

# Statistics

or “How to find answers to your questions”

Pietro Vischia<sup>1</sup>

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



LIP LHC Physics Course 2021, March 15th–17th

- I hope we go back to in-person courses soon!



- 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/4097/>
  - 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!

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

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

# 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?



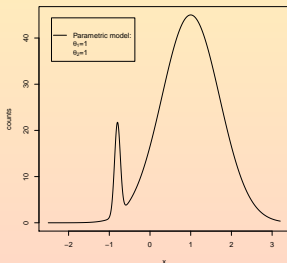
- What is the chance of obtaining a 1 when throwing a six-faced die?
  - We can throw a dice 100 times, and count how many times we obtain 1
- What is the chance of tomorrow being rainy?
  - We can try to give an answer based on the recent past weather, but we cannot – in general – *repeat tomorrow*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



Image from ["The Tiger Lillies" Facebook page](#)

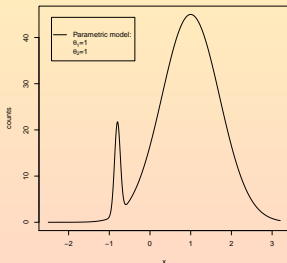
## • Theory

- Approximations
- Free parameters



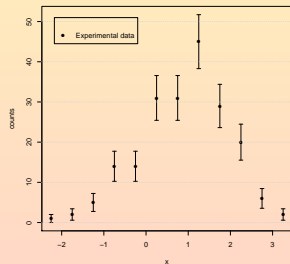
## • Theory

- Approximations
- Free parameters



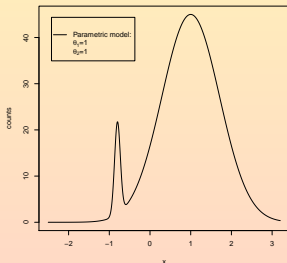
## • Experiment

- Random fluctuations
- Mismeasurements (detector effects, etc)



## • Theory

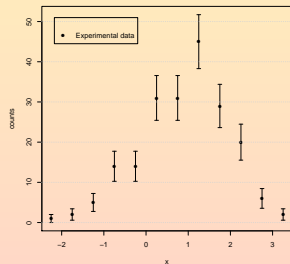
- Approximations
- Free parameters



## • Statistics!

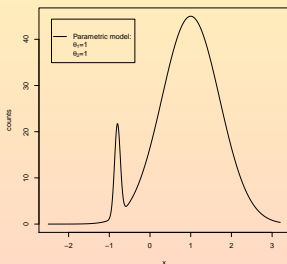
## • Experiment

- Random fluctuations
- Mismeasurements (detector effects, etc)



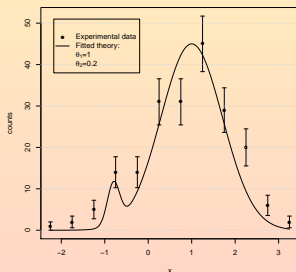
## • Theory

- Approximations
- Free parameters



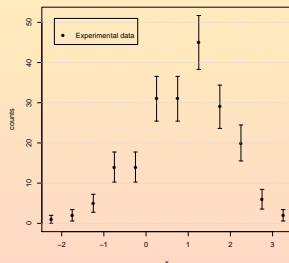
## • Statistics!

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



## • Experiment

- Random fluctuations
- Mismeasurements (detector effects, etc)



# Fundamentals

- $\Omega$ : 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

- 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:

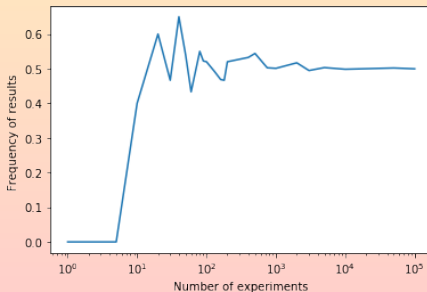
$$P(X) = \lim_{N \rightarrow \infty} \frac{n}{N}$$

- 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
<b>Royal flush</b> 	1	4	0.000154%	0.000154%	649,739 : 1	$\binom{4}{1}$
<b>Straight flush (excluding royal flush)</b> 	9	36	0.00139%	0.0014%	72,192 : 1	$\binom{10}{1}\binom{4}{1} - \binom{4}{1}$
<b>Four of a kind</b> 	156	624	0.0240%	0.0256%	4,164 : 1	$\binom{13}{1}\binom{12}{1}\binom{4}{1}$
<b>Full house</b> 	156	3,744	0.1441%	0.17%	699 : 1	$\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}$
<b>Flush (excluding royal flush and straight flush)</b> 	1,277	5,108	0.1965%	0.267%	508 : 1	$\binom{13}{5}\binom{4}{1} - \binom{10}{1}\binom{4}{1}$
<b>Straight (excluding royal flush and straight flush)</b> 	10	16,200	0.6285%	0.78%	264 : 1	$\binom{10}{1}\binom{4}{1}^5 - \binom{10}{1}\binom{4}{1}$
<b>Three of a kind</b> 	858	64,932	2.1128%	2.87%	46.2 : 1	$\binom{13}{1}\binom{4}{3}\binom{12}{2}\binom{4}{1}^2$
<b>Two pair</b> 	858	128,852	4.7633%	7.62%	29.8 : 1	$\binom{13}{2}\binom{4}{2}^2\binom{11}{1}\binom{4}{1}$
<b>One pair</b> 	2,860	1,098,240	42.2569%	49.5%	1.97 : 1	$\binom{13}{1}\binom{4}{2}\binom{12}{3}\binom{4}{1}^3$
<b>High card / High card</b> 	1,277	1,612,640	60.2177%	100%	0.996 : 1	$\left[\binom{13}{5} - 10\right] \left[\binom{4}{1}^5 - 4\right]$
<b>Wild</b>	7,462	2,598,960	100%	—	0 : 1	$\binom{52}{5}$

- 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 fixed 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 fixed 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!



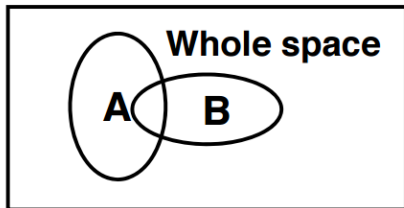
- Based on the concept of degree of belief
  - $P(X)$  is the subjective degree of belief on  $X$  being true
- De Finetti: operative definition of subjective probability, based on the concept of coherent bet
  - We want to determine  $P(X)$ ; we assume that if you bet on  $X$ , you win a fixed amount 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)



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

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

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

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

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

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

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

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

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

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

- **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/)

- 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,  $P(\text{hypothesis})$  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  $P(\text{data}|\text{model})$
- 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  $S(\text{data})$  with known distribution, and computes how likely is to see  $S(\text{data}_{\text{sampled}}) = S(\text{data}_{\text{obs}})$  when sampling from the model

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

- $\vec{X}$ , 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

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

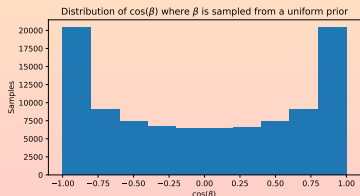
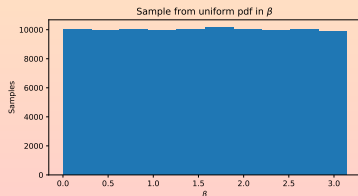
- $P(H|\vec{X})$ , the posterior probability; it is obtained as a result of an experiment
- If we parameterize  $H$  with a (continuous or discrete) parameter, we can use the parameter as a proxy for  $H$ , and instead of writing  $P(H(\theta))$  we write  $P(\theta)$  and

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

- 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”)



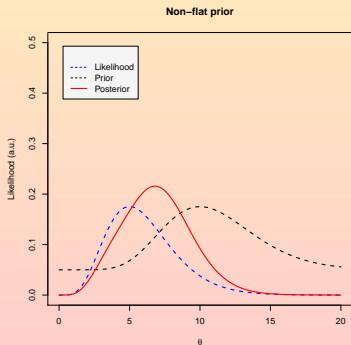
- Two ways of solving the situation
  - Objective Bayes: use a formal rule dictated by an invariance principle
  - Subjective Bayes: use something like elicitation of expert opinion
    - Ask an expert her opinion about each value of  $\theta$ , and express the answer as a curve
    - Repeat this with 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"
- Central concept: update your priors after each experiment

- 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 uniform 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

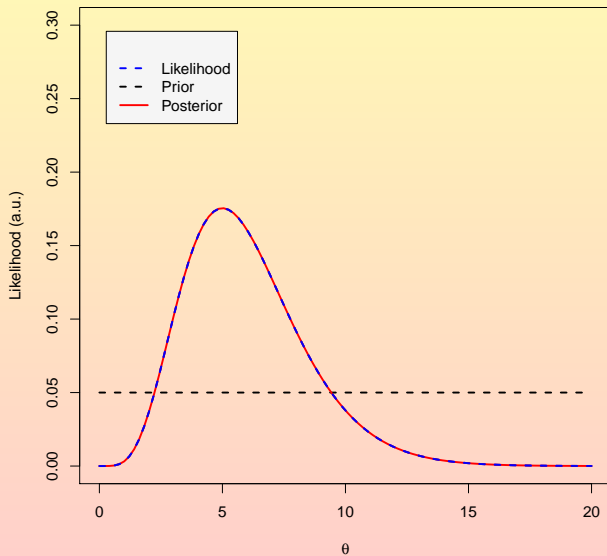
- Uniform priors must make sense
  - Uniform prior across its entire dominion: not very realistic
  - It corresponds to claiming that  $P(1 < \theta \leq 2)$  is the same as  $P(10^{41} < \theta \leq 10^{41} + 1)$
  - It's irrational to claim that a prior can cover uniformly forty orders of magnitude
  - 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

## The gist is that priors can and will impact the posterior

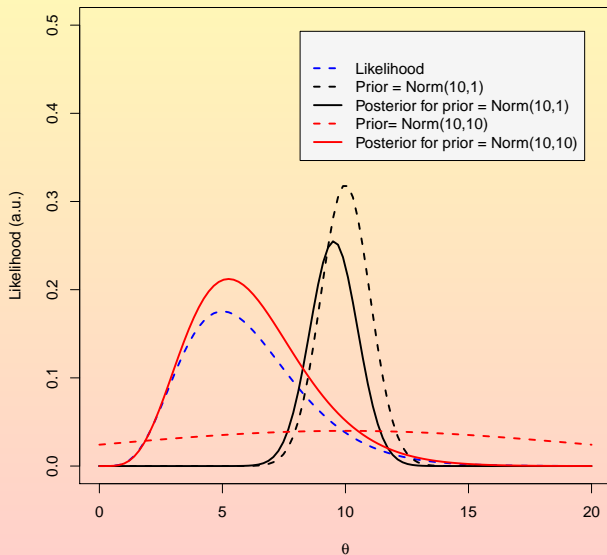
- Associating parametric priors to intervals in the parameter space corresponds to considering sets of theories
  - This is because to each value of a parameter corresponds a different theory
- In practical situations, note (Eq. 9) posterior probability is always proportional to the product of the prior 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



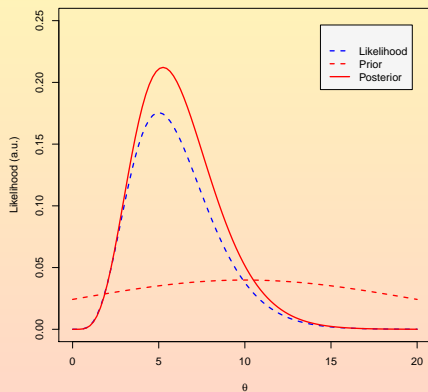
## Flat prior



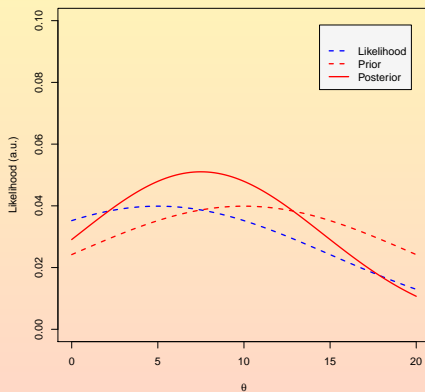
## Broad prior vs narrow prior



### 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

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

- Theory of probability originated in the context of games of chance
- Mathematical roots in the theory of Lebesgue measure and set functions in  $\mathbb{R}^n$
- Measure is something we want to define for an interval in  $\mathbb{R}^n$ 
  - 1D: the usual notion of length
  - 2D: the usual notion of area
  - 3D: the usual notion of volume
- Interval  $i = a_\nu \leq x_\nu \leq a_\nu$

$$L(i) = \prod_{\nu=1}^n (b_\nu - a_\nu).$$

- The length of degenerate intervals  $a_\nu = b_\nu$  is  $L(i) = 0$ ; it does therefore not matter the interval is closed, open, or half-open;
- We set to  $+\infty$  the length of any infinite non-degenerate interval such as  $]25, +\infty]$  or  $[-\infty, 2]$ .
- But do we connect different intervals?

- Disjoint intervals (no common point with any other)

$$i = i_1 + \dots + i_n, \quad (i_\mu i_\nu = 0 \text{ for } \mu \neq \nu);$$

- We can extend this to enumerable sequences of intervals by using the Borel lemma (ensures we can do that)
  - Crucial to move from discrete sums to integrals of continuous probability functions
- We can build a generalization of the Lebesgue measure  $L_n(S)$ : the P-measure
  - 1  $P(S)$  is non-negative,  $P(S) \geq 0$ ;
  - 2  $P(S)$  is additive,  $P(S_1 + \dots + S_n) = P(S_1) + \dots + P(S_n)$  where  $S_\mu S_\nu = 0$  for  $\mu \neq \nu$ ;
  - 3  $P(S)$  is finite for any bounded set (crucial to define the usual probability in the domain  $[0, 1]$ )
- Have we already seen these three properties today?

- Consider a class of non-negative additive set functions  $P(S)$  such that  $P(\mathbb{R}^n) = 1$ ; then

$$F(\mathbf{x}) = F(x_1, \dots, x_n) = P(\xi \leq x_1, \dots, \xi_n \leq x_n)$$

$$0 \leq F(\mathbf{x}) \leq 1$$

$$\Delta_n F \geq 0$$

$$F(-\infty, x_2, \dots, x_n) = \dots = F(x_1, \dots, x_n - 1, -\infty) = 0$$

$$F(+\infty, \dots, +\infty) = 1.$$

- We interpret  $P(S)$  and  $F(\mathbf{x})$  as distribution of a unit of mass over  $\mathbb{R}^n$ 
  - Each Borel set carries the mass  $P(S)$
  - Interpret  $\mathbf{x}$  as the quantity of mass allotted to the infinite interval  $(\xi_1 \leq x_1, \dots, \xi_n \leq x_n)$ .
  - Defining the measure in terms of  $P(S)$  or  $F(\mathbf{x})$  is equivalent
- Usually  $P(S)$  is called probability function, and  $F(\mathbf{x})$  is called distribution function
- $\sigma$ -field: a space  $\Omega$  equipped with a collection of subsets containing  $\Omega$ , closed by complement and by under countable union
  - The original Kolmogorov approach is expressed via a  $\sigma$ -field built on the space of elementary propositions (sets)

- Discrete mass point  $a$ ; a point such that the set  $\{x = a\}$  carries a positive quantity of mass.

$$P(S) = c_1 P_1(S) + c_2 P_2(S)$$

or

$$F(x) = c_1 F_1(x) + c_2 F_2(x)$$

where

$$c_\nu \geq 0, \quad c_1 + c_2 = 1,$$

- $c_1$ : component with whole mass concentrated in discrete mass points.  $c_2$ : component with no discrete mass points
- $c_1 = 1, c_2 = 0$ :  $F(x)$  is a step function, where the whole mass is concentrated in the discontinuity points
- $c_1 = 0, c_2 = 1$ , then if  $n = 1$  then  $F(x)$  is everywhere continuous, and in any dimension no single mass point carries a positive quantity of mass.



- Consider the  $n$ -dimensional interval  $i = \{x_\nu - h_\nu < \xi_\nu \leq x_\nu + h_\nu; \nu = 1, \dots, n\}$
- Average density of mass: the ratio of the P-measure of the interval—expressed in terms of the increments of the point function—to the L-measure of the interval itself

$$\frac{P(i)}{L(i)} = \frac{\Delta_n F}{2^n h_1 h_2 \dots h_n}.$$

- If partial derivatives  $f(x_1, \dots, x_n) = \frac{\partial_n F}{\partial x_1 \dots \partial x_n}$  exist, then  $\frac{P(i)}{L(i)} \rightarrow f(x_1, \dots, x_n)$  for  $h_\nu \rightarrow 0$ 
  - Density of mass at the point  $x$
  - $f$  is referred to as probability density or frequency function

- Take a distribution function  $F(x_1, \dots, x_n)$
- Let  $x_\mu \rightarrow \infty, \mu \neq \nu$
- It can be shown that  $F \rightarrow F_\nu(x_\nu)$ , and that itself is a distribution function in the variable  $x_\nu$ 
  - e.g.  $F_1(x_1) = F(x_1, +\infty, \dots, +\infty)$ .
- $F_\nu(x_\nu)$  is one-dimensional, and is called the marginal distribution of  $x_\nu$ .
  - It can be obtained by projection starting from the  $n$ -dimensional distribution
  - Shift each “mass particle” along the perpendicular direction to  $x_\nu$  until collapsing into the  $x_\nu$  axis
  - This results in a one-dimensional distribution which is the marginal distribution of  $x_\nu$ .
  - There are infinite ways of arriving to the same  $x_\nu$  starting from a generic  $n$ -dimensional distribution function
- Marginal distributions can be also built with respect to subsets of variables.

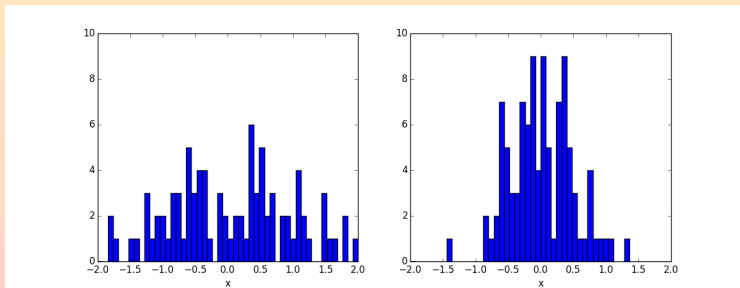
- Extend the concept of p.d.f. to an arbitrary number of variables; the joint p.d.f.  $f(X, Y, \dots)$
- If we are interested in the p.d.f. of just one of the variables the joint p.d.f. depends upon, we can compute by integration the marginal p.d.f.

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

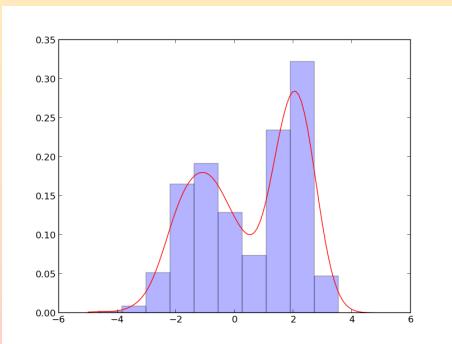
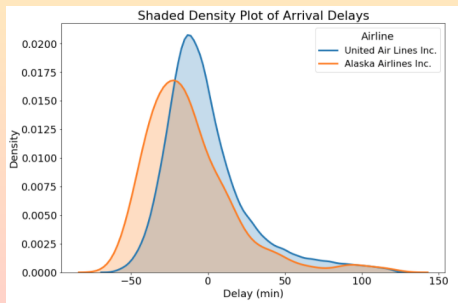
- Sometimes it's interesting to express the joint p.d.f. as a function of one variable, for a particular fixed value of the others: this is the conditional p.d.f. :

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

- 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

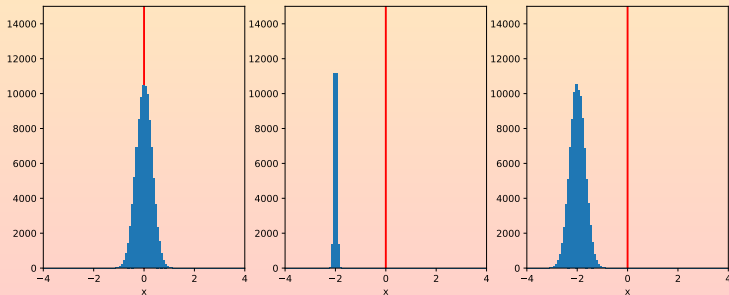


- HEP uses histograms mostly historically: counting experiments
- Statistics and Machine Learning communities typically use densities
  - Intuitive relationship with the underlying p.d.f.
  - Kernel density estimates: binning assumption  $\rightarrow$  bandwidth assumption
  - Less focused on individual 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  $g(X)$  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:  $V[g] = E[(g - \mu_g)^2] = \int (g(\vec{X}) - \mu_g)^2f(\vec{X})dX_1dX_2\dots dX_n = \sigma_g^2$

- 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)

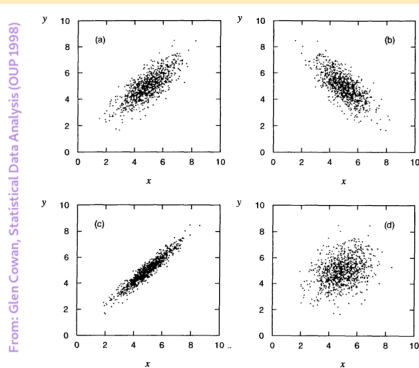


Fig. 1.9 Scatter plots of random variables  $x$  and  $y$  with (a) a positive correlation,  $\rho = 0.75$ , (b) a negative correlation,  $\rho = -0.75$ , (c)  $\rho = 0.95$ , and (d)  $\rho = 0.25$ . For all four cases the standard deviations of  $x$  and  $y$  are  $\sigma_x = \sigma_y = 1$ .

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

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

- 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

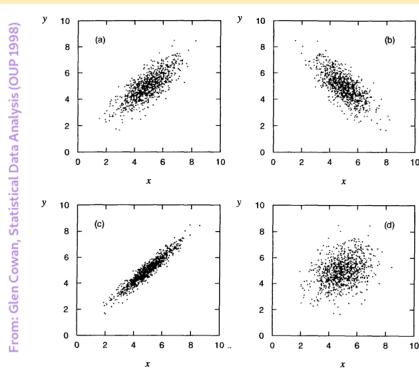


Fig. 1.9 Scatter plots of random variables  $x$  and  $y$  with (a) a positive correlation,  $\rho = 0.75$ , (b) a negative correlation,  $\rho = -0.75$ , (c)  $\rho = 0.95$ , and (d)  $\rho = 0.25$ . For all four cases the standard deviations of  $x$  and  $y$  are  $\sigma_x = \sigma_y = 1$ .

- 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  $\rho_{XY}$

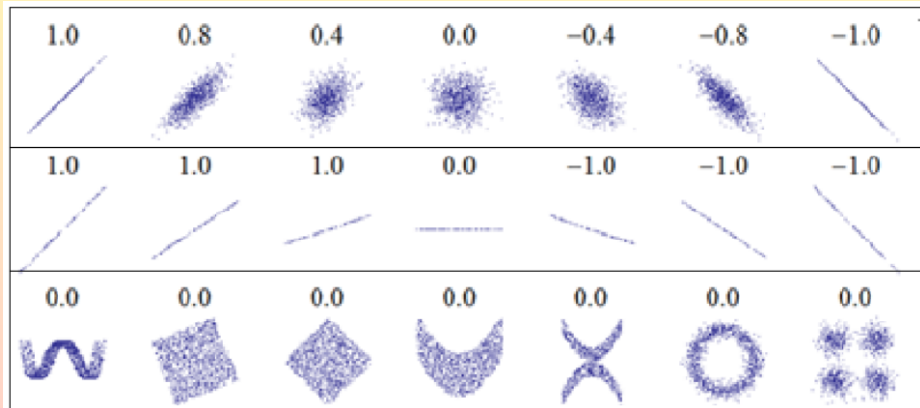


Figure from BND2010

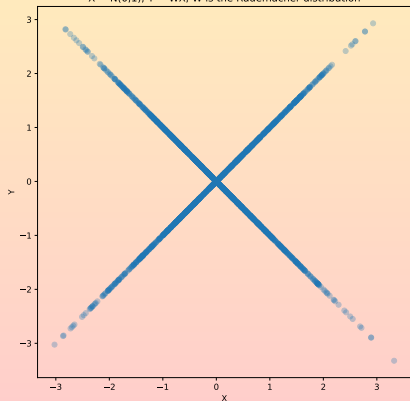
## Take it to the next level: the Mutual Information

- 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

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

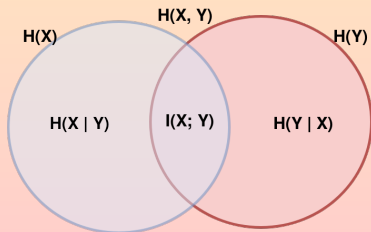
- Symmetric:  $I(X; Y) = I(Y; X)$
- $I(X; Y) = 0$  if and only if  $X$  and  $Y$  are totally independent
  - $X$  and  $Y$  can be uncorrelated but not independent; mutual information captures this!

$X = N(0,1)$ ;  $Y = WX$ ;  $W$  is the Rademacher distribution

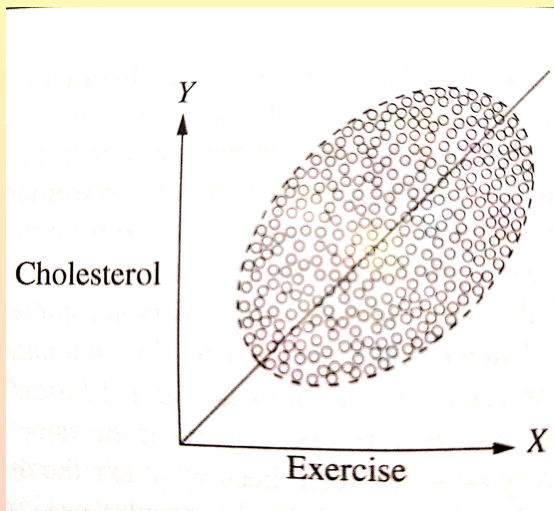


- Related to entropy

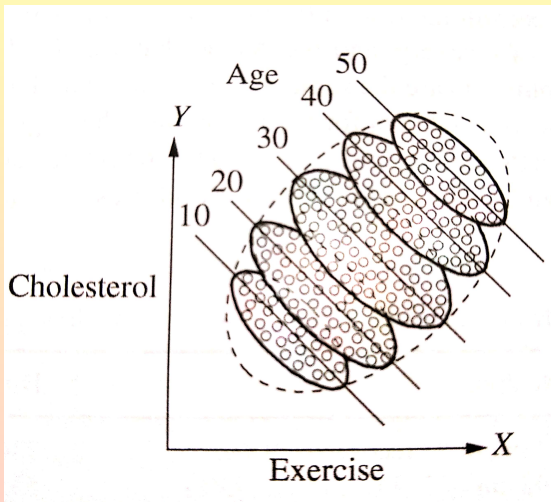
$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X, Y) \end{aligned}$$



## Does cholesterol increase with exercise?



- Question time: Cholesterol



- If we know the biological sex<sup>3</sup>, then prescribe the drug
- If we don't know the biological sex, then don't prescribe the drug

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

- Question time: DrugEffectiveness

---

<sup>3</sup>Biological sex: *anatomy of an individual's reproductive system, and secondary sex characteristics*. Gender: *either social roles based on the sex of the person (gender role) or personal identification of one's own gender based on an internal awareness* ([https://en.wikipedia.org/wiki/Sex\\_and\\_gender\\_distinction](https://en.wikipedia.org/wiki/Sex_and_gender_distinction))

- If we know the biological sex<sup>3</sup>, then prescribe the drug
- If we don't know the biological sex, then don't prescribe the drug

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

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

---

<sup>3</sup>Biological sex: *anatomy of an individual's reproductive system, and secondary sex characteristics*. Gender: *either social roles based on the sex of the person (gender role) or personal identification of one's own gender based on an internal awareness* ([https://en.wikipedia.org/wiki/Sex\\_and\\_gender\\_distinction](https://en.wikipedia.org/wiki/Sex_and_gender_distinction))



- If we know the biological sex<sup>3</sup>, then prescribe the drug
- If we don't know the biological sex, then don't prescribe the drug

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

- **Question time: DrugEffectiveness**
- Imagine we know that estrogen has a negative effect on recovery
  - Then women less likely to recovery than men
  - Table shows women are significantly more likely to take the drug
  - Consult the separate data to decide on the drug, in order not to mix effects

---

<sup>3</sup>Biological sex: *anatomy of an individual's reproductive system, and secondary sex characteristics*. Gender: *either social roles based on the sex of the person (gender role) or personal identification of one's own gender based on an internal awareness* ([https://en.wikipedia.org/wiki/Sex\\_and\\_gender\\_distinction](https://en.wikipedia.org/wiki/Sex_and_gender_distinction))

- BP = Blood Pressure

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

- Question time: DrugEffectiveness

- BP = Blood Pressure

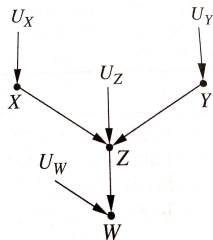
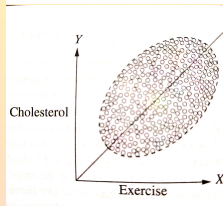
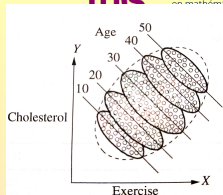
	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)

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

# The Simpson paradox: correlation is not causation

- Correlation alone can lead to nonsense conclusions
- Example 1: estrogen causes a different recovery pattern
  - Use separate data, to avoid mixing proportion of subjects with real drug effect
- Same table, different labels; must consider the combined data
  - Lowering blood pressure is actually part of the mechanism of the drug effect
- It's the causal structure that informs whether e.g. to pick the combined or separate data
- Same effect in continuous data (cholesterol vs age)
- The best solution so far are Bayesian causal networks
  - Graph theory to describe relationship between variables

Figures from Pearl, 2016



## First level of causal hierarchy: seeing

- $X$  and  $Y$  are marginally dependent, but conditionally independent given  $Z$ 
  - Same concept we have seen (with a more dramatic effect) in the cholesterol example
- Conditioning on  $Z$  blocks the path

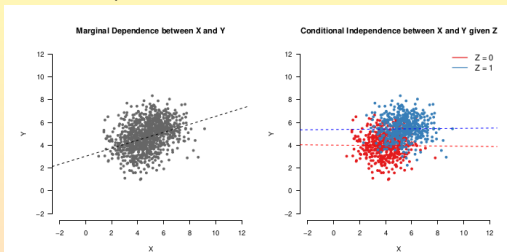


Figure 2. Left: Shows marginal dependence between  $X$  and  $Y$ . Right: Shows conditional independence between  $X$  and  $Y$  given  $Z$ .

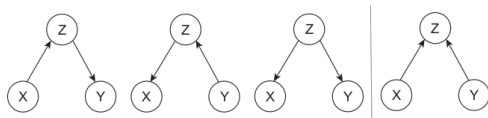


Figure 3. The first three DAGs encode the same conditional independence structure,  $X \perp\!\!\!\perp Y \mid Z$ . In the fourth DAG,  $Z$  is a collider such that  $X \not\perp\!\!\!\perp Y \mid Z$ .

Figures from Dablander, 2019

## First level of causal hierarchy: seeing

- $X$  and  $Y$  are marginally independent, but conditionally dependent given  $Z$ 
  - $Z$  is called a *collider* (not the particle physics one ☺)
- Conditioning on  $Z$  induces *collider bias*

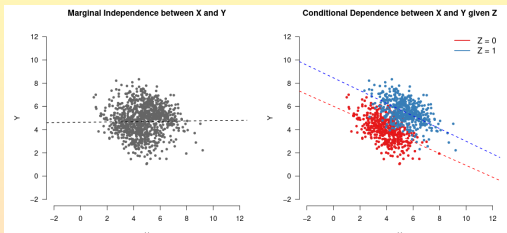


Figure 4. Left: Shows marginal independence between  $X$  and  $Y$ . Right: Shows conditional dependence between  $X$  and  $Y$  given  $Z$

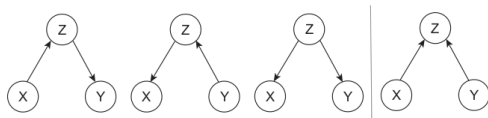


Figure 3. The first three DAGs encode the same conditional independence structure,  $X \perp\!\!\!\perp Y \mid Z$ . In the fourth DAG,  $Z$  is a collider such that  $X \not\perp\!\!\!\perp Y \mid Z$ .

Figures from Dablander, 2019

## Second level of causal hierarchy: doing

- Interventionist approach (Pearl, 2016) (not everyone agrees with this formal approach)
  - $X$  has a causal influence on  $Y$  if changing  $X$  leads to changes in (the distribution of)  $Y$
- Setting (by intervention)  $X = x$  cuts all incoming causal arrows
  - The value of  $X$  is determined only by the intervention
  - Must be able to do intervention: not mere conditioning (seeing): from  $P(Y|X = x)$  to  $P(Y|do(X = x))$
  - Difficult in social sciences
- Intervention discriminates between causal structure of different diagrams
  - Assuming that there is no unobserved confounding (i.e. all causal relationships are represented in the DAG)

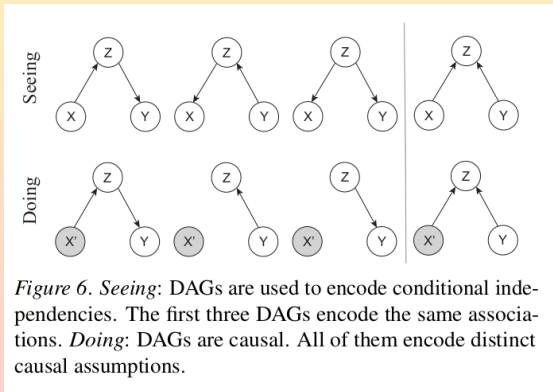


Figure 6. *Seeing*: DAGs are used to encode conditional independencies. The first three DAGs encode the same associations. *Doing*: DAGs are causal. All of them encode distinct causal assumptions.

Figures from Dablander, 2019

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

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

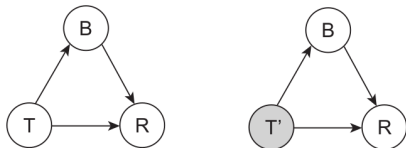


Figure 8. Underlying causal DAG of the example with treatment ( $T$ ), blood pressure ( $B$ ), and recovery ( $R$ ).

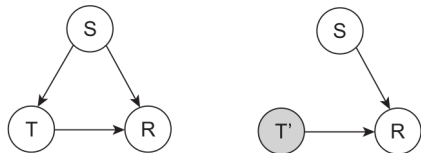


Figure 7. Underlying causal DAG of the example with treatment ( $T$ ), biological sex ( $S$ ), and recovery ( $R$ ).

Figures from Dablander, 2019



## “Doing” is for populations

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

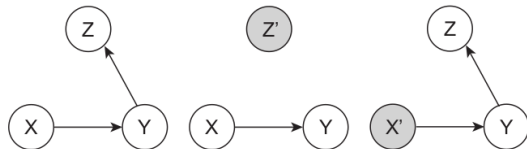


Figure 9. An excellent predictor ( $Z$ ) need not be causally effective.

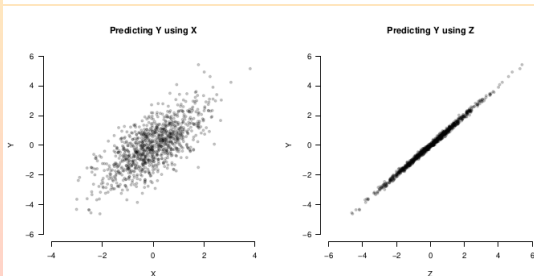


Figure 10.  $X$  is a considerably worse predictor of  $Y$  than  $Z$ .

Figures from Dablander, 2019

### Third level of causal hierarchy: imagining

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

Table 4

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

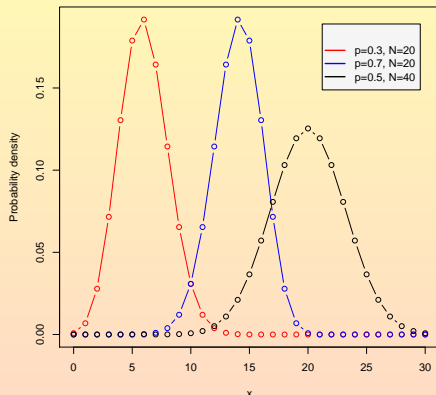
Patient	Treatment	Recovery	$\epsilon_k$
1	0	5.80	0.80
2	0	3.78	-1.22
3	1	3.68	0.68
4	1	0.74	-2.26
5	0	7.87	2.87

Figures and quote from from Dablander, 2019

## 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.

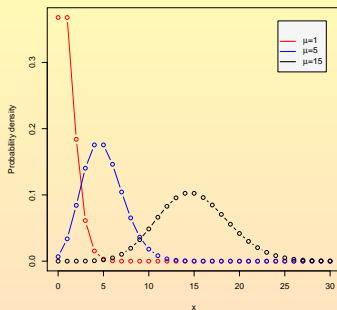


- Example: which is the probability of obtaining 3 times the number 6 when throwing a 6-faces die 12 times?

- $N = 12$ ,  $r = 3$ ,  $p = \frac{1}{6}$

- $$P(3) = \binom{12}{3} \left(\frac{1}{6}\right)^3 \left(1 - \frac{1}{6}\right)^{12-3} = \frac{12!}{3!9!} \frac{1}{6^3} \left(\frac{5}{6}\right)^9 = 0.1974$$

Poisson p.d.f.



## • 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?

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

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

## • Gaussian or Normal distribution

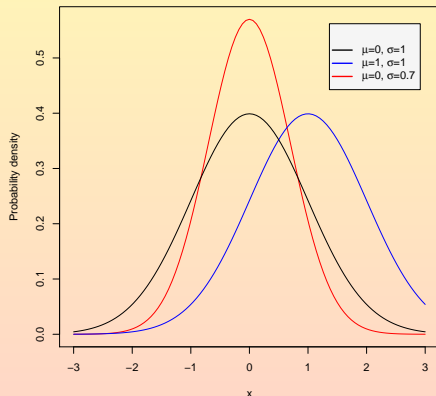
- Variable:  $X$ , real number
- Parameters:
  - $\mu$ , real number
  - $\sigma$ , positive real number

- 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

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

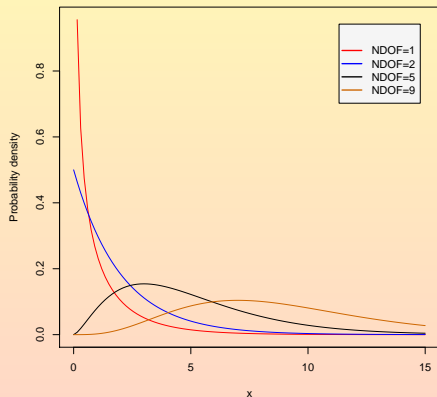
$$E[r] = N$$

$$V(r) = 2N$$

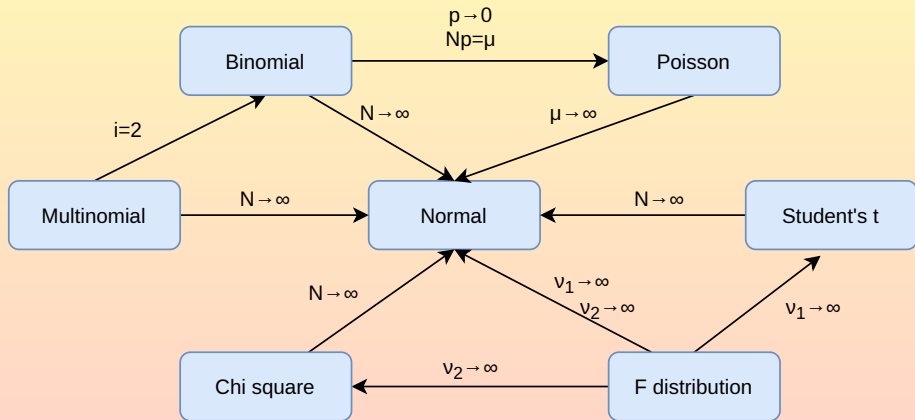
- It describes the distribution of the sum of the squares of a random variable,  $\sum_{i=1}^N X_i^2$

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

$\chi^2$  p.d.f.



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



## Point and Interval estimation



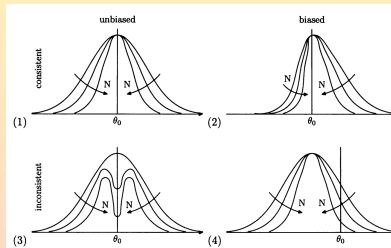
# Estimating a physical quantity

## First we need a function connecting the data and our parameter

- 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  $N$ :

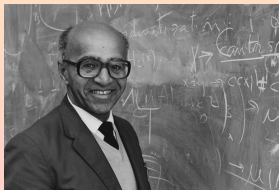
$$\lim_{N \rightarrow \infty} \hat{\vec{\theta}} = \vec{\theta}_{true}$$

- $\hat{\vec{\theta}}$  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.)
- **Once we have our estimate (the numerical value of the estimator for our data), can we throw away our data?**



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: *H T T H H H T H H T T T H T H T H*
  - **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: *HTTHHHHTHTTTHTHTH*
  - **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: *H T T H H H T H H T T T H T H T H*
  - **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

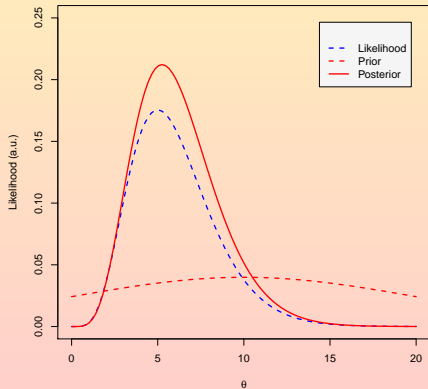
- Two likelihoods differing by only a normalization factor are equivalent
  - Implies that information resides in the shape of the likelihood
- George Bernard: replace a dataset  $D$  with a dataset  $D + Z$ , where  $Z$  is the result of tossing a coin
  - Assume that the coin toss is independent on 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

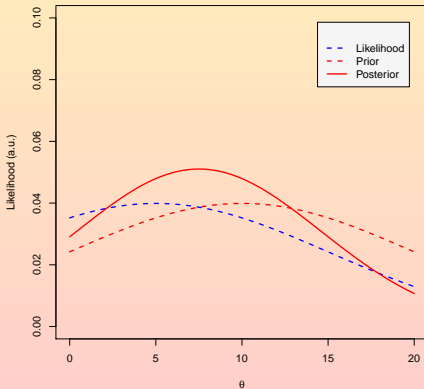
- A very narrow likelihood will provide much information about  $\theta_{true}$ 
  - The posterior probability will be more localized than the prior in the regimen in which the likelihood function dominates the product  $L(\vec{x}; \vec{\theta}) \times \pi$
  - Ideally we'd want to connect this with 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

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

- 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 to 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,  $\vec{\theta}_{ML} = \max P(\vec{\theta}|\vec{x})$ .

## Example: Nuclear Decay with Maximum Likelihood Method

- Go through it also using the exercise!
- 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.

## Example: Nuclear Decay with Maximum Likelihood Method

- Go through it also using the exercise!
- 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

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

## Example: Nuclear Decay with Maximum Likelihood Method

- Go through it also using the exercise!
- 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,  
 $L(\tau) := f(t_i; \tau)$



## Example: Nuclear Decay with Maximum Likelihood Method

- Go through it also using the exercise!
- 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

## Example: Nuclear Decay with Maximum Likelihood Method

- Go through it also using the exercise!
- 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

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

## Example: 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**

## Example: 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:

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

## Example: 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**

## Example: 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

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

## Example: 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!

	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.

## Example: 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!

	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.



## Example: 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.

## Example: 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  $\simeq$ :

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

- The variance is related to the curvature of the likelihood!!!

- Giving a point estimate without specifying a range of variation corresponding to our uncertainty is meaningless!!!
- 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%.



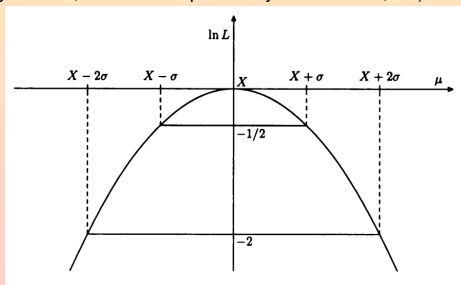
- We can therefore write

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

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

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

- 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.

- 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:

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

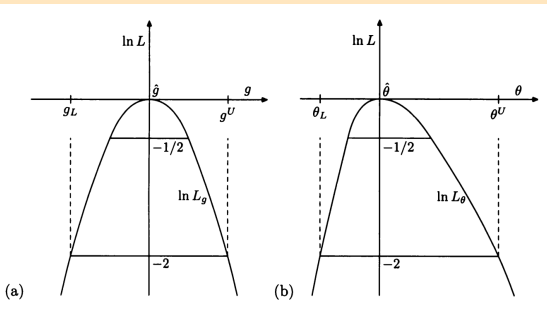
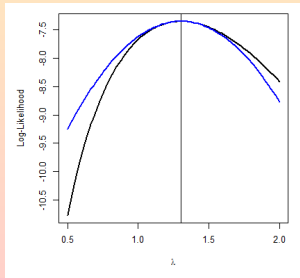
- where  $H$  is the information matrix  $I(\vec{\theta})$ .
- Under these conditions,  $V[\vec{\theta}_{ML}] \rightarrow \frac{1}{I(\vec{\theta}_{ML})}$ , and the intervals can be computed as:

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

- 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  $\vec{\theta}_{true}$

## How to extract an interval from the likelihood function /4

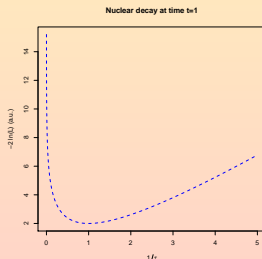
- MLE is invariant for monotonic transformations of  $\theta$ 
  - This applies not only to the maximum of the likelihood, but to all relative values
  - The likelihood ratio is therefore an invariant quantity (we'll use it for hypothesis testing)
  - Can transform the likelihood such that  $\log(L(\vec{x}; \vec{\theta}))$  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  $\mathcal{O}(\frac{1}{N})$



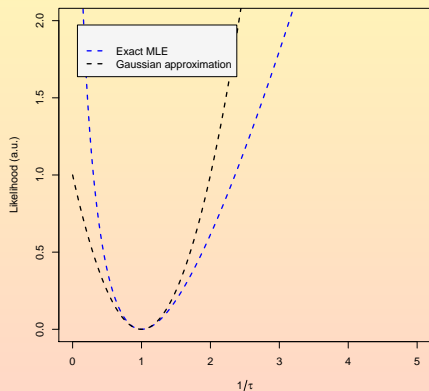
Plot from James, 2nd ed.

- $\vec{\theta}_{true}$  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  $\vec{\theta}_{true}$

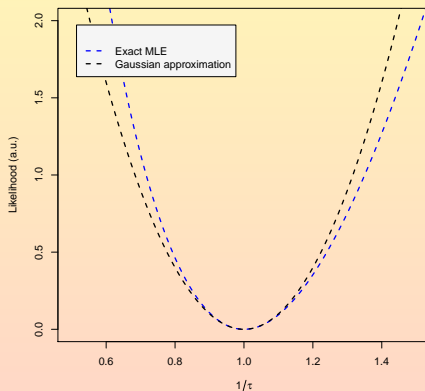
- How good is the approximation  $L(\vec{x}; \vec{\theta}) \propto \exp\left[-\frac{1}{2}(\vec{\theta} - \vec{\theta}_{MLE})^T H(\vec{\theta} - \vec{\theta}_{ML})\right]$ ?
  - Here  $H$  is the information matrix  $I(\vec{\theta})$
  - True only to  $\mathcal{O}(\frac{1}{N})$
  - In these conditions,  $V[\vec{\theta}_{ML}] \rightarrow \frac{1}{I(\vec{\theta}_{ML})}$
  - 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!**



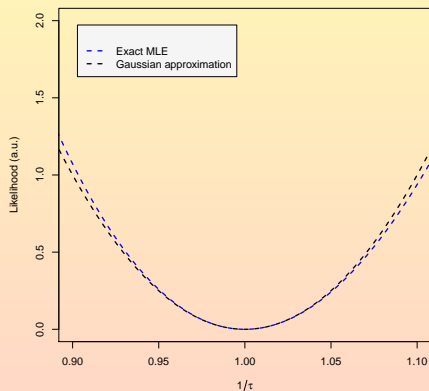
Nuclear decay at time  $t=1$  and  $N=1$



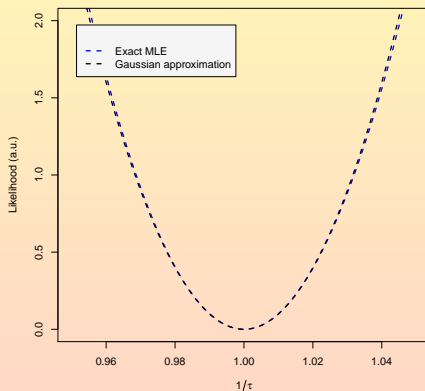
Nuclear decay at time  $t=1$  and  $N=10$



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:

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

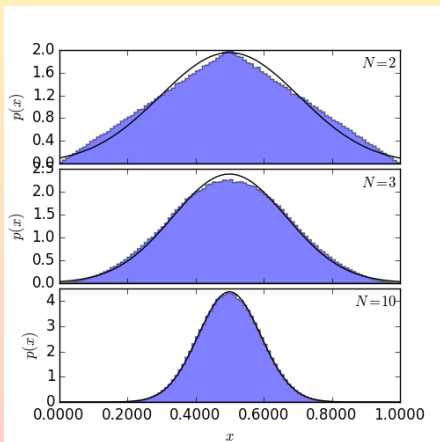
- The demonstration runs by expanding in series the characteristic function  $y_i = \frac{x_j - \mu_j}{\sqrt{\sigma_j}}$
- The theorem is valid for any p.d.f.  $f(x, \vec{\theta})$  that is reasonably peaked around its expected value.
  - If the p.d.f. has large tails, the bigger contributions from values 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!

- 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!



- Not much!

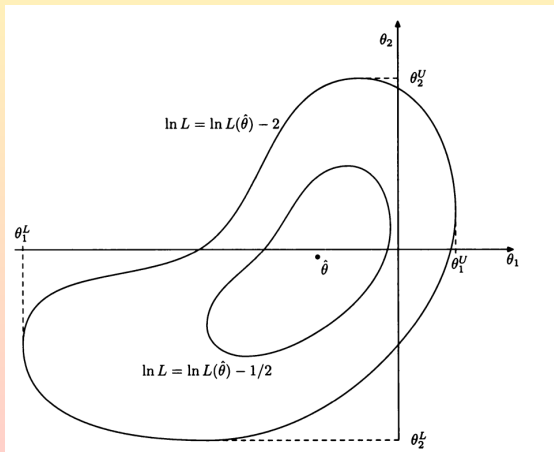
- Construct  $\log \mathcal{L}$  contours and determine confidence intervals by MINOS
- Elliptical contours correspond to gaussian Likelihoods
  - The closer to MLE, the more elliptical the contours, even in non-linear problems
  - All models are linear in a sufficiently small region
- Nonlinear regions 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*

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

Pleasant quality read: Vischia, 2019 doi:[10.1016/j.revip.2020.100046](https://doi.org/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  $\theta$

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

↓

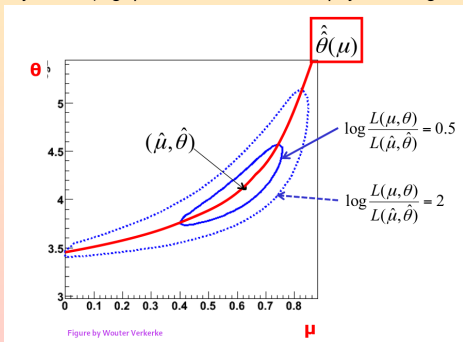
$$\mathcal{L}(\mathbf{n}, 0 | \mu, \alpha) = \prod_{i \in \text{bins}} \mathcal{P}(n_i | \mu S_i(\alpha) + B_i(\alpha)) \times \prod_{j \in \text{syst}} \mathcal{G}(0 | \alpha_j, 1)$$

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

- 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)



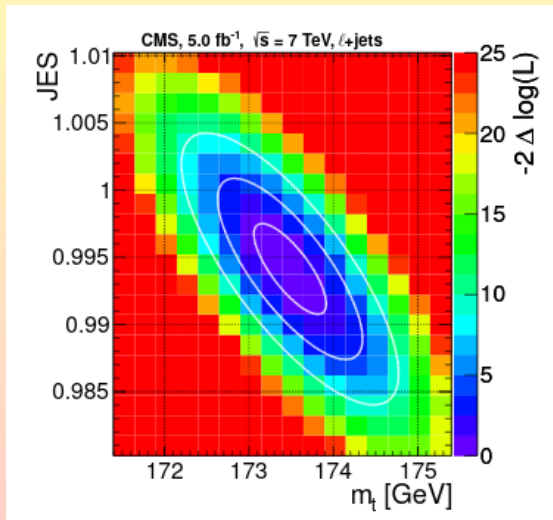
Pleasant quality read: [Vischia, 2019 doi:10.1016/j.revip.2020.100046](https://doi.org/10.1016/j.revip.2020.100046) ☺



- The likelihood ratio  $\lambda(\mu) = \frac{L(\mu, \hat{\theta}(\mu))}{L(\hat{\mu}, \hat{\theta})}$
- 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 you 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.

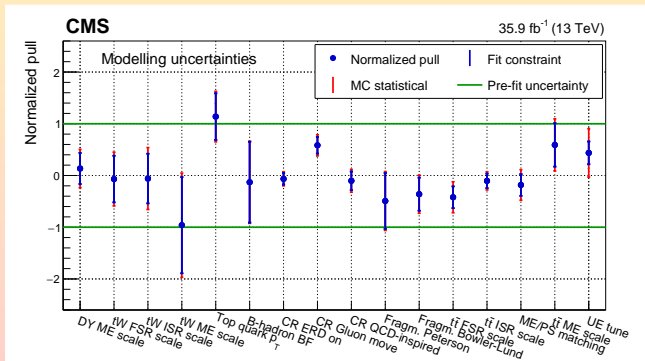


Image collected and cited in [doi:10.1016/j.revip.2020.100046](https://doi.org/10.1016/j.revip.2020.100046), references therein

- 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

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

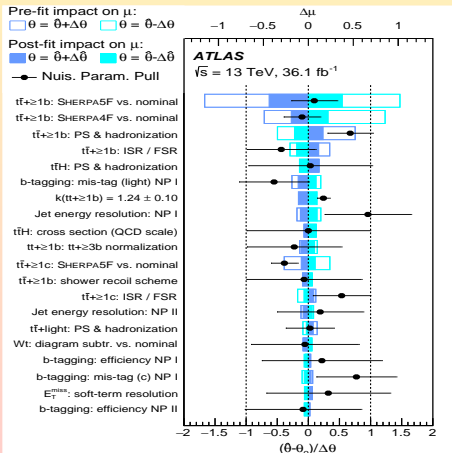
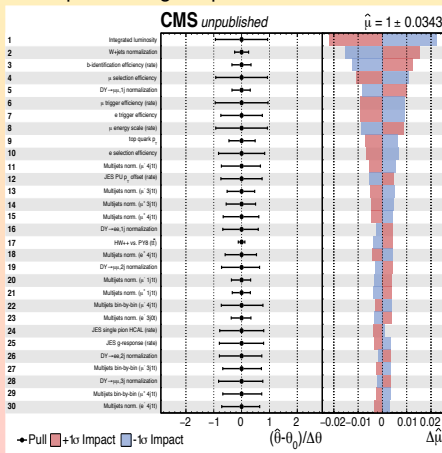
- We would conclude that the first case  $C = 0.74$ , for the second one  $C = 0.98$  (larger, still within uncertainty)
- However, these are not independent measurements!
- The formula is therefore

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

- 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!!!

## Impacts

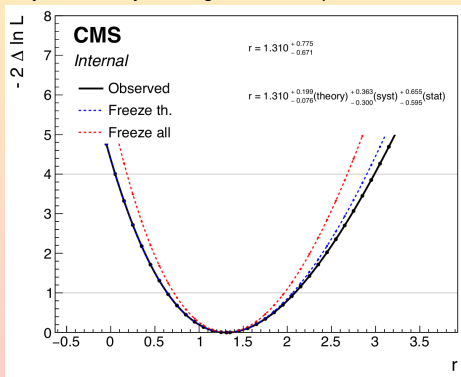
- Impact of  $\theta$  on the post-fit signal strength permits to obtain a ranking of the nuisance parameters in terms of their effect on the signal strength
  - Fix each nuisance parameter to its post-fit value  $\hat{\theta}$  plus/minus its pre-fit (post-fit) uncertainty  $\delta\theta$  ( $\delta\hat{\theta}$ )
  - 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

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

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

- Sideband measurement

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

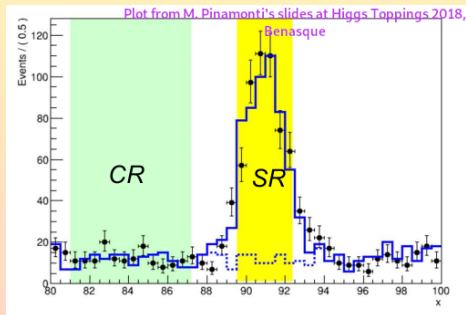
$$L_{CR}(b) = \text{Poisson}(N_{CR} | \tilde{\tau} \cdot b)$$

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

- 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

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

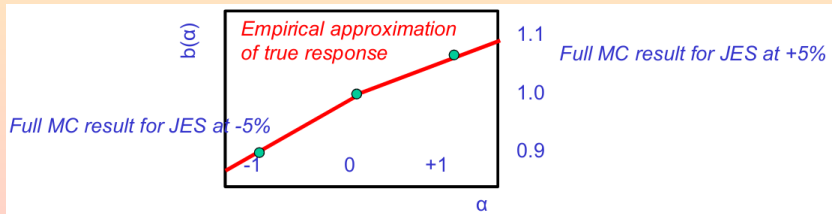


$$\mathcal{L}(\mathbf{n}, \boldsymbol{\alpha}^0 | \mu, \boldsymbol{\alpha}) = \prod_{i \in \text{bins}} \mathcal{P}(n_i | \mu S_i(\boldsymbol{\alpha}) + B_i(\boldsymbol{\alpha})) \times \prod_{j \in \text{syst}} \mathcal{G}(\alpha_j^0 | \alpha_j, \delta \alpha_j)$$



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

- Subsidiary measurement often labelled *constraint term*
- It is not a PDF in  $\alpha$ :  $\mathcal{G}(\alpha_j | 0, 1) \neq \mathcal{G}(0 | \alpha_j, 1)$
- Response function:  $\tilde{B}_i(1 + 0.1\alpha)$  (a unit change in  $\alpha$  –e.g. 5% JES– changes the acceptance by 10%)

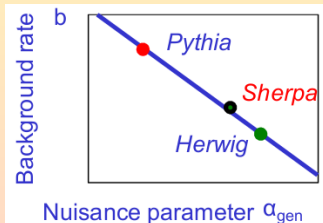


Graphics from W. Verkerke

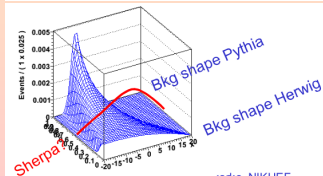
## Caveats on modelling theory uncertainties (P.V. at Benasque 2018)

- Cross section uncertainty: easy, assuming a gaussian for the constraint term<sup>\*\*\*</sup>  
 $\mathcal{L}_{full}(s, b) = \mathcal{P}(N_{SR}|s + b) \times \mathcal{G}(b|b, 0.08)$
- Factorization scale: what distribution  $\mathcal{F}$  is meant to model the constraint???  
 $\mathcal{L}_{full}(s, b) = \mathcal{P}(N_{SR}|s + b(\alpha_{FS}) \times \mathcal{F}(\alpha_{FS}|\alpha_{FS})$ 
  - “Easy” case, there is a single parameter  $\alpha_{FS}$ , clearly connected to the underlying physics model
- Hadronization/fragmentation model: run different generators, observing different results
  - Difficult! Not just one parameter, how do you model it in the likelihood?
  - 2-point systematics: you can evaluate two (three, four...) configurations, but underlying reason for difference unclear
  - Often define empirical response function

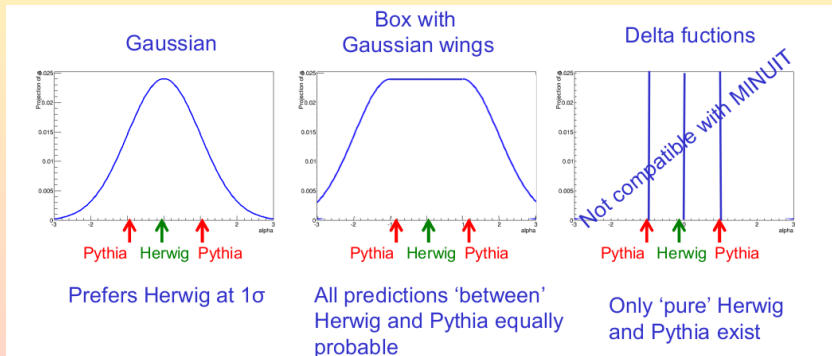
- Counting experiment: easy extend to other generators
- There must exist a value of  $\alpha$  corresponding to SHERPA



- Shape experiment: ouch!
- SHERPA is in general not obtainable as an interpolation of PYTHIA and HERWIG

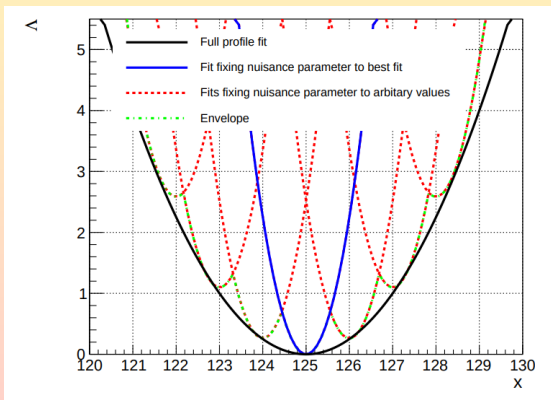


- Attempting to quantify our knowledge of the models
- There is no single parameter, difficult to model the differences within a single underlying model
- Which of these is the “correct” one?

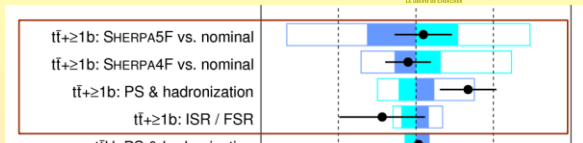


Graphics from W. Verkerke

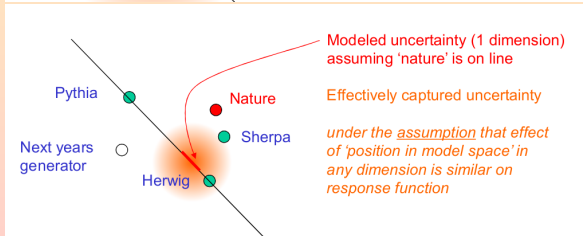
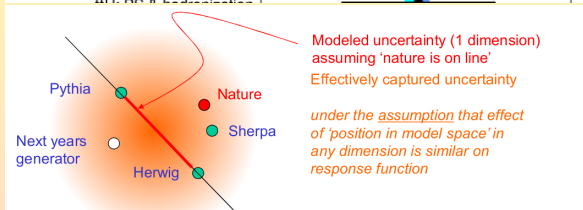
- Label each shape with an integer, and use the integer as nuisance parameter
- Can obtain the original log-likelihood as an envelope of different fixed discrete nuisance parameter values
- How do you define the various shapes?
  - Need many additional generators!
  - Interpolation unlikely to work (*SHERPA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



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



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



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

- 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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- Evidently, the time to moon and back is  $N_{EME} = N_{EM} + N_{ME}$ , and we can apply Eq. 123: **Do it!**

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

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

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

- Evidently, the time to moon and back is  $N_{EME} = N_{EM} + N_{ME}$ , and we can apply Eq. 123: **Do it!**
- Resulting estimate:

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

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

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

- We now however can argue that over the time it takes for light to go to the Moon and back any environment condition would be roughly constant
- How can we exploit this additional information? Question Time: Combining Estimates
- We can use this additional information to note that the two estimates  $N_{EM}$  and  $N_{ME}$  are independent estimates of the same physical quantity  $\frac{N_{EME}}{2}$
- Compute  $N_{EME}$  and  $\sigma(N_{EME})$  based on this reasoning
- We can therefore use Eq. 121 to compute  $\frac{N_{EME}}{2}$  and multiply the result by 2, obtaining

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

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



## What about asymmetric uncertainties?

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

## What about asymmetric uncertainties?

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

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

- Statistics is a tool to answer questions (but you must pose questions in a well-defined way)
- Mathematical definition of probability based on set theory and on the theory of Lebesgue measure
  - Frequentist and Bayesian statistics
  - Conditioning, marginalization
  - Expected values, variance
- Random variables and probability distributions
  - Correlation vs causality
- Information and likelihood principle
  - Sufficiency, ancillarity, pivoting
- Estimators
  - Point estimates with the Maximum Likelihood Estimator (MLE)
  - Interval estimates with the MLE
  - The profile likelihood ratio and modelling of systematic uncertainties

- 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