

# Statistics

or “How to find answers to your questions”

Pietro Vischia<sup>1</sup>

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



LIP Lisboa, LHC Physics Course 2022

# REMEMBER TO START RECORDING

## Why statistics?

### Fundamentals

- Set theory and measure theory
- Frequentist probability
- Bayesian probability

### Random variables and their properties

### Distributions

### Lesson 2

### Estimating a physical quantity

- Sufficiency principle
- Likelihood Principle
- Estimators and maximum likelihood
- Profile likelihood ratio

### Lesson 3

### Confidence Intervals in nontrivial cases

### Test of hypotheses

- CLs
- Significance

### Truth and models

### Summary



- Schedule: two days  $\times$  two hours
- This has evolved to being an excerpt from a longer course (about 10h)
  - You can find additional material, as well as a set of exercises, at <https://agenda.irmp.ucl.ac.be/event/4477/>
  - The exercises, in particular, are designed to show the inner workings of several techniques: try to run them! You can find them at [https://github.com/vischia/intensiveCourse\\_public](https://github.com/vischia/intensiveCourse_public)
- 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
- These slides include some material that we won't be able to cover today
  - Mostly to provide some additional details without having to refer to the full course
  - Slides with this advanced material are those with **the title in red**
- 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!



# Why statistics?

- What is the chance of obtaining a 1 when throwing a six-faced die?
- What is the chance of tomorrow being rainy?

- What is the chance of obtaining a 1 when throwing a six-faced die?
  - We can throw a dice 100 times, and count how many times we obtain 1
- What is the chance of tomorrow being rainy?

- What is the chance of obtaining a 1 when throwing a six-faced die?
  - We can throw a dice 100 times, and count how many times we obtain 1
- What is the chance of tomorrow being rainy?
  - We can try to give an answer based on the recent past weather, but we cannot – in general – *repeat tomorrow* and count

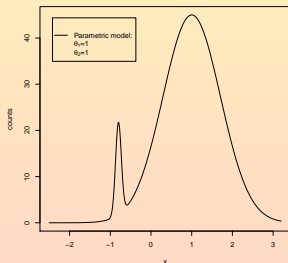
...and about making sure to be posing them in a meaningful way



Image from "The Tiger Lillies" Facebook page

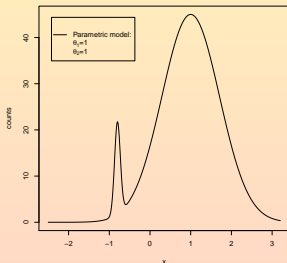
## • Theory

- Approximations
- Free parameters



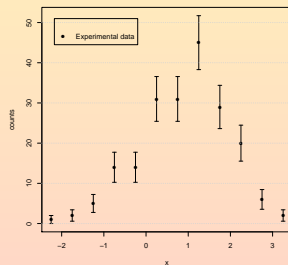
## • Theory

- Approximations
- Free parameters



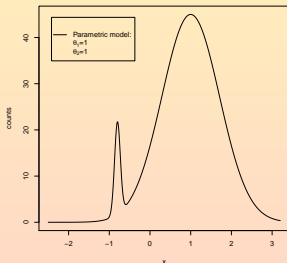
## • Experiment

- Random fluctuations
- Mismeasurements  
(detector effects, etc)



## • Theory

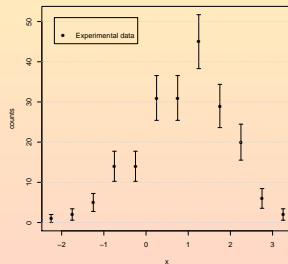
- Approximations
- Free parameters



## • Statistics!

## • Experiment

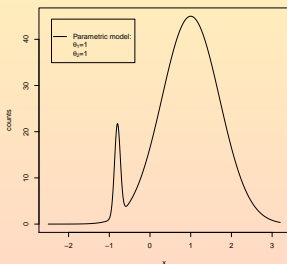
- Random fluctuations
- Mismeasurements (detector effects, etc)





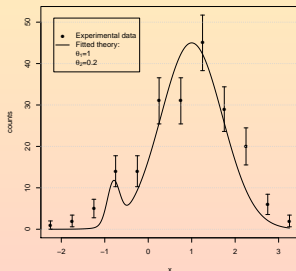
## • Theory

- Approximations
- Free parameters



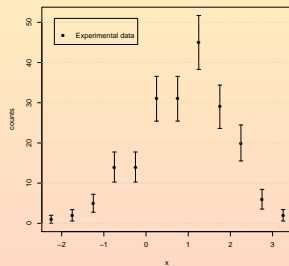
## • Statistics!

- Estimate parameters
- Quantify uncertainty in the parameters estimate
- Test the theory!



## • Experiment

- Random fluctuations
- Mismeasurements (detector effects, etc)



# Fundaments

- $\Omega$ : set of all possible elementary (exclusive) events  $X_i$
- Exclusivity: the occurrence of one event implies that none of the others occur
- Probability then is any function that satisfies the *Kolmogorov axioms*:
  - $P(X_i) \geq 0, \forall i$
  - $P(X_i \text{ or } X_j) = P(X_i) + P(X_j)$
  - $\sum_{\Omega} P(X_i) = 1$

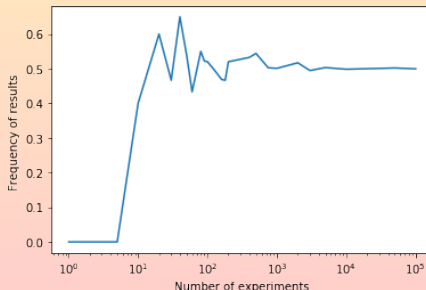


Andrey Kolmogorov.

- Cox postulates: formalize a set of axioms starting from reasonable premises
  - [doi:10.1119/1.1990764](https://doi.org/10.1119/1.1990764)
- Notation
  - $A|B$  the plausibility of the proposition  $A$  given a related proposition  $B$
  - $\sim A$  the proposition “not- $A$ ”, i.e. answering “no” to “*is  $A$  wholly true?*”
  - $F(x, y)$  a function of two variables
  - $S(x)$  a function of one variable
- The two postulates are
  - $C \cdot B|A = F(C|B \cdot A, B|A)$
  - $\sim V|A = S(B|A)$ , i.e.  $(B|A)^m + (\sim B|A)^m = 1$
- Cox theorem acts on propositions, Kolmogorov axioms on sets
- Jaynes adheres to Cox’ exposition and shows that formally this is equivalent to Kolmogorov theory
  - Kolmogorov axioms somehow arbitrary
  - A proposition referring to the real world cannot always be viewed as disjunction of propositions from any meaningful set
  - Continuity as infinite states of knowledge rather than infinite subsets
  - Conditional probability not originally defined

- Repeat a random experiment  $\xi$  (e.g. toss of a die) many times under uniform conditions
  - As uniform as possible
  - $\vec{S}$ : set of all a priori possible different results of an individual measurement
  - $S$ : a fixes subset of  $\vec{S}$
- If in an experiment we obtain  $\xi \in S$ , we will say the event defined by  $\xi \in S$  has occurred
  - We assume that  $S$  is simple enough that we can tell whether  $\xi$  is in it or not
- Throw a die:  $\vec{S} = \{1, 2, 3, 4, 5, 6\}$ 
  - If  $S = \{2, 4, 6\}$ , then  $\xi \in S$  corresponds to the event in which you obtain an even number of points
- Repeat the experiment: among  $n$  repetitions the event has occurred  $\nu$  times
  - Then  $\frac{\nu}{n}$  is the frequency ratio of the event in the sequence of  $n$  experiments
- **Question time: Frequency Ratio**












- Repeat a random experiment  $\xi$  (e.g. toss of a die) many times under uniform conditions
  - As uniform as possible
  - $\vec{S}$ : set of all a priori possible different results of an individual measurement
  - $S$ : a fixes subset of  $\vec{S}$
- If in an experiment we obtain  $\xi \in S$ , we will say the event defined by  $\xi \in S$  has occurred
  - We assume that  $S$  is simple enough that we can tell whether  $\xi$  is in it or not
- Throw a die:  $\vec{S} = \{1, 2, 3, 4, 5, 6\}$ 
  - If  $S = \{2, 4, 6\}$ , then  $\xi \in S$  corresponds to the event in which you obtain an even number of points
- Repeat the experiment: among  $n$  repetitions the event has occurred  $\nu$  times
  - Then  $\frac{\nu}{n}$  is the frequency ratio of the event in the sequence of  $n$  experiments
- Question time: Frequency Ratio
- This afternoon: obtain the answer by simulation!



- The most familiar one: based on the possibility of repeating an experiment many times
- Consider one experiment in which a series of  $N$  events is observed.
- $n$  of those  $N$  events are of type  $X$
- Frequentist probability for any single event to be of type  $X$  is the empirical limit of the frequency ratio:

$$P(X) = \lim_{N \rightarrow \infty} \frac{n}{N}$$

- The experiment must be repeatable in the same conditions
- The job of the physicist is making sure that all the *relevant* conditions in the experiments are the same, and to correct for the unavoidable changes.
  - Yes, *relevant* can be a somehow fuzzy concept
- In some cases, you can directly build the full table of frequencies (e.g. dice throws, poker)
- What if the experiment cannot be repeated, making the concept of frequency ill-defined?

Hand	Distinct Hands	Frequency	Probability	Cumulative probability	Odds	Mathematical expression of absolute frequency
Royal flush 	1	4	0.000154%	0.000154%	649,739 : 1	$\binom{4}{1}$
Straight flush (including royal flush) 	9	36	0.00139%	0.0014%	72,192 : 1	$\binom{10}{1}\binom{4}{1} - \binom{4}{1}$
Four of a kind 	156	624	0.0240%	0.0250%	4,164 : 1	$\binom{13}{1}\binom{12}{1}\binom{4}{1}$
Full house 	156	3,744	0.1441%	0.17%	693 : 1	$\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{1}$
Flush (including royal flush and straight flush) 	1,277	5,108	0.1961%	0.267%	508 : 1	$\binom{13}{5}\binom{4}{1} - \binom{10}{1}\binom{4}{1}$
Straight (including royal flush and straight flush) 	10	16,200	0.3925%	0.76%	264 : 1	$\binom{10}{1}\binom{4}{1}^5 - \binom{10}{1}\binom{4}{1}$
Three of a kind 	858	54,912	2.1128%	2.87%	46.2 : 1	$\binom{13}{1}\binom{4}{3}\binom{12}{2}\binom{4}{1}^2$
Two pair 	858	123,552	4.7838%	7.62%	29.8 : 1	$\binom{13}{2}\binom{4}{2}^2\binom{11}{1}\binom{4}{1}$
One pair 	2,860	1,098,240	42.2569%	49.5%	1.37 : 1	$\binom{13}{1}\binom{4}{2}\binom{12}{3}\binom{4}{1}^3$
High card / High card 	1,277	1,312,540	60.2177%	100%	0.896 : 1	$\left[\binom{13}{5} - 10\right] \left[\binom{4}{1}^5 - 4\right]$
Nothing 	7,462	2,598,960	100%	—	8 : 1	$\binom{52}{5}$



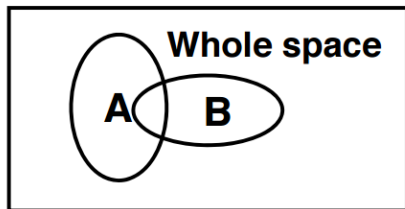
- Based on the concept of degree of belief
  - $P(X)$  is the subjective degree of belief on  $X$  being true
- De Finetti: operative definition of subjective probability, based on the concept of coherent bet
  - We want to determine  $P(X)$ ; we assume that if you bet on  $X$ , you win a fixed amount of money if  $X$  happens, and nothing (0) if  $X$  does not happen
  - In such conditions, it is possible to define the probability of  $X$  happening as

$$P(X) := \frac{\text{The largest amount you are willing to bet}}{\text{The amount you stand to win}} \quad (1)$$

- Coherence is a crucial concept
  - You can leverage your bets in order to try and not loose too much money in case you are wrong
  - Your bookie is doing a Dutch book on you if the set of bets guarantees a profit to him
  - You are doing a Dutch book on your bookie if the set of bets guarantees a profit to you
  - A bet is coherent if a Dutch book is impossible
- This expression is mathematically a Kolmogorov probability!
- Subjective probability is a property of the observer as much as of the observed system
  - It depends on the knowledge of the observer prior to the experiment, and is supposed to change when the observer gains more knowledge (normally thanks to the result of 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

- Interestingly, Venn diagrams were the basis of Kolmogorov approach (Jaynes, 2003)



$$P(A) = \frac{\text{Area of A}}{\text{Area of Whole space}}$$

$$P(B) = \frac{\text{Area of B}}{\text{Area of Whole space}}$$

$$P(A|B) = \frac{\text{Area of A} \cap B}{\text{Area of B}}$$

$$P(B|A) = \frac{\text{Area of A} \cap B}{\text{Area of A}}$$

$$P(A \cap B) = \frac{\text{Area of A} \cap B}{\text{Area of Whole space}}$$

$$P(A) \times P(B|A) = \frac{\text{Area of A}}{\text{Area of Whole space}} \times \frac{\text{Area of A} \cap B}{\text{Area of A}} = \frac{\text{Area of A} \cap B}{\text{Area of Whole space}} = P(A \cap B)$$

$$P(B) \times P(A|B) = \frac{\text{Area of B}}{\text{Area of Whole space}} \times \frac{\text{Area of A} \cap B}{\text{Area of B}} = \frac{\text{Area of A} \cap B}{\text{Area of Whole space}} = P(A \cap B)$$

$$\Rightarrow P(B|A) = P(A|B) \times P(B) / P(A)$$

$$P(A|B) = \frac{\text{small blue oval}}{\text{large blue oval}}$$

$$P(B|A) = \frac{\text{small blue oval}}{\text{small blue oval}}$$

- **Conditional probabilities are not commutative!**  $P(A|B) \neq P(B|A)$
- Example:
  - *speak English*: the person speaks English
  - *have TOEFL*: the person has a TOEFL certificate
- The probability for an English speaker to have a TOEFL certificate,  $P(\text{have TOEFL}|\text{speak English})$ , is very small ( $\ll 1\%$ )
- The probability for a TOEFL certificate holder to speak English,  $P(\text{speak English}|\text{have TOEFL})$ , is (hopefully)  $\ggggg 1\%$  ☺



From [https://www.reddit.com/r/dataisugly/comments/boo6ld/when\\_venn\\_diagram\\_goes\\_wrong/](https://www.reddit.com/r/dataisugly/comments/boo6ld/when_venn_diagram_goes_wrong/)

- Suppose you're on a game show, and you're given the choice of three doors
  - Behind one door is a car;
  - behind the others, goats.
- You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, which has a goat.
- He then says to you, "Do you want to pick door No. 2?"
- Is it to your advantage to switch your choice?

Question time: Monty Hall

- Suppose you're on a game show, and you're given the choice of three doors
  - Behind one door is a car;
  - behind the others, goats.
- You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, which has a goat.
- He then says to you, "Do you want to pick door No. 2?"
- **Is it to your advantage to switch your choice?**  
**Question time: Monty Hall**
- The best strategy is to always switch!
- The key is the presenter knows where the car is → he opens different doors
  - The picture would be different if the presenter opened the door at random

- Suppose you're on a game show, and you're given the choice of three doors
  - Behind one door is a car;
  - behind the others, goats.
- You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, which has a goat.
- He then says to you, "Do you want to pick door No. 2?"
- **Is it to your advantage to switch your choice?**  
**Question time: Monty Hall**
- The best strategy is to always switch!
- The key is the presenter knows where the car is → he opens different doors
  - The picture would be different if the presenter opened the door at random
  - **For the unconvinced: this afternoon we'll build a small simulation to check your answer!**



- Suppose you're on a game show, and you're given the choice of three doors
  - Behind one door is a car;
  - behind the others, goats.
- You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, which has a goat.
- He then says to you, "Do you want to pick door No. 2?"
- Is it to your advantage to switch your choice?

Question time: Monty Hall

- The best strategy is to always switch!
- The key is the presenter knows where the car is → he opens different doors
  - The picture would be different if the presenter opened the door at random
  - For the unconvinced: this afternoon we'll build a small simulation to check your answer!

Behind 1	Behind 2	Behind 3	If you keep 1	If you switch	Presenter opens
Car	Goat	Goat	Win car	Win goat	2 or 3
Goat	Car	Goat	Win goat	Win car	3
Goat	Goat	Car	Win goat	Win car	2

- Bayes Theorem (1763)<sup>1</sup>:

$$P(A|B) := \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

- Valid for any Kolmogorov probability
- The theorem can be expressed also by first starting from a subset  $B$  of the space
- Decomposing the space  $S$  in disjoint sets  $A_i$  (i.e.  $\cap A_i A_j = \emptyset \forall i, j$ ),  $\cup_i A_i = S$  an expression can be given for  $B$  as a function of the  $A_i$ s, the Law of Total Probability:

$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i) \quad (3)$$

- where the second equality holds only for if the  $A_i$ s are disjoint
- Finally, the Bayes Theorem can be rewritten using the decomposition of  $S$  as:

$$P(A|B) := \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)} \quad (4)$$

---

<sup>1</sup> Actually the Bayesian approach has been mainly developed and popularized by Pierre Simon de Laplace

- The Bayes theorem permits to “invert” conditional probabilities, and can be applied to any Kolmogorov probability, therefore in particular to both frequentist and Bayesian definitions
- Let's consider a mortal disease, and label the possible states of the patients
  - D: the patient is diseased (sick)
  - H: the patient is healthy
- Let's imagine we have devised a diagnostic test, characterized by the possible results
  - +: the test is positive to the disease
  - -: the test is negative to the disease
- Imagine the test is very good in identifying sick people:  $P(+|D) = 0.99$ , and that the false positives percentage is very low:  $P(+|H) = 0.01$
- You take the test, and the test is positive. Do you have the disease? Question time: Testing a Disease

- The Bayes theorem permits to “invert” conditional probabilities, and can be applied to any Kolmogorov probability, therefore in particular to both frequentist and Bayesian definitions
- Let's consider a mortal disease, and label the possible states of the patients
  - D: the patient is diseased (sick)
  - H: the patient is healthy
- Let's imagine we have devised a diagnostic test, characterized by the possible results
  - +: the test is positive to the disease
  - -: the test is negative to the disease
- Imagine the test is very good in identifying sick people:  $P(+|D) = 0.99$ , and that the false positives percentage is very low:  $P(+|H) = 0.01$
- You take the test, and the test is positive. Do you have the disease? Question time: Testing a Disease
- By the Bayes Theorem:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|H)P(H)} \quad (5)$$

- The Bayes theorem permits to “invert” conditional probabilities, and can be applied to any Kolmogorov probability, therefore in particular to both frequentist and Bayesian definitions
- Let's consider a mortal disease, and label the possible states of the patients
  - D: the patient is diseased (sick)
  - H: the patient is healthy
- Let's imagine we have devised a diagnostic test, characterized by the possible results
  - +: the test is positive to the disease
  - -: the test is negative to the disease
- Imagine the test is very good in identifying sick people:  $P(+|D) = 0.99$ , and that the false positives percentage is very low:  $P(+|H) = 0.01$
- You take the test, and the test is positive. Do you have the disease? Question time: Testing a Disease
- By the Bayes Theorem:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|H)P(H)} \quad (5)$$

- We need the incidence of the disease in the population,  $P(D)$ ! Back to question time: Testing a Disease

- The Bayes theorem permits to “invert” conditional probabilities, and can be applied to any Kolmogorov probability, therefore in particular to both frequentist and Bayesian definitions
- Let's consider a mortal disease, and label the possible states of the patients
  - D: the patient is diseased (sick)
  - H: the patient is healthy
- Let's imagine we have devised a diagnostic test, characterized by the possible results
  - +: the test is positive to the disease
  - -: the test is negative to the disease
- Imagine the test is very good in identifying sick people:  $P(+|D) = 0.99$ , and that the false positives percentage is very low:  $P(+|H) = 0.01$
- **You take the test, and the test is positive. Do you have the disease? Question time: Testing a Disease**
- By the Bayes Theorem:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|H)P(H)} \quad (5)$$

- We need the incidence of the disease in the population,  $P(D)$ ! **Back to question time: Testing a Disease**
  - It turns out  $P(D)$  is a very important to get our answer
  - $P(D) = 0.001$  (very rare disease): then  $P(D|+) = 0.0902$ , which is fairly small
  - $P(D) = 0.01$  (only a factor 10 more likely): then  $P(D|+) = 0.50$ , which is pretty high
  - $P(D) = 0.1$ : then  $P(D|+) = 0.92$ , almost certainty!

- Frequentist and Subjective probabilities differ in the way of interpreting the probabilities that are written within the Bayes Theorem
- Frequentist: probability is associated to sets of data (i.e. to results of repeatable experiments)
  - Probability is defined as a limit of frequencies
  - Data are considered random, and each point in the space of theories is treated independently
  - An hypothesis is either true or false; improperly, its probability can only be either 0 or 1. In general,  $P(hypothesis)$  is not even defined
  - “This model is preferred” must be read as “I claim that there is a large probability that the data that I would obtain when sampling from the model are similar to the data I already observed”<sup>2</sup>
  - We can only write about  $P(data|model)$
- Bayesian statistics: the definition of probability is extended to the subjective probability of models or hypotheses:

$$P(H|\vec{X}) := \frac{P(\vec{X}|H)\pi(H)}{P(\vec{X})} \quad (6)$$

---

<sup>2</sup>Typically it's difficult to estimate this probability, so one reduces the data to a summary statistic  $S(data)$  with known distribution, and computes how likely is to see  $S(data_{sampled}) = S(data_{obs})$  when sampling from the model

- Bayes' article contains the Bayes theorem (explained via a game of pool)
- A full system for subjective probabilities was (likely) independently developed by Laplace
- In a sense, Laplace is the actual father of Bayesian statistics 😊



More details in the books by Stephen M. Stigler (1996) and Sharon Bertsch McGrayne (2011)



$$P(H|\vec{X}) := \frac{P(\vec{X}|H)\pi(H)}{P(\vec{X})} \quad (7)$$

- $\vec{X}$ , the vector of observed data
- $P(\vec{X}|H)$ , the likelihood function, which fully summarizes the result of the experiment (experimental resolution)
- $\pi(H)$ , the probability of the hypothesis  $H$ . It represents the probability we associate to  $H$  before we perform the experiment
- $P(\vec{X})$ , the probability of the data.
  - Since we already observed them, it is essentially regarded as a normalization factor
  - Summing the probability of the data for all exclusive hypotheses (by the Law of Total Probability),  $\sum_i P(\vec{X}|H_i) = 1$  (assuming that at least one  $H_i$  is true).
  - Usually, the denominator is omitted and the equality sign is replaced by a proportionality sign

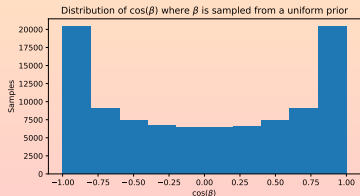
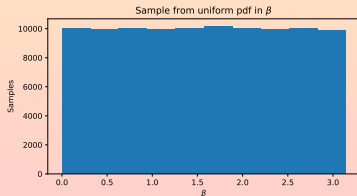
$$P(H|\vec{X}) \propto P(\vec{X}|H)\pi(H) \quad (8)$$

- $P(H|\vec{X})$ , the posterior probability; it is obtained as a result of an experiment
- If we parameterize  $H$  with a (continuous or discrete) parameter, we can use the parameter as a proxy for  $H$ , and instead of writing  $P(H(\theta))$  we write  $P(\theta)$  and

$$P(\theta|\vec{X}) \propto P(\vec{X}|\theta)\pi(\theta) \quad (9)$$

- The simplified expression is usually used, unless when the normalization is necessary
  - “Where is the value of  $\theta$  such that  $\theta_{true} < \theta_c$  with 95% probability?”; integration is needed and the normalization is necessary
  - “Which is the mode of the distribution?”; this is independent of the normalization, and it is therefore not necessary to use the normalized expression

- There is no golden rule for choosing a prior
- Objective Bayesian school: it is necessary to write a golden rule to choose a prior
  - Usually based on an invariance principle
- Consider a theory parameterized with a parameter, e.g. an angle  $\beta$
- Before any experiment, we are Jon Snow about the parameter  $\beta$ : we know nothing
  - We have to choose a very broad prior, or better uniform, in  $\beta$
- Now we interact with a theoretical physicist, who might have built her theory by using as a parameter of the model the cosine of the angle,  $\cos(\beta)$ 
  - In a natural way, she will express her pre-experiment ignorance using an uniform prior **in**  $\cos(\beta)$ .
  - This prior is not constant in  $\beta$ !!!
  - In general, there is no uniquely-defined prior expressing complete ignorance or ambivalence in both parameters ( $\beta$  and  $\cos(\beta)$ )
- We can build a prior invariant for transformations of the parameter, but this means we have to postulate an invariance principle
  - The prior already deviates from our degree of belief about the parameter (“I know nothing”)



- Two ways of solving the situation
  - Objective Bayes: use a formal rule dictated by an invariance principle
  - Subjective Bayes: use something like elicitation of expert opinion
    - Ask an expert her opinion about each value of  $\theta$ , and express the answer as a curve
    - Repeat this with many experts
    - 100 years later check the result of the experiments, thus verifying how many experts were right, and re-calibrate your prior
    - This corresponds to a IF-THEN proposition: "IF the prior is  $\pi(H)$ , THEN you have to update it afterwards, taking into account the result of the experiment"
    - Difficult in practice (query many experts, manage discussion, etc)
- Central concept: update your priors after each experiment

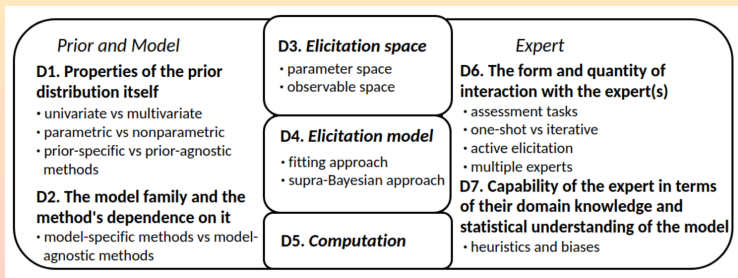


Figure from Mikkola et al. (including Aki Vehtari) [arxiv:2112.01380](https://arxiv.org/abs/2112.01380)

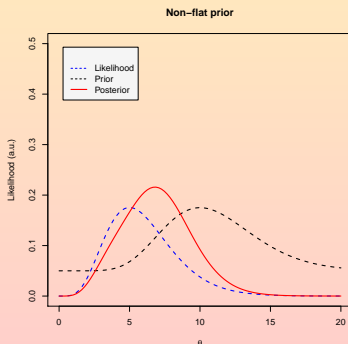
- Identifying small probabilities is also somehow difficult for individuals
  - We saw it with COVID risk vs vaccine risk
- Psychological reasons may lead to incoherences (C.P. Robert 2007, 2e, chapter 3)
  - 44% of the respondents ready to undertake cancer treatment if told *"survival probability is 68%"*
  - 18% of the respondents ready to undertake cancer treatment if told *"probability of death is 32%"*

- In particle physics, the typical application of Bayesian statistics is to put an upper limit on a parameter  $\theta$ 
  - Find a value  $\theta_c$  such that  $P(\theta_{true} < \theta_c) = 95\%$
- Typically  $\theta$  represents the cross section of a physics process, and is proportional to a variable with a Poisson p.d.f.
- An uniform prior can be chosen, eventually restricted to  $\theta \geq 0$  to account for the physical range of  $\theta$
- We can write priors as a function of other variables, but in general those variables will be linked to the cross section by some analytic transformation
  - A prior that is uniforme in a variable is not in general uniform in a transformed variable; a uniform prior in the cross section implies a non-uniform prior (not even linear) on the mass of the sought particle
- In HEP, usually the prior is chosen uniform in the variable with the variable which is proportional to the cross section of the process sought

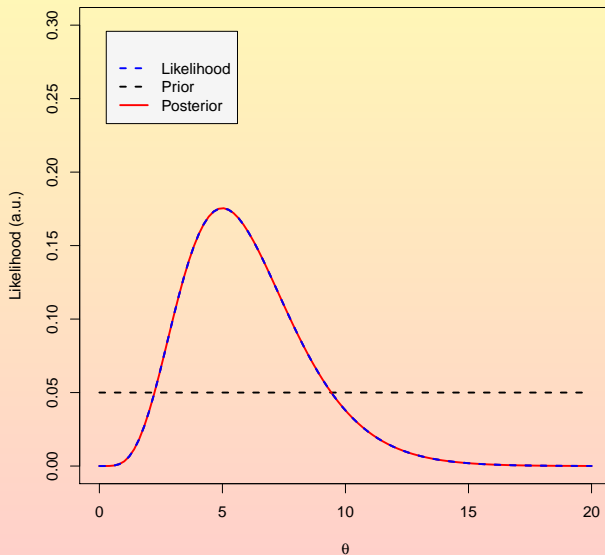
- Uniform priors must make sense
  - Uniform prior across its entire dominion: not very realistic
  - It corresponds to claiming that  $P(1 < \theta \leq 2)$  is the same as  $P(10^{41} < \theta \leq 10^{41} + 1)$
  - It's irrational to claim that a prior can cover uniformly forty orders of magnitude
  - We must have a general idea of “meaningful” values for  $\theta$ , and must not accept results forty orders of magnitude above such meaningful values
- A uniform prior often implies that its integral is infinity (e.g. for a cross section, the dominion being  $[0, \infty]$ )
  - Achieving a proper normalization of the posterior probability would be a nightmare
- In practice, use a very broad prior that falls to zero very slowly but that is practically zero where the parameter cannot meaningfully lie
  - This does not guarantee that it integrates to 1—it depends on the speed of convergence to zero
  - Improper prior

## Choosing a prior in Bayesian statistics; in practice... 3/

- Associating parametric priors to intervals in the parameter space corresponds to considering sets of theories
  - This is because to each value of a parameter corresponds a different theory
- In practical situations, note (Eq. 9) posterior probability is always proportional to the product of the prior and the likelihood
  - The prior must not necessarily be uniform across the whole dominion
  - It should be uniform only in the region in which the likelihood is different from zero
- If the prior  $\pi(\theta)$  is very broad, the product can sometimes be approximated with the likelihood,  $P(\vec{X}|\theta)\pi(H) \sim P(\vec{X}|\theta)$ 
  - The likelihood function is narrower when the data are more precise, which in HEP often translates to the limit  $N \rightarrow \infty$
  - In this limit, the likelihood is always dominant in the product
  - The posterior is independent of the prior!
  - The posteriors corresponding to different priors must coincide, in this limit

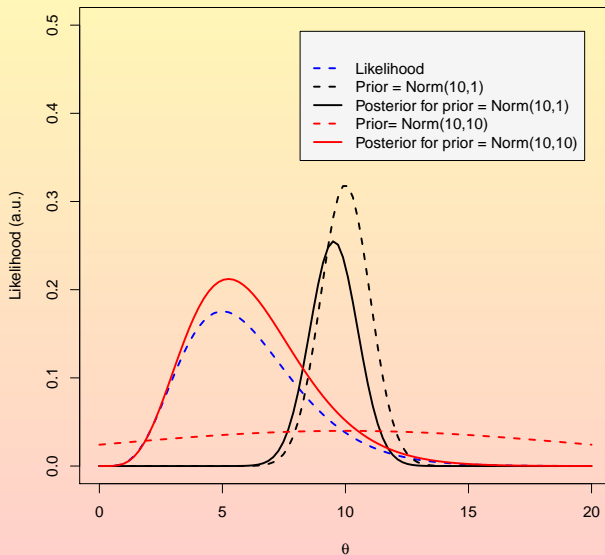


## Flat prior

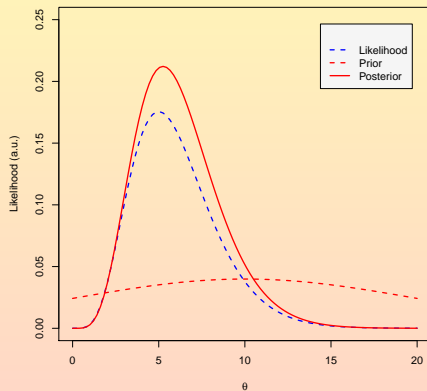




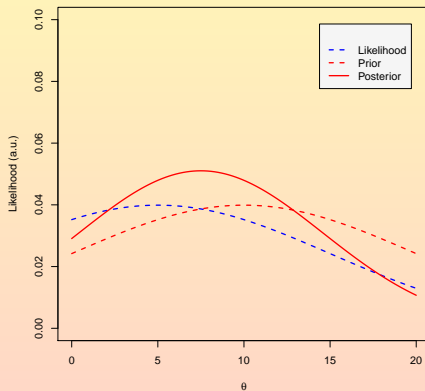
## Broad prior vs narrow prior



Broad prior vs narrow prior



Broad prior vs narrow prior



- The authors of STAN maintain a nice set of recommendations for choosing a prior distribution  
<https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>
  - It is supposed to present a balance between strongly informative priors (judged often unrealistic) and noninformative priors
- Deeply empirical recommendations
  - Give attention to computational constraints
  - A-priori dislike for invariance-principles based priors and Jeffreys priors
- Not necessarily applicable to HEP without debate, but many rather reasonable perspectives
  - Weakly/Strongly informative depends not only on the prior but also on the question you are asking  
*"The prior can often only be understood in the context of the likelihood"*
  - Weak == for a reasonably large amount of data, the likelihood will dominate  
(a "weak" prior might still influence the posterior, if the data are weak)
  - Hard constraints should be reserved to true constraints (e.g. positive-definite parameters)  
(otherwise, choose weakly informative prior on a larger range)
  - Check the posterior dependence on your prior, and perform prior predictive checks  
[doi:10.1111/rssa.12378](https://doi.org/10.1111/rssa.12378)

- Frequentists are restricted to statements related to
  - $P(data|theory)$  (kind of deductive reasoning)
  - The data is considered random
  - Each point in the “theory” phase space is treated independently (no notion of probability in the “theory” space)
  - Repeatable experiments
- Bayesians can address questions in the form
  - $P(theory|data) \propto P(data|theory) \times P(theory)$  (it is intuitively what we normally would like to know)
  - It requires a prior on the theory
  - Huge battle on subjectiveness in the choice of the prior goes here - see §7.5 of James' book

# Drawing some histograms

- **Random variable:** a numeric label for each element in the space of data (in frequentist statistics) or in the space of the hypotheses (in Bayesian statistics)
- In Physics, usually we assume that Nature can be described by continuous variables
  - The discreteness of our distributions would arise from scanning the variable in a discrete way
  - Experimental limitations in the act of measuring an intrinsically continuous variable)
- Instead of point probabilities we'll work with probabilities defined in intervals, normalized w.r.t. the interval:

$$f(X) := \lim_{\Delta X \rightarrow 0} \frac{P(X)}{\Delta X} \quad (10)$$

- Dimensionally, they are densities and they are called probability density functions (p.d.f. s)
- Inverting the expression,  $P(X) = \int f(X)dX$  and we can compute the probability of an interval as a definite interval

$$P(a < X < b) := \int_a^b f(X)dX \quad (11)$$

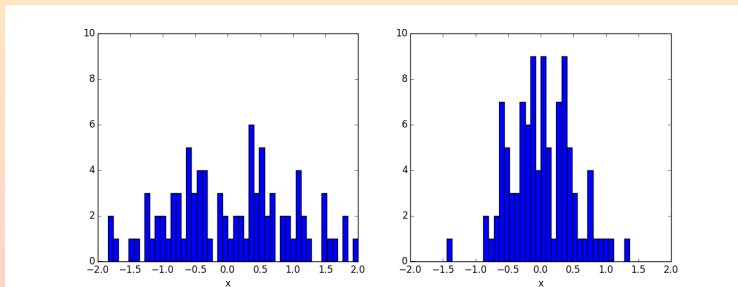
- Extend the concept of p.d.f. to an arbitrary number of variables; the joint p.d.f.  $f(X, Y, \dots)$
- If we are interested in the p.d.f. of just one of the variables the joint p.d.f. depends upon, we can compute by integration the marginal p.d.f.

$$f_X(X) := \int f(X, Y) dY \quad (12)$$

- Sometimes it's interesting to express the joint p.d.f. as a function of one variable, for a particular fixed value of the others: this is the conditional p.d.f. :

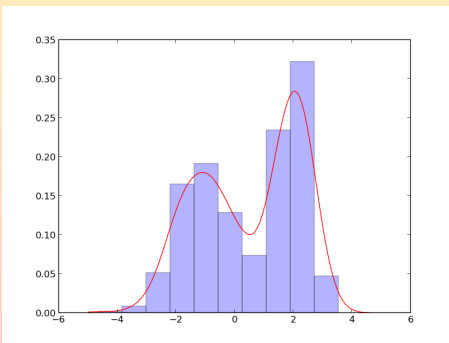
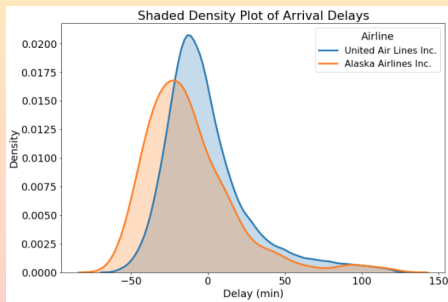
$$f(X|Y) := \frac{f(X, Y)}{f_Y(Y)} \quad (13)$$

- Repeated experiments usually don't yield the exact same result even if the physical quantity is expected to be exactly the same
  - Random changes occur because of the imperfect experimental conditions and techniques
  - They are connected to the concept of dispersion around a central value
- When repeating an experiment, we can count how many times we obtain a result contained in various intervals (e.g. how often  $1.0 \leq L < 1.1$ , how often  $1.1 \leq L < 1.2$ , etc)
  - An histogram can be a natural way of recording these frequencies
  - The concept of dispersion of measurements is therefore related to that of dispersion of a distribution
- In a distribution we are usually interested in finding a “central” value and how much the various results are dispersed around it



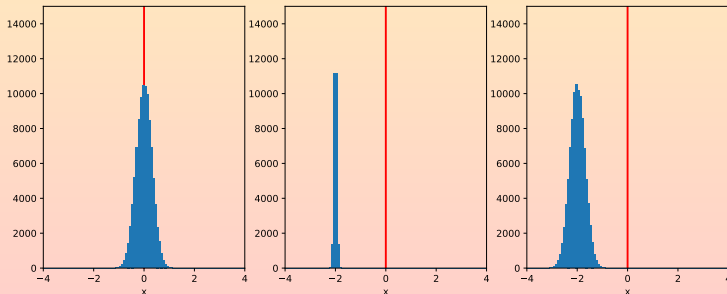


- HEP uses histograms mostly historically: counting experiments
- Statistics and Machine Learning communities typically use densities
  - Intuitive relationship with the underlying p.d.f.
  - Kernel density estimates: binning assumption  $\rightarrow$  bandwidth assumption
  - Less focused on individual bin content, more focused on the overall shape
  - More general notion (no stress about the limited bin content in tails)
- In HEP, if your events are then used “as counting experiment” it’s more useful the histogram
  - But for some applications (e.g. Machine Learning) even in HEP please consider using density estimates



Plots from TheGlowingPython and TowardsDataScience

- Two fundamentally different kinds of uncertainties
  - Error: the deviation of a measured quantity from the true value (bias)
  - Uncertainty: the spread of the sampling distribution of the measurements
- **Random (statistical) uncertainties**
  - Inability of any measuring device (and scientist) to give infinitely accurate answers
  - Even for integral quantities (e.g. counting experiments), fluctuations occur in observations on a small sample drawn from a large population
  - They manifest as spread of answers scattered around the true value
- **Systematic uncertainties**
  - They result in measurements that are simply wrong, for some reason
  - They manifest usually as offset from the true value, even if all the individual results can be consistent with each other



- We define the expected value and mathematical expectation

$$E[X] := \int_{\Omega} Xf(X)dX \quad (14)$$

- In general, for each of the following formulas (reported for continuous variables) there is a corresponding one for discrete variables, e.g.

$$E[X] := \sum_i X_i P(X_i) \quad (15)$$

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

$$E[g] := \int_{\Omega} g(X)f(X)dX \quad (16)$$

- The previous expression Eq. 14 is a special case of Eq. 16 when  $g(X) = X$
- The mean of  $X$  is:

$$\mu := E[X] \quad (17)$$

- The variance of  $X$  is:

$$V(X) := E[(X - \mu)^2] = E[X^2] - (E[X])^2 = E[X^2] - \mu^2 \quad (18)$$

- 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  $g(\vec{X})$  be a function of more variables,  $\vec{X} = (X_1, X_2, \dots, X_n)$  (with p.d.f.  $f(\vec{X})$ )

- Expected value:  $E(g(\vec{X})) = \int g(\vec{X})f(\vec{X})dX_1dX_2\dots dX_n = \mu_g$
- Variance:  $V[g] = E[(g - \mu_g)^2] = \int (g(\vec{X}) - \mu_g)^2 f(\vec{X})dX_1dX_2\dots dX_n = \sigma_g^2$

- Covariance:** of two variables  $X, Y$ :

$$V_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X\mu_Y = \int XYf(X, Y)dXdY - \mu_X\mu_Y$$

- 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!
- $\rho_{XY}$  is related to the angle in a linear regression of  $X$  on  $Y$  (or viceversa)

From: Glen Cowan, Statistical Data Analysis (OUP 1998)

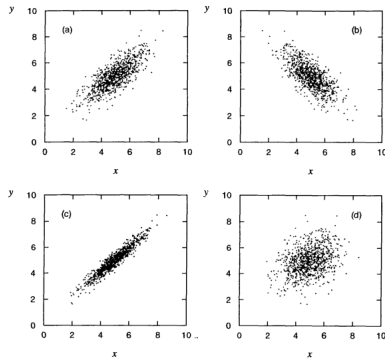


Fig. 1.9 Scatter plots of random variables  $x$  and  $y$  with (a) a positive correlation,  $\rho = 0.75$ , (b) a negative correlation,  $\rho = -0.75$ , (c)  $\rho = 0.95$ , and (d)  $\rho = 0.25$ . For all four cases the standard deviations of  $x$  and  $y$  are  $\sigma_x = \sigma_y = 1$ .

## Let's make it funnier: more variables!

- Let our function  $g(\vec{X})$  be a function of more variables,  $\vec{X} = (X_1, X_2, \dots, X_n)$  (with p.d.f.  $f(\vec{X})$ )

- Expected value:  $E(g(\vec{X})) = \int g(\vec{X})f(\vec{X})dX_1dX_2\dots dX_n = \mu_g$
- Variance:  $V[g] = E[(g - \mu_g)^2] = \int (g(\vec{X}) - \mu_g)^2 f(\vec{X})dX_1dX_2\dots dX_n = \sigma_g^2$

- Covariance:** of two variables  $X, Y$ :

$$V_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X\mu_Y = \int XYf(X, Y)dXdY - \mu_X\mu_Y$$

- 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!
- $\rho_{XY}$  is related to the angle in a linear regression of  $X$  on  $Y$  (or viceversa)
  - It does not capture non-linear correlations

Question time: CorrCoeff

From: Glen Cowan, Statistical Data Analysis (OUP 1998)

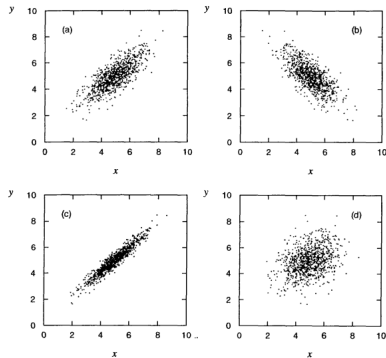


Fig. 1.9 Scatter plots of random variables  $x$  and  $y$  with (a) a positive correlation,  $\rho = 0.75$ , (b) a negative correlation,  $\rho = -0.75$ , (c)  $\rho = 0.95$ , and (d)  $\rho = 0.25$ . For all four cases the standard deviations of  $x$  and  $y$  are  $\sigma_x = \sigma_y = 1$ .

- Informs on the direction (co-increase, increase-decrease, none) of a linear correlation
- Does NOT inform on the slope of the correlation
- Several non-linear correlations yield  $\rho_{XY}$

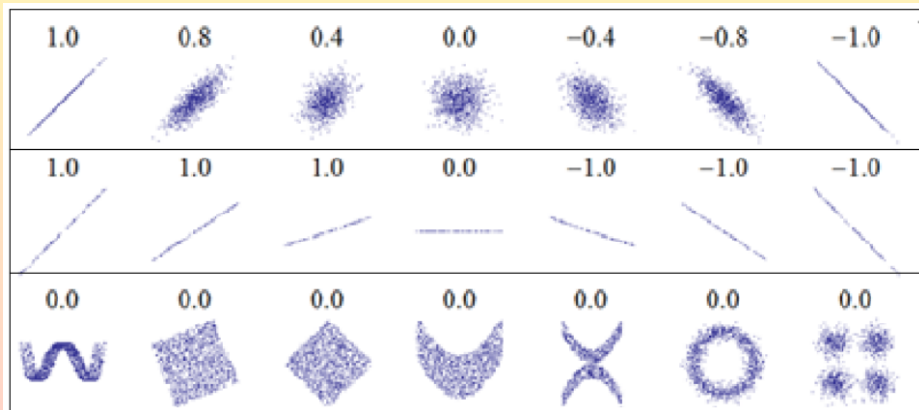


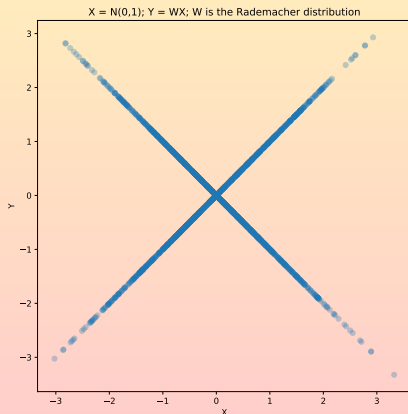
Figure from BND2010

## 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

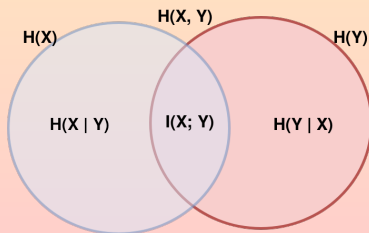
$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left( \frac{p(x, y)}{p_1(x)p_2(y)} \right)$$

- 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

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

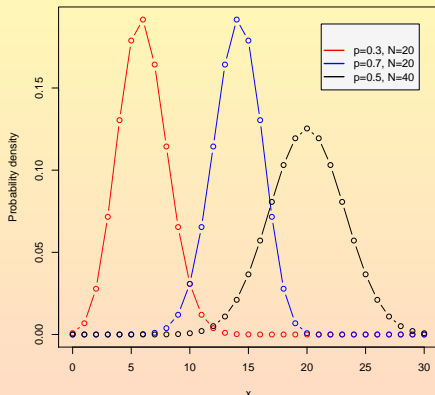




## Binomial

- Discrete variable:  $r$ , positive integer  $\leq N$
- Parameters:
  - $N$ , positive integer
  - $p$ ,  $0 \leq p \leq 1$
- Probability function:
 
$$P(r) = \binom{N}{r} p^r (1-p)^{N-r}, r = 0, 1, \dots, N$$
- $E(r) = Np$ ,  $V(r) = Np(1-p)$
- Usage: probability of finding exactly  $r$  successes in  $N$  trials. The distribution of the number of events in a single bin of a histogram is binomial (if the bin contents are independent)

Binomial p.d.f.

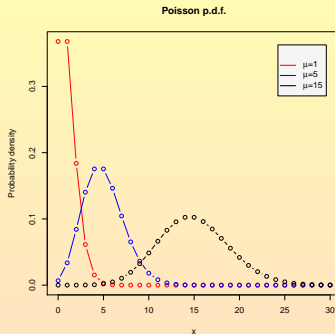


- Example: which is the probability of obtaining 3 times the number 6 when throwing a 6-faces die 12 times?
- $N = 12, r = 3, p = \frac{1}{6}$
- $$P(3) = \binom{12}{3} \left(\frac{1}{6}\right)^3 \left(1 - \frac{1}{6}\right)^{12-3} = \frac{12!}{3!9!} \frac{1}{6^3} \left(\frac{5}{6}\right)^9 = 0.1974$$

## • Poisson

- Discrete variable:  $r$ , positive integer
- Parameter:  $\mu$ , positive real number
- Probability function:  $P(r) = \frac{\mu^r e^{-\mu}}{r!}$
- $E(r) = \mu$ ,  $V(r) = \mu$
- Usage: probability of finding exactly  $r$  events in a given amount of time, if events occur at a constant rate.

- Example: is it convenient to put an advertising panel along a road?



- Probability that at least one car passes through the road on each day, knowing on average 3 cars pass each day

- $P(X > 0) = 1 - P(0)$ , and use Poisson p.d.f.

$$P(0) = \frac{3^0 e^{-3}}{0!} = 0.049787$$

- $P(X > 0) = 1 - 0.049787 = 0.95021$ .

- Now suppose the road serves only an industry, so it is unused during the weekend; Which is the probability that in any given day exactly one car passes by the road?

$$N_{\text{avg per dia}} = \frac{3}{5} = 0.6$$

$$P(X) = \frac{0.6^1 e^{-0.6}}{1!} = 0.32929$$

## • Gaussian or Normal distribution

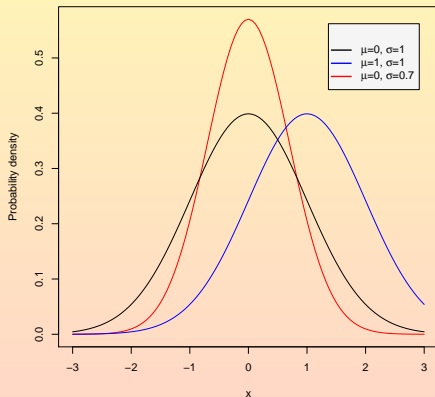
- Variable:  $X$ , real number
- Parameters:
  - $\mu$ , real number
  - $\sigma$ , positive real number

- Probability function:

$$f(X) = N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(X-\mu)^2}{\sigma^2}\right]$$

- $E(X) = \mu$ ,  $V(X) = \sigma^2$
- Usage: describes the distribution of independent random variables. It is also the high-something limit for many other distributions

Gaussian p.d.f.



- Parameter: integer  $N > 0$  degrees of freedom
- Continuous variable  $X \in \mathcal{R}$
- p.d.f., expected value, variance

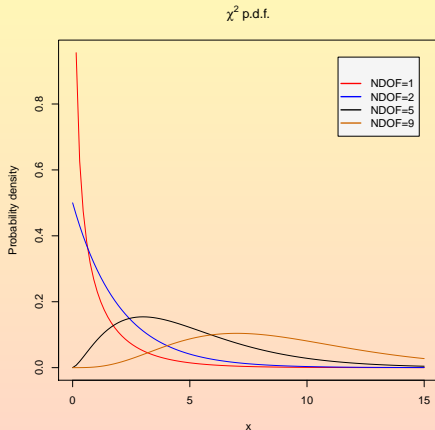
$$f(X) = \frac{\frac{1}{2} \left(\frac{X}{2}\right)^{\frac{N}{2}-1} e^{-\frac{X}{2}}}{\Gamma\left(\frac{N}{2}\right)}$$

$$E[r] = N$$

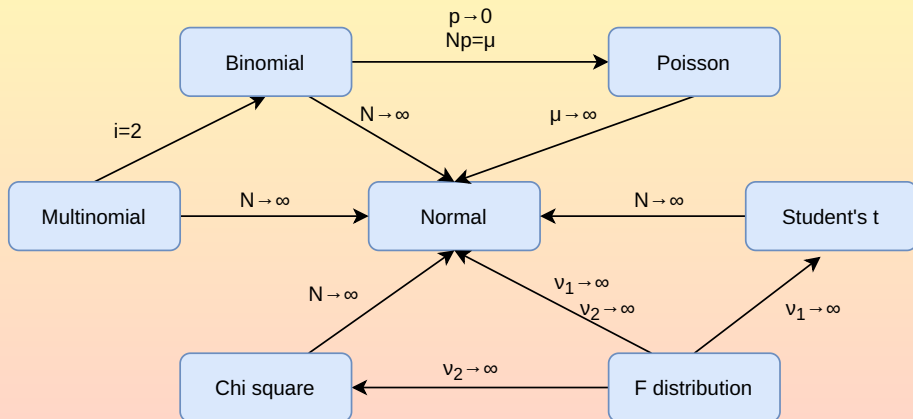
$$V(r) = 2N$$

- It describes the distribution of the sum of the squares of a random variable,  $\sum_{i=1}^N X_i^2$

Reminder:  $\Gamma() := \frac{N!}{r!(N-r)!}$



- It is often convenient to know the asymptotic properties of the various distributions



# End of Lesson 1

## Why statistics?

### Fundamentals

- Set theory and measure theory
- Frequentist probability
- Bayesian probability

### Random variables and their properties

### Distributions

### Lesson 2

#### Estimating a physical quantity

- Sufficiency principle
- Likelihood Principle
- Estimators and maximum likelihood
- Profile likelihood ratio

### Lesson 3

#### Confidence Intervals in nontrivial cases

#### Test of hypotheses

- CLs
- Significance

#### Truth and models

#### Summary

# Lesson 2

## Point and Interval estimation

# Estimating a physical quantity

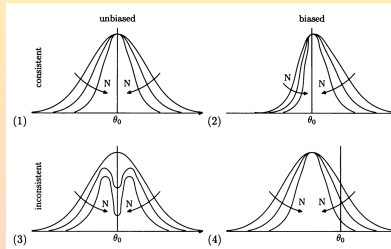


## Estimators

- Set  $\vec{x} = (x_1, \dots, x_N)$  of  $N$  statistically independent observations  $x_i$ , sampled from a p.d.f.  $f(x)$ .
- Mean and width of  $f(x)$  (or some parameter of it:  $f(x; \vec{\theta})$ , with  $\vec{\theta} = (\theta_1, \dots, \theta_M)$  unknown)
  - In case of a linear p.d.f., the vector of parameters would be  $\vec{\theta} = (\text{intercept}, \text{slope})$
- We call estimator a function of the observed data  $\vec{x}$  which returns numerical values  $\hat{\vec{\theta}}$  for the vector  $\vec{\theta}$ .
- $\hat{\vec{\theta}}$  is (asymptotically) consistent if it converges to  $\vec{\theta}_{true}$  for large  $N$ :

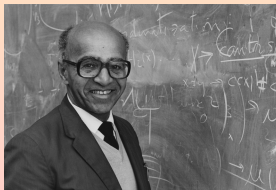
$$\lim_{N \rightarrow \infty} \hat{\vec{\theta}} = \vec{\theta}_{true}$$

- $\hat{\vec{\theta}}$  is unbiased if its bias is zero,  $\vec{b} = 0$ 
  - Bias of  $\hat{\vec{\theta}}$ :  $\vec{b} := E[\hat{\vec{\theta}}] - \vec{\theta}_{true}$
  - If bias is known, can redefine  $\hat{\vec{\theta}}' = \hat{\vec{\theta}} - \vec{b}$ , resulting in  $\vec{b}' = 0$ .
- $\hat{\vec{\theta}}$  is efficient if its variance  $V[\hat{\vec{\theta}}]$  is the smallest possible
- An estimator is robust when it is insensitive to small deviations from the underlying distribution (p.d.f.) assumed (ideally, one would want distribution-free estimates, without assumptions on the underlying p.d.f.)



Plot from James, 2nd ed.

- A test statistic is a function of the data (a quantity derived from the data sample)
- When  $X \sim f(X|\theta)$ , a statistic  $T = T(X)$  is sufficient for  $\theta$  if the density function  $f(X|T)$  is independent of  $\theta$ 
  - If  $T$  is a sufficient statistic for  $\theta$ , then also any strictly monotonic  $g(T)$  is sufficient for  $\theta$
- Minimal sufficient statistic: a sufficient statistic that is a function of all other sufficient statistics for  $\theta$
- The statistic  $T$  carries as much information about  $\theta$  as the original data  $X$ 
  - No other function can give any further information about  $\theta$
  - Same inference from data  $X$  with model  $M$  and from sufficient statistic  $T(X)$  with model  $M'$
- **Rao–Blackwell theorem**: if  $g(X)$  is an estimator for  $\theta$  and  $T$  is a sufficient statistic, then the conditional expectation of  $g(X)$  given  $T(X)$  is never a worse estimator of  $\theta$ 
  - Practical procedure: build a ballpark estimator  $g(X)$ , then condition it on a  $T(X)$  to obtain a better estimator
- **The Sufficiency Principle**: Two observations  $X$  and  $Y$  that factorize through the same value of  $T(\cdot)$ , i.e. s.t.  $T(x) = T(y)$ , must lead to the same inference about  $\theta$



Images from AmStat magazine and from Illinois.edu

- Given some data 1, 2, 3, 4, 5, you may want to estimate the population mean
  - Consider the sample mean  $\hat{x} = \frac{1+2+3+4+5}{5} = 3$  as an estimator of the sample mean (3 is the estimate)
  - Imagine we don't have the data; we only know that the sample mean is 3
  - Is the sample mean a sufficient statistic? Question time: Sufficient statistic

- Given some data 1, 2, 3, 4, 5, you may want to estimate the population mean
  - Consider the sample mean  $\hat{x} = \frac{1+2+3+4+5}{5} = 3$  as an estimator of the sample mean (3 is the estimate)
  - Imagine we don't have the data; we only know that the sample mean is 3
  - **Is the sample mean a sufficient statistic? Question time: Sufficient statistic**
  - If you only knew the sample mean of 3, you would estimate the population mean to be 3 anyway, regardless of having the data or not
  - Knowing the data (the set 1, 2, 3, 4, 5) or knowing only the sample mean does not improve our estimate for the population mean

- Given some data 1, 2, 3, 4, 5, you may want to estimate the population mean
  - Consider the sample mean  $\hat{x} = \frac{1+2+3+4+5}{5} = 3$  as an estimator of the sample mean (3 is the estimate)
  - Imagine we don't have the data; we only know that the sample mean is 3
  - **Is the sample mean a sufficient statistic? Question time: Sufficient statistic**
  - If you only knew the sample mean of 3, you would estimate the population mean to be 3 anyway, regardless of having the data or not
  - Knowing the data (the set 1, 2, 3, 4, 5) or knowing only the sample mean does not improve our estimate for the population mean
- Estimate the binomial probability of obtaining  $r$  heads in  $N$  coin tosses
  - Record heads and tails, with their order: *HTTHHHHTHTTTHTHTH*
  - **Can we somehow improve by identifying a sufficient statistic? Question time: Sufficient Statistic**

- Given some data 1, 2, 3, 4, 5, you may want to estimate the population mean
  - Consider the sample mean  $\hat{x} = \frac{1+2+3+4+5}{5} = 3$  as an estimator of the sample mean (3 is the estimate)
  - Imagine we don't have the data; we only know that the sample mean is 3
  - **Is the sample mean a sufficient statistic? Question time: Sufficient statistic**
  - If you only knew the sample mean of 3, you would estimate the population mean to be 3 anyway, regardless of having the data or not
  - Knowing the data (the set 1, 2, 3, 4, 5) or knowing only the sample mean does not improve our estimate for the population mean
- Estimate the binomial probability of obtaining  $r$  heads in  $N$  coin tosses
  - Record heads and tails, with their order: *HTTHHHTHHTTTHTHTH*
  - **Can we somehow improve by identifying a sufficient statistic? Question time: Sufficient Statistic**
  - What happens if we record only the number of heads? (remember that the binomial p.d.f. is:  
 $P(r) = \binom{N}{r} p^r (1-p)^{N-r}, r = 0, 1, \dots, N$ )

- Given some data 1, 2, 3, 4, 5, you may want to estimate the population mean
  - Consider the sample mean  $\hat{x} = \frac{1+2+3+4+5}{5} = 3$  as an estimator of the sample mean (3 is the estimate)
  - Imagine we don't have the data; we only know that the sample mean is 3
  - **Is the sample mean a sufficient statistic? Question time: Sufficient statistic**
  - If you only knew the sample mean of 3, you would estimate the population mean to be 3 anyway, regardless of having the data or not
  - Knowing the data (the set 1, 2, 3, 4, 5) or knowing only the sample mean does not improve our estimate for the population mean
- Estimate the binomial probability of obtaining  $r$  heads in  $N$  coin tosses
  - Record heads and tails, with their order: *HTTHHHHTHTTHTHTH*
  - **Can we somehow improve by identifying a sufficient statistic? Question time: Sufficient Statistic**
  - What happens if we record only the number of heads? (remember that the binomial p.d.f. is:  $P(r) = \binom{N}{r} p^r (1-p)^{N-r}$ ,  $r = 0, 1, \dots, N$ )
  - Recording only the number of heads (no tails, no order) gives exactly the same information
  - Data can be reduced; we only need to store a sufficient statistic (the distribution  $f(X|T)$  is independent of  $\theta$ )
  - **Storage needs are reduced!!!**

- Pivotal quantity: its distribution does not depend on the parameters

- For a  $Gaus(\mu, \sigma^2)$  p.d.f.,  $\frac{\bar{X} - \mu}{S/\sqrt{N}} \sim t_{student}$  is a pivot
- See exercise this afternoon



- Ancillary statistic for a parameter  $\theta$ : a statistic  $f(X)$  which does not depend on  $\theta$ 
  - Concept linked to that of *(minimal) sufficient statistic*; (maximal) data reduction while retaining all Fisher information about  $\theta$
- Can an ancillary statistic can give information about  $\theta$  even if it does not depend on it? **QT!**  
Ancillary



- Pivotal quantity: its distribution does not depend on the parameters

- For a  $Gauss(\mu, \sigma^2)$  p.d.f.,  $\frac{\bar{X} - \mu}{S/\sqrt{N}} \sim t_{student}$  is a pivot
- See exercise this afternoon



- Ancillary statistic for a parameter  $\theta$ : a statistic  $f(X)$  which does not depend on  $\theta$ 
  - Concept linked to that of (minimal) sufficient statistic; (maximal) data reduction while retaining all Fisher information about  $\theta$
- Can an ancillary statistic can give information about  $\theta$  even if it does not depend on it? **QT!**

## Ancillary

- Yes!
  - Sample  $X_1$  and  $X_2$  from  $P_\theta(X = \theta) = P_\theta(X = \theta + 1) = P_\theta(X = \theta + 2) = \frac{1}{3}$
  - Ancillary statistic:  $R := X_2 - X_1$  (no information about  $\theta$ )
  - Minimal sufficient statistic:  $M := \frac{X_1 + X_2}{2}$
  - Sample point ( $M = m, R = r$ ): either  $\theta = m$ , or  $\theta = m - 1$ , or  $\theta = m - 2$
  - If  $R = 2$ , then necessarily  $X_1 = m - 1$  and  $X_2 = m - 2$ ; Therefore necessarily  $\theta = m - 1$

- Pivotal quantity: its distribution does not depend on the parameters

- For a  $Gauss(\mu, \sigma^2)$  p.d.f.,  $\frac{\bar{X} - \mu}{S/\sqrt{N}} \sim t_{student}$  is a pivot
- See exercise this afternoon



- Ancillary statistic for a parameter  $\theta$ : a statistic  $f(X)$  which does not depend on  $\theta$ 
  - Concept linked to that of (minimal) sufficient statistic; (maximal) data reduction while retaining all Fisher information about  $\theta$
- Can an ancillary statistic can give information about  $\theta$  even if it does not depend on it? **QT!**

### Ancillary

- Yes!
  - Sample  $X_1$  and  $X_2$  from  $P_\theta(X = \theta) = P_\theta(X = \theta + 1) = P_\theta(X = \theta + 2) = \frac{1}{3}$
  - Ancillary statistic:  $R := X_2 - X_1$  (no information about  $\theta$ )
  - Minimal sufficient statistic:  $M := \frac{X_1 + X_2}{2}$
  - Sample point ( $M = m, R = r$ ): either  $\theta = m$ , or  $\theta = m - 1$ , or  $\theta = m - 2$
  - If  $R = 2$ , then necessarily  $X_1 = m - 1$  and  $X_2 = m - 2$ ; Therefore necessarily  $\theta = m - 1$
- Knowledge of  $R$  alone carries no information on  $\theta$ , but increases the precision on an estimate of  $\theta$  (Cox, Efron, Hinckley)!
- Powerful tool to improve data reduction capabilities (save money...)
- Also employed for asymptotic likelihood expressions
  - Also impact on approximate expressions for significance

- The information of a set of observations should increase with the number of observations
  - Double the data should result in double the information if the data are independent
- Information should be conditional on what we want to learn from the experiment
  - Data which are irrelevant to our hypothesis should carry zero information relative to our hypothesis
- Information should be related to precision
  - The greatest the information carried by the data, the better the precision of our result

- Common enunciation: given a set of observed data  $\vec{x}$ , the likelihood function  $L(\vec{x}; \theta)$  contains all the information that is relevant to the estimation of the parameter  $\theta$  contained in the data sample
  - The likelihood function is seen as a function of  $\theta$ , for a fixed set (a particular realization) of observed data  $\vec{x}$
  - The likelihood is used to define the information contained in a sample

- Bayesian statistics automatically satisfies the likelihood principle
  - $P(\theta|\vec{x}) \propto L(\vec{x}; \theta) \times \pi(\theta)$ : the only quantity depending on the data is the likelihood
  - *Information* as a broad way of saying *all the possible inferences about  $\theta$*
  - “Probably tomorrow will rain”
- Frequentist statistics: *information* more strictly as *Fisher information* (connection with curvature of  $L(\vec{x}; \theta)$ )
  - Usually does not comply (have to consider the hypothetical set of data that might have been obtained)
  - Need to recast question in terms of hypothetical data
  - Example: tail areas from sampling distributions obtained with toys
  - Even in forecasts: computer simulations of the day of tomorrow, or counting the past frequency of correct forecasts by the grandpa feeling arthritis in the shoulder
  - “The sentence -tomorrow it will rain- is probably true”
- The Likelihood Principle is quite vague: no practical prescription for drawing inference from the likelihood
  - Bayesian Maximum a-posteriori (MAP) estimator automatically maximizes likelihood
  - Maximum Likelihood estimator (MLE) maximizes likelihood automatically, but some foundational issues

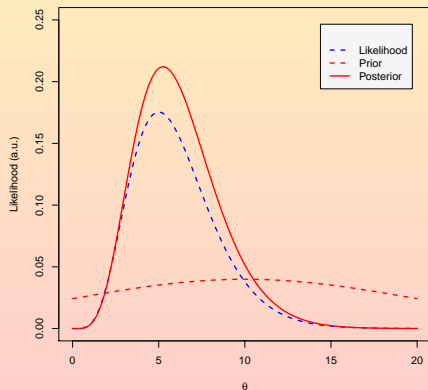
- Two likelihoods differing by only a normalization factor are equivalent
  - Implies that information resides in the shape of the likelihood
- George Bernard: replace a dataset  $D$  with a dataset  $D + Z$ , where  $Z$  is the result of tossing a coin
  - Assume that the coin toss is independent on the parameter  $\theta$  you seek to determine
  - Sampling probability:  $p(DZ|\theta) = p(D|\theta)p(Z)$
  - The coin toss tells us nothing about the parameter  $\theta$  beyond what we already learn by considering  $D$  only
  - Any inference we do with  $D$  must therefore be the same as any inference we do with  $D + Z$
  - In particular, normalizations cancel out in ratio:  $\frac{\mathcal{L}_1}{\mathcal{L}_2} = \frac{p(DZ|\theta_1 I)}{p(DZ|\theta_2 I)} = \frac{p(D|\theta_1 I)}{p(D|\theta_2 I)}$
- Do you believe probability comes from the imperfect knowledge of the observer?
  - Then the likelihood principle does not seem too profound besides the mathematical simplifications it allows
- Do you believe that probability is a physical phenomenon arising from *randomness*?
  - Then the likelihood principle has for you a profound meaning of valid principle of inference

## Likelihood and Fisher Information

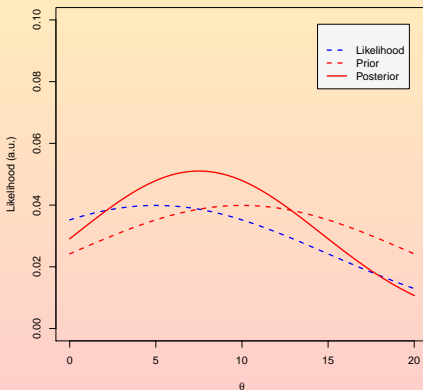
- A very narrow likelihood will provide much information about  $\theta_{true}$ 
  - The posterior probability will be more localized than the prior in the regimen in which the likelihood function dominates the product  $L(\vec{x}; \vec{\theta}) \times \pi$
  - Ideally we'd want to connect this with the Fisher Information, which therefore be large
- A very broad likelihood will not carry much information, and ideally the computed Fisher Information will be small
- What's a reasonable definition of Fisher Information based on the likelihood function?

Question time: Likelihood and Information

Broad prior vs narrow prior



Broad prior vs narrow prior



- Score:  $\frac{\partial}{\partial \theta} \ln L(X; \theta)$
- Under broad regularity conditions, if  $X \sim f(x|\theta_{true})$  the expectation of the score calculated for  $\theta = \theta_{true}$  is zero

$$E\left[\frac{\partial}{\partial \theta} \ln L(X; \theta) | \theta = \theta_{true}\right] = \frac{\partial}{\partial \theta} \int f(x|\theta_{true}) dx = \frac{\partial}{\partial \theta} 1 = 0$$

- **Fisher Information:** the variance of the score

$$I(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \ln L(X; \theta)\right)^2 | \theta_{true}\right] = \int \left(\frac{\partial}{\partial \theta} \ln f(x|\theta)\right)^2 f(x|\theta) dx \geq 0$$

- Under some regularity conditions, and when the likelihood is twice differentiable, then you can “exchange” the exponent and the number of derivations

$$I(\theta) = -E\left[\left(\frac{\partial^2}{\partial \theta^2} \ln L(X; \theta)\right) | \theta_{true}\right]$$



- The narrowness of the likelihood can be estimated by looking at its curvature
- The curvature is the second derivative with respect to the parameter of interest
- A very narrow (peaked) likelihood is characterized by a very large and positive curvature  $-\frac{\partial^2 \ln L}{\partial \theta^2}$
- The second derivative of the likelihood is linked to the Fisher Information

$$I(\theta) = -E \left[ \frac{\partial^2 \ln L}{\partial \theta^2} \right] = E \left[ \left( \frac{\partial \ln L}{\partial \theta} \right)^2 \right]$$

## Fisher Information and Jeffreys priors

- When changing variable, the change of parameterization must not result in a change of the information
  - The information is a property of the data only, through the likelihood—that summarizes them completely (likelihood principle)
- Search for a parametrization  $\theta'(\theta)$  in which the Fisher Information is constant
- Compute the prior as a function of the new variable

$$\begin{aligned}\pi(\theta) = \pi(\theta') \left| \frac{d\theta'}{d\theta} \right| &\propto \sqrt{E \left[ \left( \frac{\partial \ln N}{\partial \theta'} \right)^2 \right] \left| \frac{\partial \theta'}{\partial \theta} \right|} \\ &= \sqrt{E \left[ \left( \frac{\partial \ln L}{\partial \theta'} \frac{\partial \theta'}{\partial \theta} \right)^2 \right]} \\ &= \sqrt{E \left[ \left( \frac{\partial \ln L}{\partial \theta} \right)^2 \right]} \\ &= \sqrt{I(\theta)}\end{aligned}$$

- For any  $\theta$ ,  $\pi(\theta) = \sqrt{I(\theta)}$ ; with this choice, the information is constant under changes of variable
- Such priors are called Jeffreys priors, and assume different forms depending on the type of parametrization
  - Location parameters: uniform prior
  - Scale parameters: prior  $\propto \frac{1}{\theta}$
  - Poisson processes: prior  $\propto \frac{1}{\sqrt{\theta}}$

## The Maximum Likelihood Method 1/

- Let  $\vec{x} = (x_1, \dots, x_N)$  be a set of  $N$  statistically independent observations  $x_i$ , sampled from a p.d.f.  $f(x; \vec{\theta})$  depending on a vector of parameters
- Under independence of the observations, the likelihood function factorizes into the individual p.d.f. s

$$L(\vec{x}; \vec{\theta}) = \prod_{i=1}^N f(x_i, \vec{\theta})$$

- The maximum-likelihood estimator is the  $\vec{\theta}_{ML}$  which maximizes the joint likelihood

$$\vec{\theta}_{ML} := \operatorname{argmax}_{\theta} \left( L(\vec{x}, \vec{\theta}) \right)$$

- The maximum must be global
- Numerically, it's usually easier to minimize

$$- \ln L(\vec{x}; \vec{\theta}) = - \sum_{i=1}^N \ln f(x_i, \vec{\theta})$$

- Easier working with sums than with products
- Easier minimizing than maximizing
- If the minimum is far from the range of permitted values for  $\vec{\theta}$ , then the minimization can be performed by finding solutions to

$$- \frac{\ln L(\vec{x}; \vec{\theta})}{\partial \theta_j} = 0$$

- It is assumed that the p.d.f. s are correctly normalized, i.e. that  $\int f(\vec{x}; \vec{\theta}) dx = 1$  ( $\rightarrow$  integral does not depend on  $\vec{\theta}$ )

- Solutions to the likelihood minimization are found via numerical methods such as MINOS
  - Fred James' Minuit: <https://root.cern.ch/root/html/doc/guides/minuit2/Minuit2.html>
- $\vec{\theta}_{ML}$  is an estimator  $\rightarrow$  let's study its properties!
  - 1 **Consistent:**  $\lim_{N \rightarrow \infty} \vec{\theta}_{ML} = \vec{\theta}_{true}$ ;
  - 2 **Unbiased:** only asymptotically.  $\vec{b} \propto \frac{1}{N}$ , so  $\vec{b} = 0$  only for  $N \rightarrow \infty$ ;
  - 3 **Efficient:**  $V[\vec{\theta}_{ML}] = \frac{1}{I(\theta)}$
  - 4 **Invariant:** for change of variables  $\psi = g(\theta)$ ;  $\hat{\psi}_{ML} = g(\vec{\theta}_{ML})$
- $\vec{\theta}_{ML}$  is only asymptotically unbiased, and therefore it does not always represent the best trade-off between bias and variance
- Remember that in frequentist statistics  $L(\vec{x}|\vec{\theta})$  is not a p.d.f.. In Bayesian statistics, the posterior probability is a p.d.f.:

$$P(\vec{\theta}|\vec{x}) = \frac{L(\vec{x}|\vec{\theta})\pi(\vec{\theta})}{\int L(\vec{x}|\vec{\theta})\pi(\vec{\theta})d\vec{\theta}}$$

- Note that if the prior is uniform,  $\pi(\vec{\theta}) = k$ , then the MLE is also the maximum of the posterior probability,  $\vec{\theta}_{ML} = \max P(\vec{\theta}|\vec{x})$ .

- A nuclear decay with half-life  $\tau$  is described by the p.d.f., expected value, and variance

$$f(t; \tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$

$$E[f] = \tau$$

$$V[f] = \tau^2$$

- Sampling  $N$  independent measurements  $t_i$  from the same p.d.f. results in a set of measurements identically distributed
- **Exercise: compute the MLE for this p.d.f.**

- A nuclear decay with half-life  $\tau$  is described by the p.d.f., expected value, and variance

$$f(t; \tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$

$$E[f] = \tau$$

$$V[f] = \tau^2$$

- Sampling  $N$  independent measurements  $t_i$  from the same p.d.f. results in a set of measurements identically distributed
- **Exercise: compute the MLE for this p.d.f.**
- The joint p.d.f. can be factorized

$$f(t_1, \dots, t_N; \tau) = \prod_i f(t_i; \tau)$$

- A nuclear decay with half-life  $\tau$  is described by the p.d.f., expected value, and variance

$$f(t; \tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$

$$E[f] = \tau$$

$$V[f] = \tau^2$$

- Sampling  $N$  independent measurements  $t_i$  from the same p.d.f. results in a set of measurements identically distributed
- **Exercise: compute the MLE for this p.d.f.**
- The joint p.d.f. can be factorized

$$f(t_1, \dots, t_N; \tau) = \prod_i f(t_i; \tau)$$

- For a particular set of  $N$  measurements  $t_i$ , the p.d.f. can be written as a function of  $\tau$  only,  
 $L(\tau) := f(t_i; \tau)$

- A nuclear decay with half-life  $\tau$  is described by the p.d.f., expected value, and variance

$$f(t; \tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$

$$E[f] = \tau$$

$$V[f] = \tau^2$$

- Sampling  $N$  independent measurements  $t_i$  from the same p.d.f. results in a set of measurements identically distributed
- **Exercise: compute the MLE for this p.d.f.**
- The joint p.d.f. can be factorized

$$f(t_1, \dots, t_N; \tau) = \prod_i f(t_i; \tau)$$

- For a particular set of  $N$  measurements  $t_i$ , the p.d.f. can be written as a function of  $\tau$  only,  
 $L(\tau) := f(t_i; \tau)$
- **Now all you need to do is to maximize the likelihood**



- A nuclear decay with half-life  $\tau$  is described by the p.d.f., expected value, and variance

$$f(t; \tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$

$$E[f] = \tau$$

$$V[f] = \tau^2$$

- Sampling  $N$  independent measurements  $t_i$  from the same p.d.f. results in a set of measurements identically distributed
- **Exercise: compute the MLE for this p.d.f.**
- The joint p.d.f. can be factorized

$$f(t_1, \dots, t_N; \tau) = \prod_i f(t_i; \tau)$$

- For a particular set of  $N$  measurements  $t_i$ , the p.d.f. can be written as a function of  $\tau$  only,  $L(\tau) := f(t_i; \tau)$
- **Now all you need to do is to maximize the likelihood**
- The logarithm of the likelihood,  $\ln L(\tau) = \sum \left( \ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$ , can be maximized analytically

$$\frac{\partial \ln L(\tau)}{\partial \tau} = \sum_i \left( -\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) \equiv 0$$

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, \dots, t_N) = \frac{1}{N} \sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, \dots, t_N) = \frac{1}{N} \sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1
- The expected value is  $E[\hat{\tau}] = \tau$ , and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, \dots, t_N) = \frac{1}{N} \sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1
- The expected value is  $E[\hat{\tau}] = \tau$ , and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- What is the variance? Which is its relationship to  $N$ ? Is the estimator efficient? QT: N D 1

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, \dots, t_N) = \frac{1}{N} \sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- **What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1**
- The expected value is  $E[\hat{\tau}] = \tau$ , and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- **What is the variance? Which is its relationship to  $N$ ? Is the estimator efficient? QT: N D 1**
- The variance interestingly decreases when  $N$  increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N} \sum_i t_i\right] = \frac{1}{N^2} \sum_i V[t_i] = \frac{\tau^2}{N}$$

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, \dots, t_N) = \frac{1}{N} \sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- **What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1**
- The expected value is  $E[\hat{\tau}] = \tau$ , and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- **What is the variance? Which is its relationship to  $N$ ? Is the estimator efficient? QT: N D 1**
- The variance interestingly decreases when  $N$  increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N} \sum_i t_i\right] = \frac{1}{N^2} \sum_i V[t_i] = \frac{\tau^2}{N}$$

- The MLE is not the only estimator we can think of. **Fill the table!**

	Consistent	Unbiased	Efficient
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1 + \dots + t_N}{N}$			
$\hat{\tau} = \frac{t_1 + \dots + t_N}{N-1}$			
$\hat{\tau} = t_i$			

**Table:** Properties of different estimators of the half life for a nuclear decay.

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, \dots, t_N) = \frac{1}{N} \sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- **What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1**
- The expected value is  $E[\hat{\tau}] = \tau$ , and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- **What is the variance? Which is its relationship to  $N$ ? Is the estimator efficient? QT: N D 1**
- The variance interestingly decreases when  $N$  increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N} \sum_i t_i\right] = \frac{1}{N^2} \sum_i V[t_i] = \frac{\tau^2}{N}$$

- The MLE is not the only estimator we can think of. **Fill the table!**

	Consistent	Unbiased	Efficient
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1 + \dots + t_N}{N}$	✓	✓	✓
$\hat{\tau} = \frac{t_1 + \dots + t_N}{N-1}$			
$\hat{\tau} = t_i$			

**Table:** Properties of different estimators of the half life for a nuclear decay.

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, \dots, t_N) = \frac{1}{N} \sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- **What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1**
- The expected value is  $E[\hat{\tau}] = \tau$ , and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- **What is the variance? Which is its relationship to  $N$ ? Is the estimator efficient? QT: N D 1**
- The variance interestingly decreases when  $N$  increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N} \sum_i t_i\right] = \frac{1}{N^2} \sum_i V[t_i] = \frac{\tau^2}{N}$$

- The MLE is not the only estimator we can think of. **Fill the table! Question time: Nuclear Decay 2**

	Consistent	Unbiased	Efficient
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1 + \dots + t_N}{N}$	✓	✓	✓
$\hat{\tau} = \frac{t_1 + \dots + t_N}{N-1}$	✓	✗	✗
$\hat{\tau} = t_i$			

**Table:** Properties of different estimators of the half life for a nuclear decay.



## Nuclear Decay with Maximum Likelihood Method

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, \dots, t_N) = \frac{1}{N} \sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased? Question time: Nuclear Decay 1
- The expected value is  $E[\hat{\tau}] = \tau$ , and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- What is the variance? Which is its relationship to  $N$ ? Is the estimator efficient? QT: N D 1
- The variance interestingly decreases when  $N$  increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N} \sum_i t_i\right] = \frac{1}{N^2} \sum_i V[t_i] = \frac{\tau^2}{N}$$

- The MLE is not the only estimator we can think of. Fill the table! Question time: Nuclear Decay 2

	Consistent	Unbiased	Efficient
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1 + \dots + t_N}{N}$	✓	✓	✓
$\hat{\tau} = \frac{t_1 + \dots + t_N}{N-1}$	✓	✗	✗
$\hat{\tau} = t_i$	✗	✓	✗

**Table:** Properties of different estimators of the half life for a nuclear decay.

- Bias:  $b = E[\hat{\tau}] - \tau$ 
  - Note: if you don't know the true value, you must simulate the bias of the method
  - Generate toys with known parameters, and check what is the estimate of the parameter for the toy data
  - If there is a bias, correct for it to obtain an unbiased estimator
- $t_i$  is an individual observation, which is still sampled from the original factorized p.d.f.  
$$f(t_i; \tau) = \frac{1}{\tau} e^{-\frac{t_i}{\tau}}$$
- The expected value of  $t_i$  is therefore still  $E[\hat{\tau}] = E[t_i] = \tau$
- $\hat{\tau} = t_i$  is therefore unbiased!

	Consistent	Unbiased	Efficient
$\hat{\tau} = t_i$	✗	✓	✗

**Table:** Properties of different estimators of the half life for a nuclear decay.

- We usually want to optimize both bias  $\vec{b}$  and variance  $V[\hat{\theta}]$
- While we can optimize each one separately, optimizing them simultaneously leads to none being optimally optimized, in general
  - Optimal solutions in two dimensions are often suboptimal with respect to the optimization of just one of the two properties
- The variance is linked to the width of the likelihood function, which naturally leads to linking it to the curvature of  $L(\vec{x}; \vec{\theta})$  near the maximum
- However, the curvature of  $L(\vec{x}; \vec{\theta})$  near the maximum is linked to the Fisher information, as we have seen
- Information is therefore a limiting factor for the variance (no data set contains infinite information, variance cannot collapse to zero)
- Variance of an estimator satisfies the Rao-Cramér-Frechet (RCF) bound

$$V[\hat{\theta}] \geq \frac{1}{\hat{\theta}}$$

- Rao-Cramer-Frechet (RCF) bound

$$V[\hat{\theta}] \geq \frac{(1 + \partial b / \partial \theta)^2}{-E[\partial^2 \ln L / \partial \theta^2]}$$

- In multiple dimensions, link with the information is maintained via the full Fisher Information Matrix:

$$I_{ij} = E[\partial^2 \ln L / \partial \theta_i \partial \theta_j]$$

- Approximations

- Neglect the bias ( $b = 0$ )
- Inequality is an approximate equality (true for large data samples)

- $V[\hat{\theta}] \simeq \frac{1}{-E[\partial^2 \ln L / \partial \theta^2]}$

- Estimate of the variance of the estimate of the parameter!

- $\hat{V}[\hat{\theta}] \simeq \frac{1}{-E[\partial^2 \ln L / \partial \theta^2] |_{\theta = \hat{\theta}}}$

- For a generic unbiased estimator, can define *efficiency* of the estimator as

$$e(\hat{\theta}) := \frac{I(\theta)^{-1}}{V[\hat{\theta}]}$$

- The efficiency of a generic unbiased estimator, because of the RCF bound, is always  $e(\hat{\theta}) \leq 1$

- For multidimensional parameters, we can build the information matrix with elements:

$$\begin{aligned} I_{jk}(\vec{\theta}) &= -E \left[ \sum_i^N \frac{\partial^2 \ln f(x_i; \vec{\theta})}{\partial \theta_k \partial \theta_k} \right] \\ &= N \int \frac{1}{f} \frac{\partial f}{\partial \theta_j} \frac{\partial f}{\partial \theta_k} dx \end{aligned}$$

- (the last equality is due to the integration interval not being dependent on  $\vec{\theta}$ )

- We have calculated the variance of the MLE in the simple case of the nuclear decay
- Analytic calculation of the variance is not always possible
- Write the variance approximately as:

$$V[\hat{\theta}] \geq \frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{-E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]}$$

- This expression is valid for any estimator, but if applied to the MLE then we can note  $\vec{\theta}_{ML}$  is efficient and asymptotically unbiased
- Therefore, when  $N \rightarrow \infty$  then  $b = 0$  and the variance approximate to the RCF bound, and  $\geq$  becomes  $\simeq$ :

$$V[\vec{\theta}_{ML}] \simeq \frac{1}{-E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] \Big|_{\theta=\vec{\theta}_{ML}}}$$

- For a Gaussian p.d.f.,  $f(x; \vec{\theta}) = N(\mu, \sigma)$ , the likelihood can be written as:

$$L(\vec{x}; \vec{\theta}) = \ln \left[ - \frac{(\vec{x} - \vec{\theta})^2}{2\sigma^2} \right]$$

- Moving away from the maximum of  $L(\vec{x}; \vec{\theta})$  by one unit of  $\sigma$ , the likelihood assumes the value  $\frac{1}{2}$ , and the area enclosed in  $[\vec{\theta} - \sigma, \vec{\theta} + \sigma]$  will be—because of the properties of the Normal distribution—equal to 68.3%.

## How to extract an interval from the likelihood function 2/

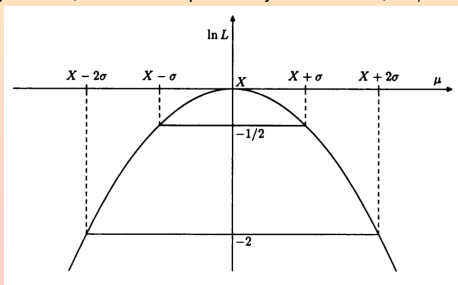
- We can therefore write

$$P\left((\vec{x} - \vec{\theta})^2 \leq \sigma\right) = 68.3\%$$

$$P(-\sigma \leq \vec{x} - \vec{\theta} \leq \sigma) = 68.3\%$$

$$P(\vec{x} - \sigma \leq \vec{\theta} \leq \vec{x} + \sigma) = 68.3\%$$

- Taking into account that it is important to keep in mind that probability is a property of sets, in frequentist statistics
  - Confidence interval: interval with a fixed probability content
- This process for computing a confidence interval is exact for a Gaussian p.d.f.
  - Pathological cases reviewed later on (confidence belts and Neyman construction)
- Practical prescription:
  - Point estimate by computing the Maximum Likelihood Estimate
  - Confidence interval by taking the range delimited by the crossings of the likelihood function with  $\frac{1}{2}$  (for 68.3% probability content, or 2 for 95% probability content— $2\sigma$ , etc)

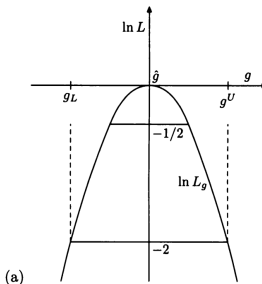
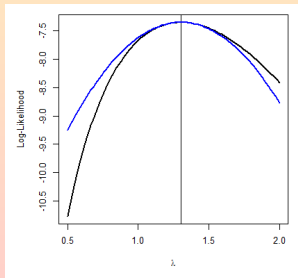


Plot from James, 2nd ed.

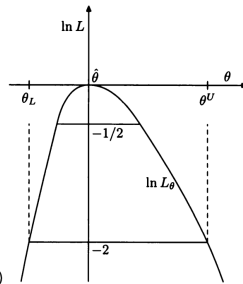


## How to extract an interval from the likelihood function 3/

- MLE is invariant for monotonic transformations of  $\theta$ 
  - This applies not only to the maximum of the likelihood, but to all relative values
  - The likelihood ratio is therefore an invariant quantity (we'll use it for hypothesis testing)
  - Can transform the likelihood such that  $\log(L(\vec{x}; \vec{\theta}))$  is parabolic, but not necessary (MINOS/Minuit)
- When the p.d.f. is not normal, either assume it is, and use symmetric intervals from Gaussian tails...
  - This yields symmetric approximate intervals
  - The approximation is often good even for small amounts of data
- ...or use asymmetric intervals by just looking at the crossing of the  $\log(L(\vec{x}; \vec{\theta}))$  values
  - Naturally-arising asymmetrical intervals
  - No gaussian approximation
- In any case (even asymmetric intervals) still based on asymptotic expansion
  - Method is exact only to  $\mathcal{O}(\frac{1}{N})$



(a)



(b)

Plot from James, 2nd ed.

- Theorem: for any p.d.f.  $f(x|\vec{\theta})$ , in the large numbers limit  $N \rightarrow \infty$ , the likelihood can always be approximated with a gaussian:

$$L(\vec{x}; \vec{\theta}) \propto_{N \rightarrow \infty} e^{-\frac{1}{2}(\vec{\theta} - \vec{\theta}_{ML})^T H(\vec{\theta} - \vec{\theta}_{ML})}$$

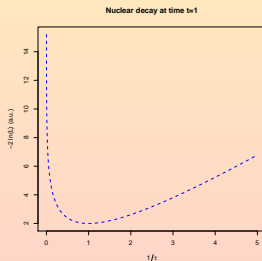
- where  $H$  is the information matrix  $I(\vec{\theta})$ .
- Under these conditions,  $V[\vec{\theta}_{ML}] \rightarrow \frac{1}{I(\vec{\theta}_{ML})}$ , and the intervals can be computed as:

$$\Delta \ln L := \ln L(\theta') - \ln L_{max} = -\frac{1}{2}$$

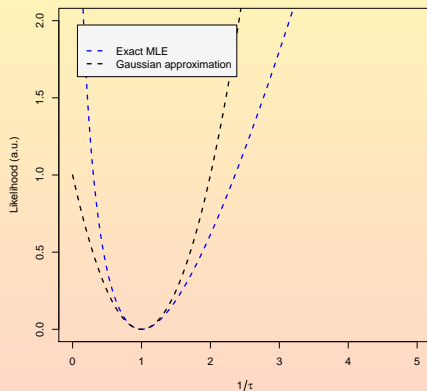
- The resulting interval has in general a larger probability content than the one for a gaussian p.d.f., but the approximation grows better when  $N$  increases
  - The interval overcovers the true value  $\vec{\theta}_{true}$

- $\vec{\theta}_{true}$  is therefore estimated as  $\hat{\theta} = \vec{\theta}_{ML} \pm \sigma$ . This is another situation in which frequentist and Bayesian statistics differ in the interpretation of the numerical result
- Frequentist:  $\vec{\theta}_{true}$  is fixed
  - “if I repeat the experiment many times, computing each time a confidence interval around  $\vec{\theta}_{ML}$ , on average 68.3% of those intervals will contain  $\vec{\theta}_{true}$ ”
  - Coverage: “the interval covers the true value with 68.3% probability”
  - Direct consequence of the probability being a property of data sets
- Bayesian:  $\vec{\theta}_{true}$  is not fixed
  - “the true value  $\vec{\theta}_{true}$  will be in the range  $[\vec{\theta}_{ML} - \sigma, \vec{\theta}_{ML} + \sigma]$  with a probability of 68.3%”
  - This corresponds to giving a value for the posterior probability of the parameter  $\vec{\theta}_{true}$

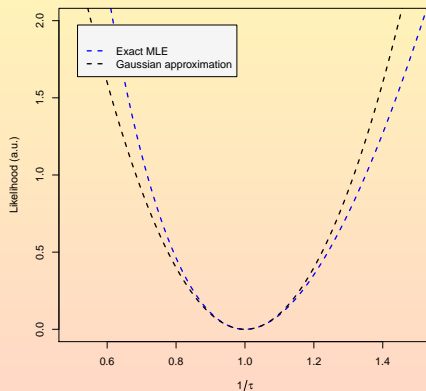
- How good is the approximation  $L(\vec{x}; \vec{\theta}) \propto \exp\left[-\frac{1}{2}(\vec{\theta} - \vec{\theta}_{MLE})^T H(\vec{\theta} - \vec{\theta}_{ML})\right]$ ?
  - Here  $H$  is the information matrix  $I(\vec{\theta})$
  - True only to  $\mathcal{O}(\frac{1}{N})$
  - In these conditions,  $V[\vec{\theta}_{ML}] \rightarrow \frac{1}{I(\vec{\theta}_{ML})}$
  - Intervals can be derived by crossings:  $\Delta \ln L = \ln L(\theta') - \ln L_{max} = k$
- This afternoon: we'll convince ourselves of how good is this approximation in case of the nuclear decay!



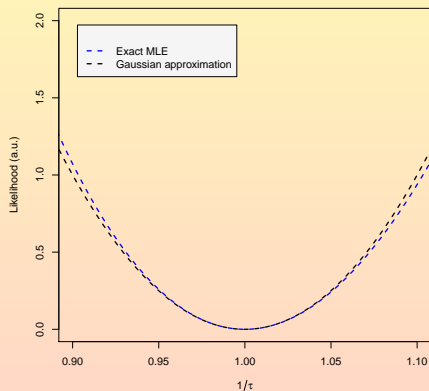
Nuclear decay at time  $t=1$  and  $N=1$



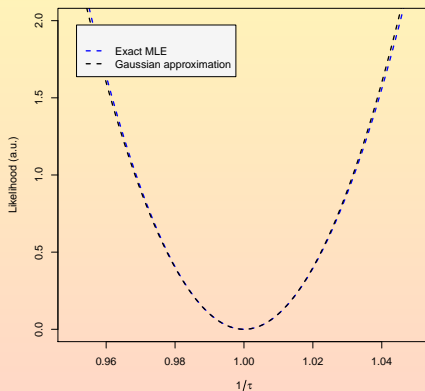
Nuclear decay at time  $t=1$  and  $N=10$



Nuclear decay at time  $t=1$  and  $N=100$



Nuclear decay at time  $t=1$  and  $N=1000$



- The convergence of the likelihood  $L(\vec{x}; \vec{\theta})$  to a gaussian is a direct consequence of the central limit theorem
- Take a set of measurements  $\vec{x} = (x_1, \dots, x_N)$  affected by experimental errors that results in uncertainties  $\sigma_1, \dots, \sigma_N$  (not necessarily equal among each other)
- In the limit of a large number of events,  $M \rightarrow \infty$ , the random variable built summing  $M$  measurements is gaussian-distributed:

$$Q := \sum_{j=1}^M x_j \sim N\left(\sum_{j=1}^M x_j, \sum_{j=1}^M \sigma_j^2\right), \quad \forall f(x, \vec{\theta})$$

- The demonstration runs by expanding in series the characteristic function  $y_i = \frac{x_j - \mu_j}{\sqrt{\sigma_j}}$
- The theorem is valid for any p.d.f.  $f(x, \vec{\theta})$  that is reasonably peaked around its expected value.
  - If the p.d.f. has large tails, the bigger contributions from values sampled from the tails will have a large weight in the sum, and the distribution of  $Q$  will have non-gaussian tails
  - The consequence is an alteration of the probability of having sums  $Q$  outside of the gaussian

- The condition  $M \rightarrow \infty$  is reasonably valid if the sum is of many small contributions.
- How large does  $M$  need to be for the approximation to be reasonably good? Question time:  
Central Limit



- The condition  $M \rightarrow \infty$  is reasonably valid if the sum is of many small contributions.
- How large does  $M$  need to be for the approximation to be reasonably good? Question time: Central Limit
- This afternoon we'll check!

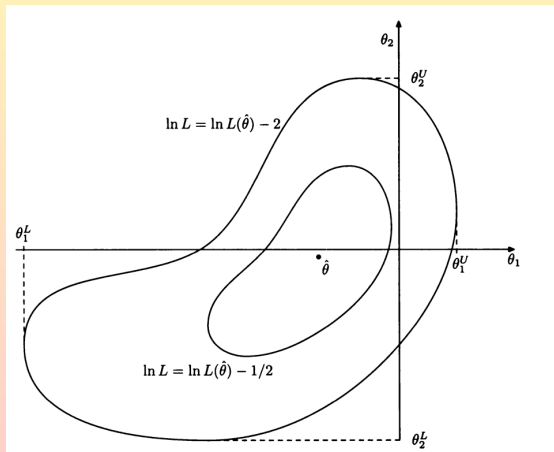
- Construct  $\log \mathcal{L}$  contours and determine confidence intervals by MINOS
- Elliptical contours correspond to gaussian Likelihoods
  - The closer to MLE, the more elliptical the contours, even in non-linear problems
  - All models are linear in a sufficiently small region
- Nonlinear regions not problematic (no parabolic transformation of  $\log \mathcal{L}$  needed)
  - MINOS accounts for non-linearities by following the likelihood contour

- Confidence intervals for each parameter

$$\max_{\theta_j, j \neq i} \log \mathcal{L}(\theta) = \log \mathcal{L}(\hat{\theta}) - \lambda$$

- $\lambda = \frac{Z_{1-\beta}^2}{2}$

- $\lambda = 1/2$  for  $\beta = 0.683$  ("1 $\sigma$ ")
- $\lambda = 2$  for  $\beta = 0.955$  ("2 $\sigma$ ")



Plot from James, 2nd ed.

## Profile likelihood ratio step by step for cross sections — Expected event



- We used to compute the total cross section of a given process by applying the naïve formula

$$\sigma = \frac{N_{data} - N_{bkg}}{\epsilon L} .$$

- $N_{sig}$  estimated from  $N_{data} - N_{bkg}$  for the measured integrated luminosity  $L$
- The acceptance  $\epsilon$  accounts for th. branching fractions fiducial region for the measurement (fiducial region: generator-level selection which defines the phase space of the measurement)
- Nowadays we model everything into the likelihood function
- $p(x|\mu, \theta)$  pdf for the observable  $x$  to assume a certain value in a single event
  - $\mu := \frac{\sigma}{\sigma_{pred}}$  (single- or multi-dimensional) *parameter of interest* (POI). A multiplier of the predicted cross section: *signal strength*
  - $\theta$  (generally multi-dimensional) *nuisance parameter* representing all the uncertainties affecting the measurement.
- Extend to a data set of many events  $X = \{x_1, \dots, x_n\}$  by taking the product of the single-event p.d.f.s.

$$\prod_{e=1}^n p(x_e|\mu, \theta)$$

## Profile likelihood ratio step by step for cross sections — Expected event

- We used to compute the total cross section of a given process by applying the naïve formula

$$\sigma = \frac{N_{data} - N_{bkg}}{\epsilon L} .$$

- $N_{sig}$  estimated from  $N_{data} - N_{bkg}$  for the measured integrated luminosity  $L$
- The acceptance  $\epsilon$  accounts for th. branching fractions fiducial region for the measurement (fiducial region: generator-level selection which defines the phase space of the measurement)
- Nowadays we model everything into the likelihood function
- $p(x|\mu, \theta)$  pdf for the observable  $x$  to assume a certain value in a single event
  - $\mu := \frac{\sigma}{\sigma_{pred}}$  (single- or multi-dimensional) *parameter of interest* (POI). A multiplier of the predicted cross section: *signal strength*
  - $\theta$  (generally multi-dimensional) *nuisance parameter* representing all the uncertainties affecting the measurement.
- Extend to a data set of many events  $X = \{x_1, \dots, x_n\}$  by taking the product of the single-event p.d.f.s.

$$\prod_{e=1}^n p(x_e|\mu, \theta)$$

- The number of events in the data set is however a random variable itself!
  - Poisson distribution with mean equal to the number of events  $\nu$  we expect from theory
- *Marked Poisson model*

$$f(X|\nu(\mu, \theta), \mu, \theta) = \text{Pois}(n|\nu(\mu, \theta)) \prod_{e=1}^n p(x_e|\mu, \theta) .$$

Pleasant quality read: Vischia, 2019 doi:10.1016/j.revip.2020.100046 ☺

- Both  $\mu$  and  $\theta$  act on the individual pdfs for the observable and on the expectation for the global amount of events
- Incorporate systematic uncertainties as nuisance parameter  $\theta$ :  
Conway, 2011 in CERN-2011-006115
  - Constrain the terms in the fit: constraint interpreted as prior coming from the auxiliary measurement
  - $\theta$  estimated with uncertainty  $\delta\theta$
  - Often Gaussian pdf, unless  $\theta$  has a physical bound at zero: then log-normal (rejects negative values)
- Likelihood  $\mathcal{L}(\mu, \theta; X)$ : take the marked Poisson model  $f(X|\nu(\mu, \theta), \mu, \theta)$  and condition on the observed value of  $X$
- MLE:  $\hat{\mu} := \operatorname{argmax}_{\mu} \mathcal{L}(\mu, \theta; X)$  still depends on the nuisance parameters  $\theta$

$$\mathcal{L}(\mathbf{n}, \boldsymbol{\alpha}^0 | \mu, \boldsymbol{\alpha}) = \prod_{i \in \text{bins}} \mathcal{P}(n_i | \mu S_i(\boldsymbol{\alpha}) + B_i(\boldsymbol{\alpha})) \times \prod_{j \in \text{syst}} \mathcal{G}(\alpha_j^0 | \alpha_j, \delta\alpha_j)$$

$$\downarrow$$

$$\mathcal{L}(\mathbf{n}, 0 | \mu, \boldsymbol{\alpha}) = \prod_{i \in \text{bins}} \mathcal{P}(n_i | \mu S_i(\boldsymbol{\alpha}) + B_i(\boldsymbol{\alpha})) \times \prod_{j \in \text{syst}} \mathcal{G}(0 | \alpha_j, 1)$$

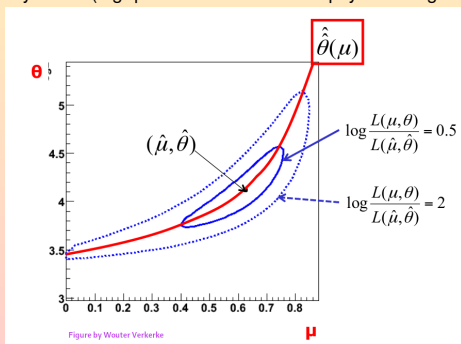
Pleasant quality read: Vischia, 2019 [doi:10.1016/j.revip.2020.100046](https://doi.org/10.1016/j.revip.2020.100046) ☺

## Eliminate dependence on the nuisance parameters

- Likelihood ratio!

$$\lambda(\mu) := \frac{\mathcal{L}(\mu, \hat{\theta})}{\mathcal{L}(\hat{\mu}, \hat{\theta})}.$$

- Denominator  $\mathcal{L}(\hat{\mu}, \hat{\theta})$  is computed for the values of  $\mu$  and  $\theta$  which jointly maximize the likelihood function.
  - *Profiling*: eliminating the dependence on the nuisance parameters by taking their conditional maximum likelihood estimate
  - Bayesians normally marginalize (integrate) rather than profiling (see [Demortier, 2002](#))
- The maximum of the likelihood ratio yields the point estimate for  $\mu$
- The second derivative of the maximum likelihood ratio yields intervals on the parameter  $\mu$ 
  - Tomorrow: the tricky cases (e.g. point estimate near the physical range allowed for the parameter)

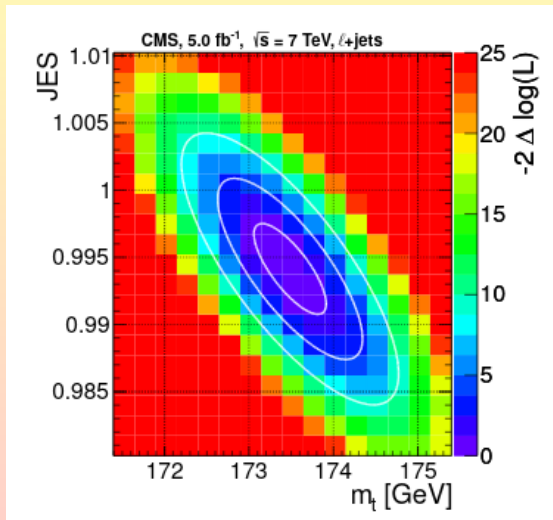


Pleasant quality read: Vischia, 2019 [doi:10.1016/j.revip.2020.100046](https://doi.org/10.1016/j.revip.2020.100046) ☺

- The likelihood ratio  $\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})}$
- Conceptually, you can run the experiment many times (e.g. toys) and record the value of the test statistic
- The test statistic can therefore be seen as a distribution
- Asymptotically,  $\lambda(\mu) \sim \exp\left[-\frac{1}{2}\chi^2\right] \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right)$  (Wilks Theorem, under some regularity conditions—continuity of the likelihood and up to 2nd derivatives, existence of a maximum, etc)
  - The  $\chi^2$  distribution depends only on a single parameter, the number of degrees of freedom
  - It follows that the test statistic is independent of the values of the nuisance parameters
  - Useful: you don't need to make toys in order to find out how is  $\lambda(\mu)$  distributed!

## What is a nuisance parameter?

- Sometimes the classification into POI and nuisance parameter washes out
- Maybe your data and your method can provide information on a systematic uncertainty



Plot from [doi:10.1007/JHEP12\(2012\)105](https://doi.org/10.1007/JHEP12(2012)105)



- More often, the analysis is not sensitive enough to treat an uncertainty as POI and measure it
- The fit can still constrain the nuisance parameter that is profiled
- Indirectly provides information about your estimate of that parameter before the fit
  - Over- or under-estimate  $\theta$  before the fit
  - See a best fit value for  $\theta$  that doesn't match very well with the prefit value
- Quote, for each nuisance parameter, two important quantities
  - **Pull**: the difference of the post-fit and pre-fit values of the parameter, normalized to the pre-fit uncertainty:  $pull := \frac{\hat{\theta} - \theta}{\delta\theta}$
  - **Constraint**: the ratio between the post-fit and the pre-fit uncertainty in the nuisance parameter.

- **Pull:** the difference of the post-fit and pre-fit values of the parameter, normalized to the pre-fit uncertainty:  $pull := \frac{\hat{\theta} - \theta}{\delta\theta}$
- **Constraint:** the ratio between the post-fit and the pre-fit uncertainty in the nuisance parameter.
- Spot easily possible issues in the fit
  - $\theta$  pulled too much may be a hint that our estimate of the pre-fit value was not reasonable
  - $\theta$  constrained too much indicates that the data contain enough information to improve the precision in the nuisance parameter with respect to our original estimate, which may or may not make sense.

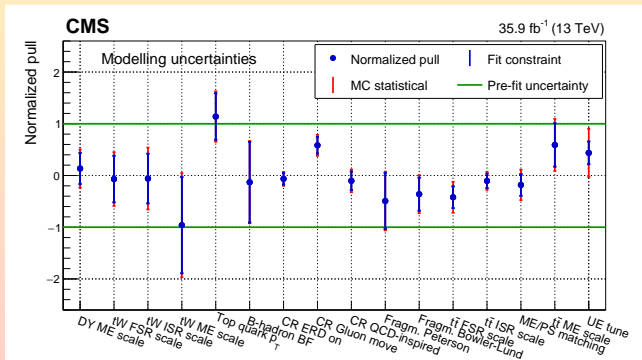


Image collected and cited in [doi:10.1016/j.revip.2020.100046](https://doi.org/10.1016/j.revip.2020.100046), references therein)

- What is more worrying, a small pull with a small constraint, or a large pull with a strong constraint? Question time: Pulls and Constraints
- A pull with very small constraint:  $\theta_{prefit} = 0 \pm 1$ ,  $\theta_{postfit} = 1 \pm 0.9$
- The same pull with a strong constraint:  $\theta_{prefit} = 0 \pm 1$ ,  $\theta_{postfit} = 1 \pm 0.2$

- What is more worrying, a small pull with a small constraint, or a large pull with a strong constraint? **Question time: Pulls and Constraints**
- A pull with very small constraint:  $\theta_{prefit} = 0 \pm 1$ ,  $\theta_{postfit} = 1 \pm 0.9$
- The same pull with a strong constraint:  $\theta_{prefit} = 0 \pm 1$ ,  $\theta_{postfit} = 1 \pm 0.2$
- A way of estimating if a shift is significant is to compare the shift with its uncertainty
- For independent measurements, the compatibility  $C$  is

$$C = \Delta\theta / \sigma_{\Delta\theta} = \frac{\theta_2 - \theta_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

- We would conclude that the first case  $C = 0.74$ , for the second one  $C = 0.98$  (larger, still within uncertainty)

- What is more worrying, a small pull with a small constraint, or a large pull with a strong constraint? Question time: Pulls and Constraints
- A pull with very small constraint:  $\theta_{prefit} = 0 \pm 1$ ,  $\theta_{postfit} = 1 \pm 0.9$
- The same pull with a strong constraint:  $\theta_{prefit} = 0 \pm 1$ ,  $\theta_{postfit} = 1 \pm 0.2$
- A way of estimating if a shift is significant is to compare the shift with its uncertainty
- For independent measurements, the compatibility  $C$  is

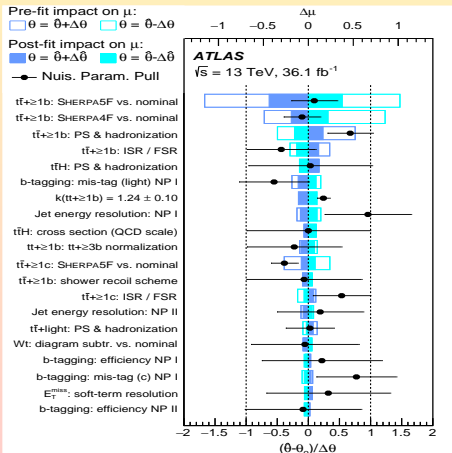
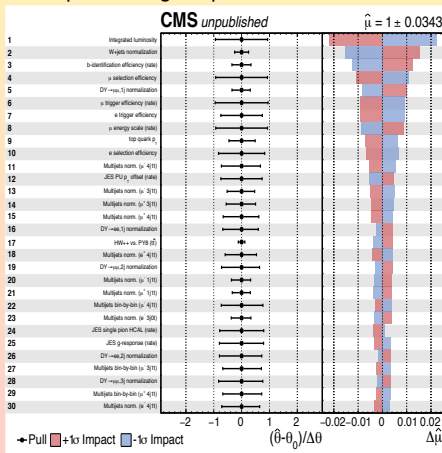
$$C = \Delta\theta / \sigma_{\Delta\theta} = \frac{\theta_2 - \theta_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

- We would conclude that the first case  $C = 0.74$ , for the second one  $C = 0.98$  (larger, still within uncertainty)
- However, these are not independent measurements!
- The formula is therefore

$$C = \Delta\theta / \sigma_{\Delta\theta} = \frac{\theta_2 - \theta_1}{\sqrt{\sigma_1^2 - \sigma_2^2}}$$

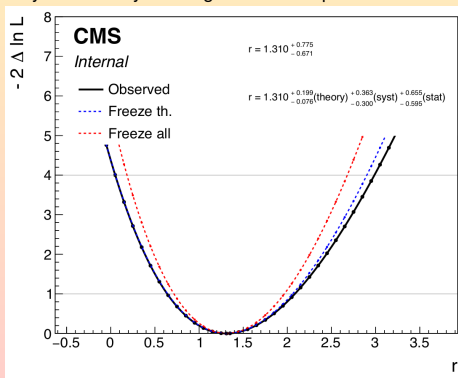
- For the first case,  $C = 2.29$ , for the second case  $C = 1.02$
- The same pull is more significant if there is (almost no) constraint!!!

- Impact of  $\theta$  on the post-fit signal strength permits to obtain a ranking of the nuisance parameters in terms of their effect on the signal strength
  - Fix each nuisance parameter to its post-fit value  $\hat{\theta}$  plus/minus its pre-fit (post-fit) uncertainty  $\delta\theta$  ( $\delta\hat{\theta}$ )
  - Reperform the fit for  $\mu$
  - Compute the impact as the difference between the original fitted signal strength and the refitted signal strength.
- Results on Asimov dataset (replacing the data with the expectations from simulated events) is expected to give “perfect” results



## Breakdown of systematic uncertainties

- What's the amount of uncertainty that is imputable to a given set of systematic effects?
  - The modern expression of Fisher's formalization of the ANOVA concept
  - *"the constituent causes fractions or percentages of the total variance which they together produce"* (Fisher, 1919)
  - *"the variance contributed by each term, and by which the residual variance is reduced when that term is removed"* (Fisher, 1921)
- Breakdown the contributions
  - Freeze a set of uncertainties  $\theta_i$  to their post-fit value
  - Repeat the fit to extract a new (smaller) uncertainty on  $\mu$
  - Obtain the contribution of  $\theta_i$  to the overall uncertainty as squared difference between the full and reduced uncertainties
  - Statistical uncertainty obtained by freezing all nuisance parameters



Toy data

Statistics for HEP

- Measure a background rate in a sideband, use the estimate in the signal region
- As described, let's model our estimation problem using profile likelihoods

$$\mathcal{L}(\mathbf{n}, \boldsymbol{\alpha}^0 | \boldsymbol{\mu}, \boldsymbol{\alpha}) = \prod_{i \in \text{bins}} \mathcal{P}(n_i | \mu S_i(\boldsymbol{\alpha}) + B_i(\boldsymbol{\alpha})) \times \prod_{j \in \text{syst}} \mathcal{G}(\alpha_j^0 | \alpha_j, \delta \alpha_j)$$

$$\lambda(\mu) = \frac{\mathcal{L}(\mu, \hat{\boldsymbol{\alpha}}_\mu)}{\mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\alpha}})}$$

- Sideband measurement

$$L_{SR}(s, b) = \text{Poisson}(N_{SR} | s + b)$$

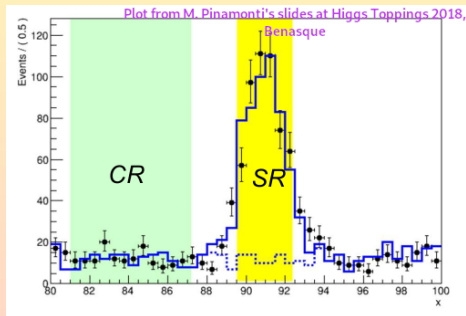
$$L_{CR}(b) = \text{Poisson}(N_{CR} | \tilde{\tau} \cdot b)$$

$$\mathcal{L}_{full}(s, b) = \mathcal{P}(N_{SR} | s + b) \times \mathcal{P}(N_{CR} | \tilde{\tau} \cdot b)$$

- Subsidiary measurement of the background rate:

- 8% systematic uncertainty on the MC rates
- $\tilde{b}$ : measured background rate by MC simulation
- $\mathcal{G}(\tilde{b} | b, 0.08)$ : our

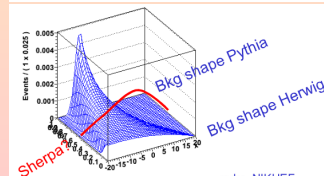
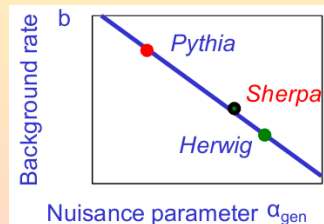
$$\mathcal{L}_{full}(s, b) = \mathcal{P}(N_{SR} | s + b) \times \mathcal{G}(\tilde{b} | b, 0.08)$$



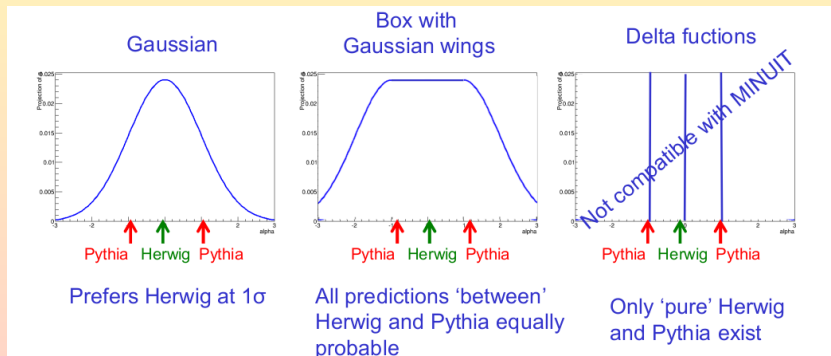


## Caveats on modelling theory uncertainties (P.V. at Benasque 2018)

- Cross section uncertainty: easy, assuming a gaussian for the constraint term<sup>1</sup>  
 $\mathcal{L}_{full}(s, b) = \mathcal{P}(N_{SR}|s + b) \times \mathcal{G}(\hat{b}|b, 0.08)$
- Factorization scale: what distribution  $\mathcal{F}$  is meant to model the constraint???
- Hadronization/fragmentation model: run different generators, observing different results
  - “Easy” case, there is a single parameter  $\alpha_{FS}$ , clearly connected to the underlying physics model
- Counting experiment: easy extend to other generators
- There must exist a value of  $\alpha$  corresponding to SHERPA
- Shape experiment: ouch!
- SHERPA is in general not obtainable as an interpolation of PYTHIA and HERWIG

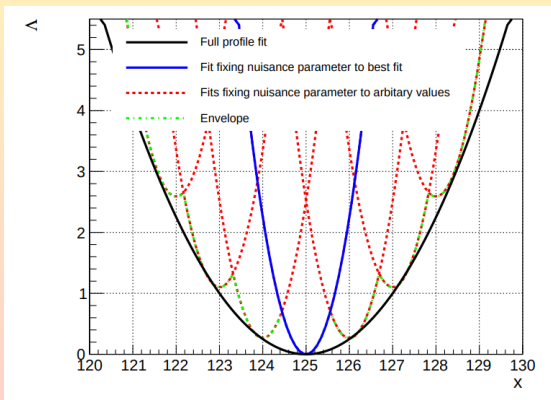


- Attempting to quantify our knowledge of the models
- There is no single parameter, difficult to model the differences within a single underlying model
- Which of these is the “correct” one?

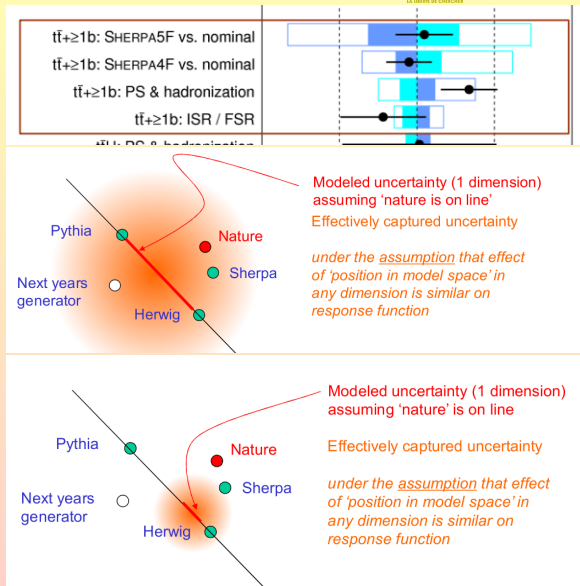


Graphics from W. Verkerke

- Label each shape with an integer, and use the integer as nuisance parameter
- Can obtain the original log-likelihood as an envelope of different fixed discrete nuisance parameter values
- How do you define the various shapes?
  - Need many additional generators!
  - Interpolation unlikely to work (*SHERPA is not midway between PYTHIA and POWHEG*)



From [arXiv:1408.6865](https://arxiv.org/abs/1408.6865)



Graphics from ATLAS and W. Verkerke, as far as I remember

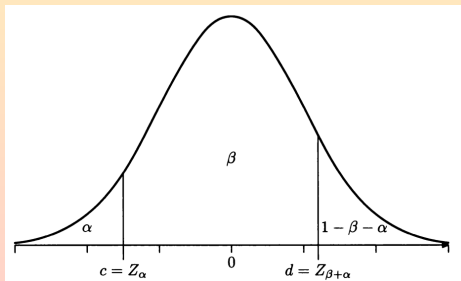
- How to interpret constraints?
- **Not as measurements**
- Correlations in the fit make interpretation complicated
- Avoid statements when profiling as a nuisance parameter

- Closure tests are alternative procedures you can use to check if your measurement is robust
  - E.g. insensitive to systematic effects
  - Usually compare alternative result with nominal result (GoF test) to decide if closure test passed
- **Closure tests are PASS/FAIL tests**
- Correct course of action: if closure test fails, then there is a mistake in the tested procedure, therefore modify/improve the procedure
  - If the alternative procedure highlights e.g. a recalibration to be done, then recalibrate (i.e. use the better procedure)
- Wrong course of action: if closure test fails, add discrepancy as uncertainty
  - The sentence “*The closure test shows a 10% discrepancy, and we consequently assign it as systematic uncertainty*” is pure BS (although you’ll sadly find it in many published papers)
- In general, if a closure test fails, always prioritize a mitigation or suppression of the effect by improving your analysis methods
  - A systematic should be added only as a very very last resort

# 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

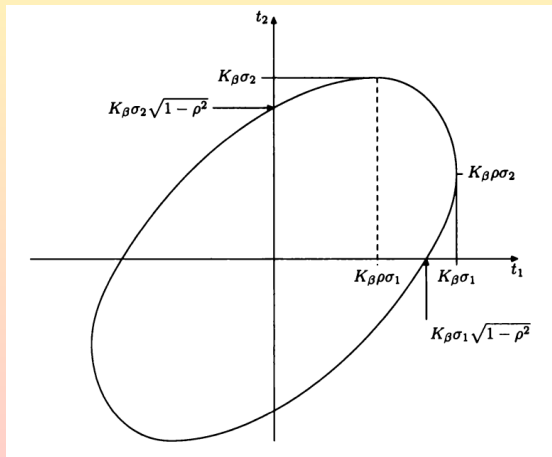
- Example: if  $f(X|\theta) = N(\mu, \sigma^2)$  with unknown  $\mu, \sigma$ , choose  $Z = \frac{X - \mu}{\sigma}$
- Find  $[c, d]$  in  $\beta = P(c \leq Z \leq d) = \Phi(d) - \Phi(c)$  by finding  $[Z_\alpha, Z_{\alpha+\beta}]$
- Infinite interval choices: here central interval  
 $\alpha = \frac{1-\beta}{2}$



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.



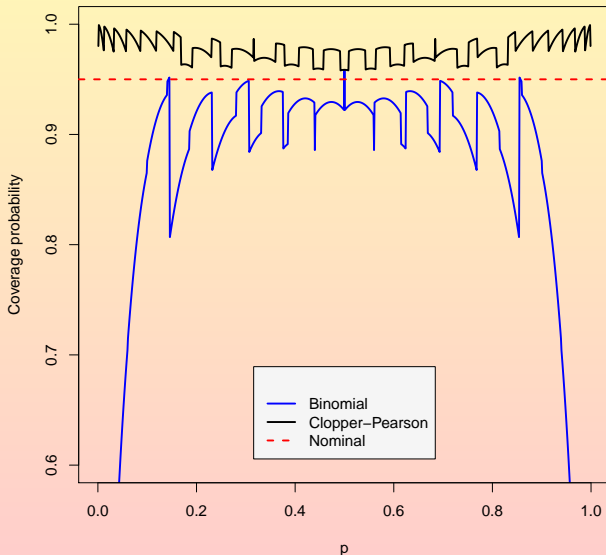
- Coverage probability of a method for calculating a confidence interval  $[\theta_1, \theta_2]$ :  
 $P(\theta_1 \leq \theta_{true} \leq \theta_2)$ 
  - Fraction of times, over a set of (usually hypothetical) measurements, that the resulting interval covers the true value of the parameter
  - Can sample 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  $\alpha$

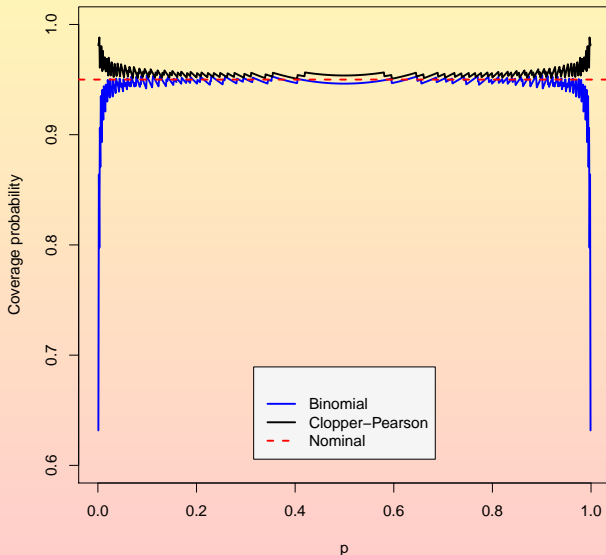
- Gaussian approximation:  $p \pm Z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{N}}$
- Clopper Pearson: invert two single-tailed binomial tests, designed to overcover
$$\sum_{r=0}^N \binom{r}{N} p^n (1 - p_{low})^{N-n} \leq \alpha/2$$
$$\sum_{r=0}^N \binom{r}{N} p^r (1 - p_{high})^{N-r} \leq \alpha/2$$
  - Single-tailed  $\rightarrow$  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

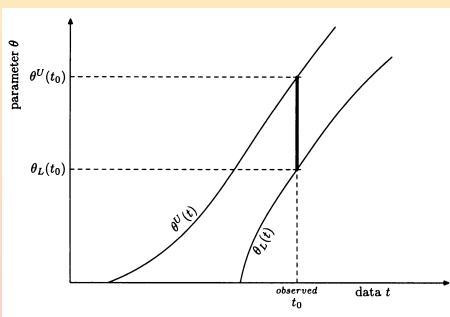
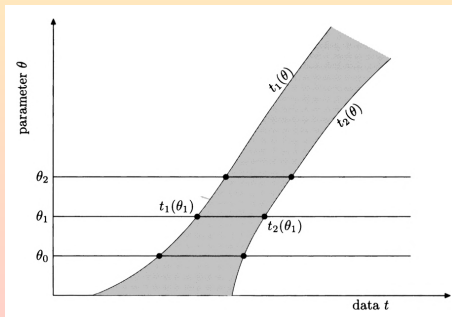


## Coverage, $N = 1000$

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



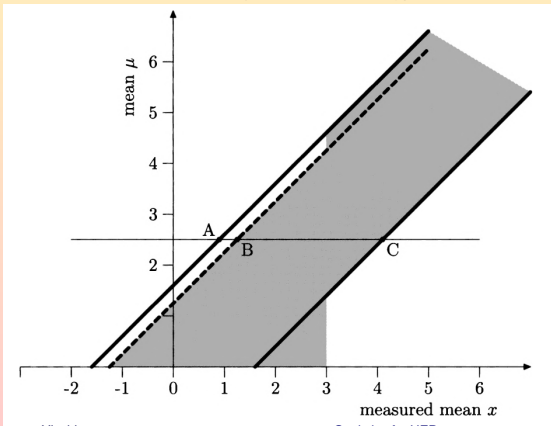
- Unique solutions to finding confidence intervals are infinite
  - Central intervals, lower limits, upper limits, etc
- Let's suppose we have chosen a way
- Build horizontally: for each (hypothetical) value of  $\theta$ , determine  $t_1(\theta)$ ,  $t_2(\theta)$  such that  $\int_{t_1}^{t_2} P(t|\theta)dt = \beta$
- 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

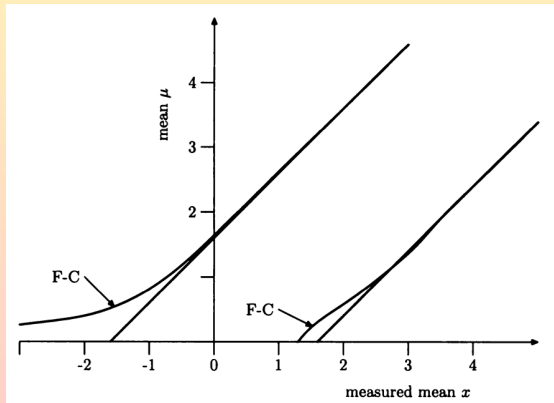
- Gaussian measurement ( variance 1) of a non-negative parameter  $\mu \sim 0$  (physical bound)
- Individual prescriptions are self-consistent
  - 90% central limit (solid lines)
  - 90% upper limit (single dashed line)
- Other choices are problematic (flip-flopping): never choose after seeing the data!
  - “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)



Plot from James, 2nd ed.

- 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.

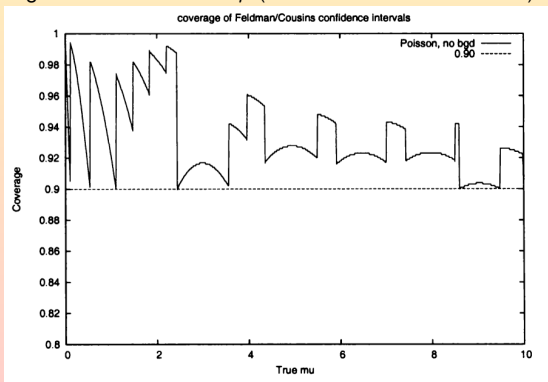


## Feldman-Cousins in HEP

- 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  $\theta$

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

- 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  $\int_{\theta_L}^{\theta^U} \pi(\theta|X) d\theta = \beta$
- 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

- 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](#)

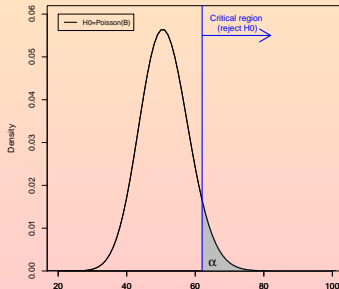
# Test of Hypotheses

- 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 $H_0$ ...

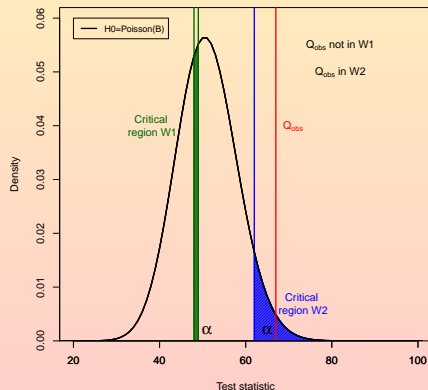
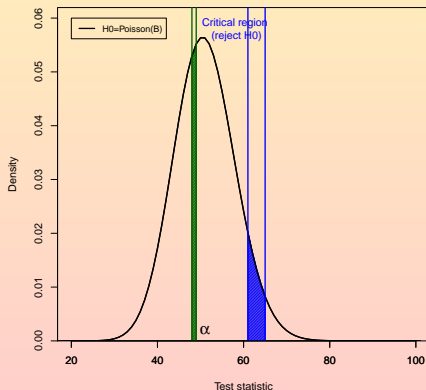
- $H_0$  is normally the hypothesis that we assume true in absence of further evidence
- Let  $X$  be a function of the observations (called “test statistic”)
- Let  $W$  be the space of all possible values of  $X$ , and divide it into
  - A critical region  $w$ : observations  $X$  falling into  $w$  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  $H_0$ !!!





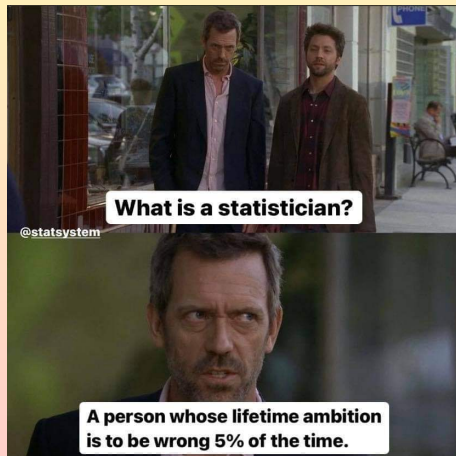
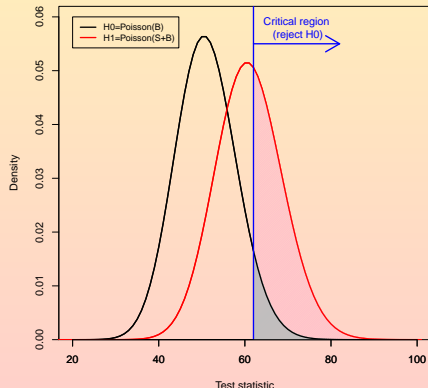
## ...while introducing some spice in it

- 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  $\vec{x}_{obs}$
  - In particular, for an infinite number of choices of  $\mathcal{W}$ , the point  $\vec{x}_{obs}$ —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,  $H_1$



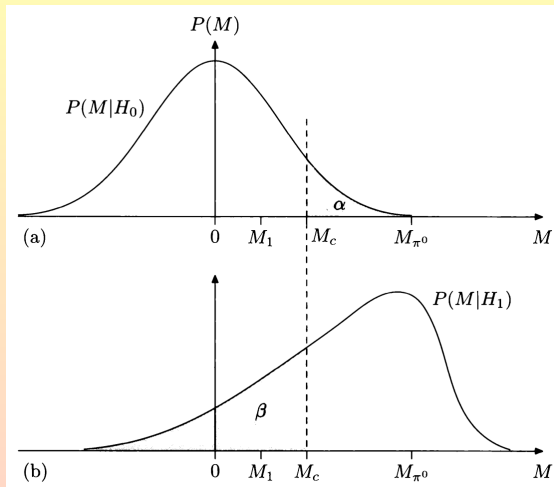
## Choose reasonable regions

- 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 $H_1$	Plot from James, 2nd ed.
$H_0$ is true	$1 - \alpha$	$\alpha$ (Type I error)	
$H_1$ 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](https://hypptest.ipynb)

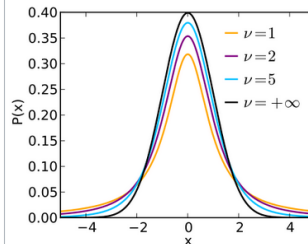
PDF

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

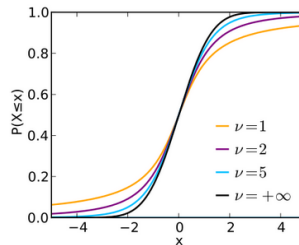


## Student's t

Probability density function

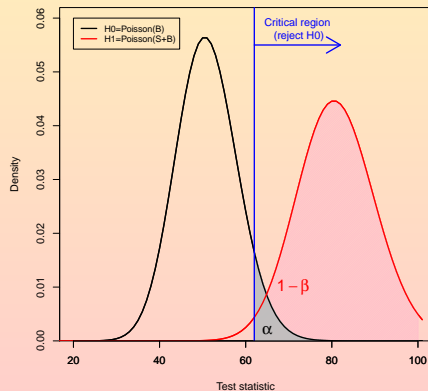
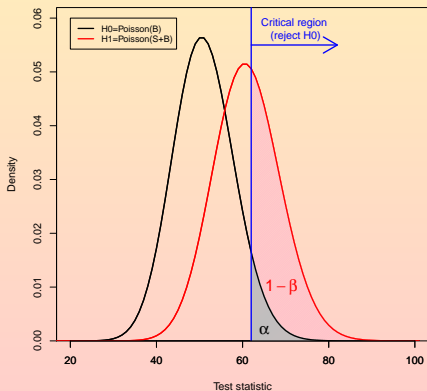


Cumulative distribution function

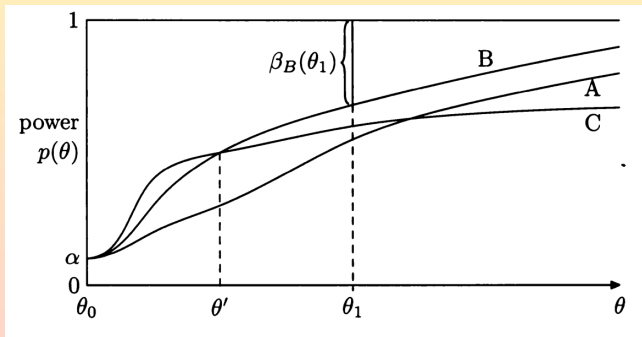


## Basic hypothesis testing – 4

- 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.



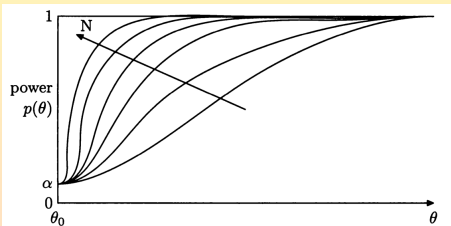
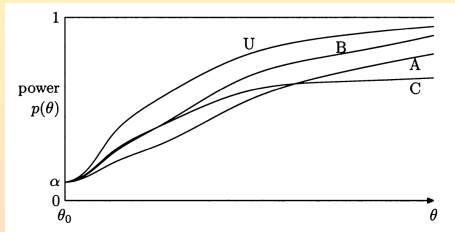
- For parametric (families of) hypotheses, the power depends on the parameter
  - $H_0 : \theta = \theta_0$
  - $H_1 : \theta = \theta_1$
  - Power:  $p(\theta_1) = 1 - \beta$
- Generalize for all possible alternative hypotheses:  $p(\theta) = 1 - \beta(\theta)$ 
  - For the null,  $p(\theta_0) = 1 -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



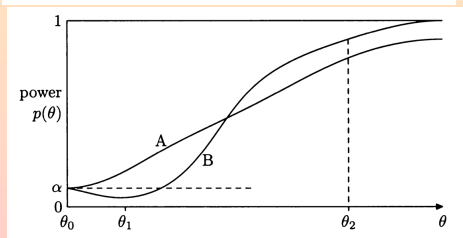
Plot from James, 2nd ed.

## 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>3</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>3</sup>Robust: a test with low sensitivity to unimportant changes of the null hypothesis

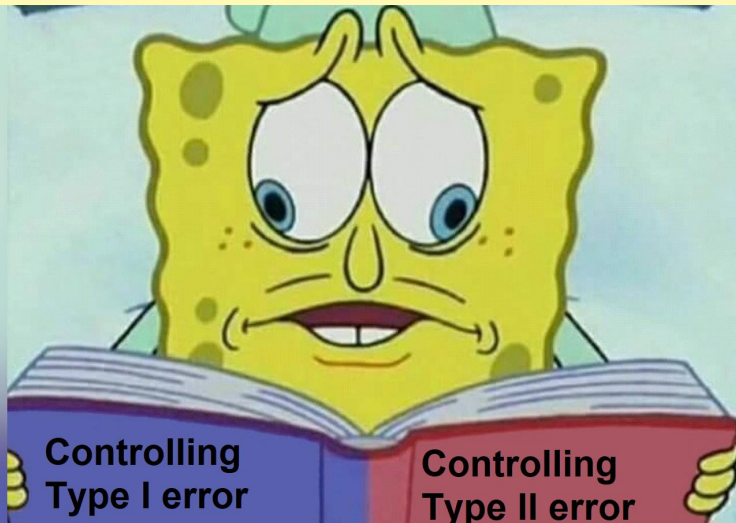


Image from the [Statistical Statistics Memes Facebook Page](#)



- 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  $\psi = \psi^*$ .

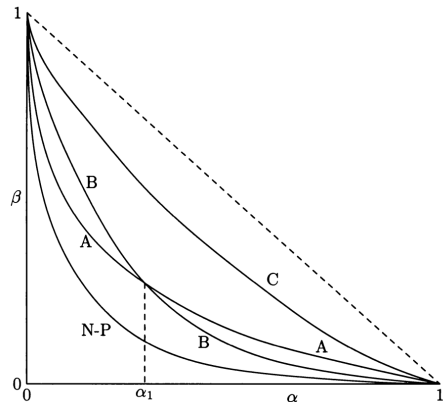


Table 10.4. A cost function.

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

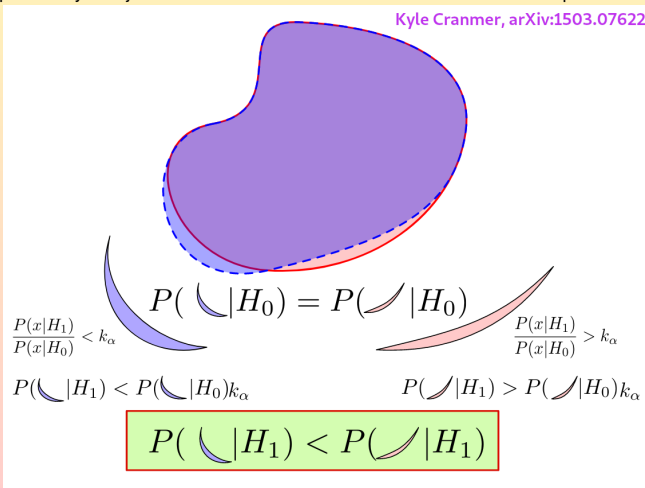


Image from Evan Vucci, Shutterstock, meme is mine

## 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

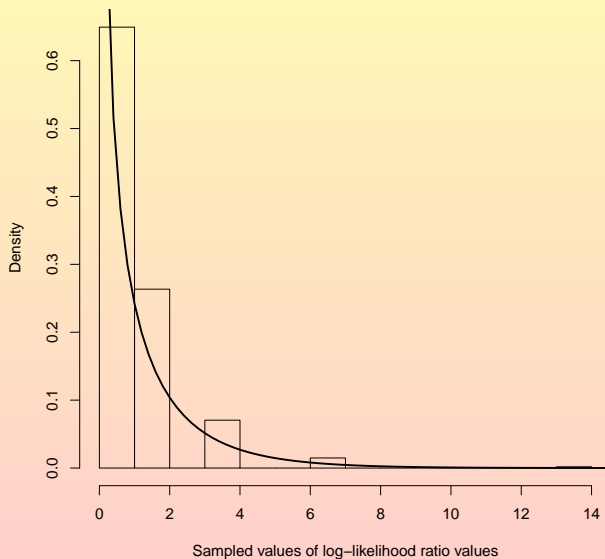


- The likelihood ratio is commonly used
- As any test statistic in the market, in order to select critical regions based on confidence levels it is necessary to know its distribution
  - Run toys to find its distribution (very expensive if you want to model extreme tails)
  - Find some asymptotic condition 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

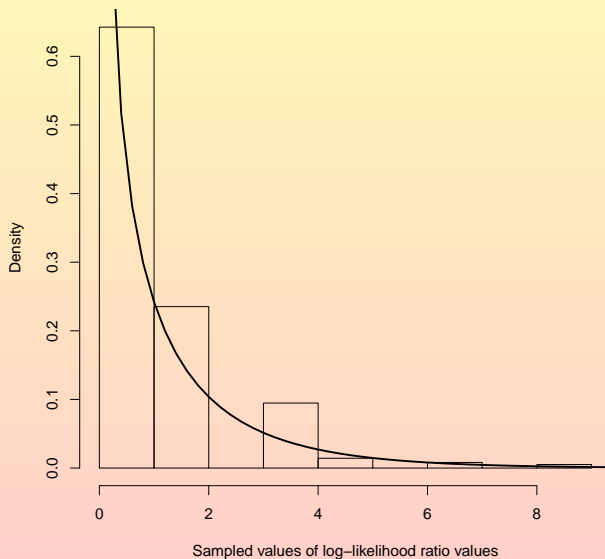
We can summarize in the

**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  $\hat{\theta}_i$  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.*

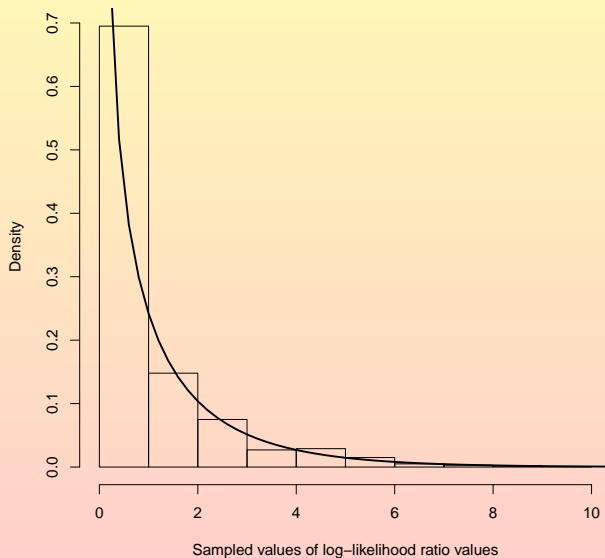
## Log-likelihood ratio



## 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
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

$ \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

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

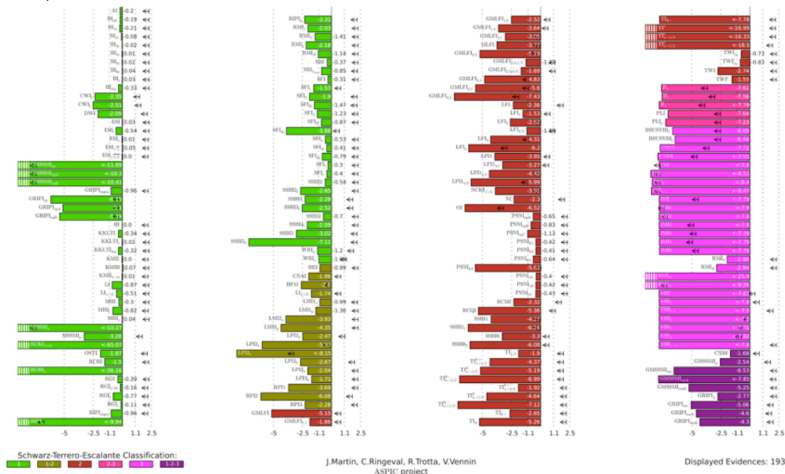
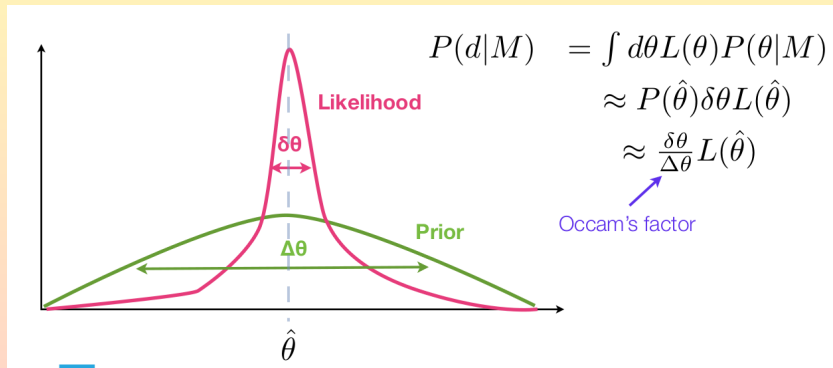


Image from Roberto Trotta, Chair Lemaître Lectures 2018

- 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 Lemaître 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  $H_1$

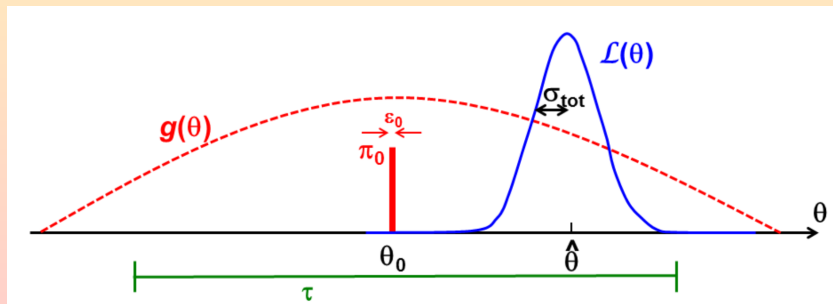


Image from Cousins, doi:10.1007/s11229-014-0525-z



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://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.

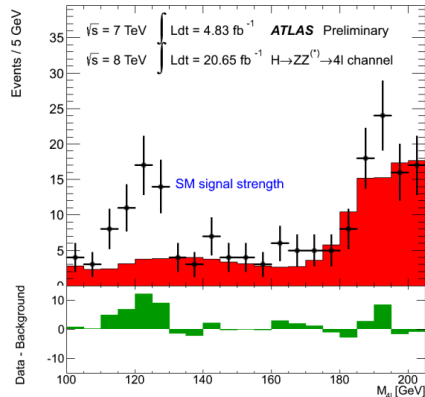
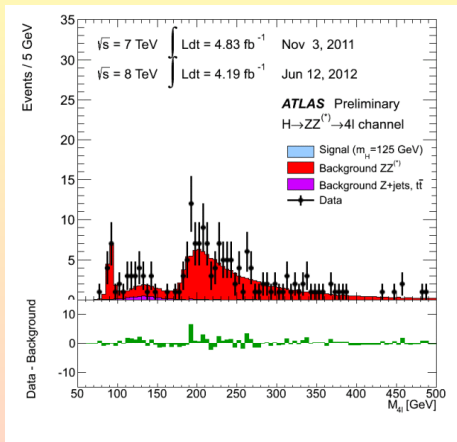
Q: Why do so many people still use  $p = 0.05$ ?

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.

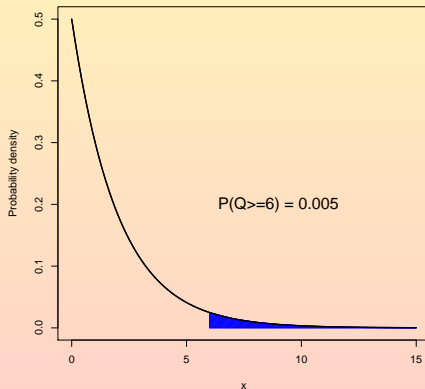
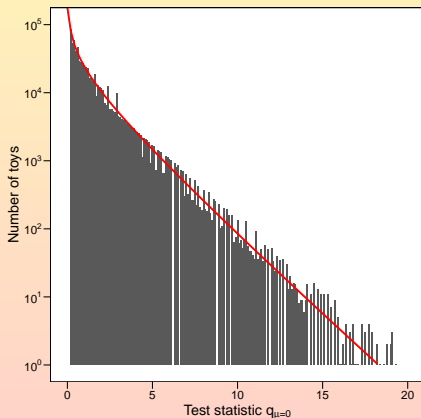
When the ASA Board decided to take up the challenge of developing a policy statement on  $p$ -values and statistical significance, it did so recognizing this was not a lightly taken step. The ASA has not previously taken positions on specific matters of statistical practice. The closest the association 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  $q \sim \chi^2(1)$ )

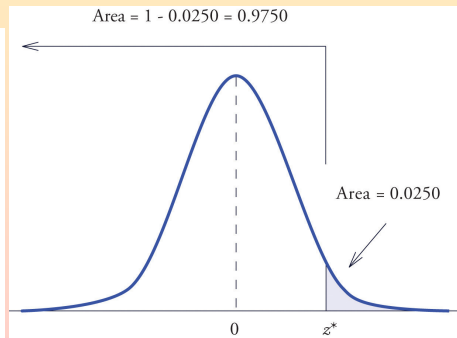
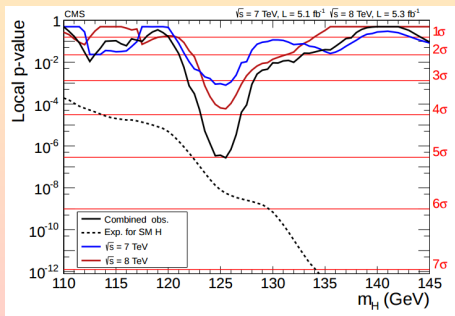
Distribution of  $q_{\mu=0}$  for  $H(\mu=0)$



Plots from Vischia—in preparation with Springer



- 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)

- 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](https://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.”

- loannidis (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>

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  $\frac{d^2 \exp(t)}{dt^2} = \exp(t)$ , 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.

$$\frac{\text{Exp}[r(1+\delta)t] + \text{Exp}[r(1-\delta)t]}{2} > \text{Exp}[rt]$$

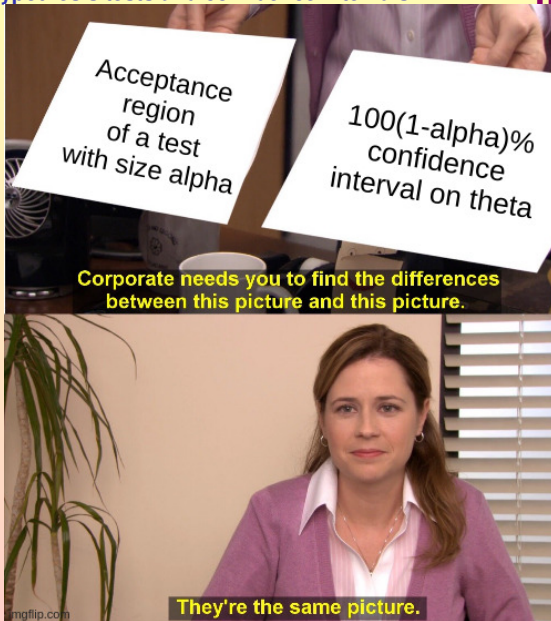
59 241 956

- 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*

- **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)



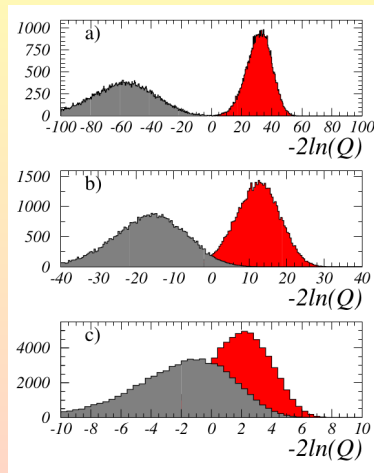


Meme generated with <https://imgflip.com/memegenerator>

- 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!

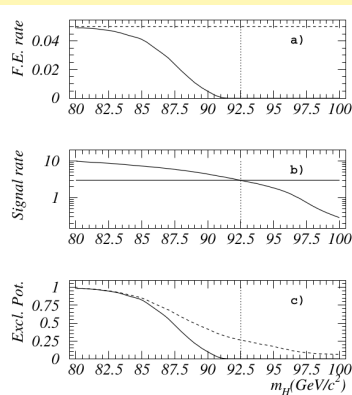
- Example: Find a monotonic  $Q$  for increasing signal-like experiments (e.g. likelihood ratio)
- $CL_{s+b} = P_{s+b}(Q \leq Q_{obs})$ 
  - Small values imply poor compatibility with  $S + B$  hypothesis, favouring  $B$ -only
- Counting experiment: observe  $n$  events
- Assume they come from Poisson 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

## From a scan of $CL_s$ to a limit on a cross section

- Scan the  $CL_s$  test statistic as a function of the POI (typically the cross section modifier  $\mu = \sigma_{obs}/\sigma_{pred}$ )
- Find its intersection with the desired confidence level
- (eventually) convert the limit on  $\mu$  back to a cross section

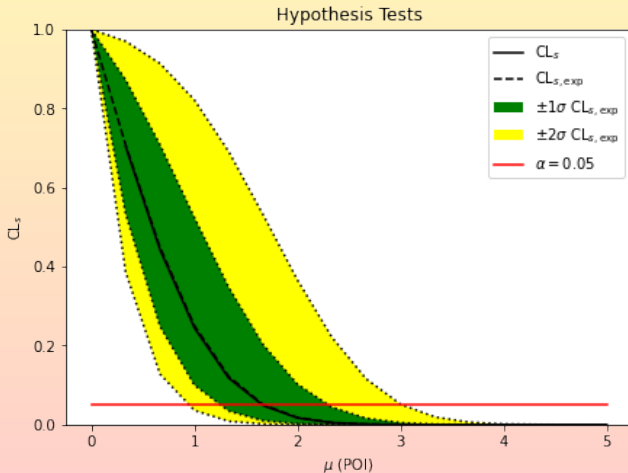
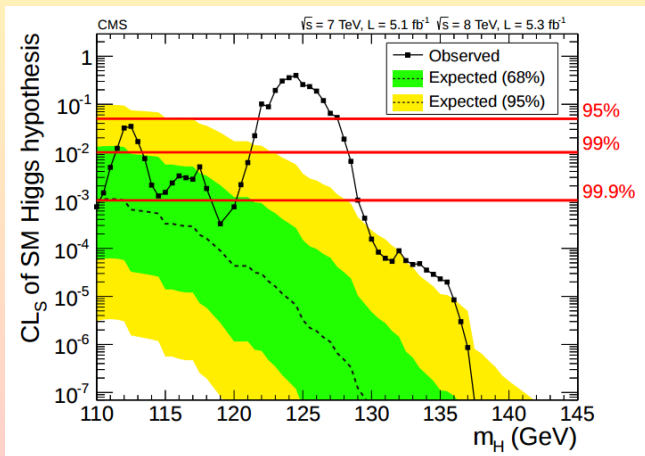


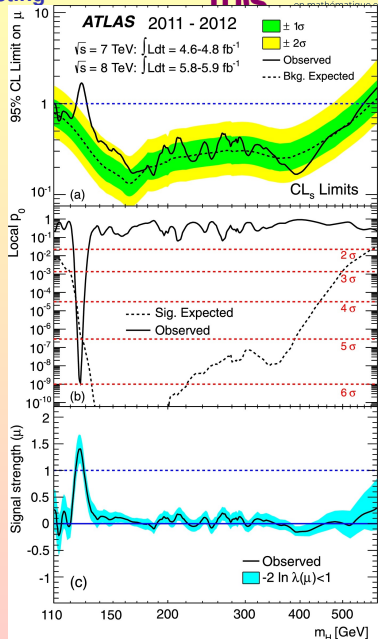
Image from the afternoon exercise on  $CL_s$

- Apply the  $CL_s$  method to each Higgs mass hypothesis
- Show the  $CL_s$  test statistic for each value of the fixed hypothesis
- 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 CMS Higgs discovery paper doi:10.1016/j.physletb.2012.08.021

- CLs limit on  $\mu$  as a function of mass hypothesis
- p-value of excess
- Fitted signal strength peaks at excess

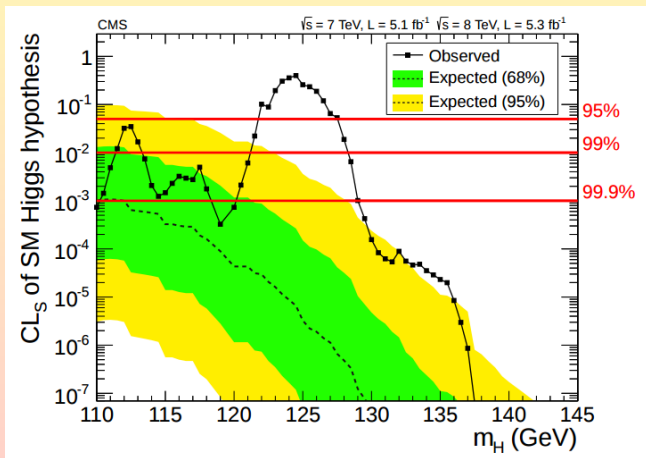


Plot from ATLAS Higgs discovery paper doi:10.1016/j.physletb.2012.08.020



## That's what we used for the Higgs discovery!

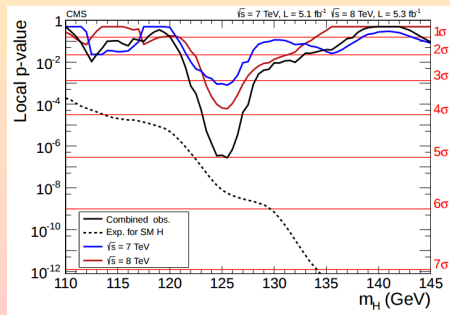
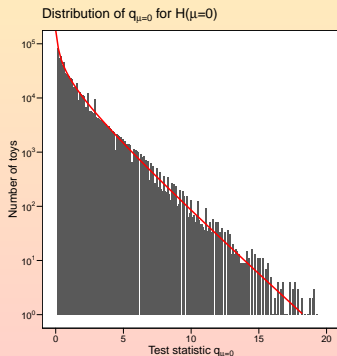
- Apply the  $CL_s$  method to each Higgs mass point
- Green/yellow bands indicate the  $\pm 1\sigma$  and  $\pm 2\sigma$  intervals for the expected values under  $B$ -only hypothesis
  - Obtained by taking the quantiles of the  $B$ -only hypothesis



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,  $q_0 = -2 \log \frac{\mathcal{L}(\text{data}|0, \hat{\theta}_0)}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\theta})}$ , 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 phoneys, 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.

Vischia

Statistics for HEP

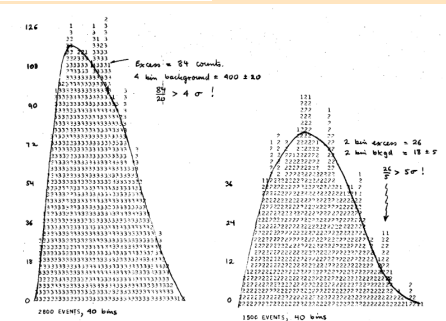
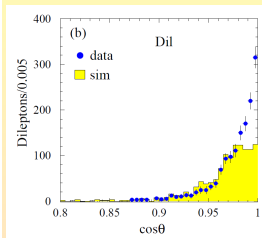
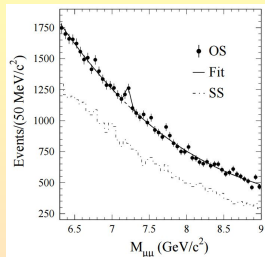


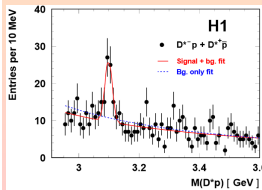
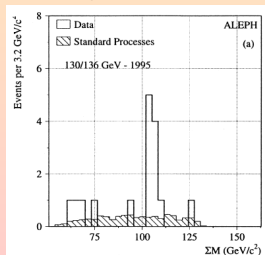
Fig. 3. Two "Las Vegas" histograms generated by G. Lynch's program GAME.

March 16th and 18th 2022

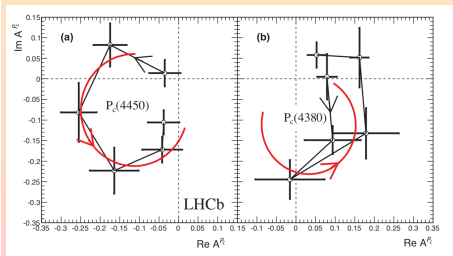
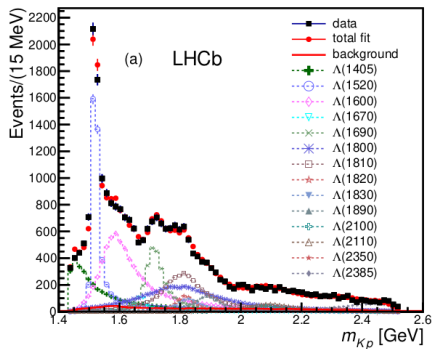
- $3.5\sigma$  (2005, CDF) in dimuon (candidate bottom squark, [doi:10.1103/PhysRevD.72.092003](https://doi.org/10.1103/PhysRevD.72.092003))



- $\sim 4\sigma$  (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.

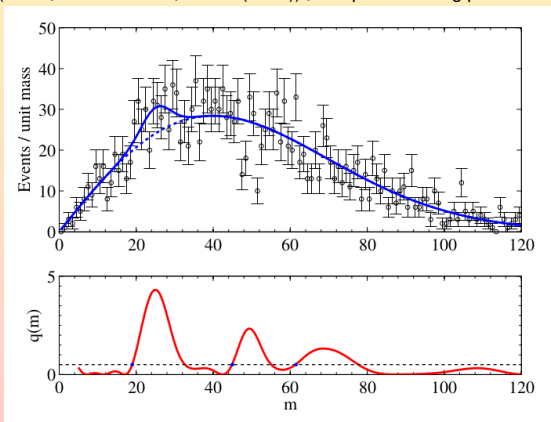


- $9\sigma$  and  $12\sigma$  (2015, LHCb): pentaquarks! ([doi:10.1103/PhysRevLett.115.072001](https://doi.org/10.1103/PhysRevLett.115.072001))
  - Several cross-checks (fit to mass spectrum, fit with non-resonant components, evolution of complex amplitude in Argand diagrams)
  - Mass measurement, soft 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

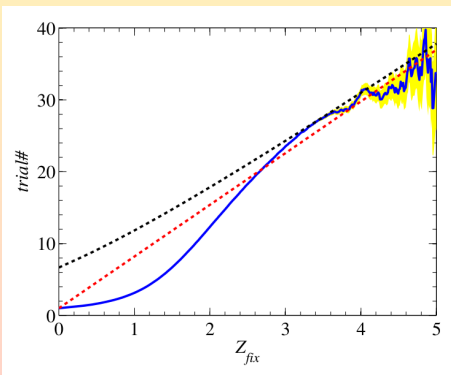
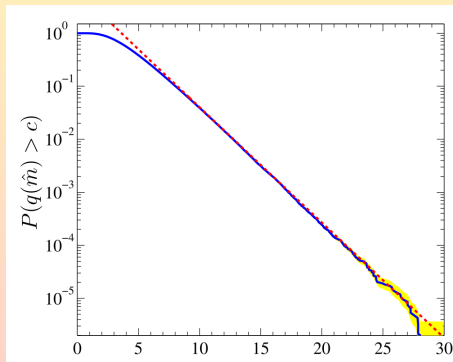
- Searching for a resonance  $X$  of arbitrary mass
  - $H_0$  = no resonance, the mass of the resonance is not defined (Standard Model)
  - $H_1 = H(M \neq 0)$ , but there are infinite possible values of  $M$
- Wilks theorem not valid anymore, no unique test statistic encompassing every possible  $H_1$
- 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  $p_b^{global} := P(q_0(\hat{m}_X) > u) \leq \langle N_u \rangle + \frac{1}{2} P_{\chi^2_1}(u)$
  - $u$  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

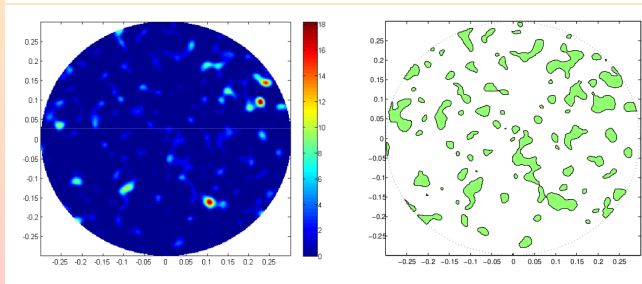
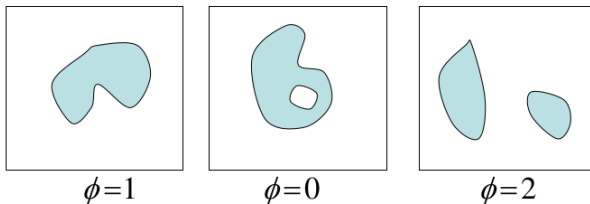


- Ratio of local (excess right here) and global (excess anywhere) p-values: trial factor
- Asymptotically linear in the number of search regions and in the fixed significance level
  - Dashed red lines: prediction based on the formula with upcrossings
  - Blue:  $10^6$  toys (pseudoexperiments)
- Here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

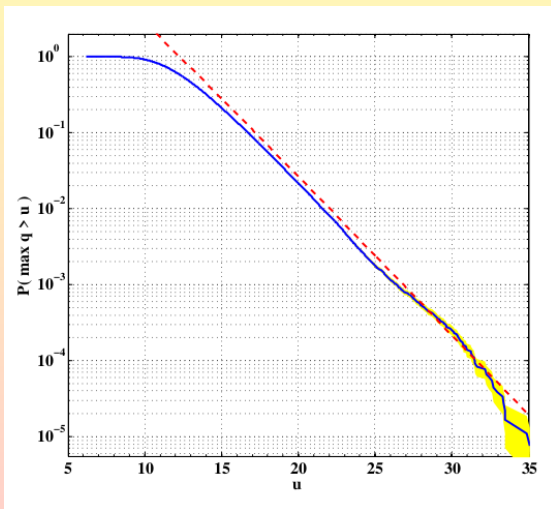


Plot from Gross-Vitells, 10.1140/epjc/s10052-010-1470-8

- 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



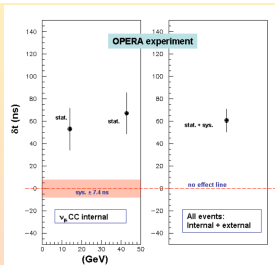
- Asymptoticity holds also for the 2D effect, as desired
  - Dashed red lines: prediction based on the formula with upcrossings
  - Blue: 200k toys (pseudoeperiments)



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'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
 
$$SEV(\mu > \mu_1) = P(Q \leq Q_{obs}; \mu > \mu_1 | \text{false}) = P(Q \leq Q_{obs}; \mu \leq \mu_1) > P(Q \leq Q_{obs}; \mu = \mu_1)$$
  - 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) 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.)

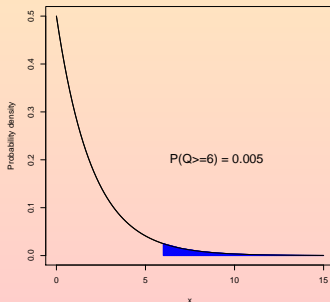
- Others (Gelman, Raftery, Berger, Bernardo) argue that a point null is impossible (at most “small”)

- 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. 18
- 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/

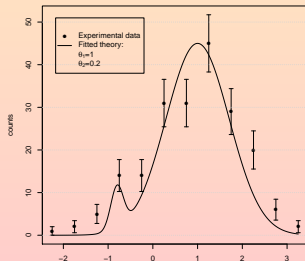
- Consider a set of  $M$  measurements  $\{(X_i, Y_i)\}$ 
  - Suppose  $Y_i$  are affected by a random error representable by a gaussian with variance  $\sigma_i$
- Consider a function  $g(X)$  with predictive capacity, i.e. such that for each  $i$  we have  $g(X_i) \sim Y_i$
- Pearson's  $\chi^2$  function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

$$\chi_P^2 := \sum_{i=1}^M \left[ \frac{Y_i - g(X_i)}{\sigma_i} \right]^2 \quad (19)$$

- 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  $\sigma_i \propto \sqrt{n_i}$

$$\chi_N^2 := \sum_{i=1}^M \frac{(n_i - g(X_i))^2}{n_i} \quad (20)$$



- 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  $\chi^2$ :

$$f(\chi^2, M) = \frac{2^{-\frac{M}{2}}}{\Gamma\left(\frac{M}{2}\right)} \chi^{N-2} e^{-\frac{\chi^2}{2}} \quad (21)$$

- The cumulative of  $f$  is

$$1 - \text{cum}(f) = P(\chi^2 > \chi_{obs}^2 | g(x) \text{ is the correct model}) \quad (22)$$

- 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  
(23)

- 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:

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

$$\frac{\chi^2}{M} \gg 1 \Rightarrow \text{poor model (increases } \chi^2), \text{ or statistically improbable fluctuation} \quad (24)$$

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

- 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:

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

$$\frac{\chi^2}{M} \gg 1 \Rightarrow \text{poor model (increases } \chi^2), \text{ or statistically improbable fluctuation} \quad (24)$$

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

- **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  $\chi^2_{obs}$  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)

- Statistics is about answering questions
  - ...and posing the questions in an appropriate way
- Foundations
  - Mathematical definition of probability
  - Bayesian and Frequentist realizations
- How wide is the table?: Point estimates and the method of maximum likelihood
- Is it really that wide, or am I somehow uncertain about it?: Interval estimates
  - Maximum likelihood
  - Neyman construction
  - Feldman-Cousins ordering
  - Coverage
- Is the table a standard-size ping-pong table or not? Testing hypotheses
  - Frequentist hypothesis testing, and some mention to the Bayesian one
  - I need no toy: the Wilks theorem
  - Upper limits and the  $CL_s$  prescription
- Can I decouple my result from my instrumentation? Unfolding
- How can I exploit learning algorithms? Machine Learning
  - Machine learning is a well defined mathematical technique
  - Used in many flavours across all the spectrum of tasks in HEP
- Are you satisfied? Check your email for the link to the questionnaire about the course!
  - This helps me a lot improving the course over the years!



- I hope this course has helped in broadening the spectrum of techniques you will consider using in the future
- Or at least that it has clarified some of the underlying concepts for techniques you already use!

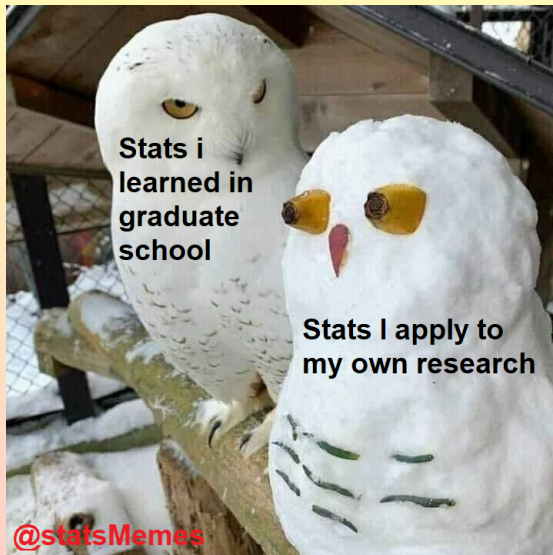


Image from the Statistical Statistics Memes Facebook Page

# THANK YOU VERY MUCH FOR ATTENDING!!

This course has already improved on the fly thanks to you!  
I'll take any further feedback and transforming into improvements for the  
next edition!

- 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