

# Statistics

or “How to find answers to your questions”

Pietro Vischia<sup>1</sup>

<sup>1</sup>CP3 — IRMP, Université catholique de Louvain



CP3—IRMP, Intensive Course on Statistics for HEP, 21/01–18/02 2022

REMEMBER TO START RECORDING



**Lesson 3**  
**Confidence Intervals in nontrivial cases**  
**Test of hypotheses**  
CLs  
Significance  
**Truth and models**

- Schedule: five days of lectures (Every Friday for the next five weeks)
  - 2h morning lecture, virtual coffee break midway (09:30–11:45)
  - 2h (probably less) afternoon exercise session, virtual coffee break midway (13:30–15:45)
- Many interesting references, nice reading list for your career
  - Papers mostly cited in the topical slides
  - Some cool books cited here and there and in the appendix
- Unless stated otherwise, figures belong to P. Vischia for inclusion in my upcoming textbook on Statistics for HEP (textbook to be published by Springer in 2021)
  - Or I forgot to put the reference, let me know if you spot any figure obviously lacking reference, so that I can fix it
  - I cannot put the recordings publicly online as “massive online course”, so I will distribute them only to registered participants, and have to ask you to not record yourself. I hope you understand.
- Your feedback is crucial for improving these lectures (a feedback form will be provided at the end of the lectures)!
  - You can also send me an email during the lectures: if it is something I can fix for the next day, I'll gladly do so!



- This course provides 3 credits for the UCLouvain doctoral school (CDD Sciences)
  - If you need it recognized by another doctoral school, you have to ask to your school
  - Besides the certificate, I am available at supplying additional information (e.g. detailed schedule) or activity (exam?)
- People connecting online: certificates will be provided by checking connection logs
  - The only way I have to check if you connected to most lectures is to check the Zoom logs
  - Make sure you connect with a recognizable email address (or let me know which unrecognizable address belongs to you)
- This course contributes to the activities of the Excellence of Science (EOS) Be.h network, <https://be-h.be/>



- I will pop up every now and then some questions
- I will open a link, and you'll be able to answer by going to [www.menti.com](https://www.menti.com) and inserting a code
- Totally anonymous (no access even for me to any ID information, not even the country): don't be afraid to give a wrong answer!
  - The purpose is making you think, not having 100% correct answers!
- First question of the day is purely a logistics matter  
**Question time: ROOT**
  - The direct links are accessible to me only: you'll see in your screens the code in a second :)
- The slides of each lecture will be available one minute after the end of the lecture
  - To encourage you to really try answering without looking at the answers

- **Lesson 1 - Fundamentals**
  - Bayesian and frequentist probability, theory of measure, correlation and causality, distributions
- **Lesson 2 - Point and Interval estimation**
  - Maximum likelihood methods, confidence intervals, most probable values, credible intervals
- **Lesson 3 - Advanced interval estimation, test of hypotheses**
  - Interval estimation near the physical boundary of a parameter
  - Frequentist and Bayesian tests, CLs, significance, look-elsewhere effect, reproducibility crisis
- **Lesson 4 - Commonly-used methods in particle physics**
  - Unfolding, ABCD, ABC, MCMC, estimating efficiencies
- **Lesson 5 - Machine Learning**
  - Overview and mathematical foundations, generalities most used algorithms, automatic Differentiation and Deep Learning

- Measure  $N$  times the same quantity: values  $x_i$  and uncertainties  $\sigma_i$ . MLE and variance are:

$$\hat{x}_{ML} = \frac{\sum_{i=1}^N \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^N \frac{1}{\sigma_i^2}}$$

$$\frac{1}{\hat{\sigma}_x^2} = \sum_{i=1}^N \frac{1}{\sigma_i^2}$$

- The MLE is obtained when each measurement is weighted by its own variance
  - This is because the variance is essentially an estimate of how much information lies in each measurement
- This works if the p.d.f. is known
  - Compare this method with an alternative one that does not assume knowledge of the p.d.f.
  - The second method will be the only one applicable to cases in which the p.d.f. is unknown

- Take a set of measures sampled from an unknown p.d.f.  $f(\vec{x}, \vec{\theta})$
- Compute the expected value and variance of a combination of such measurements described by a function  $g(\vec{x})$ .
- The expected value and variance of  $x_i$  are elementary:

$$\mu = E[x] \quad V_{ij} = E[x_i x_j] - \mu_i \mu_j$$

- If we want to extract the p.d.f. of  $g(\vec{x})$ , we would normally use the jacobian of the transformation of  $f$  to  $g$ , but in this case we assumed  $f(\vec{x})$  is unknown.

- We don't know  $f$ , but we can still write an expansion in series for it:

$$g(\vec{x}) \simeq g(\vec{\mu}) + \sum_{i=1}^N \left( \frac{\partial g}{\partial x_i} \right) \Big|_{x=\mu} (x_i - \mu_i)$$

- We can compute the expected value and variance of  $g$  by using the expansion:

$$E[g(\vec{x})] \simeq g(\mu), \quad (E[x_i - \mu_i] = 0)$$

$$\sigma_g^2 = \sum_{ij=1}^N \left[ \frac{\partial g}{\partial x_i} \frac{\partial g}{\partial x_j} \right] \Big|_{\vec{x}=\vec{\mu}} V_{ij}$$

- The variances are propagated to  $g$  by means of their jacobian!
- For a sum of measurements,  $y = g(\vec{x}) = x_1 + x_2$ , the variance of  $y$  is  $\sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2V_{12}$ , which is reduced to the sum of squares for independent measurements

- Let's compare the two ways of combining measurements, and check the role of the Fisher Information
- Let's estimate the time taken for a laser light pulse to go from the Earth to the Moon and back (in units of Earth-to-Moon-Time EMT)
  - On the Moon we have a receiver built by NASA. It's very good but placed in unfavourable conditions, yielding only a 2% precision on Earth-to-Moon
  - On Earth we have a receiver made out of scrap material. It is however placed in favourable conditions, yielding a 5% precision on Moon-to-Earth

$$N_{EM} = 0.99 \pm 0.02 \text{ EMT}$$

$$N_{ME} = 1.05 \pm 0.05 \text{ EMT}$$

- Evidently, the time to moon and back is  $N_{EME} = N_{EM} + N_{ME}$ , and we can apply Eq. 10: **Do it!**

- Let's compare the two ways of combining measurements, and check the role of the Fisher Information
- Let's estimate the time taken for a laser light pulse to go from the Earth to the Moon and back (in units of Earth-to-Moon-Time EMT)
  - On the Moon we have a receiver built by NASA. It's very good but placed in unfavourable conditions, yielding only a 2% precision on Earth-to-Moon
  - On Earth we have a receiver made out of scrap material. It is however placed in favourable conditions, yielding a 5% precision on Moon-to-Earth

$$N_{EM} = 0.99 \pm 0.02 \text{ EMT}$$

$$N_{ME} = 1.05 \pm 0.05 \text{ EMT}$$

- Evidently, the time to moon and back is  $N_{EME} = N_{EM} + N_{ME}$ , and we can apply Eq. 10: **Do it!**
- Resulting estimate:

$$\bullet N_{EME} = 0.99 + 1.05 \pm \sqrt{0.02^2 + 0.05^2} \text{ EMT} = 2.05 \pm 0.05 \text{ EMT}, \text{ corresponding to a precision of } \frac{\sigma_{N_{EME}}}{N_{EME}} \sim 2.4\%.$$



- We now however can argue that over the time it takes for light to go to the Moon and back any environment condition would be roughly constant
- **How can we exploit this additional information? Question Time: Combining Estimates**

- We now however can argue that over the time it takes for light to go to the Moon and back any environment condition would be roughly constant
- How can we exploit this additional information? Question Time: Combining Estimates
- We can use this additional information to note that the two estimates  $N_{EM}$  and  $N_{ME}$  are independent estimates of the same physical quantity  $\frac{N_{EME}}{2}$
- Compute  $N_{EME}$  and  $\sigma(N_{EME})$  based on this reasoning

- We now however can argue that over the time it takes for light to go to the Moon and back any environment condition would be roughly constant
- **How can we exploit this additional information? Question Time: Combining Estimates**
- We can use this additional information to note that the two estimates  $N_{EM}$  and  $N_{ME}$  are independent estimates of the same physical quantity  $\frac{N_{EME}}{2}$
- **Compute  $N_{EME}$  and  $\sigma(N_{EME})$  based on this reasoning**
- We can therefore use Eq. 8 to compute  $\frac{N_{EME}}{2}$  and multiply the result by 2, obtaining

$$N_{EME} = 2.00 \pm 0.03 \text{ EMT}$$

- This estimate corresponds to a precision of only 1.5%!!!
- The dramatic improvement in the precision of the measurement, from 2.4% to 1.5%, is a direct consequence of having used additional information under the form of a relationship (constraint) between the two available measurements.
- A good physicist exploits as many constraints as possible in order to improve the precision of a measurement
  - Sometimes the constraints are arbitrary or correspond to special cases
  - It is very important to explicitly mention any constraint used to derive a measurement, when quoting the result.

- Now suppose my receivers operate by taking data and performing a maximum likelihood fit to estimate  $N_{EM}$  and  $N_{ME}$
- Can I combine these two measurements with the two methods seen above?
  - $N_{EM} = 0.99 \pm 0.03$
  - $N_{ME} = 1.10^{+0.05}_{-0.01}$
- For example,  $N_{EMT} = 2.09^{+0.06}_{-0.03}$

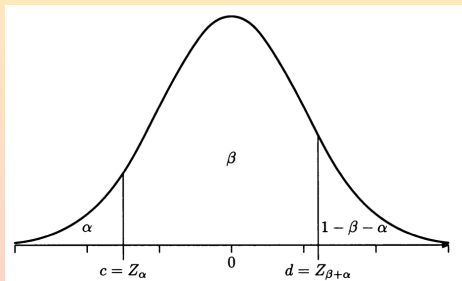
- Now suppose my receivers operate by taking data and performing a maximum likelihood fit to estimate  $N_{EM}$  and  $N_{ME}$
- Can I combine these two measurements with the two methods seen above?
  - $N_{EM} = 0.99 \pm 0.03$
  - $N_{ME} = 1.10^{+0.05}_{-0.01}$
- For example,  $N_{EMT} = 2.09^{+0.06}_{-0.03}$
- No!
- Why?

- Now suppose my receivers operate by taking data and performing a maximum likelihood fit to estimate  $N_{EM}$  and  $N_{ME}$
- Can I combine these two measurements with the two methods seen above?
  - $N_{EM} = 0.99 \pm 0.03$
  - $N_{ME} = 1.10^{+0.05}_{-0.01}$
- For example,  $N_{EMT} = 2.09^{+0.06}_{-0.03}$
- No!
- Why?
- The naïve quadrature of the two uncertainties is wrong!
  - The naïve combination is an expression of the Central Limit Theorem
  - The resulting combination is expected to be more symmetric than the measurements it originates from
  - Symmetric uncertainties usually assume a Gaussian approximation of the likelihood
  - Asymmetric uncertainties? One would need a study of the non-linearity (large biases might be introduced if ignoring this)
- Intrinsic difference between averaging and most probable value
  - Averaging results in average value and variance that propagate linearly
  - Taking the mode (essentially what MLE does) does not add up linearly!
- With asymmetric uncertainties from MLE fits, always combine the likelihoods (better in an individual simultaneous fit)

# Confidence Intervals in nontrivial cases

- Confidence interval for  $\theta$  with probability content  $\beta$ 
  - The range  $\theta_a < \theta < \theta_b$  containing the true value  $\theta_0$  with probability  $\beta$
  - The physicists sometimes improperly say the uncertainty on the parameter  $\theta$
- Given a p.d.f., the probability content is  $\beta = P(a \leq X \leq b) = \int_a^b f(X|\theta)dX$
- If  $\theta$  is unknown (as is usually the case), use auxiliary variable  $Z = Z(X, \theta)$  with p.d.f.  $g(Z)$  independent of  $\theta$
- If  $Z$  can be found, then the problem is to estimate interval  $P(\theta_a \leq \theta_0 \leq \theta_b) = \beta$ 
  - Confidence interval
  - A method yielding an interval satisfying this property has coverage

- Example: if  $f(X|\theta) = N(\mu, \sigma^2)$  with unknown  $\mu, \sigma$ , choose  $Z = \frac{X-\mu}{\sigma}$
- Find  $[c, d]$  in  $\beta = P(c \leq Z \leq d) = \Phi(d) - \Phi(c)$  by finding  $[Z_\alpha, Z_{\alpha+\beta}]$
- Infinite interval choices: here central interval  
 $\alpha = \frac{1-\beta}{2}$

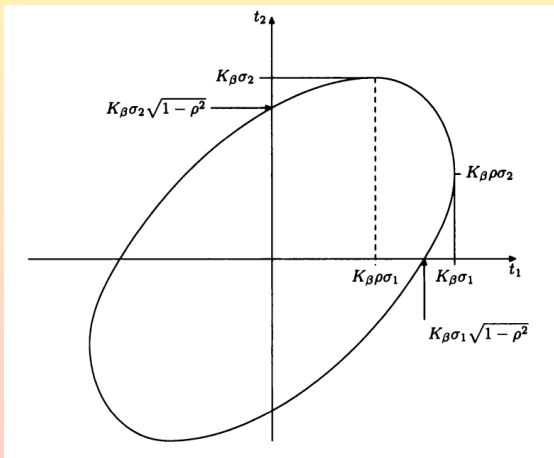


Plot from James, 2nd ed.



## Confidence intervals in many dimensions

- Generalization to multidimensional  $\theta$  is immediate
- Probability statement concerns the whole  $\theta$ , not the individual  $\theta_i$
- Shape of the ellipsoid governed by the correlation coefficient (or the mutual information) between the parameters
- Arbitrariness in the choice of the interval is still present



Plot from James, 2nd ed.

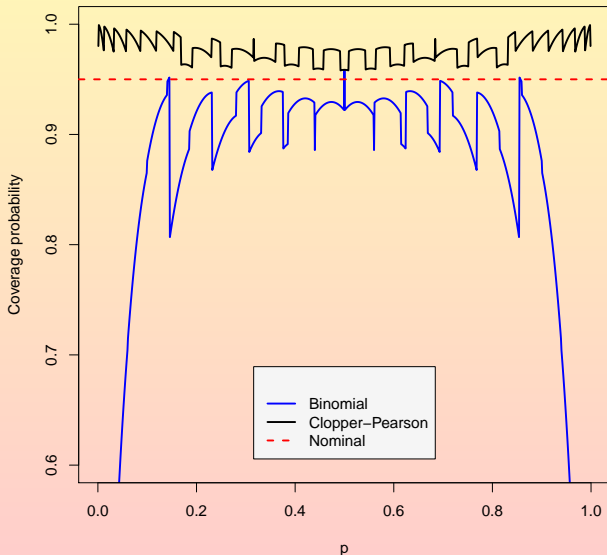
- Coverage probability of a method for calculating a confidence interval  $[\theta_1, \theta_2]$ :  
 $P(\theta_1 \leq \theta_{true} \leq \theta_2)$ 
  - Fraction of times, over a set of (usually hypothetical) measurements, that the resulting interval covers the true value of the parameter
  - Can sample with toys to study coverage
- Coverage is not a property of a specific confidence interval!
- **Coverage is a property of the method you use to compute your confidence interval**
  - It is calculated from the sampling distribution of your confidence intervals
- The nominal coverage is the value of confidence level you have built your method around (often 0.95)
- When actually derive a set of intervals, the fraction of them that contain  $\theta_{true}$  ideally would be equal to the nominal coverage
  - You can build toy experiments in each of whose you sample  $N$  times for a known value of  $\theta_{true}$
  - You calculate the interval for each toy experiment
  - You count how many times the interval contains the true value
- Nominal coverage ( $CL$ ) and the actual coverage ( $Co$ ) observed with toys should agree
  - If all the assumptions you used in computing the intervals are valid
  - If they don't agree, it might be that  $Co < CL$  (**undercoverage**) or  $Co > CL$  (**overcoverage**)
  - It's OK to strive to be conservative, but one might be unnecessarily lowering the precision of the measurement
  - When  $Co \neq CL$  you usually want at least a convergence to equality in some limit

- For discrete distributions, the discreteness induces steps in the probability content of the interval
  - Continuous case:  $P(a \leq X \leq b) = \int_a^b f(X|\theta) dX = \beta$
  - Discrete case:  $P(a \leq X \leq b) = \sum_a^b f(X|\theta) dX \leq \beta$
- Binomial: find interval  $(r_{low}, r_{high})$  such that  $\sum_{r=r_{low}}^{r=r_{high}} \binom{r}{N} p^r (1-p)^{N-r} \leq 1 - \alpha$ 
  - Also,  $\binom{r}{N}$  computationally taxing for large  $r$  and  $N$
  - Approximations are found in order to deal with the problem
- Gaussian approximation:  $p \pm Z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{N}}$
- Clopper Pearson: invert two single-tailed binomial tests
 
$$\sum_{r=0}^N \binom{r}{N} p^n (1 - p_{low})^{N-n} \leq \alpha/2$$

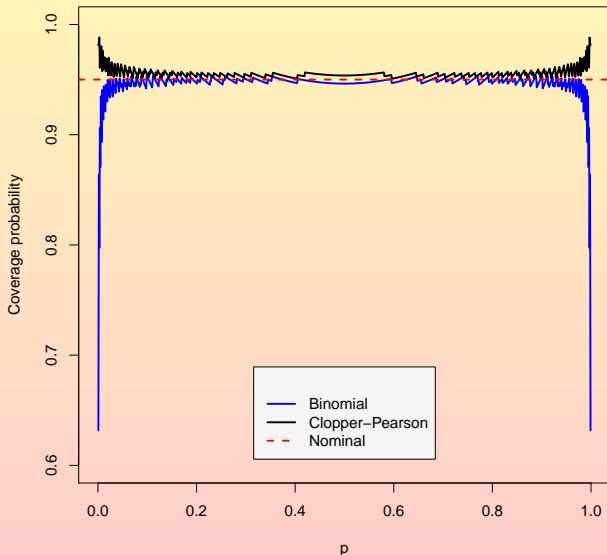
$$\sum_{r=0}^N \binom{r}{N} p^r (1 - p_{high})^{N-r} \leq \alpha/2$$
  - Single-tailed  $\rightarrow$  use  $\alpha/2$  instead of  $\alpha$

- Gaussian approximation:  $p \pm Z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{N}}$
- Clopper Pearson: invert two single-tailed binomial tests, designed to overcover
$$\sum_{r=0}^N \binom{r}{N} p^n (1 - p_{low})^{N-n} \leq \alpha/2$$
$$\sum_{r=0}^N \binom{r}{N} p^r (1 - p_{high})^{N-r} \leq \alpha/2$$
  - Single-tailed  $\rightarrow$  use  $\alpha/2$  instead of  $\alpha$
- This afternoon we will study the coverage of intervals from a gaussian approximation and from the Clopper-Pearson method
- We will also study the coverage of intervals obtained from crossings with  $\Delta \ln L$
- Question time: Coverage

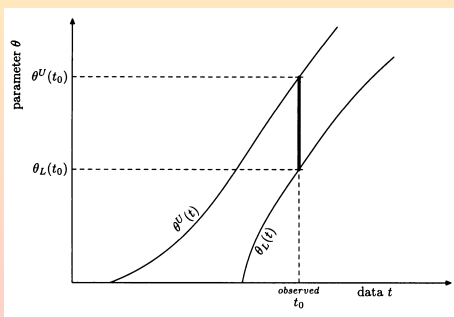
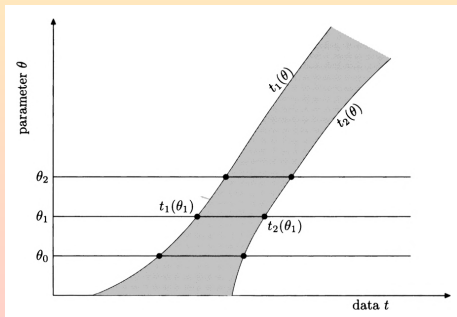
- Gaussian approximation bad for small sample sizes



- Gaussian approximation bad near  $p = 0$  and  $p = 1$  even for large sample sizes



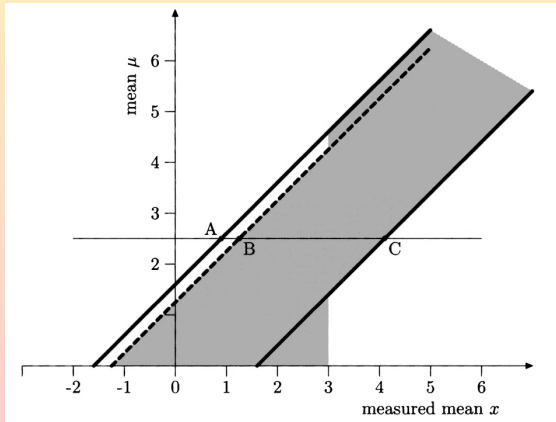
- Unique solutions to finding confidence intervals are infinite
  - Central intervals, lower limits, upper limits, etc
- Let's suppose we have chosen a way
- Build horizontally: for each (hypothetical) value of  $\theta$ , determine  $t_1(\theta)$ ,  $t_2(\theta)$  such that
 
$$\int_{t_1}^{t_2} P(t|\theta)dt = \beta$$
- Read vertically: from the observed value  $t_0$ , determine  $[\theta_L, \theta^U]$  by intersection
  - The resulting interval might be disconnected in severely non-linear cases
- Probability content statements to be seen in a frequentist way
  - Repeating many times the experiment, the fraction of  $[\theta_L, \theta^U]$  containing  $\theta_0$  is  $\beta$



Plot from James, 2nd ed.

## Upper limits for non-negative parameters

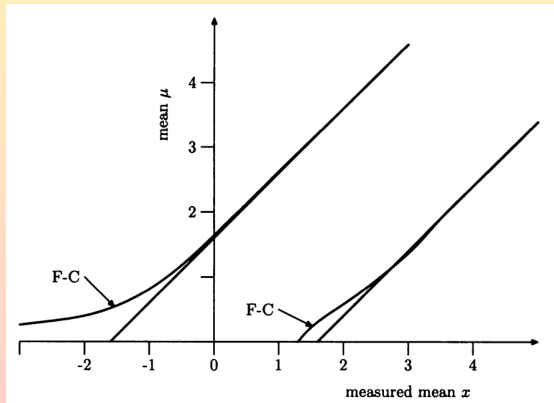
- Gaussian measurement ( variance 1) of a non-negative parameter  $\mu \sim 0$  (physical bound)
- Individual prescriptions are self-consistent
  - 90% central limit (solid lines)
  - 90% upper limit (single dashed line)
- Other choices are problematic (flip-flopping): never choose after seeing the data!
  - “quote upper limit if  $x_{obs}$  is less than  $3\sigma$  from zero, and central limit above” (shaded)
  - Coverage not guaranteed anymore (see e.g.  $\mu = 2.5$ )
- Unphysical values and empty intervals: choose 90% central interval, measure  $x_{obs} = -2.0$ 
  - Don't extrapolate to an unphysical interval for the true value of  $\mu$ !
  - The interval is simply empty, i.e. does not contain any allowed value of  $\mu$
  - The method still has coverage (90% of other hypothetical intervals would cover the true value)





- The Neyman construction results in guaranteed coverage, but choice still free on how to fill probability content
  - Different ordering principles are possible (e.g. central/upper/lower limits)
- Unified approach for determining interval for  $\mu = \mu_0$ : the likelihood ratio ordering principle
  - Include in order by largest  $\ell(x) = \frac{P(x|\mu_0)}{P(x|\hat{\mu})}$
  - $\hat{\mu}$  value of  $\mu$  which maximizes  $P(x|\mu)$  within the physical region
  - $\hat{\mu}$  remains equal to zero for  $\mu < 1.65$ , yielding deviation w.r.t. central intervals

- Minimizes Type II error (likelihood ratio for simple test is the most powerful test)
- Solves the problem of empty intervals
- Avoids flip-flopping in choosing an ordering prescription



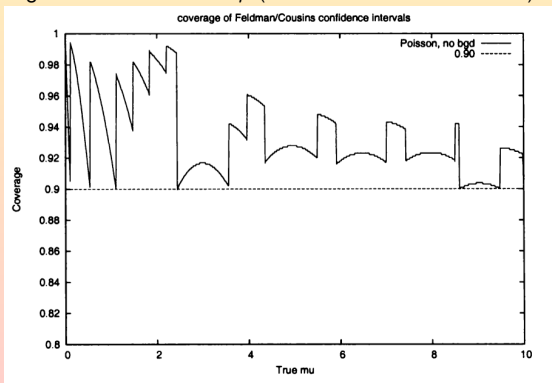
Plot from James, 2nd ed.

## Feldman-Cousins in HEP

- The most typical HEP application of F-C is confidence belts for the mean of a Poisson distribution
- Discreteness of the problem affects coverage
- When performing the Neyman construction, will add discrete elements of probability
- The exact probability content won't be achieved, must accept overcoverage

$$\int_{x_1}^{x_2} f(x|\theta)dx = \beta \quad \rightarrow \quad \sum_{i=L}^U P(x_i|\theta) \geq \beta$$

- Overcoverage larger for small values of  $\mu$  (but less than other methods)



Plot from James, 2nd ed.

- Often numerically identical to frequentist confidence intervals
  - Particularly in the large sample limit
- Interpretation is different: credible intervals
- Posterior density summarizes the complete knowledge about  $\theta$

$$\pi(\theta|\mathbf{X}) = \frac{\prod_{i=1}^N f(X_i, \theta)\pi(\theta)}{\int \prod_{i=1}^N f(X_i, \theta)\pi(\theta)d\theta}$$

- Sometimes you may want to summarize the prior with estimates of its location and of its dispersion
  - For the location, you can use mode or median (see tomorrow's lecture)
- An interval  $[\theta_L, \theta^U]$  with content  $\beta$  defined by  $\int_{\theta_L}^{\theta^U} \pi(\theta|\mathbf{X})d\theta = \beta$
- Bayesian statement!  $P(\theta_L < \theta < \theta^U) = \beta$ 
  - Again, non unique
- Issues with empty intervals don't arise, though, because the prior takes care of defining the physical region in a natural way!
  - But this implies that central intervals cannot be seamlessly converted into upper limits
  - Need the notion of shortest interval
  - Issue of the metric (present in frequentist statistic) solved because here the preferred metric is defined by the prior

## Bayesian intervals and coverage

- What about computing the frequentist coverage for Bayesian intervals?
- **Question time: Coverage Bayes**

- What about computing the frequentist coverage for Bayesian intervals?
- **Question time: Coverage Bayes**
- Even if you are not interested in frequentist methods, it can be useful! Certainly it doesn't hurt
- Knowing the sampling properties of a method can always give insights or work as a cross-check of the method
- Particularly given that typically Bayesian and frequentist answers tend to converge in the high- $N$  limit
  - Except for hypothesis tests, we'll find out later today



Image from the [Statistical Statistics Memes Facebook Page](#)

# Test of Hypotheses

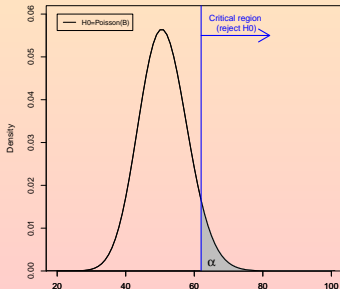
- Is our hypothesis compatible with the experimental data? By how much?
- Hypothesis: a complete rule that defines probabilities for data.
  - An hypothesis is simple if it is completely specified (or if each of its parameters is fixed to a single value)
  - An hypothesis is complex if it consists in fact in a family of hypotheses parameterized by one or more parameters
- “Classical” hypothesis testing is based on frequentist statistics
  - An hypothesis—as we do for a parameter  $\vec{\theta}_{true}$ —is either true or false. We might improperly say that  $P(H)$  can only be either 0 or 1
  - The concept of probability is defined only for a set of data  $\vec{x}$
- We take into account probabilities for data,  $P(\vec{x}|H)$ 
  - For a fixed hypothesis, often we write  $P(\vec{x}; H)$ , skipping over the fact that it is a conditional probability
  - The size of the vector  $\vec{x}$  can be large or just 1, and the data can be either continuous or discrete.

- The hypothesis can depend on a parameter
  - Technically, it consists in a family of hypotheses scanned by the parameter
  - We use the parameter as a proxy for the hypothesis,  $P(\vec{x}; \theta) := P(\vec{x}; H(\theta))$ .
- We are working in frequentist statistics, so there is no  $P(H)$  enabling conversion from  $P(\vec{x}|\theta)$  to  $P(\theta|\vec{x})$ .
- 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



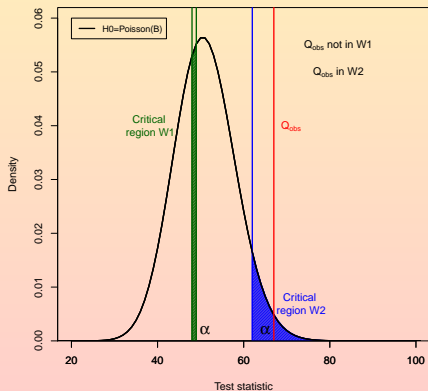
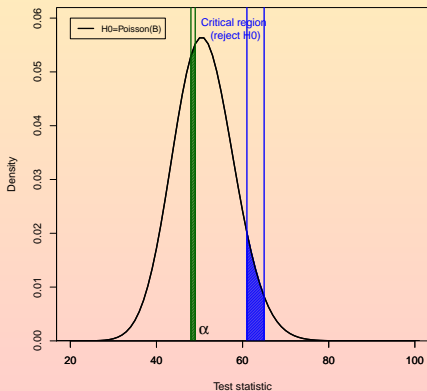
## Testing an hypothesis $H_0$ ...

- $H_0$  is normally the hypothesis that we assume true in absence of further evidence
- Let  $X$  be a function of the observations (called “*test statistic*”)
- Let  $W$  be the space of all possible values of  $X$ , and divide it into
  - A critical region  $w$ : observations  $X$  falling into  $w$  are regarded as suggesting that  $H_0$  is NOT true
  - A region of acceptance  $W - w$
- The size of the critical region is adjusted to obtain a desired *level of significance*  $\alpha$ 
  - Also called *size of the test*
  - $P(X \in w | H_0) = \alpha$
  - $\alpha$  is the (hopefully small) probability of rejecting  $H_0$  when  $H_0$  is actually true
- Once  $W$  is defined, given an observed value  $\vec{x}_{obs}$  in the space of data, we define the test by saying that we reject the hypothesis  $H_0$  if  $\vec{x}_{obs} \in W$ .
- If  $\vec{x}_{obs}$  is inside the critical region, then  $H_0$  is rejected; in the other case,  $H_0$  is accepted
  - In this context, accepting  $H_0$  does not mean demonstrating its truth, but simply not rejecting it
- Choosing a small  $\alpha$  is equivalent to giving a priori preference to  $H_0$ !!!



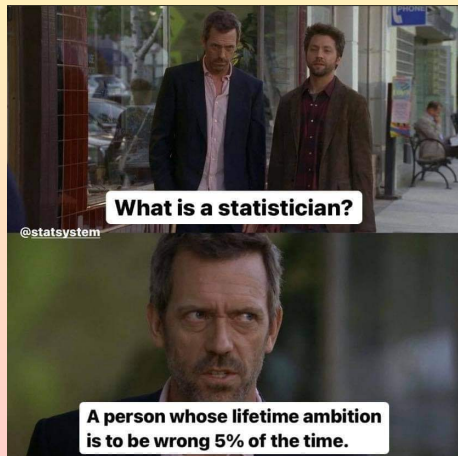
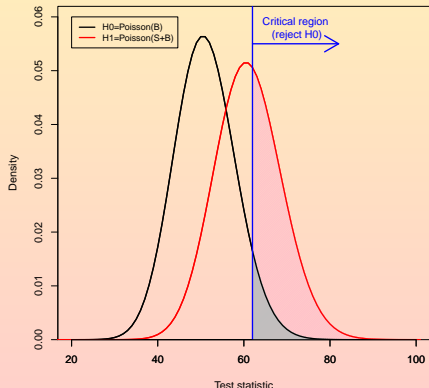
## ...while introducing some spice in it

- The definition of  $\mathcal{W}$  depends only on its area  $\alpha$ , without any other condition
  - Any other area of area  $\alpha$  can be defined as critical region, independently on how it is placed with respect to  $\bar{x}_{obs}$
  - In particular, for an infinite number of choices of  $\mathcal{W}$ , the point  $\bar{x}_{obs}$ —which beforehand was situated outside of  $\mathcal{W}$ —is now included inside the critical region
  - In this condition, the result of the test switches from accept  $H_0$  to reject  $H_0$
- To remove or at least reduce this arbitrariness in the choice of  $\mathcal{W}$ , we introduce the alternative hypothesis,  $H_1$



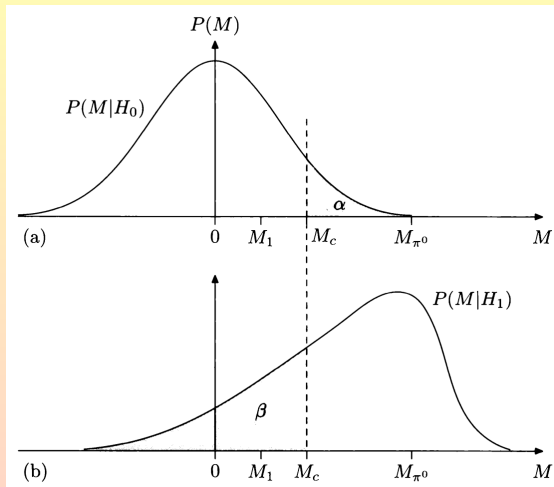
## Choose reasonable regions

- Choose a critical region so that  $P(\vec{x} \in \mathcal{W} | H_0)$  is  $\alpha$  under  $H_0$ , and as large as possible under  $H_1$
- Choice of regions is somehow arbitrary, and many choices are not more justified than others
- In Physics, after ruling out an hypothesis we aim at substituting it with one which explains better the data
  - Often  $H_1$  becomes the new  $H_0$ , e.g. from  $(H_0:\text{noHiggs}, H_1 = \text{Higgs})$  to  $(H_1:\text{Higgs}, H_1:\text{otherNewPhysics})$
  - We can use our expectations about reasonable alternative hypotheses to design our test to exclude  $H_0$



Could not find source for the meme

- $H_0: pp \rightarrow pp$  elastic scattering
- $H_1: pp \rightarrow pp\pi^0$
- Compute the missing mass  $M$  (as total rest energy of unseen particles)
- Under  $H_0$ ,  $M = 0$
- Under  $H_1$ ,  $M = 135 \text{ MeV}$



	Choose $H_0$	Choose $H_1$	
$H_0$ is true	$1 - \alpha$	$\alpha$ (Type I error)	Plot from James, 2nd ed.
$H_1$ is true	$\beta$ (Type II error)	$1 - \beta$ (power)	

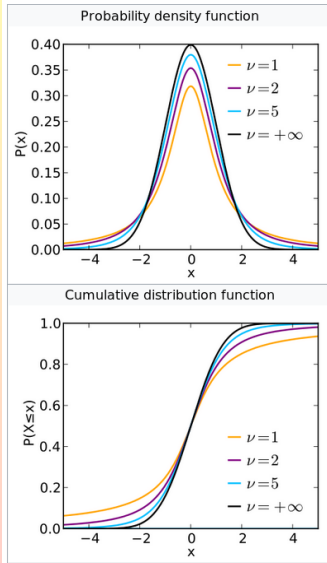
- Student's t distribution
- Test the mean!
- Will not run it this afternoon, you can check it at home [hypptest.ipynb](http://hypptest.ipynb)

PDF

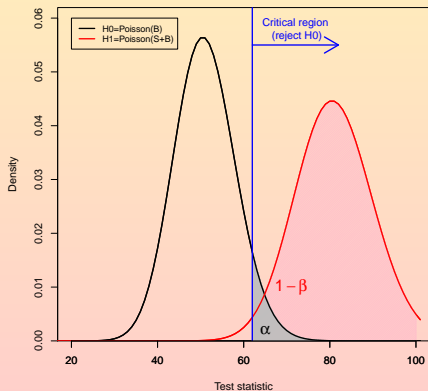
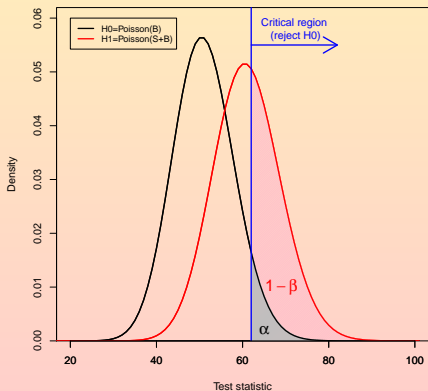
$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$



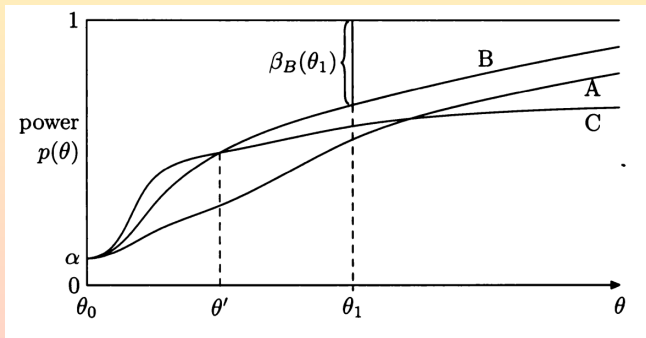
## Student's t



- The usefulness of the test depends on how well it discriminates against the alternative hypothesis
- The measure of usefulness is the *power of the test*
  - $P(X \in w | H_1) = 1 - \beta$
  - Power ( $1 - \beta$ ) is the probability of  $X$  falling into the critical region if  $H_1$  is true
  - $P(X \in W - w | H_1) = \beta$
  - $\beta$  is the probability that  $X$  will fall into the acceptance region if  $H_1$  is true
- NOTE: some authors use  $\beta$  where we use  $1 - \beta$ . Pay attention, and live with it.



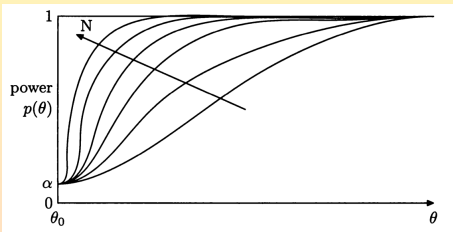
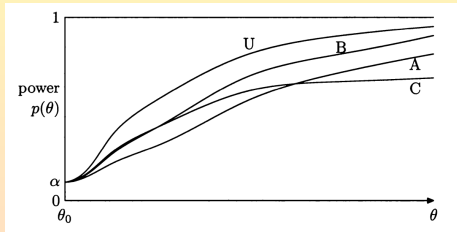
- For parametric (families of) hypotheses, the power depends on the parameter
  - $H_0 : \theta = \theta_0$
  - $H_1 : \theta = \theta_1$
  - Power:  $p(\theta_1) = 1 - \beta$
- Generalize for all possible alternative hypotheses:  $p(\theta) = 1 - \beta(\theta)$ 
  - For the null,  $p(\theta_0) = 1 - \beta(\theta_0) = \alpha$



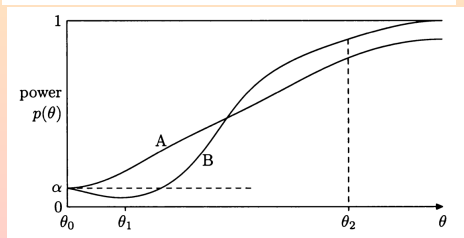
Plot from James, 2nd ed.

## Properties of tests

- More powerful test: a test which is at least as powerful as any other test for a given  $\theta$
- Uniformly more powerful test: a test which is the more powerful test for any value of  $\theta$ 
  - A less powerful test might be preferable if more robust than the UMP<sup>1</sup>
- If we increase the number of observations, it makes sense to require consistency
  - The more observations we add, the more the test distinguishes between the two hypotheses
  - Power function tends to a step function for  $N \rightarrow \infty$



- Biased test:  $\operatorname{argmin}(p(\theta)) \neq \theta_0$
- More likely to accept  $H_0$  when it is false than when it is true
- Big no-no for  $\theta_0$  vs  $\theta_1$ ]
- Still useful (larger power) for  $\theta_0$  vs  $\theta_2$



Plot from James, 2nd ed.

<sup>1</sup> Robust: a test with low sensitivity to unimportant changes of the null hypothesis



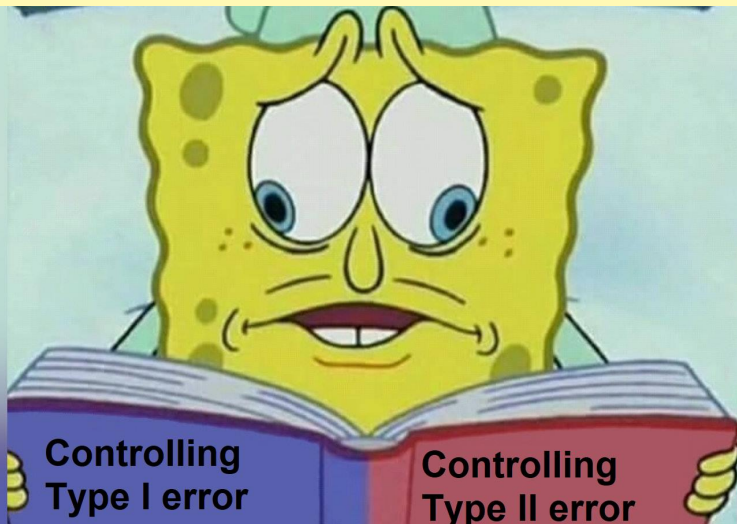
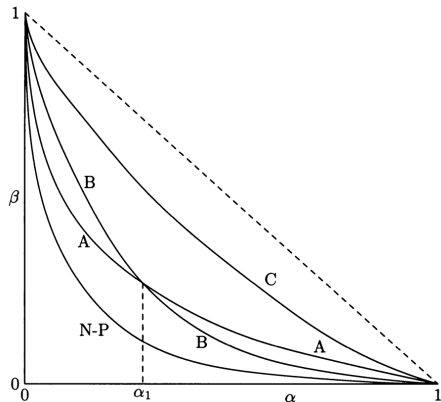


Image from the [Statistical Statistics Memes Facebook Page](#)

- Comparing only based on the power curve is asymmetric w.r.t.  $\alpha$
- For each value of  $\alpha = p(\theta_0)$ , compute  $\beta = p(\theta_1)$ , and draw the curve
  - Unbiased tests fall under the line  $1 - \beta = \alpha$
  - Curves closer to the axes are better tests
- Ultimately, though, choose based on the cost function of a wrong decision
  - Bayesian decision theory



$$h(\mathbf{X}|\theta, \phi, \psi) = \theta f(\mathbf{X}|\phi) + (1 - \theta)g(\mathbf{X}, \psi)$$

$d_0$  : No choice is possible; results are ambiguous

$d_1, \phi^*$  : Family was  $f(\mathbf{X}|\phi)$ , with  $\phi = \phi^*$

$d_2, \psi^*$  : Family was  $g(\mathbf{X}|\psi)$ , with  $\psi = \psi^*$ .

Table 10.4. A cost function.

Decisions	True state of nature	
	$\theta = \theta_1 = 1, \phi$	$\theta = \theta_2 = 0, \psi$
$d_0$	$\beta_1$	$\beta_2$
$d_1, \phi^*$	$\alpha_1(\phi^* - \phi)^2$	$\gamma_1$
$d_2, \psi^*$	$\gamma_2$	$\alpha_2(\psi^* - \psi)^2$

- Testing simple hypotheses  $H_0$  vs  $H_1$ , find the best critical region
- Maximize power curve  $1 - \beta = \int_{w_\alpha} f(\mathbf{X}|\theta_1)d\mathbf{X}$ , given  $\alpha = \int_{w_\alpha} f(\mathbf{X}|\theta_0)d\mathbf{X}$
- The best critical region  $w_\alpha$  consists in the region satisfying the likelihood ratio equation

$$\ell(\mathbf{X}, \theta_0, \theta_1) := \frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)} \geq c_\alpha$$

- The criterion, called Neyman-Pearson test, is therefore
  - If  $\ell(\mathbf{X}, \theta_0, \theta_1) > c_\alpha$  then choose  $H_1$
  - If  $\ell(\mathbf{X}, \theta_0, \theta_1) \leq c_\alpha$  then choose  $H_0$
- The likelihood ratio must be calculable for any  $\mathbf{X}$ 
  - The hypotheses must therefore be completely specified simple hypotheses
  - For complex hypotheses,  $\ell$  is not necessarily optimal

- We want to prove that  $\ell(\mathbf{X}, \theta_0, \theta_1) := \frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)} \geq c_\alpha$  gives the best acceptance region

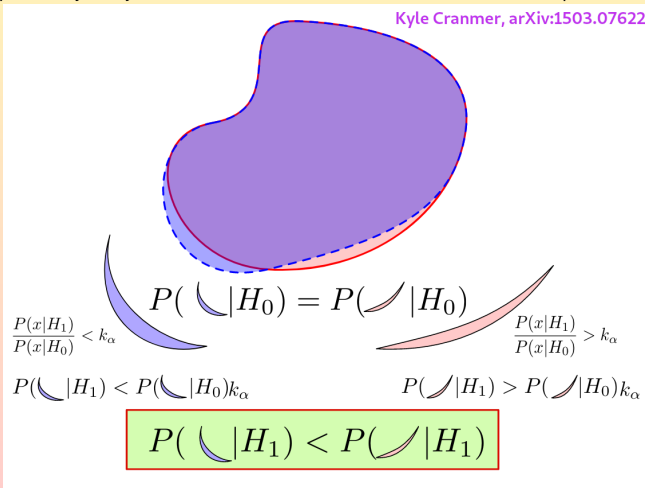


Image from Evan Vucci, Shutterstock, meme is mine

## Demonstrating the Neyman-Pearson lemma

- We want to prove that  $\ell(\mathbf{X}, \theta_0, \theta_1) := \frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)} \geq c_\alpha$  gives the best region
  - Critical region from NP (red contour), demonstrate that any other region (blue contour) has less power
  - Take out a wedge region and add it e.g. to the other side
  - Regions must have equal area under  $H_0$  (tests with same size)
  - Being on different sides of the red contour, under  $H_1$  data is less likely in the added region than in the removed one
  - Less probability to reject the null  $\rightarrow$  test based on the new contour is less powerful!

Kyle Cranmer, arXiv:1503.07622

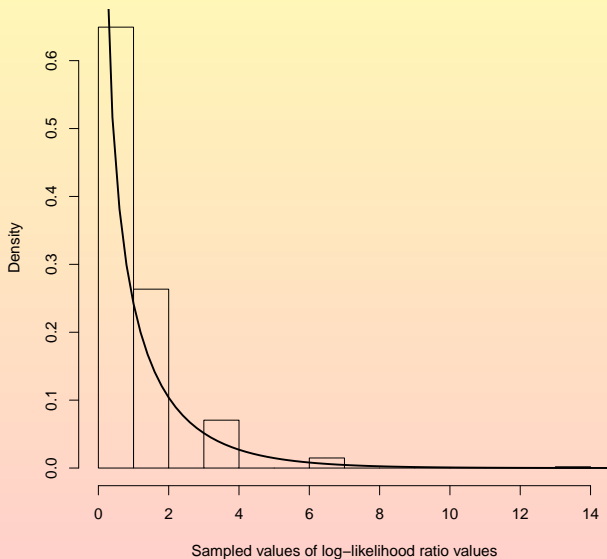


- The likelihood ratio is commonly used
- As any test statistic in the market, in order to select critical regions based on confidence levels it is necessary to know its distribution
  - Run toys to find its distribution (very expensive if you want to model extreme tails)
  - Find some asymptotic condition under which the likelihood ratio assumes a simple known form
- Wilks theorem: when the data sample size tends to  $\infty$ , the likelihood ratio tends to  $\chi^2(N - N_0)$ 
  - Exercise yesterday afternoon

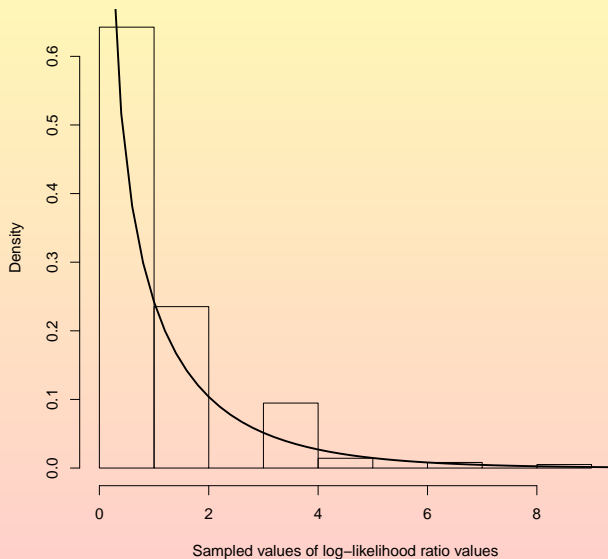
We can summarize in the

**Theorem:** *If a population with a variate  $x$  is distributed according to the probability function  $f(x, \theta_1, \theta_2 \dots \theta_h)$ , such that optimum estimates  $\bar{\theta}_i$  of the  $\theta_i$  exist which are distributed in large samples according to (3), then when the hypothesis  $H$  is true that  $\theta_i = \theta_{0i}$ ,  $i = m + 1, m + 2, \dots h$ , the distribution of  $-2 \log \lambda$ , where  $\lambda$  is given by (2) is, except for terms of order  $1/\sqrt{n}$ , distributed like  $\chi^2$  with  $h - m$  degrees of freedom.*

## Log-likelihood ratio

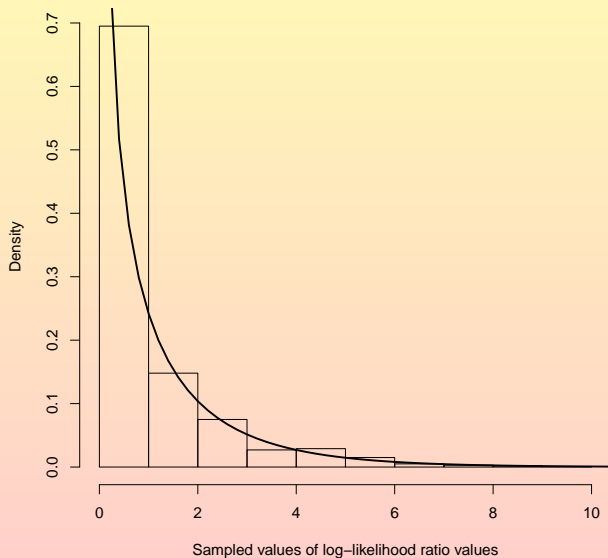


## Log-likelihood ratio





## Log-likelihood ratio



- The parameter  $\theta$  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
<b>0 to 1/2</b>	1 to 3.2	Not worth more than a bare mention
<b>1/2 to 1</b>	3.2 to 10	Substantial
<b>1 to 2</b>	10 to 100	Strong
<b>&gt; 2</b>	$> 100$	Decisive

## Trotta

$ \ln B $	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 Lemaître Lectures 2018

# Bayesian model comparison of 193 models

## Higgs inflation as reference model

Martin,RT+14

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

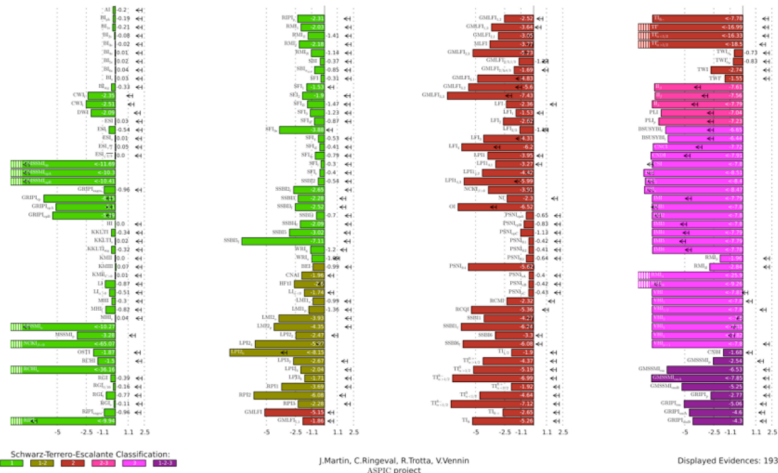
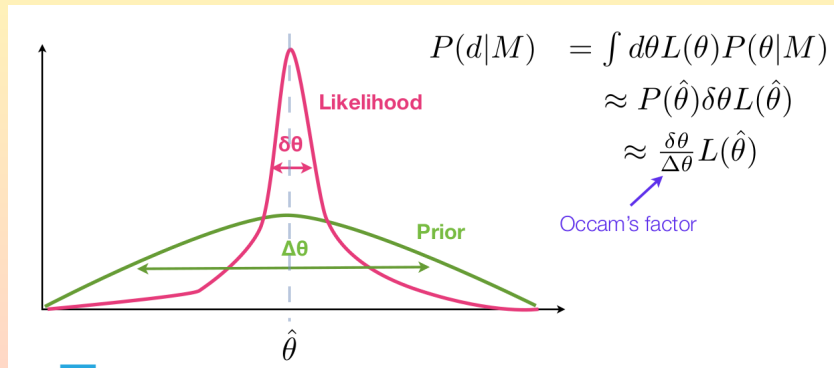


Image from Roberto Trotta, Chair Lemaître Lectures 2018

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

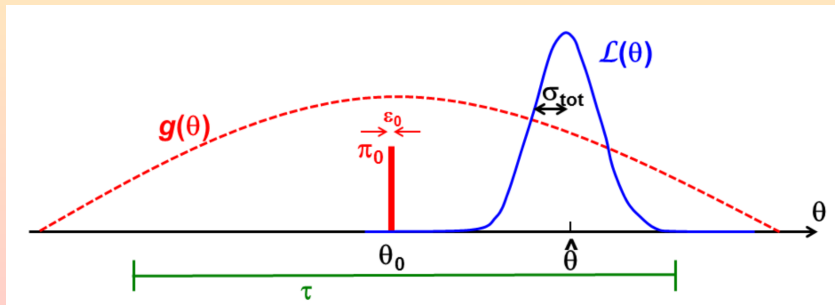


Image from Cousins, doi:10.1007/s11229-014-0525-z



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](http://www.amstat.org) • [www.twitter.com/AmstatNews](https://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](https://doi.org/10.1080/00031305.2016.1154108)

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.

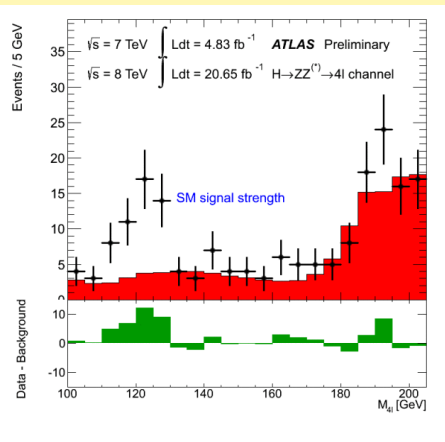
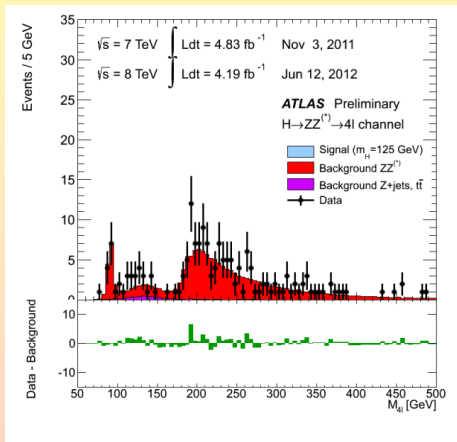
Q: Why do so many people still use  $p = 0.05$ ?

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

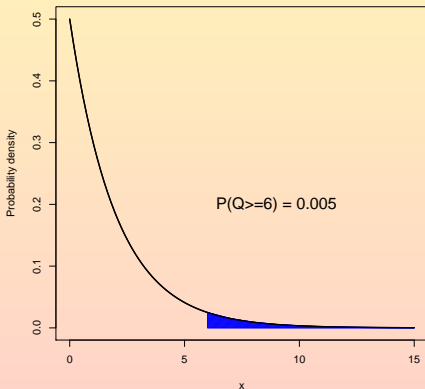
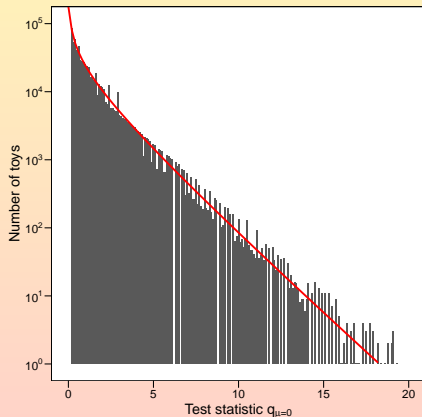


Plot from <https://cds.cern.ch/record/2230893>



- 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  $q \sim \chi^2(1)$ )

Distribution of  $q_{\mu=0}$  for  $H(\mu=0)$



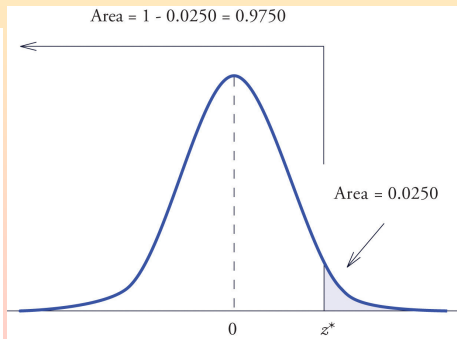
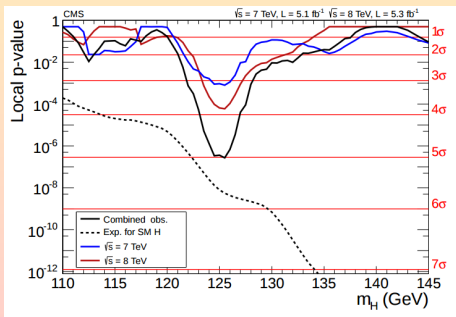
Plots from Vischia—in preparation with Springer

## And the sigmas?

- Just an artifact to convert p-values to easy-to-remember  $\mathcal{O}(1)$  numbers
  - $1\sigma: p = 0.159$
  - $3\sigma: p = 0.00135$
  - $5\sigma: p = 0.000000285$
- 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.  $2\sigma: p_{2sided} = 0.05$ )



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

- ❶ P-values can indicate how incompatible the data are with a specified statistical model.
- ❷ 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.
- ❸ 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.
- ❹ Proper inference requires full reporting and transparency
- ❺ A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- ❻ 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](https://doi.org/10.1080/00031305.2016.1154108)

## Responses to ASA statement: redefine $p$ value threshold or not use it at all

- Benjamin *et al.* ([doi:10.31234/osf.io/mky9j](https://doi.org/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](https://doi.org/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](http://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.”

- 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](#)

841 Accesses | 68 Citations | 69 Altmetric | [Metrics](#)

**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.”

- Ioannidis ([doi:10.1371/journal.pmed.0020124](https://doi.org/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

Comment on this paper

**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>

This article is a preprint and has not been certified by peer review [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.

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

**Nassim Nicholas Taleb** @nntaleb · Apr 11

John Ioannidis does not get that model uncertainty WORSENS possible outcomes under exponential growth & should lead to MORE reaction. Dangerous ignorance. Here is a derivation from Jensen's ineq.

**Ioannidis, dangerously ignorant**

WP, Apr 9 2020, Zakaria: Stanford's John Ioannidis, an epidemiologist who specializes in analyzing data, and one of the most cited scientists in the field, believes we have massively overestimated the fatality of COVID-19. "When you have a model involving exponential growth, if you make a small mistake in the base numbers, you end up with a final number that could be off 10-fold, 30-fold, even 50-fold," he told me.

That ignorant John Ioannidis said that things that grow exponentially AND are subjected to huge errors can lead to... underestimation. **He did not get that uncertainty model error WORSENS the bad outcomes.**

The intuition is that an exponential is convex to the rate of growth: simply  $\frac{d^2 \exp(x)}{dx^2} = \exp(x)$ , and that for all derivatives that remain exponential.

Consider the error rate  $\delta$ . The bias from the error assuming half the time  $r(1+\delta)$ , the other half  $r(1-\delta)$  is  $\xi$ , from Jensen's inequality.

$$\xi = \frac{1}{2} \left( \exp[r(1+\delta)t] + \exp[r(1-\delta)t] \right) - \exp[rt]$$

59 241 956

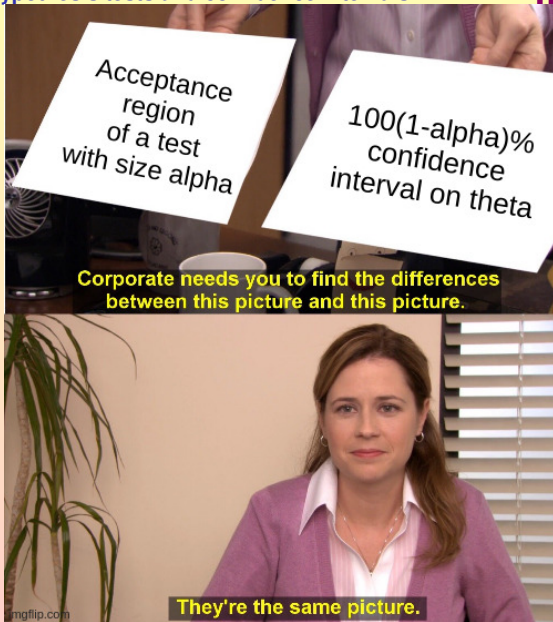
- Goal since LEP: seamless transition between exclusion, observation, discovery (historically for the Higgs)
  - Exclude Higgs as strongly as possible in its absence (in a region where we would be sensitive to its presence)
  - Confirm its existence as strongly as possible in its presence (in a region where we are sensitive to its presence)
  - Maintain Type I and Type II errors below specified (small) levels

- Composite hypothesis  $H(\theta)$  (monodimensional)
- Our tests work for simple hypotheses  $\rightarrow$  make test of simple hypothesis  $H(\theta = \theta_i)$ , scanning values  $\theta_i$  of  $\theta$ 
  - E.g., for the Higgs boson  $\theta$  can be the cross section for a given mass
- Calculate p-value for each test
- Assume our target test size  $\alpha = 0.05$
- Each hypothesis with  $p_\theta < \alpha$  can be excluded at the  $1 - \alpha = 95\%$  confidence level (CL)
  - The set of excluded hypotheses constitutes an *exclusion region*



## What's the difference between these two statements?

- **Acceptance region:** set of *values of the test statistic* for which we don't reject  $H_0$  at significance level  $\alpha$
- $100(1 - \alpha)\%$  **Confidence interval:** set of *values of the parameter  $\theta$*  for which we don't reject  $H_0$  (if  $H_0$  is assumed true)

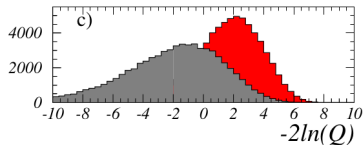
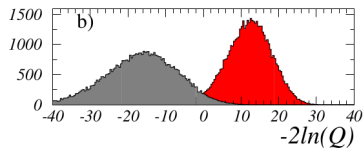
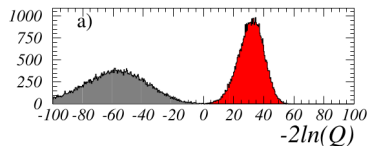


Meme generated with <https://imgflip.com/memegenerator>

- 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!

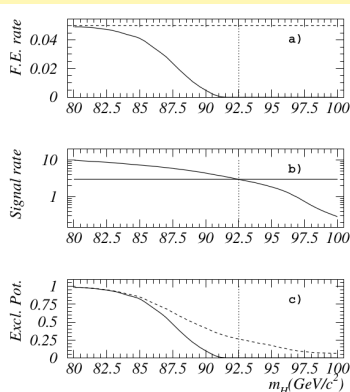
- Example: Find a monotonic  $Q$  for increasing signal-like experiments (e.g. likelihood ratio)
- $CL_{s+b} = P_{s+b}(Q \leq Q_{obs})$ 
  - Small values imply poor compatibility with  $S + B$  hypothesis, favouring  $B$ -only
- Counting experiment: observe  $n$  events
- Assume they come from Poisson processes:  $n \sim Pois(s + b)$ , with known  $b$
- Set limit on  $s$  given  $n_{obs}$
- Exclude values of  $s$  for which  $P(n \leq n_{obs} | s + b) \leq \alpha$  (guaranteed coverage  $1 - \alpha$ )
- $b = 3, n_{obs} = 0$ 
  - Exclude  $s + b \leq 3$  at 95%CL
  - Therefore excluding  $s \leq 0$ , i.e. **all** possible values of  $s$  (can't distinguish  $b$ -only from very-small- $s$ )
- Zech: let's condition on  $n_b \leq n_{obs}$  ( $n_b$  unknown number of background events)
  - For small  $n_b$  the procedure is more likely to undercover than when  $n_b$  is large, and the distribution of  $n_b$  is independent of  $s$
  - $P(n \leq n_{obs} | n_b \leq n_{obs}, s + b) = \dots = \frac{P(n \leq n_{obs} | s + b)}{P(n \leq n_{obs} | b)}$

- 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

- $CL_s := \frac{CL_{s+b}}{CL_b}$
- Exclude the signal hypothesis at confidence level CL if  $1 - CL_s \leq CL$
- Ratio of confidences is not a confidence
  - The hypothetical 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 excluded 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  $H_0 = H(\theta \neq 0)$  and test it against  $H_1 = H(\theta = 0)$ 
  - Test inversion!



Dashed:  $CL_{s+b}$

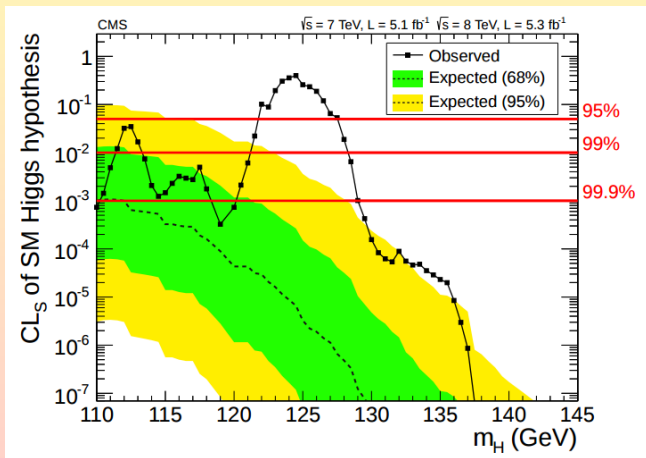
Solid:  $CL_s$

$S < 3$ : 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

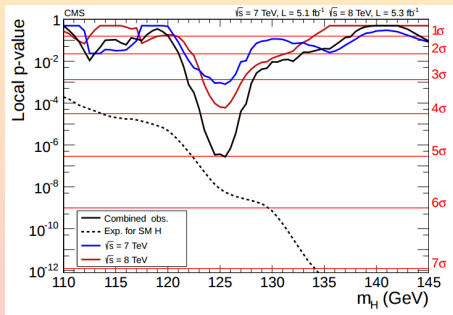
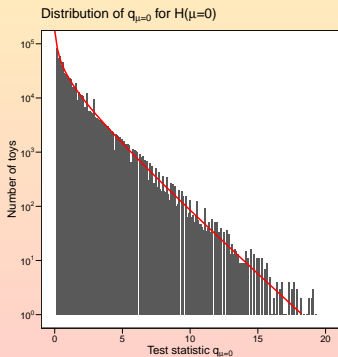


Plot from Higgs discovery paper

- This afternoon we'll play with CLs!



- Quantify the presence of the signal by using the background-only p-value
  - Probability that the background fluctuates yielding an excess as large or larger of the observed one
- For the mass of a resonance,  $q_0 = -2 \log \frac{\mathcal{L}(data|0, \hat{\theta}_0)}{\mathcal{L}(data|\hat{\mu}, \hat{\theta})}$ , with  $\hat{\mu} \geq 0$ 
  - Interested only in upwards fluctuation, accumulate downwards one to zero
- Use pseudo-data to generate background-only Poisson counts and nuisance parameters  $\theta_0^{obs}$ 
  - Use distribution to evaluate tail probability  $p_0 = P(q_0 \leq q_0^{obs})$
  - Convert to one-sided Gaussian tail areas by inverting  $p = \frac{1}{2} P_{\chi^2_1}(Z^2)$



Left plot by Pietro Vischia, right plot from ATL-PHYS-PUB-2011-011 and Higgs discovery paper

- Question time: Significance

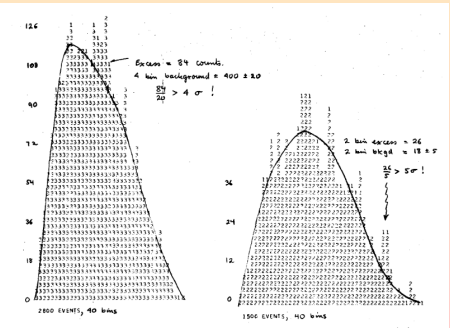
# Fluctuations in HEP? The proposal of a $5\sigma$ criterion

● Rosenfeld, 1968 (<https://escholarship.org/uc/item/6zm2636q>) *Are there any Far-out Mesons or Baryons?*

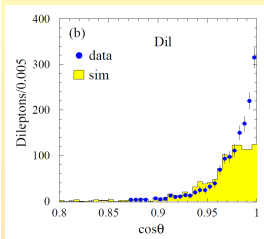
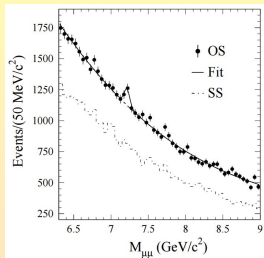
● "In summary of all the discussion above, I conclude that each of our 150,000 annual histograms is capable of generating somewhere between 10 and 100 deceptive upward fluctuations [...] (we should expect several  $4\sigma$  and hundreds of  $3\sigma$  fluctuations"

of  $3\sigma$  fluctuations. What are the implications? To the theoretician or phenomenologist the moral is simple; wait for nearly  $5\sigma$  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  $5\sigma$  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.

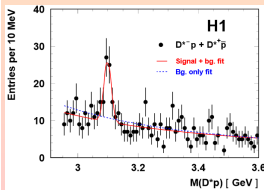
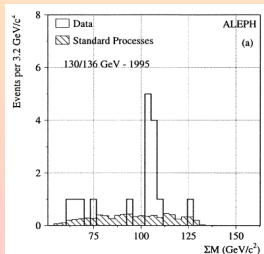
My colleague Gerry Lynch has instead tried to study this problem "experimentally" using a "Las Vegas" computer program called Game. Game is played as follows. You wait until an unsuspecting "friend" comes to show you his latest  $4\sigma$  peak. You draw a smooth curve through his data (based on the 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 phoneyes, rather than the real " $4\sigma$ " 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  $2\sigma$  or  $3\sigma$  fluctuation can easily be amplified to " $4\sigma$ " or " $5\sigma$ "; all it takes is a little enthusiasm.



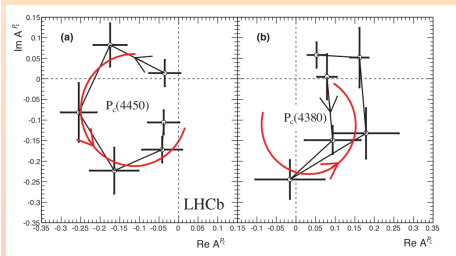
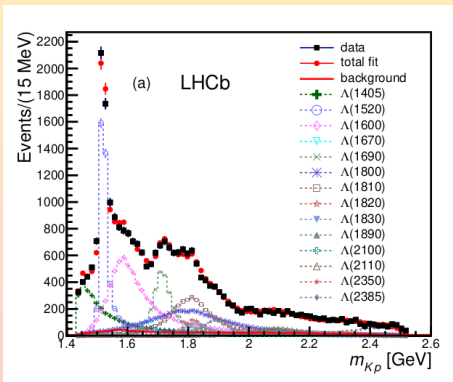
- $3.5\sigma$  (2005, CDF) in dimuon (candidate bottom squark, [doi:/10.1103/PhysRevD.72.092003](https://doi.org/10.1103/PhysRevD.72.092003))



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

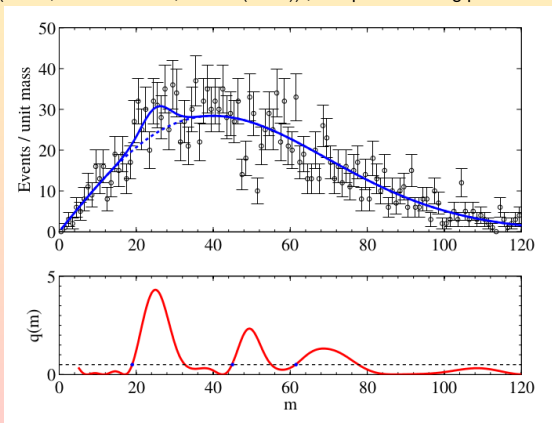


- $9\sigma$  and  $12\sigma$  (2015, LHCb): pentaquarks! ([doi:10.1103/PhysRevLett.115.072001](https://doi.org/10.1103/PhysRevLett.115.072001))
  - Several cross-checks (fit to mass spectrum, fit with non-resonant components, evolution of complex amplitude in Argand diagrams)
  - Mass measurement, soft 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  $5\text{--}6\sigma$  is meaningless
  - Asymptotic approximation not trustable (tail effects). Can run lots of toys but...
  - ...cannot possibly trust knowing your systematic uncertainties to that level



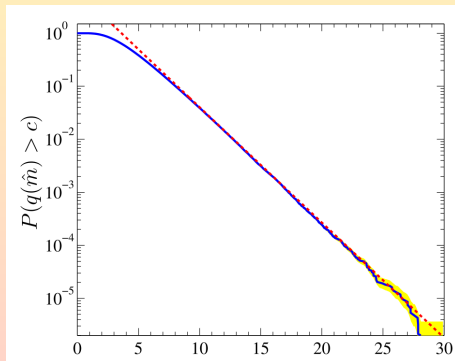
## The Look-elsewhere effect — 1

- Searching for a resonance  $X$  of arbitrary mass
  - $H_0$  = no resonance, the mass of the resonance is not defined (Standard Model)
  - $H_1 = H(M \neq 0)$ , but there are infinite 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  $p_b^{global} := P(q_0(\hat{m}_X) > u) \leq \langle N_u \rangle + \frac{1}{2} P_{\chi_1^2}(u)$
  - $u$  fixed confidence level
  - Crossings (Davis, Biometrika 74, 33–43 (1987)) , computable using pseudo-data (toys)

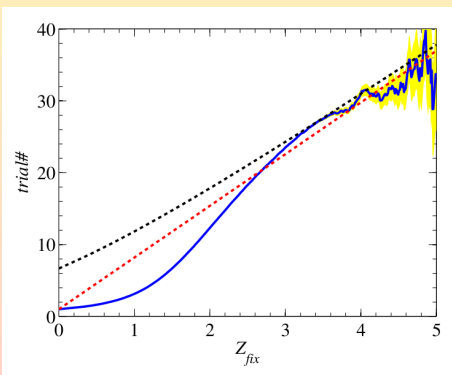


Plot from Gross-Vitells, 10.1140/epjc/s10052-010-1470-8

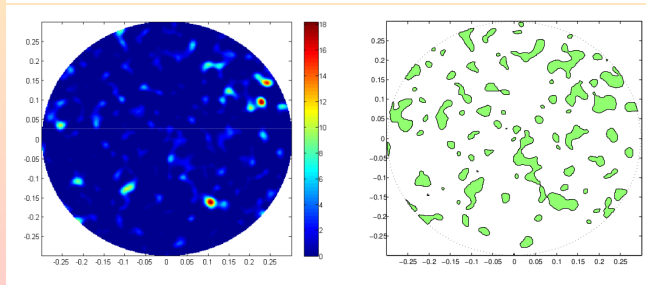
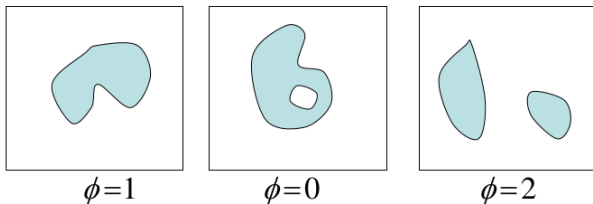
- Ratio of local (excess right here) and global (excess anywhere) p-values: trial factor
- Asymptotically linear in the number of search regions and in the fixed significance level
  - Dashed red lines: prediction based on the formula with upcrossings
  - Blue:  $10^6$  toys (pseudoexperiments)
- Here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



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



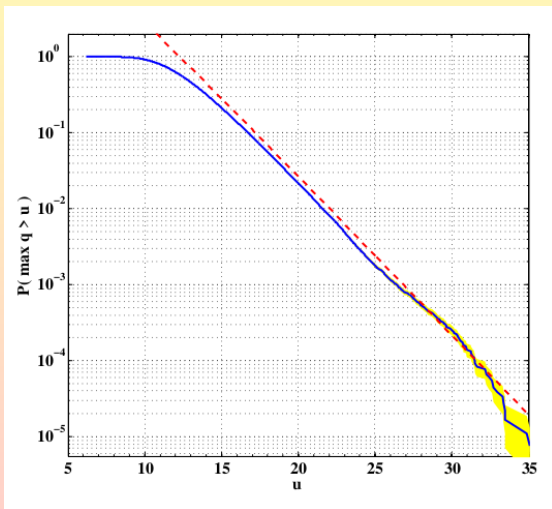
- Extension to two dimensions requires using the theory of random fields
  - Excursion set: set of points for which the value of a field is larger than a threshold  $u$
  - Euler characteristics interpretable as number of disconnected regions minus number of holes



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



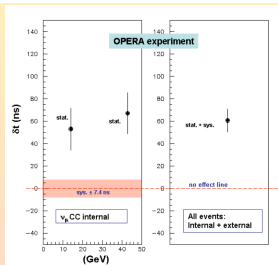
- Asymptoticity holds also for the 2D effect, as desired
  - Dashed red lines: prediction based on the formula with upcrossings
  - Blue: 200k toys (pseudoeperiments)



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

- In 2011 OPERA ([arXiv:1109.4897v1](https://arxiv.org/abs/1109.4897v1)) reported superluminal neutrino speed, with  $6.0\sigma$  significance...

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



- ...but they had a loose cable connector ([doi:10.1007/JHEP10\(2012\)093](https://doi.org/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.

- 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  
 $SEV(\mu \leq \mu_1) = P(Q > Q_{obs}; \mu \leq \mu_1 | \text{false}) = P(Q > Q_{obs}; \mu > \mu_1) > P(Q > Q_{obs}; \mu = \mu_1)$
  - Low severity: your test is not capable of detecting a discrepancy even when if it existed, therefore when not detected is a poor indication 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  
 $SEV(\mu > \mu_1) = P(Q \leq Q_{obs}; \mu > \mu_1 | \text{false}) = P(Q \leq Q_{obs}; \mu \leq \mu_1) > P(Q \leq Q_{obs}; \mu = \mu_1)$
  - Low severity: if probability of higher-than-observed  $Q_{obs}$  is fairly high, then  $Q_{obs}$  not a good indication of effect
  - High severity: if probability of smaller-than-observed  $Q_{obs}$  is very high, then such a large  $Q_{obs}$  indicates a real effect
- Cousins ([arXiv:2002.09713](https://arxiv.org/abs/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.

- Cousins ([doi:/10.1007/s11229-014-0525-z](https://doi.org/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

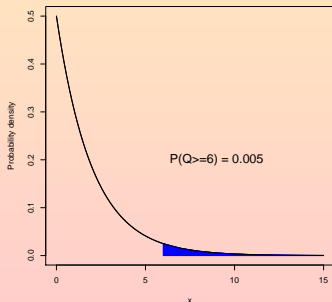
There is a deeper point to be made about core physics models concerning the difference between a model being a good “approximation” in the ordinary sense of the word, and the concept of a mathematical limit. The equations of Newtonian physics have been superseded by those of special and general relativity, 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”)

- 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

## The $\chi^2$ distribution: why degrees of freedom?

- Sample randomly from a Gaussian p.d.f., obtaining  $X_1$  y  $X_2$
- $Q = X_1^2 + X_2^2$  (or in general  $Q = \sum_{i=1}^N X_i^2$ ) is itself a random variable
  - What is  $P(Q \geq 6)$ ? Just integrate the  $\chi^2(N = 2)$  distribution from 6 to  $\infty$
- Depends only on  $N!$ 
  - If we sample 12 times from a Gaussian and compute  $Q = \sum_{i=1}^{12} X_i^2$ , then  $Q \sim \chi^2(N = 12)$
- Theorem: if  $Z_1, \dots, Z_N$  is a sequence of normal random variables, the sum  $V = \sum_{i=1}^N Z_i^2$  is distributed as a  $\chi^2(N)$ 
  - The sum of squares is closely linked to the variance  $E[(X - \mu)^2] = E[X^2] - \mu^2$  from Eq. ??
- The  $\chi^2$  distribution is useful for goodness-of-fit tests that check how much two distributions diverge point-by-point
- It is also the large-sample limit of many distributions (useful to simplify them to a single parameter)



## The $\chi^2$ distribution: goodness-of-fit tests 1/

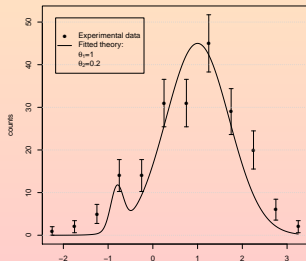
- Consider a set of  $M$  measurements  $\{(X_i, Y_i)\}$ 
  - Suppose  $Y_i$  are affected by a random error representable by a gaussian with variance  $\sigma_i$
- Consider a function  $g(X)$  with predictive capacity, i.e. such that for each  $i$  we have  $g(X_i) \sim Y_i$
- Pearson's  $\chi^2$  function related to the difference between the prediction and the experimental measurement in each point

$$\chi_P^2 := \sum_{i=1}^M \left[ \frac{Y_i - g(X_i)}{\sigma_i} \right]^2 \quad (1)$$

- Neyman's  $\chi^2$  is a similar expression under some assumptions

- If the gaussian error on the measurements is constant, it can be factorized
- If  $Y_i$  represent event counts  $Y_i = n_i$ , then the errors can be approximated with  $\sigma_i \propto \sqrt{n_i}$

$$\chi_N^2 := \sum_{i=1}^M \frac{(n_i - g(X_i))^2}{n_i} \quad (2)$$





- If  $g(X_i) \sim Y_i$  (i.e.  $g(X)$  reasonably predicts the data), then each term of the sum is approximately 1
- Consider a function of  $\chi_{N,P}^2$  and of the number of measurements  $M$ 
  - $E[f(\chi_{N,P}^2, M)] = M$
  - The function is analytically a  $\chi^2$ :

$$f(\chi^2, M) = \frac{2^{-\frac{M}{2}}}{\Gamma\left(\frac{M}{2}\right)} \chi^{M-2} e^{-\frac{\chi^2}{2}} \quad (3)$$

- The cumulative of  $f$  is

$$1 - \text{cum}(f) = P(\chi^2 > \chi_{obs}^2 | g(x) \text{ is the correct model}) \quad (4)$$

- If the p.d.f. under the correct model describes the data well, then within the measured uncertainty it should agree with the data...
  - For about 2/3 of the points, because  $\sigma_i$  represent 68.3% intervals
- ... and  $\chi^2 \simeq M$

- For a given  $M$ , the p.d.f. is known ( $\chi^2(M)$ ) and the observed value can be computed and compared with it
- Cast the problem as an hypothesis test
  - Null hypothesis: there is no difference between prediction and observation (i.e.  $g$  fits well the data)
  - Alternative hypothesis: there is a significant difference between prediction and observation
  - Under the null, the sum of squares is distributed as a  $\chi^2(M)$
  - p-values can be calculated by integration of the  $\chi^2$  distribution

$\chi^2 M \simeq 1 \Rightarrow g(X)$  approximates well the data

Large  $\chi^2 M \gg 1 \Rightarrow$  issues in data or hypothesis (increases  $\chi^2$ ), correlated measurements

Very small  $\chi^2 M \ll 1 \Rightarrow$  overestimated  $\sigma_i$ , or cherrypicked/fraudulent data, or statistically improbable

(5)

- How small/large is a small/large  $\chi^2$ ?
- Very subjective, you must decide by yourself
- Think in terms of p-value:  $P(\chi^2 > \chi_{obs}^2 | g(x) \text{ is the correct model})$ 
  - 0.001 reasonably bad,  $P(\chi^2, M) > 0.001$  expected in 1/1000 cases
  - Often, failed test defined as the infamous  $P(\chi^2, M) < 0.05$
- Problem: the p-value must be calculated by integration
- Can define *reduced*  $\chi^2$  as  $\frac{\chi^2}{M}$ , and translate the previous equation:

$$\frac{\chi^2}{M} \sim 1 \Rightarrow g(X) \text{ approximates well the data}$$

$$\frac{\chi^2}{M} \gg 1 \Rightarrow \text{poor model (increases } \chi^2\text{), or statistically improbable fluctuation} \quad (6)$$

$$\frac{\chi^2}{M} \ll 1 \Rightarrow \text{overestimated } \sigma_i, \text{ or fraudulent data, or statistically improbable fluctuation}$$

- Question time: reduced  $\chi^2$

- How small/large is a small/large  $\chi^2$ ?
- Very subjective, you must decide by yourself
- Think in terms of p-value:  $P(\chi^2 > \chi_{obs}^2 | g(x) \text{ is the correct model})$ 
  - 0.001 reasonably bad,  $P(\chi^2, M) > 0.001$  expected in 1/1000 cases
  - Often, failed test defined as the infamous  $P(\chi^2, M) < 0.05$
- Problem: the p-value must be calculated by integration
- Can define *reduced*  $\chi^2$  as  $\frac{\chi^2}{M}$ , and translate the previous equation:

$$\frac{\chi^2}{M} \sim 1 \Rightarrow g(X) \text{ approximates well the data}$$

$$\frac{\chi^2}{M} \gg 1 \Rightarrow \text{poor model (increases } \chi^2), \text{ or statistically improbable fluctuation} \quad (6)$$

$$\frac{\chi^2}{M} \ll 1 \Rightarrow \text{overestimated } \sigma_i, \text{ or fraudulent data, or statistically improbable fluctuation}$$

- **Question time: reduced  $\chi^2$**
- It's tempting but alone it's misleading! Same  $\chi^2/M$  can lead to opposite answers!
  - For a  $\chi^2/M = 7/5$ , p-value  $p = 0.22$  (reasonably good)
  - For a  $\chi^2/M = 70/50$ , p-value  $p = 0.03$  (reasonably bad)
- If you want to give the ratio, you should always either provide  $M$  or directly the p-value!

- $\chi^2(M)$  tends to a Normal distribution for  $M \rightarrow \infty$ 
  - Slow convergence
  - It is generally not a good idea to substitute a  $\chi^2$  distribution with a Gaussian
- The goodness of fit seen so far is valid only if the model (the function  $g(X)$ ) is fixed
- Sometimes the model has  $k$  free parameters that were not given and that have been fit to the data
- Then the observed value of  $\chi^2$  must be compared with  $\chi^2(N')$ , with  $N' = N - k$  degrees of freedom
  - $N' = N - k$  are called reduced degrees of freedom
  - This however works only if the model is linear in the parameters
  - If the model is not linear in the parameters, when comparing  $\chi_{obs}^2$  with  $\chi^2(N - k)$  then the p-values will be deceptively small!
- Variant of the  $\chi^2$  for small datasets: the G-test
  - $g = 2 \sum O_{ij} \ln(O_{ij}/E_{ij})$
  - It responds better when the number of events is low (Petersen 2012)

- Frederick James: Statistical Methods in Experimental Physics - 2nd Edition, World Scientific
- Glen Cowan: Statistical Data Analysis - Oxford Science Publications
- Louis Lyons: Statistics for Nuclear And Particle Physicists - Cambridge University Press
- Louis Lyons: A Practical Guide to Data Analysis for Physical Science Students - Cambridge University Press
- E.T. Jaynes: Probability Theory - Cambridge University Press 2004
- Annis?, Stuard, Ord, Arnold: Kendall's Advanced Theory Of Statistics I and II
- Pearl, Judea: Causal inference in Statistics, a Primer - Wiley
- R.J.Barlow: A Guide to the Use of Statistical Methods in the Physical Sciences - Wiley
- Kyle Cranmer: Lessons at HCP Summer School 2015
- Kyle Cranmer: Practical Statistics for the LHC - <http://arxiv.org/abs/1503.07622>
- Roberto Trotta: Bayesian Methods in Cosmology - <https://arxiv.org/abs/1701.01467>
- Harrison Prosper: Practical Statistics for LHC Physicists - CERN Academic Training Lectures, 2015 <https://indico.cern.ch/category/72/>
- Christian P. Robert: The Bayesian Choice - Springer
- Sir Harold Jeffreys: Theory of Probability (3rd edition) - Clarendon Press
- Harald Crámer: Mathematical Methods of Statistics - Princeton University Press 1957 edition

# Backup