PROCEEDINGS OF SCIENCE

Statistics for Data Analysis

Tommaso Dorigo[a,*] **and Irene Telali**[b,c]

[a] *Istituto Nazionale di Fisica Nucleare, Sezione di Padova*
Via F. Marzolo 8, 35131 Padova, Italy
[b] *Perimeter Institute for Theoretical Physics,*
31 Caroline Street North, Waterloo, Ontario, Canada N2L 2Y5
[c] *Department of Physics and Astronomy, University of Waterloo,*
200 University Avenue West, Waterloo, ON, N2L 3G1, Canada

E-mail: dorigo@pd.infn.it, etelali@perimeterinstitute.ca

This document details the contents of lectures in Statistics for the analysis of data in fundamental sciences research. They touch on some basic concepts such as point and interval estimation, error propagation, and hypothesis testing.

Corfu Summer Institute 2021 "School and Workshops on Elementary Particle Physics and Gravity"
29 August - 9 October 2021
Corfu, Greece

*Speaker

PoS(CORFU2021)315

1. Introduction

This document is a written summary of a set of lectures in statistics for data analysis offered by one of us (T.Dorigo) during the 2021 QGMM Cost action school in Corfu, Greece. Due to the limited time available for the lectures, their coverage of the considered topics was neither complete nor homogeneous. The aim of this write-up is not to fill those gaps, but rather to try and provide a set of useful entry-points for several important aspects of the practice of statistics for the analysis of experimental data in particle physics. We hope this text will also be useful for the practice of statistical inference in astro-particle and nuclear physics, as well as in other exact sciences where a precise value of the true, unknown quantities under scrutiny does exist, and where experimenters care at least as much about the intervals they quote (through interval estimation) as the central values they determine (through point estimates).

The structure of this document is organized as follows. In the remainder of this section we offer a few starting considerations on the importance of a correct education on the basic concepts of statistics for anybody who wishes to extract information from complex data sets, such as those encountered in fundamental physics research. We also argue about the importance of knowing the sampling properties of the data we are handling, and the problems that may originate from neglecting to pay attention to this topic. In Section 2 we offer a few fundamental definitions, and provide an example of the importance of pondering on the deep consequences of fundamental error propagation formulas. In Section 3 we discuss estimators, focusing in particular on the least squares and maximum likelihood techniques. This introduces a discussion of the superiority of maximum likelihood, by focusing on the problem of fitting data coming from Poisson processes, when the number of observations is insufficient to allow a safe approximation of the Poissonian with a Gaussian distribution. In Section 4 we look with more detail into the consequences of the formulas for error propagation and consider the problem of combining two measurements of the same physical quantity, in the presence of correlations. This allows us to focus on an apparent paradox of which experimental physicists need to be aware. In Section 5 we discuss the issue of choosing a model for experimental data. Section 6 provides basic concepts on confidence intervals, and discusses coverage in the context of some prototypical cases. In Section 7 we discuss the concept of ancillarity and why it is relevant for experimental measurements. Section 8 deals with a comparison of the way inference is done in a frequentist and Bayesian framework. In Section 9 we discuss hypothesis testing and provide the basic definitions of relevance for that practice. We conclude in Section 10.

1.1 The importance of a statistics background for experimental physicists

Membership of one of the large collaborations operating the detectors that analyze proton-proton collisions at the Large Hadron Collider (LHC) offers a vantage point where to observe the match and mis-match between the baggage of competences that Ph.D. students carry from their past studies and the activities they undertake in their way to a thesis on some experimental measurement. While most of them come well-equipped with knowledge in the theory of particle physics, they often lack a proper background in foundational statistics, which is often not offered during their *curriculum studiorum*. This is a pity, because as a matter of fact their knowledge of statistics plays a much deeper role in determining the success of their research than the knowledge of, say, quantum

field theory. One therefore observes how large scientific collaborations need to exert a constant steering toward acceptable statistical practice, and away from made-up, makeshift and improvised methods that have no foundations in the theory of statistics. But this needs not be so. Graduate student lectures such as those summarized in this document may help filling that gap, and sensitize a new generation of scientists toward the importance of paying more attention to correct practice in statistical inference.

1.2 Using correct language

Over time, physicists have developed their own jargon to discuss topics and ideas which were already well-known by agreed-upon terms in statistics. It is therefore useful to create a minimal dictionary of the different terms. This is provided in Table 1 below.

Physicists say	Corresponding statistics term
determine	estimate
estimate	guess
observable space	population
observe	draw a sample
data	sample
uncertainty	error
systematic	nuisance parameter

Table 1: Minimal dictionary of statistics terms and physical jargon.

Language is important. In fact, who talks bad often thinks bad, too. If one says, *e.g.*, that "the probability that the mass of the particle is below 170 GeV, given the observed data, is 0.00001", this should be immediately recognized as a probability inversion statement –one which should be anathema in a field where scientists rely on frequentist (or classical) statistics. The sentence above implies that given an observation of data connected with a parameter of nature, we are allowed to make statements on the probability distribution of that value –implicitly claiming that there exist such a thing. Classical statistics forbids this, in fact: according to it, we can only talk about the probability of the data, given parameters. Not having this distinction clear in mind is a source of potential confusion and problems.

Another example is a statement one of us once heard from a speaker at a conference, related to some astrophysical parameter estimated by an experiment: "The measurement is 0.124 ± 0.003, so the parameter is proven to be non-null at the 41-sigma level." Here the sentence shows lack of understanding of the Gaussian distribution, and of the fact that any measurement is affected by uncertainties we can estimate, but whose true probability density function is generally not Gaussian, and certainly not precisely known; as well as by uncertainties we cannot estimate or whose very existence we do not know about. The convention of quoting a number of sigma has become commonplace in fundamental science, but we should keep in mind that the number is only meant to simplify the reporting of very small tail probabilities (p-values); we thus say "five sigma" because it is easier than saying "three times ten to the minus seven", not unlike talking of gigabytes or femtobarns instead of quoting the full number of zeroes after or before the relevant digit. In the mentioned case, if the speaker mentioned above had considered that he was practically claiming

to have determined a p-value with precision well below 10^{-300}, he might have surmised that his was a ridiculous, absurd overstatement; yet only the understanding that systematic uncertainties are always present in experimental measurements, that we cannot know with arbitrary precision their distribution function, and that therefore we certainly cannot control their behaviour to levels of a part in a billion, leave alone one in ten to the three-hundredth, may prevent one from producing similar absurd claims.

Incorrect analysis practice is also unfortunately widespread in experimental physics, and much of it is due to lack of attention or insufficient training in fundamental statistics. One of the ways this manifests itself is in the liberal use of test statistics that are supposed to track the *significance* of an observation. The typical set up is the one of a counting experiment, when an excess in the rate of some observation is sought above known backgrounds. This could be the observation of photons from a source in the sky, or events with an isolated electron at a hadron collider experiment. In such cases one has to do with Poisson statistics for the sampling of the collected data. For a Poisson variable the mean equals the variance[1], so if a background count B is expected and an observation $O > B$ is obtained, the difference $O - B$ is recognized as significant if it is larger than a few times the square root of B. Armed with this observation, it is easy to be tempted to construct a figure of merit $F = S/\sqrt{B}$, where S is some expected count from a signal process that may emerge on top of the predicted background. The analyst may then seek for experimental conditions that maximize F, and to claim that the rationale be that this procedure "maximizes the expected signal significance". But $S/\sqrt{B}$ may at most be called a *pseudo-significance*[2], as it is an unsatisfying approximate tracker of the true significance of an observed excess produced by S over B, especially in the very common experimental conditions of a not very large B. An example will provide proof of the above.

1.2.1 Example 1: Is this cut advantageous?

In a real-life case from LHC analysis practice, an experimenter observed that starting from a baseline selection that provided an estimated 8 signal and 1 background events, an additional cut was capable of retaining 60 percent of the searched-for signal, while rejecting 80 percent of remaining backgrounds. In his experimental conditions, this meant that he could, by applying the selection cut, end up with a selected dataset where he could expect to see 4.8 signal and 0.2 background events. The $F = S/\sqrt{B}$ figure of merit would consequently increase from 8 to 10.7. However, if by accounting for the Poisson distribution one computes (as we argue is a more principled choice) the median of the background-only p-value distribution relative to an expected observation of 8+1=9 events given $B = 1$, this amounts to $p_{no\ cut} = 1.1 \times 10^{-6}$, which is *twice smaller* than the median p-value for an expected observation of 4.8+0.2=5 events given $B = 0.2$, $p_{cut} = 2.6 \times 10^{-6}$. The experimenter, by applying the extra selection cut, was therefore worsening the expected significance of his signal instead of improving it! The reason is that for small event counts the Gaussian approximation of the Poisson, on which the F test statistic is based, badly breaks down.

[1] Mean and variance will be properly defined in Sec. 2.1.

[2] Perhaps it is worth recalling that the suffix pseudo- comes from Greek $\pi\sigma\epsilon\phi\tau\iota\kappa o'$, which means false, fake.

1.3 Why it is crucial to know the properties of basic statistical distributions

Almost every researcher is familiar with a number of statistical distributions –the most commonly recognized are perhaps the Gaussian, the Poisson, the exponential, the Uniform, the binomial and the multinomial–, and a mediocre physicist can live a comfortable life without having other probability distributions at his or her fingertips. However, it could be argued that a good experimentalist should at the very least recognize and be familiar with the chi-square, the compound Poisson, the log-normal, the gamma, the beta, the Cauchy, the Laplace, the Fisher-Snedecor distributions –and we are certainly only quoting the first few that come to mind. Most statistics books discuss these distributions in depth, for good reason. We refer readers to any good book on statistics for a check of their knowledge of this topic.

1.3.1 Example 2: Are the data drawn from a Poisson?

A telling proof of the trouble one may end up in by not knowing the difference between a Poisson and a compound Poisson distribution, *e.g.*, comes from the search for free quarks in atmospheric showers, which was a popular research topic in the late sixties, after quarks were recognized as potential elementary constituents of hadrons. In 1968 C. Mc Cusker and I. Cairns (Univ. Sydney) observed four tracks in a Wilson chamber whose apparent ionization was compatible with the one expected from particles of fractional charge. Successively, they published an article on Physics Review Letters [1] where they showed a track which could not be anything but a fractionally charged particle: it had produced 110 droplets per unit path length in the detector, against an expectation of 229 (a number obtained from a study of all the 55,000 observed tracks). Assuming Poisson statistics, they proceeded to estimate the tail probability of observing as few as 110 droplets per unit path length in a track, as follows:

$$P(n \leq 110) = \sum_{i=0}^{110} \frac{229^i e^{-229}}{i!} \simeq 1.6 \times 10^{-18}. \quad (1)$$

Since they had observed 55,000 tracks, they then needed to account for the trial factor: this arises from having multiple chances of finding an oddly low-ionizing particle. This is easy to account for; one takes the tail probability P as computed above and computes the trial-factor-corrected $P' = 1 - (1 - P)^{55000}$, which turns out to be about 10^{-13}. From this they concluded that they had certainly observed a fractional charge particle. The horror, the horror! In truth, scattering of particles in the Wilson chamber and droplets formation are two independent Poisson processes: a single scattering on average produces $\mu = 4$ droplets. Hence one should rather compute the tail probability using the *compound* Poisson distribution, by setting $\lambda\mu = 224$, and thus:

$$P(n \leq 110) = \sum_{i=0}^{110} \sum_{N=0}^{\infty} [\frac{(N\mu)^i e^{N\mu}}{i!} \frac{\lambda^N e^{-\lambda}}{N!}] \simeq 4.7 \times 10^{-5} \quad (2)$$

from which one gets the trial factor-corrected probability of seeing at least one such track as rather $P' = 1 - (1 - P)^{55000}$, or 92.5%! In other words, the researchers published on a prestigious peer-reviewed journal a p-value which was wrong by 13 orders of magnitude... The bottom-line is that one may be strong in nuclear physics and detector building, but only knowledge of statistics may prevent one from making a fool of oneself.

2. A few Fundamental Definitions

2.1 Point estimation

What physicists commonly address as *data fitting* is the combination of two different, and potentially quite separate, tasks in statistics: the one of point estimation and the one of interval estimation. The two can still be summarized, if one does not want to emphasize their different focus, as parts of the general problem of parameter estimation. In this section we briefly introduce the fundamental concepts behind point estimation. We will later be able to use them to look with some detail into the issue of combination of different estimates of a parameter, which is indeed a special case of parameter estimation, and one of particular significance to fundamental research.

2.1.1 P.d.f., $\mathbb{E}[.]$, Mean, and Variance

The *probability density function* (p.d.f.) $f(x)$ of a random variable x is a normalized function which describes the probability to find x in a given range:

$$P(x, x + dx) = f(x)dx. \tag{3}$$

Above, $f(x)$ is defined for a continuous variable x. For discrete variables, as *e.g.* a Poisson count $P(n \mid \mu) = e^{-\mu}\mu^n/n!$, P is instead a probability *tout court*. The *expectation value* of a random variable x is then defined as:

$$\mathbb{E}[x] = \int_{-\infty}^{+\infty} x f(x)dx = \mu. \tag{4}$$

$\mathbb{E}[x]$, also called mean of x, thus depends on the distribution $f(x)$. Of crucial importance is the *second central moment* of x, given by

$$\mathbb{E}\left[(x - \mathbb{E}[x])^2\right] = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx = V(x), \tag{5}$$

and also called variance. Variance enjoys the property that

$$\mathbb{E}\left[(x - \mathbb{E}[x])^2\right] = \mathbb{E}\left[x^2\right] - \mu^2, \tag{6}$$

as is trivial to show. Also well-known is the *standard deviation* $\sigma = \sqrt{V(x)}$.

2.1.2 Estimators and bias

The parameters of a p.d.f. characterise its shape. For instance, for the following distribution,

$$f(x; \theta) = \frac{1}{\theta} e^{-x/\theta} \tag{7}$$

x is a random variable and θ is a parameter. It is important to note here that the p.d.f. must produce a probability 1.0 of x being in the sample space: that is the reason of the multiplying factor $\frac{1}{\theta}$ in the equation above. Now, suppose we have a sample of observed values: $\vec{x} = (x_1, \ldots, x_n)$. We often want to construct some function of the data which is sensitive to the value of the parameter(s), in order to estimate it (or them): $\widehat{\theta}(\vec{x})$, where the *hat* indicates that $\widehat{\theta}$ is an estimator, not a parameter.

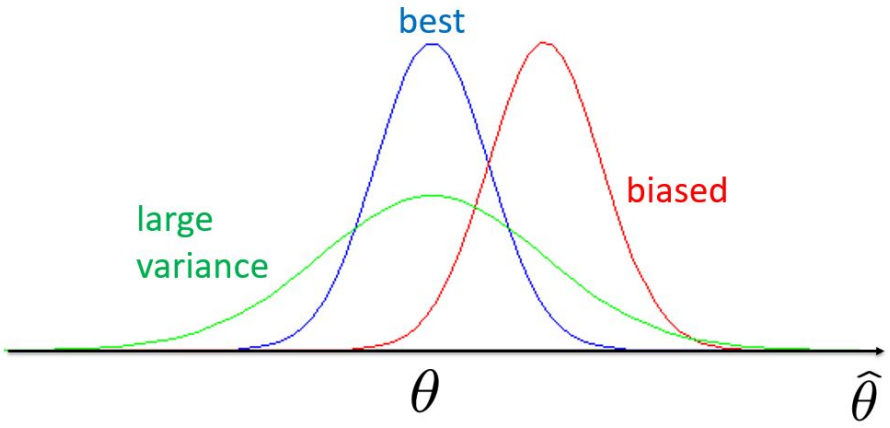

Figure 1: Possible distribution functions of an estimator $\hat{\theta}$ for a quantity with a true value θ. The estimator may have small (blue, red) or large variance (green distribution), and be biased (red) or not (blue, green distributions).

Usually we say 'estimator' for the function of $x_1, ..., x_n$. We 'estimate' the parameter using the value that the estimator takes on a particular data set.

If we were to repeat the entire measurement a large number of times, we would be able to observe how the estimates from each measurement follow a p.d.f. $g(\widehat{\theta};\theta)$ such as the one shown in Fig. 1. For estimation to be optimal we have to try and construct the estimator in such a way that it possesses two desirable characteristics. We want small (or zero) bias (which a physicist might call a systematic error):

$$b = \mathbb{E}[\hat{\theta}] - \theta; \tag{8}$$

if b is zero or negligible, the average of repeated measurements should tend to the true value. And we want a small variance (statistical error) $V(\widehat{\theta})$. Note that small bias and small variance are in general conflicting criteria: a number of statistical methods focus on finding the best trade-off between the two requirements.

2.2 Covariance and correlation

With two random variables x, y we may also define their *covariance*, defined as

$$\begin{aligned} V_{xy} &= \mathbb{E}\left[(x-\mu_x)\left(y-\mu_y\right)\right] = \mathbb{E}[xy] - 2\mu_x\mu_y + \mu_x\mu_y = \\ &= \int_{-\infty}^{+\infty} xy\, f(x,y)dxdy - \mu_x\mu_y \end{aligned} \tag{9}$$

This allows us to construct a covariance matrix V, symmetric, and with positive-defined diagonal elements, the individual variances σ_x^2, σ_y^2:

$$V = \begin{pmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & r\sigma_x\sigma_y \\ r\sigma_y\sigma_x & \sigma_y^2 \end{pmatrix} \tag{10}$$

A measure of how much x and y are correlated is given by their *correlation coefficient* r:

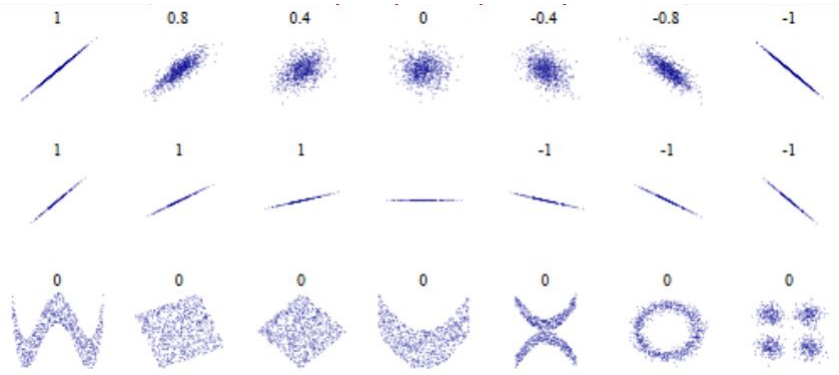

Figure 2: Values of the correlation coefficient r for various distributions (reproduced from Wikipedia).

$$r = \frac{V_{xy}}{\sigma_x \sigma_y}. \tag{11}$$

2.2.1 Uncorrelated or Independent?

Note that if two variables are independent, *i.e.*

$$f(x, y) = f_x(x) f_y(y),$$

then $r = 0$ and

$$\mathbb{E}[\mathrm{xy}] = \mathbb{E}[x]\mathbb{E}[\mathrm{y}] = \mu_x \mu_y,$$

the covariance matrix is diagonal. However, the condition $\mathbb{E}[xy] = \mathbb{E}[x]\mathbb{E}[y]$ is not sufficient for x and y to be independent. In everyday usage one speaks of "uncorrelated variables" meaning "independent". In statistical terms, uncorrelated is much weaker than independent!

On its own, $r = 0$ is a very weak condition, as a non-zero value of r only describes the tendency of the data to "line up" in a certain direction (excluding the vertical and horizontal axes). Many strictly dependent pairs of variables fulfil it. *E.g.*, a few are shown as the abscissa and ordinate of the data points in the last row of Fig. 2.

2.3 Errors

2.3.1 The error ellipse

When one measures two correlated parameters $\theta = (\theta_1, \theta_2)$, in the large-sample limit their estimators will be distributed according to a two-dimensional Gaussian centered on θ. One can thus draw an "error ellipse" (see Fig. 3) as the locus of points where the χ^2 is one unit away from its minimum value (or the log-likelihood equals $\ln(L_{\max}) - 0.5$). The location of the tangents to the axes provide the standard deviation of the estimators. The angle ϕ is given by:

$$\tan 2\phi = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_i^2 - \sigma_j^2}. \tag{12}$$

A measurement of one parameter at a given value of the other is represented by the intercept on the line connecting the two tangent points. The uncertainty of that single measurement, at a fixed value of the other parameter, is:

$$\sigma_{\text{inner}} = \sigma_i \sqrt{1 - \rho_{ij}^2}, \tag{13}$$

where the correlation coefficient ρ is the distance of each axis from the tangent point, in units of the corresponding standard deviation as shown in Fig. 3. In that case one may report

$$\hat{\theta}_i \left(\theta_j\right) \tag{14}$$

and the slope

$$\frac{d\hat{\theta}_i}{d\theta_j} = \rho_{ij} \frac{\sigma_i}{\sigma_j}. \tag{15}$$

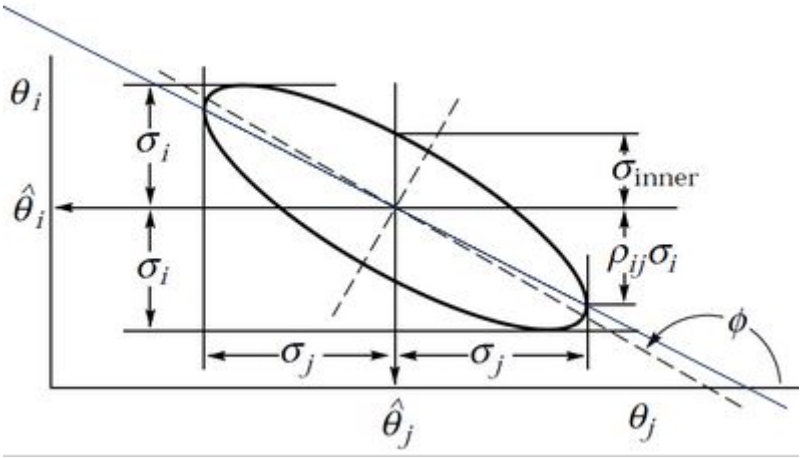

Figure 3: The error ellipse (reproduced from pdg.lbl.gov).

2.3.2 Error propagation

Imagine you have n i.i.d. (independent and identically distributed) variables x_i. Let us say you do not know their p.d.f. but at least know their mean and covariance matrix. Now let us assume that there is a function y of the x_i and you wish to determine its p.d.f.. One way to proceed is to can expand it in a Taylor series around the mean, stopping at first order:

$$y(x) \approx y(\mu) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{x=\mu} (x_i - \mu_i)\,. \tag{16}$$

From this one may show that the expectation values of y and y^2 are, to first order,

$$\begin{aligned} &\mathbb{E}[y(x)] = y(\mu) \\ &\mathbb{E}\left[y^2(x)\right] = y^2(\mu) + \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{x=\mu} V_{ij}, \end{aligned} \tag{17}$$

and the variance of y is then the second term in this expression (see Eq.(5)). In case you have a set of m functions $y(x)$, you can build their covariance matrix:

$$U_{kl} = \sum_{i,j=1}^{m} \left[\frac{\partial y_k}{\partial x_i}\frac{\partial y_l}{\partial x_j}\right]_{x=\mu} V_{ij}. \tag{18}$$

This is often expressed in matrix form once one defines a matrix of derivatives A,

$$A_{ki} = \left[\frac{\partial y_k}{\partial x_i}\right]_{x=\mu} \Rightarrow \mathbf{U} = \mathbf{AVA}^T. \tag{19}$$

The above formulas allow one to "propagate" the variances from the x_i to the y_j, but this is only valid if it is meaningful to expand linearly around the mean. So, beware of routine use of these formulas in non-trivial cases.

To see how standard error propagation works, let us use the formula for the variance of a single $y(x)$:

$$\sigma_y^2 = \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{x=\mu} V_{ij} \tag{20}$$

and consider the simplest examples we can cook up with two variables x_1, x_2: their sum and product. For the sum, the variance expression reads:

$$y = x_1 + x_2 \Rightarrow \sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2V_{12} \tag{21}$$

and for the product of the variables:

$$\begin{aligned} y = x_1 x_2 \Rightarrow \sigma_y^2 = x_2^2 V_{11} + x_1^2 V_{22} + 2x_1 x_2 V_{12} \Rightarrow \\ \frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + \frac{2V_{12}}{x_1 x_2}. \end{aligned} \tag{22}$$

One thus sees that for uncorrelated variables x_1, x_2 ($V_{12} = 0$), their variances add linearly when we compute their sum, while for their product it is the relative variances which add linearly.

2.3.3 Example 3: Error propagation and smart weighting

Thus far, we have seen how to propagate uncertainties from some measurements (random variables!) x_i to a derived quantity $y = f(x)$, Eq.(20):

$$\sigma_y^2 = \sum_i \left(\frac{\partial f(x)}{\partial x_i}\right)^2 \sigma_{x_i}^2, \tag{23}$$

which is just standard error propagation, for uncorrelated random variables x_i. What we neglect to do sometimes is to stop and think at the consequences of that simple formula, in the specific cases to which we apply it. It is instead very important to see what it really means.

To exemplify, let us take the problem of weighting two objects A and B with a two-arm scale offering a constant accuracy, say 1 gram. You have time for two weight measurements, so what do you do? Should you first weight A and then B, or should you do something else? If you weight separately A and B, by placing them in turn on one dish of the scale and finding the equilibrium by adding standard weights on the other dish, your results will be affected by the stated accuracy of the scale (the weight of the smallest reference weight): $\sigma_A = \sigma = 1g$, $\sigma_B = \sigma = 1g$. But if you instead weighted $S = A + B$, by putting them on the same dish of the scale, and then weight $D = B - A$ by putting them on different dishes, you would be able to obtain:

Figure 4: A two-arms scale. Objects to be weighted can be placed on a dish, and reference weights are placed on the other to reach equilibrium.

$$
\begin{aligned}
A = \frac{S}{2} - \frac{D}{2} \Rightarrow \sigma_A = \sqrt{\left(\frac{\sigma_S}{2}\right)^2 + \left(\frac{\sigma_D}{2}\right)^2} = \frac{\sigma}{\sqrt{2}} \\
B = \frac{S}{2} + \frac{D}{2} \Rightarrow \sigma_B = \sqrt{\left(\frac{\sigma_S}{2}\right)^2 + \left(\frac{\sigma_D}{2}\right)^2} = \frac{\sigma}{\sqrt{2}}.
\end{aligned}
\tag{24}
$$

Your uncertainties on A and B have shrunk down to 0.71 grams each, *i.e.* a value 1.41 times smaller! This is the result of having made the best out of your measurements, by making optimal use of the available information. When you placed one object on a dish, the other one was left on the table, begging to participate!

Now, let us see what happens to the previous problem if instead of a constant error of 1 gram, the scale provides measurements with accuracy of $k\%$. If we do separate weightings, of course we get $\sigma_A = kA, \sigma_B = kB$. But if we rather weight $S = B + A$ and $D = B - A$, what we get is (as $A = (S - D)/2, B = (D - S)/2$),

$$
\begin{aligned}
\sigma_\mathrm{A} = \sqrt{\frac{{\sigma_S}^2 + {\sigma_D}^2}{4}} = \sqrt{\frac{k^2(A+B)^2 + k^2(A-B)^2}{4}} = k\sqrt{\frac{A^2+B^2}{2}} \\
\sigma_B = \sqrt{\frac{{\sigma_S}^2 + {\sigma_D}^2}{4}} = \sqrt{\frac{k^2(A+B)^2 + k^2(A-B)^2}{4}} = k\sqrt{\frac{A^2+B^2}{2}}.
\end{aligned}
\tag{25}
$$

This time the procedure has shared democratically the uncertainty in the weight of the two objects. If $A = B$ we do not gain anything from our "trick" of measuring S and D: both $\sigma_A = kA$ and $\sigma_B = kB$ are the same as if you had measured A and B separately. But if they are different, we gain

accuracy on the heavier one at expense of the uncertainty on the lighter one. Of course, the limiting case of $A \gg B$ corresponds instead to a very ineffective measurement of B, while the uncertainty on A converges to what you would get if you weighted it twice.

2.3.4 Weighted average

Now suppose we need to combine two different, independent measurements with variances σ_1, σ_2 of the same physical quantity x_0, that we denote as $x_1(x_0, \sigma_1)$, $x_2(x_0, \sigma_2)$, and let the respective p.d.f.'s be $G(x_0, \sigma_i)$. We wish to combine them linearly,

$$x = cx_1 + dx_2 \tag{26}$$

to get the result with the smallest possible variance. In order to minimize variance we can play with c and d. Firstly, we note that $d = 1 - c$ if we want $\langle x \rangle = x_0$ (reason with expectation values to convince yourself of this). Then, we simply express the variance of x in terms of the variance of x_1 and x_2:

$$x = cx_1 + (1 - c)x_2, \tag{27}$$

and find the value of c which minimizes the expression. This yields:

$$\sigma_x^2 = c^2\sigma_1^2 + (1 - c)^2\sigma_2^2. \tag{28}$$

From it, we obtain:

$$\begin{aligned} x &= \frac{x_1/\sigma_1^2 + x_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2} \\ \sigma_x^2 &= \frac{1}{1/\sigma_1^2 + 1/\sigma_2^2}. \end{aligned} \tag{29}$$

We shall revisit this problem when we include the effect of correlations in a weighted average *infra* (see Sec. 4).

3. Estimators

Given a sample x_i of n observations of a random variable x, drawn from a p.d.f. $f(x)$, one may construct a statistic: a function of the data x_i containing no unknown parameters. An *estimator* is a statistic used to estimate some property of a p.d.f.. Using it on a set of data provides an estimate of the parameter.

Definition: Estimators are *consistent* if they converge to the true value for large n.

The expectation value of an estimator $\hat{\theta}$ having a sampling distribution $H(\hat{\theta}; \theta)$ is:

$$E[\hat{\theta}(x)] = \int \hat{\theta} H(\hat{\theta}; \theta) d\theta. \tag{30}$$

The most common estimators, which are ubiquitously used to obtain the location and width of a distribution, are the *sample mean* $\hat{\mu}$ and the *sample variance* $\hat{s^2}$:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{s^2} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 . \tag{31}$$

Note that the above estimators are unbiased.

Definition: The *bias* of an estimator is defined as: $b = E[\hat{\theta}] - \theta$.

An estimator can be consistent even if biased: the average of an infinite replica of experiments with finite n will not in general converge to the true value, even if $E[\hat{\theta}]$ will tend to θ as n tends to infinity.

Definition: *Mean-square error* is defined as $MSE = V[x^*] + b^2$

Notes:

- MSE is the sum of variance and squared bias, and thus gives information on the "total" error that one commits in the estimate by using a biased estimator. Given the usual trade-off between bias and variance of estimators, MSE is a good choice for the quantity to minimize when choosing an estimator, or in a regression problem.
- The Rao-Cramer-Frechet bound (see *infra*, Sec. 3.1) gives a lower limit to the variance of biased estimators, so one can take that into account in choosing an estimator.
- Consistency is an asymptotic property; *e.g.*, it does not imply that adding some more data will by force increase the precision of your estimate.
- Bias and consistency are independent properties: there are inconsistent estimators which are unbiased, and consistent estimators which are biased.

Notable examples of estimators that we will visit below are the maximum-likelihood estimator (MLE) and the least-square estimator. Asymptotically, most estimators are unbiased and normally distributed (*i.e.*, Gaussian), but the question is how far is asymptopia. Hints of the non-asymptotic regime may come from a non-parabolic nature of the Likelihood at minimum, or by the fact that two asymptotically efficient estimators provide significantly different results.

3.1 Maximum Likelihood

Take the p.d.f. of a random variable x, $f(x;\theta)$ which is analytically known, but for which the value of m parameters θ is not. The method of maximum likelihood allows us to estimate the parameters θ if we have a set of data x_i distributed according to f. The probability of our observed set x_i depends on the distribution of the p.d.f.. If the measurements are independent, we may write the probability to find x_i in $[x_i, x_i + dx_i]$ as

$$p = f\,(x_i;\theta)\,dx_i. \tag{32}$$

The likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f\,(x_i;\theta) \tag{33}$$

is then a function of the parameters θ only. It is written as the joint p.d.f. of the x_i, but we treat those as fixed. So we must be very cautious, as L is not a p.d.f.. *E.g.*, the integral under L is meaningless!

Using $L(\theta)$ one can define "maximum likelihood estimators" for the parameters θ as the values which maximize the likelihood, *i.e.* the solutions $\hat{\theta} = (\hat{\theta_1}, \hat{\theta_2}, ..., \hat{\theta_n})$ of the equation

$$\left(\frac{\partial L(\theta)}{\partial \theta_j}\right)_{\theta=\hat{\theta}} = 0 \quad \text{for } j = 1 \ldots m. \tag{34}$$

At this point let us note that the ML method requires (and exploits!) the full knowledge of the distributions. As for the variance of our MLE, in the simplest cases –*i.e.*, when one has unbiased estimates and normally distributed data– one can estimate the variance of the maximum likelihood estimate with

$$\hat{\sigma}^2_{\theta=\theta_0} = \left(-\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}_{\theta=\theta_0}. \tag{35}$$

This is also the default value used by the MIGRAD routine of MINUIT [2] to return the uncertainty of a MLE from a fit. However, note that this is only a lower limit of the variance, which applies in conditions when errors are Gaussian and when the ML estimator is unbiased. A general formula called the Rao-Cramer-Frechet inequality gives this lower bound as:

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 / \mathbb{E}\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right], \tag{36}$$

where b is the bias.

3.1.1 Example 4: The loaded die

Imagine you want to test whether a regular six-faces die is loaded –that is, somebody may have painted with lead paint the symbols on one side, to make it land more frequently on the table. Your hypothesis might be that the probabilities of the six possible occurrences of a die throw are not equal, but rather that:

$$\begin{aligned} P(1) &= 1/6 - t/2 \\ P(2) &= P(3) = P(4) = P(5) = 1/6 - t/8 \\ P(6) &= 1/6 + t \end{aligned} \tag{37}$$

The one above is just a model we might use to inject in the probability distribution of the six outcomes the flexibility of describing situations in which one of the sides (the one showing a six) has an enhanced probability of showing up; we could have coded it in a number of other ways, using one parameter or up to five independent parameters modeling the most general biases to the expectations for a regular die (one sixth on each side). With our single parameter t differing from zero, the modified probability of getting a six affects the probability of the other occurrences; in the model above we have imagined that the load decreases more significantly the probability of the throw resulting in a "1" (as that is supposedly the loaded side), and much less so the probabilities of the sides numbered 2, 3, 4, 5. Note that you should not get too enamoured with your models, and this case shows how we build one to get us started with an investigation of a universe of possibilities: the load could be such that true probabilities are more complex.

In our example, the data come from N=20 repeated throws of the die, whereupon you might get:

$$\begin{aligned} x_i &= 1 : 3 \text{ outcomes}, \\ x_i &= 2..5 : 3 \text{ outcomes}, \\ x_i &= 6 : 5 \text{ outcomes}. \end{aligned} \tag{38}$$

The likelihood is the product of probabilities, so to estimate t you write $\breve{}log(L)$ as:

$$-\log(L(t)) = -\sum_{i=1}^{N} \log\left(P\left(x_i, t\right)\right) = -3\log(1/6 - t/2) - 12\log(1/6 - t/8) - 5\log(1/6 + t) \tag{39}$$

Setting the derivative of $\breve{}log(L)$ to zero yields a quadratic equation: $360t^2 - 249t + 16 = 0$. This has one solution in the allowed range of the parameter (to avoid negative probabilities!), $t \in [-\frac{1}{6}, \frac{1}{3}]$: $\hat{t} = 0.072$. Its uncertainty can be obtained by the variance, computed as the inverse of the second derivative of the likelihood. This amounts to ± 0.084. The point estimate of the "load", the MLE, is thus different from zero, but compatible with it. We conclude that the data we collected cannot conclusively establish the presence of a bias.

The example above allows an analytical solution through the maximum likelihood method. Usually our problems are much more complex than that, and we resort to automatic methods to compute the extremum of the likelihood and its curvature at that point.

3.2 The method of least squares

Imagine you have a set of n independent measurements y_i, which we assume to be normally distributed random variables, with different unknown means λ_i and known variances σ_i^2. The y_i can be considered a vector having a joint p.d.f. which is the product of n Gaussians:

$$g\left(y_1, \ldots, y_n; \lambda_1, \ldots, \lambda_n; \sigma_1^2, \ldots \sigma_n^2\right) = \prod_{i=1}^{n} \left(2\pi\sigma_i^2\right)^{-\frac{1}{2}} e^{\frac{-(y_i-\lambda)^2}{2\sigma_i^2}}. \tag{40}$$

Let also λ be a function of x and a set of m parameters θ: $\lambda(x; \theta_1 \ldots \theta_m)$. In other words, λ is the model you want to fit to your data points $y(x)$. We want to find estimates of parameters θ of the model.

If we take the logarithm of the joint p.d.f. we get the log-likelihood function,

$$\log L(\theta) = -\frac{1}{2} \sum_{i=1}^{n} \frac{\left(y_i - \lambda\left(x_i; \theta\right)\right)^2}{\sigma_i^2} + c \tag{41}$$

(c is a constant) which is maximized by finding θ such that the following quantity is minimized:

$$\chi^2(\theta) = \sum_{i=1}^{n} \frac{\left(y_i - \lambda\left(x_i; \theta\right)\right)^2}{\sigma_i^2}. \tag{42}$$

The expression written above near the minimum follows a χ^2 distribution only if the function $\lambda(x; \theta)$ is linear in the parameters θ and if it is the true form from which the y_i were drawn. The method of least squares given above "works" also for non-Gaussian errors σ_i, as long as the y_i are

independent. But it may have worse properties than a full likelihood. If the measurements are not independent, the joint p.d.f. will be a n-dimensional Gaussian. Then the following generalization holds:

$$\chi^2(\theta) = \sum_{i,j=1}^{n} \left(y_i - \lambda\left(x_i;\theta\right)\right) \left(V_{ij}\right)^{-1} \left(y_j - \lambda\left(x_j;\theta\right)\right). \tag{43}$$

Note that unlike the maximum likelihood, writing the χ^2 only requires a unbiased estimate of the variance of a distribution, as that estimator relies on the Gaussian approximation.

3.3 The importance of knowing the properties of your estimators

Issues (and errors hard to trace) may arise in the simplest of calculations, if you do not know the properties of the tools you are working with. Take the simple problem of combining three measurements of the same quantity. Make these be counting rates, *i.e.* counts with Poisson uncertainties:

$$\begin{aligned} A_1 &= 100 \\ A_2 &= 90 \\ A_3 &= 110 \end{aligned} \tag{44}$$

The three measurements above are fully compatible with one another, given that the estimates of their uncertainties are $\sqrt{(A_i)} = 10, 9.5, 10.5$ respectively (if they were not, you should not combine them!). We may thus proceed to average them, obtaining

$$\langle A \rangle = 100.0 \pm 5.77. \tag{45}$$

Now imagine, for the sake of argument, that out of laziness, rather than do the math we used a χ^2 fit to a constant to evaluate $\langle A \rangle$. Surely we would find the same answer as the simple average of the three numbers, right? Actually, no. As is illustrated in Fig. 5, the standard χ^2 fit you can produce, *e.g.*, with call to the "Fit()" method of the ROOT analysis software does not "preserve the area" of the fitted histogram: the integral of the fitting function is lower than the integral of the data. Let us dig a little bit into this matter. This leads us to study the detailed definition of the test statistic we employ in our fits. In general, a χ^2 statistic results from a weighted sum of squares; the weights should be the inverse variances of the true values. Unfortunately, we do not know the latter!

3.3.1 Two chisquareds and a Likelihood

We can investigate why χ^2 minimization does not preserve the area of the original histogram, while a Poisson likelihood does. Firstly, we need to note that there are two possible χ^2 forms for Poisson data. The "standard" definition is called "Pearson's χ^2", which we write as:

$$\chi^2_P = \sum_{i=1}^{k} \frac{(N_i - n)^2}{n}, \tag{46}$$

where n is the best fit value, and N_i are the measurements. Note that n in the denominator introduces a variable weight in the calculation. The other (also known as "modified" χ^2) is called "Neyman's χ^2":

$$\chi^2_N = \sum_{i=1}^{k} \frac{(N_i - n)^2}{N_i}. \tag{47}$$

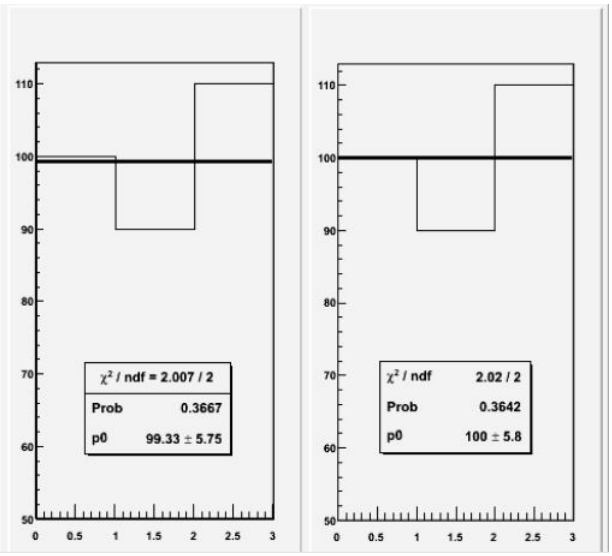

Figure 5: On the left: a χ^2 fit of a three-bin histogram to a constant, which does not 'preserve the area' of the histogram; on the right: a Likelihood fit, which does.

Once again, the weight at the denominator is variable. The difference of the two formulations, though, is that while χ^2_P uses the best-fit variances at the denominator, χ^2_N uses the individual estimated variances. Although both of these least-square estimators have asymptotically a χ^2 distribution, and display optimal properties, they use approximated weights. The result is a pathology: neither definition preserves the area in a fit. χ^2_P overestimates the area, χ^2_N underestimates it. And neither of them offers a unbiased weighted average.

The maximization of the Poisson maximum likelihood,

$$L_P = \prod_{i=1}^{k} \frac{n^{N_i} e^{-n}}{N_i!} \tag{48}$$

instead preserves the area, and obtains exactly the result of the simple average.

As previously mentioned, χ^2_P overestimates the area. In order to verify it, let us compute n from the minimum of χ^2_P, Eq.(46):

$$\begin{aligned} 0 &= \frac{\partial \chi^2_P}{\partial n} = \sum_{i=1}^{k} \frac{2n\,(n - N_i) - (N_i - n)^2}{n^2} \\ 0 &= \sum_{i=1}^{k} \left(n^2 - N_i^2\right) = kn^2 - \sum_{i=1}^{k} N_i^2 \\ \Rightarrow n &= \sqrt{\frac{\sum_{i=1}^{k} N_i^2}{k}}. \end{aligned} \tag{49}$$

n is therefore found to be the square root of the average of squares, which is by force an overestimate of the area of the data histogram. Likewise, we can prove that χ^2_N is underestimating the area. If

we minimize χ^2_N, Eq.(47), we get:

$$\begin{aligned} 0 &= \frac{\partial \chi^2_{\rm N}}{\partial n} = \sum_{i=1}^{k} \frac{2\,(N_i - n)}{N_i} \\ 0 &= \sum_{i=1}^{k} \left[(N_i - n) \prod_{j=1, j\neq i}^{k} N_j \right] = \sum_{i=1}^{k} \left[\prod_{j=1}^{k} N_j - n \prod_{j=1, j\neq i}^{k} N_j \right]. \end{aligned} \tag{50}$$

Note that, as an alternative to the above cumbersome handling of sums and products, we could have chosen to solve for n Eq. (50) above. The last line above implies that:

$$\sum_{i=1}^{k} \prod_{j=1}^{k} N_j = n \sum_{i=1}^{k} \prod_{j=1, j\neq i}^{k} N_j \tag{51}$$

from which we finally get:

$$\frac{1}{n} = \frac{\sum_{i=1}^{k} \prod_{j=1, j\neq i}^{k} N_j}{\sum_{i}^{k} \prod_{j=1}^{k} N_j} = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{N_i}. \tag{52}$$

In this case, the minimum is found for n equal to the harmonic mean of the inputs —which is an underestimate of the arithmetic mean.

At variance with the above estimators, the Poisson likelihood L_P always preserves the area, which can be shown again by minimizing L_P by first taking its logarithm, to find from Eq.(48):

$$\begin{aligned} \ln(L_P) &= \sum_{i=1}^{k} (-n + N_i \ln n - \ln N_i!) \\ 0 &= \frac{\partial \ln(L_P)}{\partial n} = \sum_{i=1}^{k} \left(-1 + \frac{N_i}{n} \right) = -k + \frac{1}{n} \sum_{i=1}^{k} N_i \\ \Rightarrow n &= \frac{\sum_{i=1}^{k} N_i}{k}. \end{aligned} \tag{53}$$

As predicted, the result for n is the arithmetic mean, so likelihood fitting preserves the area.

The three statements above can be validated by inspecting Fig. 6, which reports the results of an exercise performed with pseudo-experiments. We take a $k = 100$-bin histogram and fill each of its bins with random entries from a Poisson distribution of mean μ. Then, we fit the histogram to a constant by minimizing χ^2_P, χ^2_N, $-2ln(L_P)$ in turn. By repeating the above procedure many times, and changing μ in the displayed range, we may study the ratio of the average estimate of μ to the true μ as a function of true μ. We thus observe that the convergence to a unbiased result is slowest for Neyman's χ^2_N, but the bias is significant also for χ^2_P. This observation remains valid regardless of the value of $k > 1$. It is important to keep that in mind when you fit a histogram. For instance, standard ROOT fitting uses Neyman's definition, with $V = N_i$.

Now let us ponder over what we have just seen. What we are doing when we fit a constant through a set of k bin contents is to extract the common, unknown, true value μ from which the entries were generated, by combining the k measurements. We have k Poisson measurement of this true value. Each equivalent measurement should have the same weight in the combination, because

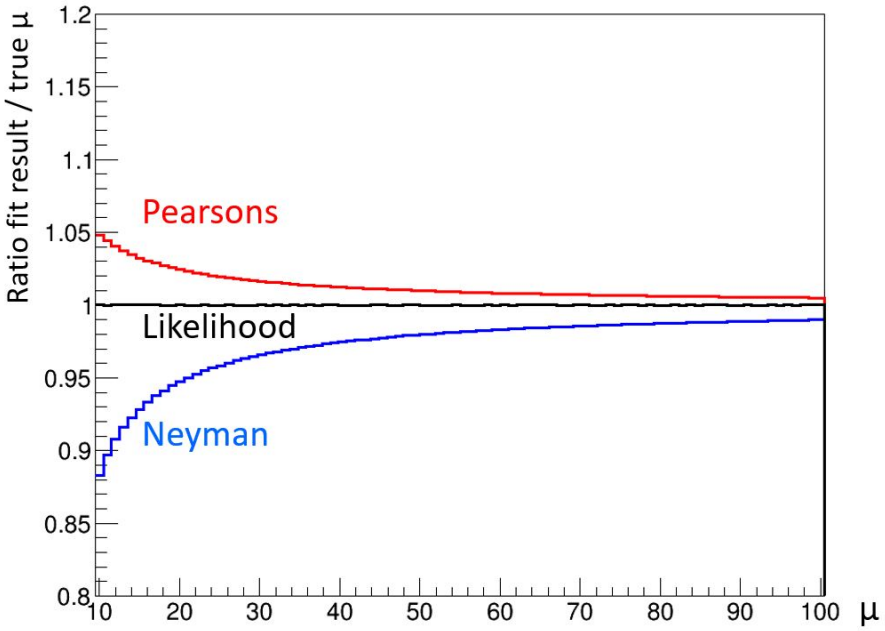

Figure 6: Comparison of the fractional bias in the estimated average content of bins of a histogram filled with counts sampled from a Poisson distribution of mean μ, for a flat distribution. The estimates result from fits with a Pearsons (red) or a Neyman chisquare (blue), or to a Poisson likelihood (black).

each is drawn from a Poisson of mean μ, whose true variance is μ. But having no μ to start with, we must use estimates of the variance as a (inverse) weight. So the χ^2_N gives different weights $1/N_i$ to the different observations. Since negative fluctuations ($N_i < \mu$) have larger weights, the result is downward biased.

What χ^2_P does is different: it uses a common weight for all measurements, but this is of course also an estimate of the true variance $V = \mu$: the denominator of χ^2_N is the fit result for the average, μ^*. Since we minimize χ^2_N to find μ^*, larger denominators get preferred, and we get a positive bias: $\mu^* > \mu$.

All methods discussed above have optimal asymptotic properties: consistency, minimum variance. However, one seldom is in that regime. χ^2_P and χ^2_N also have problems when N_i is small (on-Gaussian errors) or zero (χ^2_N undefined). These drawbacks are solved by grouping bins, at the expense of loss of information. L_P does not suffer from the approximations of the two sums of squares required by χ^2_P and χ^2_N, and it has in general better properties. Cases when the use of a log-likelihood yields problems are rare. Thus, the bottom line is that whenever possible, you should use a likelihood fit.

3.4 RCF bound, efficiency and robustness

Let us recall a few definitions for estimators. The *uniformly minimum variance unbiased estimator* (UMVU) for a parameter is the one which has the minimum possible variance, for any value of the unknown parameter it estimates. The form of the UMVU estimator depends on the distribution of the parameter. Two additional related properties of estimators are efficiency and

robustness.

- *Efficiency*: the ratio of the variance to the minimum variance bound. The smaller the variance of an estimator, the better it is in general, since we can then expect the estimator to be the closest to the true value of the parameter (if there is no bias).
- *Robustness*: more robust estimators are less dependent on deviations from the assumed underlying p.d.f..

Two classic examples of the UMVU are:

- Sample mean: as we discussed *supra*, it is the most used estimator for centre of a distribution. That is not an accident: the sample mean is the UMVU estimator of the mean, if the distribution is Gaussian. However, for non-Gaussian distributions it may not be the best choice.
- Sample mid-range (defined below): it is the UMVU estimator of the mean of a box distribution $U(x)$, defined as $U(x) = 1, x \in [-\mu/2, \mu/2]; U(x) = 0$ elsewhere.

Both sample mean and sample mid-range are efficient (asymptotically the efficiency is 1) for the quoted distribution (Gaussian and box, respectively). But for others, they are not. Robustness describes how much dependence there is in that property: robust estimators have efficiency less dependent on distribution.

3.4.1 Example 5: Choosing an estimator

Let us see what the above notes mean in a practical example: the super-luminal neutrino claim made by the OPERA collaboration in 2011 [3]. The OPERA experiment measured the travel time of 20 neutrinos produced by the CNGS beam at CERN by detecting them in the Gran Sasso cavern in central Italy (at a 730km distance). The difference between travel time of neutrinos and the ratio between distance and speed of light in vacuum, δt, is positive if a neutrino arrives earlier than predicted for light traveling in vacuum, and negative otherwise. Results are reported in Fig. 7, which shows the distribution of δt (in nanoseconds) for the 20 detected neutrinos. Because of the production mechanism from bunched proton interactions with a fixed target, you might expect this to be a box distribution.

OPERA quoted its best estimate of the δt as the sample mean of the 20 shown measurements. We note that this is not the best choice of an estimator for the location of the center of a square distribution. In the cited reference, OPERA quoted the following result: $\langle \delta t \rangle = 62.1 \pm 3.7 ns$.

As we noted already *supra*, the UMVU estimator for the box distribution is the sample mid-range, $\delta t = \frac{(t_{max}+t_{min})}{2}$. It is easy to understand why sample mid-range is better than sample mean when estimating the center of a box distribution: once you pick the *extrema* of the set of values, the rest of the data carries no information on the box location; they only add noise, to which the sample mean is exposed. Furthermore, the larger N is, the larger is the disadvantage of sample mean over sample mid-range in this case.

Now, let us show how the quoted result could have been improved by choosing the UMVU estimator, the sample mid-range. 100,000 n=20-entries histograms where used to mimic a repetition

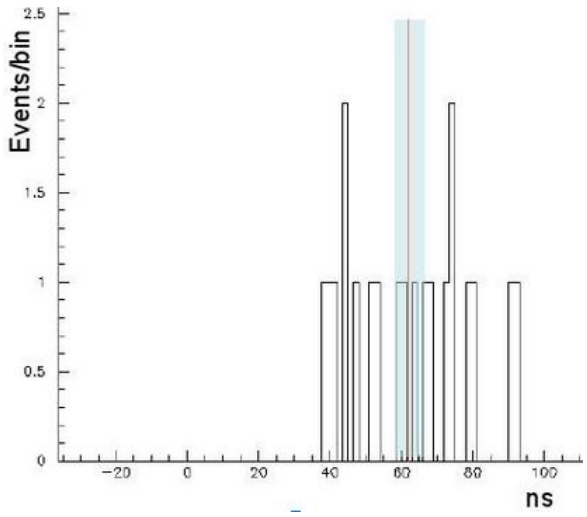

Figure 7: Distribution of δt (in nanoseconds) of individual neutrinos sent from narrow bunches by CNGS to OPERA.

of the OPERA measurement, with data distributed uniformly in $[-25 : 25]$ ns. It is important to note here that in order to get the most precise inference from the data, it is necessary to impose the condition that you include in your calculations same-sizes data sets, all of 20 elements. We shall justify this statement *infra* (Sec. 7), when we will discuss *conditioning* and *ancillary statistics*.

From the large number of toys generated as described above we may observe how the sample mean is asymptotically distributed as a Gaussian, as shown in Fig. 8. For 20 events this is already a good approximation; the expected width of the p.d.f. of the estimator is 3.24 ns. That uncertainty is consistent with the quoted OPERA result.
We can also observe that the distribution of the sample mid-point estimator $P(n\delta t)$, with the latter defined by

$$\delta t = \frac{(t_{max} + t_{min})}{2} \tag{54}$$

is asymptotically a Laplace distribution (see again Fig. 8), which consists in two negative exponentials joined at the origin. The distribution is visibly narrower than the Gaussian describing the sample mean, and its root mean square is of 1.66 ns. This suggests that if OPERA had used the UMVU estimator for their problem, the mid-point, they would have halved their statistical uncertainty:

$$\begin{aligned} \langle \delta t \rangle &= 62.1 \pm 3.7 ns \\ \langle \delta t \rangle &= 5.2 \pm 1.7 ns. \end{aligned} \tag{55}$$

We remind here the expression of the Laplace distribution:

$$f(x) = \frac{b}{2} e^{-\frac{|x-\mu|}{b}}. \tag{56}$$

Now, although the conclusions above are correct if the underlying p.d.f. of the data is exactly a box distribution, things change rapidly if we look at the real problem in more detail. Each timing measurement, before the ±25 ns random offset, is not exactly equal to the others, due to additional random smearings; in particular, the proton bunch has a peaked shape with a 3 ns FWHM (see

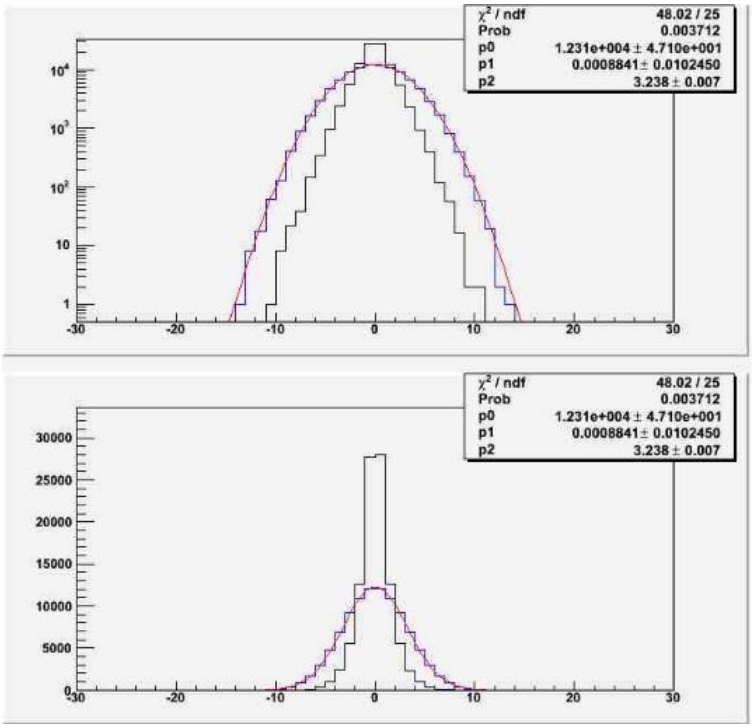

Figure 8: Sample mean and sample mid-range distributions from pseudo-experiments of 20 timing measurements sampled from a 50 ns-wide box distribution. Top: semi-log vertical axis; bottom: linear vertical axis. A fit to the sample mean distribution is well approximated by a Gaussian (red curve).

Fig. 7). Other small effects contribute to smear randomly each timing measurement. There may also be biases, such as fixed offsets due to imprecise corrections made to the δt determination. However, these systematic uncertainties do not affect our conclusions, because they do not change the shape of the p.d.f.. But the random smearings do affect our conclusions regarding the least variance estimator, since they change the p.d.f. of the data.

Thus, we may try to model the additional smearings to have more insight in the question we raised. One may assume that the smearings are Gaussian. The real p.d.f. from which the 20 timing measurements are drawn is now a convolution of a Gaussian smearing with a box distribution. By inserting that modification in the generation of toys we may study its effect. It transpires that, with 20-event samples, a Gaussian smearing with 6 ns width is enough to make the expected variance equal for the two estimators. For larger smearing the sampling distribution resembles more and more a Gaussian overall, with the result that the UMVU returns to be the sample mean. Timing smearings in OPERA were likely around 6ns or only slightly smaller. So in the OPERA experiment using the sample mean was not totally erroneous after all. What the example teaches us, however, is that one should definitely study such effects before making a decision on what estimator to employ in a given problem.

4. Error propagation and weighted averages in the presence of correlations

4.1 Common scale error between two measurements

Let us consider the least-square minimization of a combination of two measurements of the same physical quantity k, for which the covariance terms be all known. In the first case, let there be a *common offset error* σ_c. A possible parametrization of the covariance matrix corresponding to this scenario is given below. By computing its inverse, we may combine the two measurements x_1, x_2 with the least squares method:

$$V = \begin{pmatrix} \sigma_1^2 + \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_2^2 + \sigma_c^2 \end{pmatrix} \Rightarrow V^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 + \left(\sigma_1^2 + \sigma_2^2\right)\sigma_c^2} \begin{pmatrix} \sigma_2^2 + \sigma_c^2 & -\sigma_c^2 \\ -\sigma_c^2 & \sigma_1^2 + \sigma_c^2 \end{pmatrix}, \tag{57}$$

from which we write the χ^2 as

$$\chi^2 = \frac{(x_1 - k)^2\left(\sigma_2^2 + \sigma_c^2\right) + (x_2 - k)^2\left(\sigma_1^2 + \sigma_c^2\right) - 2\,(x_1 - k)\,(x_2 - k)\,\sigma_c^2}{\sigma_1^2\sigma_2^2 + \left(\sigma_1^2 + \sigma_2^2\right)\sigma_c^2}. \tag{58}$$

The minimization of the above expression leads to the following expressions for the best estimate of k and its standard deviation:

$$\begin{aligned} \hat{k} &= \frac{x_1\sigma_2^2 + x_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \\ \sigma^2(\hat{k}) &= \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} + \sigma_c^2. \end{aligned} \tag{59}$$

The best fit value does not depend on σ_c, and corresponds to the weighted average of the results when the individual variances σ_{12} and σ_{22} are used. This result is what we expected, and all is good here.

4.2 Common normalization error

In the second case we take two measurements of k having a common scale error σ_f. That means that the common error acts as a multiplier to the true value. In that case the variance, its inverse, and the least-square statistic might be written as follows:

$$\begin{aligned} V &= \begin{pmatrix} \sigma_1^2 + x_1^2\sigma_f^2 & x_1x_2\sigma_f^2 \\ x_1x_2\sigma_f^2 & \sigma_2^2 + x_2^2\sigma_f^2 \end{pmatrix} \Rightarrow V^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 + \left(x_1^2\sigma_2^2 + x_2^2\sigma_1^2\right)\sigma_f^2} \begin{pmatrix} \sigma_2^2 + x_2^2\sigma_f^2 & -x_1x_2\sigma_f^2 \\ -x_1x_2\sigma_f^2 & \sigma_1^2 + x_1^2\sigma_f^2 \end{pmatrix} \\ \chi^2 &= \frac{(x_1 - k)^2\left(\sigma_2^2 + x_2^2\sigma_f^2\right) + (x_2 - k)^2\left(\sigma_1^2 + x_1^2\sigma_f^2\right) - 2\,(x_1 - k)\,(x_2 - k)\,x_1x_2\sigma_f^2}{\sigma_1^2\sigma_2^2 + \left(x_1^2\sigma_2^2 + x_2^2\sigma_1^2\right)\sigma_f^2}. \end{aligned} \tag{60}$$

This time the minimization produces the following results for the best estimate of the weighted average and its variance:

$$\begin{aligned} \hat{k} &= \frac{x_1\sigma_2^2 + x_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2 + (x_1 - x_2)^2\,\sigma_f^2} \\ \sigma^2(\hat{k}) &= \frac{\sigma_1^2\sigma_2^2 + \left(x_1\sigma_2^2 + x_2\sigma_1^2\right)\sigma_f^2}{\sigma_1^2 + \sigma_2^2 + (x_1 - x_2)^2\,\sigma_f^2}. \end{aligned} \tag{61}$$

Before we discuss the above formulas, let us test them on a simple case:

$$\begin{aligned} x_1 &= 10 \pm 0.5 \\ x_2 &= 11 \pm 0.5 \\ \sigma_f &= 20\%. \end{aligned} \tag{62}$$

This yields the following disturbing result: $k = 9.72 \pm 0.51$! So, what is going on? Why is that outside the bounds of the two measurements? Let us shed some light on the matter, since the fact that averaging two measurements with the least-square method may yield a result outside their range requires more investigation. To try and understand what is going on, we may rewrite the result by dividing it by the weighted average result obtained ignoring the scale correlation:

$$\begin{aligned} \hat{k} &= \frac{x_1\sigma_2^2 + x_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2 + (x_1 - x_2)^2\,\sigma_f^2} \\ \bar{x} &= \frac{x_1\sigma_2^2 + x_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \\ \Rightarrow \frac{\hat{k}}{\bar{x}} &= \frac{1}{1 + \frac{(x_1-x_2)^2}{\sigma_1^2+\sigma_2^2}\sigma_f^2}. \end{aligned} \tag{63}$$

If the two measurements differ, their squared difference divided by the sum of the individual variances plays a role in the denominator. In that case the fit "squeezes the scale" by an amount modulated by σ_f^2 in order to minimize the χ^2. However, the counter-intuitive nature of the result suggests that we should dig further in the matter.

Let us consider a more general form of correlation terms in the important example of taking the average of two correlated measurements:

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{64}$$

The least-square estimators provide the following result for the weighted average [4]:

$$\hat{x} = wx_1 + (1-w)x_2 = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}x_1 + \frac{\sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}x_2 \tag{65}$$

whose inverse variance is:

$$\frac{1}{\sigma^2} = \frac{1}{1-\rho^2}\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - \frac{2\rho}{\sigma_1\sigma_2}\right) = \frac{1}{\sigma_1^2} + \frac{1}{1-\rho^2}\left(\frac{\rho}{\sigma_1} - \frac{1}{\sigma_2}\right)^2. \tag{66}$$

From the above expression we see that once we measure x_1 with variance σ_1, a second measurement x_2 of the same quantity will reduce the variance of the average, unless $\rho = \frac{\sigma_1}{\sigma_2}$; in the latter case, the second result will be entirely irrelevant to average and uncertainty. But what happens if $\rho > \frac{\sigma_1}{\sigma_2}$? In that case the weight w gets negative, and the average goes outside the "psychological" bound $[x_1, x_2]$. The reason for this behavior is that with a large positive correlation the two results are

likely to lie on the same side of the true value! On which side they are predicted to be by the least-square minimization depends on which result has the smallest variance.

It seems a paradox, but it is not. The reason why we cannot digest the fact that the best estimate of the true value μ be outside of the range of the two measurements is our incapability of understanding intuitively the mechanism of large correlations between our measurements.

4.2.1 Illustration with a conversation

In order to make sense out of the seemingly weird result we derived *supra*, let us illustrate the logic behind it with a hypothetical conversation between John and Jane.

John: "I took a measurement, got x_1. I now am going to take a second measurement x_2 which has a larger variance than the first. Do you mean to say I will more likely get $x_2 > x_1$ if $\mu < x_1$, and $x_2 < x_1$ if $\mu > x_1$?"

Jane: "That is correct. Your second measurement goes along with the first, because your experimental conditions made the two highly correlated and x_1 is more precise."

John: "But that means my second measurement is utterly useless!"

Jane: "Wrong. It will in general reduce the combined variance. Except for the very special case of $\rho = \sigma_1/\sigma_2$, the weighted average will converge to the true μ. Least-square estimators are consistent!".

John: "I still can't figure out how on earth the average of two numbers can be outside of their range. It just fights with my common sense."

Jane: "You need to think in probabilistic terms. Look at this error ellipse: it is thin and tilted (high correlation, large difference in variances)."

John: "Okay, so?"

Jane: "Please, would you pick a few points at random within the ellipse?"

John: "Done. Now what?" (see Fig. 9)

Jane: "Now please tell me whether they are mostly on the same side (orange rectangles) or on different sides (pink rectangles) of the true value."

John: "Ah! Sure, all but one are on orange areas".

Jane: "That's because their correlation makes them likely to "go along" with one another. And I can actually make it even easier for you. Take a two-dimensional plane, draw axes, draw the bisector: the latter represents the possible values of μ. Now draw the error ellipse around a point

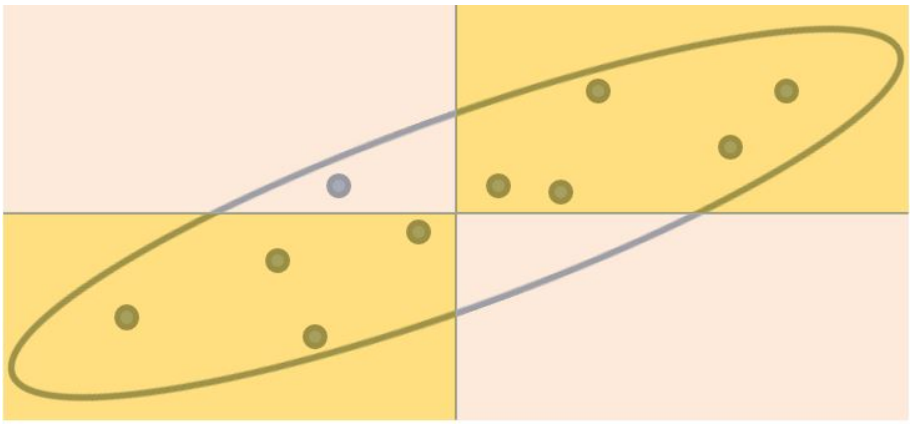

Figure 9: Two correlated measurements of the same quantity may produce the situation shown here. An overestimate in one of the measurements will likely be accompanied by an overestimate in the other.

of the diagonal. Any point, we will move it later." (see Fig. 10)

John: "Done. Now what?"

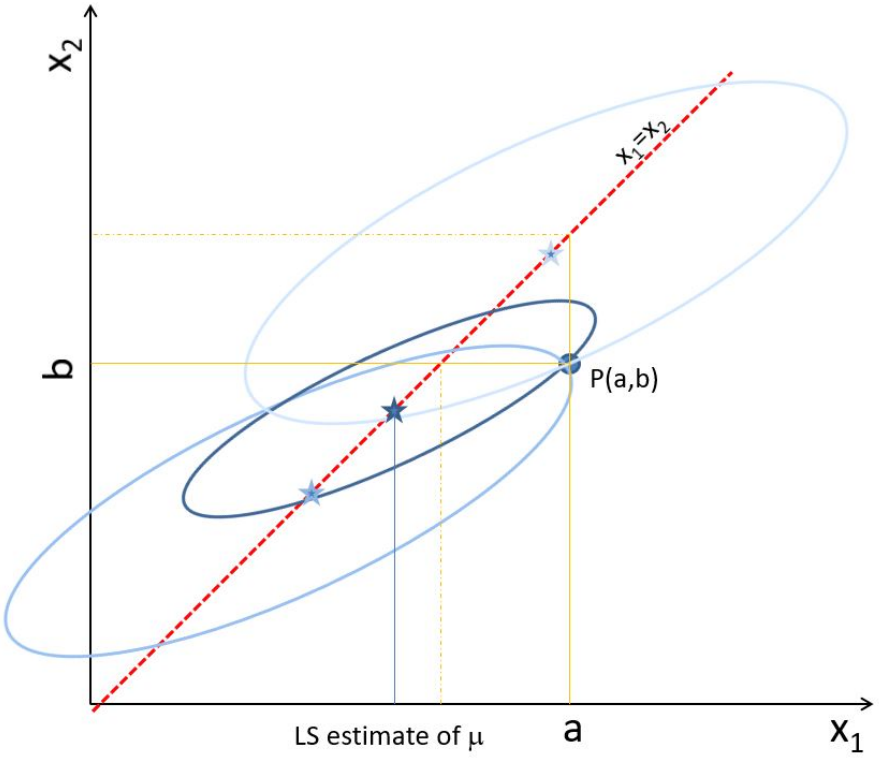

Figure 10: Geometrical construction illustrating the minimization of the χ^2 for two correlated measurements. The best-fit solution corresponds to the dark-blue star, which generates an error ellipse which passes through the measurement point (a, b), where its tangent is parallel to the bisector of the axes.

Jane: "Now enter your measurements $x = a, y = b$. That corresponds to picking a point $P(a, b)$ in the plane. Suppose you got $a > b$: you are on the lower right triangle of the plane. To find the best estimate of μ, move the ellipse by keeping its center along the diagonal, and try to scale it also, such that you intercept the measurement point P."

John: "But there's an infinity of ellipses that fulfil that requirement".

Jane: "That's correct. But we are only interested in the smallest ellipse! Its center will give us the best estimate of μ, given (a, b), the ratio of their variances, and their correlation."

John: "Oooh! Now I see it! It is bound to be outside of the interval!"

Jane: "Well, that is not true: it is outside of the interval only because the ellipse you have drawn is thin and its angle with the diagonal is significant. In general, the result depends on how correlated the measurements are (how thin is the ellipse) as well as on how different the variances are (how big is the angle of its major axis with the diagonal). Note also that in order for the "result outside bounds" to occur, the correlation must be positive!

4.2.2 Example 6: How our eye fails to fit correlated data

With the following example we want to show how our eye can be misled when looking at data points that carry correlated uncertainties. The graph in Fig. 11 shows the results of a study of data from the CDF experiment at the Fermilab Tevatron collider, a machine that collided protons with anti-protons at $\sqrt{s} = 1.8$ TeV in the 1990s. The points represent measurements of the differential cross section of processes producing a high-mass pair of hadronic jets. The prediction depends on the specific set of parton distribution functions (PDF) that was used in their calculation; the figure shows the data subtracted by predictions produced using the default set (CTEQ4M) and divided by them, so the CTEQ4M-based prediction for this fractional difference is the horizontal line at zero; other PDF sets produce different curves for the fractional difference. The data are subjected to statistical fluctuations due to Poisson statistics, which are described by the vertical bars around the data points, and by correlated systematic uncertainties due to the jet energy scale, fragmentation uncertainties, and other nuisance parameters. The latter are represented by the thick yellow band.

By looking at the figure, could you guess which of the PDF models shown in the graph is the best fit to the data: CTEQ4M (horizontal line at 0.0) or MRST (dotted curve)? It would seem as if CTEQ4M produces a better fit, as MRST is farther away from the data points. But the presence of large correlations makes the normalization much less important than the shape, as shifting all points down by the width of the yellow band causes an increase of the χ^2 by only one unit, not by a number of units equal to the number of points!

If one calculates the probabilities of the data given different models, one indeed finds that the MRST hypothesis yields a 30 times higher p-value (0.0032) of the data than the CTEQ4M hypothesis (0.00011), despite being on average "farther away" from the experimental data points. What we deduce from all this is that we must be careful with least-square fits in the presence of

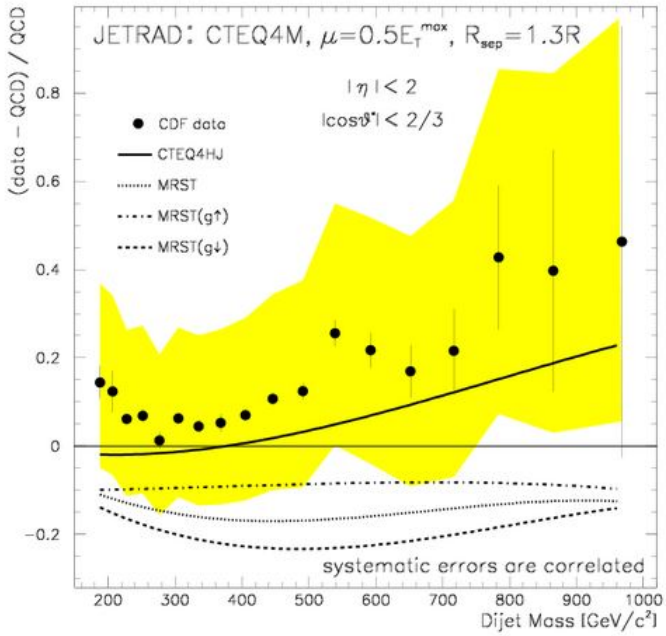

Figure 11: Residuals of measured rates of dijet events as a function of dijet mass from CTEQ4M-based theory predictions (black points). Statistical uncertainties are described by vertical bars, and correlated systematic uncertainties are shown by the yellow band [5].

large common systematics. We should, in other words, be careful to avoid trusting our eye when data points carry significant bin-to-bin correlations.

5. Choosing a model

Often in high energy physics, astro-hep, and similar fundamental research studies we do not know what is the true functional form that data are drawn from. In specific cases, we can use Monte Carlo simulations, but not always. Extracting inference from a spectrum is, thus, limited by our imprecise knowledge about background processes. If we see a deformation in the spectrum, we might want to ascribe it to the signal of some new process, but we cannot do it until we have a good understanding of the overall shape of the data in the absence of that new process. At the LHC and elsewhere we routinely use, *e.g.*, invariant mass spectra constructed with the observable characteristics of collision events to search for new particles, and to do so we sometimes employ educated guesses of the distribution of the background: when we do not have a good *a priori* model of the reconstructed mass spectrum of background processes, families of possible "background shapes" are used to investigate various possibilities for the PDF of the null hypothesis. The functional modeling of data distributions is thus a crucial problem in HEP and related fields.

5.1 Fisher's F-test

Model selection is a huge topic in Statistics, and we have no time or space to even brush the surface here. But we can at least discuss a simple yet rigorous and powerful test often used in HEP

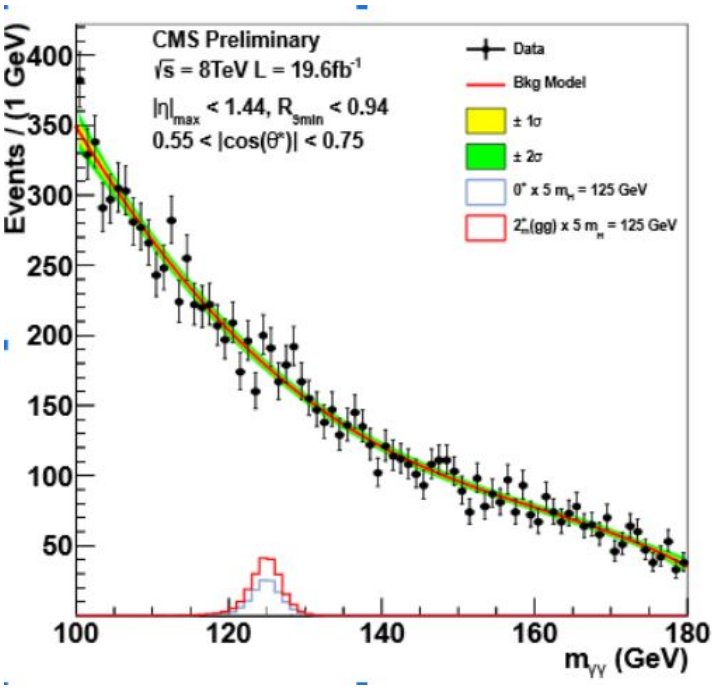

Figure 12: Example of a mass distribution where the CMS experiment at the CERN Large Hadron Collider searches for a Higgs boson decay. The data are events with two energetic photons, which –in the case a Higgs boson is present– produce a Gaussian-like bump at 125 GeV. The background, however, is not precisely known, so it is modeled with a smooth function; the green and yellow bands show uncertainties in the model.

to pick a model within a set of possibilities.

Suppose you have no clue of the real functional form followed by your data (n points), or even suppose you know only its general form (*e.g.* polynomial, but you do not know the degree). Then, you may try a function $f_1(x;p_1)$ and find it produces an overall good fit; however, you are unsatisfied about some local feature of the data that appear to be systematically missed by the model. You may be tempted to try a more complex function, usually by adding one or more parameters to f. This always improves the absolute χ^2, as long as the new model "embeds" the old one (the latter means that given any choice of $\{p_1\}$, there exists a set $\{p_2\}$ such that $f_1(x;\{p_1\}) == f_2(x;\{p_2\})$ for all x).

The question is however: how to decide whether f_2 is more motivated than f_1, or rather, that the added parameters are doing something of value to your model? It is important to not use your eye! Being led by what you consider a good fit by eye may result in choosing more complicated functions than necessary to model your data, with the result that your statistical uncertainty (*e.g.* on an extrapolation or interpolation of the function) may abnormally shrink, at the expense of a modeling systematics which you have little hope to estimate correctly. Instead, you may use the Fisher F-test. The statistic F has a Fisher distribution $f(F)$ if the added parameter is not improving the model.

The F statistic can be computed from two models f_1, f_2 involving a number v_1, v_2 of parameters respectively, by the equation below, where the y_i are the observed values of the bins of your

distributions:

$$F = \frac{\frac{\sum_i (y_i - f_1(x_i))^2 - \sum_i (y_i - f_2(x_i))^2}{p_2 - p_1}}{\frac{\sum_i (y_i - f_2(x_i))^2}{n - p_2}}. \quad (67)$$

If the more complex model f_2 is not required by the data, F distributes according to a Fisher distribution:

$$f\,(F; v_1, v_2) = \frac{v_1^{v_1/2} v_2^{v_2/2} \Gamma\left(\frac{v_1+v_2}{2}\right)}{\Gamma\,(v_1/2)\,\Gamma\,(v_2/2)} \frac{F^{\frac{v_1-1}{2}}}{(v_1 + v_2 F)^{\frac{v_1+v_2}{2}}}. \quad (68)$$

5.1.1 Example 7: Use of the F-test

Imagine we have the data shown in Fig. 13, and we need to pick a functional form to model the underlying p.d.f.. At first sight, any of the four choices shown produces a meaningful fit and P-values of the respective χ^2 are all reasonable - $(0.085, 0.485, 0.446, 0.390)$. What the F-test does is to allow us to determine whether the additional parameter in going from a constant to a line, or from a line to a quadratic, or from a quadratic to a cubic, is needed.

We need to pre-define a size α for our test[3]: we will reject the "null hypothesis" that the additional parameter is useless if $p < \alpha$. Let us pick $\alpha = 0.1$, an arbitrary but common choice. We define p as the probability that we observe a F value at least as extreme as the one in the data, assuming it is drawn from a Fisher distribution with the corresponding degrees of freedom. Note that by doing this we are implicitly also selecting a "region of interest" (large values of F). Which of the four models shown in Fig. 13 do you consider the most appropriate to fit the data? The linear? The quadratic? The cubic? And would your choice change if $\alpha = 0.318$ (1-sigma), or 0.05?

The test between constant and line (see Fig. **??** left) yields p=0.0315: there is evidence, according to our choice of α, against the null hypothesis (which implies that the additional parameter is useless, and a constant is all we need), so we reject the constant p.d.f. and take the linear fit. The test between linear and quadratic fit yields p=0.730: there is no evidence against the null hypothesis (this time corresponding to the linear model being appropriate). We therefore keep the linear model. Note that as expected, also a test of the quadratic (null hypothesis) versus the cubic model produces a high probability for the quadratic model (p=0.431); the serial application of the Fisher test to increasingly complex models however already demands us to choose the simplest model passing our requirement on α.

It is important to ensure that your set of models be *embedded* each in all other more complex ones: that means that there exists one parameter of the more complex model which, if fixed at some value, have it correspond to the less complex one, as is the case for a series of polynomials of increasing degree. If this condition is not satisfied, the Fisher test cannot be applied.

6. Confidence intervals

6.1 The simplest confidence interval: ± 1 standard deviations

The standard deviation is used in simple applications as a measure of the uncertainty of a point estimate. For example, consider N observations $X = \{x_1\}$ of a random variable x with hypothesized

[3]We discuss hypothesis tests *infra*, in Sec. 9.

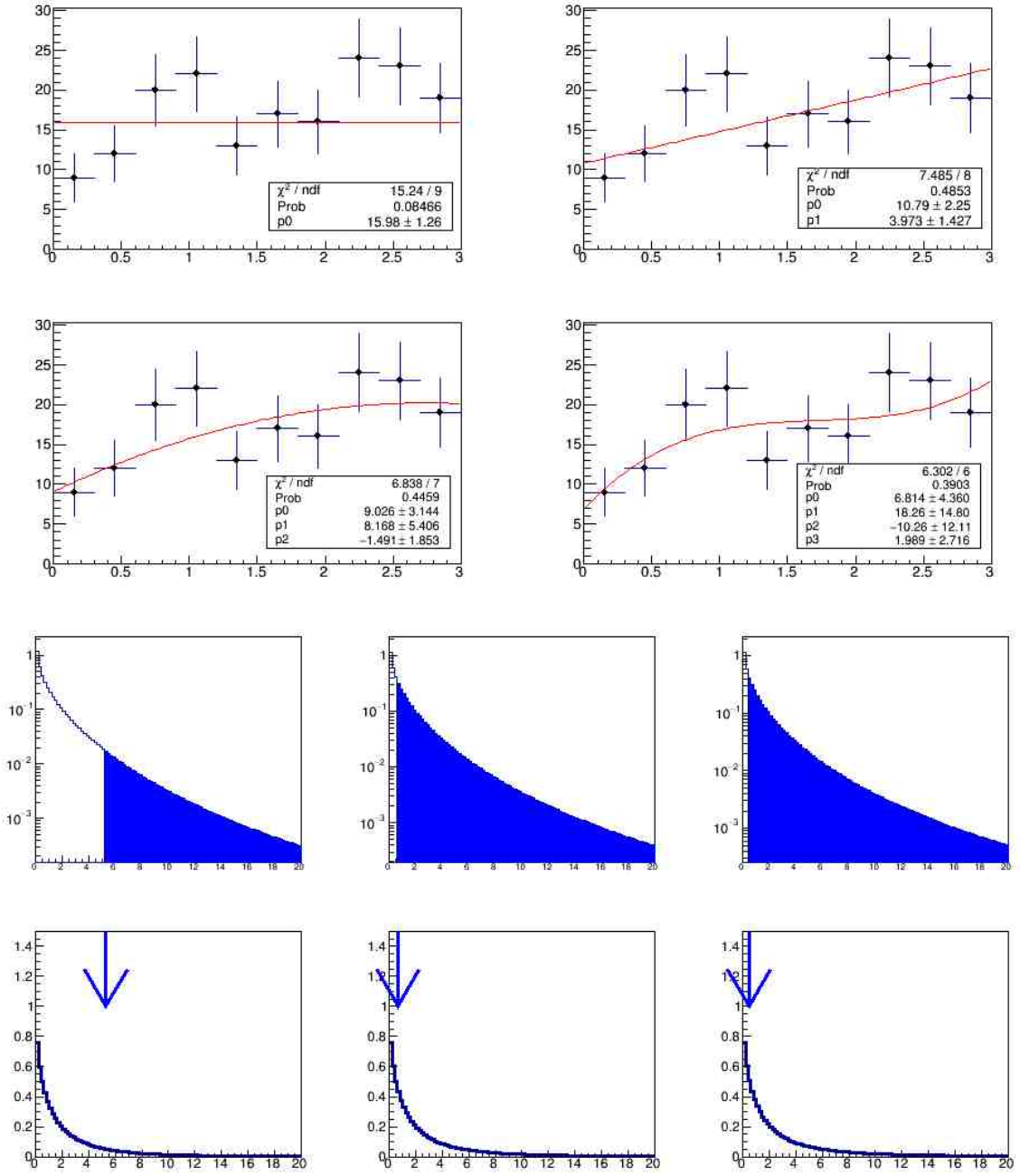

Figure 13: Top: data drawn from a linearly growing distribution in the considered interval are fit to four possible models: constant, linear, quadratic, cubic. Bottom: The two panels on the left, center, and right show respectively the results of the F test comparing constant to linear model, linear to quadratic model, and quadratic to cubic model. In all cases the top panel shows the F distribution, and the bottom panel shows its tail integral.

pdf $f(x;\theta)$, with θ unknown. The set X allows to construct an estimator $\theta^*(X)$. Using an analytic method, or the Rao-Cramer-Frechet bound, or MC sampling, one can estimate the standard deviation

of θ^*. The value $\theta^* \pm \sigma^*_{\theta^*}$ is then reported. What does this mean?

The answer to the above question is not trivial: we need to be careful. What the above means, if we want to be precise, is that in repeated estimates based on the same number N of observations of x, θ^* would distribute according to a p.d.f. $G(\theta^*)$ centered around a true value θ with a true standard deviation σ_{θ^*}, respectively estimated by θ^* and $\sigma^*_{\theta^*}$. In the large sample limit $\boldsymbol{G}()$ is a (multi-dimensional) Gaussian function.

The above would already be a complicated concoction, but it is actually made worse by the fact that in fundamental physics we often look for new, rare phenomena, and we are thus very far from the regime where good asymptotic properties of our statistics simplify matters. In most interesting cases for physics $G()$ is not Gaussian, the large sample limit does not hold, 1-sigma intervals do not cover 68.3% of the time the true parameter, and we have better be a bit more tidy in constructing intervals. But we need to have a hunch of the pdf $f(x;\theta)$ to start with!

In a typical experiment, we do not measure a physical quantity directly. *e.g.*, we do not read off the top quark mass from an analog meter. We rather cook up an estimator and extract its value (the estimate) x; we may then use a map connecting it to the true value we want, which we shall call μ in the following. We may know the distribution of x, but how to infer bounds on the true value μ from observed data x? And how do we make them sound, *i.e.*, well-behaved intervals?

6.2 Neyman's confidence interval recipe

An answer to the above questions is provided by the Neyman construction. The recipe to construct a confidence belt is the following:

1. Draw a two-dimensional graph where on the horizontal axis you put possible values of your estimator $\hat{x}$, and on the vertical axis you put the possible values of the quantity being estimated μ.

2. Specify a model which provides the probability density function of a particular observable x being found, for each value of the unknown parameter of interest: $p(x \mid \mu)$;

3. Also choose a Type-I error rate α (*e.g.* 31.8%, or 5%);

4. For each μ, draw a horizontal acceptance interval $[x_{1'}, x_2]$ such that $p(x \in [x_1, x_2] \mid \mu) = 1 - \alpha$. There are infinitely many ways of doing this: it all depends on what you want from your data, as

   - for upper limits, you choose the acceptance interval by integrating the p.d.f. from x_{low} to infinity;
   - for lower limits, you will do the opposite;
   - you might want to choose central intervals, by finding x_1 and x_2 such that the integral of the p.d.f. on the left of x_1 and on the right of x_2 equals $\alpha/2$;
   - a different possibility is to look for x_1 and x_2 such that their span the shortest connected interval such that $\int_{x_1}^{x_2} f(x \mid \mu) = 1 - \alpha$;
   - or you might want to determine the regions of the p.d.f. that have highest relative frequency, always adding them up until you get the wanted confidence level.

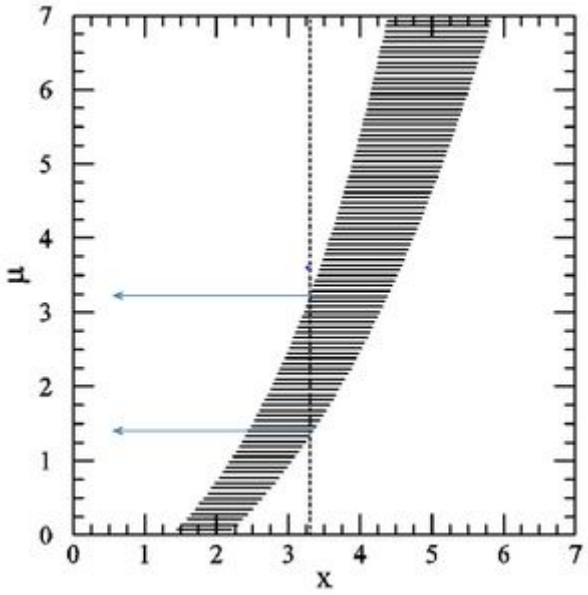

Figure 14: Neyman's confidence belt construction. See the text for detail.

In general, you have to choose an ordering principle such as one of those above, or others.

This allows you to construct what is called a confidence belt. Note that the recipe does not guarantee that this belt be simply connected: it could be composed of different areas of the plane spanned by measurements x and true values μ. But this does not change the validity of the construction, nor the relative simplicity of the procedure by means of which you can extract a confidence interval.

Upon performing an experiment, you measure $x = x^*$. Yo can then draw a vertical line through it. The vertical confidence interval $[\mu_1, \mu_2]$ (with Confidence Level (C.L.) = $1-\alpha$) is then the union of all values of μ for which the corresponding acceptance interval is intercepted by the vertical line (see Fig. 14).

The above recipe guarantees that the confidence intervals you obtain will have exact coverage only in the case when the p.d.f. $f(x \mid \mu)$ is continuous; instead, if the p.d.f. has a discrete support, it may be impossible to obtain extrema of the horizontal intervals that contain the required confidence level. In those cases, the prescription is to overrun the requested C.L. $1-\alpha$. The resulting confidence intervals will then possibly *overcover*, *i.e.* they will be more conservative than what the chosen value of α would require.

6.3 Important notions on confidence intervals

At this point in the lectures there is an important concept to stress. This can be done with an example. If we asked you to define what a vector is, you might be likely to come up with a number of possible definitions that describe the properties of a vector. But the most precise definition, which allows us to really appreciate what those properties are, involves the definition of the space of all possible vectors: A vector is an element of a vector space –a set with certain properties. It is the set that possesses those properties, and the vectors inherit them by belonging to it.

In close similarity, a confidence interval is best defined to be "an element of a confidence set", the latter being a set of intervals defined to have the property of frequentist coverage under sampling. Any given confidence interval, even ones constructed through the rigorous application of Neyman's recipe, does not necessarily cover the unknown true value of μ with any given C.L.; nor does it in fact make any sense at all, in a classical sense, to speak of the probability that a fixed interval contains μ –in frequentist statistics we cannot speak of the probability density of a parameter of nature. The property of coverage belongs to the set, not to any of its elements.

We can further stress the important point above as follows. Let the unknown true value of μ be μ_t. In repeated experiments, the confidence intervals obtained will have different endpoints $[\mu_1, \mu_2]$, depending on the random variable x. A fraction C.L. $= 1 - \alpha$ of intervals obtained by Neyman's contruction will contain ("cover") the fixed but unknown μ_t : $P(\mu_t \in [\mu_1, \mu_2]) = 1 - \alpha$. It is important thus to realize two facts:

- the random variables in the equation are μ_1 and μ_2, and not μ_t!
- Coverage is a property of the set, not of an individual interval. For a frequentist, the interval either covers or does not cover the true value, regardless of α. Thus a classic false statement you should avoid making is that "the probability that the true value is within μ_1 and μ_2 is 68%".

What we can say, instead, is that the confidence interval does consist of those values of μ for which the observed x is among the most probable (in sense specified by ordering principle) to be observed. But you should also note that when we speak of "repeated sampling" we do not require one to perform the same experiment multiple times for the confidence interval to have the stated properties. We may in fact put together different experiments and conditions in constructing the confidence belt, as long as they all have the same α; if we do it correctly, the intervals will properly cover.

6.4 The binomial ratio

As we mentioned above, over-coverage may arise when the p.d.f. has a discrete support. When constructing Neyman's confidence belt one then errors conservatively by including x values (according to one's ordering rule) until $\sum_i p(x_i \mid \mu) > 1 - \alpha$. In this case, the interval $[\mu_1,\mu_2]$ derived from an estimate x will overcover. A classic example of this are the *binomial uncertainty bars* that may be obtained from a binomial experiment in the case of a small number N of trials. The (true) variance on the binomial ratio ρ is $\sigma = \mathrm{sqrt}(\rho(1-\rho)/N)$, but if we compute its estimate from N trials, where ρ^* is the fraction of successes, we might be tempted to write is as $\sigma^* = \mathrm{sqrt}\,(\rho^*\,(1-\rho^*)\,/\mathrm{N})$. This is called the Wald interval, and it has some unwanted properties. In fact, it fails badly for small N and $\rho^* \simeq 0, 1$, when large undercoverage can be proven.

Rather than using the Wald intervals, one may apply Neyman's recipe, obtaining the so-called Clopper-Pearson intervals if a central interval ordering rule is chosen. Clopper-Pearson intervals overcover sizeably for some values of the successes/trials, but they never undercover, as expected. Fig. 15 shows the coverage probability as a function of the true value of ρ for $N = 10$.

There have been a number of studies aimed at improving the properties of confidence intervals for the binomial ratio problem. One simple recipe that avoids the problems of Wald intervals while

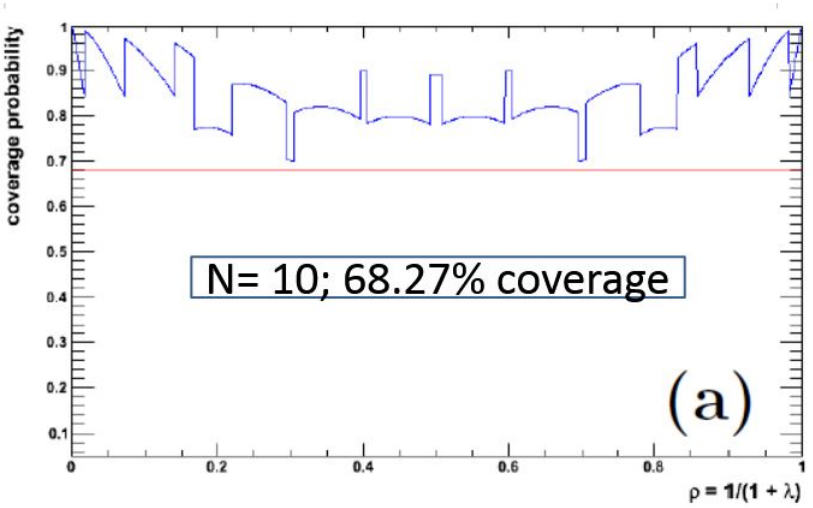

Figure 15: Coverage probability of Clopper-Pearsons error bars for a small number of trials ($N = 10$).

requiring no complicated procedure is the one of the Wilson score interval. Already in 1927, Edwin Wilson proposed the following recipe.

For the lower endpoint of the interval, one uses the lowest value ρ_1 such that

$$\rho_1 + Z_{\alpha/2}\sqrt{\rho_1\left(1-\rho_1\right)/n_{\rm tot}}$$

contains $\hat{\rho}$. Analogously for the upper endpoint, one uses the largest value ρ_2 such that

$$\rho_2 - Z_{\alpha/2}\sqrt{\rho_2\left(1-\rho_2\right)/n_{\rm tot}}$$

contains $\hat{\rho}$. Letting $T = \left(Z_{\alpha/2}\right)^2/n_{\rm tot}$, this leads to a quadratic equation in ρ for the endpoints, $(\rho-\hat{\rho})^2 = T\rho(1-\rho)$, with solutions

$$\rho = \frac{\hat{\rho}+T/2}{1+T} \pm \frac{\sqrt{\hat{\rho}(1-\hat{\rho})T+T^2/4}}{1+T}.$$

The above endpoints form the Wilson score interval; they are simple to extract and should be favored when a quick estimate of the uncertainty bar on a successes-over-trials estimate is needed.
In the formulas above, we have used the following conventions:

$$Z_{\alpha/2} = \Phi^{-1}(1-\alpha/2) = -\Phi^{-1}(\alpha/2)$$

where

$$\Phi(Z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp\left(-t^2/2\right)dt = \frac{1+\mathrm{erf}(Z/\sqrt{2})}{2}$$

so that

$$Z = \sqrt{2}\,\mathrm{erf}^{-1}(1-\alpha)$$

e.g., $Z_{\alpha/2} = 1$ for $\alpha/2 = 0.159$, and $Z_{\alpha/2} = 1.64$ for $\alpha/2 = 0.05$.

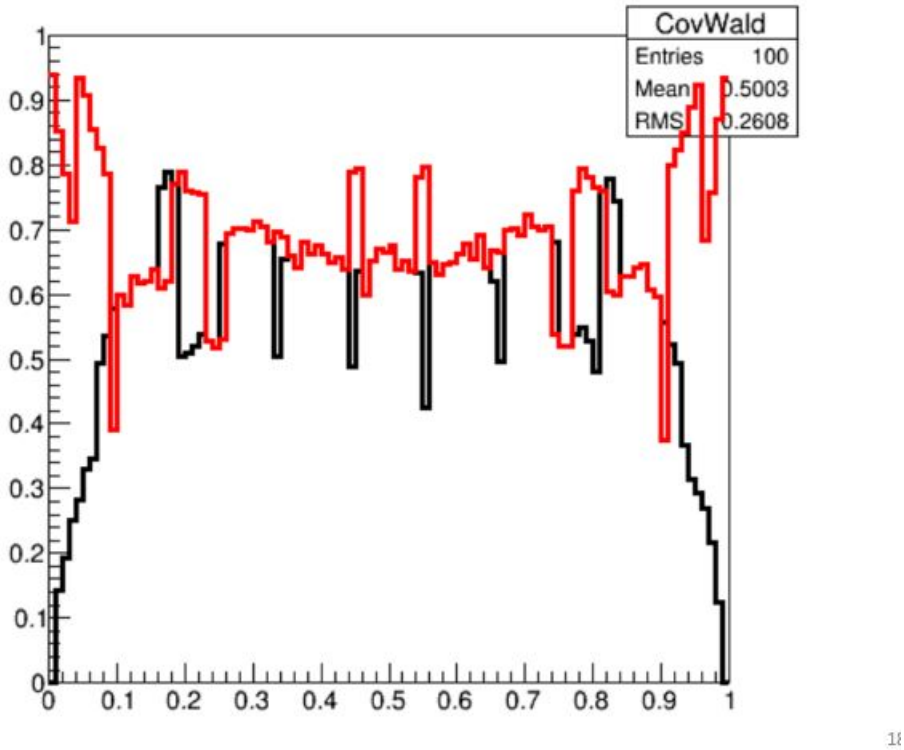

Figure 16: For a number of $N = 10$ trials, the red curve in the graph corresponds to the coverage of the Wilson score interval, while the black one shows the coverage of the Wald interval, as a function of the true value of the binomial ratio.

6.5 Undercoverage

There is a good reason why the Neyman construction is meant to ensure coverage of the resulting confidence sets: undercoverage should be avoided, and a frequentist statistician should not allow it. E.g., if you state a limit or an interval on a parameter of interest at 95% CL, and it turns out that the true coverage produced by the recipe you employed is smaller –say, 85%– even only for a subset of the possible true values of the parameter you are estimating, this corresponds to having underestimated the uncertainty bars of your measurement by a significant factor. You are thus falsely reporting the information provided by your measurement procedure.

Besides the obvious fault of assuming a wrong p.d.f. $f(x \mid \mu)$, undercoverage may result from approximate expressions for the variance, or from other specific aspects of the problem. This often has to do with our negligence in giving proper attention to specific conditions that the estimator must satisfy in order to have the good properties we routinely assume. Undercoverage may also arise from apparently innocuous procedures in the derivation of our results, such as

- deciding *a posteriori* whether to quote an upper or lower limit or instead a confidence interval around some estimated value (flip flopping);

- modifying details of your analysis because something does not look right, thus unnaturally altering the statistical properties of the data;

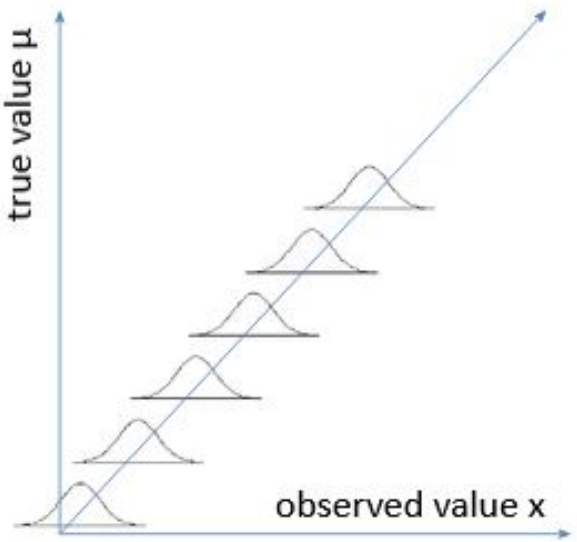

Figure 17: True mean μ of a parameter bound to be positive versus observed value of its estimate. Note that x may assume negative values.

- withholding results that look controversial.

6.6 Confidence intervals and flip-flopping

Let us now try to understand a couple of issues that the Neyman construction may run into, in the case of the measurement of a bounded parameter and the derivation of upper limits on its value. This is a frequent situation in fundamental physics. Typical observables falling in this category are cross sections for subnuclear processes, or masses of particles (*e.g.*, neutrinos).

We take the simplifying assumption that we are performing an unbiased measurement with uncertainties well described by a Gaussian distribution; we also renormalize measured values such that the variance is 1.0. In that case if μ is the true value, our experiment will return a value x which is distributed as

$$P(x \mid \mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-(x-\mu)^2/2\right) \tag{69}$$

Let us consider a Gaussian measurement with known sigma (we shall use $\sigma = 1$) of a bounded parameter $\mu >= 0$. The classical method to obtain intervals, given *e.g.* $\alpha = 0.05$, produces an upper limit in the form $\mu < x + 1.64$ (or $\mu < x + 1.28$ for $\alpha = 0.1$), as it is possible to verify by integrating a unit Gaussian distribution from 1.64 to infinity. for $x < -1.64$ the corresponding confidence interval does not contain any value of μ. Such is a violation of one of Neyman's own demands –confidence sets should not not contain empty sets.

The above problem has fostered many attempts to improve of the procedure when the parameter of interest is not defined in the full real axis. The simplest fix consists in enforcing that the upper limit be larger than 1.64 for any x.

What is called flip-flopping in this context is an unfortunately widespread practice in fundamental physics: scientists search for the signal of some physics process of interest in some data, and

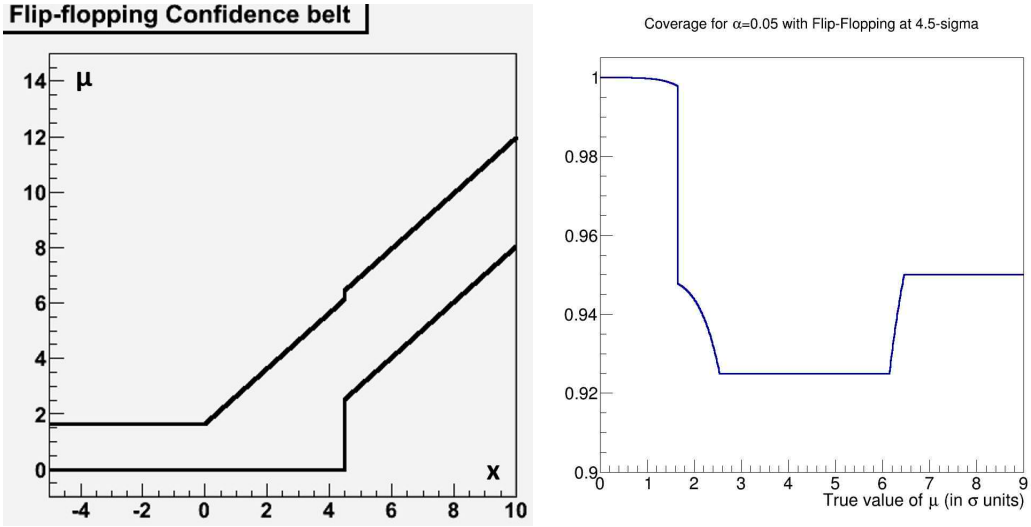

Figure 18: Left: the upper limit for a positive-defined parameter μ, fixed by enforcing $\mu^{up} = 1.64$ if the measured x is negative, may be substituted with a confidence belt if x is measured to be significantly larger than zero. The resulting coverage (right) is below the stated 95% in an intermediate region of μ values.

decide based on the results they get whether to publish an upper limit on the parameter of interest (say, a cross section for the process) or whether instead to produce a point estimate with an interval around it. As a result of this "a posteriori" procedure, we can prove that the resulting intervals will undercover. The reason why this happens is that the confidence belt is constructed with a mixed recipe, which does not guarantee all horizontal intervals to contain the required integral of the p.d.f.. Figure 18 shows the actual coverage of, *e.g.*, a flip-flopping procedure for a parameter bounded to be positive, when the result reported is an upper limit for $x < 4.5$ and an interval if $x >= 4.5$.

6.6.1 Example 8: Coverage test for likelihood intervals

In example 4 (Sec.3) we discussed the measurement of the load on a face of a cubic die. There, we solved the problem of estimating the bias in the die throw outcomes with analytical likelihood maximization. The same problem can be used to check the coverage of maximum likelihood intervals constructed with the Cramer-Rao-Frechet bound. One needs to write a simple program that generates at random a true value of the bias t (the "increase of probability" of throws yielding a six), and produces a set of N observed die throws under that hypothesis, then extracting the estimated value $\hat{t} \pm \sigma_{\hat{t}}$. One may then count how often the true value t of the load is within the 1-sigma intervals, and plot that fraction as a function of t. The result is shown in Fig. 19. By experimenting with different values of N one observes that coverage of these likelihood intervals is only approximate for small N, especially when the true value of the bias parameter t lies close to one of the boundaries $(-1/6, 1/3)$.

A note must be made about the example we studied. Our model of the bias in the probability of die throw results is what it is – a model – and it needs not represent reality. With the toy calculation we just described we cannot probe situations where the actual die has outcome probabilities that do not conform to that model; our conclusions are only valid if, as in our toys, relative frequencies of die throw outcomes conform to it. If they do not, the estimate of the bias may be plagued with

much worse coverage issues, or even be simply irrelevant –if the bias, *e.g.*, is not governed by a single parameter as encoded in our model.

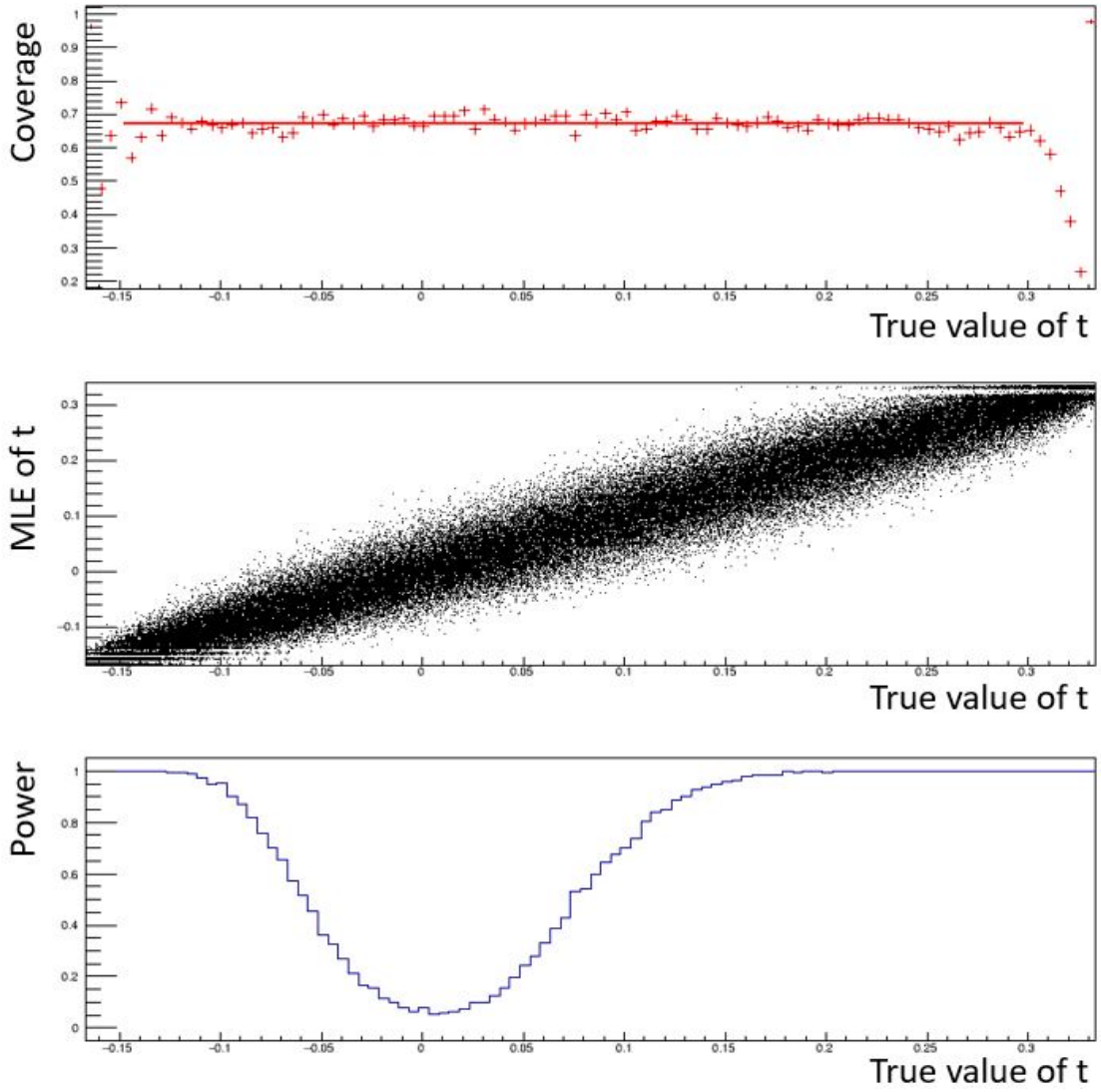

Figure 19: Top: fraction of 1-sigma likelihood intervals that include the true value of t (the load on the tested die), as a function of t, using 100 die throws per tested t value. The black line shows the average coverage in the region -0.15:0.3, which is the expected one (0.683). Significant undercoverage is observed at the boundaries. Center: dispersion of maximum likelihood estimates of t. Bottom: power of hypothesis tests for exclusion of the null hypothesis $t = 0$, at 95% C.L. (see Sec. 9).

7. Relevant subsets and ancillarity

7.1 Relevant subsets, conditioning, and ancillarity

In the previous section we have seen that Neyman's method, applied to the paradigmatic problem of a Gaussian measurement x of a parameter with known σ and unknown *positive* mean μ, yields upper limits at 95% confidence level in the form $\mu_{UL} = x + 1.64\sigma$.

We know that Neyman's procedure guarantees coverage, by construction. And yet, we can devise a betting strategy against the upper limits produced by Neyman's construction, *e.g.* at 19:1 odds (which is what can be called a "coherent bet", *i.e.* one which should pay off exactly as much as one invests, as one wins 19 times the waged amount 5% of the times, and loses the amount in the other 95% of the cases), using no more information than the observed value of x, and be guaranteed to win in the long run! The strategy consists quite simply in choosing a constant k: when an upper limit $\mu < x + 1.64$ is reported and $x < k$, one bets that the reported limit does not cover the true value of μ; otherwise, one takes no bet against it. For $k < -1.64$ this wins every bet, while for larger k the advantage is smaller but the expected win is always larger than zero. This, of course, only happens because of the constraint on $\mu > 0$; and yet, the observation has a value: if there exists a betting strategy of this kind, one has to conclude that the Neyman procedure does not make the best inference given the data.

In the above problem, the flaw of allowing winning betting strategies can be amended by adding a horizontal line at, *e.g.*, 1.64, delimiting the acceptance region of Neyman's construction for negative x, such that any confidence interval, regardless how negative x is, will contain the included values of μ (see Fig. 18, left); but this is of course an *ad-hoc* solution. More in general, in such problems one may identify an *ancillary statistic*, which can be used to partition the observation space into subspaces called *relevant subsets*. An ancillary statistic is a function of the data which yields information about the precision of the estimate of the parameter of interest, while bearing no information at all on the parameter's value.

The simplest example we can make, from experimental practice in particle physics, is the one of a branching fraction measurement for a particle decay. When N_A and N_B denote the event counts in two decay channels of the particle under study (exclusive and complementary, in the sense that the particle may only decay in those final states, which are distinguishable experimentally), one may prove that there are two equivalent ways of writing the combined probability of observing the two counts. The first,

$$P(N_A, N_B) = Poisson(N_A|\rho\,\mu_{tot}) \times Poisson(N_B|(1-\rho)\,\mu_{tot}) \tag{70}$$

describes the fact that the number of decays in each final state is a Poisson variable. The second, however, is a more attractive option:

$$P(N_A, N_B) = Poisson(N_A + N_B|\mu_{tot}) \times Binomial(N_A|\rho, N_A + N_B) \tag{71}$$

By using the second expression to estimate the branching ratio ρ, one may wholly ignore the ancillary statistic $N_A + N_B$, which provides information on the expected precision of the estimate (as the larger the sum of N_A and N_B is, the smaller will in general be the uncertainty on the binomial ratio), but which carries no information on ρ: all the information on the branching fractions in

the second expression comes from the conditional binomial factor! What this amounts to is a restriction of the sample space to a relevant subset, the one relevant for the actual observed data (the total number of decays we observed, $N_A + N_B$). By focusing on this relevant subset, which is only possible by having identified the ancillary statistic $N_A + N_B$, the problem is simplified, and the inference becomes more precise, because we may report the expected uncertainty for the subset of possible measurement outcomes that matter, rather than for the full sample space spanned by any value of N_A, N_B.

7.1.1 Cox weighting procedure

Things get even more intriguing if we consider the famous example by the late British statistician Brian Cox [6]: we consider the weighting of an object with a scale chosen between two that have different measurement precisions: the first reports weights with a 10% uncertainty, the second with a 1% uncertainty. The privilege of using the more precise scale is determined by the flipping of a coin: if you flip the coin and get tails you have to use the first scale, if you get heads you use the second. Suppose now that you do the experiment, and get to use the more precise scale, as you obtained heads. Which error do you quote for your measurement?

Of course, knowledge of your device allows you to assess a 1% precision on the measurement you report. Yet a full Neyman construction (which is unconditional on the outcomes) would require you to include the coin flipping in the procedure in order to guarantee coverage, and thus to report a much wider confidence interval. In this case, the coin flip outcome is the ancillary statistic. Its value partitions the observable space in two subsets, only one of which is relevant to the actual measurement you got to make.

7.1.2 Example 9: Locating the box

Another example we can make of ancillarity and relevant subsets is the classic problem of locating the center of a unit-box distribution using two values x_1, x_2 sampled from it, *i.e.* finding the value of μ using x_1, x_2 sampled from

$$p(x|\mu) = Uniform(\mu - 0.5, \mu + 0.5). \tag{72}$$

Suppose, *e.g.*, that $\mu = 1$, and consider in turn the following two datasets: A = $(x_1 = 0.99, x_2 = 1.01)$, and B= $(x_1 = 0.6, x_2 = 1.4)$. Neyman procedures maximizing power in the unconditional space yield the same confidence interval for both datasets A and B; however, B restricts the set of possible μ values to the range $[0.9, 1.1]$, while A only allows one to claim that μ is in the range $[0.51, 1.49]$ –a difference in interval width of a factor of five. Such a situation should ring a bell in the analyst's head: search for an ancillary statistic! There exists in fact an ancillary statistic which partitions the sample space into subsets that are characterized by different levels of expected precision on the box location. Inference can be made more powerful if we identify the proper subset for our measurement. This statistic is $\Delta = |x_1 - x_2|$, as it is easy to understand. The value of Δ affects the precision of the location estimate, but has no information on the location in itself.

The take-away bit from the above discussion is the following: search for ancillary statistics in your problem, whenever you are estimating an interval with frequentist means. The quality of your inference depends on the breadth of the whole space you are considering. The more you can restrict

it, the better (*i.e.*, the more relevant to the data you actually got, and thus to the uncertainty you face) your inference becomes. Ancillary statistics are not always easy to find, but they are quite useful, and their search is worth the time spent. If you can find one you will outperform the measurement that others using the same data and means, while still retaining frequentist coverage!

8. Frequentist and Bayesian practice

In order to discuss the two main schools of thought in professional statistics (frequentist and Bayesian), and reach the heart of the matter in a few specific problems of HEP, we need to first recall the definition of probability.

8.1 Probability

A mathematical definition is due to Kolmogorov (1933) with three axioms. Given a set of all possible elementary events X_i, mutually exclusive, we define the probability of occurrence $P(X_i)$ as obeying the following:

$$P(X_i) \geq 0 \ \ \forall i \tag{73}$$

$$P(X_i \vee X_j) = P(X_i) + P(X_j) \tag{74}$$

$$\sum_i P(X_i) = 1 \tag{75}$$

From the above we may construct more properties of the probability: If A and B are non-exclusive sets $\{X_i\}$ of elementary events, we may define

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B) \tag{76}$$

where $(A \vee B)$ describes the fact that an event X_i belongs to either set; and $(A \wedge B)$ that it belongs to both sets. A conditional probability $P(A|B)$ can then be defined as the probability that an elementary event belonging to set B is also a member of set A:

$$P(A|B) = P(A \wedge B)/P(B). \tag{77}$$

A and B are independent sets if $P(A|B) = P(A)$. In that case, from the above follows that $P(A \wedge B) = P(A)P(B)$.

8.2 Bayes theorem

The theorem linking $P(A|B)$ to $P(B|A)$ follows directly from the definition of conditional probability and the expression of $P(A \wedge B)$:

$$\left.\begin{aligned} P(A|B) &= P(A \wedge B)/P(B) \\ P(A \wedge B) &= P(B|A)P(A) \end{aligned}\right\} \quad \Rightarrow \quad P(A|B)P(B) = P(B|A)P(A)$$

If one expresses the sample space as the sum of mutually exclusive, exhaustive sets A_i (recalling the law of total probability: $P(B) = \sum_i P(B|A_i)P(A_i)$, one can rewrite the above as follows:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_i P(B|A_i)P(A_i)} \tag{78}$$

It is important to note that for probabilities to be well-defined, the "whole space" of observations needs to be defined, too. The whole space should be considered conditional on the assumptions going into the model. As we noted *supra*, restricting it to a relevant subspace (by conditioning to a specific value of an ancillary statistic) improves the quality of statistical inference.

8.3 Operational definitions of probability

We may now discuss the frequentist and Bayesian definitions of probability. A frequentist definition can be given in terms of the empirical limit of the frequency ratio between the number of successes S and trials T in an experiment repeated n times:

$$P(X) = lim_{n\to\infty}(S/T) \tag{79}$$

A definition of P as a limit is operationally sound –one can always imagine to continue sampling to obtain the required accuracy; there is nothing wrong in this, and one might compare it to the definition of electric field strength as the ratio between force on test charge and magnitude of the charge: in both cases one cannot really get to the asymptotic regime when the definition becomes exact, but in practical terms this is still viable. The real problem with the frequentist definition is that it can only be applied to repeatable experiments. This is not usually a restriction for relevant situations in fundamental science, such as those arising in particle collider experiments; but it must be kept in mind.

Within a Bayesian framework, in order to solve the problem of unrepeatable experiments we may decide to replace frequency with *degree of belief*. The best operational definition of degree of belief is the one due to de Finetti, which involves the idea of coherent bet. To define it, determine the maximum odds at which you are willing to bet that X occurs:

$$P(X) = max[expense/return] \tag{80}$$

Of course this depends on the observer as much as on the system: it is a subjective Bayesian probability. A huge body of literature exists on the subject. In scientific practice we would like our results to be coherent, in the sense that if we determine a probability of a parameter having a certain range, we would like it to be impossible for somebody knowing the procedure which yielded our result to put together a betting strategy against it, in such a way that they can on average win money! This is the heart of the matter for the use of Bayesian techniques in HEP.

We can make a few quick examples of coherent bets. *E.g.*, we might bet one dollar that we get a 6 upon throwing a die: the coherent return is $P(6) - 1 = 1/max(expense/return) = 6$ dollars. If P is known, as above, the coherent bet is a fair bet. But in general P may not be known: in that case, we may use our degree of belief. To make an example of that, you might ask yourself what would be the maximum sum you are willing to bet on your favourite soccer player scoring a goal in tomorrow's game, for a fixed return of $R = 100$ dollars in case he does. Evidently, you would not

invest $E = 100$ dollars if you believe that it is more likely that he or she will not score. If deep in your bones you feel, *e.g.*, that the chance of scoring is P=10%, you would be silly to wager more than $E_{max}(P = 0.1) = P \times R = 10$ dollars. This defines P in terms of the coherent bet. Of course, the above P is still a totally subjective prior. But it is, at least operationally, well defined through your actions.

8.3.1 Example 10: Frequentist use of Bayes theorem

Bayes theorem can be used for any P satisfying the Kolmogorov axioms, hence we can apply it to cases when no degree of belief is involved. One example, taken from HEP practice, is the following.

A b-tagging method is developed to identify hadronic jets originated from a b-quark in a collider physics experiment. One may measure:

- P(b-tag | b-jet) = 0.5 : the efficiency to identify b-quark-originated jets
- P(b-tag | !b-jet) = 0.02 : the acceptance for light-quark jets

From the above we also trivially get the following conditional probabilities:

- P(!b-tag | b-jet) = 1 − P(b-tag | b-jet) = 0.5 ,
- P(!b-tag | !b-jet) = 1 − P(b-tag | !b-jet) = 0.98.

The question is then: given a selection of b-tagged jets, what fraction of them are b-jets? *Id est*, what is P(b-jet | b-tag)? This is a trick question, as indeed, the result cannot be determined from the given information. We need, in addition, to know what is the true fraction of jets that do contain b-quarks in the total sample, P(b-jet). Take that to be P(b-jet) = 0.05; then Bayes theorem inverts the conditionality:

$$P(\text{b-jet} \mid \text{b-tag}) \propto P(\text{b-tag} \mid \text{b-jet}) P(\text{b-jet}).$$

If you then calculate the normalization factor, you get

$$P(\text{b-tag}) = P(\text{b-tag} \mid \text{b-jet}) P(\text{b-jet}) + P(\text{b-tag} \mid \text{!b-jet}) P(\text{!b-jet}) =$$

$$= 0.5 \times 0.05 + 0.02 \times 0.95 = 0.044,$$

and finally

$$P(\text{b-jet} \mid \text{b-tag}) = (0.5 \times 0.05)/0.044 = 0.568.$$

8.4 The Bayesian viewpoint

When we are dealing with hypotheses rather than events, Bayesian and frequentist schools part. Subjective probability deals with the probability of hypotheses: one may then talk of the probability of a constant of nature having a particular value. For a frequentist, speaking of the probability of a constant of nature taking on a particular value makes no sense –the constant has a fixed value, however unknown, and the probability is either zero or 1; although still a probability in Kolmogorov's sense, a frequentist cannot do anything with it. For a Bayesian that P does make sense, though; so it can be used as a factor in Bayes theorem.

If $f(x|q)$ is the p.d.f. of a random variable x, and q is a variable representing the possible values of an unknown physical parameter, then from N observations $X_{i\,i=1,..,N}$ one gets the joint density function as

$$p(X|q) = P_{i=1,...,N} f(X_i|q) \tag{81}$$

From a frequentist standpoint, q has a true, unknown, fixed value; one cannot use Bayes theorem to get $p(q|X)$ from $p(X|q)$ –it does not make sense to speak of $p(q)$. The inference Bayesians do using a particular set of data $\{X_0\}$ starts from the opposite viewpoint: $p(q)$ is a degree of belief of the parameter assuming a specific value. They can thus obtain

$$p(\theta|X_0) = \frac{p(X_0|\theta)p(\theta)}{\int p(X_0|\theta)p(\theta)d\theta}. \tag{82}$$

In the equation above, not all of the p factors are p.d.f.: $p(\theta|X_0)$ is the posterior probability density for θ; $p(X_0|\theta)$ instead is the likelihood function $L(\theta)$: this is not even a density. The data X_0 are fixed in its definition, and the factor is not defined until the data have been collected. Finally, $p(\theta)$ is the prior probability density for θ. This is a density function we cannot do without, and which determines the posterior when multiplied by the likelihood (which is sometimes called the *evidence*). The integral at the denominator in the expression then serves as a normalization factor. Summing it up, we observe that there is one, and only one, probability density in θ on each side of the equation, again consistently with the likelihood not being a density.

8.4.1 Example 11: Bayesian inference in a counting experiment

We can make an example of this construction, too. Let us consider a background-free experiment, where a theorist uses a model to predict a signal with a Poisson mean of 3 events. From the formula of the Poisson distribution, we may write:

- P(0 events | model true) = $3^0 e^{-3}/0! = 0.05$,
- P(0 events | model false) = 1.0,
- P(>0 events | model true) =0.95,
- P(>0 events | model false)= 0.0.

Now, let us imagine that the experiment is performed, and zero events are observed. What is then the probability that the model is true, P(model true | 0 evts)? Unsurprisingly, we again cannot determine it from the data alone; we need a prior degree of belief in the model, P(model true). With that in hand, we may then use Bayes theorem to invert the conditionality:

$$P(\text{model true} \mid 0) \propto P(0 \mid \text{model true}) P(\text{model true}).$$

The above highlights the necessary feature that the strength of evidence (the result of an experiment) cannot by itself overturn our beliefs on some hypothesis, if the latter has solid grounds. On the other hand, if we are testing some exotic new physics model, we may be more willing to part with it, as our prior belief in it was not strong. However, we should note that even when we reach a statement

about the probability of a given hypothesis, we need to include in the construction a cost function if we want to use it to take a decision. The cost function describes the relative costs *you* incur in depending on what action you take, given the posterior probability you have obtained. Hence decision making requires you to specify your prior probabilities along with the relative costs of the various decisions you may wish to take given the outcome (*e.g.*, publishing a paper, or designing a new experiment).

8.5 Reference priors

There are compelling arguments to prove that Bayesian reasoning with a subjective prior is the uniquely coherent way of updating personal beliefs upon obtaining new data. The question is whether the Bayesian formalism can be used by scientists to report the results of their experiments in an objective way (however one defines objectivity), and whether the result can still be coherent if we replace subjective probability with some other recipe.

In the mid-20th century Harold Jeffreys set out to identify a set of priors P(μ) that contain as little information as possible on the parameter of interest μ, so that the posterior p.d.f. will be dominated by the likelihood. Note that this is not what is unfortunately common practice in HEP, where we sometimes choose P(μ) uniform in whatever metric we happen to be using –that is a bad idea. Instead, Jeffreys' work resulted in what are today called "reference priors".

The probability integral transform assures us that we can find a metric under which the p.d.f. is uniform. Choosing the prior is then equivalent to choosing the metric under which the p.d.f. is uniform. What Jeffreys did was to choose the metric according to the Fisher information. This results in different priors depending on the problem at hand: *e.g.*,

- in a Poisson counting experiment with no background, the reference prior is found to be P(μ) = μ-0.5;
- in the case[4] of a Poisson with background b, P(μ) = $(\mu + b)^{-0.5}$;
- For a Gaussian measurement with unknown mean, P(μ) = Uniform.

We should note that what we (in HEP) call "flat priors" are not what statisticians mean with that term: flat priors for them are Jeffreys priors (flat in information metric). In general, a sensitivity analysis (effect of prior assumption on the result) should always be ran, especially in HEP. Also, in more than one dimension the problem of finding suitable priors becomes progressively harder. It is a notoriously difficult problem: you start with a prior, change variables and thus get a Jacobian, which creates structure out of nothing. *e.g.*, a uniform prior in high dimensions pushes all the probability away from the origin. Also, it is not even clear how to define subjective priors there; human intuition fails in high dimensions, and lots of arbitrariness remains in the procedure. Hence for fundamental physics these procedures should be used with high caution, and the general advice is to always compare the result of different choices for the prior.

[4] In this case, the prior belief on μ depends on b.

8.6 Likelihood ratios

Likelihood ratios, which are also constructed from the frequentist definition of probability, are the basis of a large set of techniques addressing point and interval estimation and hypothesis testing. They also do not need a prior to be constructed; we will see an example *infra*. Because of the invariance properties of the likelihood under re-parametrization, a ratio of likelihoods can be used to find the most likely values of a parameter q, given data X. A re-parametrization from q to $f(q)$ will not modify our inference: if $[q_1, q_2]$ is the interval containing the most likely values of q, $[f(q_1), f(q_2)]$ will correspondingly contain the most likely values of $f(q)$. One may find the interval by selecting all the values of q such that

$$-2[lnL(q)\check{}lnL(q_{max})] \leq Z^2 \tag{83}$$

The interval approaches asymptotically a central confidence interval with C.L. corresponding to $\pm Z$ Gaussian standard deviations. *e.g.*, if we want 68.3% C.L. intervals, we choose Z=1; for five sigma, $Z^2 = 25$, *etcetera*. But we must realize that this is an approximation! Still, it is a very good one in typical cases. Their properties depend on Wilks' theorem [7], which rests on a few regularity conditions of the problem, but even when those are violated the approximation may hold reasonably well. Likelihood ratio tests are popular in HEP because they are the default output of the "MIGRAD" function of MINUIT.

Problems with the likelihood ratio coverage arise when the parameter q approaches the boundaries of its definition; we already saw this in practice in Example 8 (Sec. 6.6.1). We can make a further example here, deriving the likelihood-ratio interval for a Poisson process with $n = 3$ events observed.

The likelihood $L(\mu) = \mu^3 e^{-\mu}/3!$ has a maximum at μ= 3. If one then computes $\Delta(2lnL) = 1$, this yields the approximate ± 1 Gaussian standard deviation interval : $[1.58, 5.08]$. For comparison, a Bayesian central interval with a flat prior yields $[2.09, 5.92]$; a Neyman central interval yields $[1.37, 5.92]$.

8.7 The likelihood principle

In both Bayesian methods and likelihood-ratio based methods, only the probability (density) for obtaining the data at hand is used: this information is contained in the likelihood function. Probabilities for obtaining other data are not used. In contrast, in typical frequentist calculations (*e.g.*, a p-value calculated as the probability of obtaining a value as extreme or more extreme than that observed), one uses probabilities of data that have not been seen.

This difference is captured by the likelihood principle: *If two experiments yield likelihood functions which are proportional, then your inferences from the two experiments should be identical*. The likelihood principle is typically built into Bayesian inference. It is instead violated (and sometimes badly) by p-values and confidence intervals. In fact, you cannot have both the likelihood principle fulfilled and coverage guaranteed.

Although practical experience indicates that the likelihood principle may be too restrictive, it is useful to keep it in mind.

8.7.1 Example 12: the likelihood principle

Imagine you expect background events sampled from a Poisson mean b, assumed known precisely. For signal mean μ, the total number of events n is then sampled from Poisson mean $\mu + b$. Thus, we may write

$$P(n) = (\mu + b)^n e^{-(\mu+b)}/n! \tag{84}$$

Now, suppose that upon performing the experiment, you observe no events at all, $n = 0$. You then write the likelihood as

$$L(\mu) = (\mu + b)^0 e^{-\mu+b}/0! = e^{-\mu} e^{-b} \tag{85}$$

Now we see that changing b from 0 to any $b^* > 0$, $L(\mu)$ only changes by the constant factor e^{-b^*}. This gets renormalized away in any Bayesian calculation, and is *a fortiori* irrelevant for likelihood ratios. So for zero events observed, likelihood-based inference about signal mean μ is independent of expected b. You immediately see the difference with frequentist inference: in the frequentist confidence interval construction, the fact that $n = 0$ is less likely for $b > 0$ than for $b = 0$ results in narrower confidence intervals for μ as b increases. That means that if you have a more background-ridden experiment, you may extract more powerful inference in this case!

8.8 In summary

We may compare three methods to compute intervals below, to stress the point that the answer one gets from statistical practice depends a lot on the way one computes it.

- Bayesian credible intervals:
  1. need a prior (can be a good thing –allows a means to put in your personal prior belief);
  2. random variable in construction is true value;
  3. usually obey the likelihood principle;
  4. can be the basis for decision theory (they provide $p(q \mid data)$);
  5. do not guarantee coverage.
- Frequentist confidence intervals:
  1. do not need a prior (can do inference reporting the result of your data keeping it objective);
  2. random variables are extrema of intervals;
  3. do not obey the likelihood principle;
  4. guarantee coverage;
  5. use p(data not obtained) for inference.
- Likelihood ratio intervals:
  1. they do not need a prior;

2. have as random variables the extrema of intervals;
3. obey the likelihood principle;
4. do not always cover.

9. A few notes on hypothesis testing

Hypothesis testing deals with a precise procedure through which data is used to accept or discard a given hypothesis on underlying parameters or models. In its essence, the procedure consists in a comparison of some test statistic, computed from observed data, with the p.d.f. that the statistic is expected to show under two different hypotheses for the parameter of interest (or the two competing theories being compared). The two hypotheses are usually labeled as H_0 (the null hypothesis, which is the "reference model" in typical cases) and H_1 (the alternate hypothesis, which in HEP is the one entailing a discovery of some new process).

The precondition to perform hypothesis testing is therefore that the hypotheses be perfectly well specified, such that a precise p.d.f. for the test statistic can be determined in advance. Also crucial is the definition of a type-I error rate (also called "size" of the test): this is the probability that the data will let us choose the alternate hypothesis H_1 when the null hypothesis H_0 is in fact true. The type-I error rate is specified by the letter α; typical values used in non-exact sciences are of 0.1 or 0.05, while in particle physics we have grown accustomed to choose a much smaller $\alpha = 2.9 \times 10^{-7}$, which corresponds to the "five-sigma criterion" (see *infra*).

Strictly connected to α is the concept of "power" $(1 - \beta)$, where β is the so-called type-2 error rate, and is defined as the probability of accepting the null hypothesis H_0 when the alternate H_1 is instead true. Power is thus the fraction of times that an alternate hypothesis is verified by the procedure. Once the test statistic has been defined, by choosing α one is automatically also choosing β. A stricter requirement on the former (a smaller fraction of times that experimenters allow themselves to incorrectly accept the alternate) causes power to decrease.

What makes the difference in power of a hypothesis test is of course the choice of a test statistic, along with the amount of data one collects. But the power of the test also depends on the value that the parameter of interest takes, if we are doing what is called a single-versus-composite hypothesis test, where the null hypothesis corresponds to a specific value of a parameter q, and the alternative hypothesis implies that the parameter has a different value. Such is a very common case in HEP and related fields: for instance, if we search for a new physics phenomenon, we identify its zero cross section to the null hypothesis, and any other value to the alternate one. Of course, in similar cases power will generally increase with the departure of q from the reference value at H_0. We saw an example of this *supra*, when we discussed the load on the die (example 8, Sec. 6.6.1).

We should also note that the choice of test statistic is very important. In simple-versus-simple hypothesis testing, the Neyman-Pearson lemma proves that there exist a uniformly most powerful test statistic: that is the likelihood ratio of the two hypotheses. However, in different setups (the common simple-versus-composite one, *e.g.*) power depends on the nuisance describing the alternative, and the situation is not any longer amenable to optimization.

9.1 Statistical significance

Statistical significance is a quantity meant to report the probability that an experiment obtains data at least as discrepant as those actually observed, under a given null hypothesis H_0. As noted above, in physics H_0 usually describes the currently accepted and established theory; there are however exceptions to the rule. One notable such instance was the search for the Higgs boson at the Large Hadron Collider, when ATLAS and CMS defined their null hypothesis to correspond to the validity of a theoretical model corresponding to the Standard Model (SM) of particle physics subtracted of the existence of a Higgs particle, and the alternate hypothesis to be the full SM endowed as expected with a Higgs boson of mass M_H. That setup, as it happens in most cases in HEP, is a simple-versus-composite test, where the compositeness is due to the unknown mass of the Higgs boson.

Given data X and a test statistic T constructed with them, one may obtain a p-value as the probability of obtaining a value of T at least as extreme as the one observed, if H_0 is true. This implies having chosen a "region of interest" as *e.g.* the one of large positive or negative values of T, departing from the values assumed for H_0. The extracted p can then be converted into the corresponding number of "sigma," *i.e.* standard deviation units from a Gaussian mean. This is done by finding x such that the integral from x to infinity of a unit Gaussian equals p:

$$\frac{1}{\sqrt{(2\pi)}} \int_x^\infty e^{-\frac{t^2}{2}} dt = p. \tag{86}$$

According to the above recipe, a 15.9% probability is a one-standard-deviation effect; a 0.135% probability is a three-standard-deviation effect; and a 0.0000287% probability corresponds to five standard deviations –"five sigma" in jargon. The convention is to use a "one-tailed" Gaussian distribution: we do not care about departures of x from the mean in the un-interesting direction, so we only integrate p-values on one side.

We note that from the construction above the conversion of p into σ units is independent of experimental detail. Using the number of sigma rather than the p-value is just a shortcut, nothing more. In particular, using "sigma" units does in no way mean we are operating some kind of Gaussian approximation anywhere in the problem! Also, it is important to note that the whole construction rests on a proper definition of the p-value. Any shortcoming of the properties of p (*e.g.*, non-flatness of its p.d.f. under the null hypothesis) totally invalidates the meaning of the derived number of sigma.

9.2 Non-Gaussian tails

The problem of "fat tails" is a recurrent one in statistics, and it has a lot of relevance also in common HEP practice. In fact, a study of the residuals of particle properties (defined as the difference between the true value of those properties, assessed by very precise recent experiments, and the measured value, in units of the quoted number of sigma) in the Review of Particle Properties in 1975 revealed that they were not Gaussian. Matts Roos *et al.* [8] considered residuals in kaon and hyperon mean life and mass measurements, and concluded that these were well described by a Student distribution $S_{10}(h/1.11)$:

$$S_{10}\left(\frac{x}{1.11}\right) = \frac{315}{256\sqrt{10}}\left(1+\frac{x^2}{12.1}\right)^{-5.5}. \quad (87)$$

While we should not allow ourselves to extrapolate to 5-sigma the behaviour found by Roos and collaborators in the bulk of the distribution, the observation is evidence that the uncertainties evaluated in experimental HEP may have a significant non-Gaussian component. In fact, a similar result was found with more statistics in a more recent study [9]. What this implies is that we should be very careful in automatically identifying a very high-significance result with the discovery of a new effect or physical process. In most cases, what is happening is that some systematic uncertainties have been underestimated, neglected, or incorrectly assessed to have a Gaussian p.d.f. when this was not the case.

10. Conclusions

In these lectures we have tried to provide a glimpse to the complexity of statistical inference. Indeed, statistics is not trivial, and even the simplest applications may hide unforeseen complexity. Part of the reason of this is that there are many different possible procedures for statistical inference, and they produce different results. Making sense of those differences requires insight in the details of the procedures. The key in HEP is in fact to try and derive results with different methods: if they do not agree, we get wary of the results, plus we learn something.

Making the right choices for what method to use is an expert-only decision, so a good particle physicist should become an expert in statistics.

We leave our readers with a short list of misguided statements they should now be able to avoid making:

- Probability inversion statements: "The probability that the SM is correct given that I see such a departure is less than x%".
- Wrong inference on true parameter values: "The top mass has a probability of 68.3% of being in the 171-174 GeV range".
- Apologetic sentences in your papers: "Since we observe no significant departure from the background, we proceed to set upper limits".
- Improper uses of the likelihood: "the upper limit can be obtained as the 95% quantile of the likelihood function".

Acknowledgments

A significant fraction of the material on which these lectures are based comes from lectures by Bob Cousins and from discussions with him; any mistake is however ours.

References

[1] C. McCusker and I. Cairns, Phys. Rev. Lett 23 (1969) 658.

[2] P. Adamson *et al.*, arxiv:1201.2631 (2011).

[3] T. Adam *et al.*, Measurement of the neutrino velocity with the OPERA detector in the CNGS beam, https://arxiv.org/abs/1109.4897v1 (2011).

[4] G.D. Cowan, Statistical Data Analysis, Clarendon press (1998).

[5] T. Affolder *et al.*, Measurement of the differential dijet mass cross section in ppbar collisions at sqrt(s)=1.8 TeV, Phys. Rev. D 61, (1999) 091101.

[6] D. Cox, Principles of Statistical Inference, Cambridge Univ. press (2006).

[7] S.S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Stat. 9 (1) (1938) 60, doi:10.1214/aoms/1177732360 (1938).

[8] M. Roos, M. Hietanen, and M.Luoma, A new procedure for averaging particle properties, Phys.Fenn. 10 (1975) 21.

[9] D. Bailey, Not Normal: the uncertainties of scientific measurements, arxiv:1612.00778 (2016).