawn through them are of course absurd; they are supposed to be the background estimates of the inexperienced experimenter. But they do illustrate that a 2σ or 3σ fluctuation can easily be amplified to "4σ" or "5σ"; all it takes is a little enthusiasm.

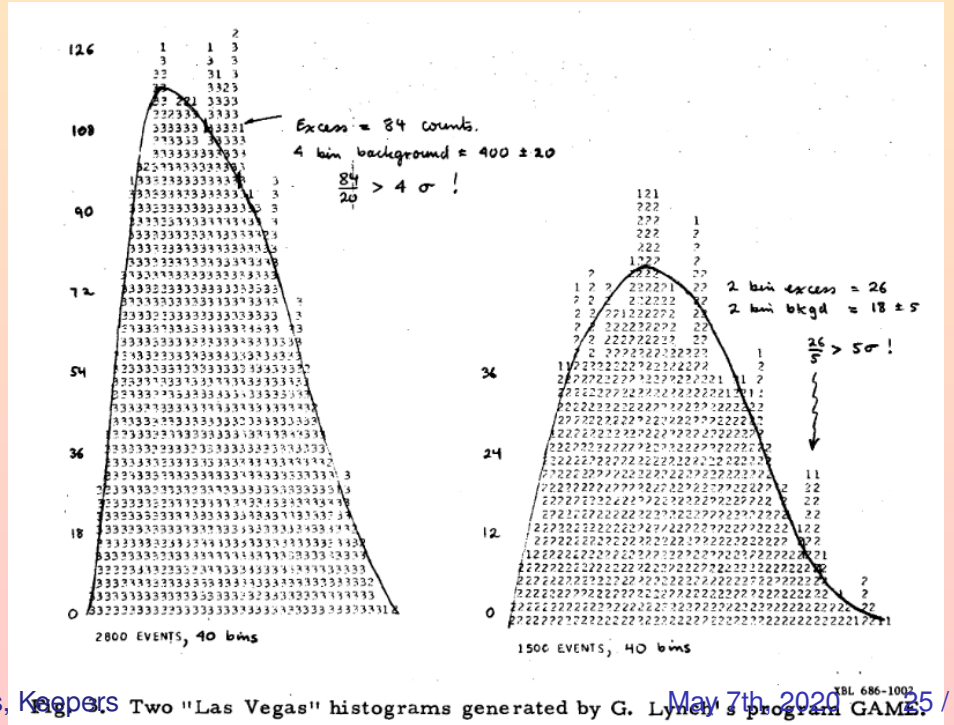
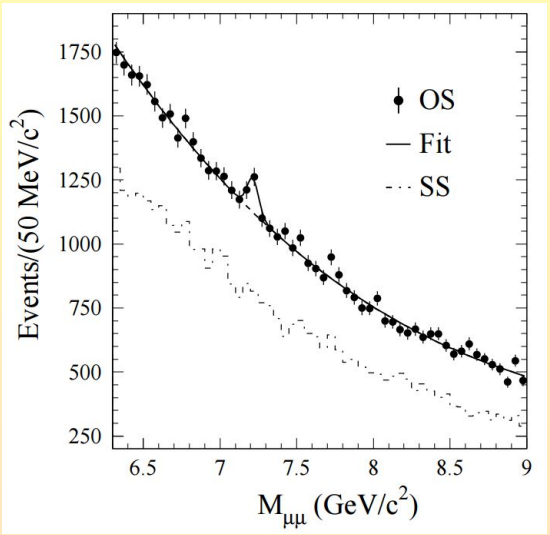
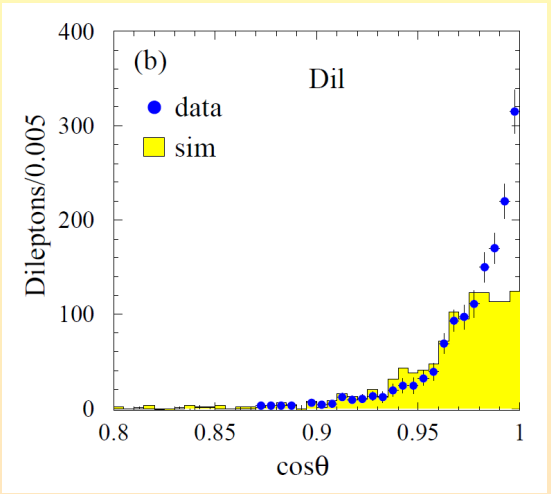


Fig. 3. Two "Las Vegas" histograms generated by G. Lynch's program GAME.

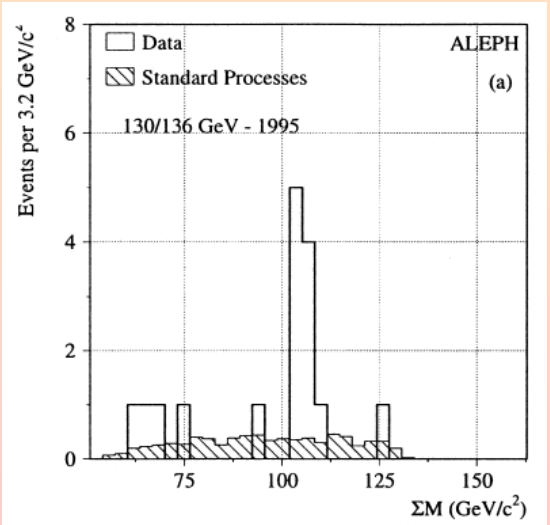
HEP has a history of unconfirmed effects

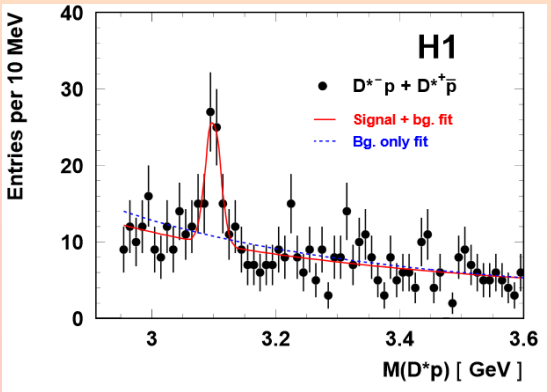
- 3.5σ (2005, CDF) in dimuon (candidate bottom squark, doi:/10.1103/PhysRevD.72.092003)





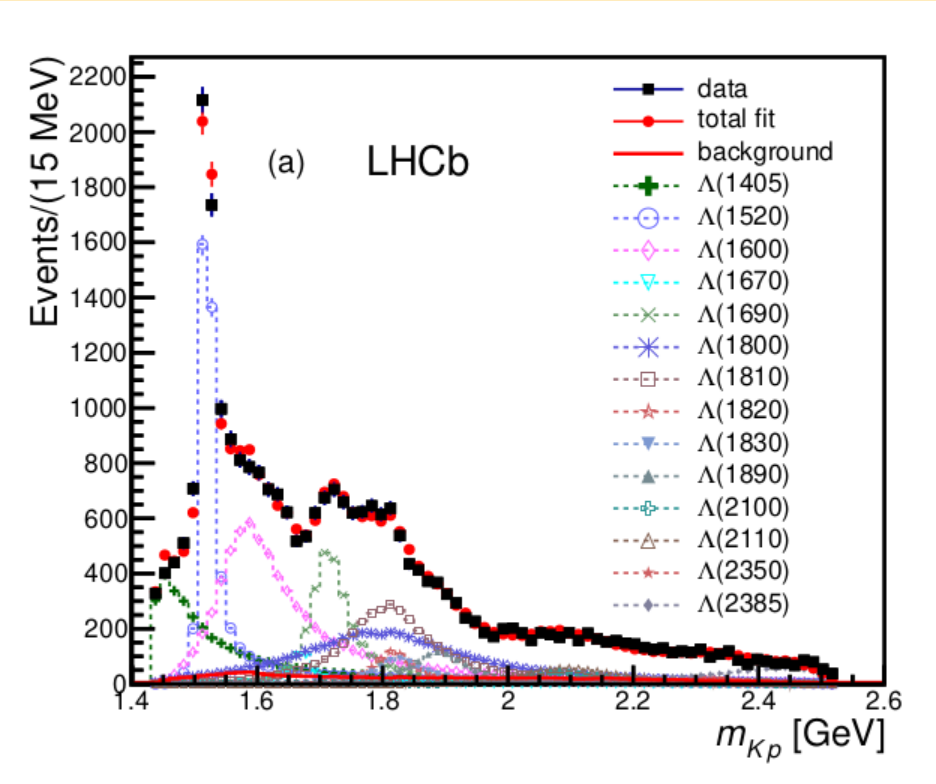
- $\sim 4\sigma$ (1996, Aleph) in four-jet (Higgs boson candidate, doi:/10.1007/BF02906976)
- 6σ (2004, H1) (narrow $\bar{c}$ baryon state, doi:/10.1016/j.physletb.2004.03.012)
  - H1 speaks of “Evidence”, not confirmed.

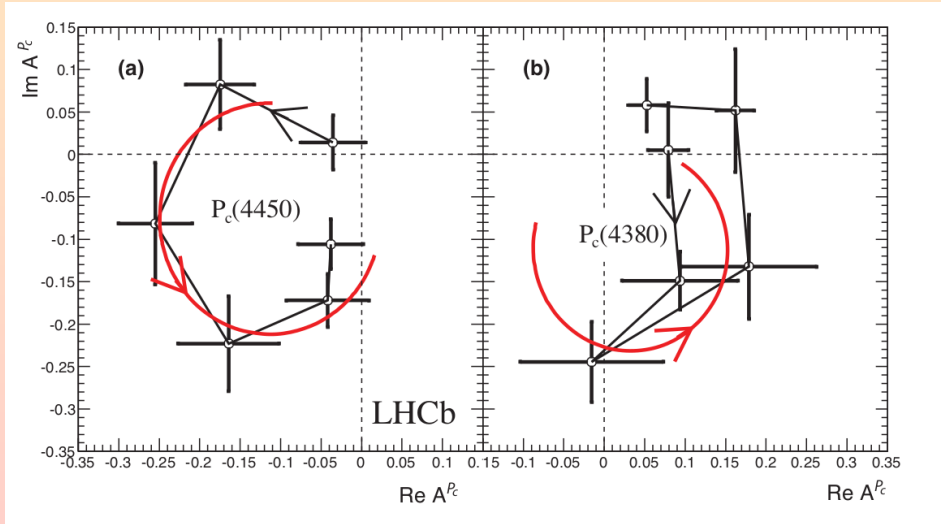




The revenge of the pentaquarks

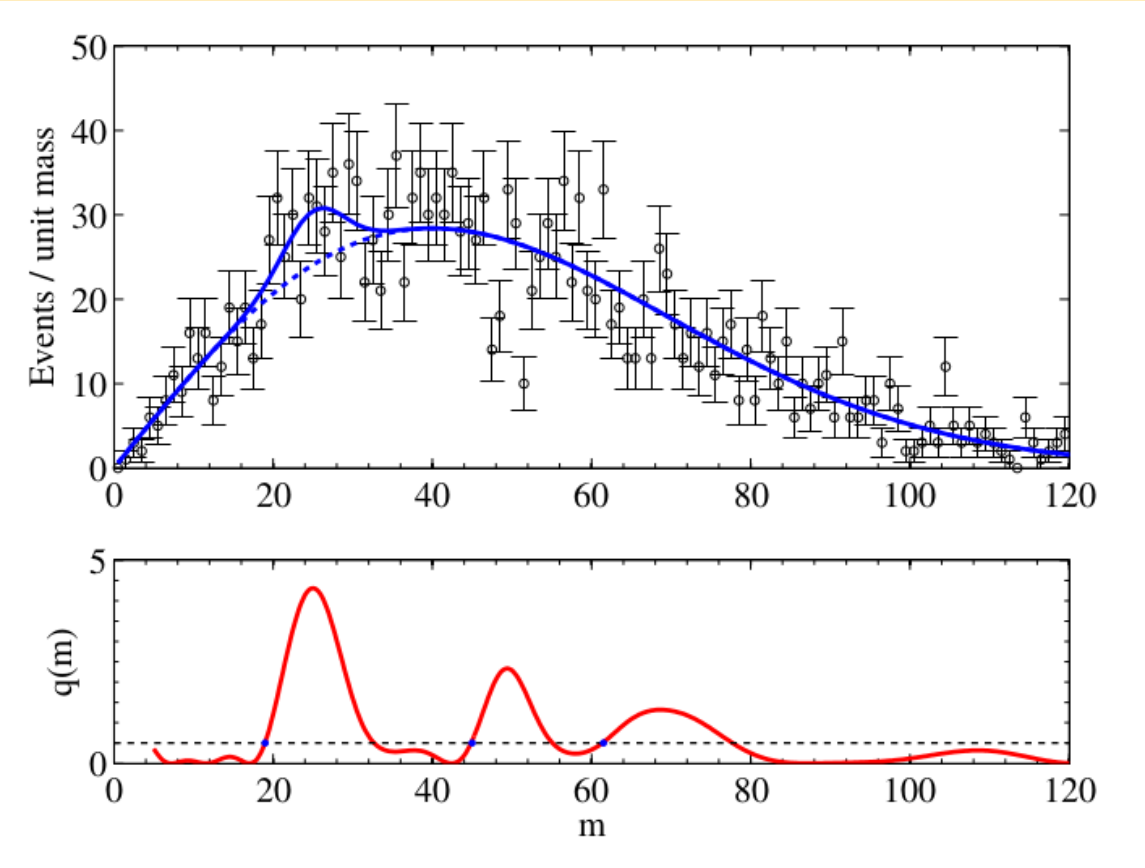
- 9σ and 12σ (2015, LHCb): pentaquarks! (doi:/10.1103/PhysRevLett.115.072001)
  - Several cross-checks (fit to mass spectrum, fit with non-resonant components, evolution of complex amplitute in Argand diagrams)
  - Mass measurement, soft statement: “Interpreted as resonant states they must have minimal quark content of $ccuud$, and would therefore be called charmonium-pentaquark states.
- One remark: quoting significances above about 5–6σ is meaningless
  - Asymptotic approximation not trustable (tail effects). Can run lots of toys but...
  - ...cannot possibly trust knowing your systematic uncertainties to that level





The Look-elsewhere effect — 1

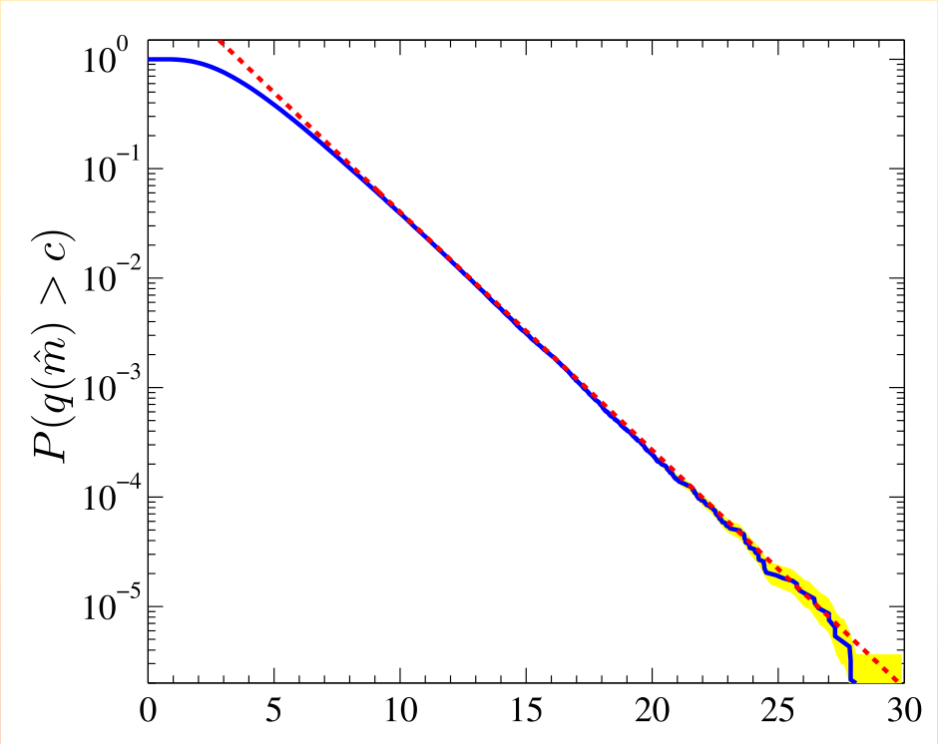
- Searching for a resonance X of arbitrary mass
  - $H_0 =$ no resonance, the mass of the resonance is not defined (Standard Model)
  - $H_1 = H(M \neq 0)$. There are many possible values of M
- Wilks theorem not valid anymore, no unique test statistic encompassing every possible H_1
- Quantify the compatibility of an observation with the B-only hypothesis
  - $q_0(\hat{m_X}) = \max_{m_X} q_0(m_X)$
  - Write a global p-value as $p_b^{global} := P(q_0(\hat{m_X}) > u) \leq \langle N_u \rangle + \frac{1}{2} P_{\chi^2_1}(u)$
  - u fixed confidence level
  - Crossings (Davis, Biometrika 74, 33–43 (1987)) , computable using pseudo-data (toys)

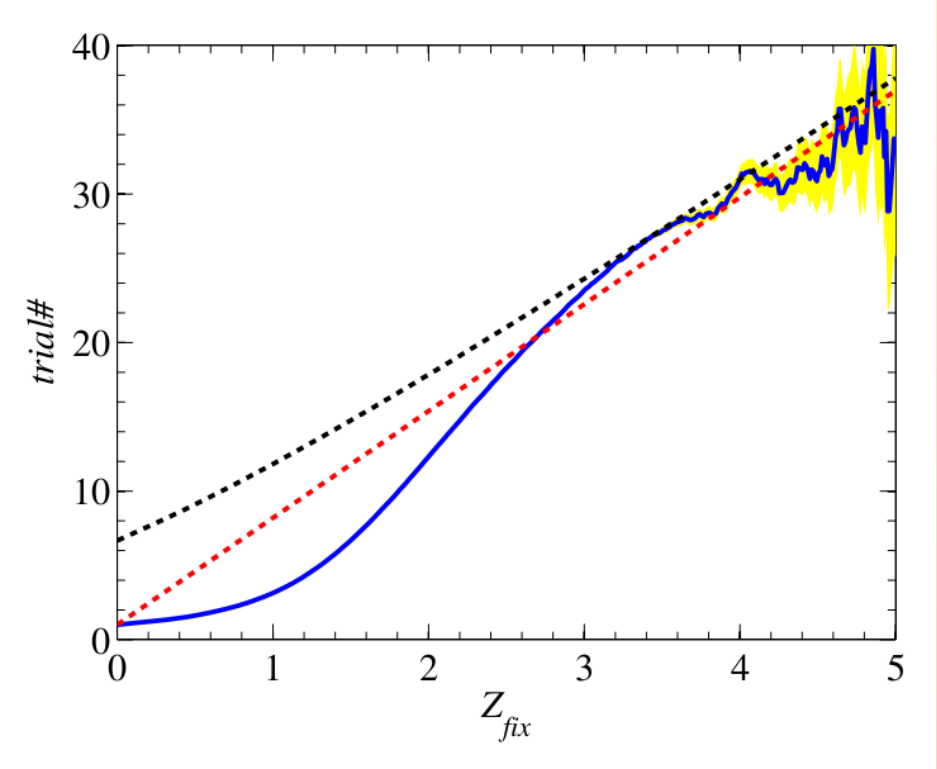


Plot from Gross-Vitells, 10.1140/epjc/s10052-010-1470-8

The Look-elsewhere effect — 2

- Ratio of local (excess right here) and global (excess anywhere) p-values: trial factor
- Asymptoticly linear in the number of search regions and in the fixed significance level
  - Dashed red lines: prediction based on the formula with upcrossings
  - Blue: 10^6 toys (pseudoexperiments)
- Here *asymptotic* means *for increasingly smaller tail probabilities*

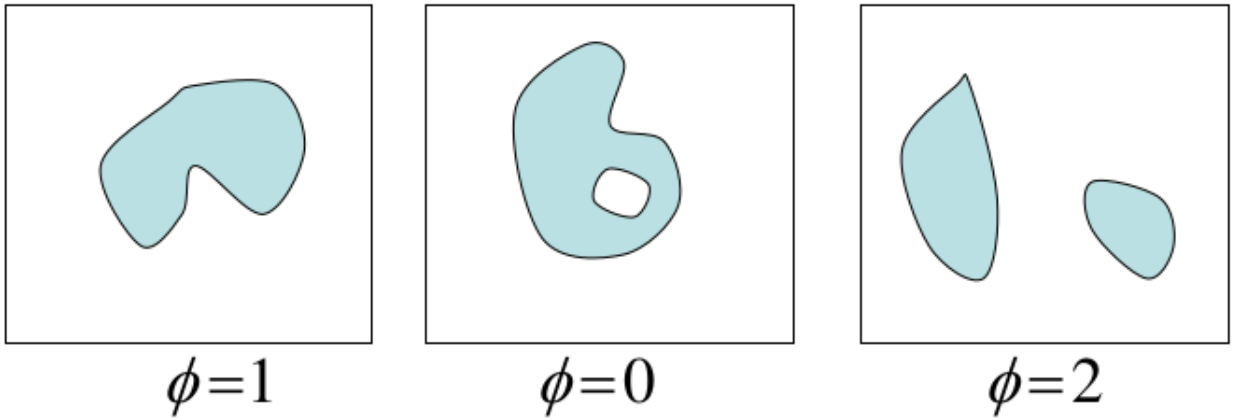


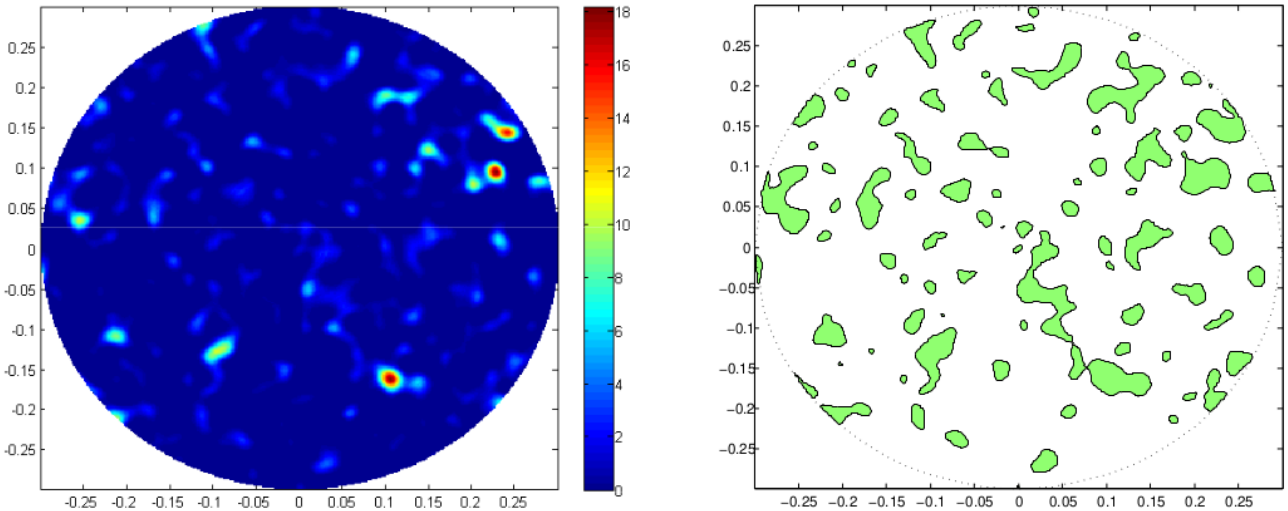


Plot from Gross-Vitells, 10.1140/epjc/s10052-010-1470-8

The Look-elsewhere effect, now also in 2D — 1

- Extension to two dimensions requires using the theory of random fields
  - Excursion set: set of points for which the value of a field is larger than a threshold u
  - Euler characteristics interpretable as number of disconnected regions minus number of holes

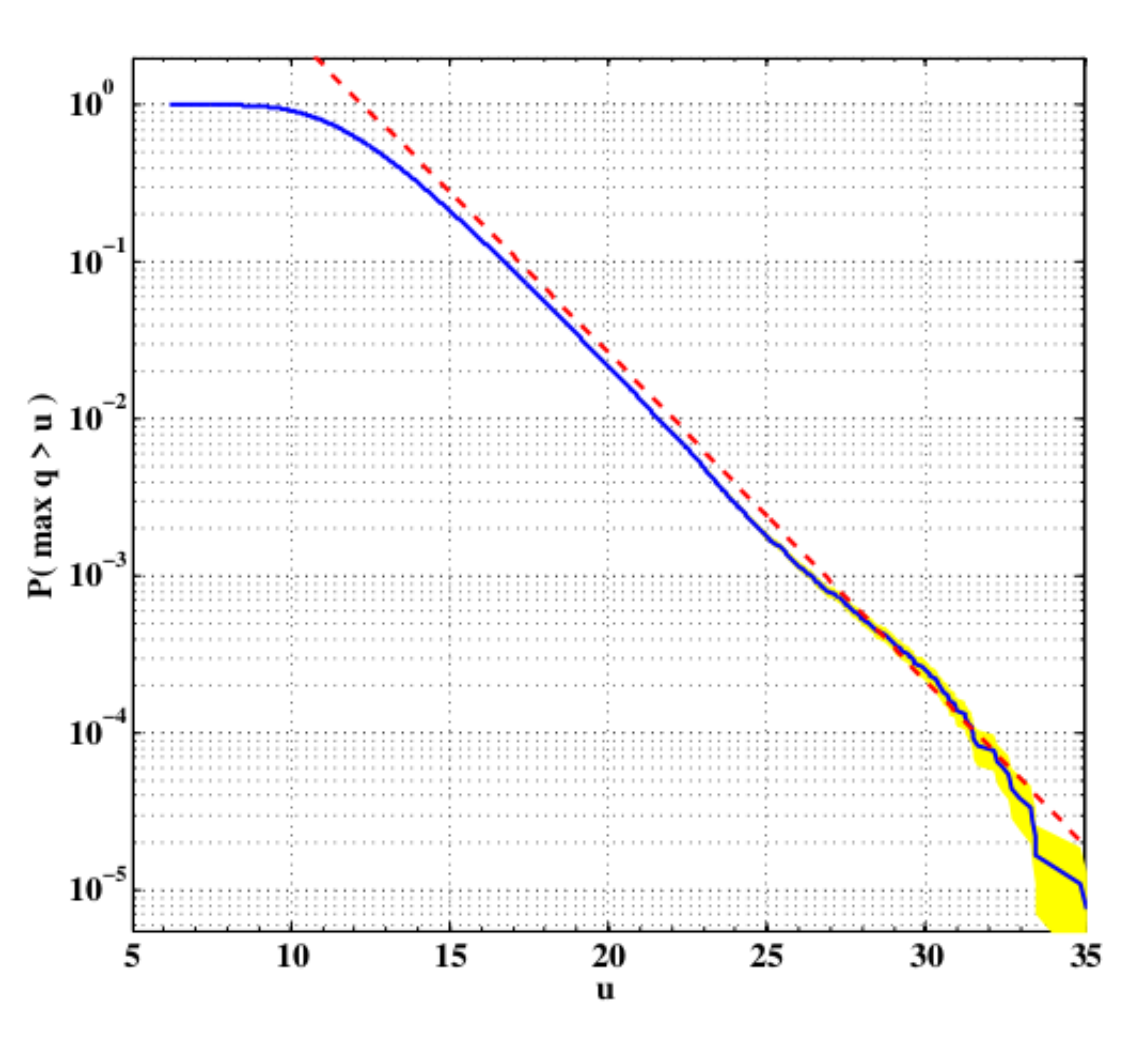




Plot from Gross-Vitells, 10.1016/j.astropartphys.2011.08.005

The Look-elsewhere effect, now also in 2D — 2

- Asymptoticity holds also for the 2D effect, as desired
  - Dashed red lines: prediction based on the formula with upcrossings
  - Blue: $200k$ toys (pseudoexperiments)

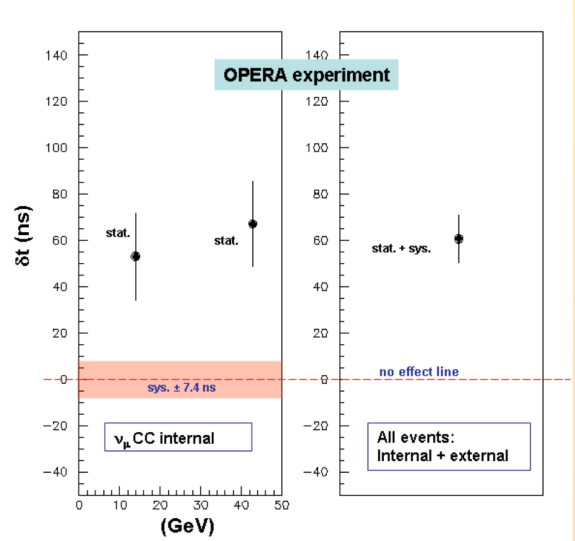


Plot from Gross-Vitells, 10.1016/j.astropartphys.2011.08.005

When there is no LEE, you still need to make sure your systematics are right

- In 2011 OPERA (arXiv:1109.4897v1) reported superluminal neutrino speed, with 6.0σ significance...

> An early arrival time of CNGS muon neutrinos with respect to the one computed assuming the speed of light in vacuum of (60.7 ± 6.9 (stat.) ± 7.4 (sys.)) ns was measured. This anomaly corresponds to a relative difference of the muon neutrino velocity with respect to the speed of light $(v\text{-}c)/c$ = (2.48 ± 0.28 (stat.) ± 0.30 (sys.)) $\times 10^{-5}$.



- ...but they had a loose cable connector (doi:/10.1007/JHEP10(2012)093)

> After several months of additional studies, with the new results reported in this paper, the OPERA Collaboration has completed the scrutiny of the originally reported neutrino velocity anomaly by identifying its instrumental sources and coming to a coherent interpretation scheme.

Deborah Mayo's Severe Testing

- Frequentist testing based on Type I and Type 2 error rates (D. Mayo "Statistical Inference as Severe Testing". Cambridge UP, 2018.)
  - Point-null avoided by considering $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$
- Generalize to test $\mu_1 = (\mu_0 + \gamma)$, $\gamma \geq 0$
- Severe interpretation of negative results (SIN)
  - When H_0 not rejected, define severity
    $SEV(\mu \leq \mu_1) = P(Q > Qobs; \mu \leq \mu_1 false) = P(Q > Qobs; \mu > \mu_1) > P(Q > Qobs; \mu = \mu_1)$
  - Low severity: your test is not capable of detecting a discrepancy even when if it existed, therefore when not detected is's poor evidence of its absence (low power)
  - High severity: your test is highly capable of detecting a discrepancy if it existed, therefore when not detected is a good indication of its absence (high power)
- Severe interpretation of rejection (SIR)
  - When H_0 rejected, define severity
    $SEV(\mu > \mu_1) = P(Q \leq Qobs; \mu > \mu_1 false) = P(Q \leq Qobs; \mu \leq \mu_1) > P(Q \leq Qobs; \mu = \mu_1)$
  - Low severity: if probability of higher-than-observed $Qobs$ is fairly high, then $Qobs$ not a good indication of effect
  - High severity: if probability of smaller-than-observed $Qobs$ is very high, then such a large $Qobs$ indicates a real effect
- Cousins (arXiv:2002.09713) seems to argue that current CL HEP practice is substantially equivalent to Mayo's severe testing
  - Very specific to HEP. Other disciplines should be worried, instead

- Box (https://www.jstor.org/stable/2286841) warns that any model is an approximation

2.3 Parsimony

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

2.4 Worrying Selectively

Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.

Truth and models: HEP is special

- Cousins (doi:/10.1007/s11229-014-0525-z) notes HEP is in a privileged position when compared with social or medical sciences

5 HEP and belief in the null hypothesis

At the heart of the measurement models in HEP are well-established equations that are commonly known as "laws of nature". By some historical quirks, the current "laws" of elementary particle physics, which have survived several decades of intense scrutiny with only a few well-specified modifications, are collectively called a "model", namely the Standard Model (SM). In this review, I refer to the equations of

There is a deeper point to be made about core physics models concerning the difference between a model being a good "approximation" in the ordinary sense of the word, and the concept of a mathematical limit. The equations of Newtonian physics have been superseded by those of special and general relativity, but the earlier equations are not just approximations that did a good job in predicting (most) planetary orbits; they are the correct *mathematical limits* in a precise sense. The kinematic

relationships. Nevertheless, whatever new physics is added, we also expect that the SM will remain a correct mathematical limit, or a correct effective field theory, within a more inclusive theory. It is in this sense of being the correct limit or correct effective field theory that physicists believe that the SM is "true", both in its parts and in the collective whole. (I am aware that there are deep philosophical questions about reality, and that this point of view can be considered "naive", but this is a point of view that is common among high energy physicists.)

- Others (Gelman, Raftery, Berger, Bernardo) argue that a point null is impossible (at most "small")

Ground truth, models, and point nulls

- I think a point or almost-point null is related to our simplifications rather than with a claim on reality
- Some disciplines deal with phenomena which cannot (yet) be explained from first principles
  - Maybe one day we will have a full quasi-deterministic model of a whole body or brain
  - Certainly so far most models are attempts at finding a functional form for the relationship between two variables
- Some disciplines (HEP) have to do with phenomena which can be explained from first principles
  - These principles are *reasonable* but not necessarily the best or the only possible ones
  - No guarantee that they reflect a universal truth
  - Arguing that the vast experimental agreement of the SM implies ground truth behaves based on our principles sounds a bit wishful thinking
  - What can be claimed is that the vast experimental agreement warrants the use of point or quasi-point nulls
- Box's view on models, and the Occam's Razor, should still lead considerations on model choices
  - A version of the Occam's Razor is even implemented in Bayesian model selection
- Still, to avoid interpreting fluctuations as real effects all disciplines should strive—when possible—to describe causal relationships rather than correlations

Generalizing expected values to functions of random variables

- Extend the concept of expected value to a generic function $g(X)$ of a random variable

$$E[g] := \int_{\Omega} g(X)f(X)dX \tag{1}$$

- The previous expression Eq. **??** is a special case of Eq. 1 when $g(X) = X$
- The mean of X is:

$$\mu := E[X] \tag{2}$$

- The variance of X is:

$$V(X) := E[(X-\mu)^2] = E[X^2] - (E[X])^2 = E[X^2] - \mu^2 \tag{3}$$

- Mean and variance will be our way of estimating a "central" value of a distribution and of the dispersion of the values around it

Let's make it funnier: more variables!

- Let our function $g(X)$ be a function of more variables, $\vec{X} = (X_1, X_2, ..., X_n)$ (with p.d.f. $f(\vec{X})$)
  - Expected value: $E(g(\vec{X})) = \int g(\vec{X}) f(\vec{X}) dX_1 dX_2 ... dX_n = \mu_g$
  - Variance: $V[g] = E\left[(g - \mu_g)^2\right] = \int (g(\vec{X}) - \mu_g)^2 f(\vec{X}) dX_1 dX_2 ... dX_n = \sigma_g^2$
- **Covariance:** of two variables X, Y:
  $V_{XY} = E\left[(X - \mu_X)(Y - \mu_Y)\right] = E[XY] - \mu_X \mu_Y = \int XY f(X, Y) dX dY - \mu_X \mu_Y$
  - It is also called "error matrix", and sometimes denoted $cov[X, Y]$
  - It is symmetric by construction: $V_{XY} = V_{YX}$, and $V_{XX} = \sigma_X^2$
  - To have a dimensionless parameter: correlation coefficient $\rho_{XY} = \frac{V_{XY}}{\sigma_X \sigma_Y}$

- V_{XY} is the expectation for the product of deviations of X and Y from their means
- If having $X > \mu_X$ enhances $P(Y > \mu_Y)$, and having $X < \mu_X$ enhances $P(Y < \mu_Y)$, then $V_{XY} > 0$: positive correlation!
- ρ_{XY} is related to the angle in a linear regression of X on Y (or viceversa)
  - It does not capture non-linear correlations

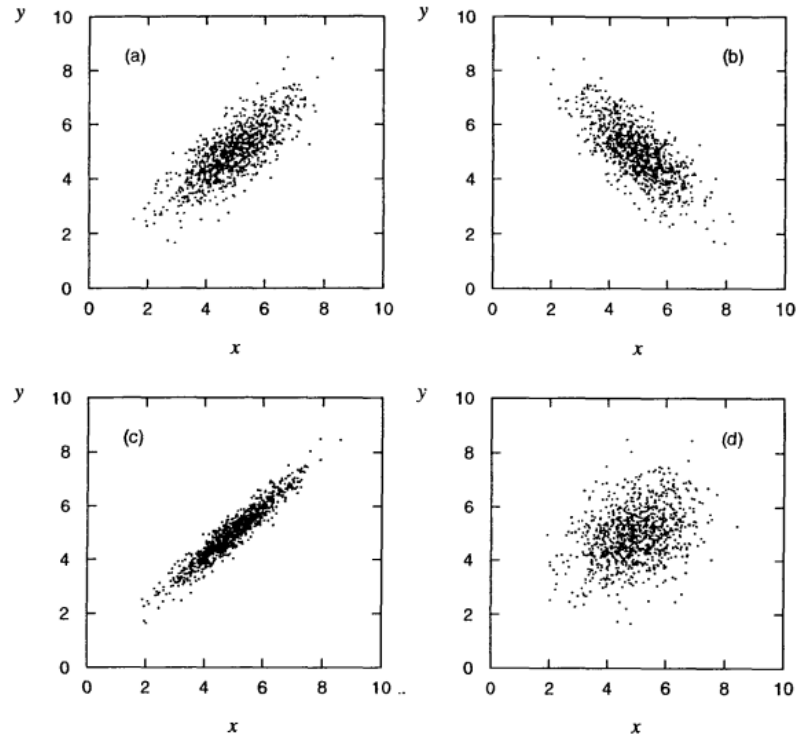


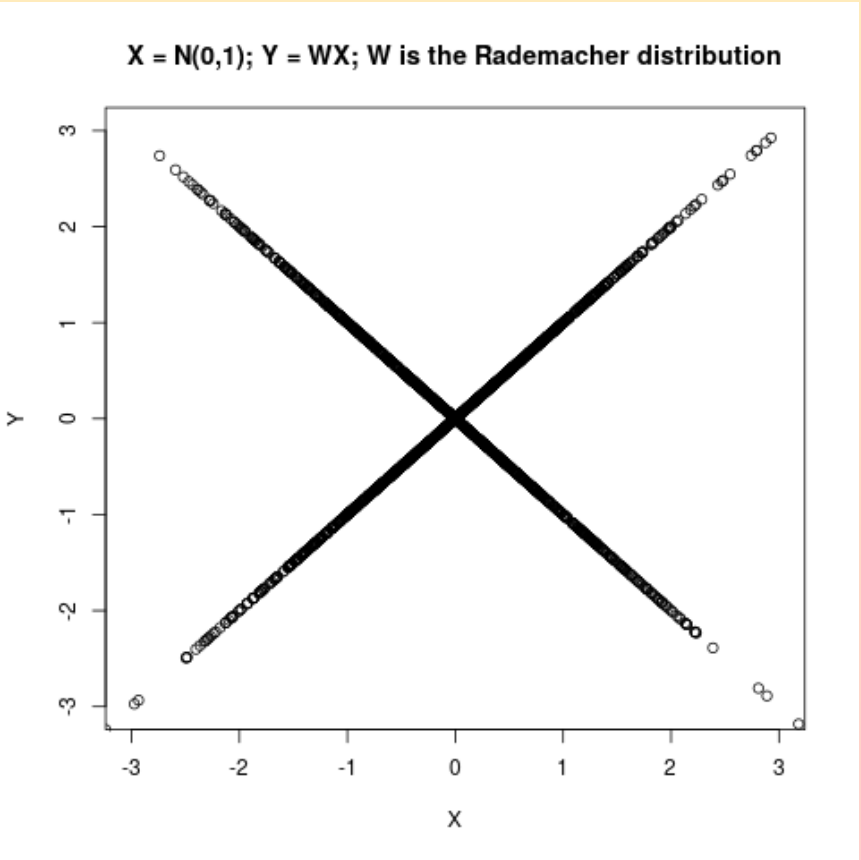
Fig. 1.9 Scatter plots of random variables x and y with (a) a positive correlation, $\rho = 0.75$, (b) a negative correlation, $\rho = -0.75$, (c) $\rho = 0.95$, and (d) $\rho = 0.25$. For all four cases the standard deviations of x and y are $\sigma_x = \sigma_y = 1$.

Take it to the next level: the Mutual Information

- Covariance and correlation coefficients act taking into account only linear dependences
- Mutual Information is a general notion of correlation, measuring the information that two variables X and Y share

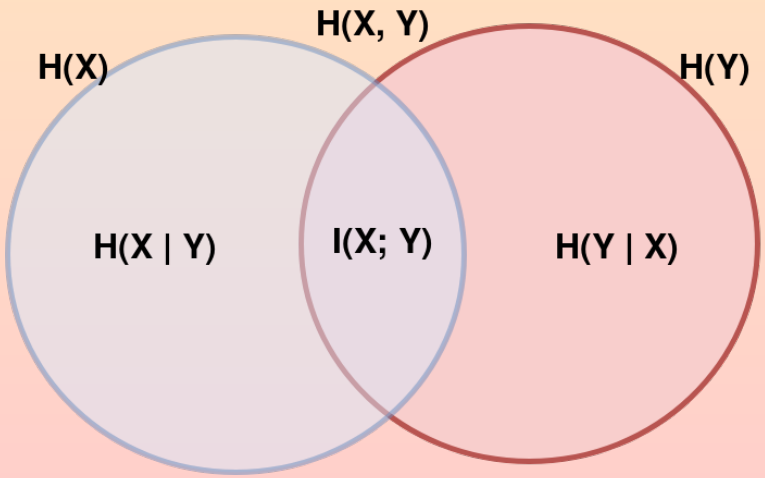
$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \left(\frac{p(x,y)}{p_1(x) p_2(y)} \right)$$

- Symmetric: $I(X;Y) = I(Y;X)$
- $I(X;Y) = 0$ if and only if X and Y are totally independent
  - X and Y can be uncorrelated but not independent; mutual information captures this!



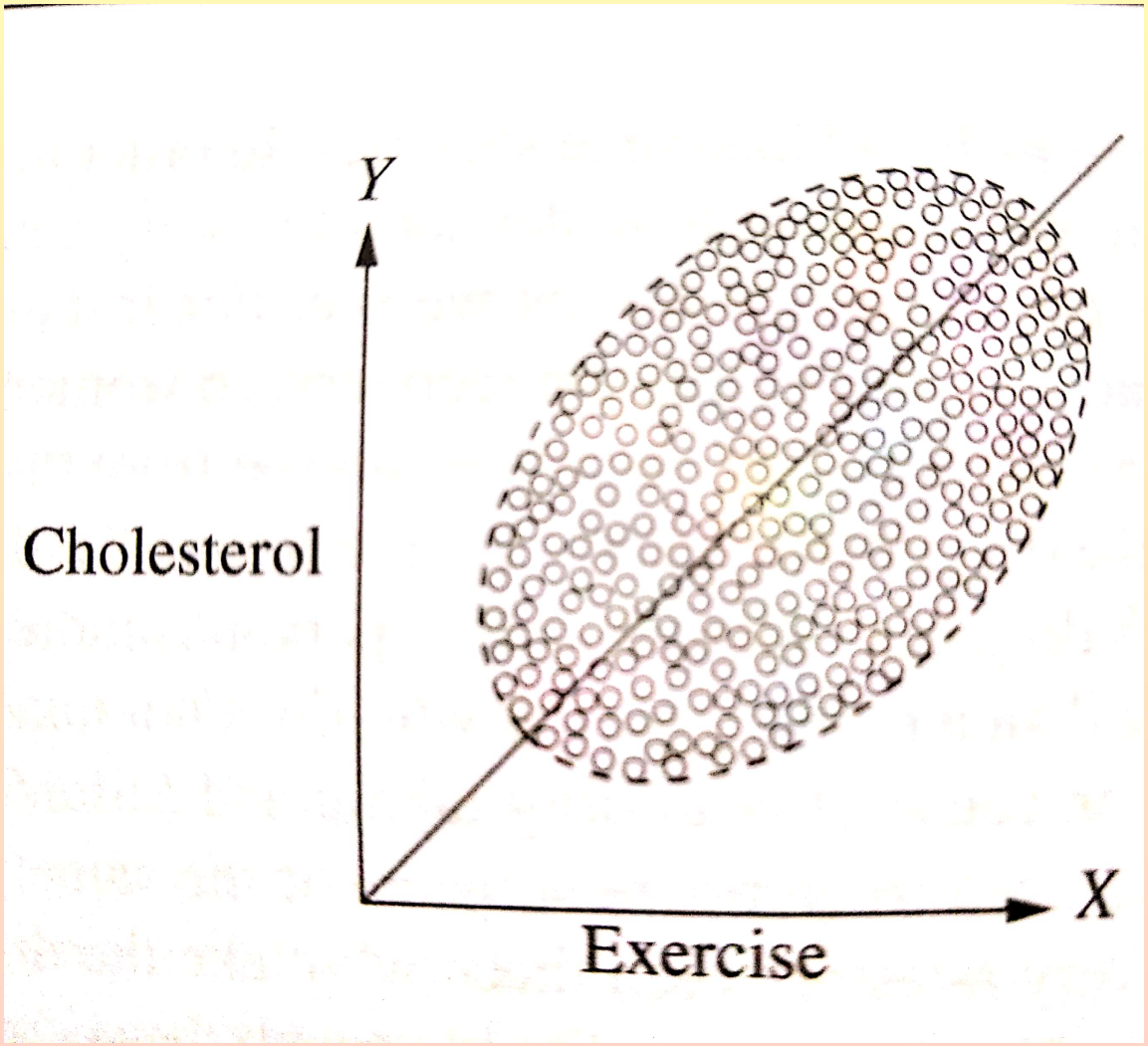
- Related to entropy

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$



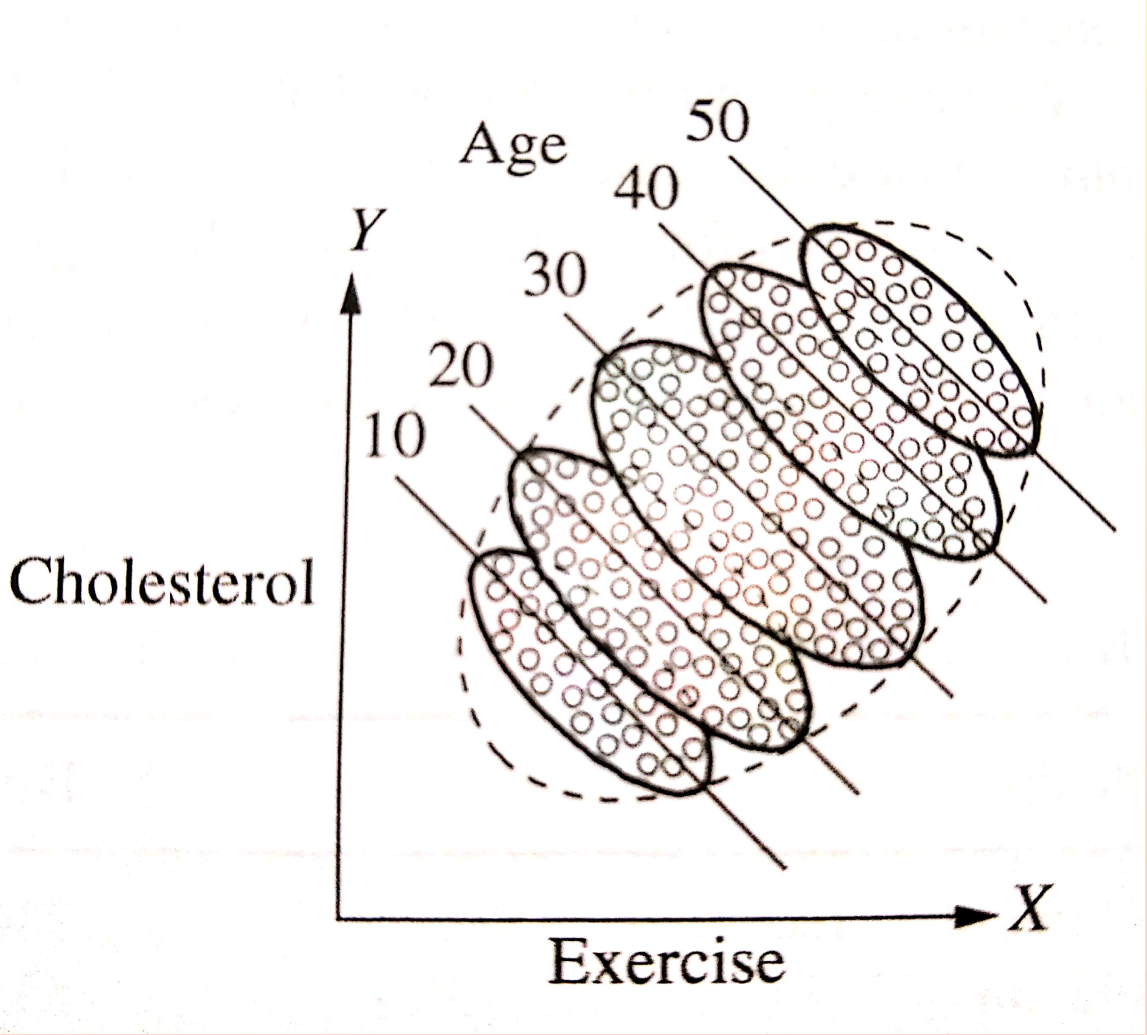
Images by Vischia (left), from Wikipedia (right)

Does cholesterol increase with exercise?



Images from Pearl, 2016

Does it, though?



Images from Pearl, 2016

Should we prescribe the drug?

- If we know the gender, then prescribe the drug
- If we don't know the gender, then don't prescribe the drug

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)

Should we prescribe the drug?

- If we know the gender, then prescribe the drug
- If we don't know the gender, then don't prescribe the drug

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)

- Imagine we know that estrogen has a negative effect on recovery
  - Then women less likely to recovery than men
  - Table shows women are significantly more likely to take the drug

Table from Pearl, 2016

Should we prescribe the drug?

- BP = Blood Pressure

	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)

Should we prescribe the drug?

- BP = Blood Pressure

	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)

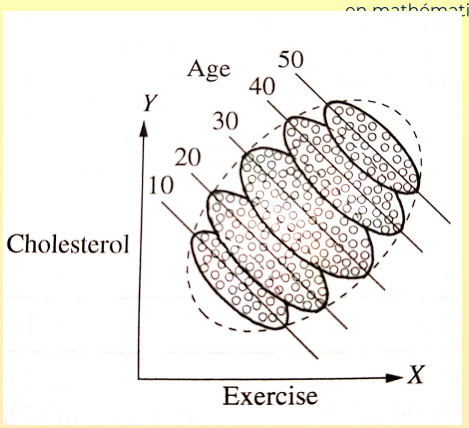
- Same table, different labels; here we must consider the combined data
  - Lowering blood pressure is actually part of the mechanism of the drug effect

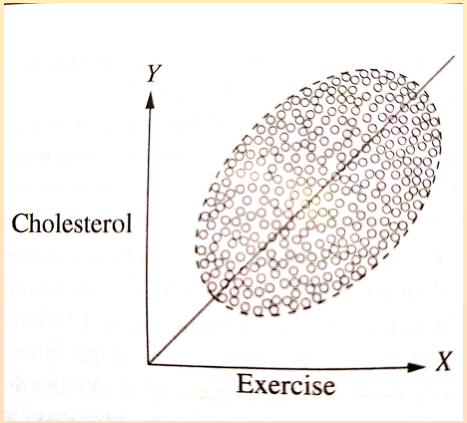
Table from Pearl, 2016

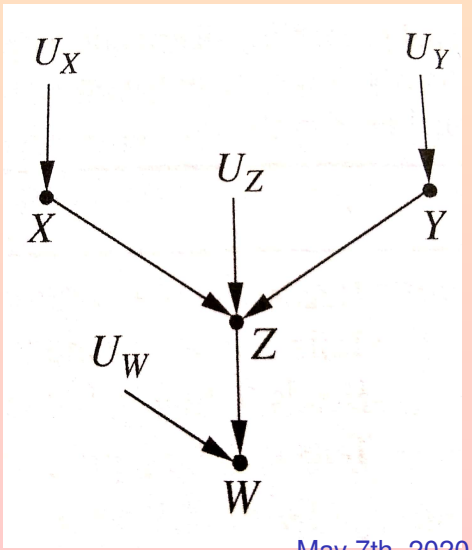
The Simpson paradox: correlation is not causation

UCLouvain
Institut de recherche en mathématique et physique

- Correlation alone can lead to nonsense conclusions
  - If we know the gender, then prescribe the drug
  - If we don't know the gender, then don't prescribe the drug
- Imagine we know that estrogen has a negative effect on recovery
  - Then women less likely to recovery than men
  - Table shows women are significantly more likely to take the drug
- Here we should consult the separate data, in order not to mix effects
- Same table, different labels; must consider the combined data
  - Lowering blood pressure is actually part of the mechanism of the drug effect
- Same effect in continuous data (cholesterol vs age)
- The best solution so far are Bayesian causal networks
  - Graph theory to describe relationship between variables







Plots from Pearl, 2016

First level of causal hierarchy: seeing

- X and Y are marginally dependent, but conditionally independent given Z
- Conditioning on Z *blocks the path*

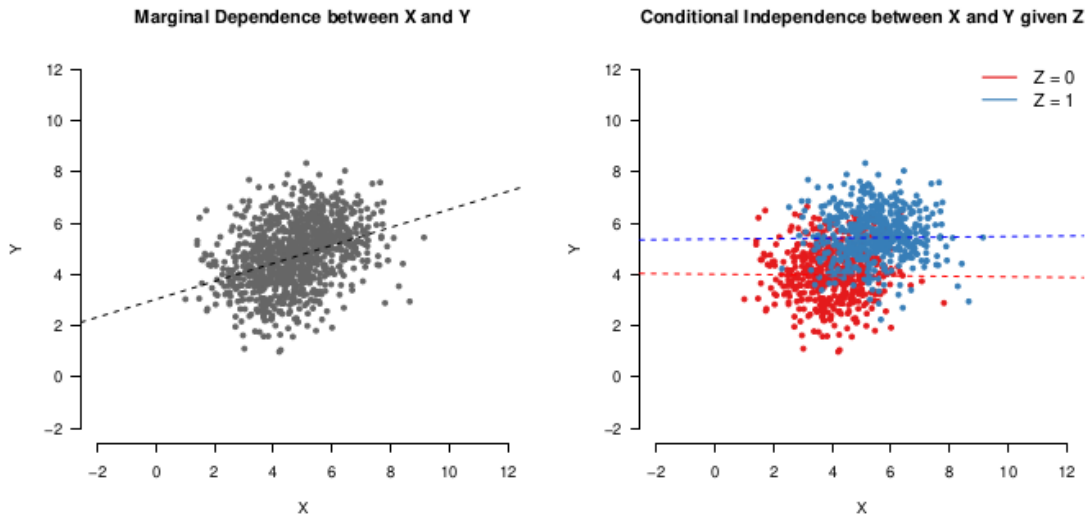


Figure 2. Left: Shows marginal dependence between X and Y. Right: Shows conditional independence between X and Y given Z.

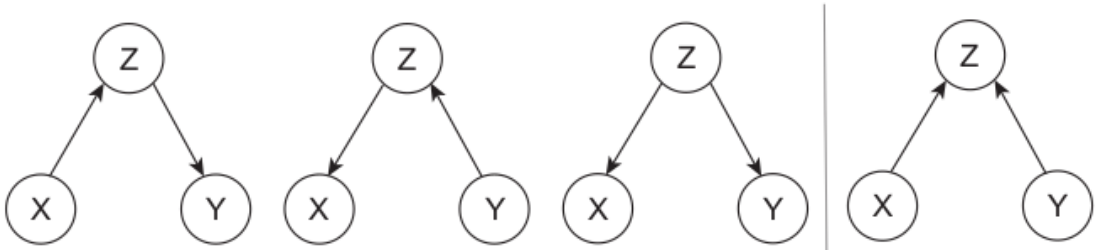


Figure 3. The first three DAGs encode the same conditional independence structure, $X \perp\!\!\!\perp Y \mid Z$. In the fourth DAG, Z is a collider such that $X \not\perp\!\!\!\perp Y \mid Z$.

Plots from Dablander, 2019

Second level of causal hierarchy: doing

- Interventionist approach (Pearl, 2016) (not everyone is onboard)
  - X has a causal influence on Y if changing X leads to changes in (the distribution of) Y
- Setting (by intervention) $X = x$ cuts all incoming causual arrows
  - The value of X is determined only by the intervention
  - Must be able to do intervention: not mere conditioning (seeing): from $P(Y|X = x)$ to $P(Y|do(X = x))$
  - Difficult in social sciences
- Intervention discriminates between causal structure of different diagrams

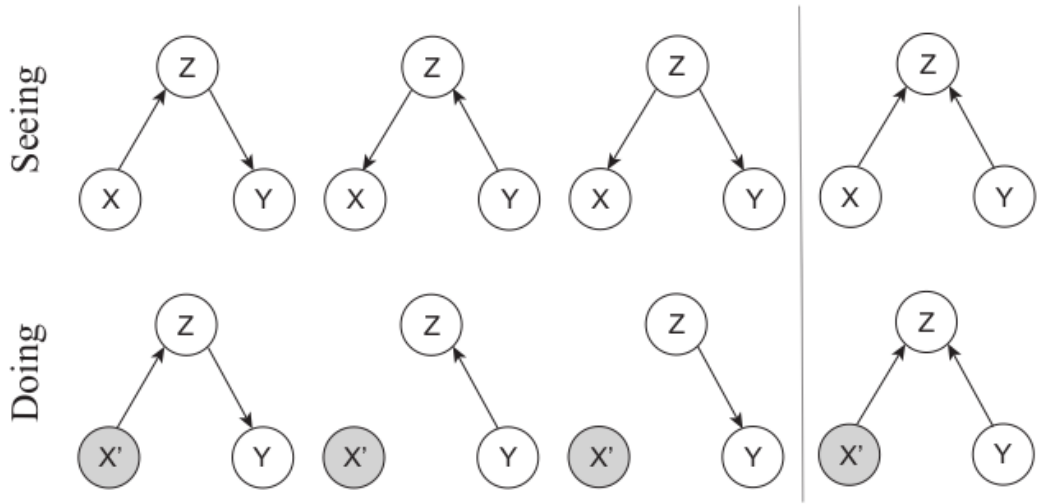


Figure 6. Seeing: DAGs are used to encode conditional independencies. The first three DAGs encode the same associations. *Doing*: DAGs are causal. All of them encode distinct causal assumptions.

Plots from Dablander, 2019

"Doing" applied to Simpson's paradox

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)

	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)

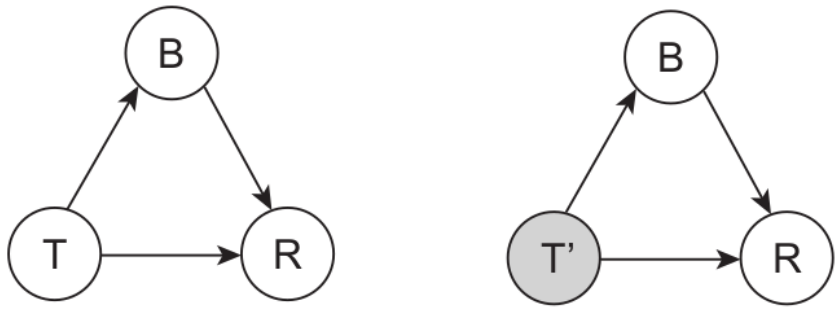


Figure 8. Underlying causal DAG of the example with treatment (T), blood pressure (B), and recovery (R).

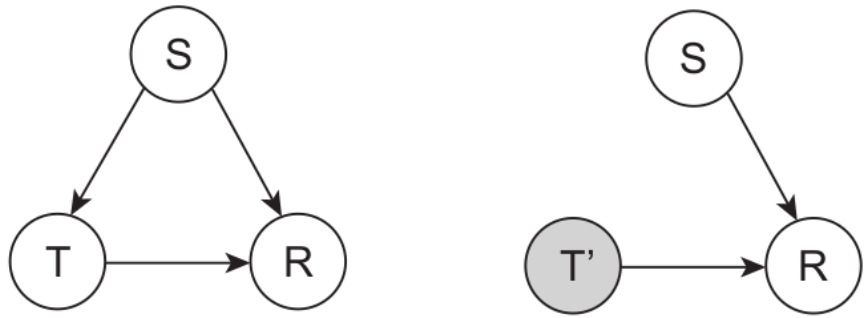


Figure 7. Underlying causal DAG of the example with treatment (T), biological sex (S), and recovery (R).

Plots from Dablander, 2019

do is for populations

- Good predictors can be causally disconnected from the effect!
- The *do* operator operates on distributions defined on populations

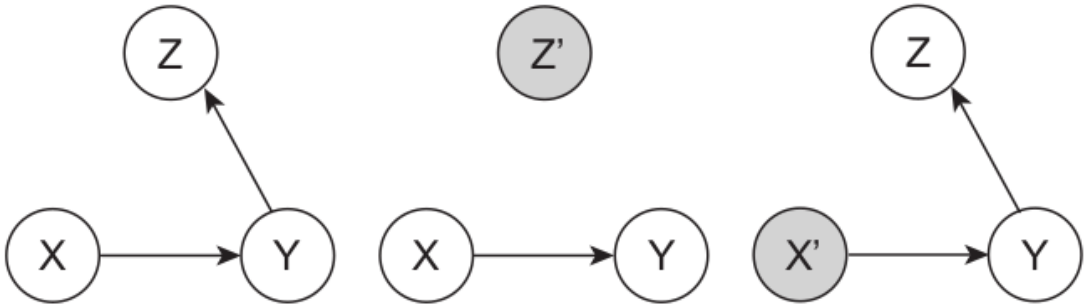


Figure 9. An excellent predictor (Z) need not be causally effective.

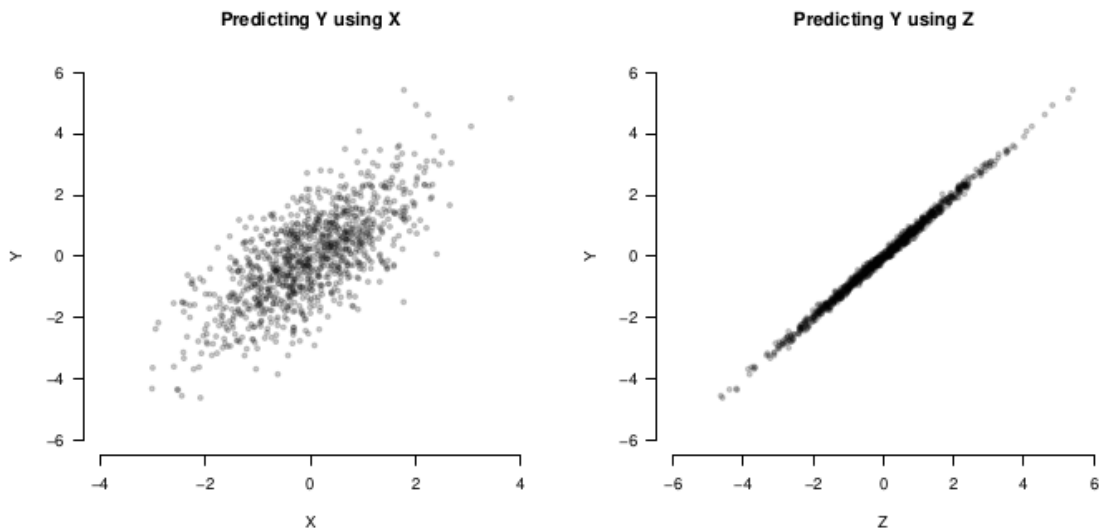


Figure 10. X is a considerably worse predictor of Y than Z.

Plots from Dablander, 2019

Third level of causal hierarchy: imagining

- The strongest level of causality acts on the individual
  - "As a matter of fact, humans constantly evaluate mutually exclusive options, only one of which ever comes true; that is, humans reason counterfactually."
- Structural Causal Models relate causal and probabilistic statements
  - $Treatment := \epsilon_T \sim N(0, \sigma)$
  - $Response := \mu + \beta Treatment + \epsilon$
  - Measure $\mu = 5, \beta = -2, \sigma = 2$
- Causal effect obscured by individual error term ϵ_i for each patient: if determined, model fully determined
- Can determine response for individual treatment!

Table 4
Data simulated from the SCM concerning grandma's treatment of the common cold.

Patient	Treatment	Recovery	ε_k
1	0	5.80	0.80
2	0	3.78	-1.22
3	1	3.68	0.68
4	1	0.74	-2.26
5	0	7.87	2.87

Plots and quote from from Dablander, 2019

Summary

- Test of hypothesis is often based on p-values
- Bayesian tests can solve some problems but still some issue with point nulls
- Even (apparently) strict 5σ criterion and severe testing still can produce false positives
  - By construction, they are supposed to.
- Interpretation of models with respect to the truth is a debatable topic
- So far, only probabilistic connections
- Causal links are needed, based on interventions
  - Often complicated in HEP

Thanks to Tommaso Dorigo for a few historical examples of flukes!

THANK YOU VERY MUCH FOR ATTENDING!!

THANKS FOR THE ATTENTION!

Backup