Elements of statistics

NPAC 2023-2024, J. Biteau

This document aims to introduce and illustrate the essential elements at Master's level for performing statistical inference. The process of inference is understood here as evaluating the relevance of a model with respect to a set of data. A concise bibliography can be found at the end of this document to start exploring the statistical approaches employed in high-energy (astro)physics and cosmology. The content of this document is largely inspired by the chapters "*Probability*" [1] and "*Statistics*" [2] of the *Particle Data Group* review, whose reading is strongly encouraged. The same applies to the reference book *Numerical Recipes* [3], in particular chapters 14 "Statistical *Description of Data*" and 15 "*Modeling of Data*". The whole book *Numerical Recipes* can be considered an essential prerequisite for a thesis in our fields. The book provides examples in C/C++, which serve as an excellent guide to understanding the analytic approaches. The present document favours the use of Python libraries, which can be tested in the suggested exercises.

1 Usual laws of probability: the ubiquity of the Gaussian

The Gaussian distribution is widely used (and sometimes overused) in physics. The classic argument justifying its use is the central limit theorem, a proof of which exploits characteristic functions. We consider the latter outside the scope of this document and restrict ourselves here to the example of distributions used in counting experiments.

1.1 Counting experiment

Consider n cells, with a binary value (0 or 1), described by independent and identically distributed random variables. The value taken by each cell follows Bernoulli's law: it takes the value 1 with probability p and the value 0 with probability $1-p$.

Let's look at the sum of the values contained in these cells, described by the random variable X. This variable X can take discrete values, $k \in [\![0, n]\!]$. For $X = k$, k cells must have a value of 1, with probability $\propto p^k$, and $n-k$ cells must have a value of 0, with probability $\propto (1-p)^{n-k}$. We need only take into account the number of possible arrangements to determine the normalisation factor, $\binom{n}{k}$. The sum of the cell values thus follows a binomial distribution.

Binomial law: $X \sim \mathcal{B}(n, p)$

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{1}$$

Mean: $\mathrm{E}(X) = np$
Variance: $\mathrm{Var}(X) = np(1-p)$.

Note 1. The binomial distribution can be generalised to cases where the value taken by each cell is not binary (e.g. dice rolls). This is known as a multinomial distribution.

1.2 When Bernoulli joins Poisson

The binomial distribution is of interest for a small number of cells (typically $n < 100$). Suppose we are interested in a light source, emitting n photons per second, with n large, and that the probability that we detect one of these photons is p, with p small. We work with a constant value of $\lambda \equiv np$, which is the mathematical expectation of the binomial distribution. So in practice, we assume that the mean value for the measurement remains constant in our thought experiment. We can then rewrite the distribution function of the variable X as follows:

$$\begin{aligned}\mathbb{P}(X=k) &= \frac{n!}{(n-k)!\ k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\dots(n-k+1)}{n^k}\left(1-\frac{\lambda}{n}\right)^{-k}\times\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^{n} \\ &\xrightarrow[n\gg 1,\ \lambda=\text{const.}]{} \frac{\lambda^k}{k!}\exp(-\lambda) \end{aligned} \tag{2}$$

Thus, when the maximum number of achievable counts n is large, the binomial distribution of parameters n and p tends towards a Poisson distribution of parameter $\lambda = np$.

Poisson law: $X \sim \mathcal{P}(\lambda)$

$$\mathbb{P}(X=k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{3}$$

Mean: $\mathrm{E}(X) = \lambda$
Variance: $\mathrm{Var}(X) = \lambda$

Note 2. The sum of independent Poisson variables with parameters λ_i follows a Poisson distribution with parameter $\sum_i \lambda_i$ (stability under summation).

1.3 When Poisson joins Gauss

Let's now assume that the average number of counts measured per second, λ, is large (typically $\lambda > 100$) and let's look at small variations around this average value, of relative amplitude $\delta \equiv k/\lambda - 1$. Using Stirling's formula, $k! \approx \sqrt{2\pi k}\ k^k \exp(-k)$, we can rewrite the Poisson distribution function as follows:

$$
\begin{aligned}
\mathbb{P}(X=k) &= \frac{1}{\sqrt{2\pi k}}\left(\frac{\lambda}{k}\right)^k \exp(k-\lambda) \\
&= \frac{1}{\sqrt{2\pi\lambda}}(1+\delta)^{-k-\frac{1}{2}}\exp(\lambda\delta) \\
&= \frac{1}{\sqrt{2\pi\lambda}}\exp\left(-\left(\lambda+\lambda\delta+\frac{1}{2}\right)\ln(1+\delta)+\lambda\delta\right) \\
&= \frac{1}{\sqrt{2\pi\lambda}}\exp\left(-\frac{\lambda\delta^2}{2}-\frac{\delta}{2}+\frac{\delta^2}{4}+\mathcal{O}(\lambda\delta^3)\right) \\
&\xrightarrow[\lambda\delta\gg 1,\ \delta\ll 1]{} \frac{1}{\sqrt{2\pi\lambda}}\exp\left(-\frac{(k-\lambda)^2}{2\lambda}\right)
\end{aligned}
\tag{4}
$$

Thus, when the average number of counts measured is large, the distribution around this average follows a normal distribution, also known as a Gaussian distribution, with average $\mu = \lambda$ and variance $\sigma^2 = \lambda$.[1]

Gauss's law: $X \sim \mathcal{N}(\mu, \sigma^2)$

$$
\mathbb{P}(X=x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{5}
$$

Mean: $\mathrm{E}(X) = \mu$
Variance: $\mathrm{Var}(X) = \sigma^2$

Figure 1 compares the binomial, Poissonian and Gaussian distributions expected in a counting experiment. These distributions of the number of measured counts can be seen as discrete probability distributions or as the histogram of the number of counts reconstructed by repeating the measurement a large number of times. For a moderately large number of cells ($n \approx 100$, top left panel), the binomial distribution describing the number of counts can be distinguished from the Poisson and Gauss distributions. If the maximum number of counts is very large ($n \sim 1000$), the number of counts is distributed according to a Poisson distribution, as illustrated by the bottom left panel in figure 1. If, in addition, the average number of counts is large ($\mu \sim 100$), the distribution around this average value follows a normal distribution with a variance equal to its average. Given the peaked nature of Poissonian and Gaussian distributions, a typical realisation of a counting experiment leading to the measurement of $k > 10$ is often assimilated to the estimation of the counting rate as $k \pm \sqrt{k}$, i.e. a Gaussian distribution centred on $\mu = k$ and of standard deviation $\sigma = \sqrt{k}$.

1.4 Gaussian probability density

The probability laws in equations (1,3,5) apply to discretely distributed variables, $k \in \mathbb{N}$. The function

$$
P : k \to \mathbb{P}(X=k) \tag{6}
$$

[1] The equality of the average and the variance might seem to conflict with usual dimensional analysis. Note that λ is an integer here. So the mean, variance and standard deviation are dimensionless quantities. In the more general case, one should make sure that the mean, μ, and the standard deviation, σ, are of the same dimension.

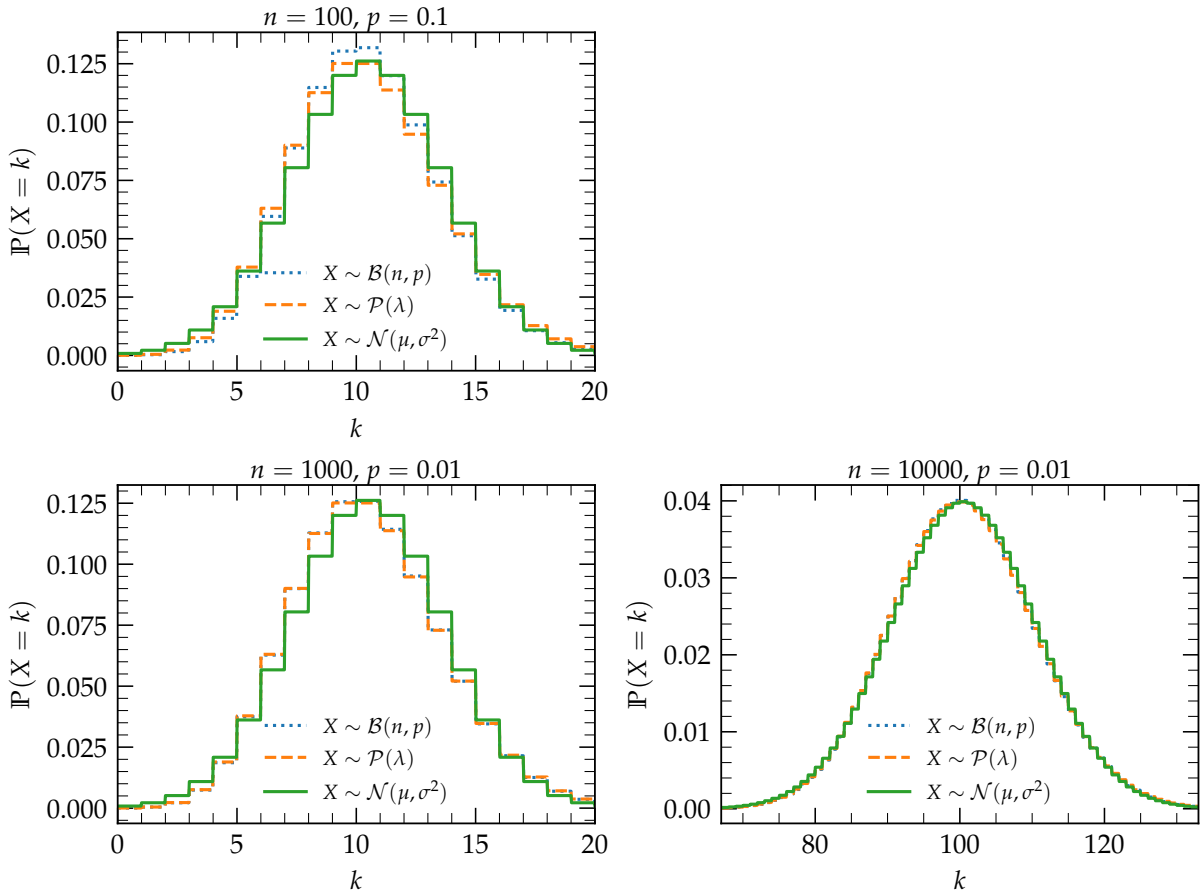

Figure 1: Binomial (dotted blue), Poisson (dashed orange) and Gaussian (solid green) distributions for different values of the parameters n and p (cf. Eq. (1)).

.

is then called *probability mass function (PMF)*. The PMF verifies $\sum_{k\in\mathbb{N}} P(k) = 1$.

The Gaussian distribution is also defined for $x \in \mathbb{R}$. The function

$$p : x \to \mathbb{P}(X = x) \tag{7}$$

is then called *probability density function (PDF)*. The PDF verifies $\int_{x\in\mathbb{R}} p(x)\mathrm{d}x = 1$. The properties of the Gaussian distribution of zero mean and unit standard deviation, known as the standard normal distribution, are illustrated in figure 2.

The integral of the PDF up to x,

$$\begin{aligned} F : x &\to \mathbb{P}(X \leq x) \\ x &\to \int_{-\infty}^{x} p(t)\mathrm{d}t \end{aligned} \tag{8}$$

is called *cumulative distribution function (CDF)*. The CDF varies between 0 and 1. Its complementary function, $1 - F(x)$, is called *survival function (SF)*:

$$\begin{aligned} S : x &\to \mathbb{P}(X > x) \\ x &\to \int_{x}^{\infty} p(t)\mathrm{d}t \end{aligned} \tag{9}$$

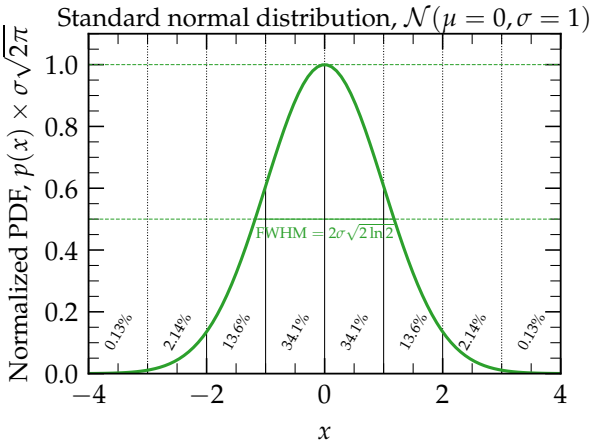

Figure 2: Gauss distribution with zero mean and unit standard deviation. The PDF is here normalised by a pre-factor $\sigma\sqrt{2\pi}$. The *full width at half maximum (FWHM)*, indicated by a horizontal line, is $2\sigma\sqrt{2\ln 2} \approx 2.35\sigma$. The area under the curve in the intervals $[-4\sigma; -3\sigma]$, $[-3\sigma; -2\sigma]$, etc., is indicated as a percentage in each of the zones

.

The CDF and SF of the Gaussian distribution can be estimated using the tabulated function erf, known as the Gaussian error function:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t. \tag{10}$$

and its complementary function, $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$. Note in particular that the SF of the Gaussian distribution is

$$\begin{aligned} \int_x^\infty \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x'-\mu)^2}{2\sigma^2}\right) \mathrm{d}x' &= \frac{1}{\sqrt{\pi}} \int_{\frac{x-\mu}{\sigma\sqrt{2}}}^\infty \mathrm{e}^{-t^2} \mathrm{d}t \\ &= \frac{1}{2}\,\mathrm{erfc}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right) \end{aligned} \tag{11}$$

Note 3. A numerical implementation of the usual PMF, PDF, CDF, SF is available in many languages, for example in the `stats` library of `scipy` in Python[2] and in the `MathCore` library from `ROOT` in C/C++.[3]

2 Statistical inference: from Bayes to χ^2 via likelihood

Let's suppose we are dealing with a set of measured number of counts, $y_i = k_i$, as a function of a relevant variable, x_i. This could, for example, be a spectrum in high-energy (astro)physics, i.e. a number of events as a function of a measured energy. Another example in the context of NPAC's laboratory work could be the calibration of the number of ADC (analog-to-digital converter) counts from a detector as a function of the energy associated with nuclear lines.

How can we identify the most appropriate model $f(x; \boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$ (e.g. for an affine relationship $f(x_i; \boldsymbol{\theta} = \{a, b\}) = a + bx_i$) for representing the evolution of $\{y_i\}$ as a function of $\{x_i\}$?

[2] `https://docs.scipy.org/doc/scipy/reference/stats.html`
[3] `https://root.cern/doc/master/group__Math.html`

2.1 The general case: Bayesian approach

The theory of probability built up since Laplace enables us to rigorously define the process of inferring the parameters $\boldsymbol{\theta}$ of a hypothesized model, H, from a set of data, D, and possibly from prior information about the model, I. The degree of plausibility of H knowing D and I is given by Bayes' theorem.

Bayes' theorem

$$\mathbb{P}(H|D,I) = \frac{\mathbb{P}(D|H,I)\mathbb{P}(H|I)}{\mathbb{P}(D|I)} \tag{12}$$

$\mathbb{P}(H|D,I)$: probability *a posteriori*, or **posterior**, describing the credible parameters of the hypothesis H given prior information I and new data D;

$\mathbb{P}(D|H,I)$: **likelihood**, describing the probability of data occurrence for a model;

$\mathbb{P}(H|I)$: probability *a priori*, or **prior**, describing the initial degree of plausibility of the model parameters, for example the range in which they can vary;

$\mathbb{P}(D|I)$: Bayesian **evidence**, this normalization factor is the integral of the numerator over the entire parameter space, so that the posterior is a probability.

Note 4. Bayes' theorem derives directly from the axioms of probability (or Kolmogorov's axioms). Indeed, note that for two non-disjoint sets of events A and B, $\mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$. Thus, the intersection is obtained either by taking from A and then taking from B knowing that we took from A (left-hand term), or by taking from B and then taking from A knowing that we took from B (right-hand term).

By design, the Bayesian approach is perfectly adapted to refining a measurement using new data, D_1. The information acquired from previous measurements, D_1, is encoded in the posterior $\mathbb{P}(H|D_1,I)$, which can be used as a prior to determine the posterior $\mathbb{P}(H|D_2,D_1,I)$. For a sufficiently large number n of measurements, the choice of initial prior, $\mathbb{P}(H|I)$, most often has a relatively small impact on the final posterior $\mathbb{P}(H|\{D_i\}_{1\leq i\leq n},I)$. For a small number of measurements, the choice of the initial prior legitimately raises questions. In the context of this introductory document, we favour the use of simple priors such as flat priors (constant function of the parameters).

The Bayesian approach is particularly well-suited to inferring the best parameters and their credible intervals. We are most often interested in cases where the tested hypothesis is entirely determined by a set of continuous parameters, $\boldsymbol{\theta}$. The prior is then characterized by a PDF $\pi(\boldsymbol{\theta})$, the likelihood of the data $\mathbf{d}$ by the PDF $p(\mathbf{d}|\boldsymbol{\theta})$, so we can write the posterior, or the PDF of the parameters *a posteriori*, as

$$p(\boldsymbol{\theta}|\mathbf{d}) = \frac{p(\mathbf{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(\mathbf{d}|\boldsymbol{\theta'})\pi(\boldsymbol{\theta'})\mathrm{d}\boldsymbol{\theta'}} \tag{13}$$

The likelihood function of the data is sometimes written as $\mathcal{L}_{\mathbf{d}}(\boldsymbol{\theta}) \equiv p(\mathbf{d}|\boldsymbol{\theta})$. Leaving aside the constant normalization term, we can then write the posterior as

$$p(\boldsymbol{\theta}|\mathbf{d}) \propto \mathcal{L}_{\mathbf{d}}(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{14}$$

Let's suppose we are dealing with a set of parameters $\boldsymbol{\theta}$ and we are only interested in the parameter θ_0 (e.g. invariant mass of two photons from a Higgs boson). The other parameters $\{\theta_i\}_{i\geq 1}$ are called nuisance parameters. Their values are less interesting to us, but they can still have an impact on the estimate of θ_0 (the nuisance parameter could describe a systematic error induced e.g. by the energy reconstruction bias of an electromagnetic calorimeter). In other words, θ_0 can be correlated with the $\{\theta_i\}_{i\geq 1}$. In the Bayesian approach, we determine the marginal posterior of θ_0 by marginalizing over the nuisance parameters, i.e. by integrating over all the values that $\{\theta_i\}_{i\geq 1}$ can take. The marginalized posterior of the parameter of interest θ_0 is then

$$p(\theta_0|\mathbf{d}) \propto \int \mathcal{L}_{\mathbf{d}}(\theta_0, \{\theta_i\}_{i\geq 1})\pi(\theta_0, \{\theta_i\}_{i\geq 1})\mathrm{d}\theta_1..\mathrm{d}\theta_{n-1} \tag{15}$$

For a given prior $\pi(\boldsymbol{\theta})$, e.g. a constant function of the parameters, and a given likelihood $\mathcal{L}_{\mathbf{d}}(\boldsymbol{\theta})$, which we will make explicit in section 2.2, the estimation of the best parameters and associated credible intervals derives directly from the posterior in equation (15). Several estimators derived from the posterior can be chosen for the best parameters (e.g. mode also known as maximum a posteriori, mean, median) and for the credible interval (symmetric or asymmetric interval with respect to the best parameter, smallest interval containing a fraction of the integral of the posterior, bounds defined by an equal value of the posterior). In the case of a Gaussian posterior, the various estimators proposed in the literature yield consistent estimates. For the purposes of this introductory document, we suggest using the median as the best parameter, and the percentiles at 16% and 84% as the bounds of the credible interval at 68%, that is Gaussian bounds within $\pm 1\sigma$ as illustrated in figure 2.

The complete posterior in equation (14) and the marginalized posteriors in equation (15) can be determined analytically in some simple cases, e.g. for Gaussian likelihoods and priors. Analytical solutions should always be preferred when the problem is solvable. However, in most cases, there is no closed form for the posterior, so numerical sampling and integration methods are employed, taking advantage of the Metropolis-Hastings algorithm for example. A particularly widespread implementation of such algorithms employs Markov chain Monte Carlo (MCMC) methods. This is the case, for example, of the sampler `emcee`[4] [4], which we will use in the following numerical exercises.

2.2 Likelihood: Bayesian and frequentist approaches

The likelihood $\mathcal{L}_{\mathbf{d}}(\boldsymbol{\theta}) \equiv p(\mathbf{d}|\boldsymbol{\theta})$ describes the probability of occurrence of observed data $\mathbf{d}$ for a set of fixed parameters $\boldsymbol{\theta}$. Let's assume a set of statistically independent measurements $\mathbf{d} = \{d_i\}$, where the PDF of d_i is described by $p_i(d_i|\boldsymbol{\theta})$. The independence of the measurements means that the likelihood can be written as $\mathcal{L}_{\mathbf{d}}(\boldsymbol{\theta}) = \prod_{i=1}^{n_{\mathrm{pts}}} p_i(d_i|\boldsymbol{\theta})$.

In a counting experiment, for example, we could measure $\{k_i\}$ as a function of a quantity $\{x_i\}$ and model these counts by a function f such that the Poisson expectation of the number of counts is given by $\lambda = f(x;\boldsymbol{\theta})$. The likelihood of the data is then described by the product of Poisson distributions described in equation (3):

$$\mathcal{L}_{\{x_i,k_i\}}(\boldsymbol{\theta}) = \prod_{i=1}^{n_{\mathrm{pts}}} \frac{\Big(f(x_i;\boldsymbol{\theta})\Big)^{k_i} \exp\Big(-f(x_i;\boldsymbol{\theta})\Big)}{k_i!} \tag{16}$$

[4] `https://emcee.readthedocs.io`

For example, for the simple model where f is a constant function of parameter $\boldsymbol{\theta} = \mu$, the likelihood reads $\mathcal{L}_{\{x_i,k_i\}}(\mu) = \mu^{\sum k_i} \exp(-n_{\text{pts}}\mu) / \prod k_i!$. The likelihood is maximum for $\hat{\mu} = k_{\text{tot}}/n_{\text{pts}}$ (average number of counts), where $k_{\text{tot}} = \sum k_i$ is the total number of counts. For a flat prior on μ, the value $\hat{\mu}$ is also the maximum a posteriori, relatively close to the mean and median values of μ for $k_{\text{tot}} \gg 1$, as can be verified numerically.

The Poisson example above illustrates that in the most regular cases (unimodal likelihood, sufficiently peaked and relatively symmetric), we can approximate the best parameter a posteriori using the maximum likelihood, $\hat{\theta} = \arg\max \mathcal{L}_{\mathbf{d}}$. This is the basis of the so-called *frequentist* approach, which consists in determining the parameter values for which the model takes most frequently the observed values. The range of acceptable parameters is then estimated using confidence intervals, whose unambiguous definition is trickier than for Bayesian credible intervals (see [2], in particular the methods of Feldman & Cousins and Rolke et al. for counting experiments). For a sufficiently strong signal (i.e. when the determination of upper limits is not necessary), the bounds of the 68% confidence interval $[\theta_-; \theta_+]$ around the best parameter $\hat{\theta}$ are often estimated from the likelihood profile $\mathcal{L}_{\mathbf{d}}(\theta)$. The bounds at 1σ can thus be estimated using the equation $\mathcal{L}_{\mathbf{d}}(\theta_\pm) = \mathcal{L}_{\mathbf{d}}(\hat{\theta}) \exp(-1/2)$, which corresponds to $x = \mu \pm 1\sigma$ in equation (5). So the confidence interval in a one-dimensional parameter space is given by the equation $\ln \mathcal{L}_{\mathbf{d}}(\theta) \geq \ln \mathcal{L}_{\mathbf{d}}(\hat{\theta}) - 1/2$. Note that defining the confidence interval using $\Delta \ln \mathcal{L}_{\mathbf{d}} = 1/2$ is only valid for a 1D likelihood profile.[5] For a multi-dimensional parameter space, we can build a 1D profile for parameter θ_0 by profiling the other parameters, i.e. by maximizing the likelihood at θ_0 fixed with respect to the other parameters. This is the approach used by the `minos` method of the minimizer `Minuit`.[6] This approach makes it possible to define asymmetric error bars, $\hat{\theta} - \theta_- \neq \theta_+ - \hat{\theta}$, for general likelihood profiles.

In the case of symmetric error bars, we can also define the uncertainty $\sigma_\theta \equiv \hat{\theta} - \theta_- = \theta_+ - \hat{\theta}$ from a Gaussian property derived from the equation (5):

$$\sigma_\theta^2 = \left[-\frac{\partial^2 \ln \mathcal{L}_{\mathbf{d}}}{\partial^2 \theta} \bigg|_{\hat{\theta}} \right]^{-1} \tag{17}$$

In the simplest cases, the best parameter and its confidence interval can be determined analytically. Returning to the constant modeling of a counting experiment discussed at the beginning of this section 2.2, $\mathcal{L}_{k_{\text{tot}},n_{\text{pts}}}(\mu) \propto \mu^{k_{\text{tot}}} \exp(-n\mu)$, the zero of the first derivative gives a maximum likelihood in $\hat{\mu} = k_{\text{tot}}/n_{\text{pts}}$ and the calculation of the second derivative gives $\sigma_\mu = \hat{\mu}/\sqrt{k_{\text{tot}}}$.

The equation (17) is easily generalized to several dimensions using the second partial derivative matrix, called the Hessian matrix, H, such that

$$\mathrm{H}_{ij} = -\left. \frac{\partial^2 \ln \mathcal{L}_{\mathbf{d}}}{\partial\theta_i \partial\theta_j} \right|_{\hat{\boldsymbol{\theta}}} \tag{18}$$

Note 5. The mathematical expectation of the Hessian matrix, i.e. its mean value over all possible parameters, provides the Fisher information matrix, which plays an important role in Bayesian analysis.

The inverse of the Hessian matrix, $\mathrm{V} = \mathrm{H}^{-1}$, provides the covariance matrix of the parameters, whose diagonal gives the squared uncertainties on the parameters, $\mathrm{V}_{ii} = \sigma_{\theta_i}^2$, and the non-diagonal terms of the covariance matrix, $\mathrm{V}_{ij} = \rho_{ij}\sigma_{\theta_i}\sigma_{\theta_j}$ for $i \neq j$, determine the correlation

[5] Other values of $\Delta \ln \mathcal{L}_{\mathbf{d}}$ must be used to identify the confidence region at 68% e.g. in 2D. To convince ourselves, we can check that, for a 2D Gaussian of equal widths along x and y, the integral under the curve at 1σ of the maximum is not 68% but only 39% (see e.g. `https://corner.readthedocs.io/en/latest/pages/sigmas/`).

[6] `https://iminuit.readthedocs.io/en/stable/about.html`

between parameters, $-1 \leq \rho_{ij} \leq 1$. Inversion of the Hessian matrix is the approach used by the `hesse` method of the `Minuit` minimizer. The covariance between pairs of parameters is often represented using the uncertainty ellipse (or standard error ellipse) shown in figure 3. In this figure, the two parameters are anti-correlated, $\rho_{ij} < 0$, in a non-maximum way, i.e. $\rho_{ij} > -1$.

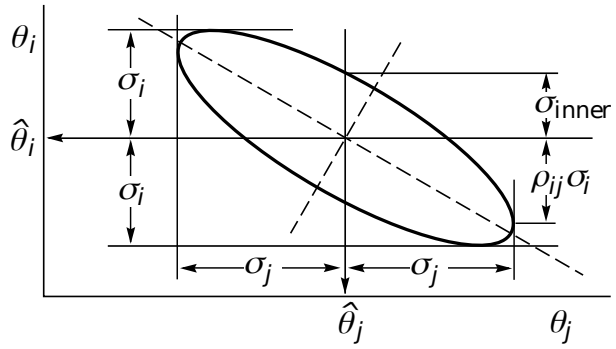

Figure 3: Ellipse of uncertainty for two parameters θ_i and θ_j. *Adapted from [2].*

2.3 Gaussian likelihood: frequentist approach and χ^2 minimum

Let's now assume, in a frequentist approach, that we have access to a set of independent measurements $\{y_i\}$ with uncertainties $\{\sigma_i\}$ as a function of $\{x_i\}$. We are no longer limited to the simple counting experiment, for which $y_i = k_i$ and $\sigma_i = \sqrt{k_i}$. Summarizing the values to their best estimate and standard deviation, $y_i \pm \sigma_i$, implies that the likelihood terms are approximated by Gaussian functions, i.e.

$$\mathcal{L}_{\{x_i,y_i,\sigma_i\}}(\boldsymbol{\theta}) = \prod_{i=1}^{n_{\text{pts}}} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{\big(y_i - f(x_i;\boldsymbol{\theta})\big)^2}{2\sigma_i^2} \right) \tag{19}$$

Maximizing the likelihood is equivalent to minimizing deviance $D = -2\ln\mathcal{L}$, i.e.

$$D = \sum_{i=1}^{n_{\text{pts}}} \frac{\big(y_i - f(x_i;\boldsymbol{\theta})\big)^2}{\sigma_i^2} + \sum_{i=1}^{n_{\text{pts}}} \ln\left(2\pi\sigma_i^2\right) \tag{20}$$

In particular, we can define the saturated deviance, $D_{\text{sat}} = -2\ln\mathcal{L}_{\text{sat}}$, which corresponds to an ideal model that perfectly reproduces the data, i.e. $f_{\text{sat}}(x_i;\boldsymbol{\theta}) = y_i$. In the Gaussian equation (20), $D_{\text{sat}} = \sum_{i=1}^{n_{\text{pts}}} \ln\left(2\pi\sigma_i^2\right)$, which is independent of the model parameters $\boldsymbol{\theta}$ by construction. The remaining term is used to define the quantity to be minimized, i.e. $\chi^2 \equiv D - D_{\text{sat}}$.

χ^2 estimator or least-squares method

The χ^2 is the sum of the squared differences between the model and the data, weighted by the inverse of the squared uncertainties:

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{n_{\text{pts}}} \frac{\big(y_i - f(x_i;\boldsymbol{\theta})\big)^2}{\sigma_i^2} \tag{21}$$

y_i: value measured in x_i;

$f(x_i;\boldsymbol{\theta})$: value taken by the model in x_i for a set of parameters $\boldsymbol{\theta}$;

σ_i: uncertainty of measurement y_i.

Note 6. The χ^2 is an adimensioned quantity. Weighting by the inverse of the squared uncertainty follows naturally from the Gaussian likelihood. This is a good property: the larger the uncertainty on a measurement, the less the impact of this term on the sum; the smaller the uncertainty, the more important the associated measurement.

Maximum-likelihood estimation of the best parameters and their uncertainties is naturally extended to the least-squares method, which can be implemented using the `LeastSquares` class of `Minuit`. Thus, we find as best parameters $\hat{\boldsymbol{\theta}} = \arg\min \chi^2$ with uncertainties provided by the diagonal terms of the covariance matrix $\boldsymbol{\sigma}_\theta = \sqrt{\text{diag V}}$. The covariance matrix is defined using the Hessian matrix by $\left(\text{V}^{-1}\right)_{ij} = \frac{1}{2}\left.\frac{\partial\chi^2}{\partial\theta_i\partial\theta_j}\right|_{\hat{\boldsymbol{\theta}}}$ (`hesse` method from `Minuit`). Asymmetric error bars can also be reconstructed using the χ^2 profile as a function of each parameter (`minos` method from `Minuit`), which follows a parabola for a Gaussian likelihood profile. The 68% confidence interval is then determined by the equation $\chi^2(\theta) \leq \chi^2(\hat{\theta}) + 1$, where the value of $\Delta\chi^2 = 1$ corresponds to the interval at $\pm 1\sigma$ for a 1D parameter space. This value of $\Delta\chi^2$ is derived from $\Delta \ln \mathcal{L}_\mathbf{d} = 1/2$, with $\chi^2 = -2 \ln \mathcal{L}/\mathcal{L}_\text{sat}$.

Note 7. The least-squares method described in equation (21) is also the one followed by the `curve_fit` function in `scipy.optimize`, provided one specifies the σ_i uncertainties (parameter `sigma = sigma_i` and option `absolute_sigma = True`). Without these inputs, which are not the default ones (!), the parameter uncertainties returned by `curve_fit` are meaningless in the context described in this document.

In simple cases, an analytical solution can be obtained by derivation around the maximum likelihood or minimum χ^2. Let's take the simplest case, the constant model $f(x) = a$. The best value of a corresponds to the one for which the χ^2 is minimum and therefore for which $\partial\chi^2/\partial a = 0$ and $\partial^2\chi^2/\partial a^2 > 0$. In our case,

$$\begin{aligned}\frac{\partial\chi^2}{\partial a} &= \sum_i \frac{\partial}{\partial a}\left[\frac{(y_i - a)^2}{\sigma_i^2}\right] \\ &= -2\sum_i \frac{y_i - a}{\sigma_i^2}\end{aligned} \tag{22}$$

$$\begin{aligned}\frac{\partial^2\chi^2}{\partial^2 a} &= -2\sum_i \frac{\partial}{\partial a}\left[\frac{y_i - a}{\sigma_i^2}\right] \\ &= 2\sum_i \frac{1}{\sigma_i^2} > 0\end{aligned} \tag{23}$$

According to Eq. (22), the best value of a corresponds to

$$\hat{a} = \frac{\sum_i y_i/\sigma_i^2}{\sum_i 1/\sigma_i^2} \tag{24}$$

The average value of a set of data with uncertainties is therefore an average weighted by the inverse square of the uncertainties.

The uncertainty σ_a on $\hat{a}$ can be determined from the second derivative, $\sigma_a = \left[\frac{1}{2}\frac{\partial^2\chi^2}{\partial^2 a}\right]^{-1/2}$, i.e.

$$\sigma_a = \frac{1}{\sqrt{\sum_i 1/\sigma_i^2}} \tag{25}$$

If the uncertainties are the same for all points ($\sigma_i = \sigma$), then according to the equation (25) $\sigma_a = \frac{\sigma}{\sqrt{n_{\text{pts}}}}$.

In practice, fitting algorithms minimize a χ^2 either analytically (e.g. for polynomial models) or numerically to find the best parameters and associated uncertainties. Note that parameters must be initialized to relevant values to avoid getting stuck in a local minimum. Note also that parameter uncertainty only makes sense if the model is a "good" model, which we will further discuss in section 3.

Numerical exercises The exercises suggested here are based on the `emcee`[a] and `iminuit`[b] documentations. The two sets of tutorial can be used to develop Bayesian and frequentist analyses, respectively.

Exercise 1. Bayesian analysis

- Create a jupyter notebook named Bayesian Analysis.
- Generate a set of test data, comprising 21 points equally spaced between $x_0 = 100$ and $x_1 = 200$, following a normal distribution centred on $y = a_{\text{true}} + b_{\text{true}}x$ with $a_{\text{true}} = 10$ and $b_{\text{true}} = 0.1$ and standard deviation $\sigma_y = 2$.
- Define the functions `log_likelihood` (Gaussian likelihood) and `log_prior` (constant prior) using the tutorial *Fitting a model to data* from `emcee`.
- Using the same tutorial (subsections *Marginalization & uncertainty estimation* and *Results*), display a corner plot to determine the credible intervals of the fitted parameters a and b.

Exercise 2. Frequentist analysis

- Create a jupyter notebook called Frequentist Analysis
- Generate a test data set as in exercise "Bayesian analysis".
- Define a function `least_square` using the tutorial *Basics* from `iminuit`.
- Using the same tutorial (subsections *Quick start*, *Quick access to fit results* and *Plotting / Drawing confidence regions*), display a corner plot to determine the confidence intervals of the fitted parameters a and b.

Exercise 3. Comparison of approaches

- Create a jupyter notebook named Comparative analysis.
- Create functions `bayesian_fit`, `frequentist_fit` and `analytic_fit` returning the best parameters of an affine model and associated uncertainties for a data set `x_i, y_i, sigma_i`. For the analytical approach, the use of pen and paper is strongly encouraged before coding!

- Generate n test data sets as before, for various a and b parameters, and compare the best parameters obtained with the three approaches.
- Generate n test data sets as before, for fixed a and b parameters, and compare the 68% intervals obtained with the three approaches.

Exercise 4. Towards a good parametrisation of the model

- Repeat the previous exercises for data points equally spaced between $x_0 = -50$ and $x_1 = 50$. Compare the covariance matrices.
- Comment on the difference with the results obtained for $x_0 = 100$ and $x_1 = 200$. Can the model be re-parameterised to get results comparable to those obtained for data centered around $x = 0$?

[a] `https://emcee.readthedocs.io`
[b] `https://iminuit.readthedocs.io`

3 Overfitting and underfitting

We examined in section 2 how to identify the best parameters $\hat{\boldsymbol{\theta}}$ of a model and how to estimate the associated uncertainties $\boldsymbol{\sigma}_\theta$. However, we have overlooked a key question: what model should we propose for the data? How many free parameters are relevant?

For n_{pts} measurement points, $\{x_i, y_i, \sigma_i\}$, we could naively propose a polynomial model with n_{pts} parameters. By implementing a least-squares method, we could then obtain a model passing perfectly through each of the points, i.e. a best-fit model yielding a null χ^2. But in this case, what use would measurement uncertainties be? Furthermore, the relevance of such a model in terms of interpolation between measurement points and extrapolation beyond them would be highly debatable.

In this section, we discuss how to determine whether a model $f(x; \hat{\boldsymbol{\theta}})$, whose parameters have been optimized to best fit a dataset, is a "good" model. This way, we can determine whether the number of parameters employed is appropriate. The discussion applies more naturally to frequentist approaches. Note, however, that the notions discussed here can be extended to the Bayesian formalism.

3.1 Goodness of fit

The suitability of a model is linked to the question of the expected value for the statistical estimator, in this case the χ^2. If the model "passes through the error bars", then the ratios $(y_i - f(x_i; \hat{\boldsymbol{\theta}}))^2/\sigma_i^2$ are of the order of 1 and the χ^2, which is the sum of these ratios, is of the order of n_{pts}. In practice, we need to penalize the χ^2 estimate for the dimension of the parameter space, i.e. determine the number of truly independent terms in the sum. The number of degrees of freedom (d.o.f. or n.d.f.) is defined as $\nu = n_{\text{pts}} - n_{\text{par}}$, where n_{par} is the number of free parameters. It can be shown that, if the data points are a realization of the proposed model, the value of χ^2 tends towards $\nu \pm \sqrt{2\nu}$ for a large number of points.

A model with a reduced χ^2, χ^2/ν, of the order of $1 \pm \sqrt{2/\nu}$ could therefore be considered satisfactory. A model with a reduced χ^2 close to zero would be "too good to be true", and

a model with a reduced χ^2 greater than say 2 would not "pass through the error bars". In practice, a low χ^2 value corresponds either to overfitting, i.e. too many free parameters, or to overestimating the uncertainties of the measurement points. A large χ^2 value corresponds in practice either to an underfitting, i.e. an overly simple model, or to an underestimation of the uncertainties on the measurement points.

A more quantitative goodness-of-fit estimator than the reduced χ^2 can be determined based on the expected distribution of the χ^2. If the measurement points are a Gaussian realization of the model, then for a number ν of degrees of freedom the χ^2 follows the PDF:

$$\mathbb{P}(X = x) = \frac{x^{\nu/2-1} \exp(-x/2)}{2^{\nu/2}\Gamma(\nu/2)}, \tag{26}$$

where Γ is the gamma function. The integral of the PDF above the obtained value χ^2_0, i.e. the SF of X evaluated in χ^2_0 provides the p-value: $p = S_{\chi^2}(\chi^2_0, \nu)$. If the measurements are a realization of the model, the p-value follows a uniform distribution between 0 and 1. In practice, for a given dataset, the fit is usually considered "correct" if the p-value lies between 10% and 90%. The question of underfitting arises for a p-value below 10% and overfitting for a p-value above 90%.

Note 8. The p-value for the χ^2 distribution can be estimated in Python using the functions `stats.chi2.sf(chi2, n.d.f.)` from `scipy` in Python or `TMath::Prob(chi2, n.d.f.)` in the C/C++ version of `ROOT`.

Note 9. The uniform distribution of the p-value can be used to adjust the thresholds to be considered for overfitting and underfitting when fitting models to multiple datasets. So, if you run 100 fits in a row, you should not be surprised to get p-values of the order of 1% or 99% if the datasets are a realization of the tested model.

Note that the notion of a "good" model is somewhat misleading. By definition, the estimation of the p-value answers the question: with what confidence level can we reject the observation of a χ^2 value at least as high as the one obtained? By negation, we consider in practice that, for a "good" model, the χ^2 value obtained is far from being rejected.

3.2 Model rejection and statistical significance

We can compare the 10% and 90% thresholds discussed earlier with the number of Gaussian standard deviations, as shown in figure 2. For $Z = \frac{x-\mu}{\sigma}$ in equation (11), the integral of the distribution beyond $Z\sigma$ is equal to

$$p = \text{erfc}(Z/\sqrt{2}), \tag{27}$$

or $p \approx 5\%$ beyond 2σ.

The equation (27) introduces the notion of statistical significance, Z, expressed as the number of Gaussian standard deviations or the number of σ. This notion is frequently used in our scientific fields to qualify the probability that a signal does or does not originate from background noise. Assuming that the background noise is Gaussian distributed around a mean μ with a standard deviation σ, then the p-value p is the probability of obtaining, through statistical fluctuations, a signal at least $Z\sigma$ away from the mean noise value (two-sided interval test). If a strictly positive signal is expected, the probability of obtaining a signal greater than $\mu + Z\sigma$ from the background noise is $p/2$ (one-sided interval test).

Given the number of experiments carried out and the number of configurations tested to discover e.g. the Higgs boson or gravitational waves, a p-value of the order of one percent is not sufficient to ensure a discovery.[7] We usually speak of a hint for a significance above 3σ or $p \lesssim 0.3\%$, an evidence above 4σ or $p \lesssim 0.6 \times 10^{-4}$ and detection or discovery (first detection) above 5σ or $p \lesssim 0.6 \times 10^{-6}$.

Note 10. The functions `special.erfcinv` in `scipy` and `TMath::ErfcInv` in `ROOT` can be used to convert a p-value into a Gaussian significance, Z, by inverting the equation (27).

3.3 Choosing the model

Let's assume that we have found a model resulting in an acceptable fit to the data, i.e. one for which the χ^2 p-value is between 10% and 90%. Couldn't a more complex model (e.g. linear model instead of constant) provide an even better fit?

If the simpler H_0 model is nested within the more complex H_1 model (e.g. the constant model is nested within the linear, which is nested within the quadratic), the χ^2 value obtained for H_1 is necessarily lower than that obtained for H_0: $\chi^2_1 \leq \chi^2_0$. However, the number of parameters of H_1, $n_{\mathrm{par},1}$, is greater than that of H_0, $n_{\mathrm{par},0}$, so that $\nu_1 < \nu_0$. The two effects may counteract each other when estimating the p-values for H_0 and H_1, so the model complexity needed to reproduce the data may be hard to determine.

Wilks' theorem provides an objective selection criterion for frequentist analyses (likelihood ratio test or χ^2 difference test). If the parameter space $\boldsymbol{\theta}_0$ of dimension $n_{\mathrm{par},0}$ is included in the larger space $\boldsymbol{\theta}_1$ of dimension $n_{\mathrm{par},1}$, i.e. if H_0 and H_1 are nested, then the difference of the best χ^2 in the spaces $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, $\Delta\chi^2 = \chi^2_0 - \chi^2_1$, follows a χ^2 distribution with $\Delta\nu = n_{\mathrm{par},1} - n_{\mathrm{par},0}$ degrees of freedom.

We can therefore evaluate $p = S_{\chi^2}(\Delta\chi^2, \Delta\nu)$ to estimate the probability that the data are better reproduced by H_1 than by H_0. This p-value can be considered as the degree of rejection of the H_0 hypothesis in favor of H_1.[8]

Note 11. An interesting mathematical property can be used to quickly determine the rejection significance of a simple model in favor of one with a single additional parameter ($n_{\mathrm{par},1} = n_{\mathrm{par},0} + 1$). Indeed, by inverting the equation (27), the significance simplifies as follows:

$$\begin{aligned} Z &= \sqrt{2}\ \mathrm{erfc}^{(-1)}\big(S_{\chi^2}(\Delta\chi^2, \Delta\nu)\big) \\ &= \sqrt{\Delta\chi^2} \quad \text{for } \Delta\nu = 1 \end{aligned} \tag{28}$$

The threshold number of $Z\sigma$ used in practice to parameterize a background model is often less restrictive than that used to claim the detection of a new effect. For such background modeling, we recommend adopting the more complex model if it is favored over the simpler model by 2σ or 3σ.

[7] Such an assertion illustrates the main Bayesian criticism of frequentist approaches: p-values behave as probabilities only within a restricted framework, and do not necessarily take into account all the prior measurements made, e.g. by other experiments. In the frequentist literature, one can find the terms *pre-trial* and *post-trial* p-values, which allow the authors to distinguish between taking into account or not the fact that multiple tests have been run for several configurations. Taking into account this "look-elsewhere effect" addresses the problem of multiple comparisons, without requiring the full Bayesian formalism.

[8] A test statistic similar to the likelihood ratio exists in Bayesian analysis: the Bayes factor.

3.4 Model validation

The statistical approaches described here can be used to quantitatively assess the ability of models to reproduce data. However, it may be that the type of models evaluated is not appropriate, or that the data present one or more irregularities that are poorly taken into account by Gaussian uncertainties (outliers).

One should always make a point of evaluating the distance between the best-fit model $f(x, \hat{\boldsymbol{\theta}})$ and the data $\{x_i, y_i, \sigma_i\}$ by plotting the residuals, $y_i - f(x_i, \hat{\boldsymbol{\theta}})$, and normalized residuals or pulls, $\big(y_i - f(x_i, \hat{\boldsymbol{\theta}})\big)/\sigma_i$, as a function of x_i.

The pulls are often used to identify the points or group of points contributing most to the χ^2 value. These normalized residuals follow a standard normal distribution if the data are a realization of the model. The points must also be randomly distributed on either side of the model. The residuals can be used to assess large absolute deviations between model and data. Comparing residuals and pulls often helps identifying poorly estimated uncertainties.

4 Propagation of uncertainties and model parameterization

We now know how to determine whether a model satisfactorily matches a data set, using a test statistic such as χ^2 in frequentist analysis. For a model fitting well the data, we are also able to determine the best parameters $\hat{\boldsymbol{\theta}}$ and their covariance matrix $\mathrm{V}_{\hat{\boldsymbol{\theta}}}$, which contains the square of the uncertainties on its diagonal and the correlation terms outside. This representation, using the first two moments of $\boldsymbol{\theta}$, fully defines the Gaussian likelihood. Using the first two moments is merely a practical approximation to the multi-dimensional posterior that would be obtained in Bayesian analysis.[9]

Suppose the physical quantity of interest Ω is a function of $\boldsymbol{\theta}$, $\Omega = g(\boldsymbol{\theta})$. We wish to determine the PDF of Ω from the multi-dimensional distribution of $\boldsymbol{\theta}$. In Bayesian analysis, we often keep trace of the path followed by Monte Carlo Markov chains along $\boldsymbol{\theta}$, so that the distribution of Ω can easily be reconstructed by evaluating the function g at each point of the parameter space explored by the chains. In frequentist analysis, the PDF of Ω is approximated by a Gaussian centered on $g(\hat{\boldsymbol{\theta}})$ and of variance determined by the propagation of uncertainties.

4.1 Propagation of uncertainties

The variance of the variable Ω can be determined using a linearization around $\hat{\Omega} = g(\{\hat{\theta}_i\})$:

$$\Omega \approx g(\{\hat{\theta}_i\}) + \sum_i \left.\frac{\partial g}{\partial \theta_i}\right|_{\hat{\theta}_i} \big(\theta_i - \hat{\theta}_i\big) \tag{29}$$

This is a strict equality if g is a linear function of $\{\theta_i\}$. Otherwise, the Taylor expansion in equation (29) is based on small variations around $\{\hat{\theta}_i\}$, i.e. small uncertainties on these

[9] For flat priors, this simplification is akin to the so-called Laplace approximation in Bayesian analysis.

parameters. The above equation can be written as

$$\Omega - \hat{\Omega} \approx \sum_i \left.\frac{\partial g}{\partial \theta_i}\right|_{\hat{\theta}_i} (\theta_i - \hat{\theta}_i), \tag{30}$$

By squaring both sides and averaging over the $\{\theta_i\}$, we obtain the following formula.

Propagation of uncertainties

The variance of the variable $\Omega = g(\{\theta_i\})$, linearized around the $\{\hat{\theta}_i\}$, is

$$\sigma_\Omega^2 = \sum_i \left.\frac{\partial g}{\partial \theta_i}\right|_{\hat{\boldsymbol{\theta}}}^2 \sigma_{\theta_i}^2 + 2 \sum_{i>j} \left.\frac{\partial g}{\partial \theta_i}\right|_{\hat{\boldsymbol{\theta}}} \left.\frac{\partial g}{\partial \theta_j}\right|_{\hat{\boldsymbol{\theta}}} \rho_{ij} \sigma_{\theta_i} \sigma_{\theta_j} \tag{31}$$

$\mathrm{V}_{ii} = \sigma_{\theta_i}^2$: variance or squared uncertainty of θ_i;

$\mathrm{V}_{ij} = \rho_{ij} \sigma_{\theta_i} \sigma_{\theta_j}$: covariance of $\{\theta_i\}$.

This equation is homogeneous: $\sigma_{\theta_i}^2$ has the same dimension as $(\partial \theta_i)^2$ and σ_Ω^2 has the same dimension as $(\partial g)^2$. We have established the propagation formula for a one-dimensional variable Ω, but the reasoning can easily be generalized to $\boldsymbol{\Omega} = \{\Omega_k\}$. Uncertainty propagation, involving the Jacobian matrix $\mathrm{J} = \left[\frac{\partial \Omega_k}{\partial \theta_i}\right]$, then reduces to $\mathrm{V}_{\boldsymbol{\Omega}} = \mathrm{J} \mathrm{V}_{\boldsymbol{\theta}} \mathrm{J}^\top$.

Note 12. The `propagate` method in the Python library `jacobi` can easily be used to evaluate the above equation numerically.[10]

Let's study two practical cases analytically:

- Weighted sum: $\Omega = \sum_i p_i \times \theta_i$ where p_i are constant numerical values (no associated uncertainty). We then obtain:

  $$\sigma_\Omega^2 = \sum_i p_i^2 \sigma_{\theta_i}^2 \tag{32}$$

  If we average quantities with the same uncertainty, i.e. $\sigma_{\theta_i} = \sigma$ and $p_i = 1/n_{\mathrm{pts}}$, we find $\sigma_\Omega = \sigma / \sqrt{n_{\mathrm{pts}}}$, as in equation (25).

- Weighted product: $\Omega = \prod_i \theta_i^{p_i}$ where the p_i are constant numerical values (no associated uncertainty). We then obtain:

  $$\sigma_\Omega^2 = \sum_i \left(p_i \times \theta_i^{p_i - 1} \times \prod_{j \neq i} \theta_j^{p_j} \right)^2 \sigma_{\theta_i}^2$$

  $$\text{i.e.} \quad \left(\frac{\sigma_\Omega}{\Omega}\right)^2 = \sum_i p_i^2 \left(\frac{\sigma_{\theta_i}}{\theta_i}\right)^2 \tag{33}$$

  The latter expression shows that it is often useful to push the analytical calculation a little further in order to rewrite the solution in a simple way.

[10] See `https://iminuit.readthedocs.io/en/stable/notebooks/error_bands.html`.

4.2 Minimizing covariance between model parameters

Let's now apply the uncertainty propagation formula in equation (31) to our initial problem. We have been able to determine the parameters $\boldsymbol{\theta}$ and the associated covariance matrix, so that the function $f(x;\hat{\boldsymbol{\theta}})$ best fits the data $\{x_i, y_i, \sigma_i\}$. The model can be seen as a function of the parameters, with an uncertainty

$$\sigma_f^2 = \sum_i \left(\frac{\partial f}{\partial \theta_i}\right)^2 \sigma_{\theta_i}^2 + 2\sum_{i>j} \frac{\partial f}{\partial \theta_i}\frac{\partial f}{\partial \theta_j}\rho_{ij}\sigma_{\theta_i}\sigma_{\theta_j} \tag{34}$$

Just as the model f depends on the variable x, the model uncertainty σ_f depends on this same variable. The constraints on the model therefore depend on where it is evaluated, with smaller uncertainties in the range covered by the data and larger uncertainties outside the data range.

Let's illustrate this point with a linear model, $f(x;\{a,b\}) = a + bx$. The uncertainty on the model induced by the uncertainties on $\{a,b\}$ is then

$$\sigma_f^2 = \sigma_a^2 + x^2\sigma_b^2 + 2x\rho_{ab}\sigma_a\sigma_b \tag{35}$$

This uncertainty is minimal for $x = x_0$ such that $\partial\sigma_f^2/\partial x = 0$, i.e $x_0 = -\rho_{ab}\sigma_a/\sigma_b$. We can then rewrite the model to minimize the correlation between its parameters, i.e. $g(x;\{a_0,b_0\}) = a_0 + b_0(x - x_0)$ with $b_0 = b$ and $a_0 = a + bx_0$, the value x_0 being fixed. The covariance between a_0 and b_0 is then

$$\begin{aligned}
\mathrm{Cov}(a_0, b_0) &= \mathrm{Cov}(a + bx_0, b) \\
&= \mathrm{Cov}(a,b) + x_0\mathrm{Cov}(b,b) \\
&= \rho_{ab}\sigma_a\sigma_b + x_0\sigma_b^2 \\
&= 0
\end{aligned} \tag{36}$$

The value x_0 is called the decorrelation point or pivot point. For a wide class of models (e.g. linear, exponential, power law), we can find a point of strict decorrelation, i.e. such that the covariance between parameters is strictly zero. This point is generally located in the middle of the range covered by the $\{x_i\}$.

From an analytical point of view, expressing the model as $f(x) = a + bx$ or $g(x) = a_0 + b_0(x - x_0)$ may seem equivalent. However, let's not forget that the propagation of uncertainties in equation (31) is derived from a Taylor expansion, which is all the more valid (for a non-linear model) the smaller the uncertainties. From a numerical point of view, the parameter space is easier to explore for decorrelated parameters, which speeds up the fitting procedure. Whenever possible, therefore, we should parametrize the model in a way adapted to the data, i.e. so as to minimize correlations between parameters.

5 Conclusion

This document provides an introduction to the inference methods used in the scientific themes covered by the NPAC master's program. Comparing a model to a data set provides probabilistic inferences about the validity of the model and the values of its parameters. The most appropriate framework for inferring parameters and their credibility intervals is the Bayesian formalism (see equation (12)). In the latter, the probability distribution functions of parameters are determined by means of integration, a process known as marginalization. In the frequentist formalism, which only covers the likelihood term introduced in the Bayesian approach, probability is interpreted as the frequency of the outcome of a repeated experiment. The frequentist approach is therefore particularly well suited to determining the statistical significance of a signal. In the case of Gaussian distributions, likelihood maximization adopted in the frequentist formalism reduces to a least-squares approach, in which one minimizes the χ^2. The latter is the sum of the squared differences between model and data weighted by the squared uncertainties on the data (see equation (21)). Where a Bayesian approach integrates, a frequentist approach finds the zero-derivative point and estimates uncertainties by profiling or second-derivative calculation around this point. The uncertainties can be propagated naturally in the Bayesian approach, while the formula for propagating uncertainties in equation (31) can be used in the frequentist approach. In both cases, the models to be fitted to the data are better formulated by minimizing correlations between their parameters. When comparing results from multiple experiments, frequentist methods provide a natural framework for estimating the degree of tension between results, while Bayesian methods allow us to properly combine information and determine the degree of credibility of parameter values.

Although they answer different questions, frequentist and Bayesian approaches often give similar results. In the simplest cases, they can be implemented analytically, and more generally numerically using well-established libraries. Interested readers are encouraged to practice both approaches in order to grasp their limitations and interests. The bibliography at the end of this document will provide a deeper insight into the foundations and methods of statistical inference.

Numerical exercises

Exercise 5. Fitting a model to a histogram

- Create a jupyter notebook called Histogram Analysis.
- Using the introduction of the *Cost Functions* tutorial from `iminuit`, generate a test data set, comprising 1000 background events distributed according to an exponential of slope 1 and 100 signal events distributed according to a Gaussian of mean $\mu = 1$ and width $\sigma = 0.1$.
- The *Maximum-likelihood fits/Binned fit* sub-section of this tutorial illustrates how to analyze binned data with a Poissonian maximum-likelihood method. Using this method, fit a signal + background model to the generated data.
- Perform a similar fit using a least-squares method (what uncertainty should be used for the points?). Compare with results obtained using the binned analysis. Evaluate the quality of the fit using the χ^2 and the associated p-value.

- Plot residuals and pulls for both fits, using the tutorial *RooFit tutorials/109: chi-square residuals and pulls* as a guide.

Exercise 6. Statistical significance

- Create a sub-section of the jupyter notebook named Statistical significance.
- Using the method of least squares, fit to the previous histogram:
  - A model H_0 described by an exponential with two free parameters. Determine the χ^2 obtained, i.e. χ^2_0.
  - A model H_1 described by an exponential with two free parameters plus a Gaussian of free amplitude but with mean and width fixed at $\mu = 1$ and $\sigma = 0.1$. Determine the χ^2 obtained, i.e. χ^2_1.
- By how many degrees of freedom do H_0 and H_1 differ? Are they nested? Using the equation (27), determine the statistical significance at which the signal is detected.
- Perform a similar fit for a model H_2 described by an exponential with two free parameters plus a Gaussian with three free parameters. At which significance level is this model preferred to H_0? to H_1?

Exercise 7. Model error band

- Create a jupyter notebook named Error band.
- As in Exercise 1, generate a data set following a linear model.
- Fit a linear model with a least-squares method to this data set.
- Using the tutorial *How to draw error bands* from `iminuit`, plot the best model and its 68% confidence interval.
- As in Exercise 1, fit a linear model to the data in a Bayesian framework.
- For a given value of x, plot the distribution of expected model values $y = f(x; \{\mathbf{a}, \mathbf{b}\})$. Return the quantiles at 16% and 84% of this distribution.
- Plot the 68% credibility interval of the model and compare it with the 68% confidence interval.

Exercise 8. Bonus

- Create a jupyter notebook called Ultimate Analysis.
- Generate a data set as in Exercise 5.
- Build an analysis to identify the best model and its parameters, determine the goodness of fit and residuals, the degree of statistical significance of the signal and the error band on the model.

References

[1] Glen Cowan for the Particle Data Group. "Review of Particle Physics. 39. Probability." In: *Progress of Theoretical and Experimental Physics* 2022.8, 083C01 (Aug. 2022), p. 083C01. DOI: 10.1093/ptep/ptac097. URL: https://pdg.lbl.gov/2023/web/viewer.html?file=../reviews/rpp2022-rev-probability.pdf.

[2] Glen Cowan for the Particle Data Group. "Review of Particle Physics. 40. Statistics." In: *Progress of Theoretical and Experimental Physics* 2022.8, 083C01 (Aug. 2022), p. 083C01. DOI: 10.1093/ptep/ptac097. URL: https://pdg.lbl.gov/2023/web/viewer.html?file=../reviews/rpp2022-rev-statistics.pdf.

[3] William H. Press et al. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. 3rd ed. USA: Cambridge University Press, 2007. ISBN: 0521880688. URL: http://numerical.recipes/book.html.

[4] Daniel Foreman-Mackey et al. "emcee: The MCMC Hammer". In: *Pub. of the ASP* 125.925 (Mar. 2013), p. 306. DOI: 10.1086/670067. arXiv: 1202.3665 [astro-ph.IM].

Glossary

CDF cumulative distribution function. 4, 5

FWHM full width at half maximum. 5

PDF probability density function. 4–7, 13, 15

PMF probability mass function. 4, 5

SF survival function. 4, 5, 13