

Statistics

or “How to find answers to your questions”

Pietro Vischia¹

¹CP3 — IRMP, Université catholique de Louvain



LIP-Lisboa, Statistics Lectures (March 16th and 18th, 2020), Course on Physics at the LHC
2020

Why statistics?

Fundamentals

Games, weather

Random variables and distributions

Random variables and their properties

Distributions

Estimating a physical quantity

Likelihood Principle

Estimators and maximum likelihood

Profile likelihood ratio



- Schedule: two lessons
 - Monday 16.03, 17h (this lesson)
 - Wednesday 18.03, 17h (unless you prefer e.g. Tuesday)
- The slides contain links to a few exercises and examples
 - In a longer course there is time to go through them, not in two lessons
 - You are encouraged to play with the exercises offline
- Many interesting references
 - Papers mostly in each slide
 - Some cool books after the summary slide of the second lesson
- Unless stated otherwise, figures belong to P. Vischia, *****
(textbook to be published by Springer in 2021)
- Your feedback is crucial for improving these lectures!

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?

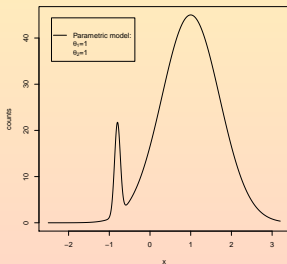
- What is the chance of obtaining a 1 when throwing a six-faced die?
 - We can throw a dice 100 times, and count how many times we obtain 1
- What is the chance of tomorrow being rainy?
 - We can try to give an answer based on the recent past weather, but we cannot – in general – *repeat tomorrow* and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



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

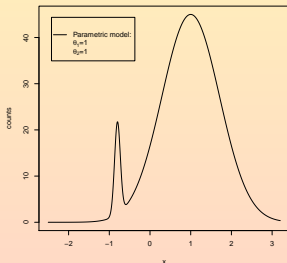
• Theory

- Approximations
- Free parameters



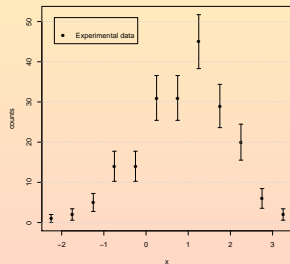
• Theory

- Approximations
- Free parameters



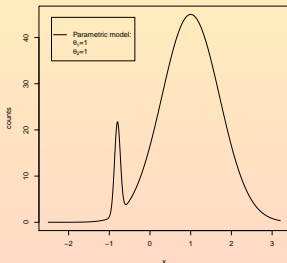
• Experiment

- Random fluctuations
- Mismeasurements (detector effects, etc)



• Theory

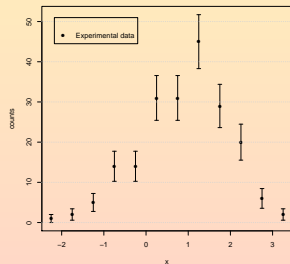
- Approximations
- Free parameters



• Statistics!

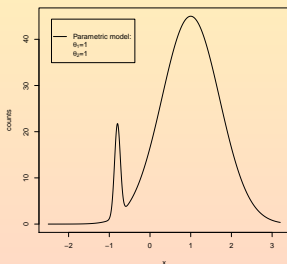
• Experiment

- Random fluctuations
- Mismeasurements (detector effects, etc)



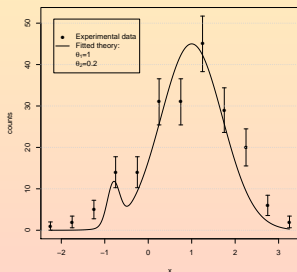
• Theory

- Approximations
- Free parameters



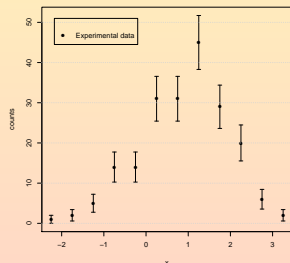
• Statistics!

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



• Experiment

- Random fluctuations
- Mismeasurements (detector effects, etc)



Fundamentals

- Ω : 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 theorem (1946): formalize a set of axioms starting from reasonable premises¹
 - $c * b|a = F(c|b * a, b|a)$
 - $\sim b|a = S(b|a)$, 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

¹ $a|b$ = the occurrence of event a conditioned on the occurrence of event b

- 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?

- In \mathbb{R}^1 , an interval $[a, b]$ has length:

$$\begin{aligned}L(i) &= b - a \\L(a, a) &= 0 \\L(\infty) &= \infty.\end{aligned}$$

- Disjoint intervals (no common point with any other)

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

- Define the sum as $L(i) := L(i_1) + \dots + L(i_n)$
 - Extendable to an enumerable sequence of intervals (crucial for defining continuous density functions)
- **Borel lemma:** we consider a finite closed interval $[a, b]$ and a set of Z intervals such that every point of $[a, b]$ is an inner point of at least one interval belonging to Z .
 - Then there is a subset Z' of Z containing only a finite number of intervals, such that every point of $[a, b]$ is an inner point of at least one interval belonging to Z' .
- Generalizable to N dimensions, with $L(i)$ additive function of i : $i = \sum i_n \Rightarrow L(i) = \sum L(i_n)$

- $L(i)$ is a non-negative additive function (finite- or infinite-valued): a measure
- Definition extendable from intervals to complex sets:
 - $L(S) \geq 0$
 - If $S = S_1 + \dots + S_n$, where $S_\mu S_\nu = 0$ for $\mu \neq \nu$ then $L(S) = L(S_1) + \dots + L(S_n)$
 - If S is an interval i , then the set function $L(S)$ reduces itself to the interval function $L(i)$, $L(S) = L(i)$
- True only for Borel sets
 - In layman's terms, sets that can be constructed by taking countable unions or intersections (and their respective complements) of open sets
- $L(S)$ is a measure and it's called Lebesgue measure
 - The extension from $L(i)$ to $L(S)$ is unique (the only set function defined on the whole \mathcal{B}_1 satisfying the properties above)
 - Extension to \mathbb{R}^n is immediate: $L_n(S)$

- Generalization of $L_n(S)$: 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 $[0, 1]$)

- Associate to any $P(S)$ a point function $F(\mathbf{x}) = F(x_1, \dots, x_n)$

$$F(\mathbf{x}) = F(x_1, \dots, x : n) := P(\xi_1 \leq x_1, \dots, \xi_n \leq x_n).$$

- Trivial in one dimension. $P(S)$ must have an upper bound!
- Map $F(a) = F(b)$ to set of null P-measure, $P(a < x \leq b) = 0$
- $F(\mathbf{x})$ is in each point a non-decreasing function everywhere-continuous to the right

$$P(a < x \leq a + h) = \Delta F(a) = F(a + h) - F(a),$$

- 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

- What about individual points?

- 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

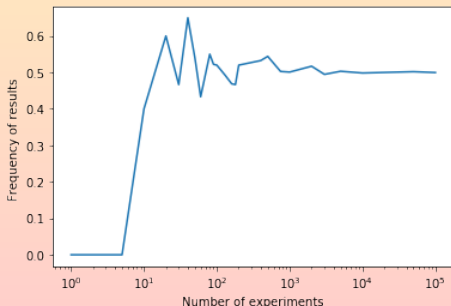
$$\frac{P(i)}{L(i)} = \frac{\Delta_n F}{2^n h_1 h_2 \dots h_n}.$$

- 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_ν
 - 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_ν .
 - It can be obtained by projection starting from the n -dimensional distribution
 - Shift each “mass particle” along the perpendicular direction to x_ν until collapsing into the x_ν axis
 - This results in a one-dimensional distribution which is the marginal distribution of x_ν .
 - There are infinite ways of arriving to the same x_ν starting from a generic n -dimensional distribution function
- Marginal distributions can be also built with respect to subsets of variables.

Random experiment

- Repeat a random experiment ξ (e.g. toss of a die) many times under uniform conditions
 - As uniform as possible
 - \vec{S} : set of all a priori possible different results of an individual measurement
 - S : a fixes subset of \vec{S}
- If in an experiment we obtain $\xi \in S$, we will say the event defined by $\xi \in S$ has occurred
 - We assume that S is simple enough that we can tell whether ξ 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 ν times
 - Then $\frac{\nu}{n}$ is the frequency ratio of the event in the sequence of n experiments
- **EXERCISE:** For a fixed event, how does the frequency ratio behave for increasing n ?
 wget <https://raw.githubusercontent.com/vischia/statex/master/frequencyRatio.ipynb>











- 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:

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

Frequentist probability - 2

- 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	Combinatoric probability	Odds	Mathematical expression of absolute Frequency
Royal flush 	1	4	0.000154%	0.000154%	649,739 : 1	$\binom{4}{1}$
Straight flush (including royal flush)	3	36	0.06139%	0.0012%	72,192 : 1	$\binom{10}{1}\binom{4}{1} - \binom{4}{1}$
Four of a kind 	156	624	0.0240%	0.0156%	4,164 : 1	$\binom{13}{1}\binom{12}{1}\binom{4}{1}$
Full house 	156	3,744	0.1441%	0.17%	699 : 1	$\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}$
Flush (including royal flush and straight flush)	1,277	5,108	0.1965%	0.267%	508 : 1	$\binom{13}{5}\binom{4}{1} - \binom{10}{1}\binom{4}{1}$
Straight (including royal flush and straight flush)	10	16,200	0.3925%	0.76%	264 : 1	$\binom{10}{1}\binom{4}{1}^5 - \binom{10}{1}\binom{4}{1}$
Three of a kind 	858	54,912	2.1128%	2.87%	46.2 : 1	$\binom{13}{1}\binom{4}{3}\binom{12}{2}\binom{4}{1}^2$
Two pair 	858	128,852	4.7639%	7.62%	29.8 : 1	$\binom{13}{2}\binom{4}{2}\binom{11}{1}\binom{4}{1}$
One pair 	2,860	1,098,240	42.2569%	49.5%	1.97 : 1	$\binom{13}{1}\binom{4}{2}\binom{12}{3}\binom{4}{1}^3$
Big pair / High card 	1,277	1,382,540	60.2177%	100%	0.996 : 1	$\left[\binom{13}{5} - 10\right] \left[\binom{4}{1}^3 - 4\right]$
Five 	7,462	2,598,960	100%	—	0 : 1	$\binom{52}{5}$

Subjective (Bayesian) probability

- 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

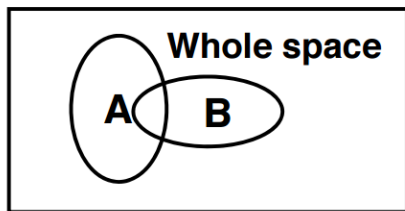
$$P(X) := \frac{\text{The largest amount you are willing to bet}}{\text{The amount you stand to win}} \quad (1)$$

- 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
 - 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

Conditional probabilities: Bayes theorem

- Probabilities can be combined to obtain more complex expressions



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

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

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

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

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

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

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

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

A word of advice about conditional probabilities

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

- **Conditional probabilities are not commutative!** $P(A|B) \neq P(B|A)$
- Example:
 - A : speaking English
 - B : having a TOEFL certificate
- The probability for an English speaker to have a TOEFL certificate, $P(\text{have TOEFL}|\text{speaking English})$, is very small (say $\sim 1\%$ very roughly)
- The probability for a TOEFL certificate holder to speak English, $P(\text{speaking English}|\text{have TOEFL})$, is unarguably $\ggggg 3\%$ ☺

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

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

A trickier example of conditional probability: the Monty Hall problem

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

A trickier example of conditional probability: the Monty Hall problem

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

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

- Bayes Theorem (1763):

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

- 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. $\cap A_i A_j = 0 \forall i, j$), $\cup_i A_i = S$ 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:

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

- 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?

A Diagnosis problem

- 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?**
- By the Bayes Theorem:

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

A Diagnosis problem

- 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?**
- 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, $P(D)$! 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 $P(D|+) = 0.4977$, which is pretty high (and substantially higher than the previous one)

- 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” **fix**
 - 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:

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

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

- \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 θ 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. the ratio of vacuum expectation values v in a quantum field theory, $\beta := \frac{v_1}{v_2}$
- Before any experiment, we are Jon Snow about the parameter β : we know nothing
 - We have to choose a very broad prior, or better uniform, in β
- Now we interact with a theoretical physicist, who might have built her theory by using as a parameter of the model the tangent of the ratio, $\tan\beta$
 - In a natural way, she will express her pre-experiment ignorance using an uniform prior **in** $\tan\beta$.
 - This prior is not constant in β !!!
 - In general, there is no uniquely-defined prior expressing complete ignorance or ambivalence in both parameters (β and $\tan\beta$)
- 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”)

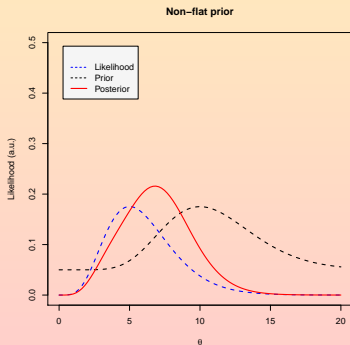
- Two ways of solving the situation
 - Objective Bayes: use a formal rule dictated by an invariance principle
 - Subjective Bayes: use something like elicitation of expert opinion
 - Ask an expert her opinion about each value of θ , and express the answer as a curve
 - Repeat this with many experts
 - 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 θ
 - Find a value θ_c such that $P(\theta_{true} < \theta_c) = 95\%$
- Typically θ 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 θ
- 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

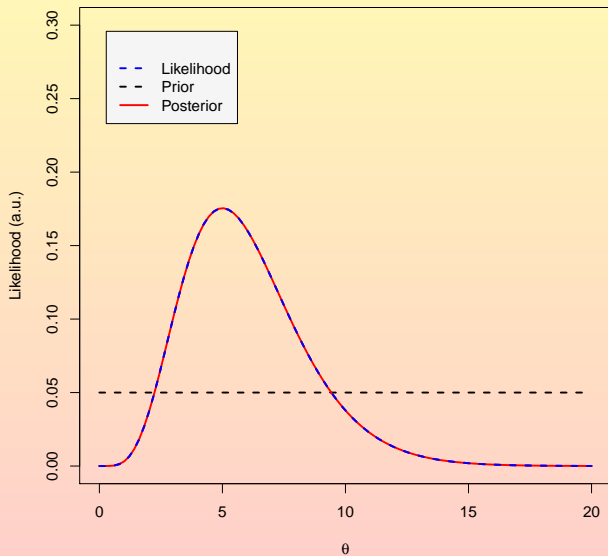
- Uniform priors must make sense
 - Uniform prior across its entire dominion: not very realistic
 - It corresponds to claiming that $P(1 < \theta \leq 2)$ is the same as $P(10^{41} < \theta \leq 10^{41} + 1)$
 - It's irrational to claim that a prior can cover uniformly forty orders of magnitude
 - We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

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

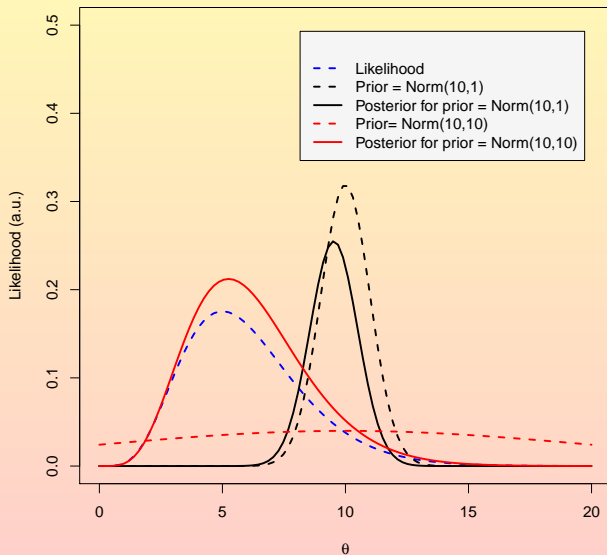
- Associating parametric priors to intervals in the parameter space corresponds to considering sets of theories
 - This is because to each value of a parameter corresponds a different theory
- In practical situations, note (Eq. 9) posterior probability is always proportional to the product of the prior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



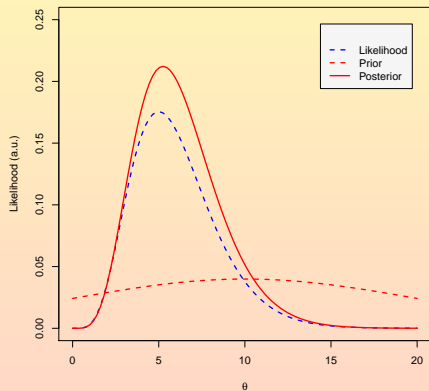
Flat prior



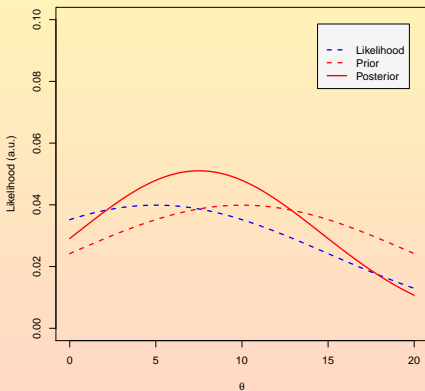
Broad prior vs narrow prior



Broad prior vs narrow prior



Broad prior vs narrow prior



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

Drawing some histograms

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

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

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

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

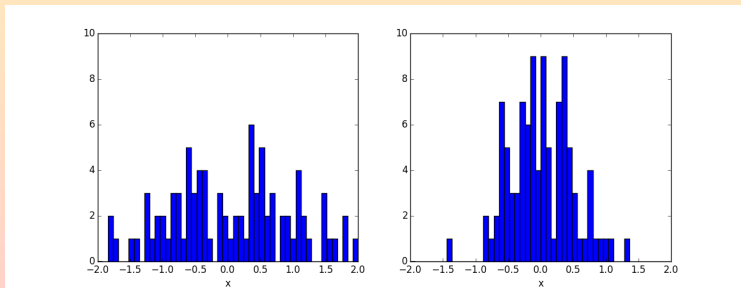
- Extend the concept of p.d.f. to an arbitrary number of variables; the joint p.d.f. $f(X, Y, \dots)$
- If we are interested in the p.d.f. of just one of the variables the joint p.d.f. depends upon, we can compute by integration the marginal p.d.f.

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

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

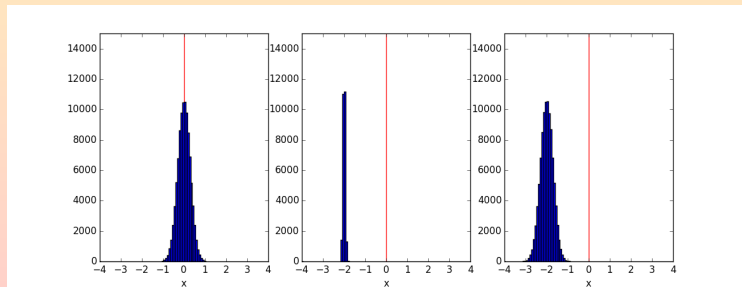
$$f(X|Y) := \frac{f(X, Y)}{f_Y(Y)} \quad (13)$$

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



Sources of uncertainty (errors?)

- 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:

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

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

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

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

- It is also called "error matrix", and sometimes denoted $cov[X, Y]$
- It is symmetric by construction: $V_{XY} = V_{YX}$, and $V_{XX} = \sigma_X^2$
- To have a dimensionless parameter: correlation coefficient $\rho_{XY} = \frac{V_{XY}}{\sigma_X\sigma_Y}$

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

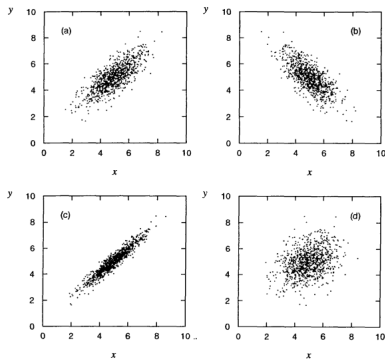


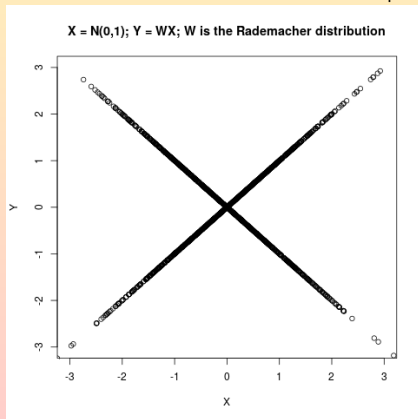
Fig. 1.9 Scatter plots of random variables x and y with (a) a positive correlation, $\rho = 0.75$, (b) a negative correlation, $\rho = -0.75$, (c) $\rho = 0.95$, and (d) $\rho = 0.25$. For all four cases the standard deviations of x and y are $\sigma_x = \sigma_y = 1$.

Take it to the next level: the Mutual Information

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

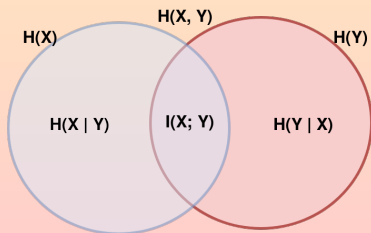
$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p_1(x)p_2(y)} \right)$$

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

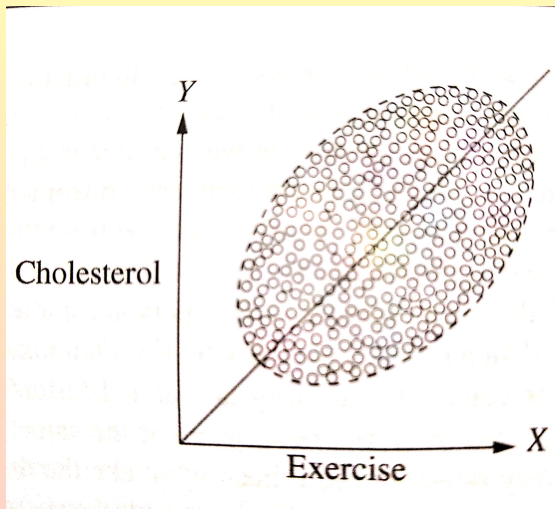


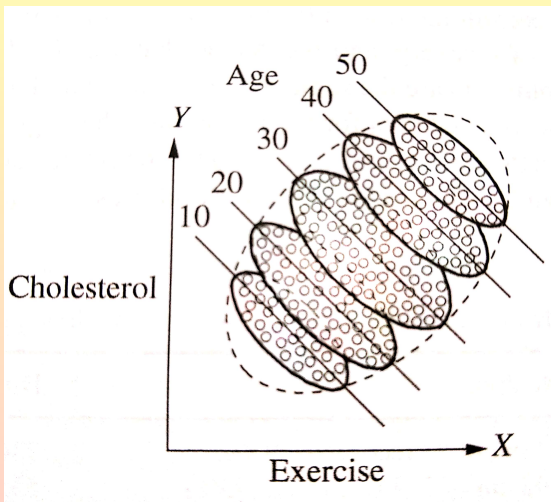
- Related to entropy

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



Does cholesterol increase with exercise?





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

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

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

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

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

- 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%)

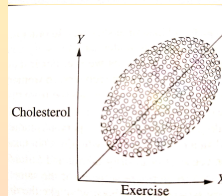
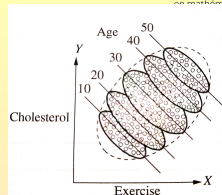
- BP = Blood Pressure

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

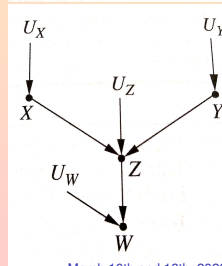
- Same table, different labels; here we must consider the combined data
 - Lowering blood pressure is actually part of the mechanism of the drug effect

The Simpson paradox: correlation is not causation

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

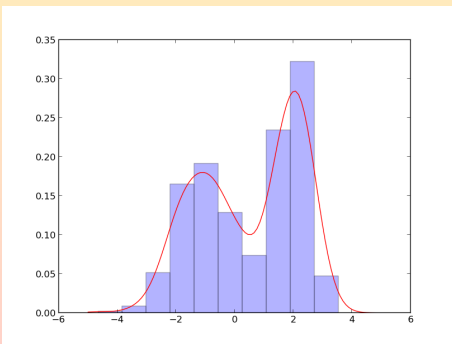
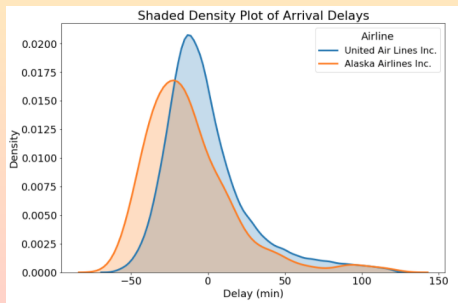


	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 250 recovered (83%)
	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP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



Distributions... or not?

- 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



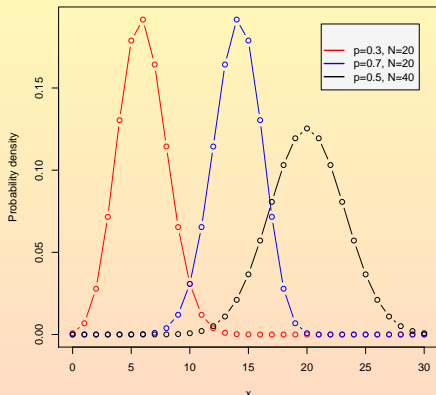
Plots from TheGlowingPython and TowardsDataScience

Binomial

- Discrete variable: r , positive integer $\leq N$
- Parameters:
 - N , positive integer
 - p , $0 \leq p \leq 1$
- Probability function:

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

Binomial p.d.f.



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

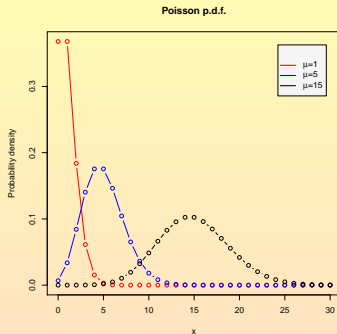
- $N = 12$, $r = 3$, $p = \frac{1}{6}$

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

• Poisson

- Discrete variable: r , positive integer
- Parameter: μ , 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

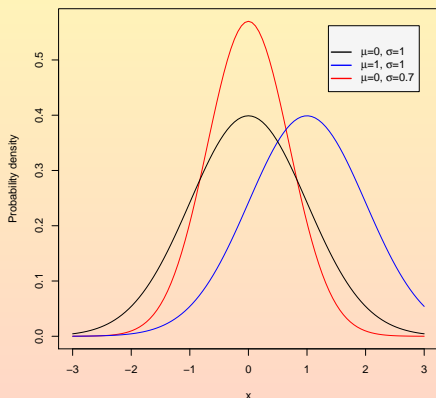
- Variable: X , real number
- Parameters:
 - μ , real number
 - σ , positive real number

- 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

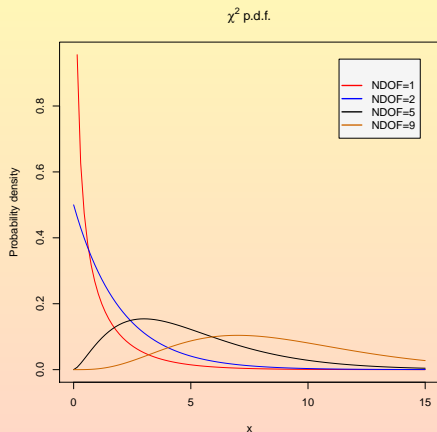
$$f(X) = \frac{1}{2} \left(\frac{X}{2}\right)^{\frac{N}{2}-1} e^{-\frac{X}{2}} \frac{1}{\Gamma(\frac{N}{2})}$$

$$E[r] = N$$

$$V(r) = 2N$$

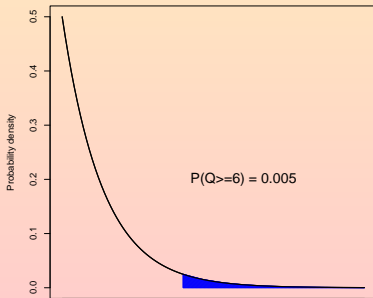
- It describes the distribution of the sum of the squares of a random variable, $\sum_{i=1}^N X_i^2$

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



The χ^2 distribution: why degrees of freedom?

- Sample randomly from a Gaussian p.d.f., obtaining X_1 y X_2
- $Q = X_1^2 + X_2^2$ (or in general $Q = \sum_{i=1}^N X_i^2$) is itself a random variable
 - What is $P(Q \geq 6)$? Just integrate the $\chi^2(N = 2)$ distribution from 6 to ∞
- Depends only on $N!$
 - If we sample 12 times from a Gaussian and compute $Q = \sum_{i=1}^{12} X_i^2$, then $Q \sim \chi^2(N = 12)$
- Theorem: if Z_1, \dots, Z_N is a sequence of normal random variables, the sum $V = \sum_{i=1}^N Z_i^2$ is distributed as a $\chi^2(N)$
 - The sum of squares is closely linked to the variance $E[(X - \mu)^2] = E[X^2] - \mu^2$ from Eq. 18
- The χ^2 distribution is useful for goodness-of-fit tests that check how much two distributions diverge point-by-point
- It is also the large-sample limit of many distributions (useful to simplify them to a single parameter)



The χ^2 distribution: goodness-of-fit tests 1/

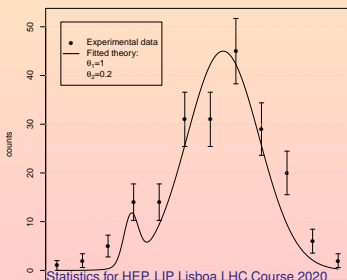
- Consider a set of M measurements $\{(X_i, Y_i)\}$
 - Suppose Y_i are affected by a random error representable by a gaussian with variance σ_i
- Consider a function $g(X)$ with predictive capacity, i.e. such that for each i we have $g(X_i) \sim Y_i$
- Pearson's χ^2 function related to the difference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

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

- Neyman's χ^2 is a similar expression under some assumptions

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

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



The χ^2 distribution: goodness-of-fit tests 2/

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

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

- The cumulative of f is

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

- Comparing χ^2 with the number of degrees of freedom M , we therefore have a criterion to test for goodness-of-fit
 - For a given M , the p.d.f. is known ($\chi^2(M)$) and the observed value can be computed and compared with it
 - Null hypothesis: there is no difference between prediction and observation (i.e. g fits well the data)
 - Alternative hypothesis: there is a significant difference between prediction and observation
 - Under the null, the sum of squares is distributed as a $\chi^2(M)$
 - p-values can be calculated by integration of the χ^2 distribution

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

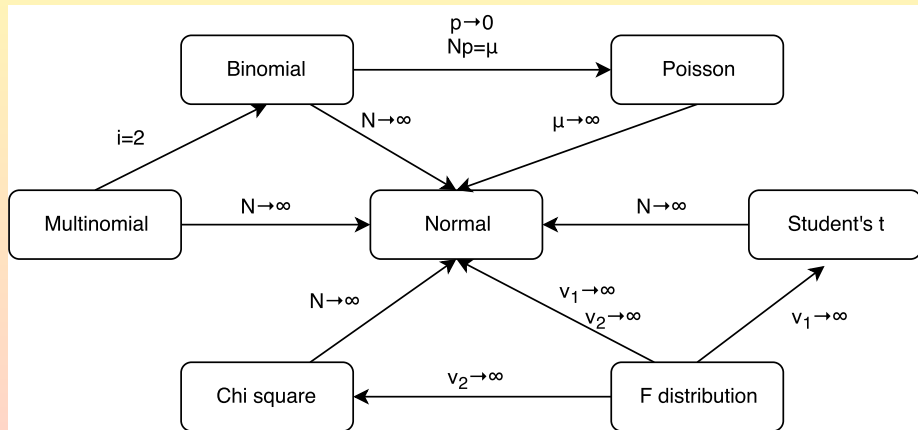
$$\frac{\chi^2}{M} \gg 1 \Rightarrow \text{poor model (increases } \chi^2), \text{ or statistically improbable fluctuation} \quad (23)$$

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

The χ^2 distribution: goodness-of-fit tests 3/

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

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



Estimating a physical quantity

- 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 θ contained in the data sample
 - The likelihood function is seen as a function of θ , 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 it
 - $P(\theta|\vec{x}) \propto L(\vec{x}; \theta) \times \pi(\theta)$: the only quantity depending on the data is the likelihood
 - *Information* as a broad way of saying *all the possible inferences about θ*
 - “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
 - 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 θ you seek to determine
 - Sampling probability: $p(DZ|\theta) = p(D|\theta)p(Z)$
 - The coin toss tells us nothing about the parameter θ 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

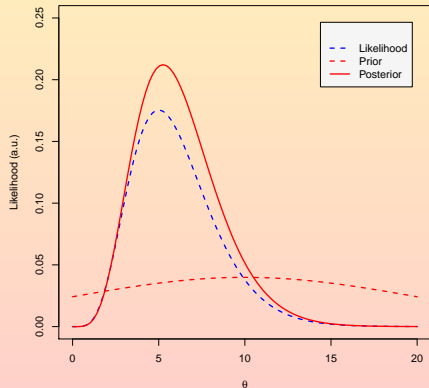
- The narrowness of the likelihood can be estimated by looking at its curvature
- The curvature is the second derivative with respect to the parameter of interest
- A very narrow (peaked) likelihood is characterized by a very large and positive $-\frac{\partial^2 \ln L}{\partial \theta^2}$
- The second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

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

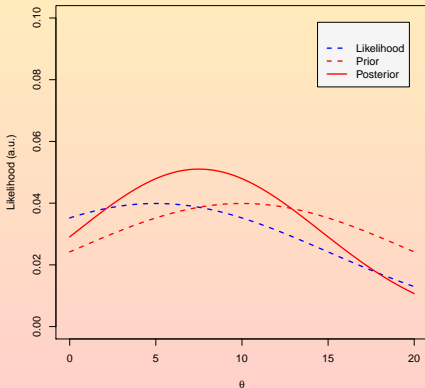
Likelihood and Fisher Information

- A very narrow likelihood will provide much information about θ_{true}
 - The posterior probability will be more localized than the prior in the regimen in which the likelihood function dominates the product $L(\vec{x}; \vec{\theta}) \times \pi$
 - The Fisher Information will be large
- A very broad likelihood will not carry much information, and in fact the computed Fisher Information will turn out to be small

Broad prior vs narrow prior



Broad prior vs narrow prior



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 θ , $\pi(\theta) = \sqrt{I(\theta)}$; 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 $\propto \frac{1}{\sqrt{\theta}}$

- A test statistic is a function of the data (a quantity derived from the data sample)
- A statistic $T = T(X)$ is sufficient for θ if the density function $f(X|T)$ is independent of θ
 - If T is a sufficient statistic for θ , then also any strictly monotonic $g(T)$ is sufficient for θ
- The statistic T carries as much information about θ as the original data X
 - No other function can give any further information about θ
 - Same inference from data X with model M and from sufficient statistic $T(X)$ with model M'

- Example: data 1, 2, 3, 4, 5; sample mean (estimate of population mean) $\hat{x} = \frac{1+2+3+4+5}{5} = 3$
 - Imagine we don't have the data; we only know that the sample mean is 3
 - **Is the sample mean a sufficient statistic?**

Example: is it sufficient?

- Example: data 1, 2, 3, 4, 5; sample mean (estimate of population mean) $\hat{x} = \frac{1+2+3+4+5}{5} = 3$
 - Imagine we don't have the data; we only know that the sample mean is 3
 - **Is the sample mean a sufficient statistic?**
 - Since the sample mean is 3, we also estimate the population mean to be 3
 - 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

Example: is it sufficient?

- Example: data 1, 2, 3, 4, 5; sample mean (estimate of population mean) $\hat{x} = \frac{1+2+3+4+5}{5} = 3$
 - Imagine we don't have the data; we only know that the sample mean is 3
 - **Is the sample mean a sufficient statistic?**
 - Since the sample mean is 3, we also estimate the population mean to be 3
 - 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
- Binomial test in coin toss
 - Record heads and tails, with their order: *HTTHHHHTHTTTHTHTH*
 - **Can we somehow improve by identifying a sufficient statistic?**

Example: is it sufficient?

- Example: data 1, 2, 3, 4, 5; sample mean (estimate of population mean) $\hat{x} = \frac{1+2+3+4+5}{5} = 3$
 - Imagine we don't have the data; we only know that the sample mean is 3
 - **Is the sample mean a sufficient statistic?**
 - Since the sample mean is 3, we also estimate the population mean to be 3
 - 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
- Binomial test in coin toss
 - Record heads and tails, with their order: *HTTHHHHTHHTTTHTHTH*
 - **Can we somehow improve by identifying a sufficient statistic?**
 - **What happens if we record only the number of heads? (remember that the binomial p.d.f. is:**

$$P(r) = \binom{N}{r} p^r (1-p)^{N-r}, r = 0, 1, \dots, N$$

Example: is it sufficient?

- Example: data 1, 2, 3, 4, 5; sample mean (estimate of population mean) $\hat{x} = \frac{1+2+3+4+5}{5} = 3$
 - Imagine we don't have the data; we only know that the sample mean is 3
 - **Is the sample mean a sufficient statistic?**
 - Since the sample mean is 3, we also estimate the population mean to be 3
 - 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
- Binomial test in coin toss
 - Record heads and tails, with their order: *HTTHHHHTHHTTTHTH*
 - **Can we somehow improve by identifying a 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
 - Storage needs are reduced

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

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



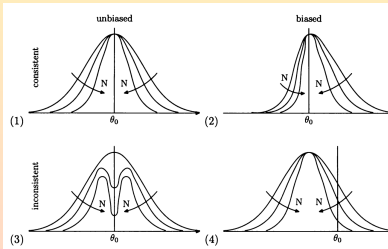
- Ancillary statistic for a parameter θ : a statistic $f(X)$ which does not depend on θ
 - Concept linked to that of (*minimal*) *sufficient statistic*; (maximal) data reduction while retaining all Fisher information about θ
- An ancillary statistic can give information about θ even if it does not depend on it!
 - 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 θ)
 - 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 θ , but increases the precision on an estimate of θ (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 (evolution of my proceedings in preparation as paper)

Estimators

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

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

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



Plot from James, 2nd ed.

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

Nuclear Decay with Maximum Likelihood Method

- A nuclear decay with half-life τ 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.**

Nuclear Decay with Maximum Likelihood Method

- A nuclear decay with half-life τ 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

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

Nuclear Decay with Maximum Likelihood Method

- A nuclear decay with half-life τ 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 τ only,
 $L(\tau) := f(t_i; \tau)$

Nuclear Decay with Maximum Likelihood Method

- A nuclear decay with half-life τ 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 τ only,
 $L(\tau) := f(t_i; \tau)$
- **Now all you need to do is to maximize the likelihood**

Nuclear Decay with Maximum Likelihood Method

- A nuclear decay with half-life τ 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 τ 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

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

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?**

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?**
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

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

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?**
- 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?**

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?**
- 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?**
- The variance interestingly decreases when N increases, and it is possible to demonstrate that the estimator is efficient

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

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?
- 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?
- 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!

	Consistente	Insegado	Eficiente
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1 + \dots + t_N}{N}$			
$\hat{\tau} = \frac{t_1 + \dots + t_N}{N-1}$			
$\hat{\tau} = t_i$			

Table: Propiedades de diferentes estimadores de la vida media de un decaimiento nuclear.

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?
- 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?
- 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!

	Consistente	Insegado	Eficiente
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1 + \dots + t_N}{N}$	✓	✓	✓
$\hat{\tau} = \frac{t_1 + \dots + t_N}{N-1}$			
$\hat{\tau} = t_i$			

Table: Propiedades de diferentes estimadores de la vida media de un decaimiento nuclear.

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?
- 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?
- 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!

	Consistente	Insegado	Eficiente
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1 + \dots + t_N}{N}$	✓	✓	✓
$\hat{\tau} = \frac{t_1 + \dots + t_N}{N-1}$	✓	✗	✗
$\hat{\tau} = t_i$			

Table: Propiedades de diferentes estimadores de la vida media de un decaimiento nuclear.

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?
- 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?
- 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!

	Consistente	Insegado	Eficiente
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1 + \dots + t_N}{N}$	✓	✓	✓
$\hat{\tau} = \frac{t_1 + \dots + t_N}{N-1}$	✓	✗	✗
$\hat{\tau} = t_i$	✗	✓	✗

Table: Propiedades de diferentes estimadores de la vida media de un decaimiento nuclear.

Why $\hat{\tau} = t_i$ is unbiased

- 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!

	Consistente	Inssegado	Eficiente
$\hat{\tau} = t_i$	✗	✓	✗

Table: Propiedades de diferentes estimadores de la vida media de un decaimiento nuclear.

- 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, this is linked with the 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!

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

- 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 $\vec{\theta}$)

Estimating variance non-analytically

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

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

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

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

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

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

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

How to extract an interval from the likelihood function 2/

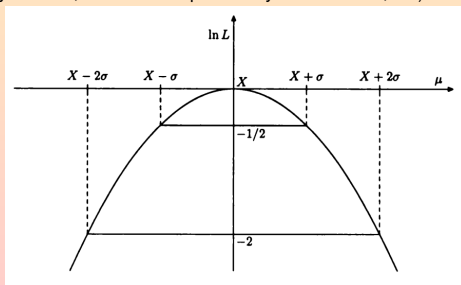
- We can therefore write

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

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

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

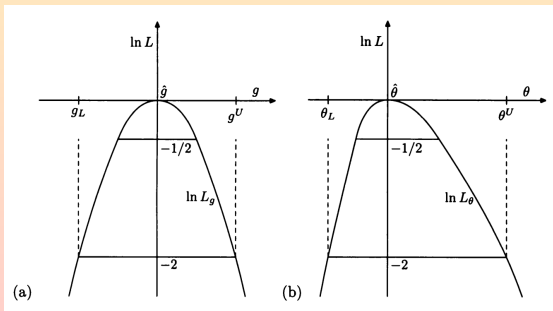
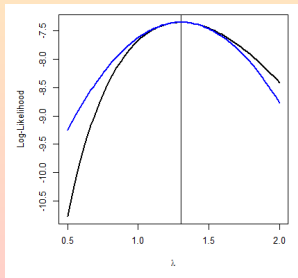
- Taking into account that it is important to keep in mind that probability is a property of sets, in frequentist statistics
 - Confidence interval: interval with a fixed probability content
- This process for computing a confidence interval is exact for a Gaussian p.d.f.
 - Pathological cases reviewed 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— 2σ , etc)



Plot from James, 2nd ed.

How to extract an interval from the likelihood function 3/

- MLE is invariant for monotonic transformations of θ
 - 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.

And in many dimensions...

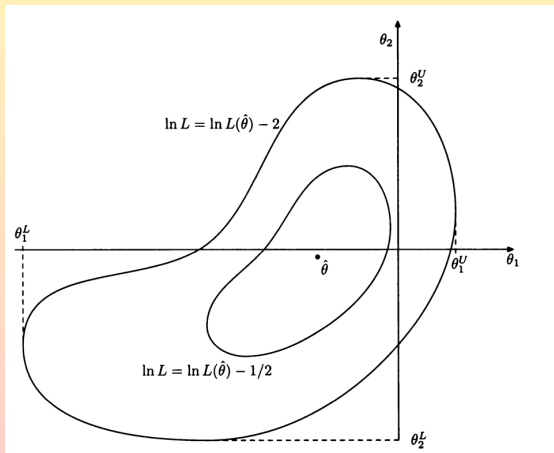
- Construct $\log \mathcal{L}$ contours and determine confidence intervals by MINOS
- Elliptical contours correspond to gaussian Likelihoods
 - The closer to MLE, the more elliptical the contours, even in non-linear problems
 - All models are linear in a sufficiently small region
- Nonlinear regions 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 σ ")
- $\lambda = 2$ for $\beta = 0.955$ ("2 σ ")



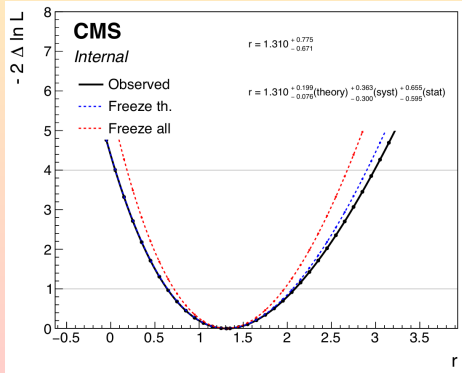
Plot from James, 2nd ed.

What if I have systematic uncertainties? /1

- Parametrize them into the likelihood function; conventional separation of parameters in two classes
 - the Parameter(s) of Interest (POI), often representing σ/σ_{SM} and denoted as μ (*signal strength*)
 - the parameters representing uncertainties, *nuisance parameters* θ
- $H_0: \mu = 0$ (Standard Model only, no Higgs)
- $H_1: \mu = 1$ (Standard Model + Standard Model Higgs)
- Find the maximum likelihood estimates (MLEs) $\hat{\mu}, \hat{\theta}$
- Find the conditional MLE $\hat{\hat{\theta}}(\mu)$, i.e. the value of θ maximizing the likelihood function for each fixed value of μ
- Write the test statistics as $\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})}$
 - Independent on the nuisance parameters (profiled, i.e. their MLE has been taken as a function of each value of μ)
 - Can even “freeze” them one by one to extract their contribution to the total uncertainty
- To model the nuisance parameters you can reparameterize them as $\alpha(\theta)$ introducing an explicit “p.d.f.” for them $\mathcal{L}(\mathbf{n}, \alpha^0 | \mu, \alpha) = \prod_{i \in bins} \mathcal{P}(n_i | \mu S_i(\alpha) + B_i(\alpha)) \times \prod_{j \in syst} \mathcal{G}(\alpha_j^0 | \alpha_j, \delta \alpha_j)$
 - The likelihood ratio is then $\lambda(\mu) = \frac{\mathcal{L}(\mu, \hat{\hat{\alpha}}(\mu))}{\mathcal{L}(\hat{\mu}, \hat{\alpha})}$

What if I have systematic uncertainties? /2

- The likelihood ratio $\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})}$
- Conceptually, you can run the experiment many times (e.g. toys) and record the value of the test statistic
- The test statistic can therefore be seen as a distribution
- Asymptotically, $\lambda(\mu) \sim \exp\left[-\frac{1}{2}\chi^2\right] \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right)$ (Wilks Theorem, under some regularity conditions—continuity of the likelihood and up to 2nd derivatives, existence of a maximum, etc)
 - The χ^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!



Toy data

How to extract an interval from the likelihood function

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

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

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

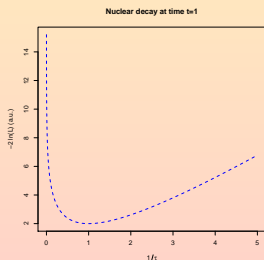
$$\Delta \ln L := \ln L(\theta') - \ln L_{max} = -\frac{1}{2}$$

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

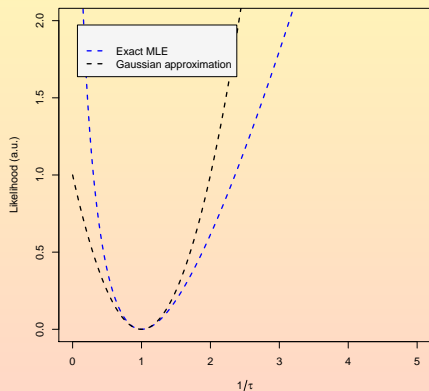
- $\vec{\theta}_{true}$ is therefore estimated as $\hat{\theta} = \vec{\theta}_{ML} \pm \sigma$. This is another situation in which frequentist and Bayesian statistics differ in the interpretation of the numerical result
- Frequentist: $\vec{\theta}_{true}$ is fixed
 - “if I repeat the experiment many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

Non-normal likelihoods and Gaussian approximation — 1

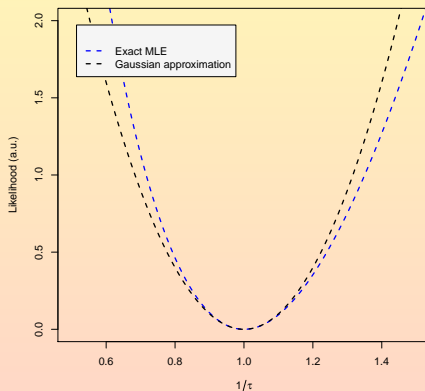
- How good is the approximation $L(\vec{x}; \vec{\theta}) \propto \exp\left[-\frac{1}{2}(\vec{\theta} - \vec{\theta}_{MLE})^T H(\vec{\theta} - \vec{\theta}_{ML})\right]$?
 - Here H is the information matrix $I(\vec{\theta})$
 - True only to $\mathcal{O}(\frac{1}{N})$
 - In these conditions, $V[\vec{\theta}_{ML}] \rightarrow \frac{1}{I(\vec{\theta}_{ML})}$
 - Intervals can be derived by crossings: $\Delta \ln L = \ln L(\theta') - \ln L_{max} = k$
- Convince yourselves of how good is this approximation in case of the nuclear decay (simplified case of N measurements in which $t_i = 1$)!
[wget https://raw.githubusercontent.com/vischia/statex/master/nuclearDecay.R](https://raw.githubusercontent.com/vischia/statex/master/nuclearDecay.R)



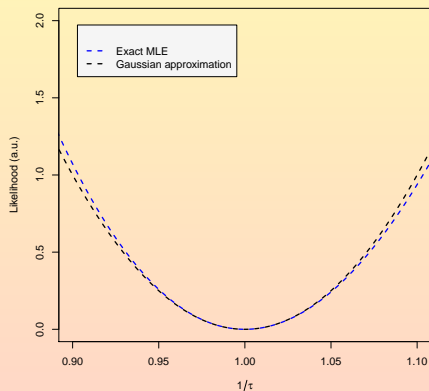
Nuclear decay at time $t=1$ and $N=1$



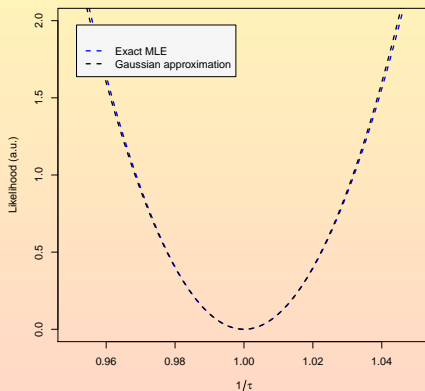
Nuclear decay at time $t=1$ and $N=10$



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

Asymptoticity of the Central limit theorem

- 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?

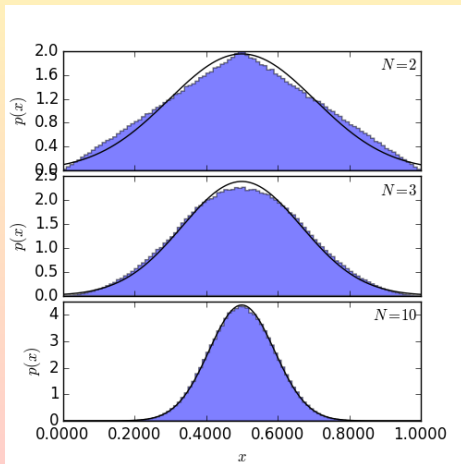
Asymptoticity of the Central limit theorem

- 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?
- Download the file and check!

wget <https://raw.githubusercontent.com/vischia/statex/master/centrallimit.py>

Asymptoticity of the Central limit theorem

- 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?
- Download the file and check!
 wget <https://raw.githubusercontent.com/vischia/statex/master/centrallimit.py>
- Not much!



From sidebands to systematic uncertainties

- 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

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

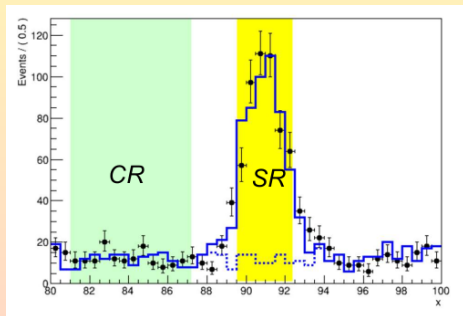
$$L_{CR}(b) = \text{Poisson}(N_{CR} | \tilde{\tau} \cdot b)$$

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

- 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

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



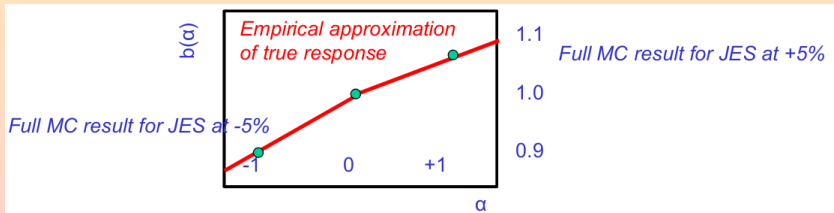
Renormalization of the subsidiary measurement

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



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

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

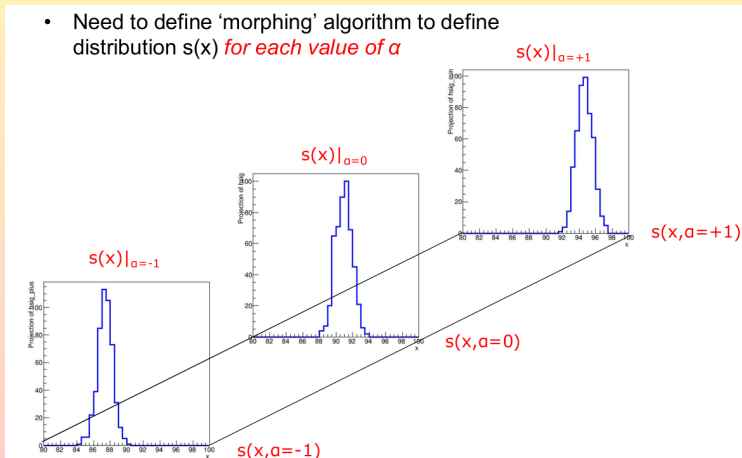


Graphics from W. Verkerke

Interpolation needed between template models

- Conditional density $f(x|\alpha)$ constructed by some means for a discrete set of values $\alpha_1, \dots, \alpha_N$
- The exact dependence of $f(x|\alpha)$ on α is unknown
 - In practice $f(x|\alpha_i)$ often nonparametric density estimates in the x space (e.g. histograms)
- Problem: determine $f(x|\alpha)$ for arbitrary α_i
 - Typically α_i within the cloud of $\alpha_1, \dots, \alpha_N$, and direct calculation too expensive
 - Need to keep the densities normalized: $\int f(x|\alpha) dx = 1, \forall \alpha$

- Need to define 'morphing' algorithm to define distribution $s(x)$ *for each value of α*



Graphics from W. Verkerke

Horizontal or vertical morphing?

- Vertical interpolation of single-parameter 1D densities:

$$f(x|\alpha) = w_1 f(x|\alpha_1) + (1 - w_1) f(x|\alpha_2),$$

$$w_1 = \frac{\alpha_2 - \alpha}{\alpha_2 - \alpha_1}, \alpha \in [\alpha_1, \alpha_2]$$

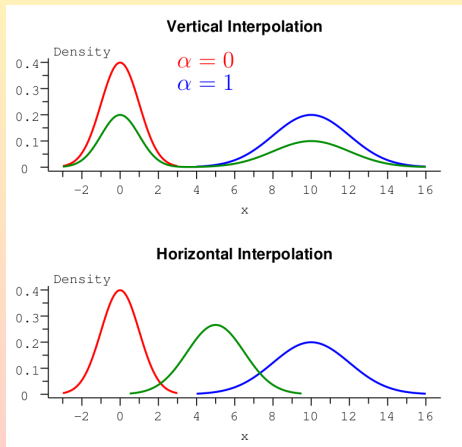
- Horizontal interpolation: identical parameter dependence, but interpolate quantile function

$$q(y|\alpha) = w_1 q(y|\alpha_1) + (1 - w_1) q(y|\alpha_2),$$

$$q(y|\alpha) := F^{-1}(y|\alpha)$$

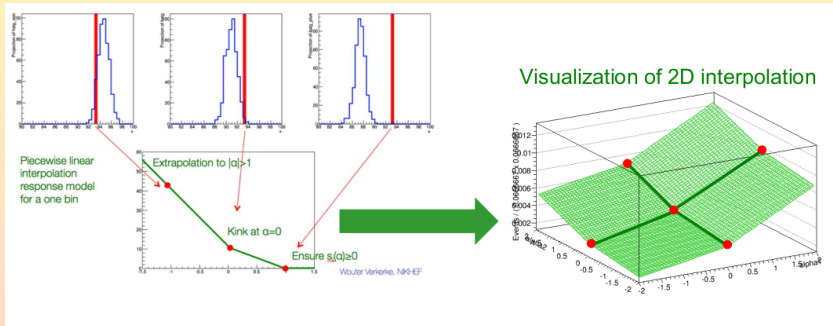
- Have to solve $q(y|\alpha) = x$ numerically
- Difficult to evaluate numerically around $y = 0$ and $y = 1$

- Vertical interpolation is often not what you want
 - Except some cases, e.g. interpolation of detector efficiency curves



Horizontal interpolation/morphing in one dimension

- For HEP application and univariate densities, reasonable solution is linear interpolation
 - A.L. Read, Linear interpolation of histograms, NIM A 425, 357 (1999)
 - Can fail dramatically if the change in shape is comparable with or smaller than MC statistical fluctuations
 - Sometimes we may want to avoid adding this new degree of freedom in the model
 - Decoupling rate and shape effects is always possible, even when not neglecting the shape ones)

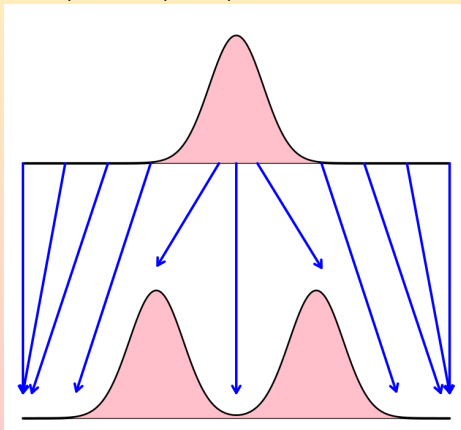


Graphics from W. Verkerke

- The cases $f(\vec{x}|\alpha)$ and $f(\vec{x}|\vec{\alpha})$ remain delicate
- Multivariate parameters: $g(\cdot|\vec{\alpha}) = \sum_{i=1}^N w_i(\vec{\alpha}, \vec{\alpha}_1, \dots, \vec{\alpha}_N)g(\cdot|\vec{\alpha}_i)$
 - $g(\cdot|\vec{\alpha})$ either density function (x) or quantile function (y)
 - Non-negative weights summing up to 1; many techniques (polynomial, local poly, spline best used in 1D)
 - Lack of generality because assumes Euclidean space

What if our metric is not Euclidean?

- Given two distributions P_0 and P_1 , define an *optimal map* T transforming $X \sim P_0$ into $T(X) \sim P_1$ (Monge, 1781)
- Define a geodesic path between P_0 and P_1 in the space of the distributions, according to a given metric
 - Shape-preserving notion of averages of distributions
 - Distance based on transport along geodesic paths
- Let $X \sim P_0$, and find T by minimizing $\mathbb{E} \left[\| X - T(X) \|^p \right] = \int \| x - T(x) \|^p dP_0(x)$
 - Minimization over all T s.t. $T(X) \sim P_1$. Can replace Euclidean distance with any distance
 - The minimizer is called *optimal transport map*

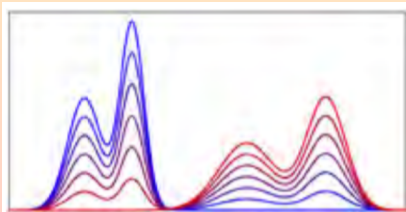


Generalize to arbitrary metric

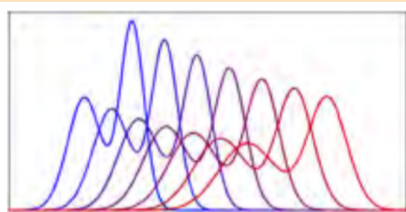
- Formally a minimization of the weighted average distance:

$$S(f, \vec{\alpha}, \vec{\alpha}_1, \vec{\alpha}_N) = \sum_{i=1}^N w_i(\vec{\alpha}, \vec{\alpha}_1, \vec{\alpha}_N) \left[D(f(x|\vec{\alpha}), f(x|\vec{\alpha}_i)) \right]^p$$

- $D(f(x), g(x))$ is a distance (metric functional in the space of distributions)
- Every metric generates an interpolation method (see Chap. 14 of *Encyclopedia of Distances*, Deza and Deza, 4ed., Springer, 2016)
- L^2 distance generates vertical morphing (with $p = 2$, $[D(\cdot)]^p$ is the integrated squared error)
- Wasserstein distance generates horizontal morphing (p=1 Earth Mover distance)
 - $W_p(X, Y) := W_p(P_0, P_1) = \left(\int \|x - T^*(x)\|^p dP_0(x) \right)^{1/p}$, T^* optimal transport map
 - Works well in defining a metric in the space of almost all distributions
 - The set of distributions equipped with Wasserstein distance is a geodesic space (Riemannian if $p = 2$)
 - Given P_0 and P_1 there is always a shortest path (geodesic) between them, and its length is the Wasserstein distance $W(P_0, P_1)$

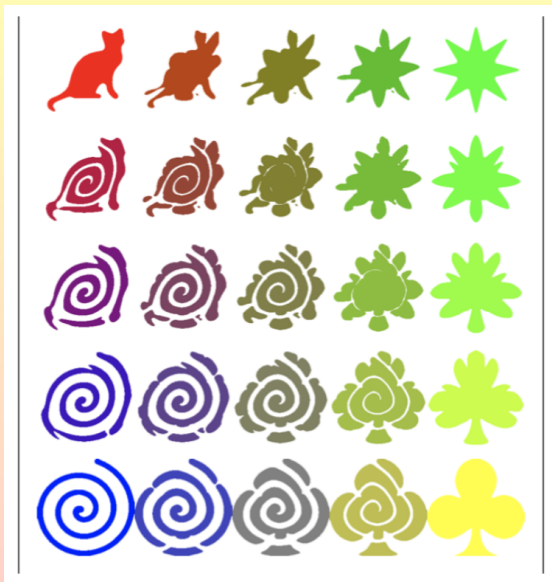


ℓ_2 interpolation



Wasserstein interpolation

Graphics from Bonneel, Peyre, Cuturi, 2016



Graphics from Peyre, Cuturi, 2019

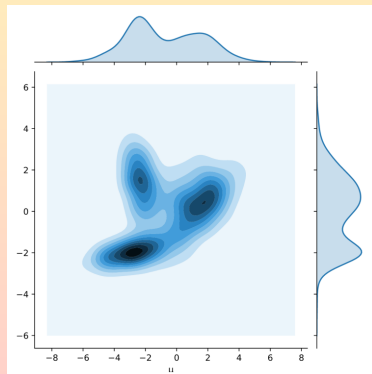
What if a transport map from P_0 to P_1 does not exist?

- Example: $P = \delta_0$ (point mass at 0), $Q = \text{Gaussian}$
- Kantorovich relaxation: take the mass at x and split it into small components
- \mathcal{J} set of all joint distributions J for (X, Y) with marginals P and Q (coupling between P and Q)
- Find J to minimize $\mathbb{E}_J \left[\|X - Y\| \right] = \left(\int \|x - y\|^p dJ(x, y) \right)^{\frac{1}{p}}$
- Wasserstein distance: $W(P, Q) = W(X, Y) = \left(\inf_J \int \|x - y\|^2 dJ(x, y) \right)^{\frac{1}{2}}$

- If an optimal transport T exists, then the optimal J is degenerate and supported on the curve $(x, T(x))$
- Regularization possible by adding term:

$$\mathbb{E}_J \left[\|X - Y\| \right] = \left(\int \|x - y\|^p dJ(x, y) \right)^{\frac{1}{p}} + \lambda f(J)$$

- $f(J)$ e.g. entropy
- Fast, and easier inference
- How to choose λ ? Not clear effect of regularization

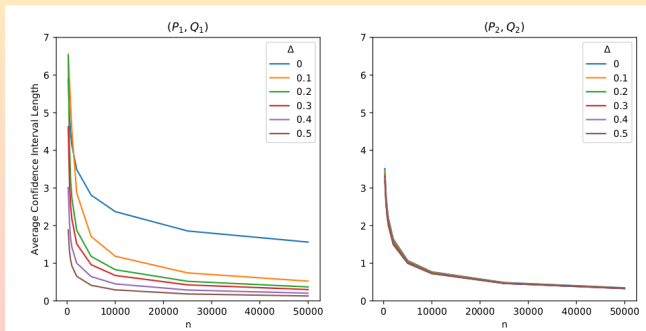


Graphics from Wikipedia

Uncertainty quantification

- These methods introduce an uncertainty in the morphed shape determination
- \hat{T} estimate of T based on samples $X_1, \dots, X_N \sim P_0, Y_1, \dots, Y_N \sim P_1$
- Closeness of \hat{T} to T ($\hat{W}(P_0, P_1)$ to $W(P_0, P_1)$) depends on number of dimensions

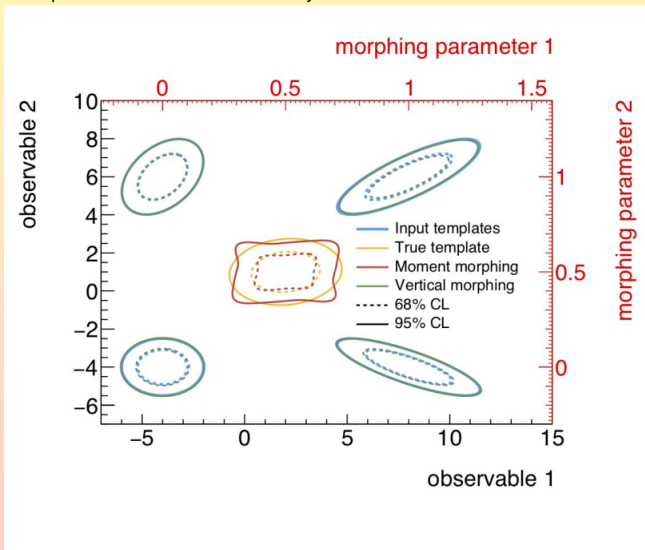
$$\mathbb{E} \int \|\hat{T}(x) - T(x)\|^2 dP_0(x) \approx \left(\frac{1}{N}\right)^{\frac{1}{d}}$$
 (curse of dimensionality)
- Getting confidence intervals very hard, solved only for special cases
 - 1D (Munck, Czado, Sommerfeld)
 - MultiD: sliced Wasserstein distance (average W between 1D projections of P_0 and P_1)
 - Under this approximation (weaker metric), can derive confidence regions by a minimax game on the L^1 norm of quantile functions of P_0 and P_1 for a fixed confidence level
 - Coverage guaranteed by construction



Graphics from [arXiv:1909.07862](https://arxiv.org/abs/1909.07862). Here P_0 is P and P_1 is Q , indices refer to two example cases, $n = 100$

Moment morphing

- Moment morphing: morph standardized densities instead of densities
 - Useful for models with well-behaved first moments (mean and variance)
 - Not as good as horizontal morphing in 1D (inefficient version of it), good approximation in N
 - How to morph the covariance matrix? Many choices available



Graphics from Lydia Brenner

The Inverse Rosenblatt Transformation

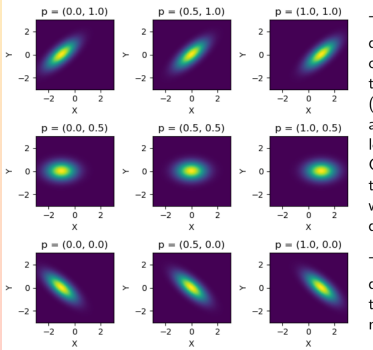
- Devise a multi-D equivalent of quantile function: the *Inverse Rosenblatt transformation* (Ann. Math. Statist. 23, 470 (1952)).
- The inverse Rosenblatt transformation $x_1 = F_1^{-1}(z_1), x_2 = F_2^{-1}(z_2 | z_1)$ uses conditional quantile functions: we know how to interpolate them!
- Computationally intensive (k non-linear equations to be solved numerically, N calls to root-finding, etc)

Let $X = (X_1, \dots, X_k)$ be a random vector with distribution function $F(x_1, \dots, x_k)$. Let $z = (z_1, \dots, z_k) = TX = T(x_1, \dots, x_k)$, where T is the transformation considered. Then T is given by

$$\begin{aligned} z_1 &= P\{X_1 \leq x_1\} = F_1(x_1), \\ z_2 &= P\{X_2 \leq x_2 \mid X_1 = x_1\} = F_2(x_2 \mid x_1), \\ &\vdots \\ z_k &= P\{X_k \leq x_k \mid x_{k-1} = z_{k-1}, \dots, X_1 = x_1\} = F_k(x_k \mid x_{k-1}, \dots, x_1). \end{aligned}$$

One can readily show that the random vector $Z = TX$ is uniformly distributed over the k -dimensional unit cube.

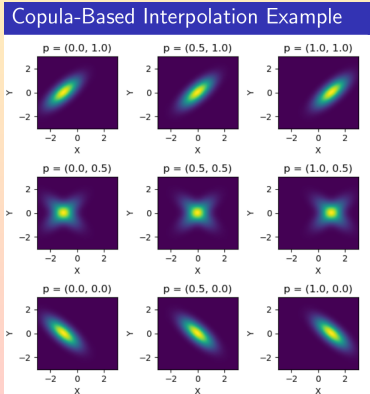
Inverse Rosenblatt Interpolation Example



Graphics by Igor Volobouev

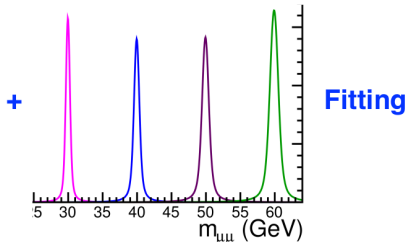
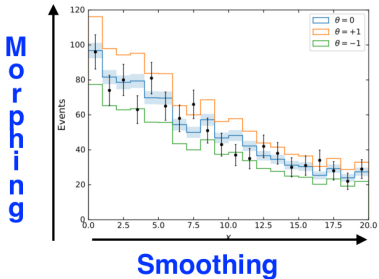
Copula morphing

- Probability integral transforms of marginals of $f(\vec{x})$: $z_1 = F_1(x_1), \dots, z_k = F_k(x_k)$
- Copula density $c(\vec{z})$ is density of the vector of z_k , captures mutual information (and $c(\vec{z})$ uniform if and only if all X_i independent)
- Given the marginal densities $f_i(x) = \frac{dF_i(x)}{dx}$, then $f(\vec{x}) = c(F_1(x_1), \dots, F_k(x_k)) \prod_{i=1}^k f_i(x_i)$
- Now do horizontal morphing on the marginals separately in each variable, then interpolate vertically the copula density
- Much faster than Inverse Rosenblatt transformation
- Results intuitively more “reasonable”

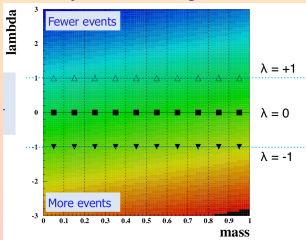


Graphics by Igor Volobouev

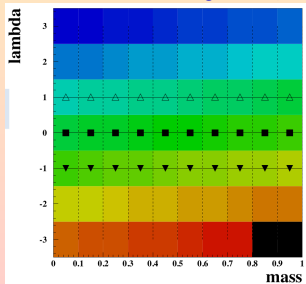
How we tend to call things in CMS



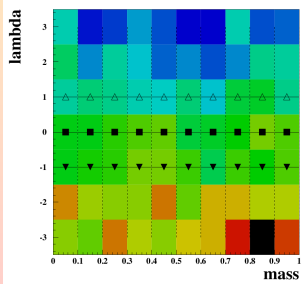
Analytic knowledge on λ, m

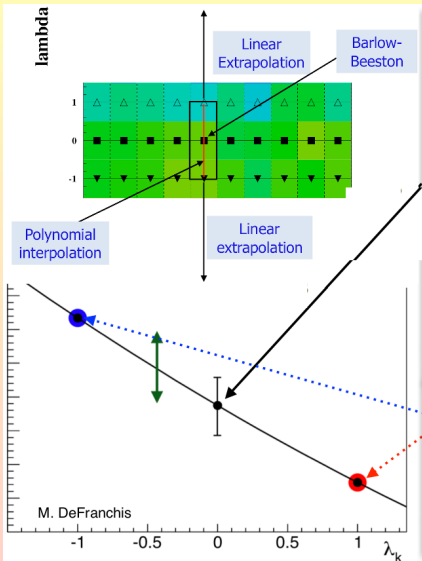


Discretized knowledge on λ, m



Statistical fluctuations





Statistical uncertainty of nominal templates taken into account in Poisson based template fits to data

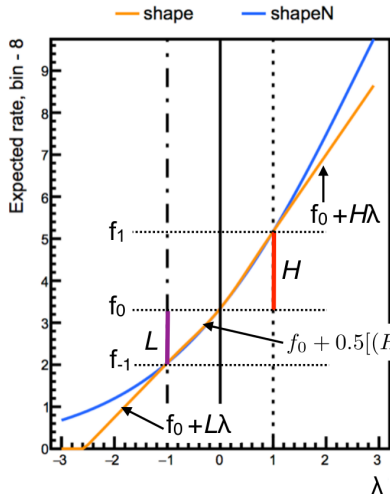
- 'Barlow Beeston': one additional nuisance parameter per contributing template J. Barlow, C. Beeston, CPC 77 (1993) 219-228
- 'Barlow Beeston lite': one additional nuisance parameter for templates sum \rightarrow **Standard Procedure in CMS**

John Conway, arXiv1103.0354

Statistical uncertainty of $\pm 1\sigma$ Templates usually neglected \rightarrow can lead to fake constraints for λ , see https://indico.cern.ch/event/761804/contributions/3160985/attachments/1733339/2802398/Defranchis_template_constraints.pdf

https://indico.cern.ch/event/761804/contributions/3160985/attachments/1733339/2802398/Defranchis_template_constraints.pdf

Morphing in the Higgs Combination Tool



<https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/>

Shape morphing:

- First normalise templates dividing by the sum over all horizontal (e.g. mass) bins → obtain fractions
- shape: morph bin-wise fractions vs λ
- shapeN: morph log of bin-wise fractions vs λ

Interpolation for $-1 < \lambda < 1$:

$$f_0 + 0.5[(H - L)\lambda + 1/8(H + L)[3\lambda^6 - 10\lambda^4 + 15\lambda^2]]$$

Fulfills $f''=0$ for $\lambda=-1$ and $\lambda=1$

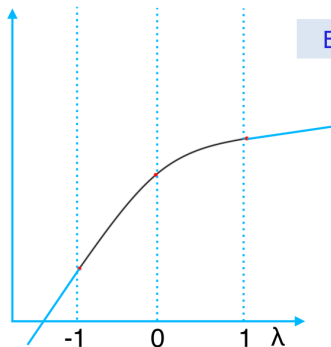
Slide by Olaf Behnke

Cubic spline interpolation + straight line extrapolation

Used in  Tool

<http://www.eikp.physik.uni-karlsruhe.de/~cttheta/testing.html#index.html>
<http://www.eikp.physik.uni-karlsruhe.de/~cttheta/theta-auto/index.html>

Template variation
in a bin



Example



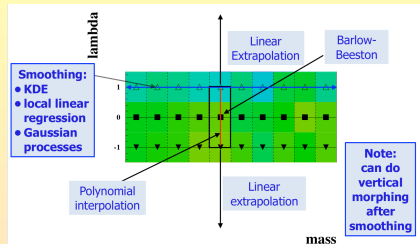
Alternative: **overall straight line** → need to symmetrise uncertainties
 Could be tested with additional templates for $\lambda = -3, 2, 2, 3$ etc.

10

Slide by Olaf Behnke

Horizontal smoothing

- Horizontal smoothing with well-established methods in literature
- Kernel-based methods depend on choice of bandwidth
 - Discussed in detail last week (Nick McColl)
- Local linear regression depends on locality window



Kernel Density Estimation (KDE)

Material © Chad Shafer:
<https://indico.in2p3.fr/event/19290/contributions/75800>

- Sample n independent points X_i from unknown distribution f
- KDE estimate:
 - Example: Gaussian Kernel

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

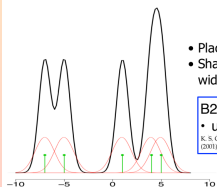
$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

- Places a smoothed-out lump over each data
- Shape of 'lumps' is controlled by $K(\cdot)$; their width controlled by h

B2G-18-008:

- use adaptive width $h \sim 1/\sqrt{f(x)}$

K. S. Coombes, "Kernel estimation in high-energy physics", *Comput. Phys. Commun.* **136** (2001) 198, doi:10.1016/S0010-4655(00)00243-5, arXiv:hep-ex/0011057.



13

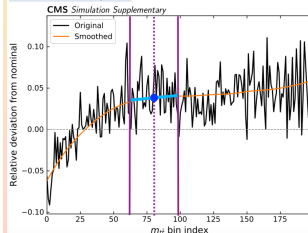
Slide by Olaf Behnke

Local Linear Regression (LOWESS)

CMS HIG-17-027

See talk A. Popov

<https://indico.in2p3.fr/event/19290/contributions/75800>



- Use points in sliding window
- Give points near centre larger weights
- Fit straight line
- Move window →
- Connect fitted window centre points

Optimise hyper-pars with cross-validation

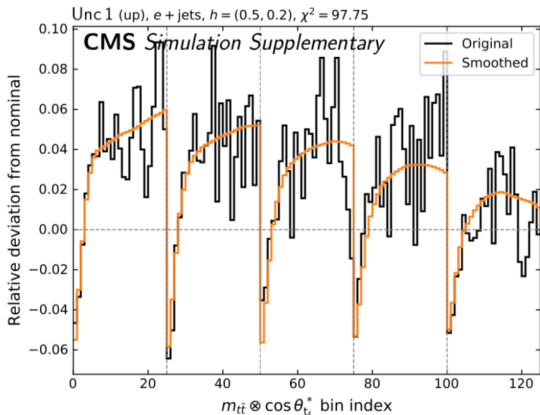
15

Smoothing and Goodness-of-Fit tests

- To compare the smoothed and unsmoothed templates it's tempting to use χ^2
- However, χ^2 not well defined; by construction, smoothing alters number of degrees of freedom
- You have first to treat your smoothing method as a linear filter, and calculate NDoF (in KDE, related to autocorrelation of the kernels used)
 - Somehow related to time series analysis: reduction of NDoF
 - There is literature on this, we can put it in twiki; in the meantime, ask Igor Volobouev ☺

Local Linear Regression (LOWESS)

CMS HIG-17-027



Example for
Final s/b
discriminator
Smoothing

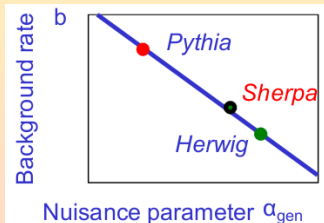
χ^2 GOF Tests



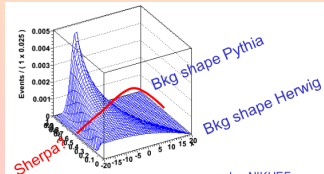
Caveats on modelling theory uncertainties (P.V. at Benasque 2018)

- Cross section uncertainty: easy, assuming a gaussian for the constraint term
 $\mathcal{L}_{full}(s, b) = \mathcal{P}(N_{SR}|s + b) \times \mathcal{G}(b|0.08)$
- Factorization scale: what distribution \mathcal{F} is meant to model the constraint???
- Hadronization/fragmentation model: run different generators, observing different results
 - “Easy” case, there is a single parameter α_{FS} , clearly connected to the underlying physics model
- 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 α corresponding to SHERPA



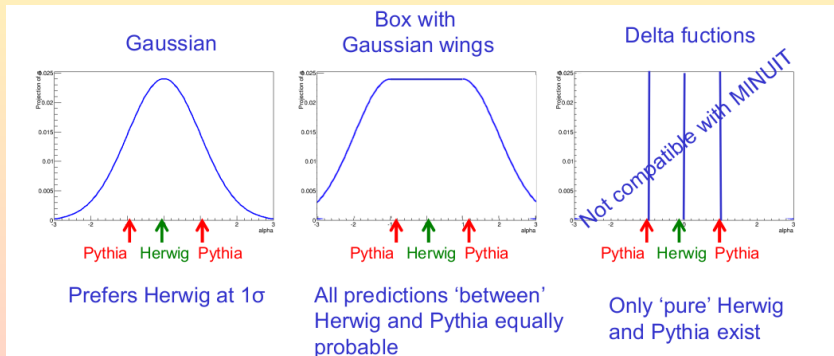
- Shape experiment: ouch!
- SHERPA is in general not obtainable as an interpolation of PYTHIA and HERWIG



Graphics from W. Verkerke

Define a constraint term

- 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?



Prefers Herwig at 1σ

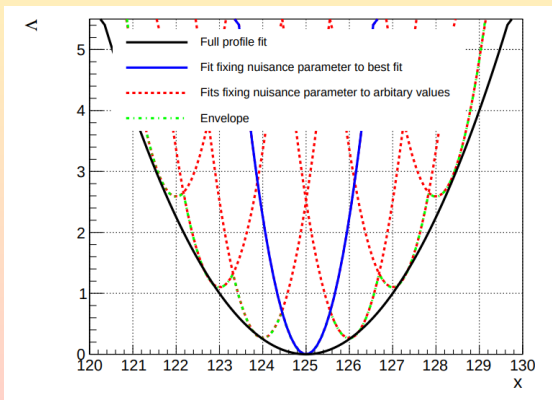
All predictions 'between'
 Herwig and Pythia equally
 probable

Only 'pure' Herwig
 and Pythia exist

Graphics from W. Verkerke

Solving the delta functions issue: discrete profiling

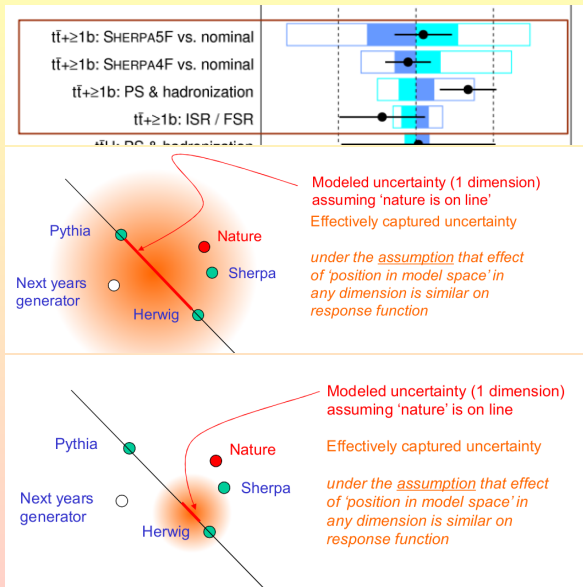
- Label each shape with an integer, and use the integer as nuisance parameter
- Can obtain the original log-likelihood as an envelope of different fixed discrete nuisance parameter values
- How do you define the various shapes?
 - Need many additional generators!
 - Interpolation unlikely to work (*SHERPA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



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

The issue of over-constraining

- 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

- 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

THANKS FOR THE ATTENTION!

Backup