Statistical methods for data treatment with and without signal (ATLAS+CMS)

Pietro Vischia[1]
(on behalf of the ATLAS and CMS Collaborations)

[1] CP3 — IRMP, Université catholique de Louvain

UCLouvain
Institut de recherche
en mathématique et physique

Kolumbari, ICFNP2019 Lecture

The purpose of this lecture

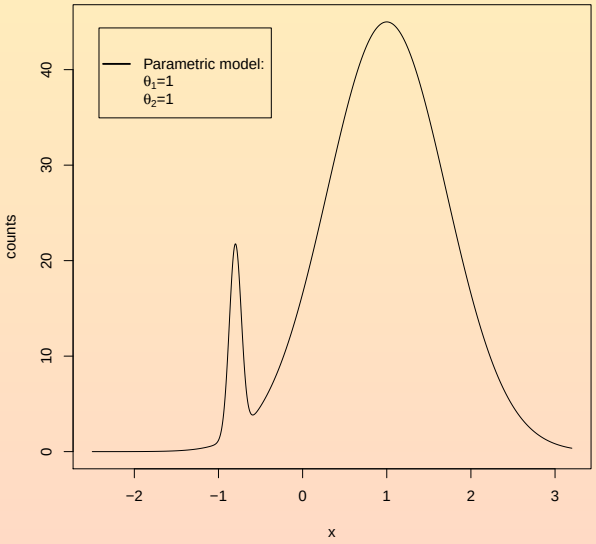
- *The lecture is supposed to help non experts to understand what is shown in the other talks of CMS and ATLAS on the new searches.*
- Difficult to resume everything in a single talk
- Will focus on some fundaments and on a few selected topics
  - Definition of probability
  - Point estimation
  - Interval estimation
  - Hypothesis testing
  - Unfolding
  - Gaussian processes for unknown signals

Fundaments

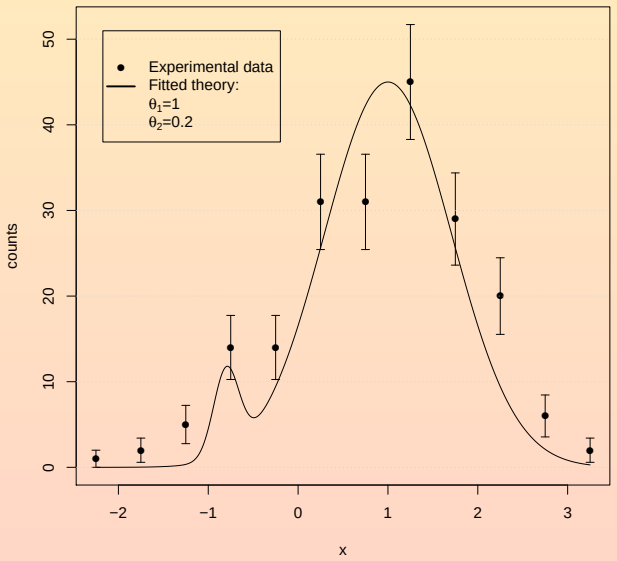
Where does statistics live

UCLouvain
Institut de recherche
en mathématique et physique

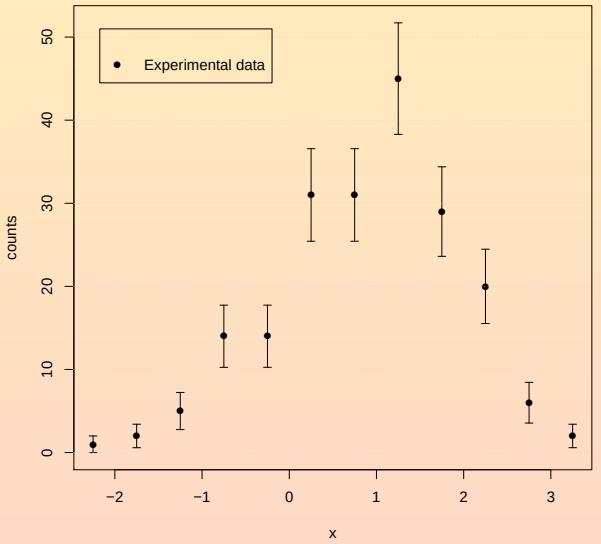
- **Theory**
  - Approximations
  - Free parameters



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



- **Experiment**
  - Random fluctuations
  - Mismeasurements (detector effects, etc)



What is a "probability"? — Kolmogorov

- 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.

What is a "probability"? — Cox and Jaynes

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

Probability in the Theory of Measure — What's a length?

- 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 defined for an interval in $\mathbb{R}^n$ (e.g. in 3D it's 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 $[-\infty, 2]$.
- Connect different intervals via the Borel lemma
  - In layman's terms, sets that can be constructed by taking countable unions or intersections (and their respective complements) of open sets
  - **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: $i = \sum i_n \Rightarrow L(i) = \sum L(i_n)$, a *measure*
- Definition extendable from intervals to complex sets (true only for Borel sets): *Lebesgue measure*
  - $L(S) \geq 0$
  - If $S = S_1 + \ldots + S_n$, where $S_\mu S_\nu = 0$ for $\mu \neq \nu$ then $L(S) = L(S_1) + \ldots + L(S_n)$
  - If S is an interval i, then the set function $L(S)$ reduces itself to the interval function $L(i)$, $L(S) = L(i)$

Probability measure!

- Generalization of $L_n(S)$: the P-measure
  1. $P(S)$ is non-negative, $P(S) \geq 0$;
  2. $P(S)$ is additive, $P(S_1 + \ldots + S_n) = P(S_1) + \ldots + P(S_n)$ 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 $F(\boldsymbol{x}) = F(x_1, \ldots, x_n) := P(\xi_1 \leq x_1, \ldots, \xi_n \leq x_n)$
  - Trivial in one dimension. $P(S)$ must have an upper bound!
  - Map $F(a) = F(b)$ to set of null P-measure, $P(a < x \leq b) = 0$
- $F(\boldsymbol{x})$ is in each point a non-decreasing function everywhere-continuous to the right

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

- We interpret $P(S)$ and $F(\boldsymbol{x})$ as distribution of a unit of mass over $\mathbb{R}^n$
  - Each Borel set carries the mass $P(S)$
  - Interpret ($\boldsymbol{x}$ as the quantity of mass allotted to the infinite interval $(\xi_1 \leq x_1, \ldots, \xi_n \leq x_\nu)$.
  - Defining the measure in terms of $P(S)$ or $F(\boldsymbol{x})$ is equivalent
- Usually $P(S)$ is called probability function, and $F(\boldsymbol{x})$ is called distribution function

Discreteness, probability density function

- What about individual points?
  - Discrete mass point $\boldsymbol{a}$; a point such that the set $\{\boldsymbol{x} = \boldsymbol{a}\}$ carries a positive quantity of mass.

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

or

$$F(\boldsymbol{x}) = c_1 F_1(\boldsymbol{x}) + c_2 F_2(\boldsymbol{x})$$

where

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

  - c_1: component with whole mass concentrated in discrete mass points. c_2: component with no discrete mass points
  - $c_1 = 1, c_2 = 0$: $F(\boldsymbol{x})$ 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 $F(\boldsymbol{x})$ is everywhere continuous, and in any dimension no single mass point carries a positive quantity of mass.
- Probability density function
  - Consider the n-dimensional interval $i = \{x_\nu - h_\nu < \xi_\nu \leq x_\nu + h_\nu; \nu = 1, \ldots, n\}$
  - 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

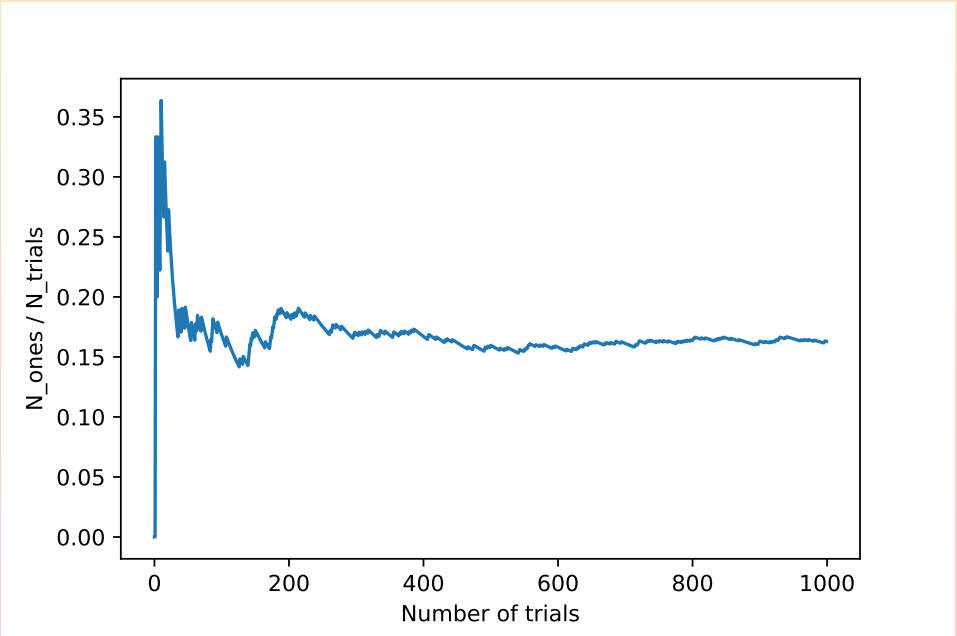
$$\frac{P(i)}{L(i)} = \frac{\Delta_n F}{2^n h_1 h_2 \ldots h_n}.$$

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

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
  - If $S = \{1\}$, then $\xi \in S$ corresponds to the event in which you score 1 with the die
- 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
- Frequentist probability for any single event to be of type X is the empirical limit of the frequency ratio:

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



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

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

- 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	$30 + 10 = 40$
		$0.5 + 0.25 = 0.75$	30	40

- Bayes Theorem (1763):

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

- 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_is, the Law of Total Probability:

$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i) \tag{3}$$

- where the second equality holds only for if the A_is are disjoint
- Finally, the Bayes Theorem can be rewritten using the decomposition of S as: $P(A|B) := \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$

- Bayesian statistics: the definition of probability is extended to the subjective probabilty of models or hypotheses:

$$P(H|\vec{X}) := \frac{P(\vec{X}|H)\pi(H)}{P(\vec{X})} \tag{4}$$

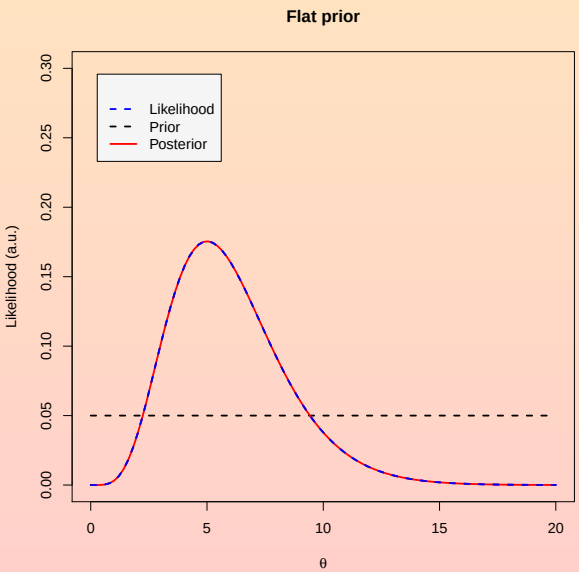
- $\vec{X}$, the vector of observed data
- $P(\vec{X}|H)$, the likelihood function, which fully summarizes the result of the experiment (experimental resolution)
- $\pi(H)$, the probability of the hypothesis H. It represents the probability we associate to H before we perform the experiment
- $P(\vec{X})$, 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).
  - When integration not needed (e.g. to find the mode rather than areas), omit denominator

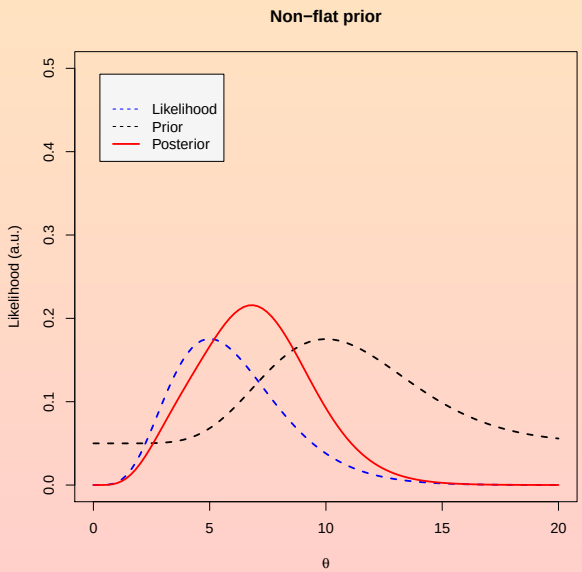
$$P(H|\vec{X}) \propto P(\vec{X}|H)\pi(H) \tag{5}$$

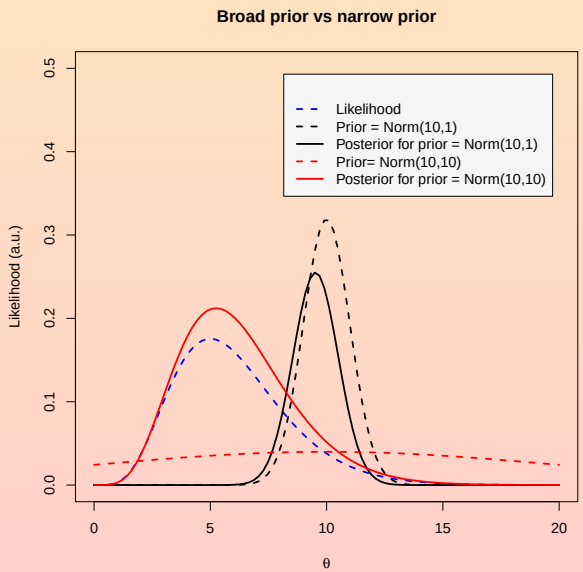
- $P(H|\vec{X})$, the posterior probability; it is obtained as a result of an experiment

Choosing a prior in Bayesian statistics

- 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
- The posterior probability is proportional to the product of the prior and the likelihood
  - The prior doesn't necessarily have to 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 $N \to \infty$
  - In this limit, the likelihood is always dominant in the product
  - The posterior is indipendent of the prior!
  - The posteriors corresponding to different priors must coincide, in this limit





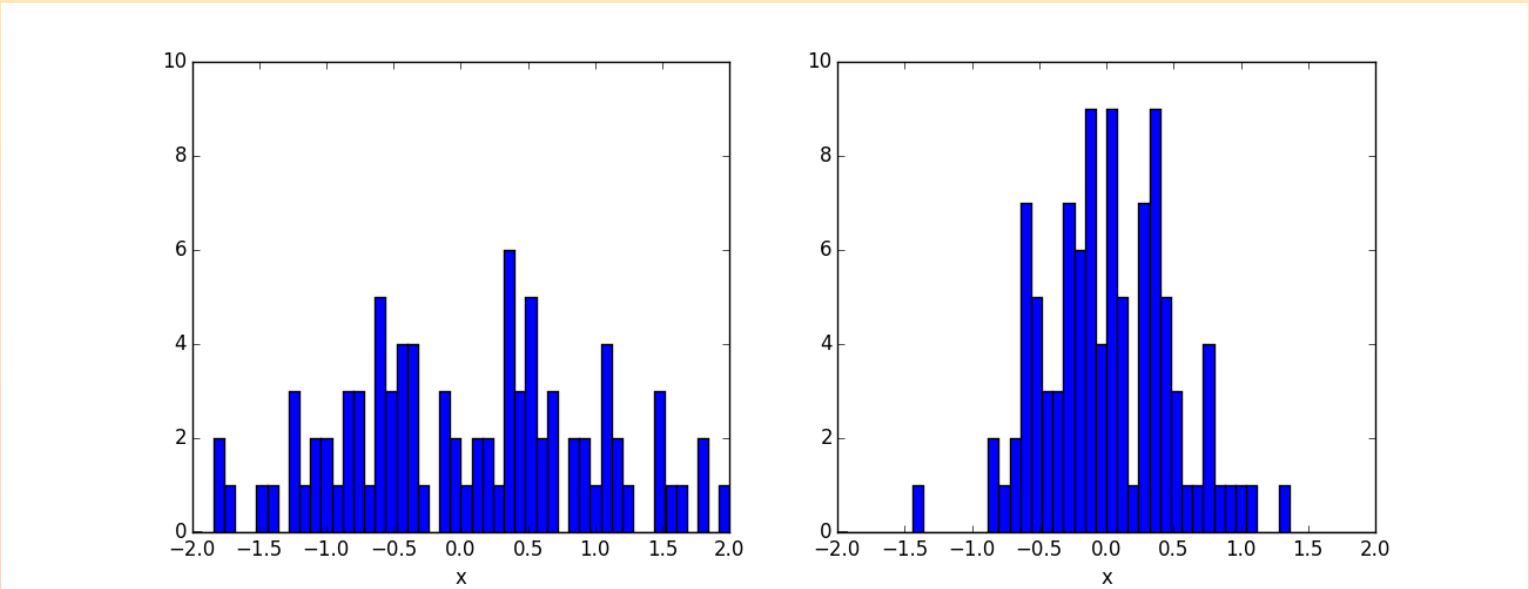


Short summary on bayesian vs. frequentist

- 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
  - Desirable property in HEP: indipendence of the result on the choice of the prior

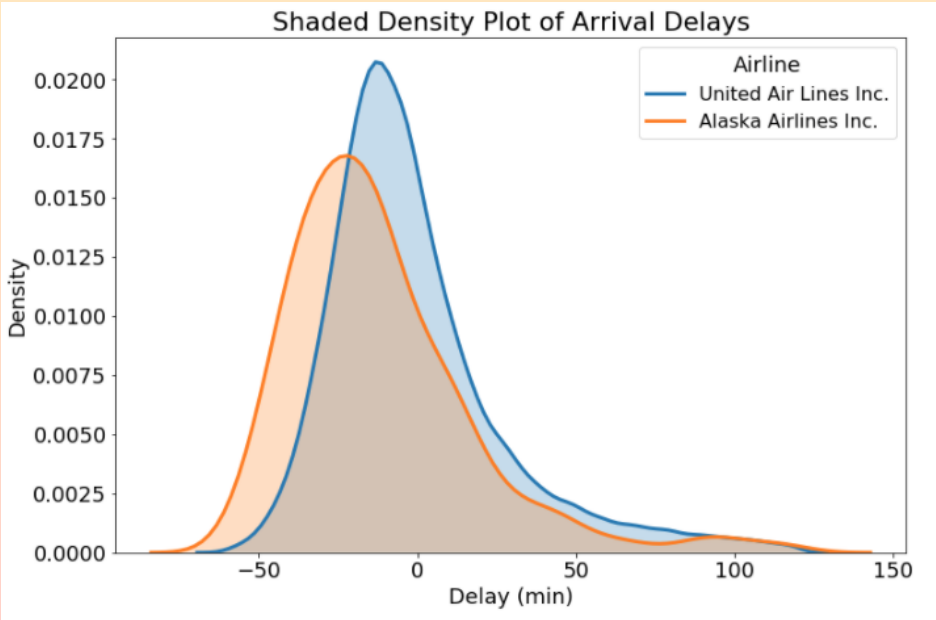
Dispersion and distributions

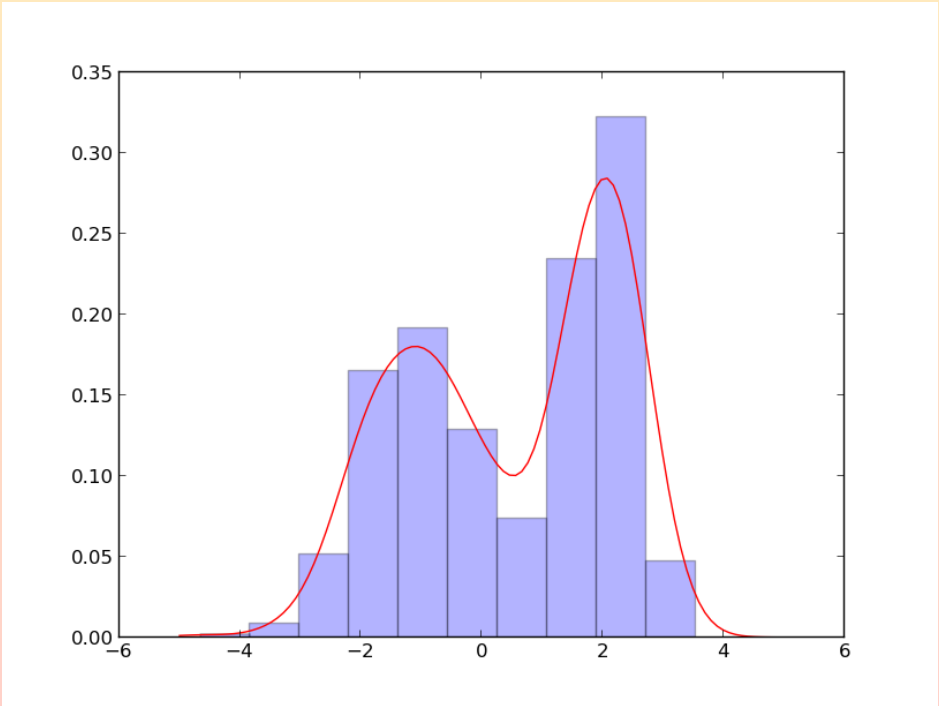
- Repeated experiments usually don't yield the exact same result even if the physical quantity is expected to be exactly the same
  - Random changes occur because of the imperfect experimental conditions and techniques
  - They are connected to the concept of dispersion around a central value
- When 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



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 binned data are the result of counting experiments $\rightarrow$ histograms
  - New physics often searched for in the tail of distributions
  - But for some applications (e.g. Machine Learning) even in HEP please consider using density estimates





Plots from TheGlowingPython and TowardsDataScience

Expected values

- We define the expected value and mathematical expectation

$$E[X] := \int_{\Omega} Xf(X)dX \tag{6}$$

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

$$E[X] := \sum_{i} X_i P(X_i) \tag{7}$$

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

$$E[g] := \int_{\Omega} g(X)f(X)dX \tag{8}$$

- The previous expression Eq. 6 is a special case of Eq. 8 when $g(X) = X$
- The mean of X is:

$$\mu := E[X] \tag{9}$$

- The variance of X is:

$$V(X) := E[(X-\mu)^2] = E[X^2] - \mu^2 \tag{10}$$

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

Let's make it funnier: more variables!

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

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

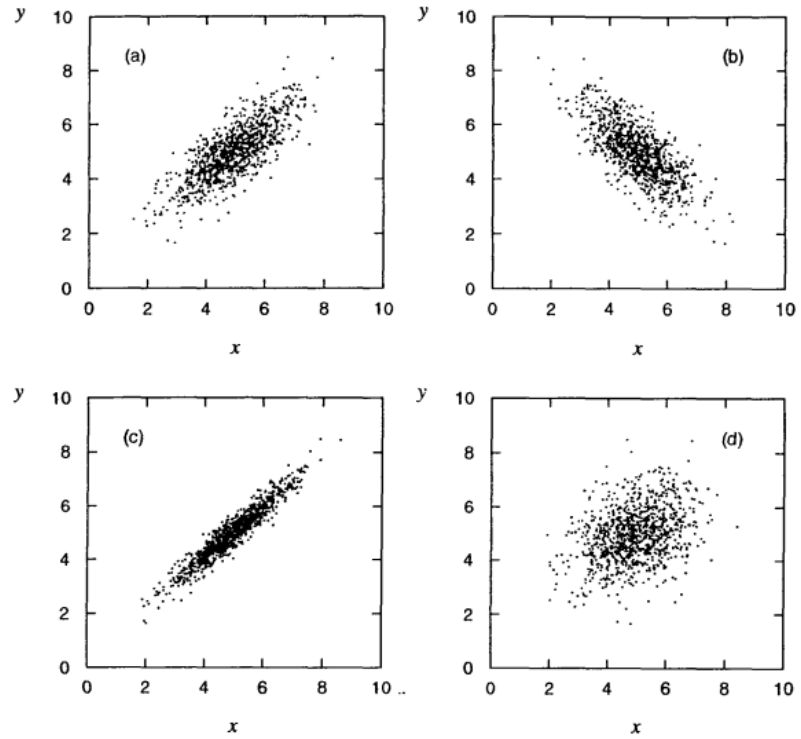


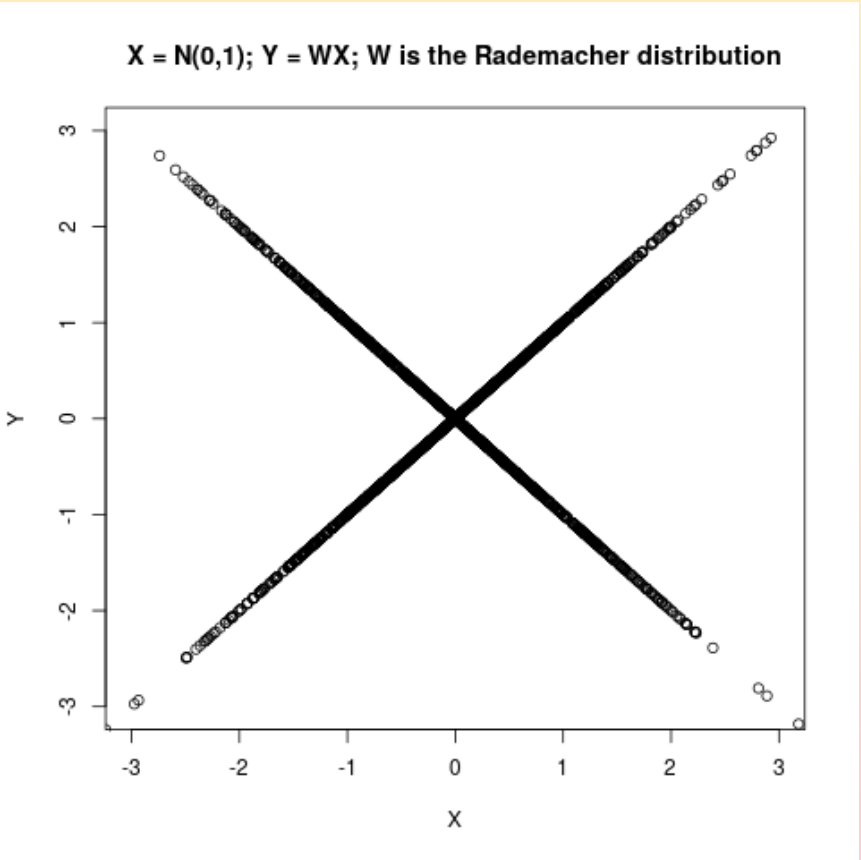
Fig. 1.9 Scatter plots of random variables x and y with (a) a positive correlation, $\rho = 0.75$, (b) a negative correlation, $\rho = -0.75$, (c) $\rho = 0.95$, and (d) $\rho = 0.25$. For all four cases the standard deviations of x and y are $\sigma_x = \sigma_y = 1$.

Mutual information: take it to the next level

- 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

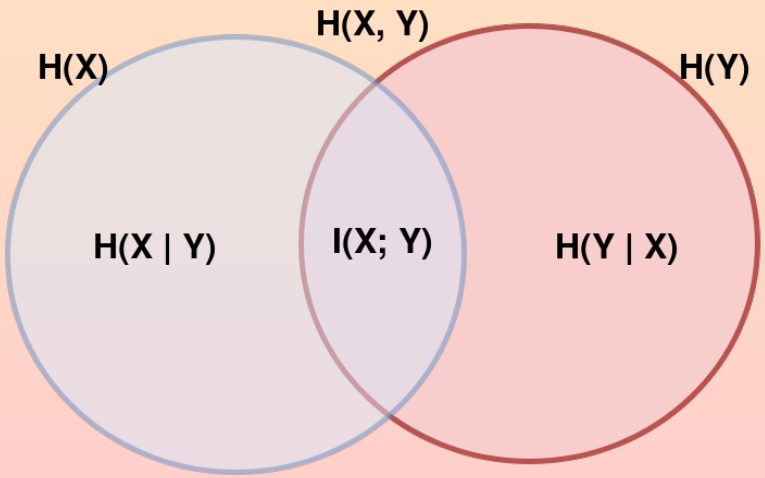
$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y) log\left(\frac{p(x,y)}{p_1(x)p_2(y)}\right)$$

- 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

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



Estimating a physical quantity

Information, the Fisher way

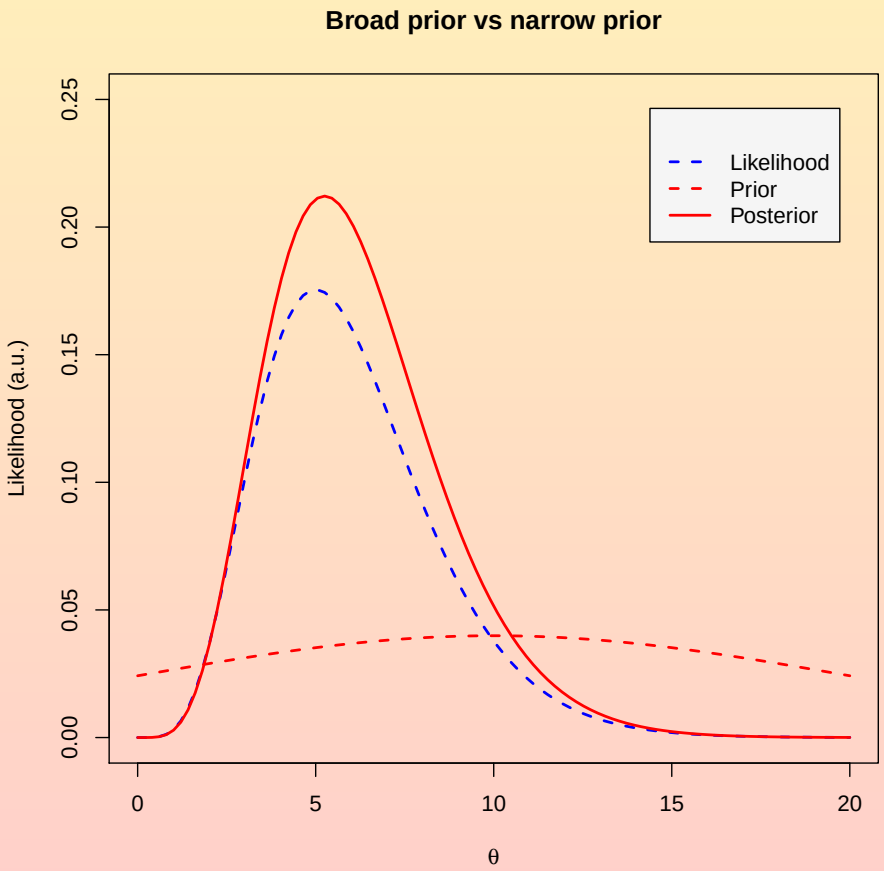
- The information of a set of observations should increase with the number of observations
  - Double the data should result in double the information if the data are independent
- Information should be conditional on what we want to learn from the experiment
  - Data which are irrelevant to our hypothesis 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

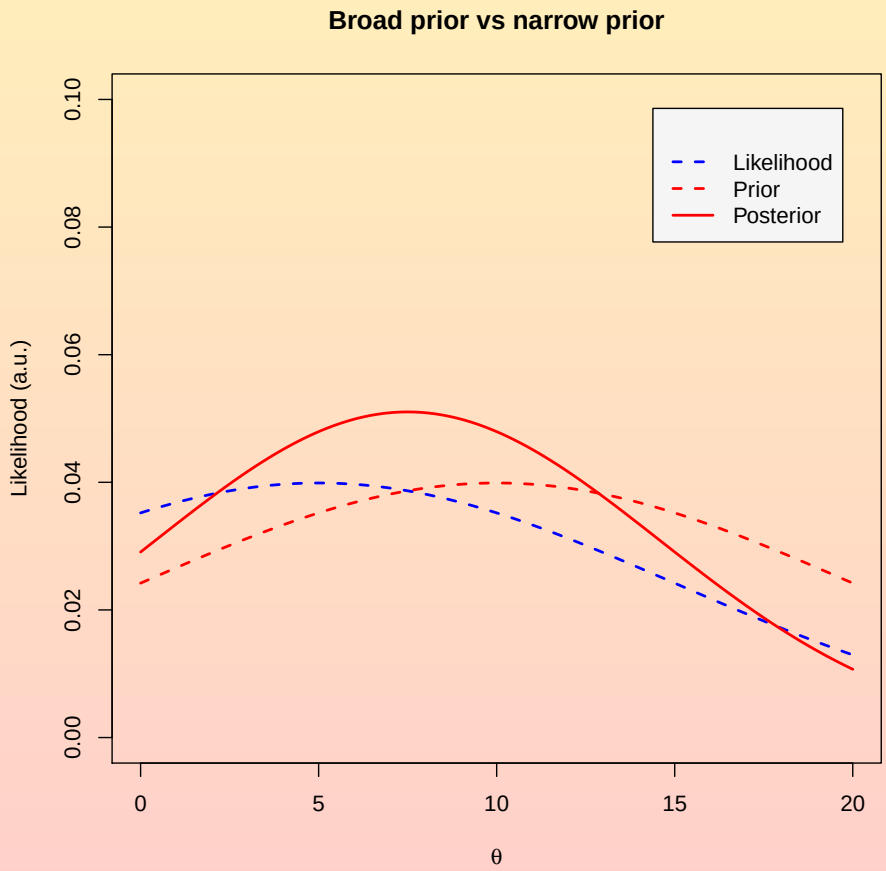
- The narrowness of the likelihood can be estimated by looking at its curvature
- The curvature is the second derivative with respect to the parameter of interest
- A very narrow (peaked) likelihood is characterized by a very large and positive $-\frac{\partial^2 lnL}{\partial \theta^2}$
- The second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

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

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





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

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

- 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 $\propto \frac{1}{\sqrt{\theta}}$

Sufficient statistic and data reduction

- A test statistic is a function of the data (a quantity derived from the data sample)
- A statistic $T = T(\boldsymbol{X})$ is sufficient for θ if the density function $f(\boldsymbol{X}|T)$ 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 $\boldsymbol{X}$
  - No other function can give any further information about θ
  - Same inference from data $\boldsymbol{X}$ with model M and from sufficient statistic $T(\boldsymbol{X})$ with model M'
- Example: binomial test in coin toss
  - Record heads and tails, with their order: $HTTHHHTHHTTTHTHTH$
  - 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: $P(r) = \binom{N}{r} p^r (1-p)^{N-r}$, $r = 0, 1, \ldots, N$
  - 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

- Common enunciation: given a set of observed data $\vec{x}$, the likelihood function $L(\vec{x};\theta)$ contains all the information relevant to the measurement of 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 $\vec{x}$
  - As we have seen, the likelihood is used to define the information contained in a sample
- Bayesian statistics normally complies, frequentist statistics usually does not, because a frequentist has to consider the hypothetical set of data that might have been obtained.
- This on one side implies that a frequentist always needs multiple sets of observations
  - 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
- On the other side a Bayesian would say “Probably tomorrow will rain”, a frequentist “the sentence -tomorrow it will rain- is probably true”

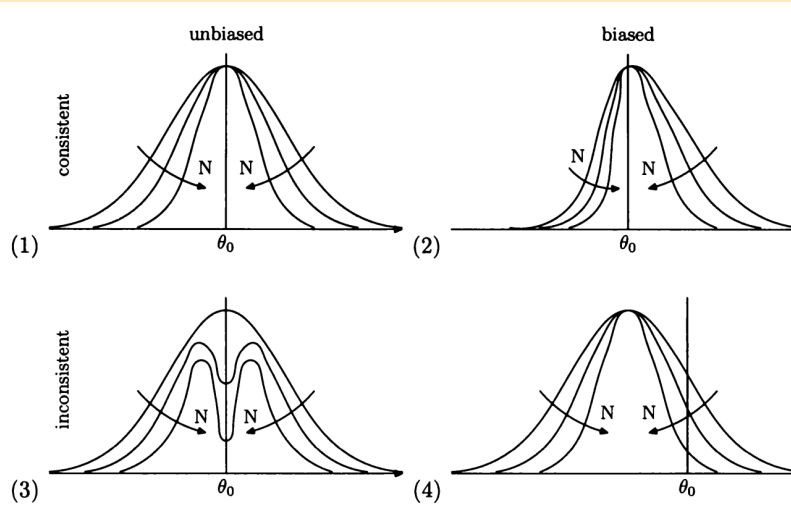
Estimators

- Set $\vec{x} = (x_1, ..., x_N)$ 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 $\vec{\theta} = (\theta_1, ..., \theta_M)$ unknown)
  - In case of a linear p.d.f., the vector of parameters would be $\vec{\theta} = (intercept, slope)$
- 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:

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

- $\hat{\vec{\theta}}$ 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



Plot from James, 2nd ed.

- 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.)

The Maximum Likelihood Method 1/

- Let $\vec{x} = (x_1, ..., x_N)$ 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

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

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

$$- lnL(\vec{x}; \vec{\theta}) = - \sum_{i=1}^{N} lnf(x_i, \vec{\theta})$$

  - 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

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

  - 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/htmldoc/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, $\vec{\theta}_{ML} = maxP(\vec{\theta}|\vec{x})$.

Bias-variance tradeoff and optimal variance 1/

- We usually want to optimize both bias $\vec{b}$ and variance $V[\hat{\vec{\theta}}]$
- While we can optimize each one separately, optimizing them simultaneously leads to none being optimally optimized, in genreal
  - 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

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

Information Inequality – 1

- Rao-Cramer-Frechet (RCF) bound
  $V[\hat{\theta}] \geq \frac{(1+\partial b/\partial\theta)^2}{-E\left[\partial^2 lnL/\partial\theta^2\right]}$
  - In multiple dimensions, this is linked with the Fisher Information Matrix:
    $I_{ij} = E\left[\partial^2 lnL/\partial\theta_i\partial\theta_j\right]$
- Approximations
  - Neglect the bias ($b = 0$)
  - Inequality is an approximate equality (true for large data samples)
- Variance: $V[\hat{\theta}] \simeq \frac{1}{-E\left[\partial^2 lnL/\partial\theta^2\right]}$
- Estimate of the variance: $\hat{V}[\hat{\theta}] \simeq \frac{1}{-E\left[\partial^2 lnL/\partial\theta^2\right]|_{\theta=\hat{\theta}}}$
- For multidimensional parameters, we can build the information matrix with elements:

$$I_{jk}(\vec{\theta}) = -E\Big[\sum_i^N \frac{\partial^2 lnf(x_i;\vec{\theta})}{\partial\theta_k\partial\theta_k}\Big]$$
$$= N\int \frac{1}{f}\frac{\partial f}{\partial\theta_j}\frac{\partial f}{\partial\theta_k}dx$$

- (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:

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

- 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 $N \to \infty$ then $b = 0$ and the variance approximate to the RCF bound, and $\geq$ becomes $\simeq$:

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

How to extract an interval from the likelihood function 1/

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

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

- 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/

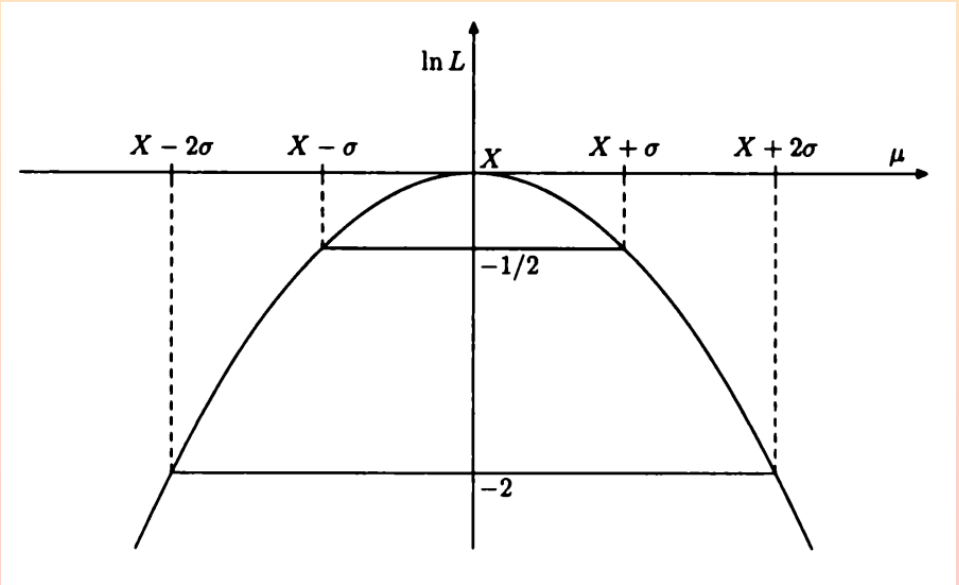
- We can therefore write

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

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

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

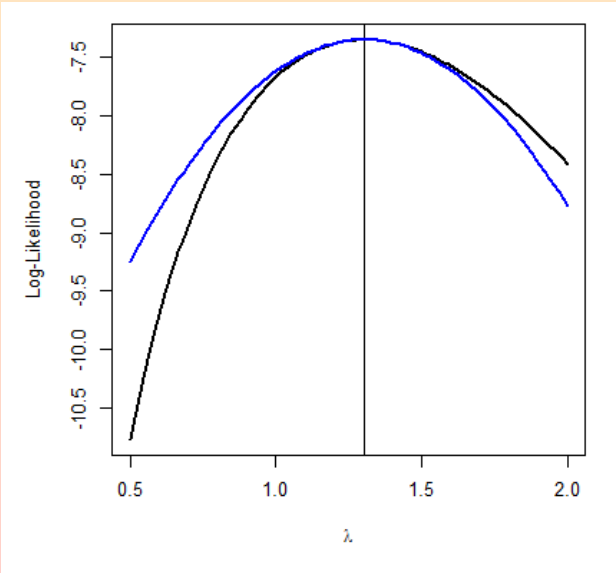
- Taking into account that it is important to keep in mind that probability is a property of sets, in frequentist statistics
  - Confidence interval: interval with a fixed probability content
- This process for computing a confidence interval is exact for a Gaussian p.d.f.
  - Pathological cases reviewed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

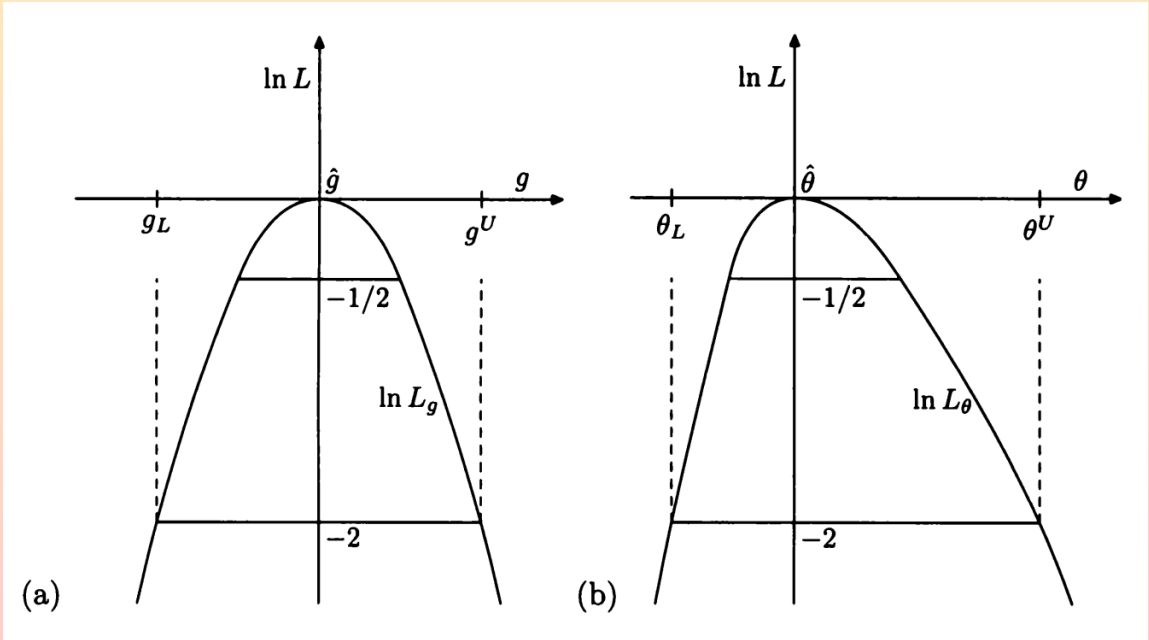


Plot from James, 2nd ed.

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.

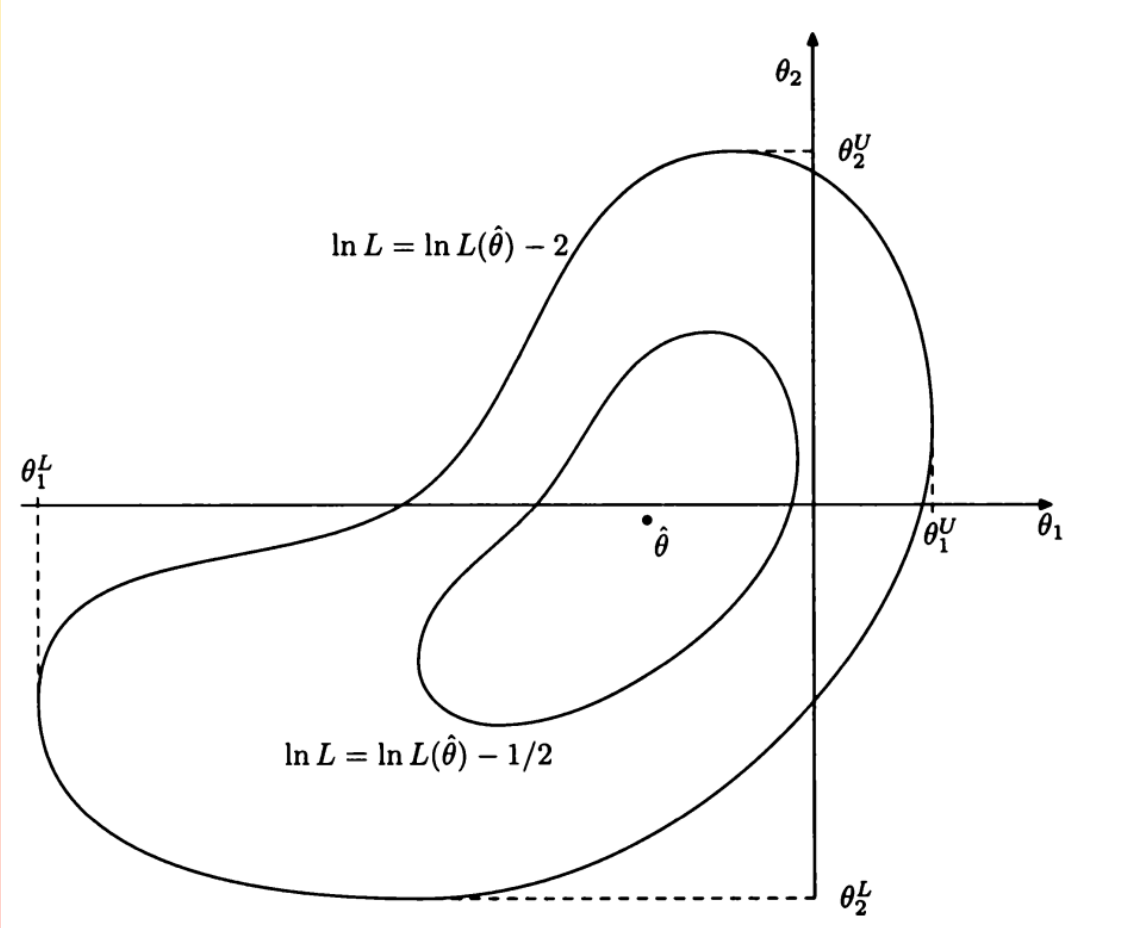
And in many dimensions...

- 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

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

- $\lambda = \frac{Z^2_{1-\beta}}{2}$
  - $\lambda = 1/2$ for $\beta = 0.683$ ("1σ")
  - $\lambda = 2$ for $\beta = 0.955$ ("2σ")



Plot from James, 2nd ed.

What if I have systematic uncertainties?

- 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*)
  - $\mu = 0$ (Standard Model only, no new particle), $\mu = 1$ (Standard Model + new particle)
  - the parameters representing uncertainties, *nuisance parameters* θ
- 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
- 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, $\lambda(\mu) \sim exp\left[-\frac{1}{2}\chi^2\right]\left(1+\mathcal{O}(\frac{1}{\sqrt{N}})\right)$ (Wilks Theorem, under some regularity conditions—likelihood continuous 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 and has a known form
    - Pivot quantity
  - Useful: you don't need to make toys in order to find out how is $\lambda(\mu)$ distributed!
- See talk by Paolo Francavilla for details on how we apply these notions in Higgs Measurements

Summary: how to extract an interval from the likelihood function

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

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

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

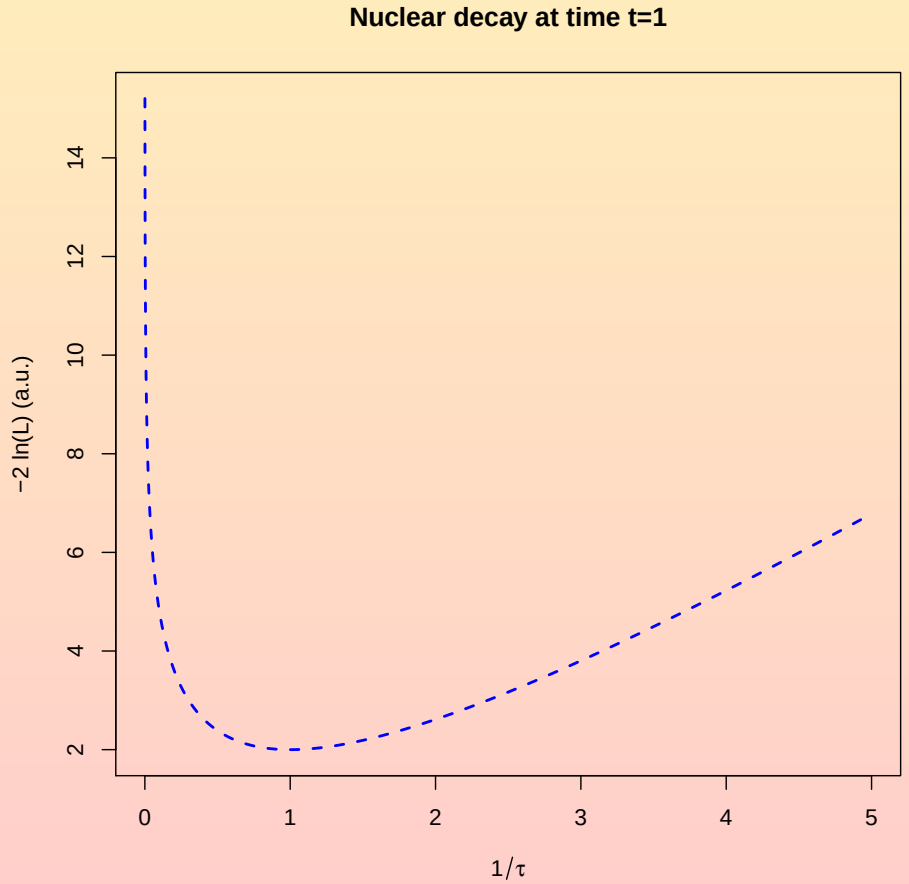
$$\Delta lnL := lnL(\theta') - lnL_{max} = -\frac{1}{2}$$

- 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 stimated 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 probabilty 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

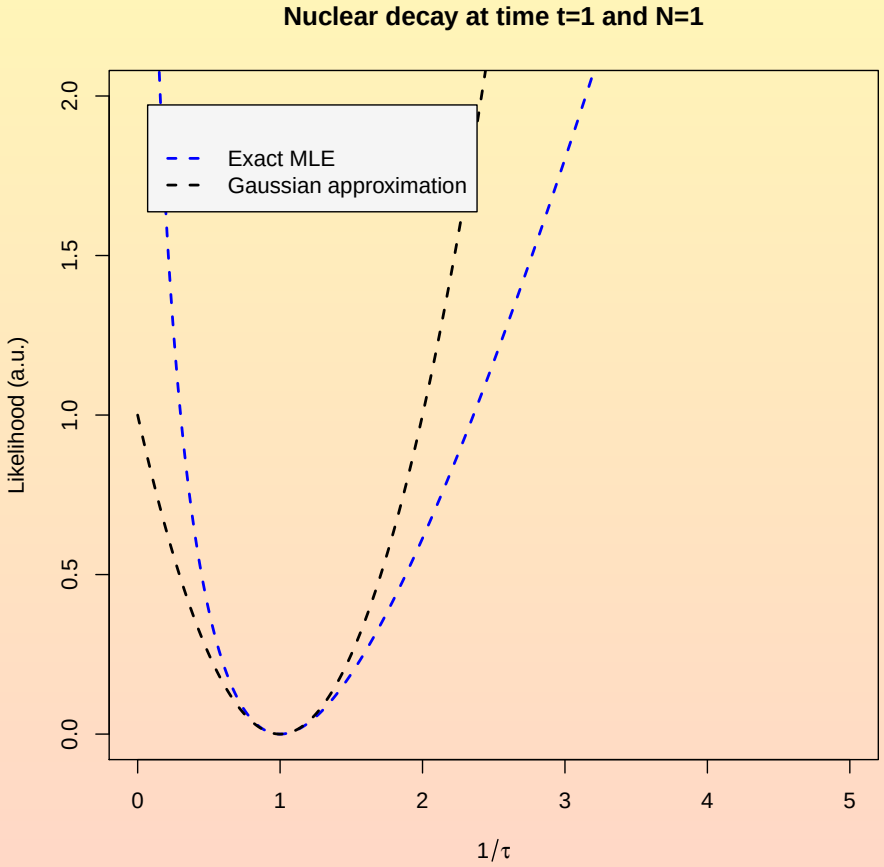
- How good is the approximation $L(\vec{x};\vec{\theta}) \propto exp\Big[-\frac{1}{2}(\vec{\theta}-\vec{\theta}_{MLE})^T H(\vec{\theta}-\vec{\theta}_{ML})\Big]$?
  - Here H is the information matrix $I(\vec{\theta})$
  - True only to $\mathcal{O}(\frac{1}{N})$
  - In these conditions, $V[\vec{\theta}_{ML}] \to \frac{1}{I(\vec{\theta}_{ML})}$
  - Intervals can be derived by crossings: $\Delta lnL = lnL(\theta') - lnL_{max} = k$
- Nuclear decay with half-life τ: N measurements in which $t_i = 1$, each measurement t sampled from the same p.d.f.

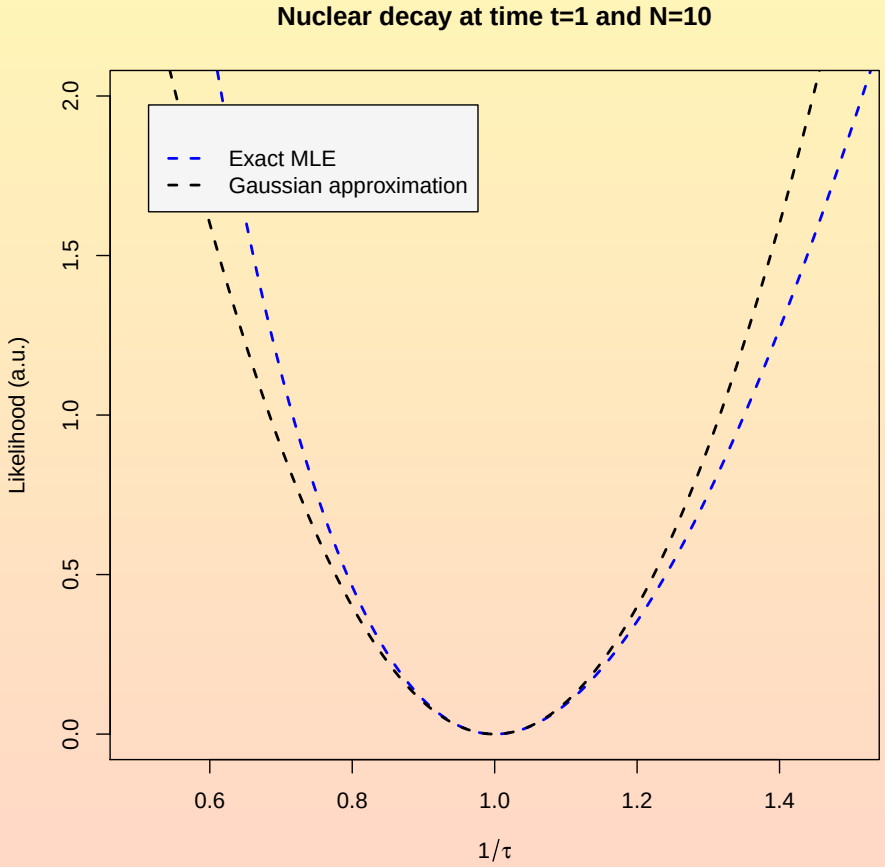
$$f(t;\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$
$$E[f] = \tau$$
$$V[f] = \tau^2$$



Non-normal likelihoods and Gaussian approximation — 2

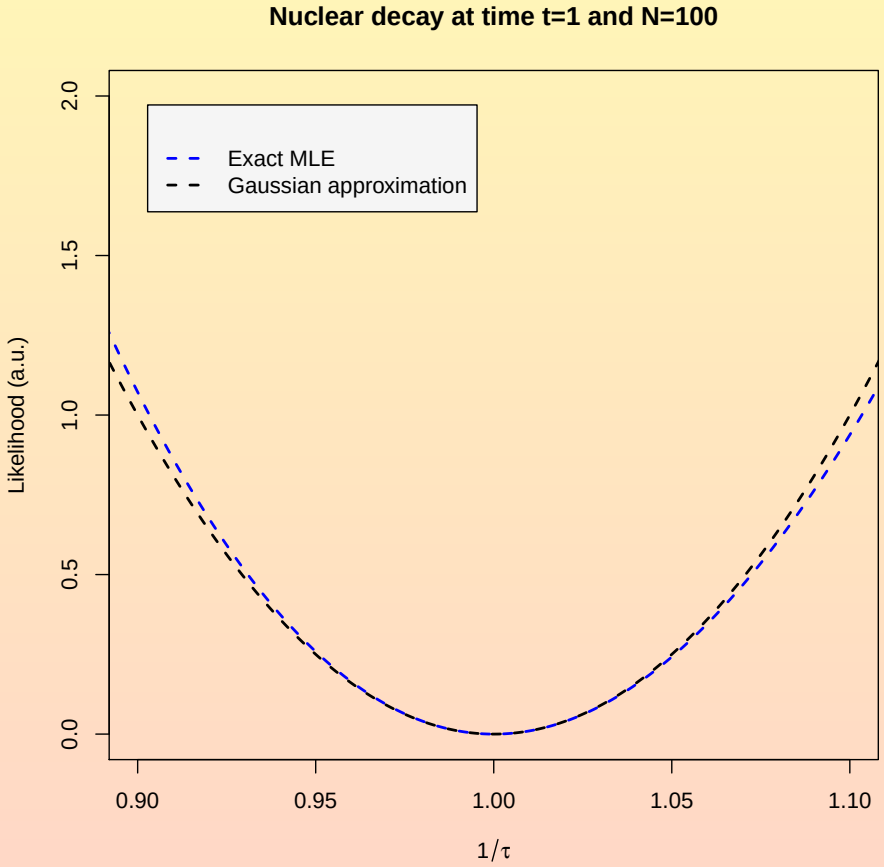
UCLouvain
Institut de recherche
en mathématique et physique

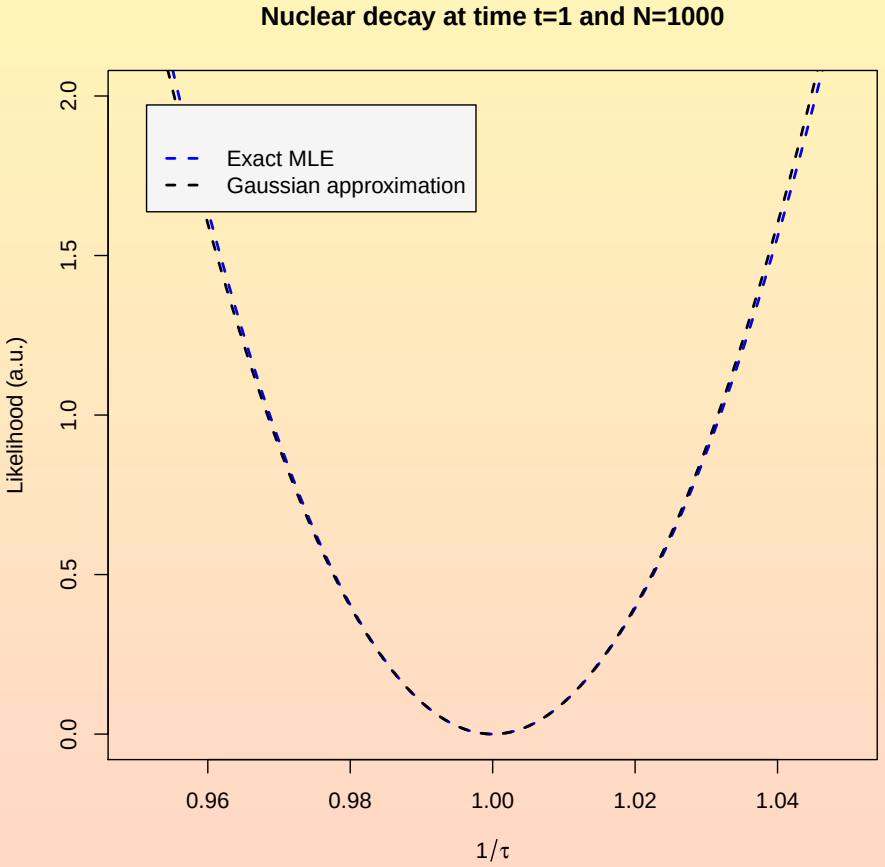




Non-normal likelihoods and Gaussian approximation — 3

UCLouvain
Institut de recherche
en mathématique et physique





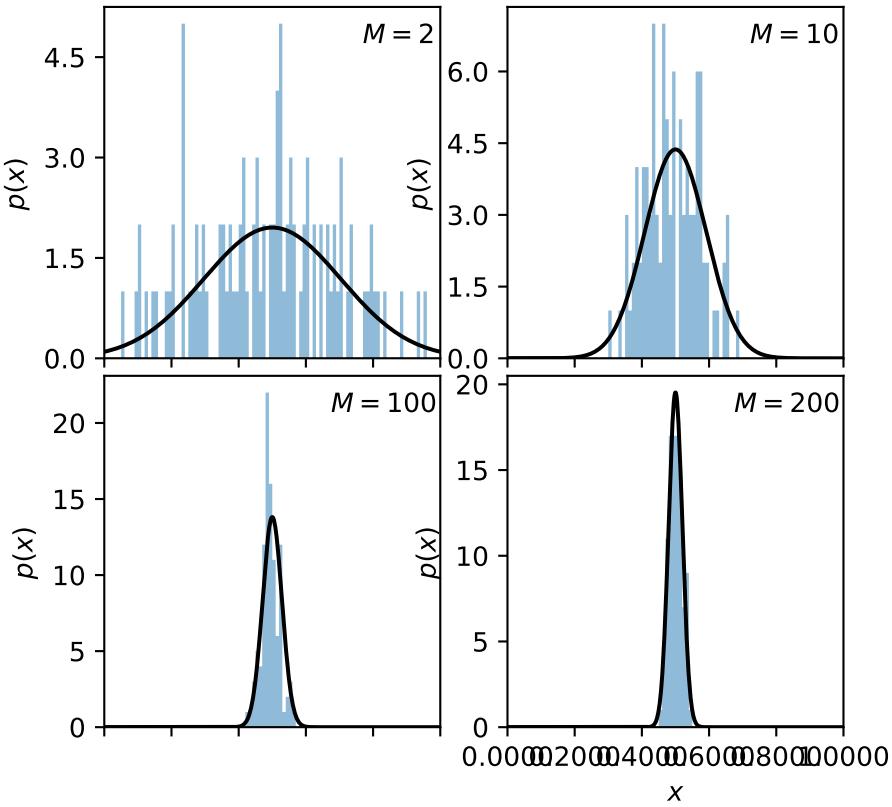
The Central Limit Theorem

- $L(\vec{x};\vec{\theta}) \to Gaus$ is direct consequence of the central limit theorem
- Take a set of measurements $\vec{x} = (x_i, ..., x_N)$ affected by experimental errors that results in uncertainties $\sigma_1, ..., \sigma_N$ (not necessarily equal among each other)
- For $M \to \infty$, the random variable built summing M measurements is gaussian-distributed:

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

- Valid for any p.d.f. $f(x,\vec{\theta})$ that is reasonably peaked around its expected value.
  - Large tails in the p.d.f. lead to non-gaussian tails in Q

- The condition $M \to \infty$ is reasonably valid if the sum is of many small contributions.
- M does not need to be very large for the approximation to be reasonably good!



Combination of measurements

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

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

- If we want to extract the p.d.f. of $g(\vec{x})$, we would normally use the jacobian of the transformation of f to g, but in this case we assumed $f(\vec{x})$ is unknown.
- We don't know f, but we can still write an expansion in series for it:

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

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

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

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

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

What about asymmetric uncertainties?

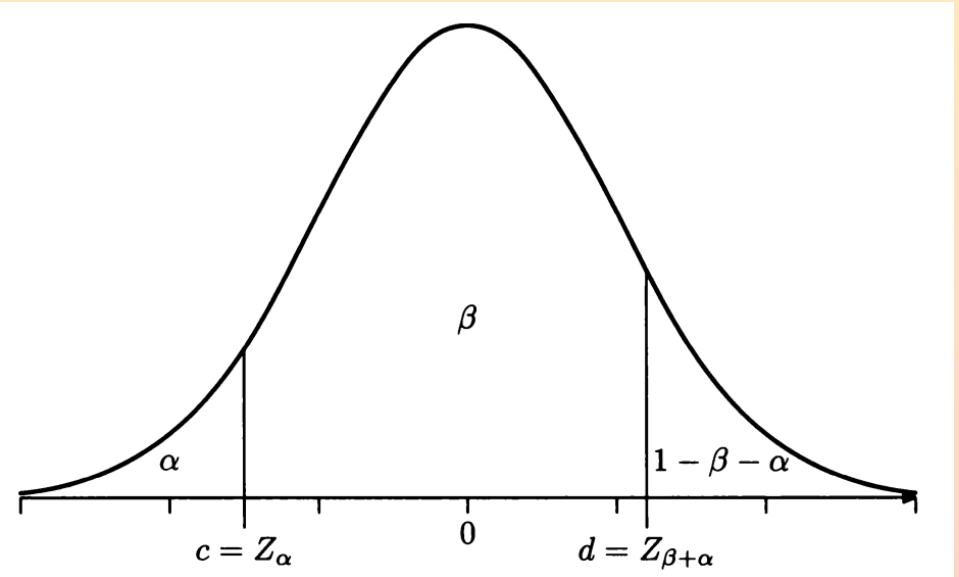
- Suppose I have two measurements affected by very asymmetric uncertainties
  - $N_1 = 0.99 \pm 0.03$
  - $N_2 = 1.10^{+0.05}_{-0.01}$
- For example, $N_{comb} = 2.09^{+0.06}_{-0.03}$. NO!!!
- The naïve quadrature of the two uncertainties is wrong!
  - The naïve combination is an expression of the Central Limit Theorem
  - The resulting combination is expected to be more symmetric than the measurements it originates from
  - Symmetric uncertainties usually assume a Gaussian approximation of the likelihood
  - Asymmetric uncertainties? One would need a study of the non-linearity (large biases might be introduced if ignoring this)
- Intrinsic difference between averaging and most probable value
  - Averaging results in average value and variance that propagate linearly
  - Taking the mode (essentially what MLE does) does not add up linearly!
- With asymmetric uncertainties from MLE fits, always combine the likelihoods (better in an individual simultaneous fit)
  - See talk by Paolo Francavilla on the details of combinations in the Higgs group

Confidence Intervals in nontrivial cases

Confidence intervals!

- Confidence interval for θ with probability content β
  - The range $\theta_a < \theta < \theta_b$ containing the true value θ_0 with probability β
  - The physicists sometimes improperly say the uncertainty on the parameter θ
- Given a p.d.f., the probability content is $\beta = P(a \leq X \leq b) = \int_a^b f(X|\theta)dX$
- If θ is unknown (as is usually the case), use auxiliary variable $Z = Z(X, \theta)$ with p.d.f. $g(Z)$ independent of θ
- If Z can be found, then the problem is to estimate interval $P(\theta_a \leq \theta_0 \leq \theta_b) = \beta$
  - Confidence interval
  - A method yielding an interval satisfying this property has coverage

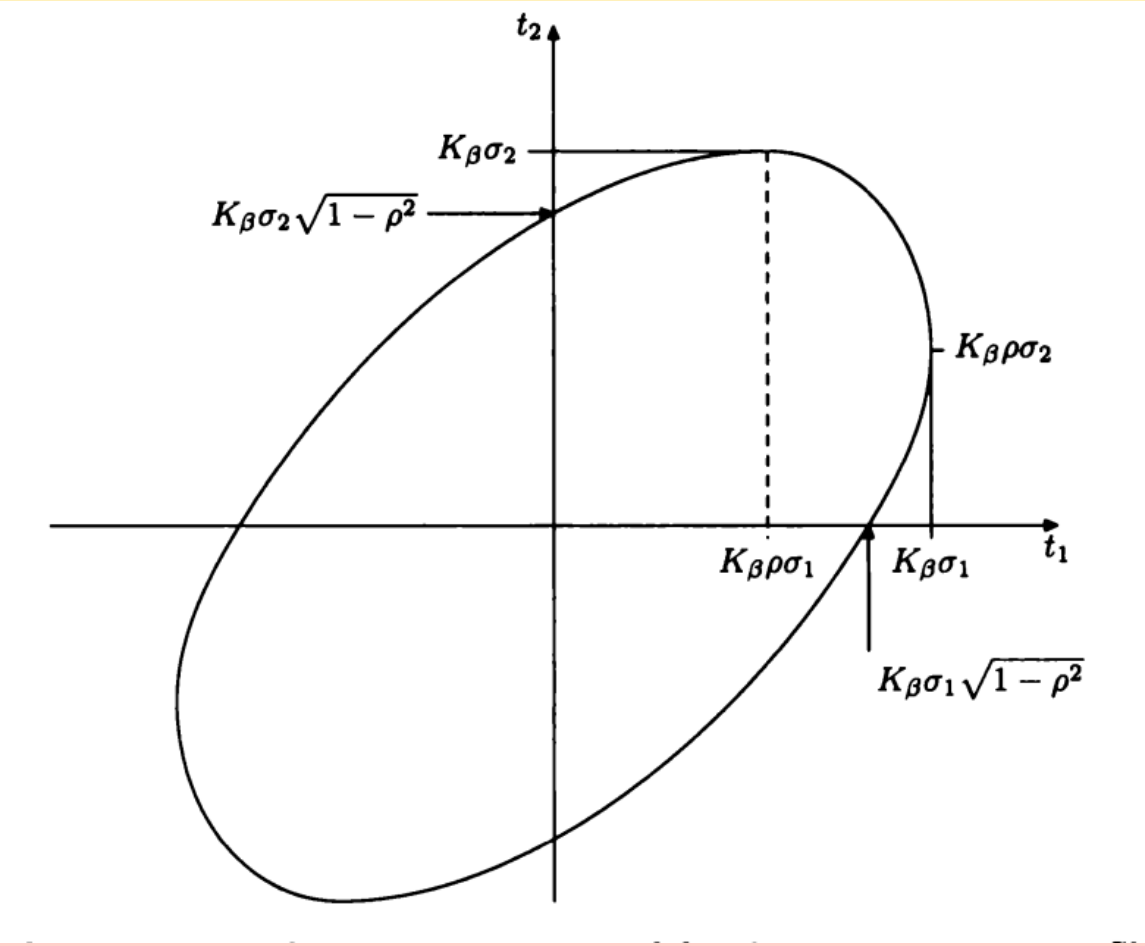
- Example: if $f(X|\theta) = N(\mu, \sigma^2)$ with unknown μ, σ, choose $Z = \frac{X-\mu}{\sigma}$
- Find $[c, d]$ in $\beta = P(c \leq Z \leq d) = \Phi(d) - \Phi(c)$ by finding $[Z_\alpha, Z_{\alpha+\beta}]$
- Infinite interval choices: here central interval $\alpha = \frac{1-\beta}{2}$



Plot from James, 2nd ed.

Confidence intervals in many dimensions

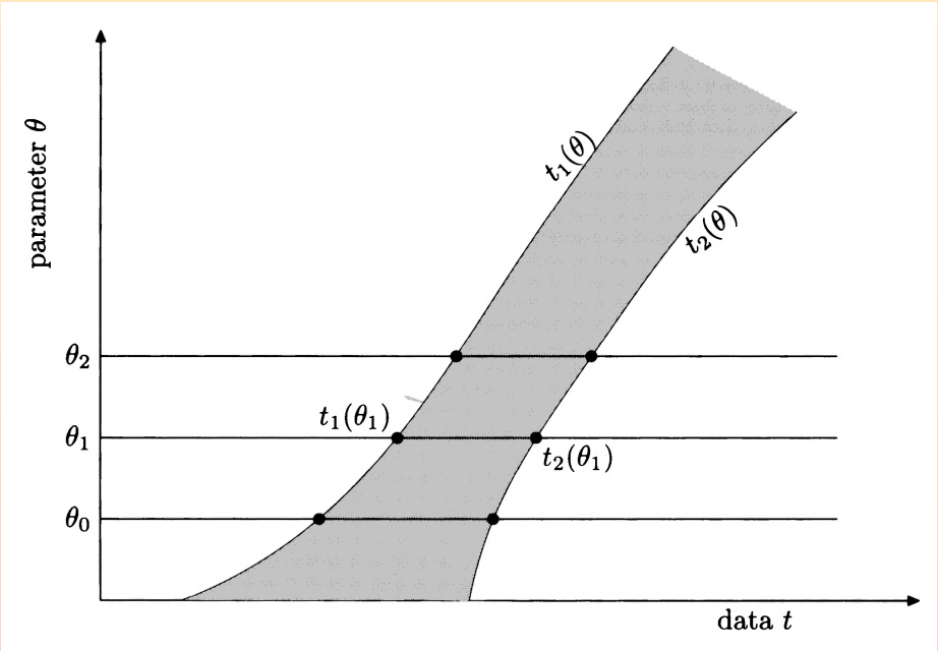
- Generalization to multidimensional $\boldsymbol{\theta}$ is immediate
- Probability statement concerns the whole $\boldsymbol{\theta}$, not the individual θ_i
- Shape of the ellipsoid governed by the correlation coefficient (or the mutual information) between the parameters
- Arbitrariety in the choice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

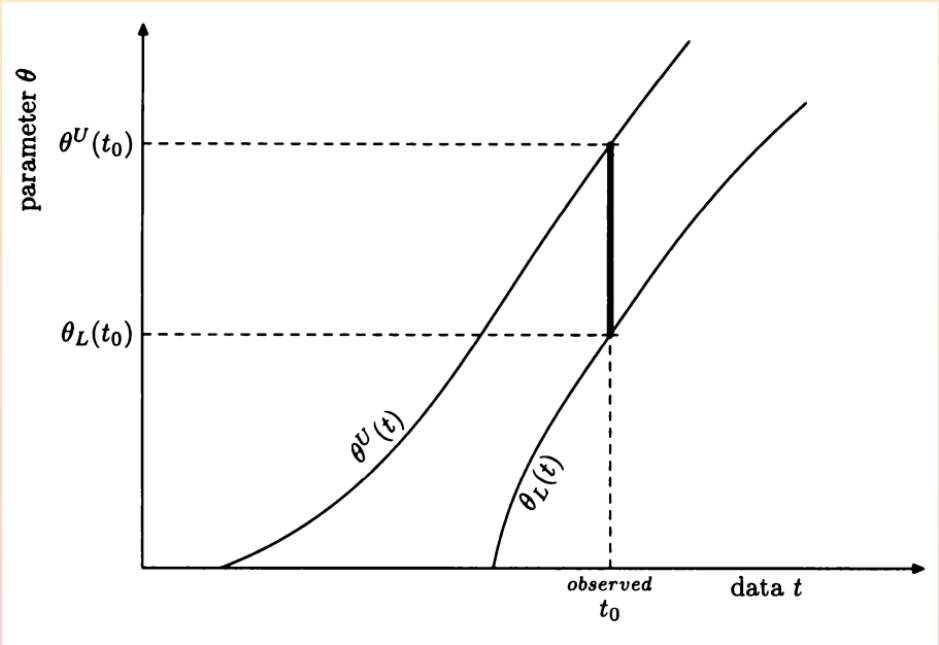


Plot from James, 2nd ed.

Confidence belts: the Neyman construction

- Unique solutions to finding confidence intervals are infinite
  - Central intervals, lower limits, upper limits, etc
- Let's suppose we have chosen a way
- Build horizontally: for each (hypothetical) value of θ, determine $t_1(\theta)$, $t_2(\theta)$ such that $\int_t 1^t 2P(t|\theta)dt = \beta$
- Read vertically: from the observed value t_0, determine $[\theta_L, \theta^U]$ by intersection
  - The resulting interval might be disconnected in severely non-linear cases
- Probability content statements to be seen in a frequentist way
  - Repeating many times the experiment, the fraction of $[\theta_L, \theta^U]$ containing θ_0 is β





Plot from James, 2nd ed.

Coverage

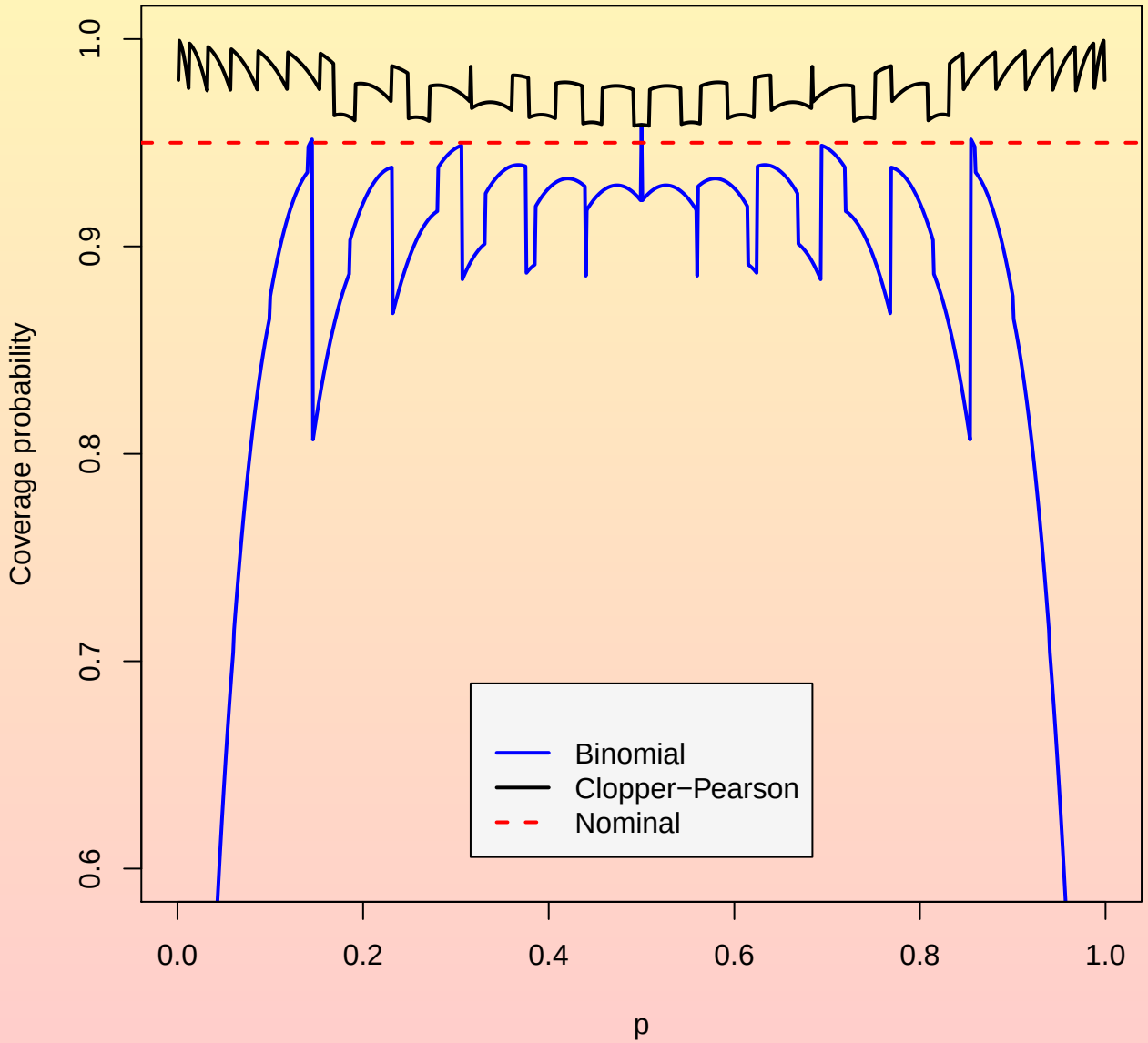
- Coverage probability of a method for calculating a confidence interval $[\theta_1, \theta_2]$: $P(\theta_1 \leq \theta_{true} \leq \theta_2)$
  - Fraction of times, over a set of (usually hypothetical) measurements, that the resulting interval covers the true value of the parameter
  - Can sample with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

Coverage: the binomial case

- For discrete distributions, the discreteness induces steps in the probability content of the interval
  - Continuous case: $P(a \leq X \leq b) = \int_a^b f(X|\theta)dX = \beta$
  - Discrete case: $P(a \leq X \leq b) = \sum_a^b f(X|\theta)dX \leq \beta$
- Binomial: find interval (r_{low}, r_{high}) such that $\sum_{r=r_{low}}^{r=r_{high}} \binom{r}{N} p^r (1-p)^{N-r} \leq 1-\alpha$
  - Also, $\binom{r}{N}$ computationally taxing for large r and N
  - Approximations are found in order to deal with the problem
- Gaussian approximation: $p \pm Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{N}}$
- Clopper Pearson: invert two single-tailed binomial tests, designed to overcover
  $\sum_{r=0}^{N} \binom{r}{N} p^n (1-p_{low})^{N-n} \leq \alpha/2$
  $\sum_{r=0}^{N} \binom{r}{N} p^r (1-p_{high})^{N-r} \leq \alpha/2$
  - Single-tailed $\rightarrow$ use $\alpha/2$ instead of α
- Calculating coverage of the two methods
  - For a given N, calculate intervals for various numbers of successes r, and plot the intervals of p as a function of r
  - Do a coverage test by using the procedure outlined in the previous slide
  - Draw the coverage probability as a function of p

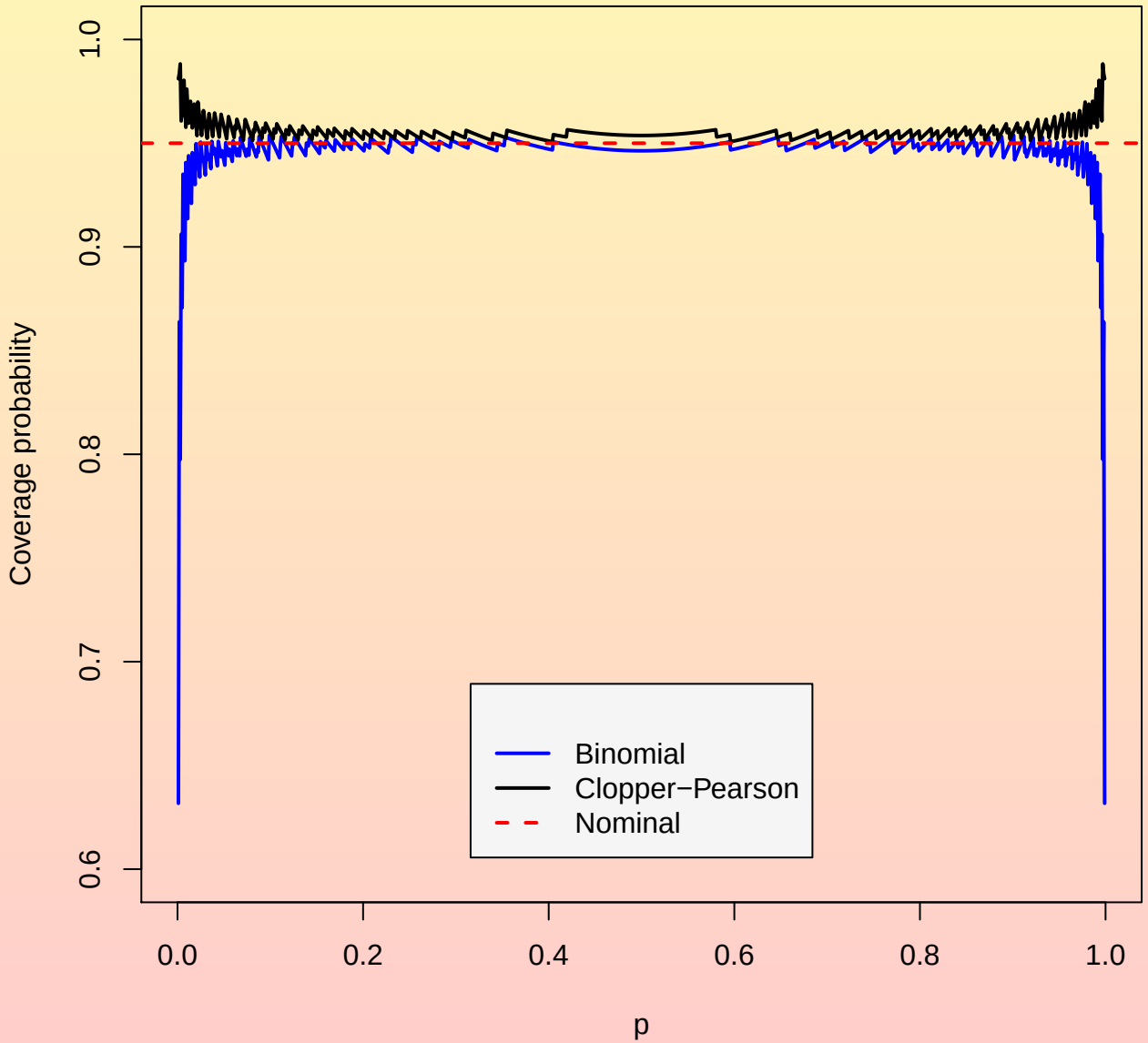
Coverage, $N = 20$

- Gaussian approximation bad for small sample sizes



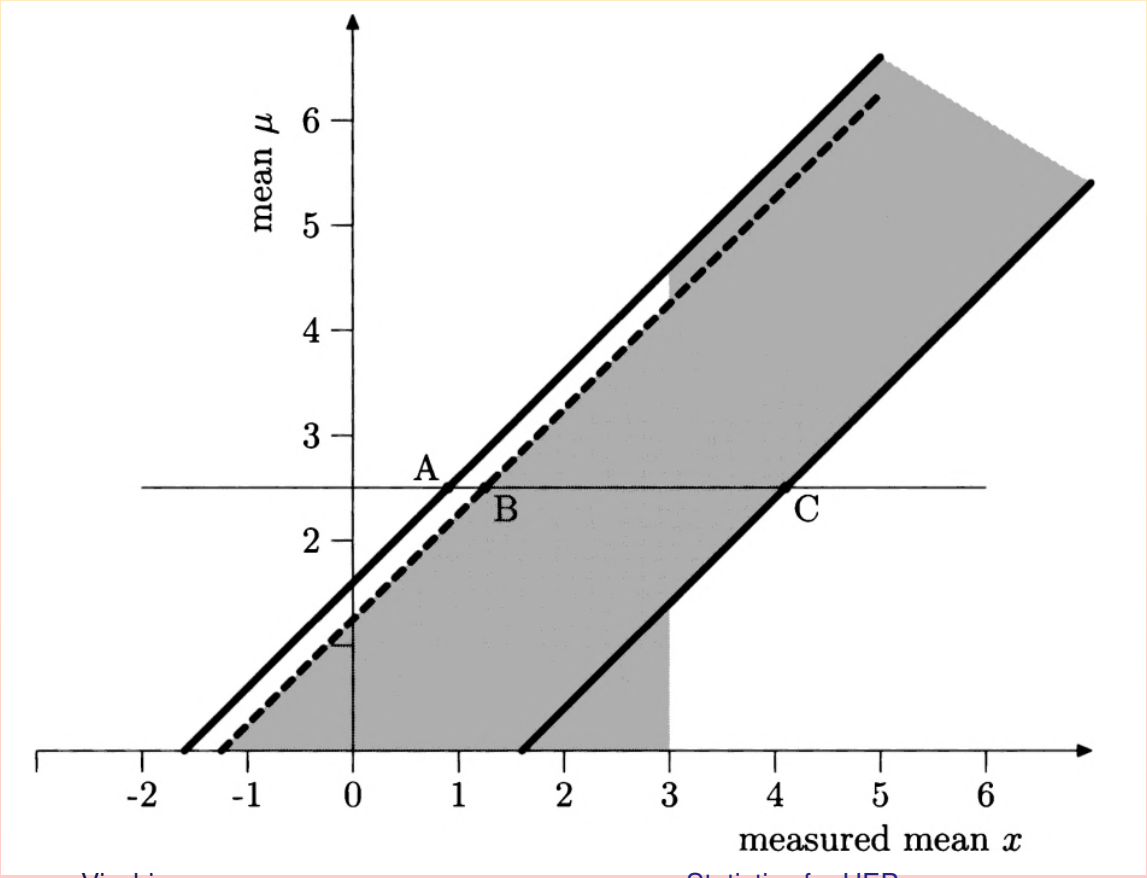
Coverage, $N = 1000$

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



Upper limits for non-negative parameters

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

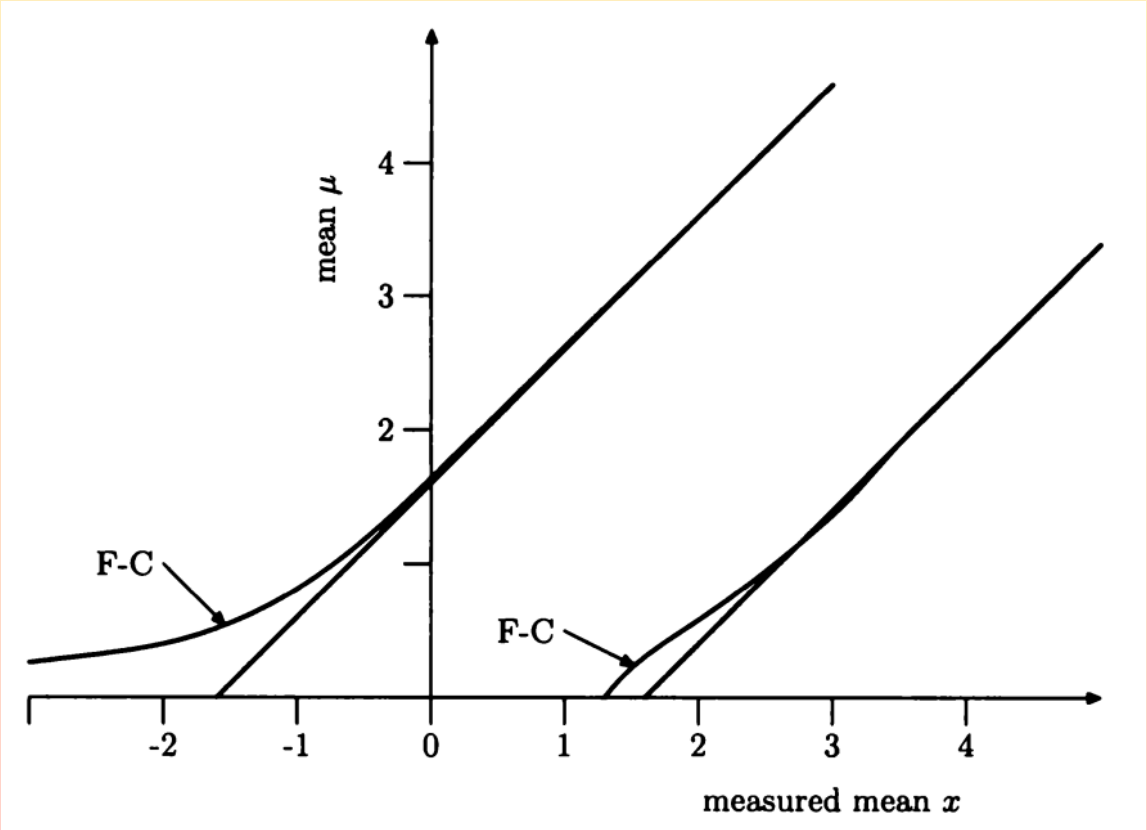


Plot from James, 2nd ed.

Unphysical values: Feldman-Cousins

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

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



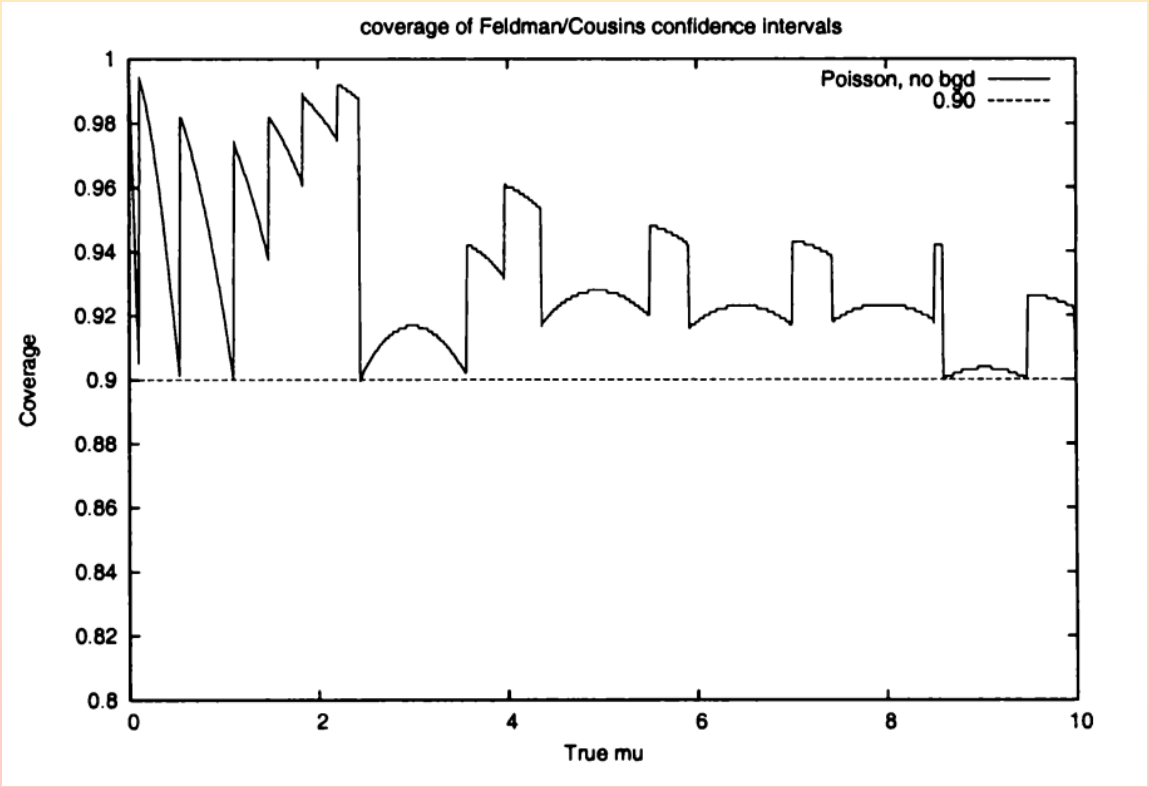
Plot from James, 2nd ed.

Feldman-Cousins in HEP

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

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

- Overcoverage larger for small values of μ (but less than other methods)



Plot from James, 2nd ed.

Bayesian intervals

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

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

- An interval $[\theta_L, \theta^U]$ with content β defined by $\int_{\theta_L}^{\theta^U} \pi(\theta|\boldsymbol{X}) d\theta = \beta$
- Bayesian statement! $P(\theta_L < \theta < \theta^U) = \beta$
  - Again, non unique
- Issues with empty intervals don't arise, though, because the prior takes care of defining the physical region in a natural way!
  - But this implies that central intervals cannot be seamlessly converted into upper limits
  - Need the notion of shortest interval
  - Issue of the metric (present in frequentist statistic) solved because here the preferred metric is defined by the prior

What is an hypothesis...

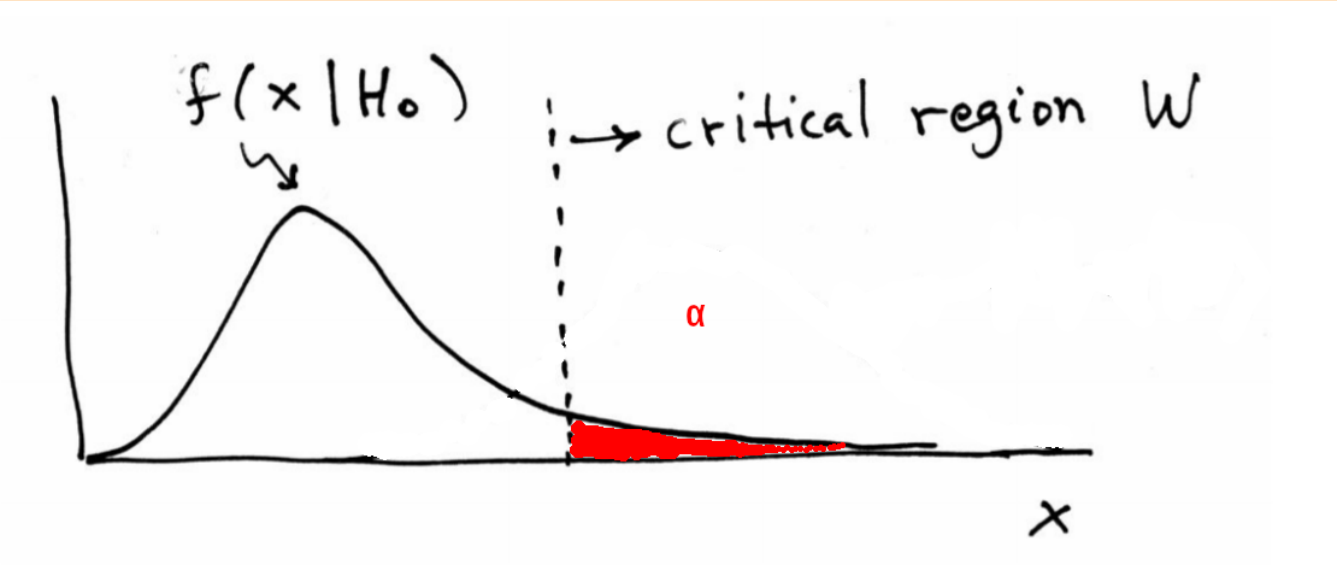
- Is our hypothesis compatible with the experimental data? By how much?
- Hypothesis: a complete rule that defines probabilities for data.
  - An hypothesis is simple if it is completely specified (or if each of its parameters is fixed to a single value)
  - An hypothesis is complex if it consists in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

...and how do we test it?

- The hypothesis can depend on a parameter
  - Technically, it consists in a family of hypotheses scanned by the parameter
  - We use the parameter as a proxy for the hypothesis, $P(\vec{x};\theta) := P(\vec{x};H(\theta))$.
- We are working in frequentist statistics, so there is no $P(H)$ enabling conversion from $P(\vec{x}|\theta)$ to $P(\theta|\vec{x})$.
- Statistical test
  - A statistical test is a proposition concerning the compatibility of H with the available data.
  - A binary test has only two possible outcomes: either accept or reject the hypothesis

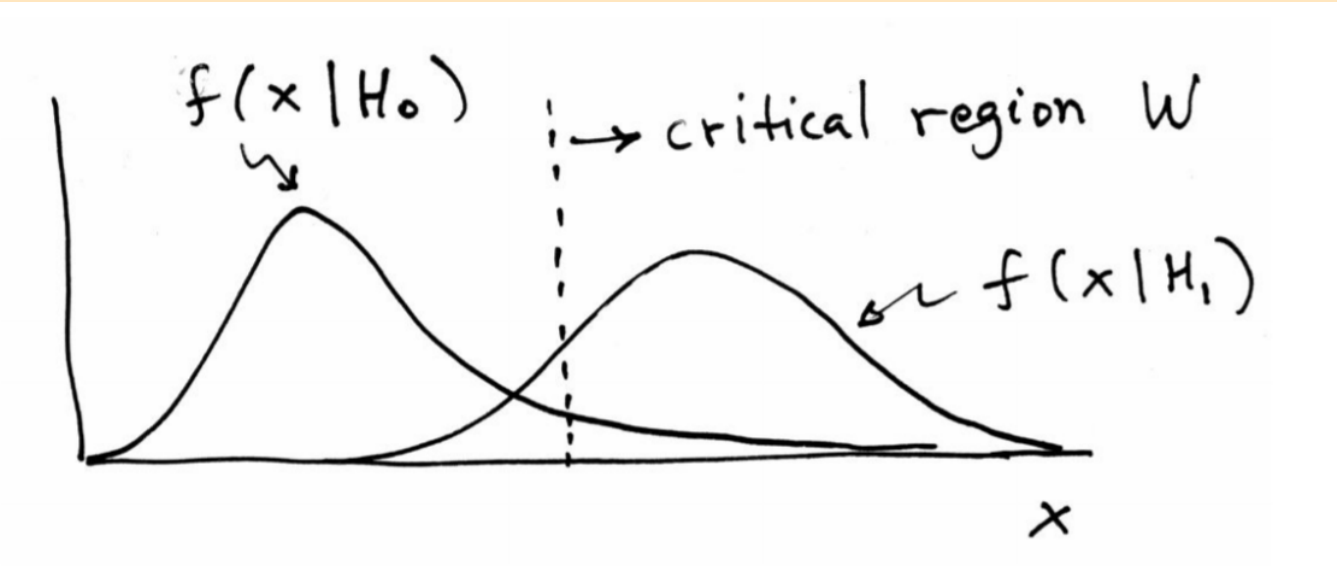
Testing the world as we know it...

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



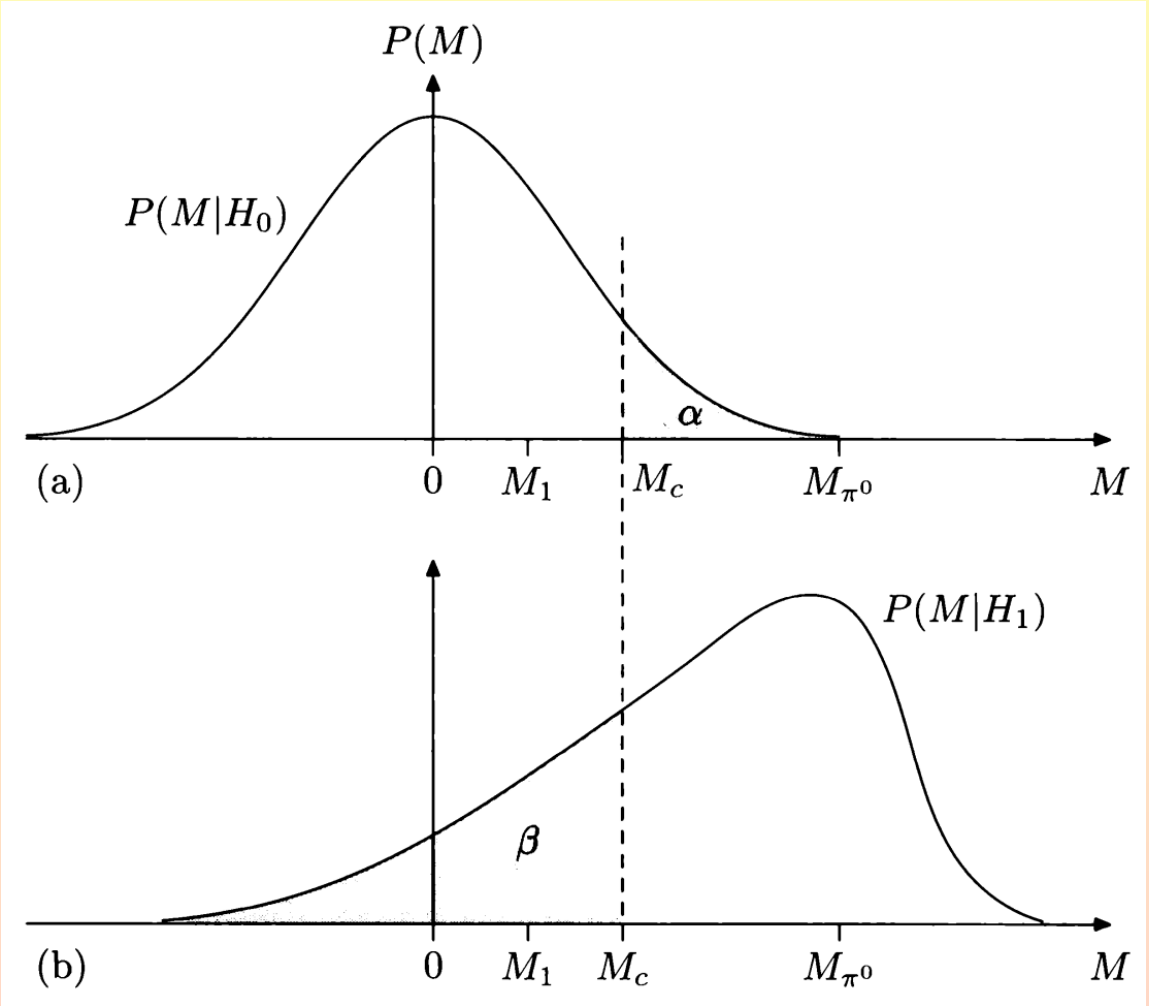
...while introducing some spice in it

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



A small example

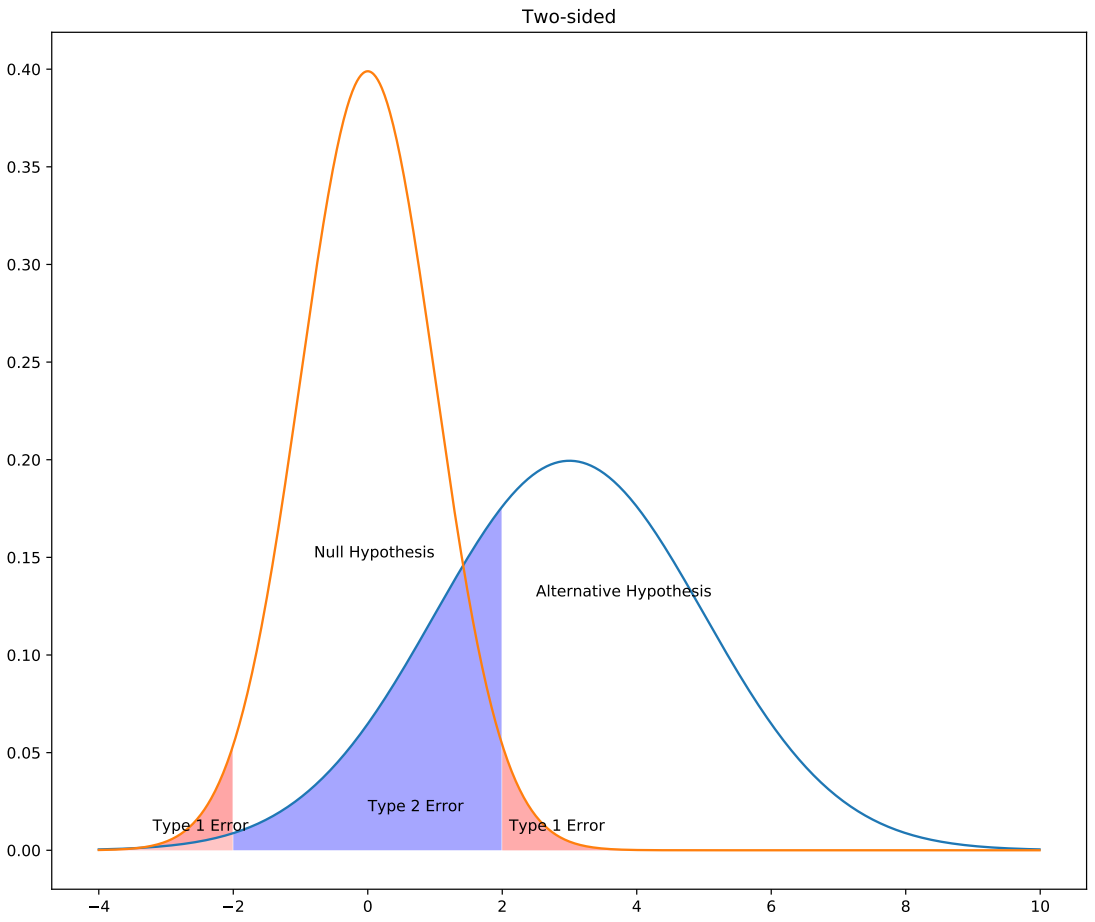
- H_0: $pp \to pp$ elastic scattering
- H_1: $pp \to pp\pi^0$
- Compute the missing mass M (as total rest energy of unseen particles)
- Under H_0, $M = 0$
- Under H_1, $M = 135\ MeV$



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

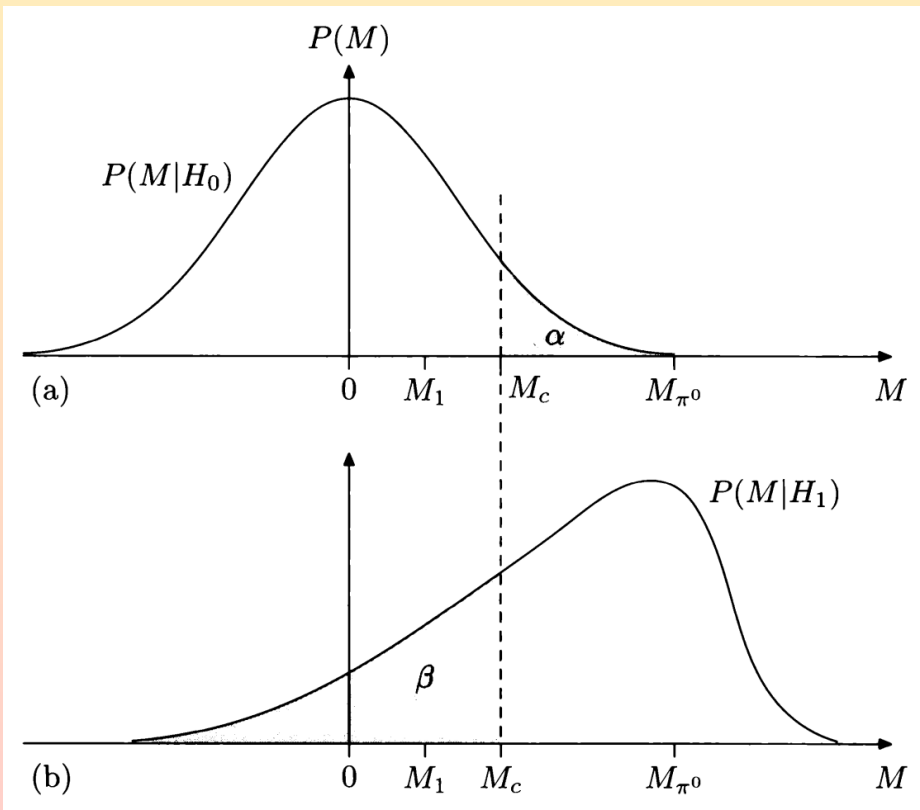
Plot from James, 2nd ed.

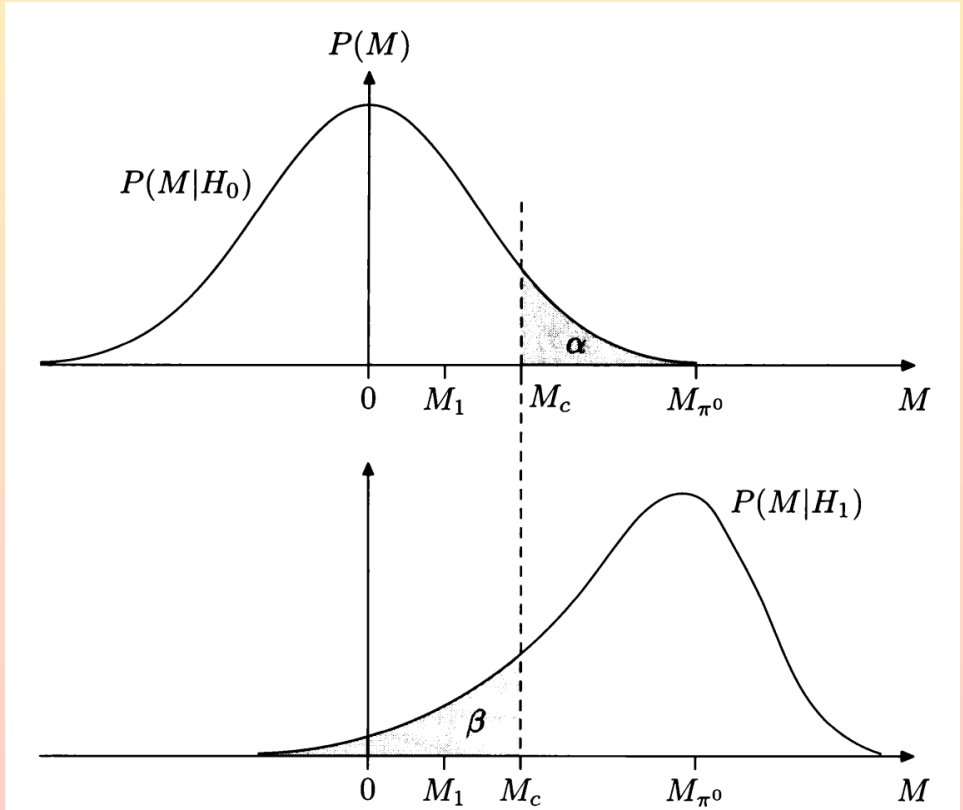
Works for both one-sided and two-sided tests



Power

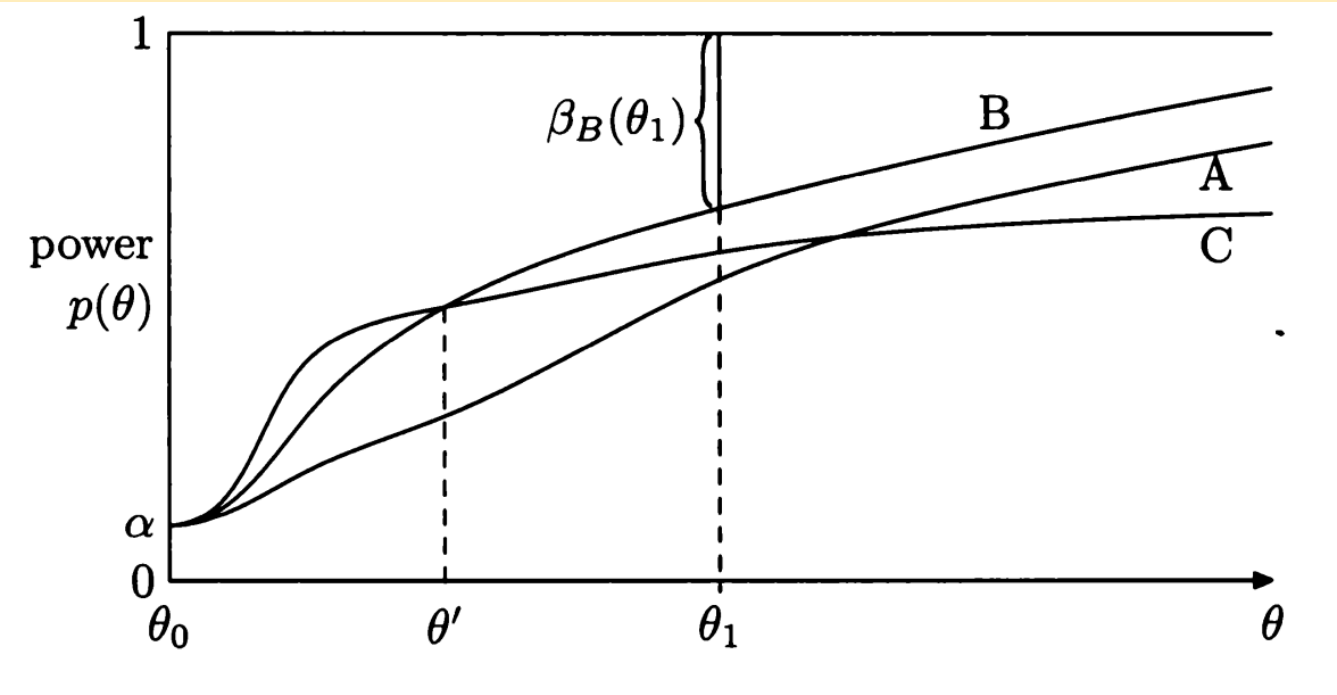
- The usefulness of the test depends on how well it discriminates against the alternative hypothesis
- The measure of usefulness is the *power of the test*
  - $P(X \in w|H_1) = 1 - \beta$
  - Power $(1 - \beta)$ is the probabiliity of X falling into the critical region if H_1 is true
  - $P(X \in W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





Plots from James, 2nd ed.

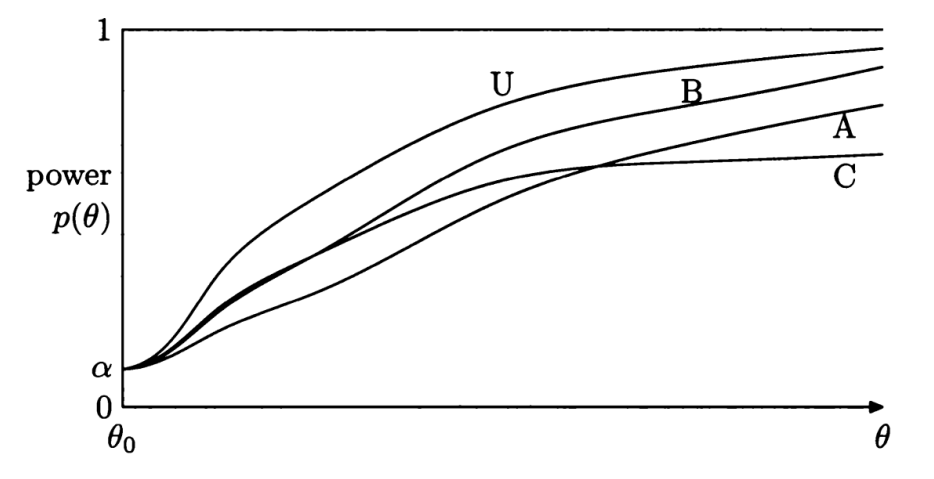
- For parametric (families of) hypotheses, the power depends on the parameter
  - $H_0 : \theta = \theta_0$
  - $H_1 : \theta = \theta_1$
  - Power: $p(\theta_1) = 1 - \beta$
- Generalize for all possible alternative hypotheses: $p(\theta) = 1 - \beta(\theta)$
  - For the null, $p(\theta_0) = 1 - \beta(\theta_0)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

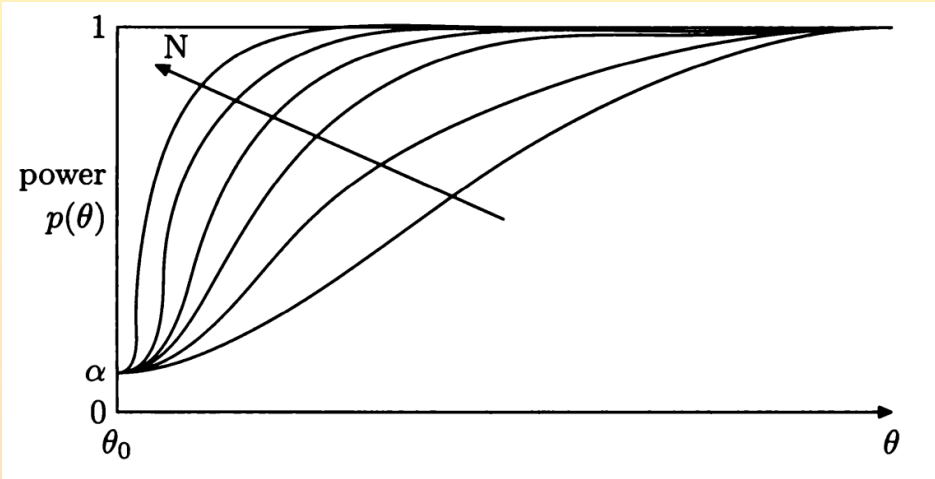


Plot from James, 2nd ed.

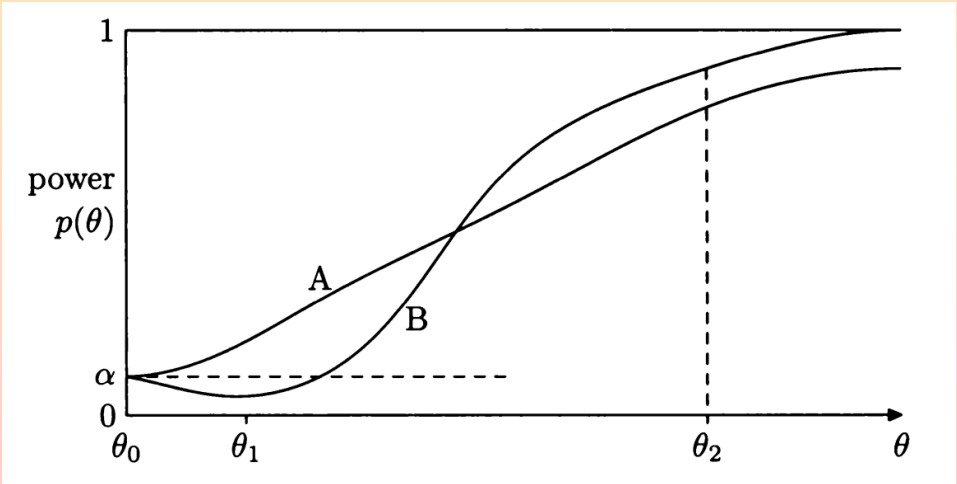
Properties of tests

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





- Biased test: $argmin(p(\theta)) \neq \theta_0$
- More likely to accept H_0 when it is false than when it is true
- Big no-no for θ_0 vs θ_1]
- Still useful (larger power) for θ_0 vs θ_2



Plot from James, 2nd ed.

[1] Robust: a test with low sensitivity to unimportant changes of the null hypothesis

Play with Type I (α) and Type II (β) errors freely

UCLouvain
Institut de recherche en mathématique et physique

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

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

d_0 : No choice is possible; results are ambiguous
d_1, ϕ^* : Family was $f(\mathbf{X}|\phi)$, with $\phi = \phi^*$
d_2, ψ^* : Family was $g(\mathbf{X}|\psi)$, with $\psi = \psi^*$.

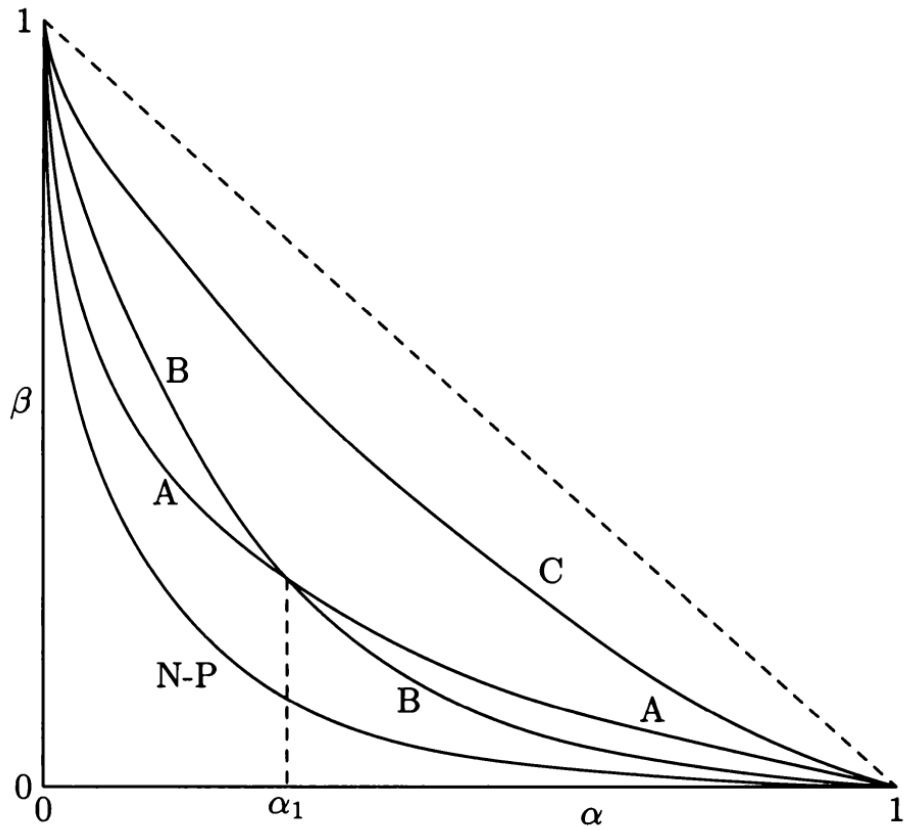


Table 10.4. A cost function.

Decisions	True state of nature	
	$\theta = \theta_1 = 1, \phi$	$\theta = \theta_2 = 0, \psi$
d_0	β_1	β_2
d_1, ϕ^*	$\alpha_1(\phi^* - \phi)^2$	γ_1
d_2, ψ^*	γ_2	$\alpha_2(\psi^* - \psi)^2$

Plot from James, 2nd ed.

Find the most powerful test

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

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

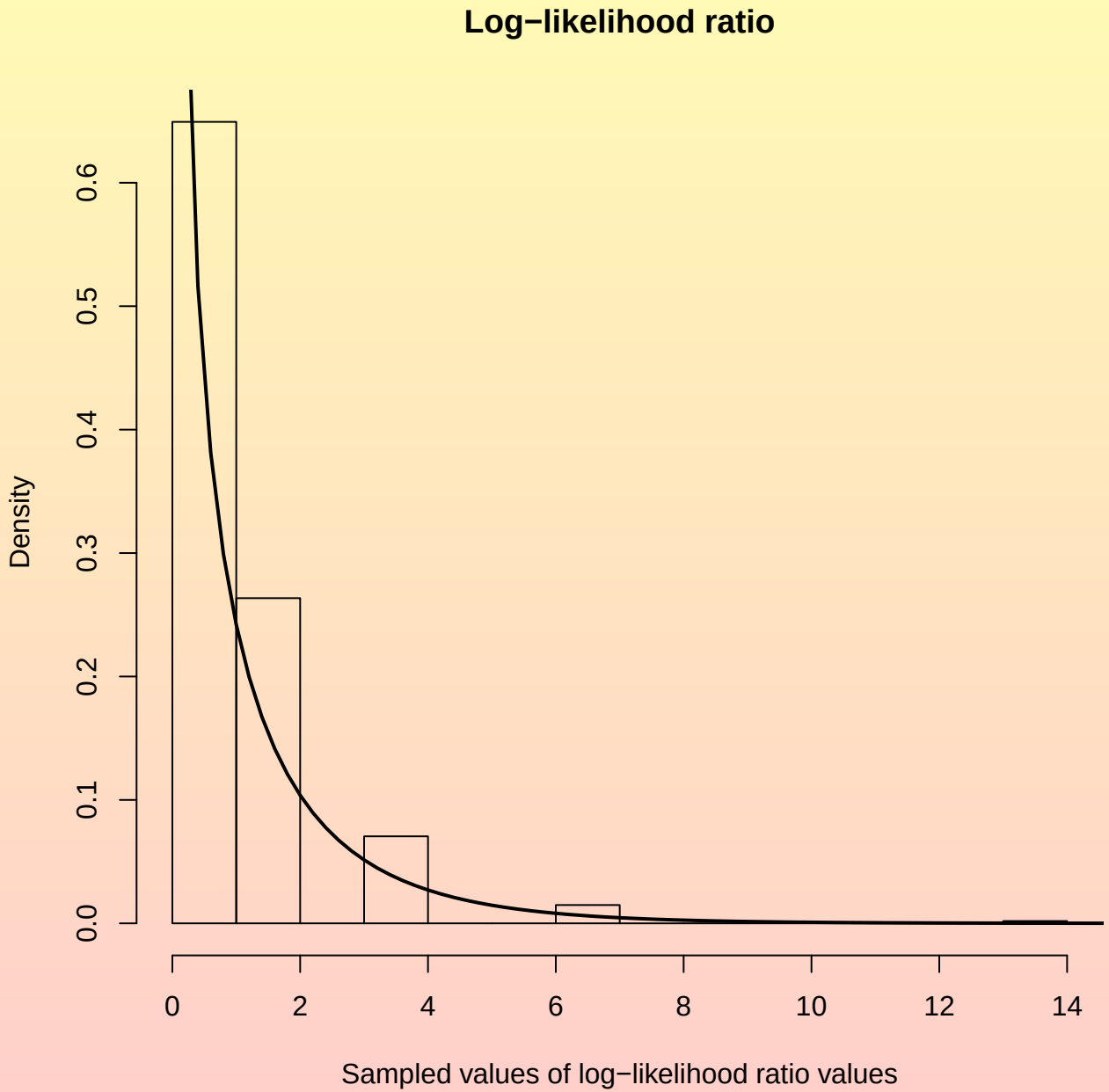
- The criterion, called Neyman-Pearson test is therefore
  - If $\ell(\mathbf{X}, \theta_0, \theta_1) > c_\alpha$ then choose H_1
  - If $\ell(\mathbf{X}, \theta_0, \theta_1) \leq c_\alpha$ then choose H_0
- The likelihood ratio must be calculable for any $\mathbf{X}$
  - The hypotheses must therefore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

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

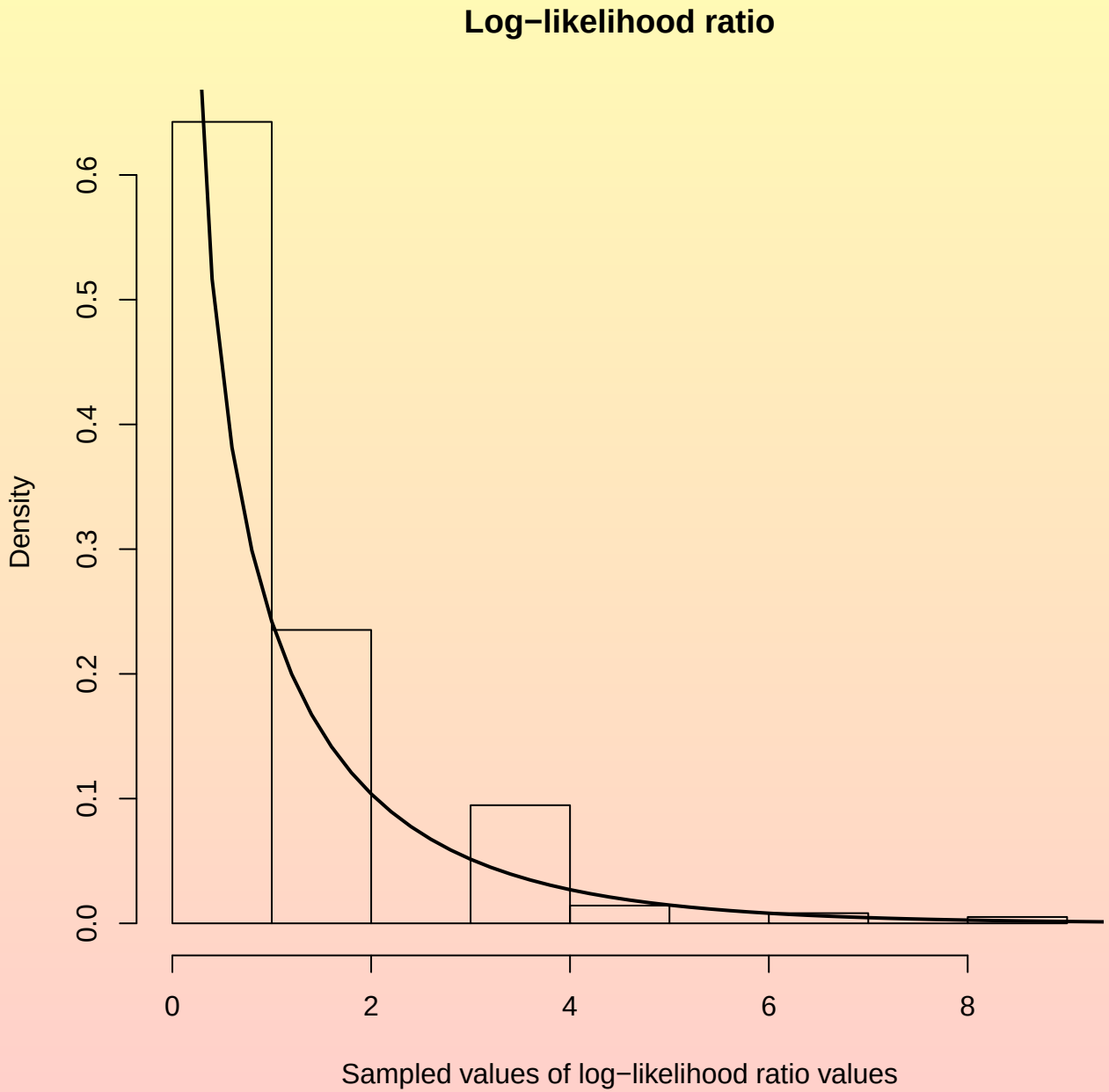
We can summarize in the

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

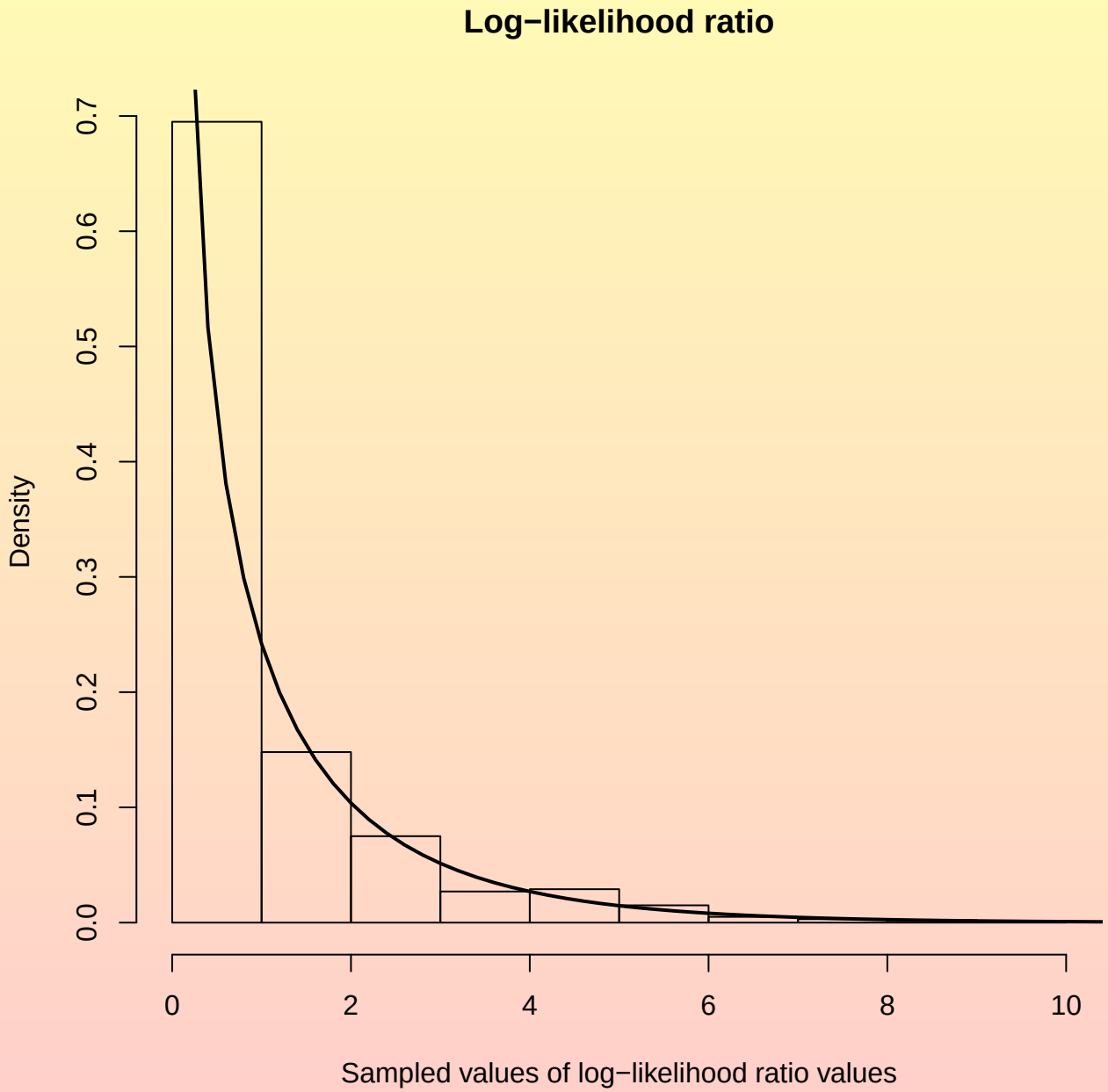
Verifying the Wilks theorem: N=2



Verifying the Wilks theorem: N=10



Verifying the Wilks theorem: N=100



Testing hypotheses near the boundary: Zech

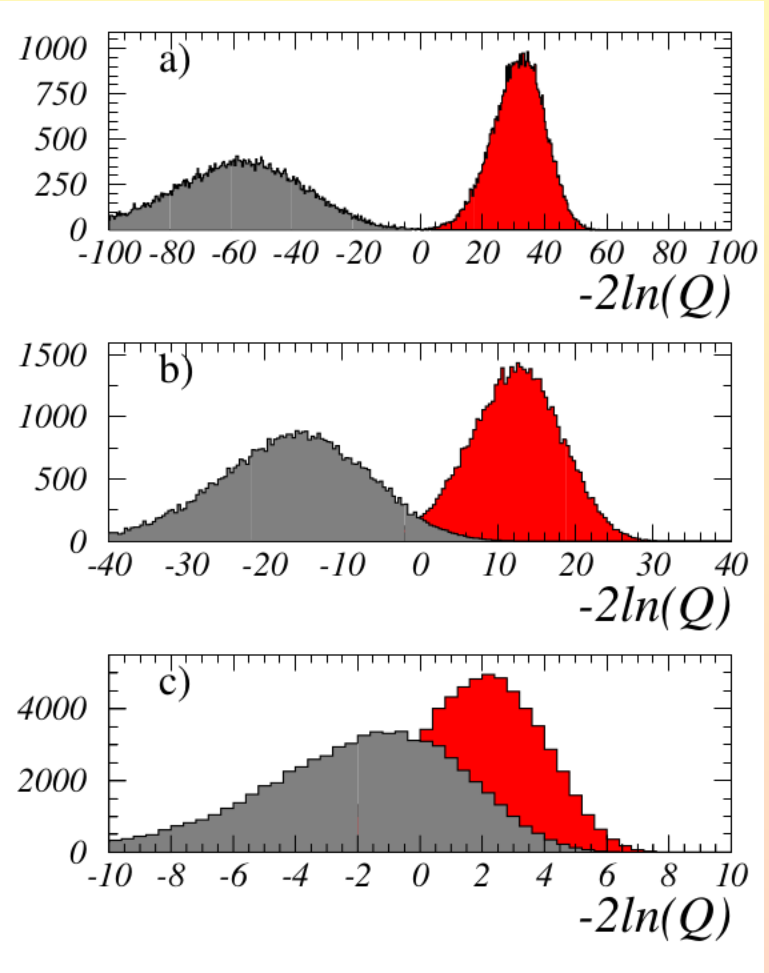
- Counting experiment: observe n events
- Assume they come from Poisson processes: $n \sim Pois(s+b)$, with known b
- Set limit on s given n_{obs}
- Exclude values of s for which $P(n \leq n_{obs}|s+b) \leq \alpha$ (guaranteed coverage $1-\alpha$)
- $b = 3$, $n_{obs} = 0$
  - Exclude $s+b \leq 3$ at 95%CL
  - Therefore excluding $s \leq 0$, i.e. **all** possible values of s (can't distinguish b-only from very-small-s)
- Zech: let's condition on $n_b \leq n_{obs}$ (n_b unknown number of background events)
  - For small n_b the procedure is more likely to undercover than when n_b is large, and the distribution of n_b is independent of s
  - $P(n \leq n_{obs}|n_b \leq n_{obs}, s+b) = \ldots = \frac{P(n \leq n_{obs}|s+b)}{P(n \leq n_{obs}|b)}$

Beyond coverage: CLs

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

Get your confidence levels right

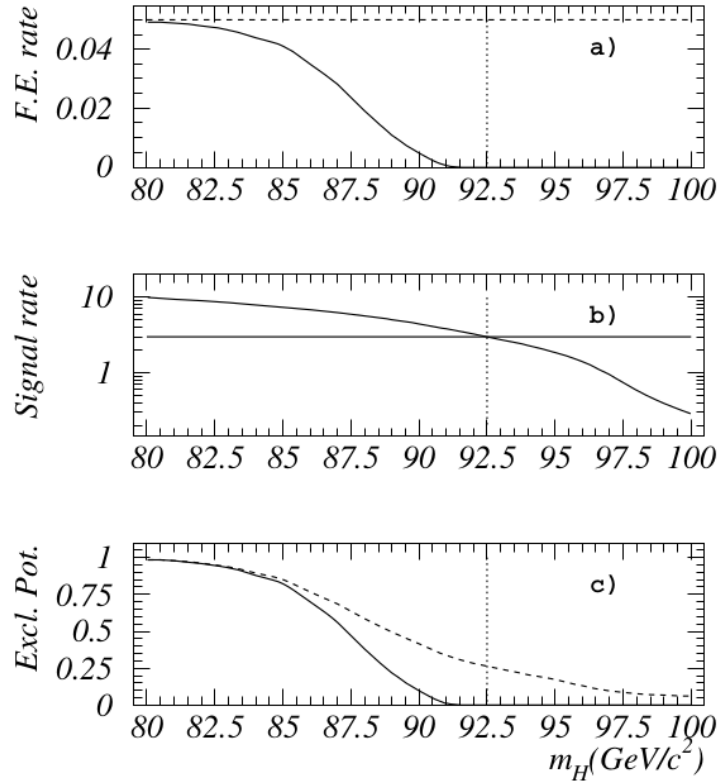
- Find a monotonic Q for increasing signal-like experiments (e.g. likelihood ratio)
- $CL_{s+b} = P_{s+b}(Q \leq Q_{obs})$
  - Small values imply poor compatibility with $S + B$ hypothesis, favouring B-only
- $CL_b = P_b(Q \leq Q_{obs})$
  - Large (close to 1) values imply poor compatibility with B-only,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



Plot from Read, CERN-open-2000-205

Avoid issues at low signal rates

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

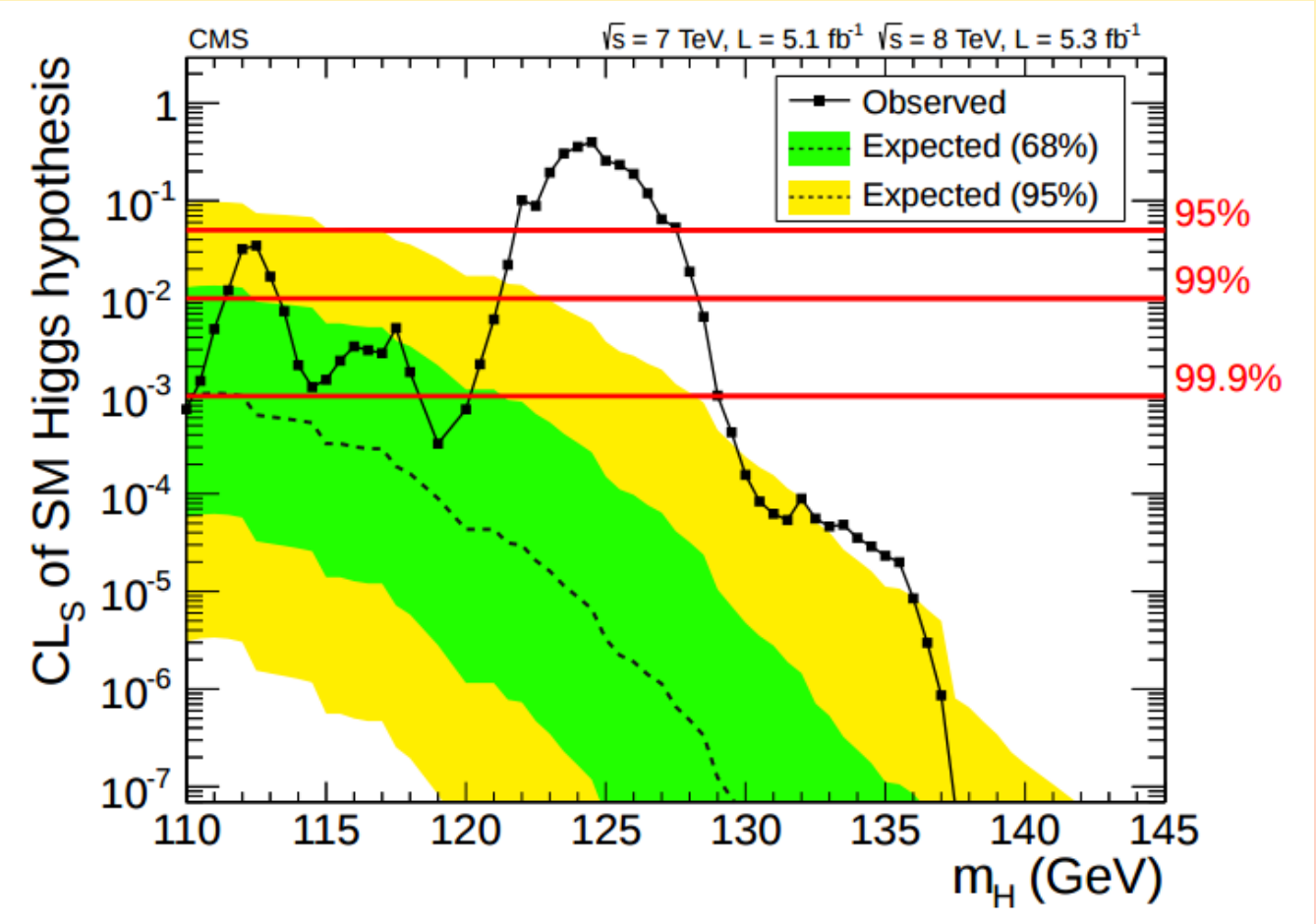


Dashed: CL_{s+b}
Solid: CL_s
$S < 3$: exclusion for a B-free search $\equiv 0$

Plot from Read, CERN-open-2000-205

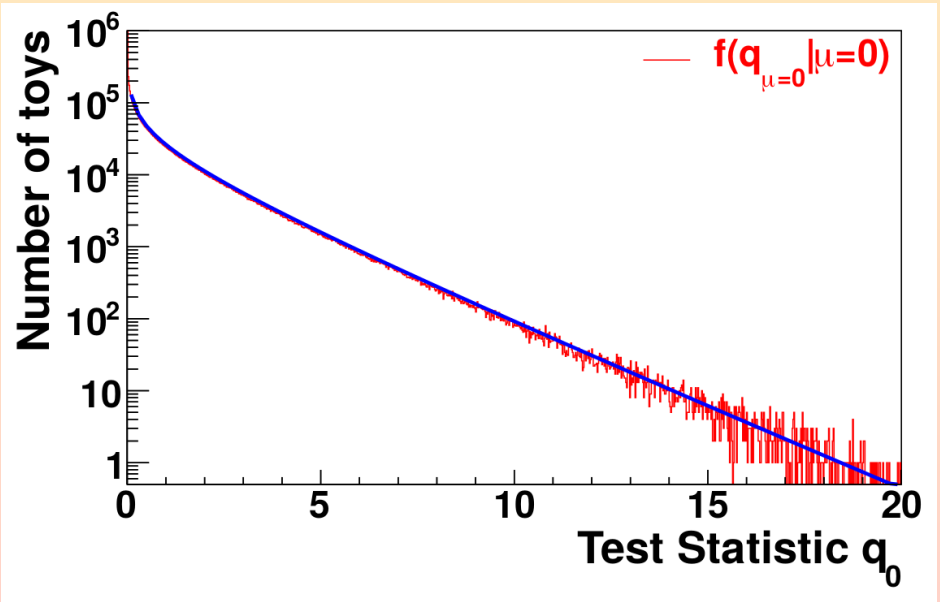
A practical example: Higgs discovery - 1

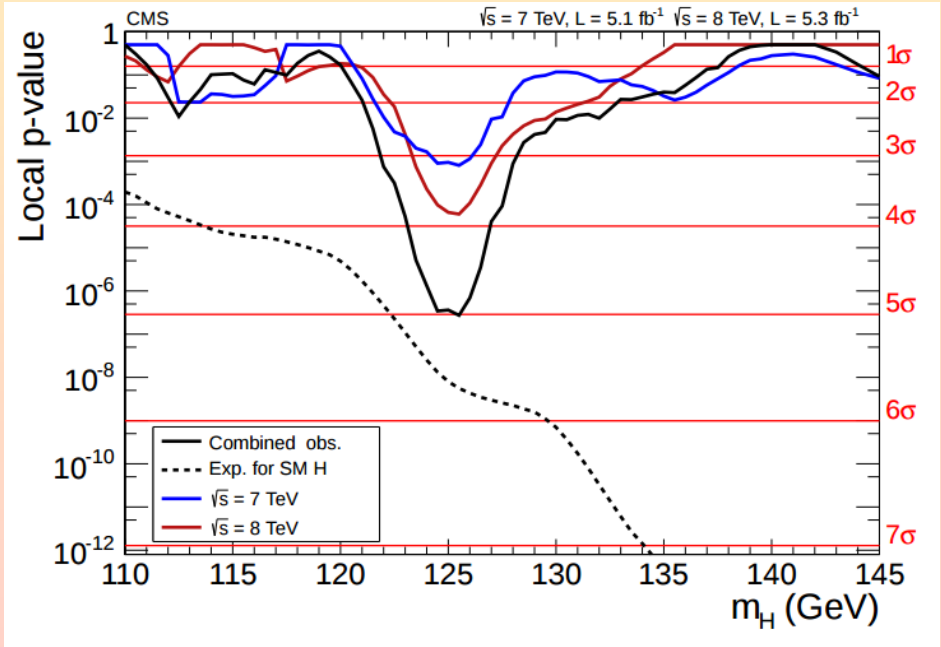
- Apply the CL_s method to each Higgs mass point
- Green/yellow bands indicate the $\pm 1\sigma$ and $\pm 2\sigma$ intervals for the expected values under B-only hypothesis
  - Obtained by taking the quantiles of the B-only hypothesis



Quantifying excesses

- Quantify the presence of the signal by using the background-only p-value
  - Probability that the background fluctuates yielding and excess as large or larger of the observed one
- For the mass of a resonance, $q_0 = -2log\frac{\mathcal{L}(data|0,\hat{\theta}_0)}{\mathcal{L}(data|\hat{\mu},\hat{\theta})}$, with $\hat{\mu} \geq 0$
  - Interested only in upwards fluctuation, accumulate downwards one to zero
- Use pseudo-data to generate background-only Poisson counts and nuisance parameters θ_0^{obs}
  - Use distribution to evaluate tail probability $p_0 = P(q_0 \leq q_0^{obs})$
  - Convert to one-sided Gaussian tail areas by inverting $p = \frac{1}{2}P_{\chi_1^2}(Z^2)$: *Significance*

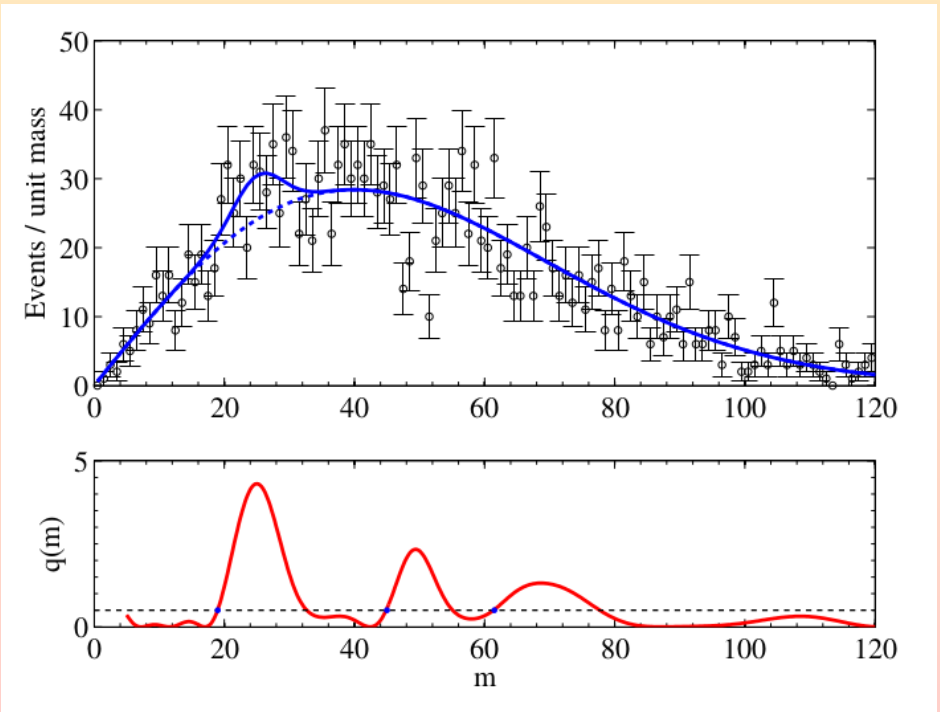


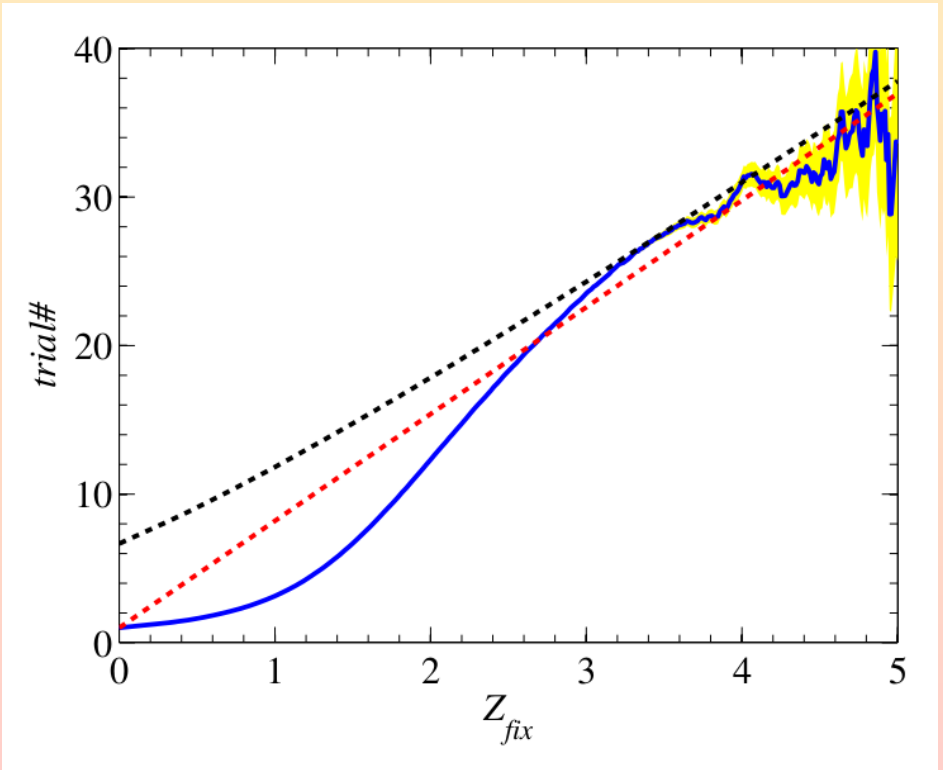


Plots from ATL-PHYS-PUB-2011-011 and from Higgs discovery

The Look-elsewhere effect

- Searching for a resonance X of arbitrary mass
  - $H_0 =$ no resonance, the mass of the resonance is not defined (Standard Model)
  - $H_1 = H(M \neq 0)$, but there are infinte possible values of M
- Wilks theorem not valid anymore, no unique test statistic encompassing every possible H_1
- Quantify the compatibility of an observation with the B-only hypothesis
  - $q_0(\hat{m_X}) = \max_{m_X} q_0(m_X)$
  - Write a global p-value as $p_b^{global} := P(q_0(\hat{m_X}) > u) \leq \langle N_u \rangle + \frac{1}{2} P_{\chi^2_1}(u)$
  - u fixed confidence level
  - Crossings computable using pseudo-data (toys)
  - Ratio of global and local p-value: trial factor
  - Asymptoticly linear in the number of search regions and in the fixed significance level





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

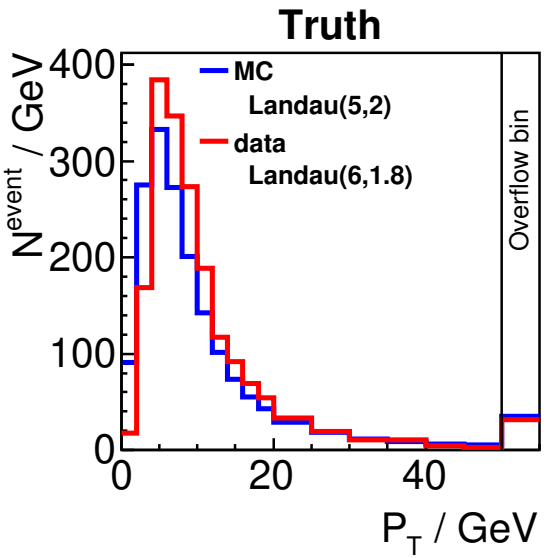
Measuring differential distributions

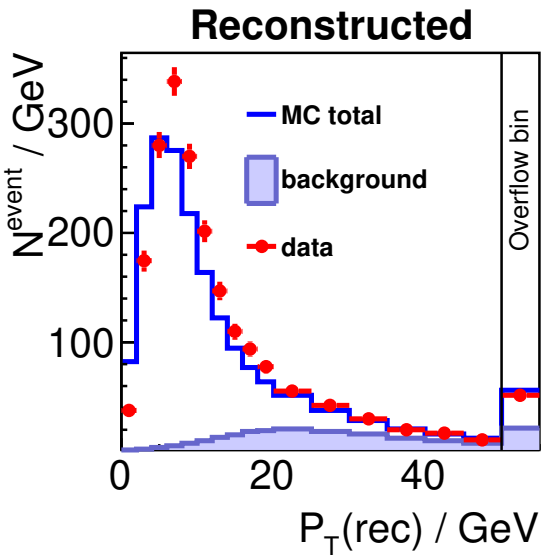
Unfolding: the problem

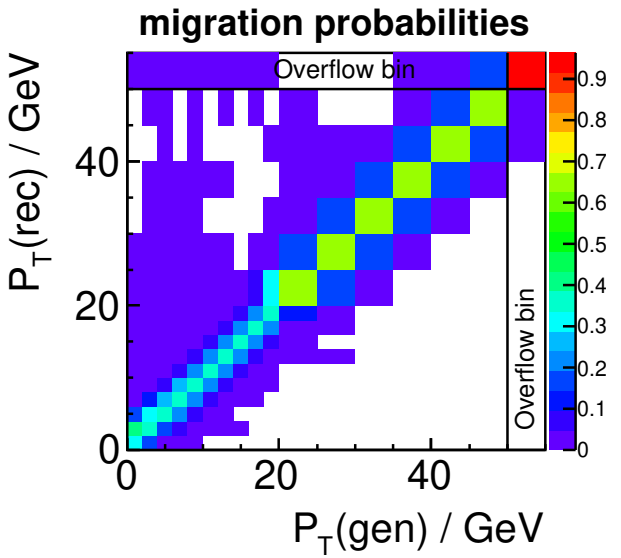
- Unfolding it's about how to invert a matrix that should not be inverted

$$\mathcal{L} = (\mathbf{y} - \mathbf{A}\mathbf{x})^T \mathbf{V}_{\mathbf{yy}} (\mathbf{y} - \mathbf{A}\mathbf{x}),$$

- Observations y, to be transformed in the theory space into x
  - Model the detector as a response matrix
  - Invert the response to convert experimental data to theory space distributions
  - Usually to compare with models in the theory space
- Dealing with model uncertainties is quite delicate, usually better to fold any new theory and make comparisons in the experimental data space
  - Data preservation and future reinterpretability of the result motivate unfolding even in the most delicate cases







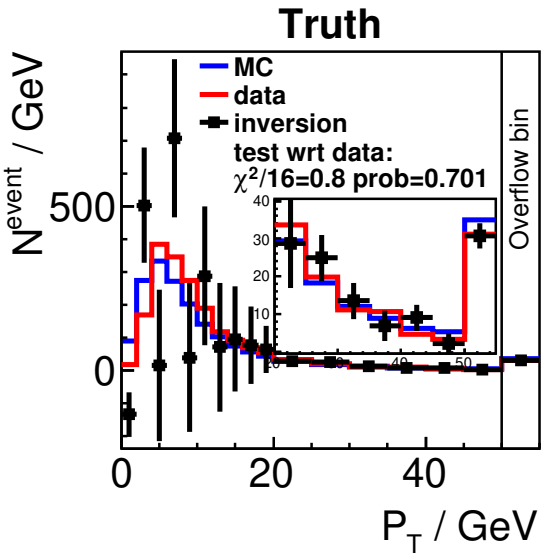
Plot from ArXiv:1611.01927

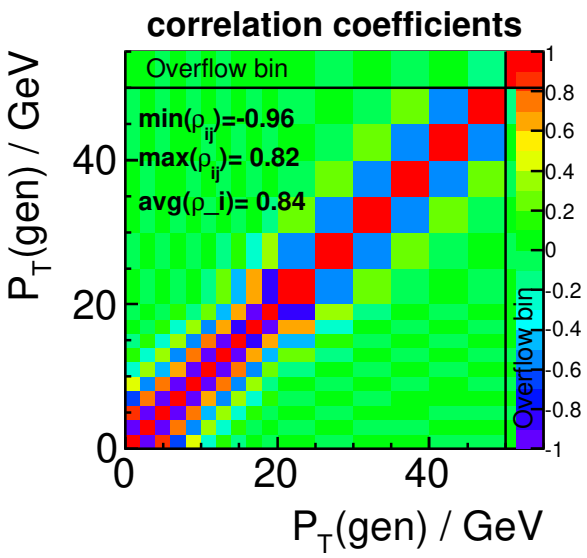
Unfolding: naïve solutions

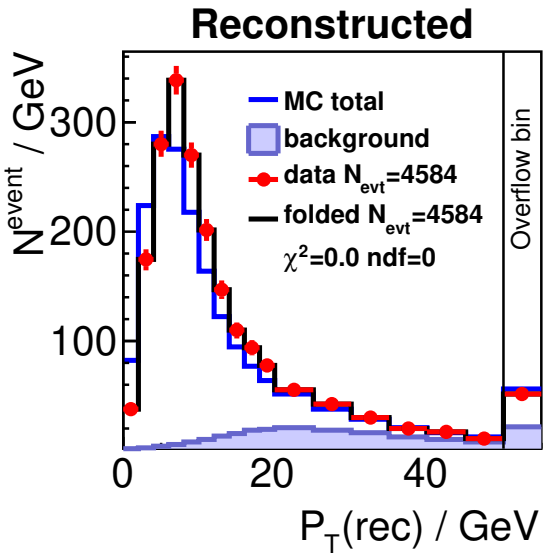
- Bin-by-bin correction factors $\hat{x}_i = (y_i - b_i)\frac{N_i^{\text{gen}}}{N_i^{\text{rec}}}$; disfavoured
  - Heavy biases due to the underlying MC truth
  - Yields the wrong normalization for the unfolded distribution
- Invert the response matrix $\hat{\boldsymbol{x}} = \boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{b})$
  - Only for square matrices, but always unbiased
  - Oscillation patterns (small determinants in matrix inversion)
  - Patterns also seen as large negative $\rho_{ij} \sim -1$ near diagonal
  - Result is correct within uncertainty envelope given by $\boldsymbol{V}_{xx}$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

determinant



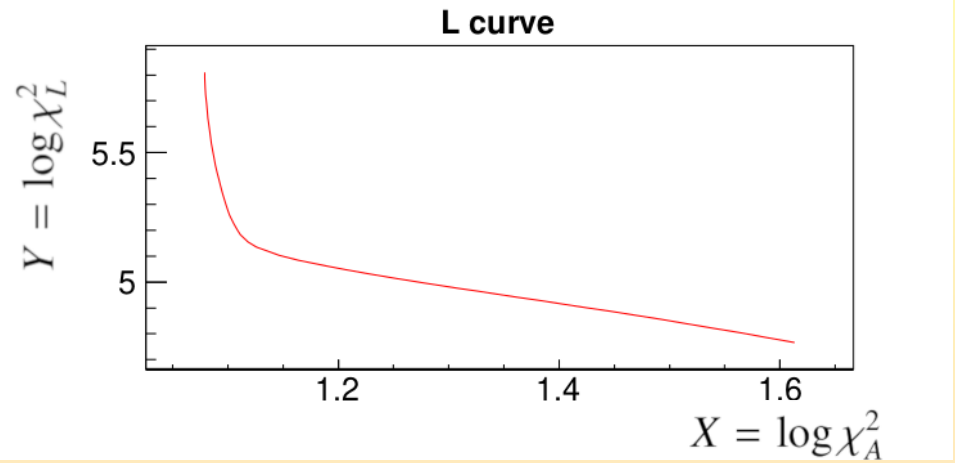




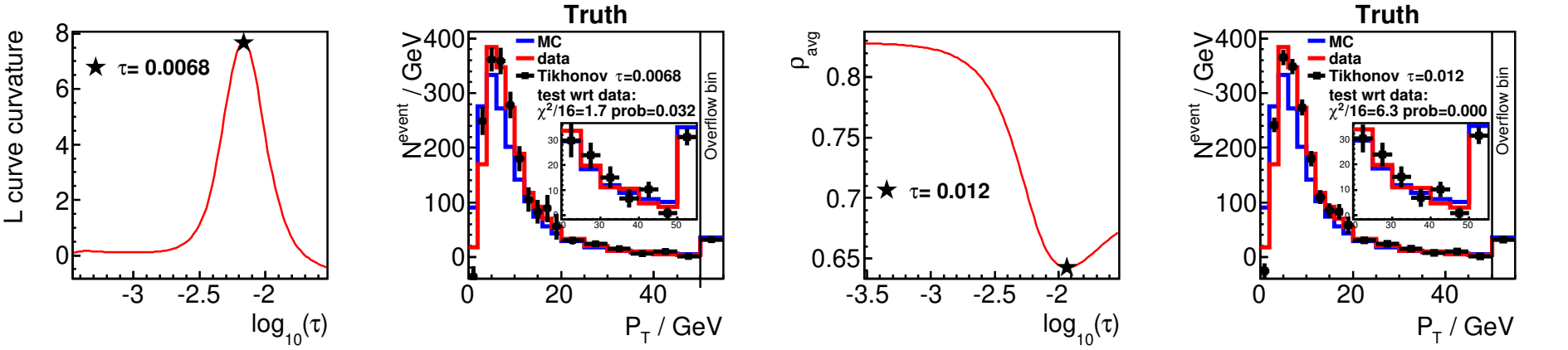
Cartoon from https://www.mathsisfun.com/algebra/matrix-inverse.html, plots from ArXiv:1611.01927

Unfolding: regularization 1/

$$\chi^2_{\text{TUnfold}} = \chi^2_A + \tau^2 \chi^2_L$$
$$\chi^2_A = (\boldsymbol{A}\hat{\boldsymbol{x}} + \boldsymbol{b} - \boldsymbol{y})^{\mathsf{T}} (\boldsymbol{V_{yy}})^{-1} (\boldsymbol{A}\hat{\boldsymbol{x}} + \boldsymbol{b} - \boldsymbol{y})$$
$$\chi^2_L = (\hat{\boldsymbol{x}} - \boldsymbol{x_B})^{\mathsf{T}} \boldsymbol{L}^{\mathsf{T}} \boldsymbol{L} (\hat{\boldsymbol{x}} - \boldsymbol{x_B})$$



- Choose τ corresponding to maximum curvature of L-curve
- Or minimize the global $\rho_{\text{avg}} = \frac{1}{M_x} \sum_{j=1}^{M_x} \rho_j$
  - Often results in stronger regularization than L-curve

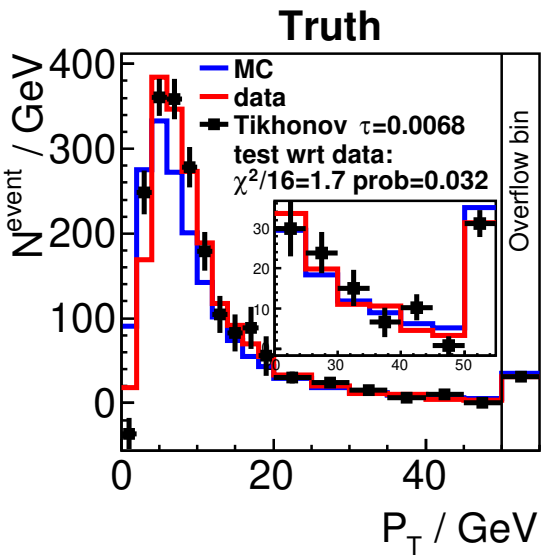


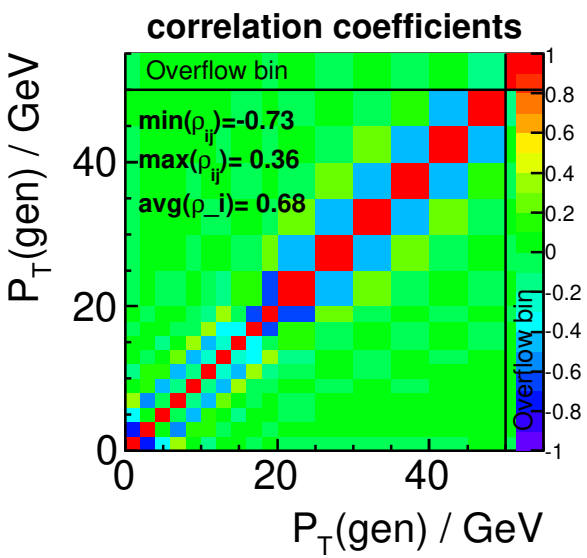
Plots from ArXiv:1611.01927

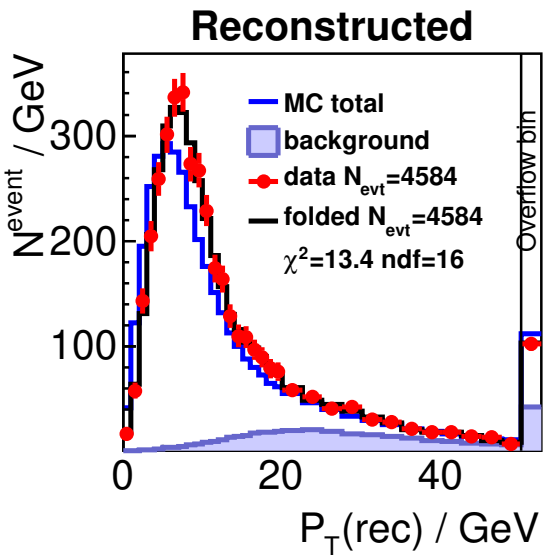
Unfolding: regularization 2/

$$\mathcal{L}(\mathbf{x},\lambda) = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3,$$
$$\mathcal{L}_1 = (\mathbf{y}-\mathbf{A}\mathbf{x})^T \mathbf{V_{yy}} (\mathbf{y}-\mathbf{A}\mathbf{x}),$$
$$\mathcal{L}_2 = \tau^2 (\mathbf{x}-f_b\mathbf{x_0})^T (\mathbf{L}^T\mathbf{L})(\mathbf{x}-f_b\mathbf{x_0}),$$
$$\mathcal{L}_3 = \lambda(Y - \mathbf{e}^T\mathbf{x}),$$
$$Y = \sum_i y_i,$$
$$e_j = \sum_i A_{ij}.$$

- $\mathbf{y}$: observed yields
- $\mathbf{A}$: response matrix
- $\mathbf{x}$: the unfolded result
- $\mathcal{L}_1$: least-squares minimization ($V_{ij} = e_{ij}/e_{ii}e_{jj}$ correlation coefficients)
- $\mathcal{L}_2$: regularization with strength τ
- Bias vector $f_b\mathbf{x_0}$: reference with respect to which large deviations are suppressed
- $\mathcal{L}_3$; area constraint (bind unfolded normalization to the total yields in folded space)







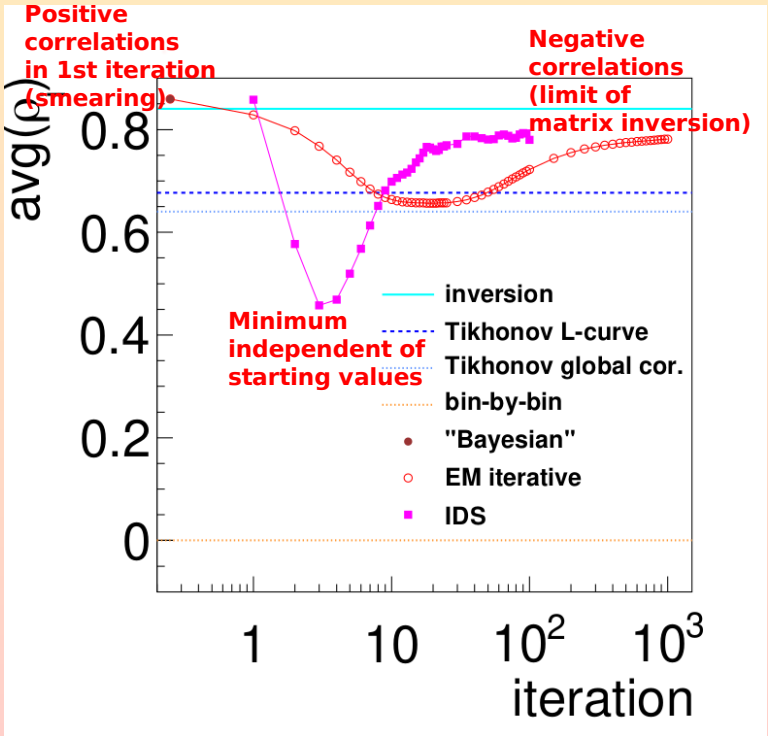
Plots from ArXiv:1611.01927

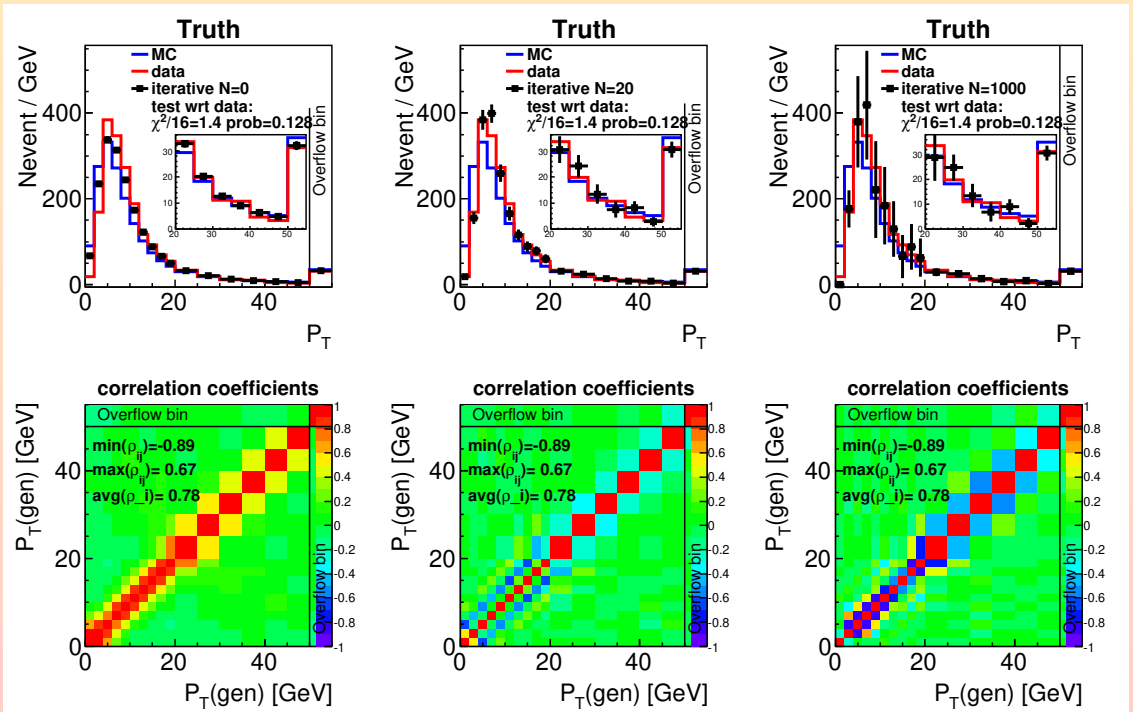
Unfolding: Iterative Unfolding

- Iterative improvement over the result of a previous iteration;

$$x_j^{(n+1)} = x_j^{(n)} \sum_{i=1}^{M} \frac{A_{ij}}{\epsilon_j} \frac{y_i}{\sum_{k=1}^{N} A_{ik} x_k^{(n)} + b_i}$$

  - It converges (slowly, $N_{iter} \sim N_{bins}^2$) to the MLE of the likelihood for independent Poisson-distributed y_i
  - Not necessarily unbiased for correlated data (does not make use of covariance of input data V_{yy})
- In HEP most people don't iterate until convergence
  - Fixed N_{iter} is often used; the dependence on starting values provides regularization
- Intrinsically frequentist method
  - for $N_{iter} \to \infty$ converges to matrix inversion, if all $\hat{x}_j$ from matrix inversion are positive
  - $N_{iter} = 0$ sometimes called improperly "Bayesian" unfolding (the author, D'Agostini, is Bayesian)
- Don't use software defaults!!! (e.g. some software has $N_{iter} = 4$)
  - Minimizing the global ρ is a good objective criterion, but there are others (Akaike information, etc)





Plots from ArXiv:1611.01927

When the signal is unknown: Gaussian Processes in HEP

Gaussian processes in HEP

- GP: associate a multivariate gaussian to a set of random variables ($N_{dim} = N_{random\ variables}$)

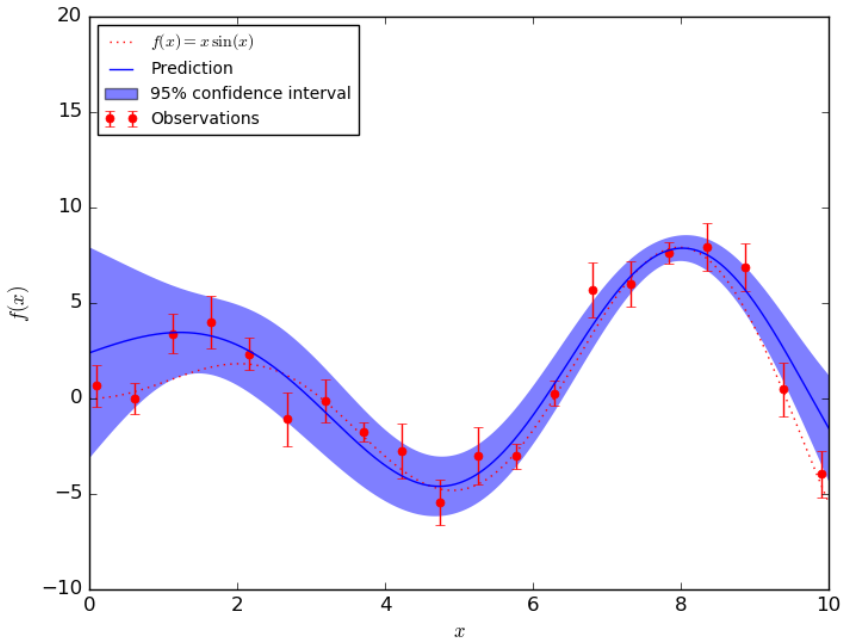
$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_N \end{bmatrix} \sim \text{Gaus}(\boldsymbol{\mu}, \Sigma)$$

- From bin counts $\boldsymbol{y}$ to a multivariate gaussian
- Infer new values y by extending the dimension of Gaus$(\boldsymbol{\mu}, \Sigma)$
- Use a *kernel* or measure of similarity between bin centers (counts) and a *mean function*
  - E.g. Exponential square kernel $k(x_i, x_j) = Aexp\left(-\frac{(x_i - x_j)^2}{2l^2} \right)$, with A and l hyperparameters

- Infer a new $\boldsymbol{y}_*$, located in $\boldsymbol{x}_*$, using

$$p(\boldsymbol{y}_*|\boldsymbol{x}_*, \boldsymbol{x}, \boldsymbol{y}) = \text{Gaus}(\boldsymbol{y}_*|\boldsymbol{\mu}_*, \Sigma_*)$$
$$\boldsymbol{\mu}_* = m(\boldsymbol{x_*}) + \boldsymbol{K}_*^T \Sigma^{-1}(\boldsymbol{y} - m(\boldsymbol{x}))$$
$$\Sigma_* = \boldsymbol{K}_{**} - \boldsymbol{K}_*^T \Sigma^{-1} \boldsymbol{K}_*$$

- Optimize the hyperparameters (e.g. maximum likelihood): "nonparametric" description of a function
- Can model pretty well models with zero mean



AMVA4NewPhysics deliverable 2.5 public report (F. Jimenez et al), plot from Scikit learn documentation

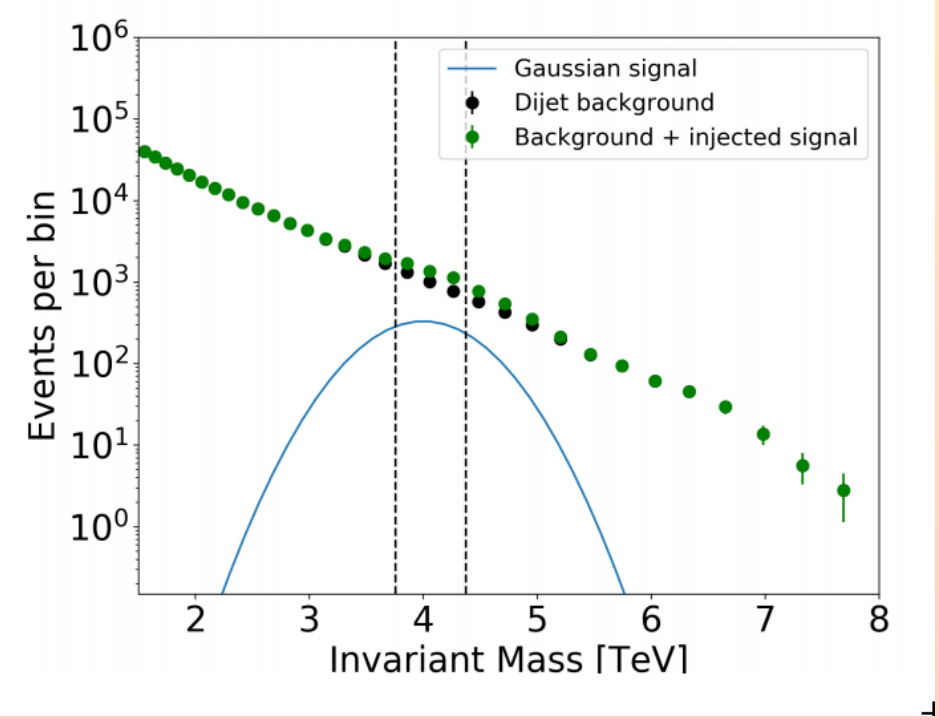
Modelling background and signal distributions

- First step
  - GP fit is performed on a pure background distribution (e.g. from simulation) using the background kernel Σ_B
  - Obtain posterior mean and covariance (that comes from the kernel with optimized hyperparams θ_B
- Second step
  - θ_B (and possibly the posterior mean) of the background fit are used in performing another fit using $\Sigma_{SB} = \Sigma_B + \Sigma_S$
  - Obtain a GP for the full model, including the hyperparameters θ_S of the signal
- Goal: accomodate the background, and identify a possible unknown signal
- Example: ATLAS dijet general search MonteCarlo dataset

$$\mu(x) = 0\,, \tag{9}$$

$$\Sigma_B(x,x') = A\ \exp\left(\frac{d-(x+x')}{2a}\right)\sqrt{\frac{2l(x)l(x')}{l(x)^2+l(x')^2}}\exp\left(\frac{-(x-x')^2}{l(x)^2+l(x')^2}\right), \tag{10}$$

$$\Sigma_S(x,x') = C\ \exp\left(-\frac{1}{2}(x-x')^2/k^2\right)\exp\left(-\frac{1}{2}\left((x-m)^2+(x'-m)^2\right)/t^2\right), \tag{11}$$

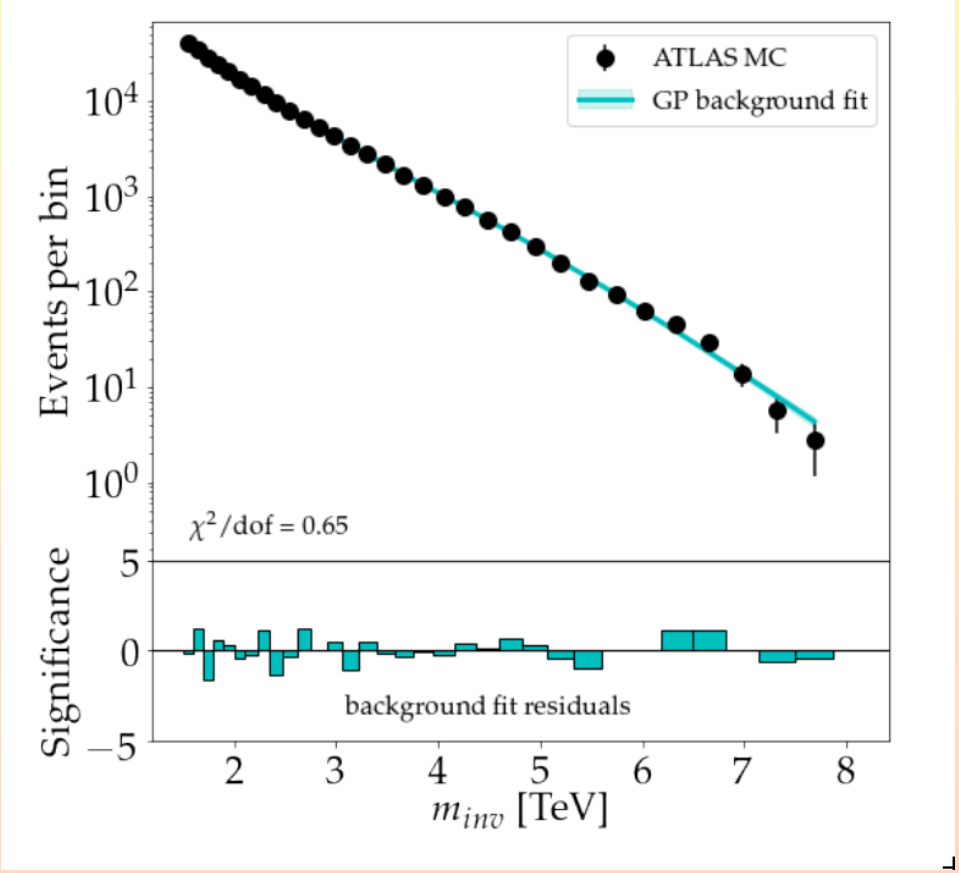


AMVA4NewPhysics deliverable 2.5 public report (F. Jimenez et al)

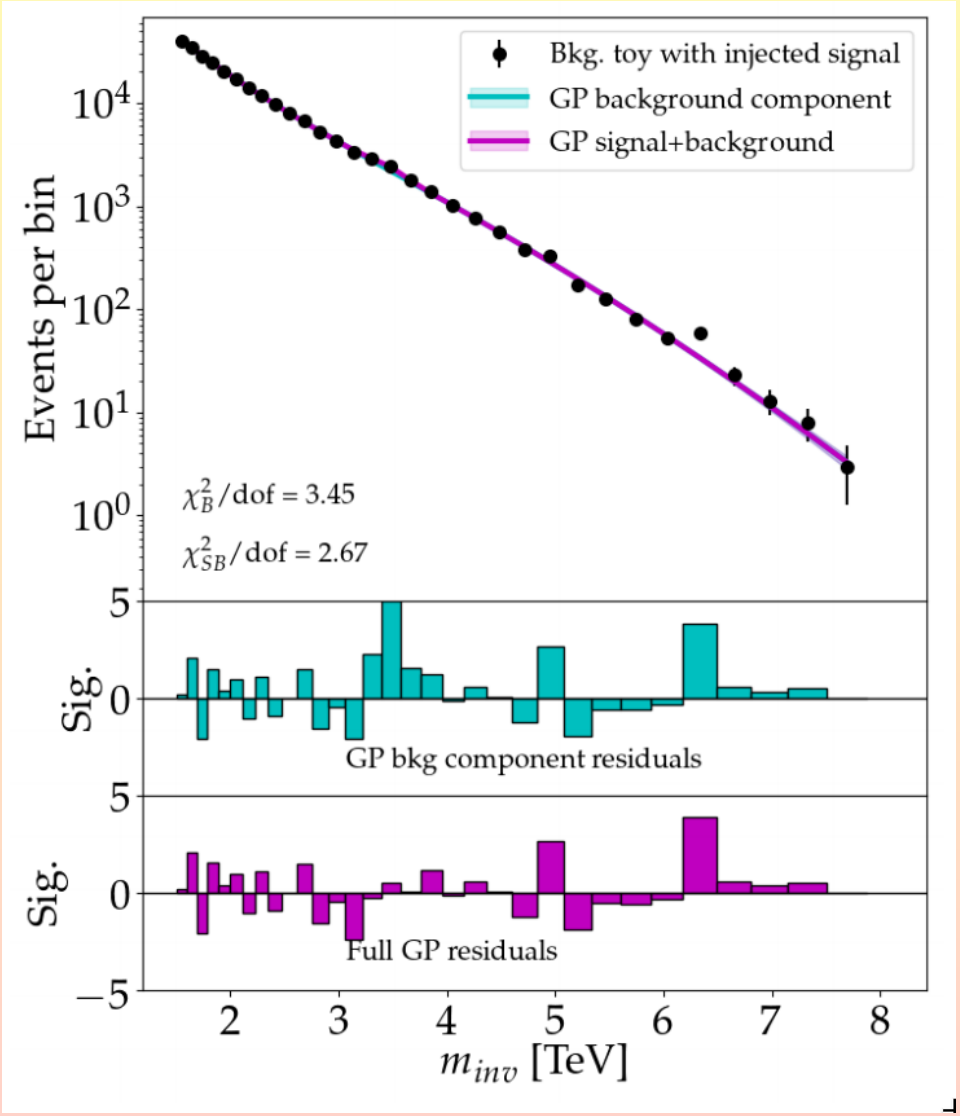
Apply the procedure

UCLouvain
Institut de recherche
en mathématique et physique

First step, bkg-only GP fit



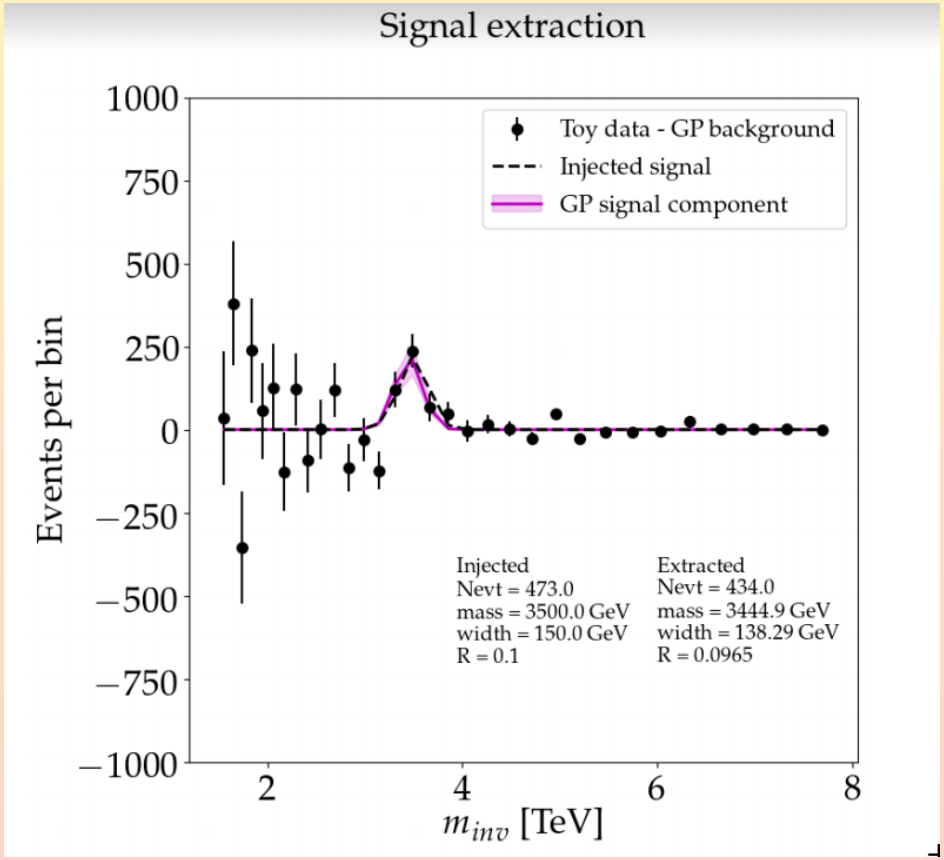
Second step, S+B GP fit using bkg-only result

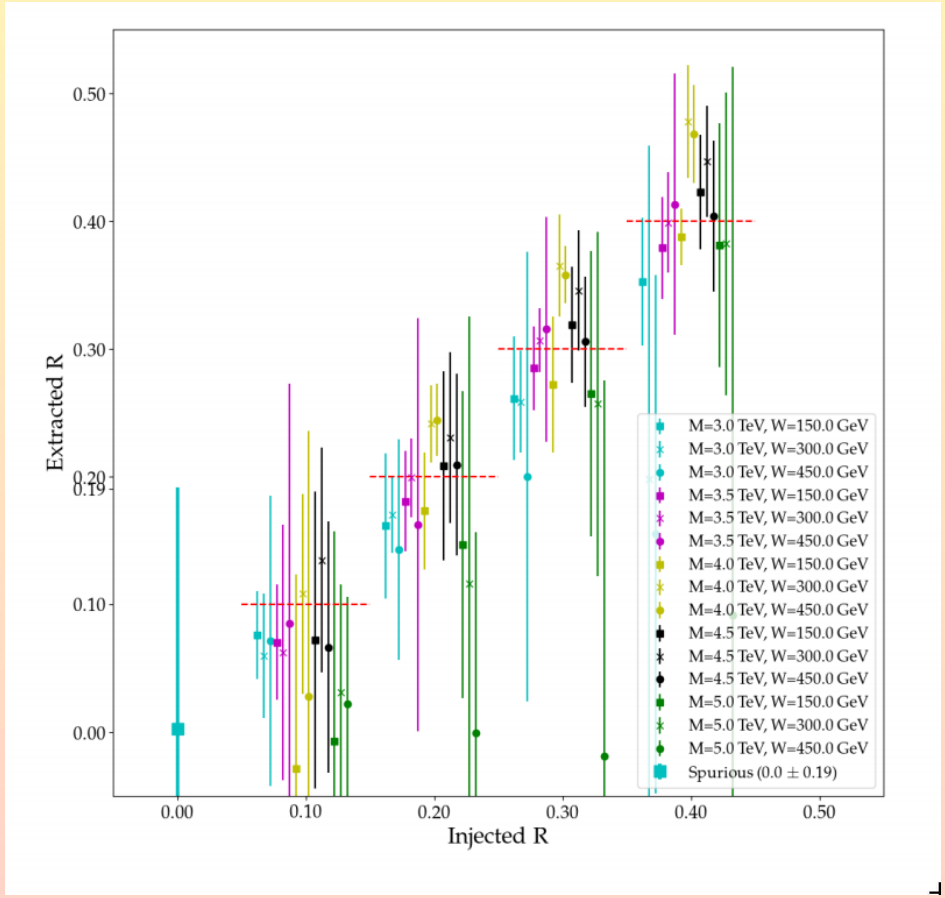


AMVA4NewPhysics deliverable 2.5 public report (F. Jimenez et al)

Signal extraction and performance

- Residuals w.r.t. bkg-only GP shows good performance in recovering unknown signal
- Poorer performance at higher mass hypotheses (due to non-uniform binning in the mass spectrum, and low event counts)





AMVA4NewPhysics deliverable 2.5 public report (F. Jimenez et al)

Summary

- Some highlights on statistical methods for data analysis in ATLAS and CMS
- Necessary selection of topics
  - Definition of probability
  - Point estimation
  - Interval estimation
  - Hypothesis testing
  - Unfolding
  - Gaussian processes for unknown signals
- Machine learning covered in my talk on the 28th!!!

Non-exhaustive list of references

- Glen Cowan, *Statistical Data Analysis*, Oxford Science Publications
- Louis Lyons, *Statistics for Nuclear And Particle Physicists*, Cambridge University Press
- Louis Lyons, *A Practical Guide to Data Analysis for Physical Science Students*, Cambridge University Press
- R.J.Barlow, *A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley
- Kyle Cranmer, *Lessons at HCP Summer School 2015*
- Kyle Cranmer, *Practical Statistics for the LHC*, http://arxiv.org/abs/1503.07622
- Harrison Prosper, *Practical Statistics for LHC Physicists*, CERN Academic Training Lectures, 2015 https://indico.cern.ch/category/72/
- Harald Cramér, *Mathematical Methods of Statistics*, Princeton UP
- Frederick James, *Statistical Methods in Experimental Physics* (2nd Edition), World Scientific
- D. R. Cox, *Principles of Statistical Inference*, Cambridge University Press
- E.T. Jaynes, *Probability Theory: the Logic Of Science*, Cambridge University Press
- Jim Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer
- Annis, Stuard, Ord, Arnold, *Kendall's Advanced Theory Of Statistics I and II*, Wiley
- Pearl, Judea, *Causal inference in Statistics, a Primer*, Wiley
- Behnke, Kröninger, Schott, Schörner-Sadenius, *Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods*
- Narsky, Porter, *Statistical Analysis Techniques in Particle Physics*, Wiley
- Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning*, Springer

THANK YOU FOR YOUR ATTENTION!!

Backup