Practical statistics excerpts from 2023 European School of High Energy Physics

Troels C. Petersen[a]

[a]Niels Bohr Institute, Copenhagen, Denmark

These lecture notes provide an overview of fundamental statistical concepts used in high-energy physics research. The discussion begins with the philosophy of statistics, emphasizing its role in analyzing experimental data, correcting biases, and ensuring precision. Key estimators such as mean, standard deviation, skewness, and kurtosis, which help summarize datasets, are introduced. Various probability density functions (PDFs), including Binomial, Poisson, and Gaussian distributions, are explored to demonstrate their relevance in data modeling. Methods for analyzing data, such as maximum likelihood estimation and ChiSquare tests, are presented, offering techniques for extracting meaningful insights from observations. The notes also cover hypothesis testing, explaining concepts like false positive rates, p-values, and significance levels for evaluating scientific claims. Practical applications include setting observational limits and improving statistical methodology. By applying these techniques, researchers can ensure rigorous data analysis, enabling reliable conclusions and impactful discoveries in physics.

This chapter should be cited as: Practical statistics excerpts from 2023 European School of High Energy Physics, Troels C. Petersen, DOI: 10.23730/CYRSP-2025-009.103,
in: Proceedings of the 2023 European School of High-Energy Physics,
CERN Yellow Reports: School Proceedings, CERN-2025-009, DOI: 10.23730/CYRSP-2025-009, p.103.

1 Philosophy of statistics

When confronted with hard-won data, one of course wants to analyse it in the most optimal way. However, this is often a bit harder than one might think.

> "The art of drawing conclusions from experiments and observations consists in evaluating probabilities and in estimating whether they are sufficiently great or numerous enough to constitute proofs. This kind of calculation is more complicated and more difficult than it is commonly thought to be."
>
> [Antoine Lavoisier, French chemist 1743-1794]

1.1 Why Statistics?

We collect data in order to see trends and compare it to our expectations (various theories in physics). However, experiments are inherently non-deterministic due to both quantum and chaos effects. Even without these, experiments are limited in precision by cost, time, etc. We are thus limited to finite samples with limited resolution that typically need to be corrected for all sorts of nuisance effects in order to be both accurate (i.e. non-biased) and precise (i.e. with a small uncertainty). Statistics is the tool for this process, and will—along with domain knowledge—be needed for extracting good estimates with correct uncertainties.

And given the hard work that lies behind planning, building, and running large detectors in order to give birth to cutting edge data, it would be near criminal to analyse it with anything but the most powerful data analysis methods available.

A common misconception is that statistics provides a straight path forward in all situations. This is not so. There are of course great examples of clear cut cases, but more often it is wise to follow the insights of John Tukey (US statistician, 1915-2000) who argued "the need for statisticians to reject the role of 'guardian of proven truth', and to resist attempts to provide once-for-all solutions and tidy over-unifications of the subject".

1.2 Why Uncertainties?

A single number without any indication of the size of its uncertainty is useless. To see this, imagine that you had measured the speed of gravity to be 2.91×10^8 m/s. Perhaps surprisingly, such a measurement would tell you…nothing! For depending on the size of the uncertainty, you might reach three very different conclusions:

$(2.91 \pm 7.36) \times 10^8$ **m/s.** In this case, the speed of gravity could be pretty much anything, also far exceeding the speed of light or even negative.[1]

$(2.91 \pm 0.07) \times 10^8$ **m/s.** This result is consistent with the speed of light and not much else.

$(2.908 \pm 0.007) \times 10^8$ **m/s.** Here, the small uncertainty (high precision) of the measurement shows that it is NOT exactly the speed of light, and hence a new discovery of some phenomena that slows down the speed of gravity by about 2.5%.

One of the main goals of doing good experiments and, hence. physics research is to minimise the uncertainties on the results. One of the main goals of data analysis and, hence, statistics is to get the uncertainties right! Obtaining credible uncertainties is hard work bordering on an art form. While statistical methods yield known and well understood uncertainties, it is your job as an experimenter and expert to check that all sources of uncertainty are accounted for. In the end, the target is to estimate the values, distributions, and principles behind the origin of the data, thereby "luring" the truth out of nature (in science), the body (medicine), and the markets (economy).

2 The basics of statistics

The simplest data is a series of N independent observations or measurements of the same quantity, $x_1, \ldots, x_n$. It could consist of anything from heights of people to invariant masses of decayed elementary particles. The first thing we want to do is inspect, describe, and summarize the data.

2.1 Printing and plotting data

The simplest way that we can "see" the data visually is by printing a few values and plotting the data in a histogram. When reading a new data file, it is worthwhile to print the first 10–20 entries along with the total number of entries, just to get a crude impression of the data types and sizes. Make sure that you know if the data is integer, discrete, rounded, continuous, or contains Not-a-Number (NaN) or outlier values.

The obvious next step is plotting the data. For a single data series (i.e. 1D data) a histogram is the obvious choice. Always make sure that you control the histogram range and number of bins, and do so cleverly! Algorithms (ROOT or MatPlotLib) are notoriously poor at doing this, as they do not have any sense of what is important. You should for integers make sure, that the bin width is an integer, and that it is shifted by a half to ensure that the middle of the bin matches the average of the bin range (Example: For the number of hits in a certain detector with up to 20 hits, choose the range to be $[-0.5, 20.5]$ with 21 bins). The histograms shows the data distribution, and "transforms" the (potentially very long) list into x-values (middle of the bins), y-values (number of entries in each bin), and σ_y-values (uncertainty on the y-values, approximate by the square root of the y-values themselves), which in turn can be fitted (see Fitting Data).

[1] Given the size of the uncertainty, one would also report the result as $(3 \pm 9) \times 10^8$ m/s.

More advanced is plotting pairs of data points (y-values vs. x-values, times series, etc.) producing either a graph (ordered data) to see trends or a scatter plot (unordered data) to see correlations.

2.2 Estimators

While printing and plotting data is imperative, statistics is about quantification: We want to assign numerical values to data, that in simple terms describe the data beyond the (important) number of data points. For this, we use estimators. These are functions of the data, which yield a single value output (the estimate). The typical notation for an estimator is the quantity it is trying to estimate (e.g. μ) with a hat above (as in $\hat{\mu}$).

In order to describe the distribution that has given rise to a dataset, we first of all want to convey the typical value of x. This can be done in many different ways, most often based on the (arithmetic) mean defined as:

$$\hat{\mu}(x) = \frac{1}{N} \sum_i x_i \ . \tag{1}$$

While the formula for the mean is of course well known to all, it is worth noting that this formula arises from the principle of maximum likelihood (as most basic formulae do in statistics) and also while it is a great estimator, it is not perfect for all purposes, as it is not very robust to far outliers (imagine values around 0–1 with a single way-off value of 1000). There are many alternatives: Median, Mode, Geometric Mean, Harmonic Mean, and Truncated Mean. The median and truncated mean are both robust to outliers, while the others hold other virtues. Yet, the usual (arithmetic) remains the most commonly used.

In addition to a typical value, the next thing to consider would be the typical variation of the values, thus also giving an idea of the range of values. This is described by the variance V, or rather the square root of it, called the *Standard Deviation* (SD or Std.),[2] defined as:

$$\widehat{Std}(x)^2 = V(x) = \frac{1}{N} \sum_i (x_i - \mu)^2 \ . \tag{2}$$

Given a dataset, we don't know the true mean μ. Naively, we could calculate the sample Std. using the estimated mean from above instead of the true mean. However, then we have used some information (one degree of freedom) for estimating this mean, and hence one should correct the formula (known as Bessel's correction) for the Std. as follows:

$$\widehat{Std}(x)^2 = V(x) = \frac{1}{N-1} \sum_i (x_i - \mu)^2 \ . \tag{3}$$

This is an unbiased estimator of the Std., as can be proven mathematically or shown by example using a simple simulation (e.g. take the Std. of N=3 random unit Gaussian numbers many times, and consider the distribution mean of the Std. values when including the correction or not).

Given repeated measurements of the same quantity, the Std. describes the typical (Gaussian) uncertainty (denoted $\sigma(x)$) on *each measurement*. However, as the precision of the mean improves with the square

[2] The reason why both the variance and the square root of it (the Std.) has a specific name in statistics is likely, that the variance has theoretical importance and is used in many calculations, while the Std. has the same unit as the values we consider and is therefore often the number that we seek, understand easiest, and therefore quote.

root of the number of measurements, the uncertainty *on the estimated mean* ($\sigma(\hat{\mu}_x)$) is given as:

$$\sigma(\hat{\mu}_x) = \sigma(x)/\sqrt{N} \ . \tag{4}$$

Make absolutely sure that you understand the difference between the uncertainty on a single measurement (σ_x) and the uncertainty on an estimated mean ($\sigma(\hat{\mu}_x)$), as this is a very common (and grave) mistake to make!

EXAMPLE:

In 1797–98 Henry Cavendish famously made 29 independent measurements of the average density (and thus the total mass) of the Earth:
5.5, 5.61, 5.88, 5.07, 5.26, 5.55, 5.36, 5.29, 5.58, 5.65, 5.57, 5.53, 5.62, 5.29, 5.44, 5.34, 5.79, 5.1, 5.27, 5.39, 5.42, 5.47, 5.63, 5.34, 5.46, 5.3, 5.75, 5.68, and 5.85.
After having inspected the values and plotted them in a histogram (e.g. 10 bins in the range [5.0, 6.0], as the extreme values are 5.1 and 5.85), we calculate the mean to be 5.448 and the Std. to be 0.333. This in turn gives an uncertainty on the mean of $0.33/\sqrt{29} = 0.06$, and we would thus give the result as 5.448 ± 0.06g/cm^3.
Note: It took more than a century to improve on this great measurement, which was within 1% of the modern day value.

The Std. has many names, among others the Root Mean Square Error (RMSE, where the name contains the recipe for its calculation), and simply the width of a distribution. For a Gaussian distribution, which is given by:

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

the Std. coincides with the σ in the formula.
The uncertainty on the Std. (“the uncertainty on the uncertainty”) is rarely considered, as quantifying the uncertainty alone is typically sufficient. However, for adequate statistics, it can be estimated as:

$$\sigma(\text{Std.}) = \text{Std.}/\sqrt{2(n-1)} \tag{6}$$

The mean and width of a distribution can be seen as the 1st and 2nd (central) moment of data. Naturally, higher moments exists, named the Skewness (3rd) and Kurtosis (4th), defined as:

$$\begin{aligned} \text{Skewness} &= \sum_i (x_i - \mu)^3/\sigma^3 && (7) \\ \text{Kurtosis} &= \sum_i (x_i - \mu)^4/\sigma^4 - 3 && (8) \end{aligned}$$

The Skewness measures the asymmetry of a distribution, while the Kurtosis quantifies how long the tails of a distribution are. Both are 0 for the Gaussian distribution, which (of course) holds a special place in statistics.
Taking a step back, and thinking about estimators, these are meant to be good prediction for a true value

that we don't know. But what do we mean by "good"? This is defined by three criteria:

Consistency. The estimator must converge towards the true value for large statistics.

Asymptotic normality. The estimator should obtain a normal distribution around the true value for large statistics.

Efficiency. The estimator must have the minimal possible error, defined by the so-called Rao–Cramer-bound.

While estimators can provide key values that quantify distributions, this does not necessarily give any insight into the type and hence origin of the distributions at hand. This requires that we recognize certain typical Probability Distributions Functions (PDFs), which are the result of some (typically simple) underlying principles.

2.3 Covariance and correlation

When given two variables x and y, one can estimate the covariance between them V_{xy} as follows:

$$V_{xy} = \frac{1}{N-1} \sum_i (x_i - \mu_x)(y_i - \mu_y)\ . \tag{9}$$

The covariance quantifies to which degree x and y are *linearly* correlated. If y is high (above its mean) when x is high (above its mean) and similarly for low values, then the covariance obviously comes out positive. This indicates a positive correlation. Conversely, if y is high when x is low and vice versa, then V_{xy} becomes negative, indicating a negative correlation.

Extending Eq. (9) to include all combination of variables, this produces a matrix of variances, the so-called covariance matrix, which is symmetric, and where the diagonal carries the variances of each variable. The correlation matrix plays a central role in statistics. It encapsulates to what degree the different variables are related to one another, both for measurements (as Eq. (9) describes), but also for e.g. fit parameters.

While the numerical value of the covariance can be hard to interpret, the correlation coefficient ρ_{xy} defined below is much easier to understand, as it only takes values in the range [-1,1]:

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}\ . \tag{10}$$

It is important to notice that in the above only linear correlations can be determined. Non-linear correlations are not included. Thus, if $\rho \neq 0$ one can conclude that there is a correlation. However, if $\rho = 0$, then one can *not* be sure that there are no correlations. If $\rho = \pm 1$ then x and y are completely correlated.

Correlation can also be measured based on rank (i.e. position when sorted) rather than value. One example is Spearman's rank correlation, which follows the same formula as for ρ, but where the values (x,y) in the covariance formula have been substituted with their ranks ($R(x)$, $R(y)$), that is the (integer) position in a list sorted by value. The philosophy behind rank correlation is that it measures the degree of bijectivity, i.e. if there is a monotonic relation between values of x and y, even if it is not linear.

There exists more complex measures of correlation, which also includes non-linear relations through the amount of information that x and y share. Some of these measures are Distance Correlation, Correlation Ratio, and Mutual Information (based on the Kullback–Leibler divergence—beyond the scope of these notes).

3 Probability Density Functions

A Probability Density Function (PDF) is a function, $f(x)$, that describes the probabilities of specific outcomes. They can be discrete (throw of a die: $X \in [1, 2, 3, 4, 5, 6]$) or continuous (random number from computer: $x \in [0, 1]$). The value of a PDF at a specific point, x_0, can be interpreted as a relative likelihood for a random value from $f(x)$ to take on the value x_0. Thus, $f(x_0)dx$ is the probability of x falling in the infinitesimal interval $[x_0, x_0 + dx]$.

In order for a function to be a PDF, it must never produce negative values (no negative probabilities) and must have a unit integral, so that probability is conserved (i.e. the probability of any outcome is one): $f(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f(x)dx = 1$. For discrete PDF distributions these formulae can be considered as sums.
The mean and variance of a PDF are calculated much like their corresponding estimators (note: These are the true values, not estimates from a sample, and hence there are no hats):

$$\mu = \int_{-\infty}^{\infty} x f(x) dx \;, \tag{11}$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx \;. \tag{12}$$

Since PDFs "only" provides values proportional to probabilities, the integral of PDFs is often of interest. Specifically, the Cumulative Distribution Function (CDF) and Survival Function (SF) are defined as:

$$CDF_{F(x)} = \int_{-\infty}^{x} f(t) dt \;, \tag{13}$$

$$SF_{f(x)} = \int_{x}^{\infty} f(t) dt \;. \tag{14}$$

Obviously, CDF(x) + SF(x) = 1, and both are just a simple way of writing the specific integrals in a simple way. We will return to the SF, when we get to the ChiSquare method.

3.1 The Central Limit Theorem and the Gaussian distribution

Imagine that you add some random numbers and repeat the process (in the same way) many times. What distribution of sums should you expect? While both the question and the answer may seem a bit arbitrary, the answer and the implications are rather surprising: The distribution of sum resembles a Gaussian distribution! And that is why uncertainties tends to be Gaussian! More specifically, it can be proven that "the sum of N independent continuous random variables x_i with means μ_i and variances σ_i^2 becomes a Gaussian random variable with mean $\mu = \sum_i \mu_i$ and variance $\sigma^2 = \sum_i \sigma_i^2$ in the limit that N approaches infinity."
This is called the Central Limit Theorem (CLT), and holds a prominent place in probability theory, as it is much of the foundation behind our thinking and working with random variations, uncertainties, and the Gaussian as the "unit distribution" of statistics.

You might think, that adding random numbers of various unknown distributions is a special case, but it turns out, that most continuous measurements are exactly a result of such a process: You start with a specific "true" value (e.g. energy of a photon from a specific decay), which is perturbed through interactions before reaching your detector, state of your detector at the time of arrival, random processes in relation to the photon interactions in the detector, distortions from the read-out electronics, etc. to end at a value around the true value, but not quite. The CLT is the reason why you obtain a distribution resembling a Gaussian, when repeating a measurement many times. And if it is not Gaussian, but exhibits a clear structure, then typically some effect is at play, which you will want to find, understand, and correct for.

While the Gaussian distribution holds a central place in statistics, there are several other fundamental distributions, which are important.

3.2 Discrete PDFs

3.2.1 Binomial distribution

Consider a random process (e.g. roll of a die) with probability of success (e.g. rolling a 6) p repeated independently N times. The distribution of number of successes n (denoted $P(X = n)$) will then follow a Binomial distribution:

$$f(n, N, p) = P(X = n) = \left(\frac{N!}{n!(N-n)!}\right) p^n (1-p)^{N-n} \tag{15}$$

The formula is relatively straightforward to understand. The n successes have probability p^n while the $N - n$ failures have probability $(1-p)^{N-n}$ and the number of ways these can be obtained is the binomial coefficient, defined as:

$$\binom{N}{n} = \left(\frac{N!}{n!(N-n)!}\right) \tag{16}$$

If you sum all the terms in the Binomial distribution, you should see that you get 1 no matter the values of the two parameters p and N, so the PDF is normalised, as it should be.

The mean and variance of the Binomial are $\mu = Np$ and $\sigma^2 = Np(1-p)$, respectively. As the Binomial is often used in relation to fractions $f = n/N$ (e.g. efficiencies), the formula for the uncertainty on a fraction is useful to remember: $\sigma(f) = \sigma(n)/N = \sqrt{Np(1-p)}/N = \sqrt{p(1-p)/N}$.

The requirements for being binomially distributed are a fixed number of trials (N) that are independent and with only two outcomes of constant probability. Thus, e.g. the cards of a poker hand or rolling a die until a 6 appears are not binomially distributed. If there are more than two outcomes, the formula expands to become the multinomial distribution.

3.2.2 Poisson distribution

Sometimes, one does not know p and N separately for a process, but only the rate of success (whatever that defines), given as $\lambda = Np$, and then the Binomial distribution can not be used. But if N is large (which also makes the calculation of the binomial coefficients challenging) and p is small, then the Binomial distribution of number of successes approaches the Poisson distributed (by Stirling's formula):

$$f(n, \lambda) = P(X = n) = \frac{\lambda^n e^{-\lambda}}{n!} \tag{17}$$

Notice how the Poisson distribution only has one parameter, λ, and that both the mean and the variance is λ. The latter is of tremendous importance to remember (i.e. commit to memory now!), as this means that the uncertainty on a (Poisson) number is the square root of that number. You may think that being a Poisson case is rare, but in fact most numbers are:

Number of entries in a single bin in a histogram. This is why we assign bin uncertainties as the square root of the number of entries in a bin. It is of course an approximation that these uncertainties are Gaussian, but if $\lambda > 20$ then it is a very good approximation.

Number of people killed in traffic a year. Many venture into traffic many times in a year, and the probability each time is (fortunately) very small. The probability is not constant, but the sum of numbers from two Poisson distributions (λ_a and λ_b) is also Poisson distributed ($\lambda = \lambda_a + \lambda_b$).

Number of die hard left-wings voting for the conservatives. Just as another example, where p is very likely small, as is required.

In fact, both the Binomial and the Poisson distributions tend to take the shape of a (discrete) Gaussian distribution for large values of Np, as do many other continuous PDFs in other limits corresponding to high statistics.

EXAMPLE:

The square root uncertainty on a counting number has a significant and growing impact. Take the number of people killed in traffic a year, which in 2022 was 155 in Denmark. The statistical uncertainty is $\sqrt{155} = 12.4$ or about 8%. Thus, in trying to improve traffic safety, only effects larger than this can be seen with a single year's data.

Consider then the situation for commercial flights ending in deaths. Perhaps there are say four each year, that is with a 50% uncertainty! How will you ever know if new initiatives to improve aviation safety have any impact? Wait 50 years?

The answer was to change the legislation completely! By now, everyone is obligated to report near-misses, faulty equipment, etc., at the risk of getting fired if NOT reporting. There are 10000s of such cases, giving plenty of statistics to monitor developments in detail. Such changes in legislation is becoming more and more common (e.g. entering hospitals), all because of the Poisson distribution.

3.3 Continuous PDFs

3.3.1 Uniform

The uniform distribution is perhaps best known for being the random numbers that computers provide, but it can of course be made more general between a and b:

$$f(a,b) = \frac{1}{|b-a|} \text{ for } x \in [a,b], \text{ else } 0 \tag{18}$$

The unit uniform PDF has mean 0.5 and variance 1/12. You can calculate these values from Eqs. (12), which is a good exercise to check that you have understood the concept. Given that uniform numbers are the typical computer output, this is also those to be transformed into other distributions.

3.3.2 Exponential distribution

The exponential distribution is single parameter PDF given by:

$$f(\tau) = \frac{1}{\tau} \exp(-t/\tau) \;\; \text{for } x \in [0, \infty] \; . \tag{19}$$

The distribution is in particle physics typically used for lifetime measurements and background modelling.

3.3.3 ChiSquare distribution

The ChiSquare distribution is in fact a family of distributions defined by the number of degrees of freedom (N_{DOF} or ν). It is essentially the distribution obtained from adding ν unit Gaussian numbers squared.

$$f(\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} \exp(-x/2) \;\; \text{for } x \in [0, \infty] \; . \tag{20}$$

The distribution is interesting, since it reflects the expected χ^2 values from e.g. a comparison or fit with the corresponding N_{DOF}. This is central for testing goodness-of-fit in χ^2-fits and in hypothesis testing. Note that the mean value is ν.

3.3.4 Student's t distribution

The Student's t distribution is the curious result of beer brewing and industrial secrecy, and it is an example of Stigler's Law of Eponymy. Like the ChiSquare distribution, it only has ν as a parameter:

$$f(\nu) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)} \left(1 + t^2/\nu\right)^{-(\nu+1)/2} \;\; \text{for } t \in [-\infty, \infty] \; . \tag{21}$$

The distribution is a generalisation of the Gaussian distribution, which it approximates for large values of ν, while it is the "infinite tailed" Cauchy distribution for $\nu = 1$. It is used when doing hypothesis testing with small samples, where the mean and variance are not well known but estimated. Once the statistics gets significant (typically above 10), the Gaussian distribution becomes a good approximation for the hypothesis testing.

3.3.5 Other distributions

There is a wealth of different PDF distributions, many of which are used for very special situations. For this reason note that polynomials are typically not included as PDFs, as these can take on negative values and are usually not normalised. These deficiencies can be rectified considering special classes of polynomials, but natural phenomena rarely follow (higher order) polynomial distributions. Polynomials are therefore rarely used for modelling histogram data, but rather for fitting points with uncertainties. Of course, occasionally one resorts to using a polynomial modelling of the background distribution, especially when the shape is complicated, statistics is high, and the range of interest is finite (e.g. $H \to \gamma\gamma$).

3.4 The philosophy of PDFs

While the basic PDFs often serve well as the building blocks for models, they certainly have their limitations, famously summarised as:

> “Essentially, all models are wrong, but some are useful”
>
> [George E. P. Box, British Statistician 1919–2013]

The point is, that real data is never ideal and hardly ever follows these simple PDFs perfectly. Especially, when considering a lot of data, models often tend to have a hard time mimicking the data perfectly, as the high statistics and corresponding small uncertainties makes all the minuscule variations and unknown effects stand out.

This is expected, and the reader should ensure awareness and understanding of the imperfect modelling, rather than introducing new ad hoc fitting parameters. No need to chase mice, when there are tigers around!

4 Fitting data

Given data, we typically want to extract information from it. Going beyond the estimators introduced in section 2.2, this would typically consist of fitting the data with a functional form.

Data typically consist of points (x, y) with an uncertainty on y, $\sigma(y)$ (and potentially also on x, $\sigma(x)$), where the uncertainty is considered Gaussian. Alternatively, the data is a series of measurements, which can be put into a histogram, typically in 1 dimension (1D), but potentially in 2D or even 3D. We will in the following only consider fitting in 1D, though the arguments and methods extend to multiple dimensions. However, be warned that the complexity of fits grows fast with increasing dimensionality, due to possible correlations, large number of fit parameters, and an explosion in the number of bins.

Testing to what extent the fit model—and more generally any model, from fitting or not—actually matches the data is a core theme in statistics. For now, we will start by considering how to perform a fit, and how to extract results and uncertainties from it. Towards the end, we will start asking to what extent the model is reasonable, naturally leading to the next chapter on hypothesis testing.

4.1 Principle of Maximum Likelihood

Fitting data means obtaining the fit parameter values such that the function “best matches” the data. Obviously, we need to define what we mean by “best matches”. Here the Principle of Maximum Likelihood (PML) comes into play. Essentially it states that “best matches” is the model that has the highest likelihood, where likelihood for a dataset (measurements x_i) is defined as a function of the fit parameters as follows:

$$\mathcal{L}(\theta) = \prod_i PDF(x_i, \theta) \ . \tag{22}$$

In all its simplicity, it says that given data and a PDF one should choose the parameters (θ) of the PDF such that the likelihood ($\mathcal{L}$) is maximal. What is astounding is the variety of results that can be deduced from this principle (e.g. estimators in previous chapter).

Before applying this principle in more detail, let us just have a look at why it works at all. Imagine for example, that we have a series of repeated measurements x_i and are fitting these with a Gaussian distribution. If the mean was (far) off, then the value of the PDF would be rather small at each of the measurement values x_i and so would the resulting product and hence value of the likelihood $\mathcal{L}$. Thus, the mean should match the mean of the data to yield a higher likelihood value.

Likewise, if the Gaussian PDF is too narrow, it will of course yield high likelihood values for the central points, but those on the tail will have very small values, leading to a lower overall likelihood value $\mathcal{L}$. Conversely, if the Gaussian is too wide, some of the PDF will have significant values where there are no measurement points, again leading to a lower value of $\mathcal{L}$. Similarly for the normalisation.

Thus, the result is that the principle of maximum likelihood "forces" the PDF parameters to take on the values that best match the data. As it turns out, this definition of "best" is at the core of statistics, and the principle of maximum likelihood rules supremely. All the formulae for means, standard deviation, ChiSquare test (below), etc. can be derived from this principle.

In practice, it turns out that multiplying many numbers tends to lead to numerical problems, because the product either becomes very large or very small. For this reason, it is customary to take the logarithm of the likelihood (making the product a sum), and also multiplying by -2 to make it a minimisation problem. We'll get back to this factor two in a moment.

4.2 The ChiSquare method

The PDF of a measurement with uncertainties is always assumed Gaussian (if nothing else is stated). Thus, if we are fitting a series of measurements, that is x_i, y_i, and $\sigma(y_i)$, with a function $f(x,\theta)$, then ($-2\ln$ of) the likelihood would be:

$$\begin{aligned} -2\ln(\mathcal{L}(x,\theta)) &= -2\sum_i \ln\left(\frac{1}{\sqrt{2\pi}\sigma(y_i)}\exp\left(-\frac{1}{2}\left(\frac{y_i - f(x_i,\theta)}{\sigma(y_i)}\right)^2\right)\right) \\ &= -2\sum_i \ln\left(\frac{1}{\sqrt{2\pi}\sigma(y_i)}\right) - 2\sum_i -\frac{1}{2}\left(\frac{y_i - f(x_i,\theta)}{\sigma(y_i)}\right)^2 \\ &= C + \sum_i \left(\frac{y_i - f(x_i,\theta)}{\sigma(y_i)}\right)^2 . \end{aligned} \tag{23}$$

Now we want to minimise this with respect to the fit parameters θ (e.g. a and b, if fitting a line), which is why the first term can be considered a constant. It does not depend on the fit parameters θ, only the data points.

At this point, I hope the reader recognise the formula for the ChiSquare, which also explains the factor 2 in front of the negative log likelihood. When uncertainties are Gaussian (as they often are), minimising the likelihood is equivalent to minimising the ChiSquare. However, the ChiSquare comes with an advantage, namely a goodness-of-fit measure. That is, from the ChiSquare value (at the minimum) one can evaluate if the fit matches the data well. Sure, from minimising θ these are the parameters for which the fit function best fits the data, but that in itself does not tell you, if "best" is also good. It can be horrible. For the same reason, always inspect a fit visually, if you can, as your eyes are sharp at fitting and evaluating a fit ("Chi-by-eye" is the semi-technical term).

4.2.1 ChiSquare goodness-of-fit measure

So how to evaluate the ChiSquare value from a ChiSquare fit? Well, if —NOTE "IF"—your data is good and has correct Gaussian uncertainties, and if the model you fit with actually correctly describes the data (the ChiSquare assumptions), then each data point should contribute with a unit Gaussian number squared. And the distribution of such a sum is known: It is of course the ChiSquare distribution. The number of degrees of freedom ν for the ChiSquare distribution is the number of data points fitted, but adjusted for the number of fitting parameters:[3]

$$N_{\text{DoF}} = N_{\text{Data Points}} - N_{\text{Fit Parameters}} \ . \tag{24}$$

Note that even if there is no fit (i.e. when the model is given/fixed for example when comparing two histograms), one can still calculate the ChiSquare value, simply setting $N_{\text{Fit Parameters}} = 0$.
With the knowledge of how the ChiSquare value *should* distribute itself for a given situation (i.e. N_{DoF}), one can calculate **the probability of obtaining a certain ChiSquare value or worse** (the p value) from the integral of the ChiSquare distribution from the ChiSquare value obtained and to infinity (i.e. the survival function). If the assumptions are fulfilled, then these p-values will distribute themselves as a uniform distribution. This means, that when things are as they should be, all p-values are equally likely. However, when these assumptions are *not* fulfilled, in particular when the data does not follow the fit model, this increases the ChiSquare value, and in turn lowers the p-values obtained. The more statistically powerful the data (typically by high statistics and sharp observables), the more it diminishes. At some point it gets so low, that one can simply not uphold the hypothesis that the model matches the data.
The p-value of a ChiSquare probability can roughly be interpreted as follows:

If $0.01 < \textbf{Prob}(\chi^2, N_{\textbf{DoF}}) < 0.99$, then the fit is typically good. Where exactly to draw the limits can vary, but remember that if you do say 20 (independent) fits, where everything is perfectly in place (i.e. the conditions are fulfilled), then there should still on average be one p-value below 5%.

If $\textbf{Prob}(\chi^2, N_{\textbf{DoF}}) < 0.01$, then you have either been very unlucky (1:100) or (more likely) something is wrong. Assuming your data and uncertainties are perfect, then the model is unlikely.

If $0.99 < \textbf{Prob}(\chi^2, N_{\textbf{DoF}})$, then your fit is *too* good. The typical causes are overestimated uncertainties on the data or correlations between the points.

The calculation of a p-value assumes that the data is trustworthy, the uncertainties correct and Gaussian, and that the model matches the data (the "if" above).

> "It is however, more than doubtful whether the conditions for the theoretical validity of the method are realised in statistical practice, and the student would do well to regard the method as recommended chiefly by its comparative simplicity and by the fact that it has stood the test of experience."
>
> [G.U. Yule and M.G. Kendall 1958]

Having stood more than an additional half century, this is a testament to the robustness of the ChiSquare.

[3] This is related to the fact that a polynomial with N parameters can be made to go through N points.

EXAMPLE:

Consider a linear fit (2 parameters) to 9 data points, minimized to yield a ChiSquare value of 9.1. The number of degrees of freedom is $9 - 2 = 7$ and the calculated p-value (the probability of obtaining this χ^2 value or greater given the conditions) is Prob$(\chi^2 = 9.1, N_{\text{DoF}} = 7) = 0.246$. The conclusion is in this case is that the linear model fits the data well.

However, if the ChiSquare value had been 19.1, the p-value obtained would have been Prob$(\chi^2 = 19.1, N_{\text{DoF}} = 7) = 0.0079$, indicating that the model is unlikely to be matching the data. If the ChiSquare value instead had been 1.1, the p-value obtained would have been Prob$(\chi^2 = 1.1, N_{\text{DoF}} = 7) = 0.993$, pointing to the uncertainties being too large or a correlation between the points "damping" the expected statistical fluctuations.

Note that the p-value has a direct correspondence with "number of σs", in that a p-value of 0.0027 corresponds to a (two-sided) "3σ" observation.

4.2.2 Versions of the ChiSquare calculation

Several versions of the ChiSquare exists, mainly differing by how the uncertainty is calculated: Based on the Expected (E) or the Observed (O) number of events:

$$\sum_i \frac{(O_i - E_i)^2}{E_i} \qquad \text{vs.} \qquad \sum_i \frac{(O_i - E_i)^2}{O_i} \,. \tag{25}$$

In both cases, the uncertainty is naturally taken to be the square root of the number of entries. In the first (Pearson) case all bins typically have non-zero values, and the question is "how many bins to include?". In the second case (implemented in Minuit) the ChiSquare sum can only be taken over the non-zero entries. This means that the fit doesn't "feel" the bins with zero entries, even if these contain valuable information (e.g. tails of distributions).

However, since the ChiSquare should only be used, when there is significant statistics (so that the uncertainties are Gaussian), the difference between these two approaches is rarely significant.

4.3 Binned likelihood fit

In addition to the unbinned likelihood fit, a binned version exists, building on the likelihood principle. Given a histogram where the number of observed (expected) entries are O_i (E_i), the binned likelihood is:

$$\mathcal{L}(\theta) = \prod_i \text{Poisson}(O_i, E(\theta)_i) = \prod_i \frac{E(\theta)_i^{O_i} e^{-E(\theta)_i}}{O_i!} \tag{26}$$

where the second equation is obtained simply by inserting into the Poisson distribution Eq. (17), which is the expected distribution for each bin. The advantage of an unbinned likelihood fit over a ChiSquare fit is, that the unbinned likelihood fit also works for low statistics. The downside is, that there is no simple goodness-of-fit test.

4.4 Which fitting method to use?

The unbinned likelihood fit is alluring, as it is the "best possible" in terms of requiring the least assumptions (e.g. Gaussian errors for the ChiSquare fit) and producing the results with the smallest uncertainties. But in terms of convergence, robustness, and not the least goodness-of-fit measure, the ChiSquare fit fares best.

The choice should thus reflect the situation. *If statistics is high* (e.g. all bins above 10 entries), and the Gaussian approximation of the uncertainties is thus valid, then the ChiSquare is recommendable due to its direct goodness-of-fit measure. *If statistics is low*, then the unbinned likelihood fit is probably the best choice.

In particle physics, the binned likelihood fits have been preferred as part of more complex fits across many samples and dimensions (with e.g. RooFit or HistFactory within). Since some channels are likely to have little statistics in some bins, the likelihood is chosen.

4.5 A Goodness-of-fit measure for likelihood fits

While likelihood values may take on essentially any range (also negative), there is no inherent way of testing if a likelihood fit actually yields a model that fits data well. However, it can be done, using simulation.

Imagine having performed a likelihood fit, obtaining the likelihood value $\mathcal{L}_{\text{fit}}$ and fit parameters $\hat{\theta}$. If the fit model reflects the data, then producing a new similar dataset based on the fit parameters should yield a new (simulated) dataset, for which the likelihood value should be similar. Repeating the simulation, fitting, and recording of the resulting likelihood value thus produces the expected distribution of the likelihood values, against which the likelihood value obtained from the fit to the true data $\mathcal{L}_{\text{fit}}$ can be compared.

5 Hypothesis Testing

Suppose in a beer tasting, that someone gets 9 out of 10 right. Does that prove that the person can taste the differences between beers? The slightly surprising answer is "No".

What we can say is that the result is inconsistent (at some significance level) with the hypothesis that the person chooses at random. This leaves us with the alternative hypotheses, that the person can taste the differences or has cheated (consciously or unconsciously).

In statistics one can never prove a hypothesis directly. However, one can set up alternative hypotheses and disprove these. That is how one works in statistics...

5.1 Nomenclature in Hypothesis Testing

Hypothesis testing is like a criminal trial. The basic "null" hypothesis is Innocent (denoted H_0) and this is the hypothesis we want to test, compared to an "alternative" hypothesis, Guilty (denoted H_1). Innocence ("negative") is initially assumed, and this hypothesis is only rejected, if enough evidence proves otherwise, i.e. that the probability of innocence is very small ("beyond reasonable doubt").

Given two possible truths (innocent or guilty) and two possible verdicts (acquittal or conviction), there are four outcomes: True positive (TP, guilty and convicted), false positive (FP, innocent but convicted), false negative (FN, guilty but acquitted), and true negative (TN, innocence and acquitted). Given these four numbers, one can determine two rates:

False Positive Rate (FPR) is defined as $FPR = FP/(FP + TN)$ is the probability of rejecting H_0, when it is true (e.g. the rate of convicting the innocents).

False Negative Rate (FNR) is defined as $FNR = FN/(FN+TP)$ is the probability of accepting H_0, when it is false (e.g. the rate of acquitting the guilty).

For a given test, these two rates are inter-dependent. Take for example electron identification: If you lower the threshold for being an electron H_0, then you will also lower the FPR (rejecting a true electron), but you will at the same time increase the FNR (accepting a non-electron). Similarly, TPR and TNR can be defined.

The purpose of a hypothesis test is to yield (calculable/predictable) distributions of a test statistic t for the Null (H_0) and Alternative (H_1) hypotheses, which are as separated from each other as possible (in order to minimise FPR and FNR). The way to determine the separation is to plot the TPR against the FPR as a function of the possible selection criteria. This produces the Receiver Operating Characteristic (ROC) curve, which incorporates essentially everything about a given test.

Take again the court example. If we required impossible amounts of evidence, then no one would be convicted, and both the FPR (the innocent) and the TPR (the guilty) would be 0. Conversely, if we required no evidence at all, then everyone would be convicted, and the FPR and TPR would be 1. However, somewhere in between these extremes lies the power of the courts: Obtaining a high TPR while maintaining a low FPR.

Note that there can be several hypothesis, and that hypothesis testing may exclude any number of hypothesis ranging from none to all! Testing multiple hypothesis simultaneously is more complex.

5.2 Steps in Hypothesis Testing

Consider a case for which you want to do a hypothesis testing, and state a null hypothesis along with an alternative hypothesis. Think about the statistical assumptions you are making (independence, distributions, etc.), and then decide for an appropriate statistical test. Then define the relevant test statistic t, which could be anything from a counting number to a machine learning output. Next, derive/calculate the test statistic distribution under null and alternative hypothesis. In standard cases, these are well known distributions (Poisson, Gaussian, Student's t, etc.).

Before you do the actual test, you should select a significance level (α) that is a probability threshold below which the null hypothesis will be rejected. In particle physics we typically use 0.0027 (3σ) for "evidence", 0.000063 (4σ) for "observation", and 0.00000057 (5σ) for "discovery". Other sciences tend to use slightly lower values of α (e.g. 0.05 (2σ) in biology and medicine). There is no universally accepted value, as there should not be as "the weight of evidence for an extraordinary claim must be proportioned to its strangeness" [Pierre-Simon Laplace]. For an excellent discussion of this related to particle physics see Louis Lyons, "Discovering the significance of 5σ".

Once all the pieces above are in place, compute from (otherwise blinded) observations/data the value of the test statistic t, and use this to calculate the probability of observation under null hypothesis (p-value). Reject the null hypothesis for the alternative if the p-value is below the significance level.

5.3 The Neyman–Pearson lemma and Wilk's theorem

While a single likelihood value says little, the ratio of likelihood values (after minimisation) between two competing hypothesis (the null $\mathcal{L}_0$ and the alternative $\mathcal{L}_1$) plays a central role in hypothesis testing. If the two hypothesis are simple (i.e. have no free parameters) then the Neyman–Pearson Lemma (loosely) states that the $(-2\ln)$ likelihood ratio D defined as:

$$D = -2\ln\left(\frac{\mathcal{L}_0}{\mathcal{L}_1}\right) = -2\ln\mathcal{L}_0 + 2\ln\mathcal{L}_1 \tag{27}$$

is the best possible test statistic that exists. Now, that is a strong statement, which is why likelihood ratios are often used. An example use case is the determination of the Higgs particle spin, where the null hypothesis was 0, while the alternative hypothesis was 2 (for theoretical reasons spin 1 is not (well, hardly) possible). In this case none of the hypothesis have any free parameters, and the lemma applies. The challenge lies in determining the expected distributions of D for the two hypothesis.

Most often though, the hypotheses and associated likelihoods have free parameters. If the null hypothesis is nested in the alternative hypothesis (i.e. that the alternative hypothesis contains the null hypothesis), then the very nice Wilk's Theorem states that in the limit of large statistics the likelihood ratio D approximately follows a Chi-Square distribution with $N_{\text{DoF}} = N_{\text{DoF}}() - N_{\text{DoF}}()$, if H_0 is true.

The reason for the nested requirement is that it ensures that the alternative hypothesis will always have a higher likelihood value than the null, and hence D will always be positive, as it should be to reflect a ChiSquare value. If uncertainties are Gaussian (e.g. high statistics histograms) then Wilk's theorem extends to include Chi-Square differences also.

EXAMPLE:

Imagine that you are fitting a (high statistics) data sample which is an exponentially falling spectrum of background events with a potential Gaussian peak in the middle. In order to test the null hypothesis H_0 that there is only background (2 parameter fit) against the alternative hypothesis H_1 that is a Guassian peak in addition (5 parameter fit), you fit the data for each case.

The resulting fits (without and with a Gaussian peak component) yields likelihood values of $-2\ln\mathcal{L}_0 = -10017$ and $-2\ln\mathcal{L}_1 = -10000$. The likelihood values in themselves do not reveal much, but their difference yields $D = 17$ for $5 - 2 = 3$ degrees of freedom, which results in a p-value of 0.00071 (approximately 3.2σ for a one-sided test).

You thus conclude, that you have found evidence for a peak.

5.4 Various hypothesis tests

Many different types of statistical tests exists. Below a few of the more frequently used ones are listed.

5.4.1 One-sample test

Used when comparing e.g. the mean of a sample to a known value. Example: Comparing mean of measurements to known constant: $\mu_{\text{exp}} = 2.91 \pm 0.02$ vs. $c = 2.99$, which is thus a 4σ difference.

5.4.2 Two-sample test

Used when comparing e.g. the means of two samples. If the samples are high statistics (or the sample σs are known), then the test works much like the one-sample test. Example: Comparing a sample to control sample: $\mu_{\text{exp}} = 4.01 \pm 0.12$ vs. $\mu_{\text{control}} = 3.88 \pm 0.05$, which is thus a $(4.01 - 3.88)/\sqrt{0.12^2 + 0.05^2} = 1.0\sigma$ difference.
In case there is a pairing between the two samples (e.g. twins), this may reduce the test to a paired test, which is a one-sample test, the advantage being that the pairing will cancel out much of the variance from nuisance parameters (e.g. genetic biases in the case of twins).

5.4.3 Chi-squared test

This test evaluates the adequacy of a model compared to data (or between two datasets) through the χ^2 value. Example: Model fitted to (possibly binned) data, yielding p-value = Prob($\chi^2 = 45.9, N_{\text{DoF}} = 36) = 0.125$.
More generally, this test can be used for determining parameters of and over-determined system of equations (i.e. with more equations than unknowns), which is a powerful way of solving many averaging, calibration, and combinatorial challenges.

5.4.4 Wald–Wolfowitz runs test

The WW runs test is a binary check for independence between entries in a series of values. Imagine a fit, where you want to check if the N data points lie randomly above and below the fit or rather in collected groups (islands). This can be tested, since the number of expected islands $\mu(N_{\text{runs}})$ and its variation $\sigma(N_{\text{runs}})$ can be determined from the number of entries passing some binary requirement (e.g. above or below fit) N_+ and N_-:

$$\begin{aligned}
\mu(N_{\text{runs}}) &= 1 + \frac{2N_+N_-}{N}\,, &\quad (28)\\
\sigma(N_{\text{runs}}) &= \sqrt{\frac{2N_+N_-(2N_+N_- - N)}{N^2(N-1)}}\,. &\quad (29)
\end{aligned}$$

In a sense, the WW runs test is complimentary to the ChiSquare test in that it considers the sign rather than the size of a deviations. As such, it serves well to check a fit, where the result despite a good ChiSquare p-value (possibly due to overestimated systematic uncertainties) does not seem to follow the trends of the data well. Note that $\sigma(N_{\text{runs}})$ is only approximately Gaussian, when N_+ and N_- are greater than about 10.
Example: You fit $N = 25$ data points, and the resulting distribution of data points above/below the fit is $+++++-------++++++-------$++. Thus $N_+ = 13$ and $N_- = 12$, which yields $\mu = 13.5 \pm 2.4$ islands. You observe $\mu(N_{\text{runs}}) = 5$, which is suspiciously ($(13.5 - 5)/2.4 = 3.5\sigma$) low.

5.4.5 Kolmogorov–Smirnov test

This test compares the degree to which two 1D distributions are compatible, i.e. a p-value for the distributions being samples of the same underlying PDF. Example: Compatibility between data and MC sample, yielding p-value = 0.87 (thus the distributions are compatible). The Kolmogorov–Smirnov test does not make any assumption about distributions, which makes it very powerful.

6 Short note on setting limits

If you observe 0 events of type X in an experiment, then where to set the limit? The way you should think about it is as follows: The number of X-events is Poisson distributed, because you probably made many attempts (N) to produce and observe X-events, but the probability of doing so (p) was rather low. If this is the case, then how large could the value of $\lambda = Np$ be, such that it is still consistent to have say 5% chance of observing N = 0?

The answer is obtained from increasing λ until the observed is Poisson$(N_{\text{obs}} = 0, \lambda = N_{\text{obs}} = 2.996) = 0.05$. Thus the 95% Upper Limit is (rounded to be) 3.0. In case you wanted the 90% Upper Limit, the answer would be 2.30.

7 Final words of advice

Statistics and the general understanding of numbers and their relations play a central role in particle physics. Embrace it! Doing so makes you a member of the much sought-after class of persons, who are hyper literate in numbers. This skill will take you far, not just in particle physics but in most endeavours encompassing numericals.

At times you will of course encounter difficult situations, where you are unsure of what the correct statistical approach is. You can of course test this numerically with simulations for closure tests, but the doubt might remain. This happens to all experts in all fields, the author included. In that case, do not be too troubled, but rather lean on the wise words of a mentor and fellow statistician:

> “Don’t worry too much about statistics. Just tell us what you do and do what you tell us.”
> [Roger Barlow, ICHEP conference 2006 in Moscow]