Statistics

or “How to find answers to your questions”

Pietro Vischia[1]

[1] CP3 — IRMP, Université catholique de Louvain

UCLouvain
Institut de recherche
en mathématique et physique

CP3, Lectures on Statistics for HEP

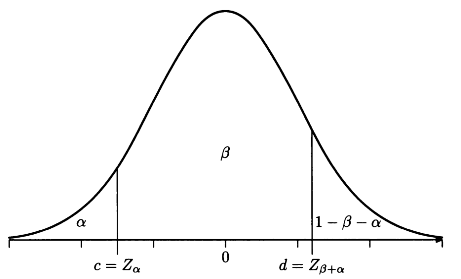
Contents

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 $\alpha = \frac{1-\beta}{2}$

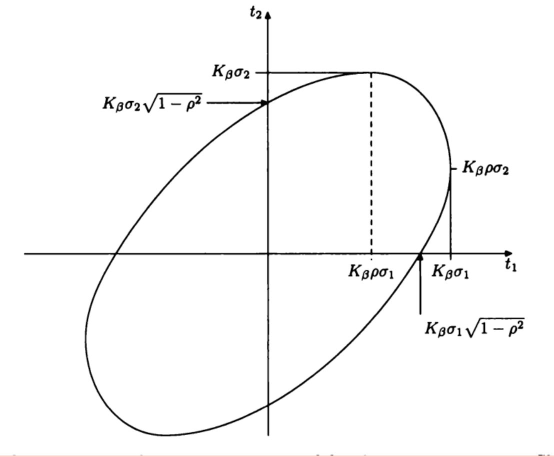


Plot from James, 2nd ed.

Confidence intervals in many dimensions

UCLouvain
Institut de recherche
en mathématique et physique

- 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

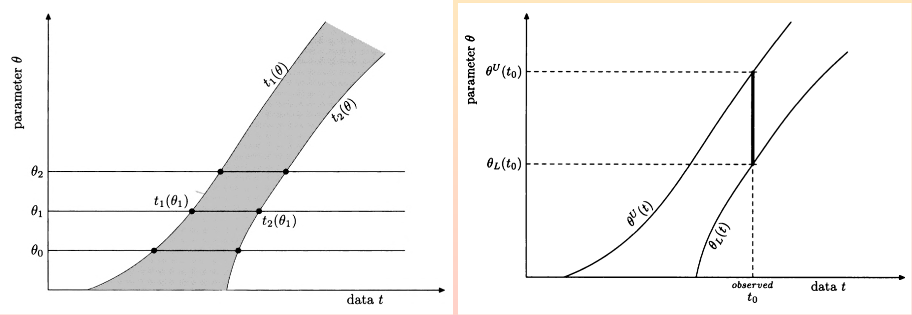


Plot from James, 2nd ed.

Confidence belts: the Neyman construction

UCLouvain
Institut de recherche
en mathématique et physique

- 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

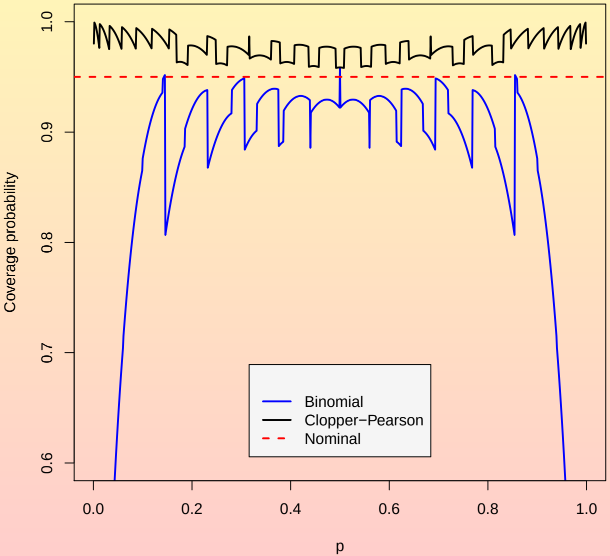
- For discrete distributions, the discreteness induces steps in the probability content of the interval
  - Continuous case: $P(a \leq X \leq b) = \int_a^b f(X|\theta)dX = \beta$
  - Discrete case: $P(a \leq X \leq b) = \sum_a^b f(X|\theta)dX \leq \beta$
- Binomial: find interval (r_{low}, r_{high}) 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

Coverage: the binomial case

- 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
- Study coverage of intervals from a gaussian approximation and from the Clopper-Pearson method
  wget https://raw.githubusercontent.com/vischia/statex/master/coverageTest.R
  wget https://raw.githubusercontent.com/vischia/statex/master/coverageTest.py
  wget https://raw.githubusercontent.com/vischia/statex/master/coverageTest.ipynb
  - 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
  - Find the issue with the Clopper Pearson implementation in python
  - What happens for different sample sizes N?

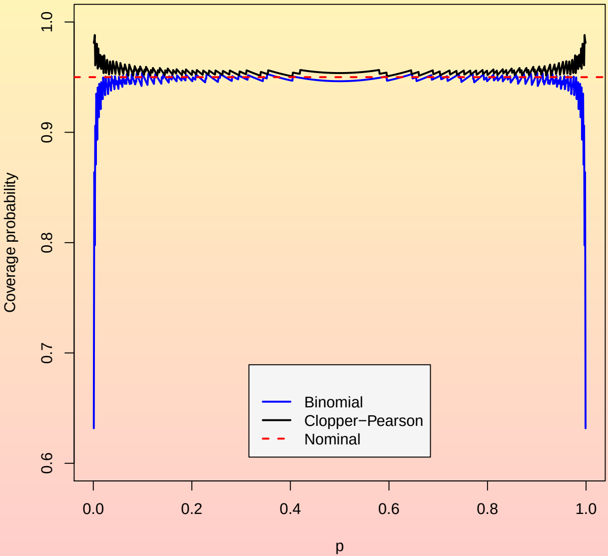
Coverage, $N = 20$

- Gaussian approximation bad for small sample sizes



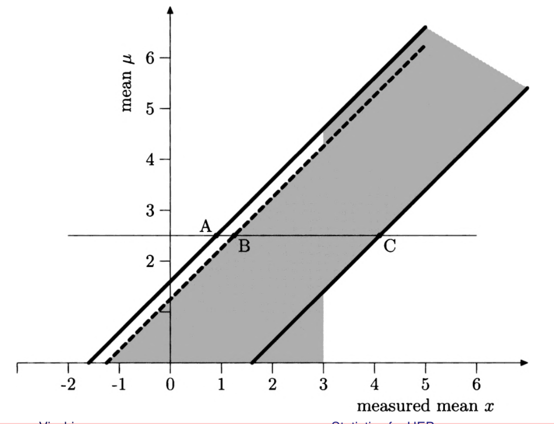
Coverage, $N = 1000$

- 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)

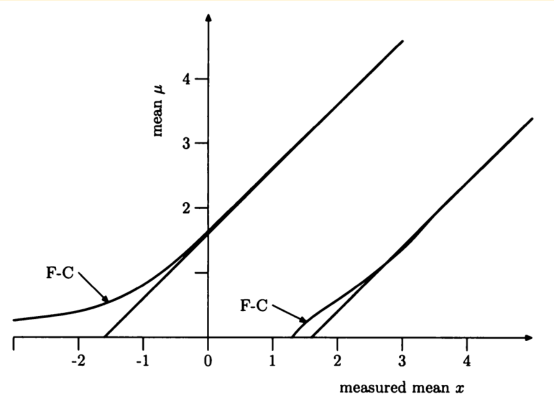


Plot from James, 2nd ed.

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.

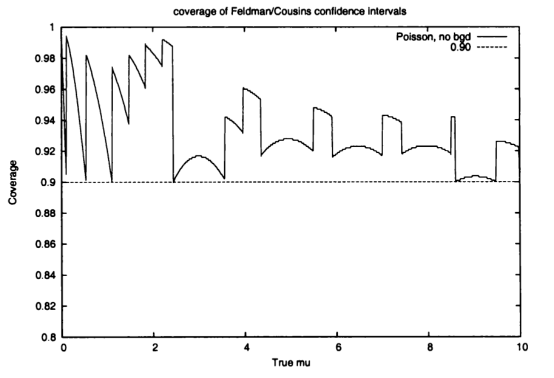
Feldman-Cousins in HEP

UCLouvain
Institut de recherche
en mathématique et physique

- 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.

UCLouvain
Institut de recherche
en mathématique et physique

- 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...

UCLouvain
Institut de recherche
en mathématique et physique

- 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.

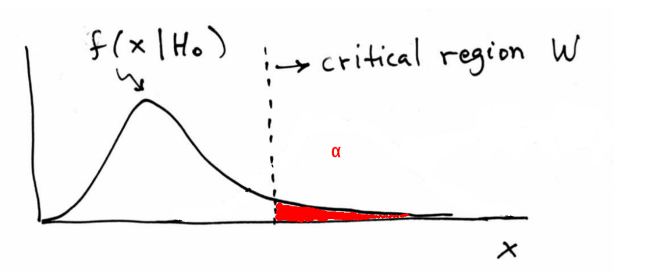
...and how do we test it?

- 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...

UCLouvain
Institut de recherche
en mathématique et physique

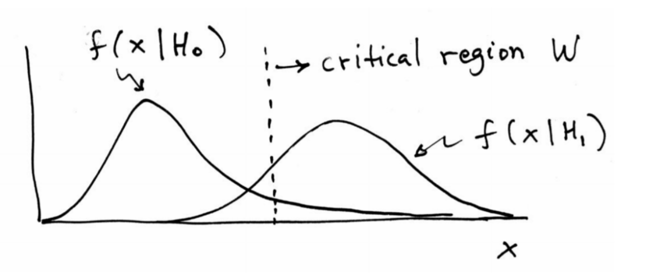
- Suppose we want to test an hypothesis H_0
- H_0 is normally the hypothesis that we assume true in absence of further evidence
- Let $\mathbf{X}$ be a function of the observations (called "*test statistic*")
- Let W be the space of all possible values of $\mathbf{X}$, and divide it into
  - A critical 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

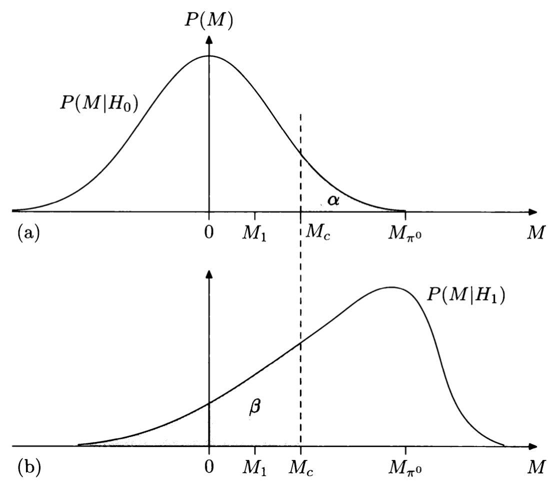
UCLouvain
Institut de recherche
en mathématique et physique

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



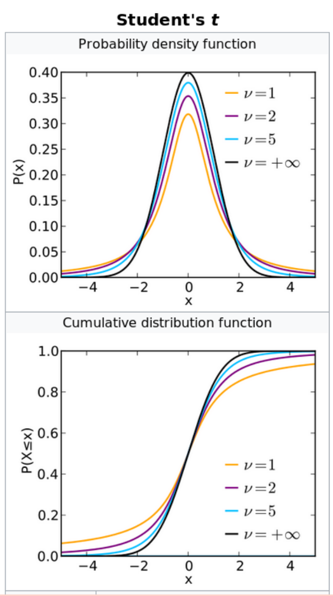
	Choose H_0	Choose H_1
H_0 is true	$1-\alpha$	α (Type I error)
H_1 is true	β (Type II error)	$1-\beta$

Plot from James, 2nd ed.

A longer example

- Student's t distribution
- Test the mean!
- wget hyptest.ipynb

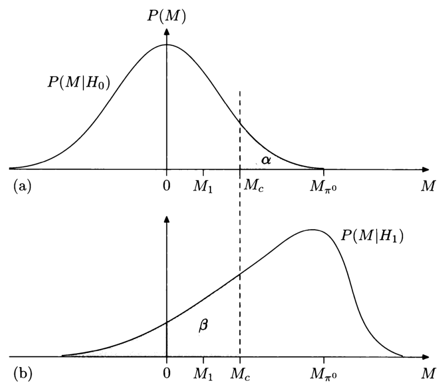
PDF	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$

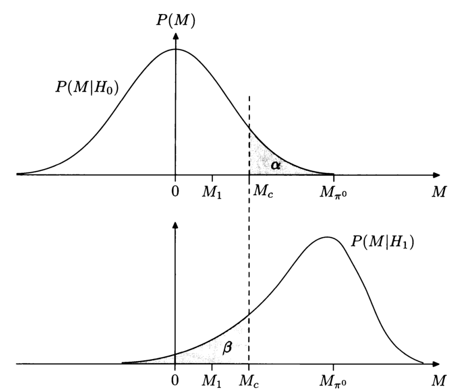


Basic hypothesis testing – 4

UCLouvain
Institut de recherche
en mathématique et physique

- 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.

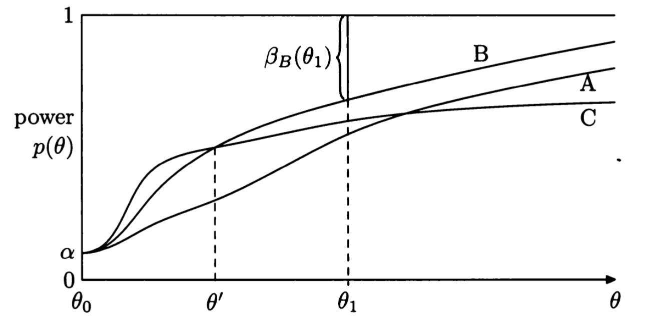




Plots from James, 2nd ed.

Comparing tests

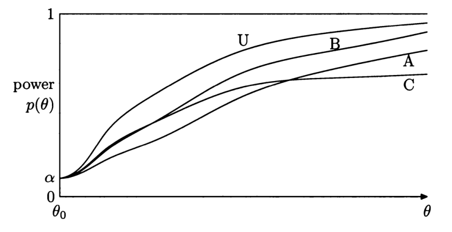
- For parametric (families of) hypotheses, the power depends on the parameter
  - $H_0 : \theta = \theta_0$
  - $H_1 : \theta = \theta_1$
  - Power: $p(\theta_1) = 1 - \beta$
- Generalize for all possible alternative hypotheses: $p(\theta) = 1 - \beta(\theta)$
  - For the null, $p(\theta_0) = 1 - \beta(\theta_0)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

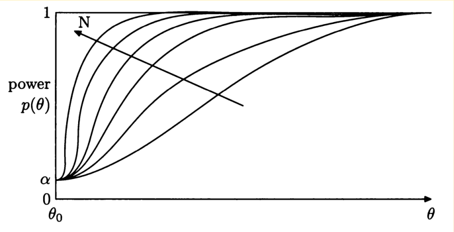


Plot from James, 2nd ed.

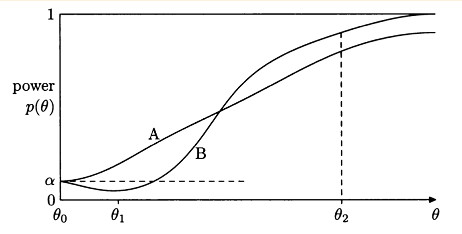
Properties of tests

- 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 $\theta_1]$
- Still useful (larger power) for θ_0 vs θ_2

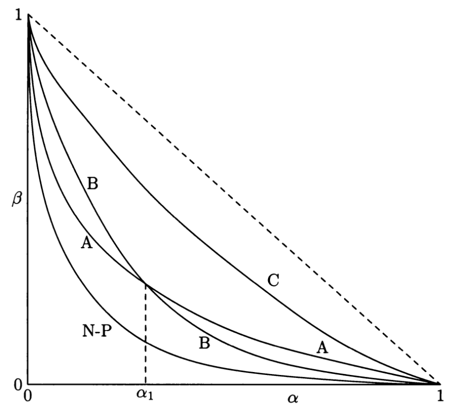


Plot from James, 2nd ed.

[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

- 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

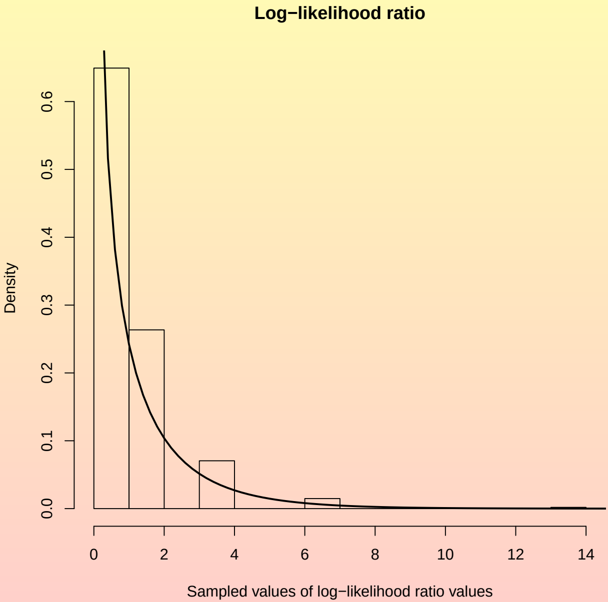
Intermezzo: the Wilks theorem

- The likelihood ratio is commonly used
- As any test statistic in the market, in order to select critical regions based on confidence levels it is necessary to know its distribution
  - Run toys to find its distribution (very expensive if you want to model extreme tails)
  - Find some asymptotic condition under which the likelihood ratio assumes a simple known form
- Wilks theorem: when the data sample size tends to ∞, the likelihood ratio tends to $\chi^2(N - N_0)$
  - Check if it's actually true!
    wget https://raw.githubusercontent.com/vischia/statex/master/wilks.R
    wget https://raw.githubusercontent.com/vischia/statex/master/wilks.ipynb

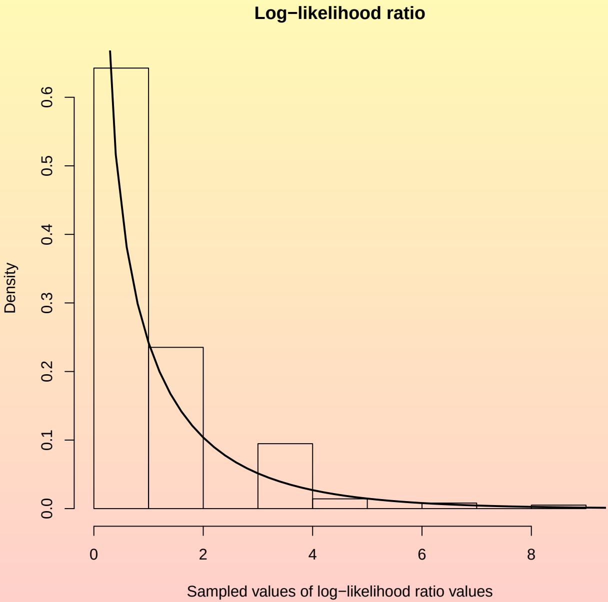
We can summarize in the

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.

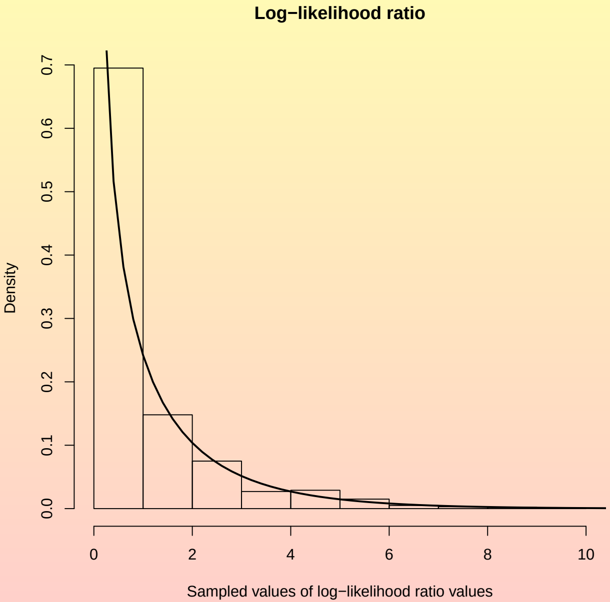
Verifying the Wilks theorem: N=2



Verifying the Wilks theorem: N=10



Verifying the Wilks theorem: N=100



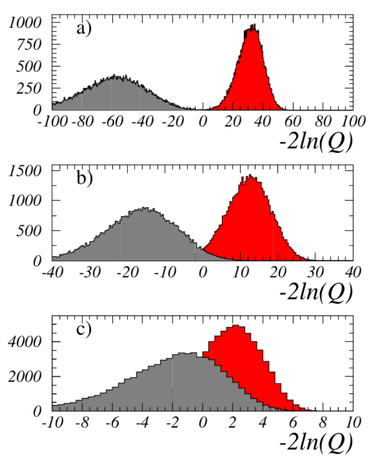
Testing hypotheses near the boundary: Zech

- 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 <u>confidence level</u> (CL)
- <u>Confidence limit</u>: 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!

- 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!

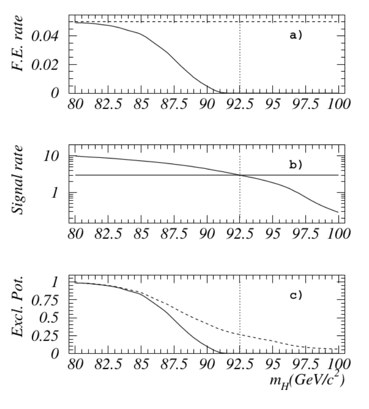


Plot from Read, CERN-open-2000-205

Avoid issues at low signal rates

UCLouvain
Institut de recherche
en mathématique et physique

- $CL_s := \frac{CL_{s+b}}{CL_b}$
- 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!

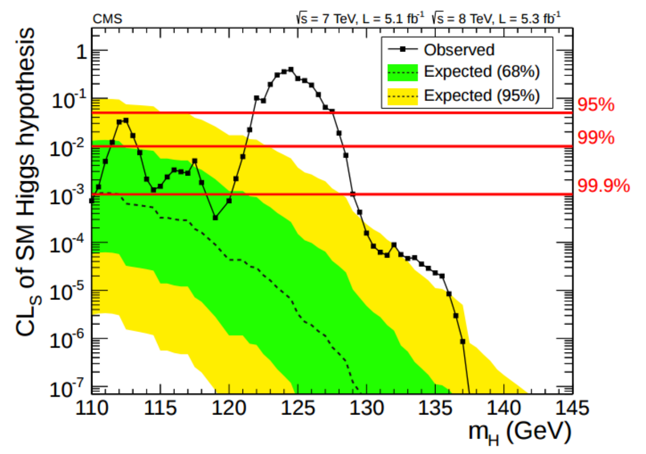


Dashed: CL_{s+b}
Solid: CL_s
$S < 3$: exclusion for a B-free search $\equiv 0$

Plot from Read, CERN-open-2000-205

A practical example: Higgs discovery - 1

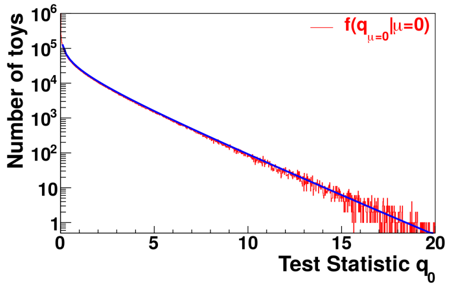
- Apply the CL_s method to each Higgs mass point
- Green/yellow bands indicate the $\pm 1\sigma$ and $\pm 2\sigma$ intervals for the expected values under B-only hypothesis
  - Obtained by taking the quantiles of the B-only hypothesis

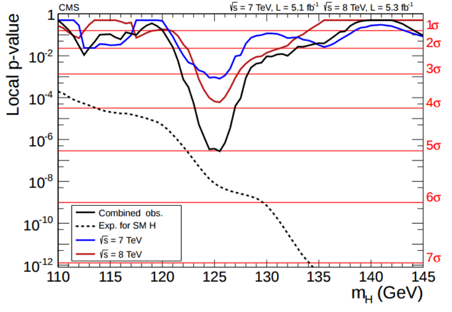


- Now let's play with CLs!
- wget https://raw.githubusercontent.com/vischia/statex/master/cls_counting.ipynb
- You will need to install the first two (the other two are for the next exercises)
  - `pip3 install pyhf –user`
  - `pip3 install uproot –user`
  - `pip3 install –user pyunfold`
  - `pip3 install –user seaborn`

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 $p = \frac{1}{2}P_{\chi^2_1}(Z^2)$

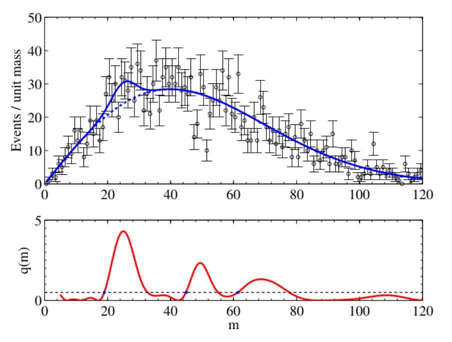


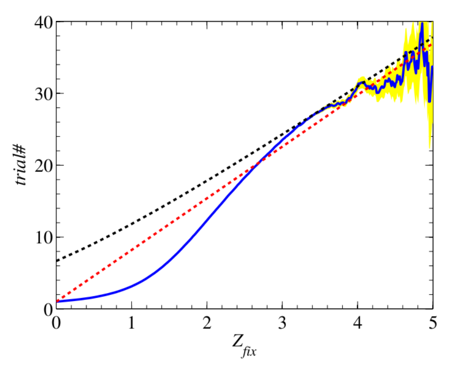


Plots from ATL-PHYS-PUB-2011-011 and from Higgs discovery

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 $p_b^{global} := P(q_0(\hat{m_X}) > u) \leq \langle N_u \rangle + \frac{1}{2} P_{\chi_1^2}(u)$
  - 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

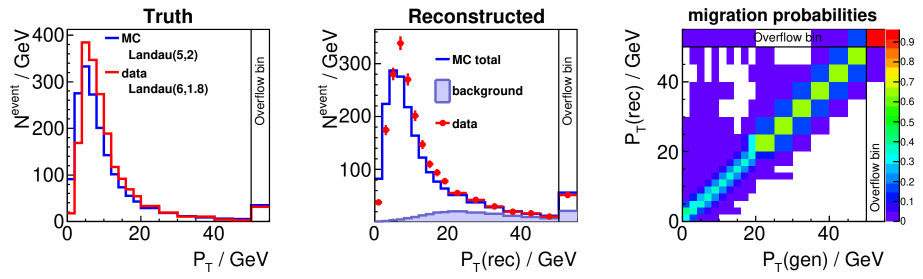
Unfolding: the problem

UCLouvain
Institut de recherche
en mathématique et physique

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

$$\mathcal{L} = (\mathbf{y} - \mathbf{Ax})^T \mathbf{V_{yy}} (\mathbf{y} - \mathbf{Ax}),$$

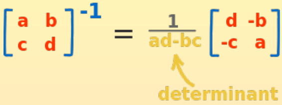
- Observations y, to be transformed in the theory space into x
  - Model the detector as a response matrix
  - Invert the response to convert experimental data to theory space distributions
  - Usually to compare with models in the theory space
- The best solution is to fold any new theory and make comparisons in the experimental data space

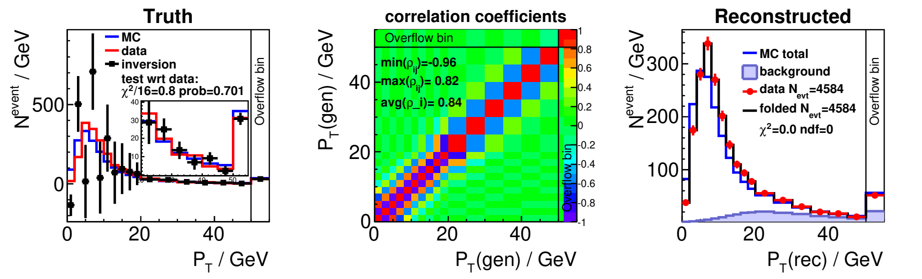


Plot from ArXiv:1611.01927

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 $\boldsymbol{V_{xx}}$

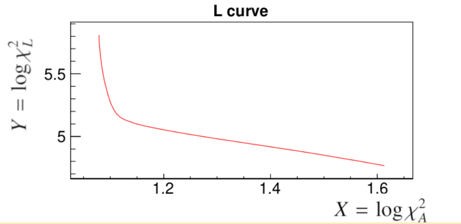




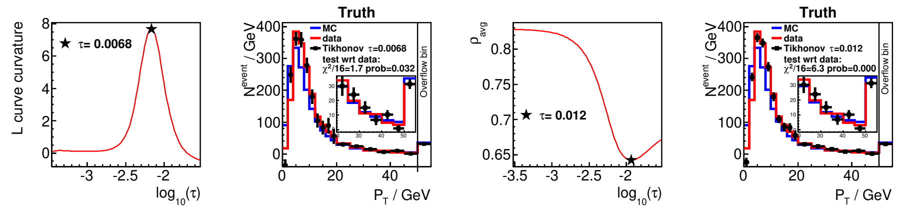
Cartoon from https://www.mathsisfun.com/algebra/matrix-inverse.html, plots from ArXiv:1611.01927

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

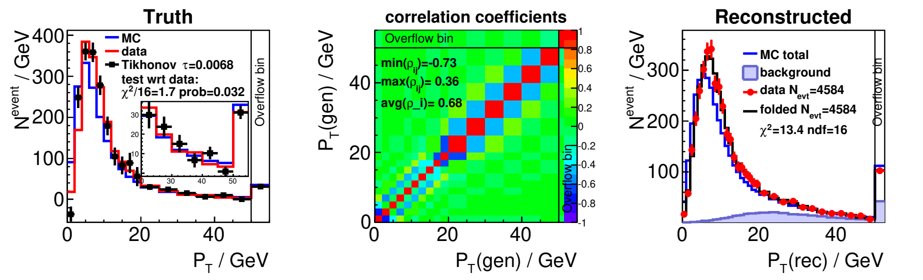


Plots from ArXiv:1611.01927

Unfolding: regularization 2/

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

- $\mathbf{y}$: 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

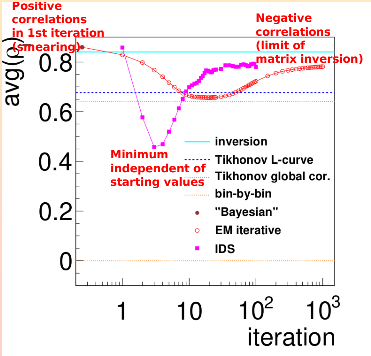
Unfolding: Iterative Unfolding

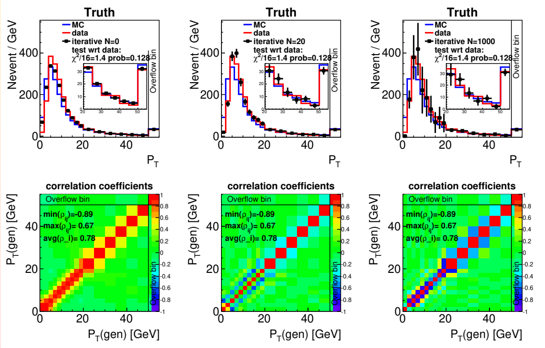
UCLouvain
Institut de recherche
en mathématique et physique

- 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 $\boldsymbol{V_{yy}}$)
- 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

Exercise time!

- I don't really have to add anything to the wonderful `pyunfold` tutorials: https://github.com/jrbourbeau/pyunfold/tree/master/docs/source/notebooks
- Basic unfolding
  wget tutorial.ipynb
- Change your prior!
  wget user_prior.ipynb
- Regularization
  wget regularization.ipynb
- Multivariate unfolding
  wget multivariate.ipynb
- You can get them all by running
  the pyunfold/https://raw.githubusercontent.com/vischia/statex/master/pyunfold/get.sh script from the exercises repository

- Statistics is about answering questions
  - ...and posing the questions in an appropriate way
- Foundations
  - Mathematical definition of probability
  - Bayesian and Frequentist realizations
- How wide is the table?: Point estimates and the method of maximum likelihood
- Is it really that wide, or am I somehow uncertain about it?: Interval estimates
  - Maximum likelihood
  - Neyman construction
  - Feldman-Cousins ordering
  - Coverage
- Is the table a standard-size ping-pong table or not? Testing hypotheses
  - Frequentist hypothesis testing, and some mention to the Bayesian one
  - I need no toy: the Wilks theorem
  - Upper limits and the CL_s prescription
- Can I decouple my result from my instrumentation? Unfolding
- What we did not go through (but I am happier having provided more detail about core methods)
  - A couple experimental methods (ABCD and the like)
  - Machine learning
- Thanks to Cristian for having written a notebook with the first non-notebook exercises!
  - If it's fine with you, I'll check it and upload it with your name on it
- Are you satisfied? Tell me more by clicking here https://forms.gle/T4XbmZXLEi6KL8rN7 (or taking the link from the indico of the last lecture)

THANK YOU VERY MUCH FOR ATTENDING!!

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

Non-exhaustive list of references

- Frederick James: Statistical Methods in Experimental Physics - 2nd Edition, World Scientific
- Glen Cowan: Statistical Data Analysis - Oxford Science Publications
- Louis Lyons: Statistics for Nuclear And Particle Physicists - Cambridge University Press
- Louis Lyons: A Practical Guide to Data Analysis for Physical Science Students - Cambridge University Press
- Annis?, Stuard, Ord, Arnold: Kendall's Advanced Theory Of Statistics I and II
- Pearl, Judea: Causal inference etc etc, a Primer (add full details)
- 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/

THANKS FOR THE ATTENTION!

Backup