Introduction to Statistical Data Analysis in Particle Physics

Sino-Thai International Summer School on HEP and Astrophysics in 2026

July 19, 2026

Contents

1 Preface

Particle physics, also known as high-energy physics, is the discipline that studies the most fundamental constituents of matter—elementary particles—their properties, and their interactions. Its theoretical framework is built upon quantum mechanics and quantum field theory. Because the quantum behavior of elementary particles is intrinsically random and their interaction processes obey probabilistic laws, the analysis of experimental data must rely on rigorous statistical methods. **Statistics plays an irreplaceable core role in particle physics experiments.**

Particle physics experiments, according to their respective physics goals, can be broadly divided into two categories: **search/test-type** and **measurement-type**. The objective of a search/test-type experiment is to look for new particles or new interactions not yet confirmed to exist in the experimental sample, or to test whether a certain theory or hypothesis is supported by experiment. A measurement-type experiment, on the other hand, aims to extract the properties of particles or the parameters of interactions from the experimental sample, under the premise that a certain particle or interaction has already been confirmed to exist.

These two types of experiments roughly correspond to two major themes in statistics: **hypothesis testing** and **parameter estimation**. The goal of hypothesis testing is to examine whether a data sample provides sufficient support for a specific hypothesis (testing whether the data are consistent with a given theory). For example, in the experiment that searched for the Higgs boson (Fig. 1), the data sample consisted of particle events reconstructed from detector signals, and the hypotheses to be tested were "the Higgs boson exists" or "the Higgs boson does not exist." The goal of parameter estimation is to estimate the values of probability distribution parameters from a data sample (using data to estimate the numerical values of free parameters in a theoretical model). For example, in a particle lifetime measurement, the data sample consists of the decay times of particles, the probability distribution of the parameter being estimated can be an exponential distribution, and the parameter to be estimated is the particle's lifetime.

The goal of this course note is not to present a comprehensive account of all important knowledge about statistical analysis of particle physics experimental data; rather, it is to help students understand the fundamental applications of statistics in particle physics, grasp several important statistical methods and the core ideas behind them, fully appreciate the rigor required in the statistical analysis of experimental data, and, on that basis, eventually be able to carry out rigorous experimental data analysis themselves. This note will first introduce some background knowledge, and then build upon it to present maximum likelihood estimation and likelihood ratio tests, together with their applications in particle physics experiments. Along the way, the note will interweave an introduction to the basic usage of ROOT Brun et al. (1997), the most widely used data analysis framework in high-energy physics experiments, and provide some simple data analysis examples for students to study and explore. The note is aimed at senior undergraduate students who have a preliminary understanding of the research goals of particle physics and are familiar with the fundamental concepts of probability theory and mathematical statistics. Running the data analysis scripts provided in this note requires that students have already properly installed the ROOT software.

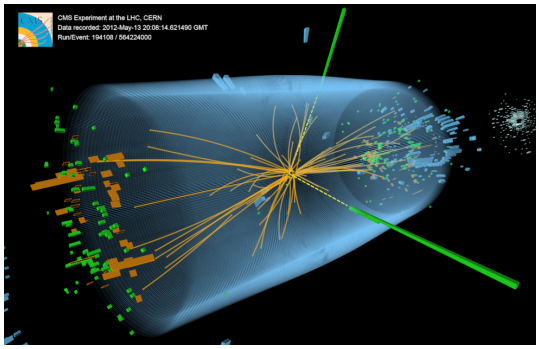
2 Preliminary Knowledge and a Few Probability Distributions

2.1 Particle Decay

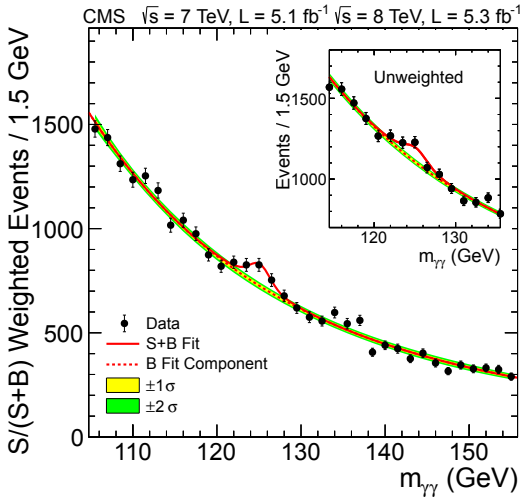
The vast majority of elementary particles[1] or composite particles[2] that constitute the material world of our daily lives are stable, such as protons, electrons, carbon atoms, etc. However, there also exist many unstable particles that undergo decay, transforming into other stable or unstable particles, such as muons, neutrons, cesium-137 nuclei, the Higgs boson, etc. Taking the muon as an example, the **muon** is an elementary particle belonging to the charged lepton family.

[1] Elementary particles are the most fundamental units of matter, and it cannot be confirmed whether they are composed of other, even more fundamental particles. Examples include electrons, quarks, photons, etc.

[2] Composite particles are microscopic particles formed by the combination of elementary particles. Examples include protons, atoms, etc.



(a) A Higgs boson decay event in CMS Anon (2025).



(b) The diphoton invariant mass spectrum measured by CMS. The bump on the fitted red curve is the invariant mass peak at 125 GeV discovered experimentally, corresponding to the decay of the Higgs boson.

Figure 1: In 2012, the CMS experiment at the Large Hadron Collider (LHC) at CERN discovered the Higgs boson, thereby completing the experimental observation of all particles in the Standard Model Chatrchyan et al. (2012). This is a classic search-type experiment. The other LHC experiment, ATLAS, also independently discovered the Higgs boson Aad et al. (2012).

The muon has a mass of 105.6583755(23) MeV/c^2 and a lifetime of 2.1969811(22) μs Navas et al. (2024). The muon decays into an electron, an electron antineutrino, and a muon neutrino ($\mu^- \to e^- \bar{\nu}_e \nu_\mu$).

2.1.1 Statistical Law of Particle Decay

At the high-school level we have learned that the decay of unstable particles follows an exponential law. For N_0 unstable particles, the expected average number surviving evolves with time as

$$\langle N(t) \rangle = N_0 e^{-t/\tau} \ , \quad t \geq 0 \ . \tag{1}$$

We define the particle's decay time as the proper time[3] elapsed from a given moment until the particle decays. It can be shown that this law implies that the decay time of a particle obeys the **exponential distribution**

$$\boxed{f(t) = \frac{1}{\tau} e^{-t/\tau} \ , \quad t \geq 0 \ .} \tag{2}$$

where τ is the lifetime of the particle. We commonly use the **lifetime** τ or the **half-life** $T_{1/2}$ to characterize the degree of stability of a particle[4]. For a collection of unstable particles of the same species, after each lifetime has elapsed the average number of particles decreases by a factor of $1/e$; after each half-life, the average number decreases by a factor of one half. The longer the particle's lifetime, the more stable the particle, and vice versa. It is straightforward to show that **the particle's lifetime is also the expectation of the particle's decay time**:

$$\langle t \rangle = \int_0^{+\infty} t f(t) = \int_0^{+\infty} \frac{t}{\tau} e^{-t/\tau} = \tau \ . \tag{3}$$

[3]Proper time is the time interval in the rest frame of the particle. If the reference frame is not the particle's rest frame and the relative velocity between the particle and the reference frame is high, then due to relativistic time dilation the observed lifetime of the particle in that frame will be longer.

[4]In particle physics, lifetime is more commonly used than half-life.

2.1.2 Particle Decay in Quantum Mechanics

Let us revisit the topic of particle decay from the perspective of quantum theory. In quantum mechanics, we can obtain the wave function of an unstable particle by introducing an imaginary part into the Hamiltonian. First, we review the case of a stable particle. For a stable particle at rest, its Hamiltonian is its total energy, i.e., the rest energy of the particle:

$$H_{\text{stable}} = mc^2 \ . \tag{4}$$

where m is the mass of the particle and c is the speed of light. In particle physics, we often adopt natural units ($c = \hbar = 1$). In natural units, the Schrödinger equation for a stable particle is

$$i\frac{\partial \phi}{\partial t} = m\phi \ . \tag{5}$$

Solving the equation shows that the wave function of a stable particle is

$$\phi(t) = \frac{1}{(2\pi)^{3/2}} e^{-imt} \ . \tag{6}$$

It is easy to see that its probability amplitude $|\phi(t)|^2$ is a constant, time-independent, and thus manifests as a stable particle.

For an unstable particle, its Hamiltonian can be written as

$$H_{\text{unstable}} = m - i\frac{\Gamma}{2} \ . \tag{7}$$

Compared to the stable particle, there is an additional imaginary part. The parameter Γ here is called the **decay width** of the particle, or simply the **width**, which has dimensions of energy. Hence, the Schrödinger equation for an unstable particle can be written as

$$i\frac{\partial \psi}{\partial t} = \left(m - i\frac{\Gamma}{2}\right)\psi \ . \tag{8}$$

The wave function of the unstable particle is then

$$\psi(t) = \frac{1}{(2\pi)^{3/2}} e^{-i(m - i\Gamma/2)t} = \frac{1}{(2\pi)^{3/2}} e^{-imt} e^{-\Gamma t/2} \ . \tag{9}$$

We can see that compared to the stable particle, the wave function contains an additional exponentially decaying factor, and its probability amplitude $|\psi(t)|^2 \propto e^{-\Gamma t}$ decays exponentially. Thus it manifests as an unstable particle. Comparing this with the decay-time distribution of unstable particles (Eq. (2)), we obtain

$$\begin{aligned} \Gamma &= \frac{1}{\tau} \quad \text{(natural units)}, \\ \Gamma &= \frac{\hbar}{\tau} \quad \text{(SI units)}. \end{aligned} \tag{10}$$

Here, $\hbar$ is the reduced Planck constant.[5] It follows that the newly introduced particle width Γ and the particle lifetime τ are intimately related: they are inversely proportional to each other. The more stable the particle, the longer its lifetime and the narrower its width; the more unstable the particle, the shorter its lifetime and the broader its width.

The reader may wonder: since the lifetime and the width are directly related, why is the width needed? What physical significance does it carry? A rigorous theoretical answer to this question is somewhat involved, but a key insight worth revealing is that **the mass of an unstable particle does not have a definite value; rather, it obeys a specific random distribution.** This distribution is called the **relativistic Breit–Wigner distribution**, or

[5] Note that the reduced Planck constant $\hbar$ has dimensions of [time] × [energy], while the width has dimensions of energy; hence the two formulas differ by only a factor of $\hbar$ when converting between unit systems.

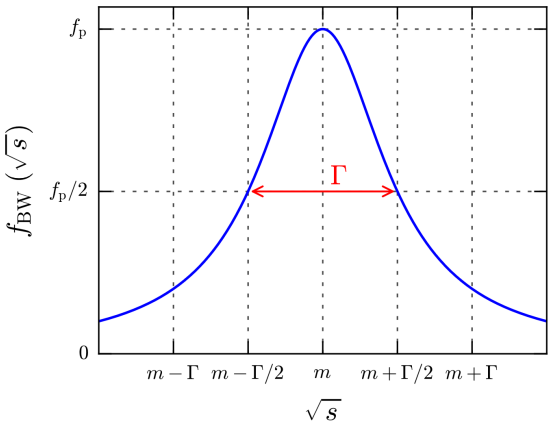


Figure 2: Distribution of the mass of an unstable particle Yu (2025). This distribution is the relativistic Breit–Wigner distribution.

simply the **Breit–Wigner distribution**.[6] The distribution takes the form

$$\boxed{f\left(\sqrt{s}\right) = \frac{\Gamma}{2\pi} \frac{1}{\left(\sqrt{s}-m\right)^2 + \Gamma^2/4} \,.} \tag{11}$$

As stated above, **the physical mass m of the unstable particle is the central value of this distribution, while Γ is its full width at half maximum (FWHM).** In the expression, $\sqrt{s}$ is the invariant mass (i.e., the center-of-mass energy) of the unstable particle.

We can gain a more intuitive understanding of the origin of the particle mass broadening through the energy–time uncertainty relation in quantum mechanics,

$$\Delta E \Delta t \geq \frac{\hbar}{2} \,, \tag{12}$$

We already know that the survival time of an unstable particle has a finite uncertainty $\Delta t = \tau$, so according to the energy–time uncertainty relation above, the total energy (mass) of the particle will also have a finite uncertainty $\Delta E \simeq \hbar/2\tau = \Gamma/2$. Thus, the mass of an unstable particle possesses an intrinsic uncertainty, and the magnitude of this uncertainty is inversely proportional to the particle's lifetime.

Experimentally, the mass distribution of unstable particles means that when we measure the products of their decay, their invariant mass does not fall precisely at the physical mass m, but exhibits a broadening. If the particle's width is very broad, much broader than the broadening introduced by detector resolution, we can directly measure the width by measuring the width of the distribution. However, if the particle's width is relatively narrow (i.e., the lifetime is relatively long), it is more precise to measure the lifetime by directly counting the particle's flight distance or decay time and then convert it to a width. Although superficially, measuring the decay-time spectrum and the invariant-mass spectrum of a particle appear to be completely different things, it is interesting that they can lead to the same measurement goal.

It is worth noting that although the lifetimes of many particles appear very short, their widths may be much narrower than one might imagine, because the Planck constant is so small. For example, the muon lifetime is about 2.197 μs, corresponding to a width of merely 3.119 $\times$

[6] A rigorous derivation of the mass distribution of particles involves some subtleties. In fact, this conclusion follows directly from the form of the two-point correlation function of unstable particles. Taking the two-point correlation function of an unstable real scalar boson as an example,

$$\tilde{G}^{(2)}(p) \simeq \frac{i}{p^2 - m^2 + im\Gamma} \,.$$

The Breit–Wigner distribution of the mass of an unstable particle actually originates from the squared modulus of this expression. A relatively rigorous derivation can be found in the *Quantum Field Theory Lecture Notes* by Zhaohuan Yu of Sun Yat-sen University Yu (2025). In fact, the two-point correlation function can be roughly understood as the propagation of a particle, so one can understand why the observed mass of an unstable particle is not fixed through the energy–time uncertainty relation.

10^{-10} eV, nearly ten orders of magnitude lower than the energy scales of interest in condensed-matter physics. The π^0 meson lifetime of 8.43×10^{-17} s already seems extremely short, yet its corresponding width of 8.13 eV has only just reached the electron-volt scale. But there also exist many particles with rather broad widths, which are mainly heavier elementary particles or composite particles, such as the W boson (width 2.085 GeV), the Z boson (width 2.4955 GeV), $\psi(3770)$ (width 27.2 MeV), $\rho(770)$ (width 147.4 MeV), etc. Generally speaking, for these very broad particles the direct significance of lifetime in experiment is no longer important. Through the comparison of particle widths and lifetimes, one can gain an intuitive sense of the order of magnitude of the Planck constant.

2.2 Counting Experiments

In this section we focus on **counting experiments**. The term "counting experiment" here refers to a class of experiments in which the experimental sample consists of simple counts, leaning more towards a statistical experimental method rather than any specific type of experiment. **Counting may be the most important experimental method in particle physics. Many concrete experiments in particle physics can be directly or indirectly reduced to a counting experiment.**

Examples of experiments that can generally be directly reduced to counting experiments include: measuring the cosmic-ray flux in a region over a certain period of time (measurement-type experiment), measuring the activity of a radioactive source by decay counting (measurement-type experiment), searching for beyond-the-Standard-Model physics events in low-background experiments (search-type experiment), searching for ultra-high-energy cosmic-ray gamma photons (search-type experiment), etc. An experiment that could achieve its goal through other measurement methods but can also be accomplished via counting is, for example, Chien-Shiung Wu's experiment that verified parity non-conservation in weak interactions (test-type experiment). That experiment could be carried out by measuring the angular distribution of decay products with a spectrometer, but could also be accomplished by counting the decay-electron yield in a given direction.

Because of the universality and simplicity of the statistical laws governing counting, the advantages of counting experiments lie in their broad applicability, simple statistical analysis, and the direct connection between statistical and physical results. In our laboratory course, if students are unsure about the statistical method to use when analyzing results, they can try to interpret the experiment they have carried out as one or several counting experiments and perform the statistical analysis accordingly; this usually yields good results as well.

2.2.1 Statistical Law of Counting

The Poisson distribution is used to describe the probability distribution of the number of independent random events occurring with a constant event rate per unit time. The Poisson distribution is a discrete random distribution; the probability mass function for a Poisson distribution with expected number of events n is

$$P(k = N) = \frac{n^N}{N!} e^{-n} \ . \tag{13}$$

A criterion for judging whether a counting variable follows a Poisson distribution: **If the events being counted satisfy (1) a fixed expected number of events and (2) independence between any two events, then the count of such events follows a Poisson distribution.** Almost all counting samples involved in particle physics experimental data follow or approximately follow a Poisson distribution. The following are some classic examples:

- The number of unstable-particle decay events within a fixed time interval.
- The number of muons passing through a small cosmic-ray detector within a fixed time interval.
- The number of events in a given bin of a histogram.

Note that if the expected number of events is not constant (e.g., a time-varying event rate) or if the events are not independent (correlated), then their count no longer strictly follows a Poisson distribution. In such cases a statistical correction usually needs to be introduced. The Fano factor involved in particle detector technology is a typical example of correcting the distribution of a counting random variable.

An important property of Poisson variables is **the addition theorem of the Poisson distribution: the sum of mutually independent Poisson variables still follows a Poisson distribution**.

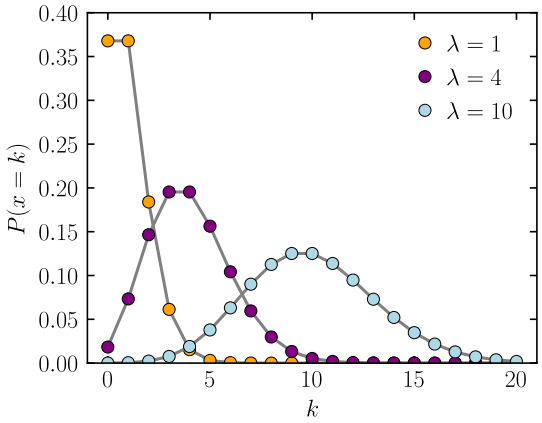


Figure 3: Probability mass function of the Poisson distribution Skbkekas (2010).

2.2.2 Asymptotic Properties of the Poisson Distribution

When the expectation n of the Poisson distribution is large, the Poisson distribution converges asymptotically to a Gaussian distribution with mean n and standard deviation $\sqrt{n}$:

$$\boxed{\text{Pois}(n) \to \mathcal{N}\left(n, \sqrt{n}\right), \quad n \to \infty .} \tag{14}$$

This asymptotic property can be proved directly by evaluating the limit, or it can be regarded as a natural corollary of Lyapunov's central limit theorem. Therefore, we often say that **the uncertainty of a count N is $\sqrt{N}$**. In particle physics experiments, it is generally considered that when $n \gtrsim 20$, the Gaussian distribution already provides a good approximation to the Poisson distribution.

3 Parameter Estimation

We first study the topic of **parameter estimation** in statistics and its applications in particle physics experiments. Parameter estimation is the main branch of statistics involved in measurement-type experiments in particle physics. The goal of parameter estimation is to estimate the values of probability distribution parameters from a data sample; in particle physics experiments this can be concretely expressed as estimating one or more parameters of a theoretical model from experimental data.

3.1 Introduction: How to Measure the Muon Lifetime?

3.1.1 Principle of the Muon Lifetime Measurement Experiment

In Section 2.1 we introduced particle decay and its related theory, but the reader may still be puzzled about how to measure particle lifetimes experimentally. We mentioned earlier that for particles with relatively long lifetimes, the lifetime can be determined by directly counting their decay times. The muon lifetime (about 2.197 μs) is exceptionally long among elementary particles, which allows scientists to determine the muon lifetime directly through time measurements. The basic experimental principle of measuring the muon lifetime is relatively simple. In

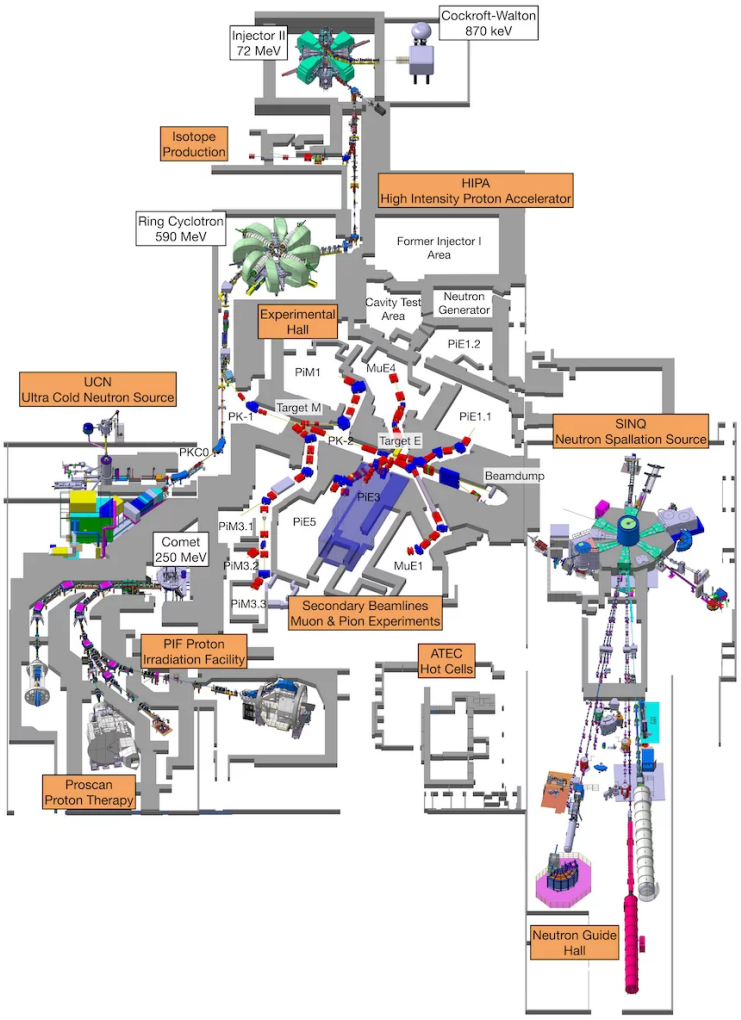


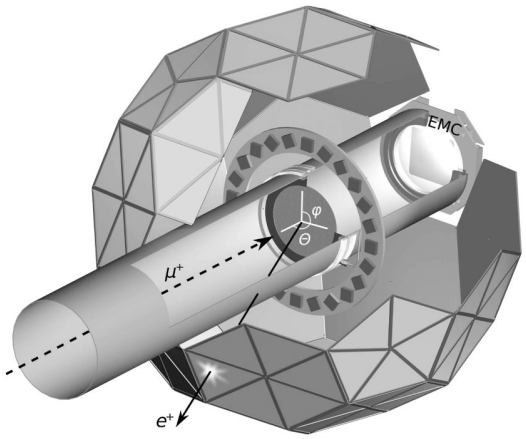
Figure 4: Plan view of the PSI accelerator and beamline facility in Switzerland Anon (2025).

the experiment, a detector is used to record the decay time of each muon, yielding a series of decay-time samples $t_1, \cdots, t_N$; the exponentially distributed probability distribution parameter is then estimated from this data set.

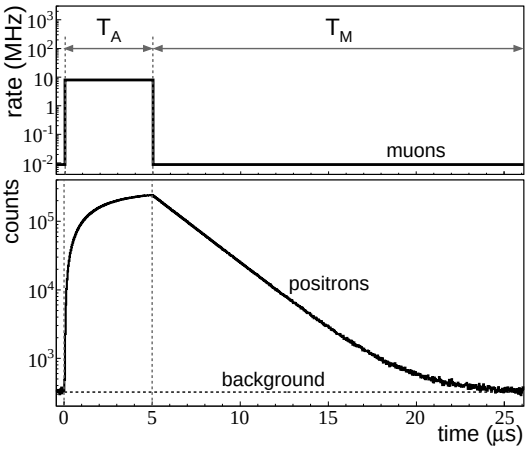
In experiments it is generally necessary to use a stopping target to slow down and stop the muons before measuring the lifetime. Since muons decay inside the material, one generally needs to use positively charged muons (μ^+) to measure the muon lifetime. The reason for typically not using negatively charged muons (μ^-) is that μ^- will be attracted to the atomic nucleus and fall into atomic orbits to form a muonic atom ($\mu^- N$); the relativistic effect arising from the muon's orbital motion causes the observed decay time of the muonic atom into an electron to be longer, requiring corrections in the data analysis that introduce certain systematic uncertainties. Positively charged muons, on the other hand, are not affected by this effect, and their lifetime in the material is the same as that of a free muon. Therefore, using positively charged muons to measure the muon lifetime yields higher-precision results.

3.1.2 Accelerator Muon Sources

Precise measurement of the muon lifetime requires a vast number of muon decay samples. The larger the sample size, the smaller the statistical uncertainty of the measurement, and therefore a very high-flux muon source is needed to accomplish this task. Only accelerator-based muon sources can provide the muon intensity required for this purpose. Existing accelerator muon sources worldwide produce muons by bombarding targets with protons. The basic principle is to use a high-intensity proton beam to collide with nuclei to produce π mesons; the π mesons



(a) Detector schematic Webber et al. (2011).



(b) Time structure of the beam at the experimental terminal. During the accumulation period (T_A) the beam is on, muons accumulate on the stopping target, and the detector trigger is disabled; during the measurement period (T_M) the beam is off, muons gradually decay, and the detector detects the produced positrons and records their timing Tishchenko et al. (2013).

Figure 5: Detector and beam time structure of the MuLan experiment.

then decay to produce muons, which are collected to form a muon beam for experimental use. Currently operating muon sources exist at PSI in Switzerland (Fig. 4), J-PARC in Japan, Fermilab in the United States, TRIUMF in Canada, and ISIS in the United Kingdom. At present no muon source is under construction in China, but the China Spallation Neutron Source (CSNS) in Dongguan, the Heavy-Ion Accelerator Facility (HIAF) under construction in Huizhou, and the Accelerator-Driven Subcritical System (CiADS) all have clear plans for muon source construction and will deliver beams successively within the next few years.

3.1.3 The MuLan Experiment

The most precise muon lifetime measurement to date was obtained in 2011 by the MuLan experiment based on the πE3 beamline at PSI. MuLan used an array of plastic scintillator detectors to record over 2×10^{12} muon decay events, achieving a precision at the parts-per-million level Webber et al. (2011). MuLan aimed to measure the muon lifetime with high precision, thereby determining the Fermi constant (G_F) with high precision. The experiment determined the muon lifetime by measuring the time at which positrons[7] produced by stopped muon decays hit the scintillators. The detector configuration used by MuLan is shown in Fig. 5(a); the detector array is highly symmetric to reduce systematic uncertainties arising from muon spin polarization and detector asymmetry. The experiment used a low-energy muon beam with a periodic time structure; the data-acquisition cycle was divided into a 5 μs beam-on accumulation period and a 22 μs beam-off measurement period. During the accumulation period the beam is on, muons accumulate on the stopping target, and the detector trigger is disabled; during the measurement period the beam is off, muons gradually decay, and the detector detects the produced positrons and records their timing (Fig. 5(b)). After a series of timing corrections, the positron hit times can be converted into muon decay times and used for the final statistical analysis.

3.1.4 Statistical Analysis in the Lifetime Measurement

After obtaining the muon decay-time data from the experiment, we must consider how to extract the muon lifetime from them. We already know that the probability density function (p.d.f.) of the decay time is

$$f(t) = \frac{1}{\tau} e^{-t/\tau} \ , \tag{15}$$

[7] $\mu^+ \to e^+ \nu_e \bar{\nu}_\mu$.

and since the lifetime is the mathematical expectation of the decay time, it would seem that we could estimate the muon lifetime by simply averaging over all time samples:

$$\hat{\tau} = \frac{1}{N}\sum_{k=0}^{N} t_k \ . \tag{16}$$

But this does not seem quite right. The premise for doing so is that the experimental data conform well to an exponential distribution, and we will find that this premise is not satisfied: in addition to a reasonably good exponential spectrum, the experiment also has a uniform-distribution background superimposed, as shown in Fig. 5(b). The origin of this uniform background is mainly accidental coincidences caused by radiation in the experimental hall, natural radioactive background, detector noise, etc. Hence, the distribution to be considered is

$$\boxed{f(t) = \frac{r}{\tau}e^{-t/\tau} + \frac{1-r}{T_{\mathrm{M}}} \ .} \tag{17}$$

This distribution contains two terms, being the superposition of two distributions. The first term is an exponential distribution, representing the physical signal of muon decays; the second term is a uniform distribution, representing the background, where T_{M} is the width of the measurement time window (see Fig. 5(b)). The entire distribution contains two parameters to be estimated, τ and r. Among them, τ is the lifetime parameter of interest, while r represents the ratio of signal to background.

Now a problem arises. There are now two parameters to estimate, and it is easy to show that τ is no longer the expectation of this distribution, so the lifetime cannot be estimated by simple averaging. What should we do then?

3.2 Parameter Estimation

For ease of exposition, let us first give a general definition of parameter estimation. Suppose the directly observable quantity in the experiment is the random variable x, with probability density function $f(x;\boldsymbol{\theta})$, where $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_n\}$ are the physical quantities of interest to be measured. After N measurements of the random variable x, we obtain a sample $\boldsymbol{x} = \{x_1, \cdots, x_N\}$ of size N. The process of estimating the parameter $\boldsymbol{\theta}$ from the sample $\boldsymbol{x}$ is called **parameter estimation**, or more strictly, **point estimation**.[8] The parameter $\boldsymbol{\theta}$ is estimated using a function $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ of the sample $\boldsymbol{x}$; $\hat{\boldsymbol{\theta}}$ is called an **estimator**[9] of the parameter $\boldsymbol{\theta}$, and a particular value of $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ is called an **estimate** of $\boldsymbol{\theta}$.

To help the reader understand the definition of parameter estimation, let us recast it formally using the background-free muon lifetime measurement as an example. The directly observable quantity of this experiment is the muon survival time t, whose probability density function is $f(t;\tau) = e^{-t/\tau}/\tau$, where τ is the muon lifetime to be measured. Experimentally, the muon decay time is measured N times, yielding a decay-time data set $\{t_1, \cdots, t_N\}$ of sample size N.[10] Hence, $\hat{\tau} = \frac{1}{N}\sum_{k=0}^{N} t_k$ is an estimator of the muon lifetime τ.

Returning to the problem posed earlier: for a general probability density function $f(x;\boldsymbol{\theta})$, the problem of estimating parameters from experimental data in fact reduces to the problem of finding an estimator. The maximum likelihood estimator is the most important estimator in estimation theory. The next subsection introduces the general definition and usage of maximum likelihood estimation.

---

[8] Parameter estimation also includes **interval estimation**, one of whose main objectives is to estimate the **confidence interval** of a point estimator (roughly speaking, to estimate the uncertainty of the point estimator).

[9] In fact, properties of estimators such as unbiasedness, consistency, and efficiency are also very important. However, due to space limitations they cannot be discussed here; the reader may consult other relevant literature Cowan (1997); Zhu (2016).

[10] That is, the sample size. This term is commonly used in statistical analysis of particle physics experiments.

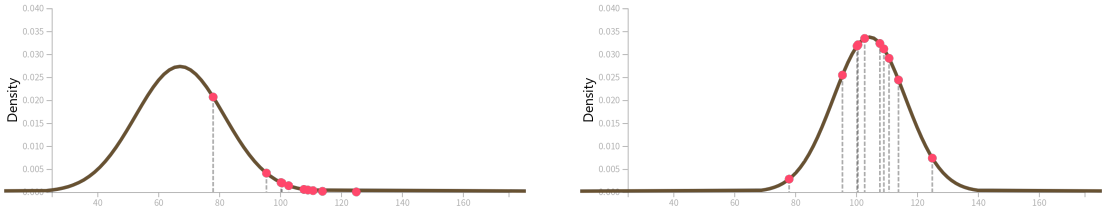


Figure 6: Left: a situation with a small likelihood function value. Right: the situation where the likelihood function value is maximized. Each point in the figure is a sample, and the curve is the Gaussian distribution with the parameter to be estimated.

3.3 Maximum Likelihood Estimation

3.3.1 Maximum Likelihood Estimation

The core idea of maximum likelihood estimation is: **given the observed data, choose as the estimate the parameter value that maximizes the probability of observing these data.** For a given set of independent and identically distributed (i.i.d.) samples $\boldsymbol{x} = \{x_1, \cdots, x_N\}$, if one needs to estimate a probability density function $f(x; \boldsymbol{\theta})$ with parameter $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_n\}$ that best matches the sample, then intuitively the optimal parameter should make the probability density function satisfy:

- Where the probability density is large, the number of samples should be relatively large;
- Where the probability density is small, the number of samples should be relatively small;
- Where the probability density is zero, no samples should appear.

We can try to construct a function $L(\boldsymbol{\theta}; \boldsymbol{x})$ of $\boldsymbol{\theta}$ such that when this function attains its maximum, the sample $\boldsymbol{x}$ has the highest degree of match with the probability density function $f(x; \boldsymbol{\theta})$. In this way, we can take the maximum point $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \boldsymbol{x})$ as the estimator of $\boldsymbol{\theta}$. In fact, this function is precisely the **likelihood function** in estimation theory:

$$\boxed{L(\boldsymbol{\theta}; \boldsymbol{x}) = \prod_{k=1}^{N} f(x_k; \boldsymbol{\theta}) \ .} \tag{18}$$

It is not hard to see that this function satisfies the three requirements mentioned above.[11] **The likelihood function quantifies the degree of match between the parameter-$\boldsymbol{\theta}$ probability density function $f(x; \boldsymbol{\theta})$ and the sample $\boldsymbol{x}$.** The larger the value of the likelihood function, the more likely it is that the sample $\boldsymbol{x}$ was produced by $f(x; \boldsymbol{\theta})$ with the corresponding parameter. Therefore, when the likelihood function is maximized, the corresponding parameter value has the highest probability of being the optimal estimate of the parameter to be estimated.[12]

---

[11] If you cannot imagine it, you may try this demonstration: https://rpsychologist.com/likelihood/

[12] One may sense that this contains some ideas from **Bayesian statistics**. In fact, maximum likelihood estimation is equivalent to **maximum *a posteriori* estimation**

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}; \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{x})} \ ,$$

in the special case where the prior probability distribution $p(\boldsymbol{\theta})$ is uniform (note that the denominator $p(\boldsymbol{x})$ is independent of the parameter to be estimated and is equivalent to a constant. Since $\boldsymbol{x}$ is an i.i.d. sample, $f(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{k=1}^{N} f(x_k; \boldsymbol{\theta})$ is exactly the likelihood function. If $p(\boldsymbol{\theta})$ is constant, then maximizing $p(\boldsymbol{\theta}|\boldsymbol{x})$ is equivalent to maximizing $L(\boldsymbol{\theta}; \boldsymbol{x})$). In other words, if we have no prior knowledge about the parameter $\boldsymbol{\theta}$, then maximum *a posteriori* estimation reduces to maximum likelihood estimation.

Interestingly, from the Bayesian statistics viewpoint, in many experimental data analyses it would seem more appropriate to use maximum *a posteriori* estimation rather than maximum likelihood estimation, because others have always previously carried out related experiments and we always have some prior knowledge about the object of study. However, the pain point of maximum *a posteriori* estimation is that the form of the prior probability distribution $p(\boldsymbol{\theta})$ is often not easy to determine; even if it can be determined, the estimation result obtained

The likelihood function itself is in the form of a product, so operating on its logarithm is more convenient than working with the likelihood function itself. The natural logarithm of the likelihood function is called the **log-likelihood**:

$$\log L(\boldsymbol{\theta}; \boldsymbol{x}) = \sum_{k=1}^{N} \log f(x_k; \boldsymbol{\theta}) \ . \tag{19}$$

It is easy to see that the maximum point of the log-likelihood is the same as that of the likelihood function. Therefore, as long as we find the maximum of the log-likelihood, we have completed the maximum likelihood estimation of the distribution parameters. Formally, **the maximum point of $L(\boldsymbol{\theta}; \boldsymbol{x})$ or $\log L(\boldsymbol{\theta}; \boldsymbol{x})$,**

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}; \boldsymbol{x}) \ , \tag{20}$$

is called the maximum likelihood estimator of $\boldsymbol{\theta}$. This procedure is called **maximum likelihood estimation**. In practice, after estimating the central values of the parameters, one usually also needs to estimate the uncertainties of the parameters and check the quality of the fit. The complete steps of maximum likelihood estimation are as follows:

1. Write down the log-likelihood function $\log L(\boldsymbol{\theta}; \boldsymbol{x})$.
2. Find the maximum point of $\log L(\boldsymbol{\theta}; \boldsymbol{x})$ to obtain the estimator $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}; \boldsymbol{x})$.
3. Estimate the uncertainty of the estimator (see Section 3.3.3).
4. Usually one should also check the goodness of fit, which requires a goodness-of-fit test.

In some simple cases the maximum likelihood estimator admits an analytic solution. Let us take the background-free muon lifetime measurement experiment as an example. As described above, we already know that $\hat{\tau} = \frac{1}{N}\sum_{k=0}^{N} t_k$ is an estimator of the lifetime τ. What then is the maximum likelihood estimator of the lifetime? First we write down the log-likelihood function for the measurement data $\{t_1, \cdots, t_N\}$:

$$\log L(\tau; \boldsymbol{t}) = \sum_{k=1}^{N} \log \left(\frac{1}{\tau} e^{-t_k/\tau} \right) = -N \log \tau - \frac{1}{\tau} \sum_{k=1}^{N} t_k \ . \tag{21}$$

To find the maximum point, we differentiate the above expression with respect to τ and set it to zero:

$$\frac{\partial}{\partial \tau} \log L(\tau; \boldsymbol{t}) = -\frac{N}{\tau} + \frac{1}{\tau^2} \sum_{k=1}^{N} t_k = 0 \ . \tag{22}$$

This is easily solved to give

$$\hat{\tau} = \arg\max_{\tau} \log L(\tau; \boldsymbol{t}) = \frac{1}{N} \sum_{k=0}^{N} t_k \ . \tag{23}$$

This is the same as the estimator we found earlier.

However, only in a few simple cases does the maximum likelihood estimator have an analytic solution; in most cases the estimator does not admit a simple analytic form. Take the muon lifetime measurement with background as an example. The model to be fitted is

$$f(t; \tau, r) = \frac{r}{\tau} e^{-t/\tau} + \frac{1-r}{T_{\mathrm{M}}} \ . \tag{24}$$

---

by feeding in a prior parameter distribution constructed from previous experiments is to some extent already a **combination** of experimental results and may affect the independence of the final result. Some reflections on Bayesian statistics can be found in Ref. Vallverdú (2016).

The corresponding log-likelihood function is

$$\log L(\tau, r; \boldsymbol{t}) = \sum_{k=1}^{N} \log \left(\frac{r}{\tau} e^{-t_k/\tau} + \frac{1-r}{T_\mathrm{M}} \right) . \tag{25}$$

Thus, the equations to be solved are

$$\begin{aligned} \frac{\partial}{\partial \tau} \log L(\tau, r; \boldsymbol{t}) &= \sum_{k=1}^{N} \frac{\frac{r t_k e^{-\frac{t_k}{\tau}}}{\tau^3} - \frac{r e^{-\frac{t_k}{\tau}}}{\tau^2}}{\frac{r e^{-\frac{t_k}{\tau}}}{\tau} + \frac{1-r}{T_\mathrm{M}}} = 0 , \\ \frac{\partial}{\partial r} \log L(\tau, r; \boldsymbol{t}) &= \sum_{k=1}^{N} \frac{\frac{e^{-\frac{t_k}{\tau}}}{\tau} - \frac{1}{T_\mathrm{M}}}{\frac{r e^{-\frac{t_k}{\tau}}}{\tau} + \frac{1-r}{T_\mathrm{M}}} = 0 . \end{aligned} \tag{26}$$

This system of equations is very likely highly transcendental. Even if a closed-form solution exists, its usefulness may not be great.

Faced with a general model, maximum likelihood estimation requires numerical solution. This reduces mathematically to a numerical **optimization** problem, i.e., the problem of finding the extremum of a function. Today's numerical optimization methods are quite mature, and most can be applied to the solution of maximum likelihood estimation. We will not expand on optimization methods here. Interested readers may consult the relevant literature Ma (2001); Nocedal et al. (2006).

The process of estimating the parameters of a probability density function via maximum likelihood estimation is also commonly referred to as fitting. In fact, the **least squares method** familiar from curve fitting is a special case of maximum likelihood estimation.

Maximum likelihood estimation that treats each event as an individual sample is sometimes called "event-by-event fit" in particle physics data analysis. The contrasting method is to first fill the events into a histogram, and then treat the counts in each bin of the histogram as samples for maximum likelihood estimation. This method is called "binned likelihood." The advantage of this histogram-based maximum likelihood estimation is that it can significantly speed up the computation when the sample size is very large, but it can affect the precision of the estimation when the histogram binning is too coarse. Due to space limitations, the theoretical framework of maximum likelihood estimation for fitting histograms will not be presented in detail here. Interested readers may consult Ref. Zhu (2016).

The high-energy physics data analysis framework ROOT includes the RooFit library Verkerke et al. (2003). The basic purpose of RooFit is to fit user-specified probability density functions to user-supplied data. RooFit essentially implements the construction of the likelihood function and the numerical solution of maximum likelihood estimation, but it provides a concise programming interface for users, so that users do not need to manually construct the likelihood function or directly invoke the underlying optimization algorithms to carry out data fitting.

To use and understand RooFit correctly, we first need to understand the complete statistical analysis procedure. Therefore, before introducing the usage of RooFit, we first briefly present the extended maximum likelihood estimation, the uncertainty of the maximum likelihood estimator, and the method for reporting results.

3.3.2 Extended Maximum Likelihood Estimation

In particle physics experimental data analysis, in many cases we need to fit the number of background events and the number of signal events separately, rather than merely the ratio of signal to background. In other words, one usually needs to fit a composite model:

$$\boxed{\rho(x; \boldsymbol{\theta}_\mathrm{sig}, \boldsymbol{\theta}_\mathrm{bkg}, n_\mathrm{sig}, n_\mathrm{bkg}) = n_\mathrm{sig} f_\mathrm{sig}(x; \boldsymbol{\theta}_\mathrm{sig}) + n_\mathrm{bkg} f_\mathrm{bkg}(x; \boldsymbol{\theta}_\mathrm{bkg}) .} \tag{27}$$

where the parameters n_sig and n_bkg are the numbers of signal and background events, respectively, and $f_\mathrm{sig}(x; \boldsymbol{\theta}_\mathrm{sig})$ and $f_\mathrm{bkg}(x; \boldsymbol{\theta}_\mathrm{bkg})$ are the probability density functions of the signal and background, respectively. For instance, in the muon lifetime measurement, the model we generally want to fit is

$$\boxed{\rho(t; \tau, n_\mathrm{sig}, n_\mathrm{bkg}) = \frac{n_\mathrm{sig}}{\tau} e^{-t/\tau} + \frac{n_\mathrm{bkg}}{T_\mathrm{M}} .} \tag{28}$$

It should be noted that $\rho(x;\theta_{\text{sig}},\theta_{\text{bkg}},n_{\text{sig}},n_{\text{bkg}})$ is only a general density function, not a probability density function. Therefore, one cannot blindly treat it as a probability density function to construct the likelihood function; otherwise, very strange results will be obtained.[13] To correctly fit n_{sig} and n_{bkg}, we first consider the following transformation:

$$\begin{aligned} n_{\text{sig}} &= rn \ , \\ n_{\text{bkg}} &= (1-r)n \ . \end{aligned} \tag{29}$$

where $r = n_{\text{sig}}/(n_{\text{sig}}+n_{\text{bkg}})$ is the fraction of signal among the total number of events, and $n = n_{\text{sig}} + n_{\text{bkg}}$ is the total number of events. Eq. (27) is then transformed into

$$\rho(x;\boldsymbol{\theta}_{\text{sig}},\boldsymbol{\theta}_{\text{bkg}},n,r) = n\left(rf_{\text{sig}}(x;\boldsymbol{\theta}_{\text{sig}}) + (1-r)f_{\text{bkg}}(x;\boldsymbol{\theta}_{\text{bkg}})\right) = nf(x;\boldsymbol{\theta}) \ . \tag{30}$$

In the above, $f(x;\boldsymbol{\theta}) = rf_{\text{sig}}(x;\boldsymbol{\theta}_{\text{sig}}) + (1-r)f_{\text{bkg}}(x;\boldsymbol{\theta}_{\text{bkg}})$ is the combined probability density function of background and signal, with $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\text{sig}},\boldsymbol{\theta}_{\text{bkg}},r\}$. It is important to note that **at this point the total number of events is also a random variable** (otherwise the problem reduces to fitting only the signal fraction as before). Hence, we merely need to fit one more parameter—the total number of events n—in addition to estimating the combined probability density function $f(x;\boldsymbol{\theta})$ (note that the signal fraction r is already fitted together within $f(x;\boldsymbol{\theta})$).

Concretely, this means treating the measured total number of events as a random variable as well, and estimating its expectation n during the fit. As described in Section 2.2.1, the event count, being a counting random variable, satisfies a Poisson distribution. In order to simultaneously estimate the parameter n in the maximum likelihood estimation, we consider incorporating the Poisson distribution of the total event count into the likelihood function:

$$\boxed{L(n,\boldsymbol{\theta};N,\boldsymbol{x}) = \frac{n^N}{N!}e^{-n}\prod_{k=1}^{N} f(x_k;\boldsymbol{\theta}) \ .} \tag{31}$$

This is the **extended likelihood function**. Since $N!$ does not affect the maximum point of the extended likelihood function, it can be ignored in the computation. The corresponding log-likelihood function is

$$\boxed{\begin{aligned} &\log L(n,\boldsymbol{\theta};N,\boldsymbol{x}) = -n + N\log n - \log N! + \sum_{k=1}^{N}\log f(x_k;\boldsymbol{\theta}) \ , \\ \Leftrightarrow\quad &\log L(n,\boldsymbol{\theta};N,\boldsymbol{x}) = -n + N\log n + \sum_{k=1}^{N}\log f(x_k;\boldsymbol{\theta}) \ . \end{aligned}} \tag{32}$$

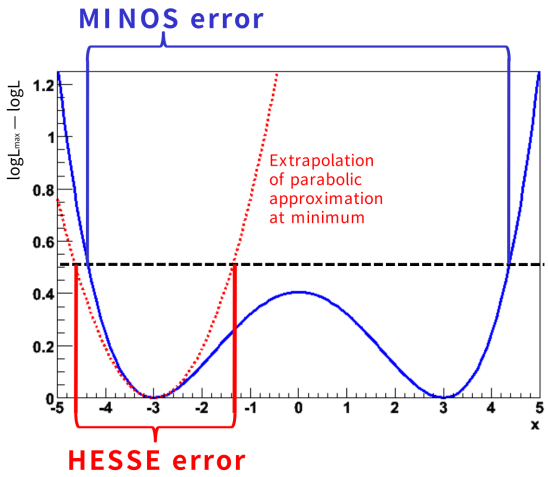
Taking the maximum point of $L(n,\boldsymbol{\theta};N,\boldsymbol{x})$ or $\log L(n,\boldsymbol{\theta};N,\boldsymbol{x})$ as the estimator $\hat{n}$ and $\hat{\boldsymbol{\theta}}$ is called **extended maximum likelihood estimation. Its key distinction from ordinary maximum likelihood estimation is that the sample size is also treated as a random variable.** Its procedural steps are the same as those of ordinary maximum likelihood estimation and will not be repeated here.

In RooFit, for ordinary probability density functions, RooFit defaults to ordinary maximum likelihood estimation. When fitting superimposed probability density functions (`RooAddPdf`), RooFit defaults to extended maximum likelihood estimation. Users can also manually choose whether to use extended maximum likelihood estimation at the time of fitting.

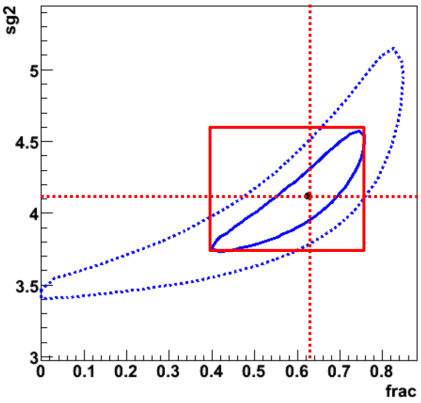
3.3.3 Estimating the Statistical Uncertainty of the Maximum Likelihood Estimator

After obtaining the maximum likelihood estimates of the parameters, we need to assess the statistical uncertainty of the estimators. The statistical uncertainty reflects the uncertainty of the parameter estimation and is an indispensable part of reporting experimental results. Estimating the uncertainty of an estimator is more accurately described as estimating the **confidence interval** of the estimator, which is an **interval estimation** problem, a dedicated topic in mathematical statistics. A proper understanding of interval estimation requires considerable prerequisite knowledge; due to space limitations, we will not delve deeply here but will only briefly introduce some terminology and two basic methods.

[13] You can try it if you don't believe it.



(a) Comparison of confidence intervals given by the information matrix method (HESSE) and the profile method (MINOS). The figure shows an example where the results of the two methods differ considerably.



(b) Confidence intervals given by the profile method for a two-parameter estimation (red box).

Figure 7: Schematic illustration of parameter uncertainty estimation methods Verkerke (2008).

(1) Information matrix method This method is based on the asymptotic normality of the maximum likelihood estimator. When the sample size N is sufficiently large, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ converges in distribution to a multivariate normal distribution:

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}\left(\boldsymbol{\theta}_0, \boldsymbol{V}\right) \ , \tag{33}$$

where $\boldsymbol{\theta}_0$ is the true value of the parameter, and the covariance matrix $\boldsymbol{V}$ is given by the inverse of the Fisher information matrix:

$$\boldsymbol{V} = \boldsymbol{\mathcal{I}}^{-1}, \quad \mathcal{I}_{ij} = -\left\langle \frac{\partial^2 \log L}{\partial\theta_i \partial\theta_j} \right\rangle . \tag{34}$$

The diagonal elements V_{ii} of the covariance matrix are the **variances** of the parameters θ_i, and their square roots are the **standard deviations** $\sigma_{\theta_i} = \sqrt{V_{ii}}$. The off-diagonal elements V_{ij} are the **covariances** $\mathrm{cov}\left(\theta_i, \theta_j\right)$ between parameters θ_i and θ_j. For i.i.d. samples, the Fisher information matrix can be estimated by computing the Hessian matrix of the sample log-likelihood function, and the covariance matrix can be estimated by inverting the information matrix:

$$\hat{\mathcal{I}}_{ij} = -\left.\frac{\partial^2 \log L}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \quad \hat{\boldsymbol{V}} = \hat{\boldsymbol{\mathcal{I}}}^{-1} \ . \tag{35}$$

The confidence interval for each parameter is $(\hat{\theta}_i - \hat{\sigma}_{\theta_i}, \hat{\theta}_i + \hat{\sigma}_{\theta_i})$. Confidence intervals based on the information matrix are always symmetric, as they rest on the large-N asymptotic behavior of the estimator. When analytic calculation is not feasible, one may approximate the second derivatives via numerical differentiation.

(2) Profile method The uncertainty estimates provided by the information matrix are accurate only when the sample size is relatively large. When the sample size is relatively small, the likelihood function departs from a Gaussian shape, and the confidence interval is generally not symmetric. A more precise interval can be constructed through the ratio of the maximum of the likelihood function to the maximum of the **profile likelihood** (i.e., the likelihood ratio). Fixing θ_i in the likelihood function and re-optimizing with respect to the remaining parameters yields the maximum of the profile likelihood; one then computes the log-likelihood ratio statistic between the profile likelihood and the full likelihood:

$$\lambda(\theta_i) = -2\left(\max_{\boldsymbol{\theta}, \theta_i=\hat{\theta}_i} \log L(\boldsymbol{\theta}, \boldsymbol{x}) - \max_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}, \boldsymbol{x})\right), \tag{36}$$

For a single-parameter fit, when $\lambda(\theta_i) < 1$, θ_i approximately falls within the 68.3% confidence interval. In other words, given the experimental sample, we believe that the true value of θ_i has a probability of about 68.3% of lying within the interval $(\hat{\theta}_i - \hat{\sigma}^-_{\theta_i}, \hat{\theta}_i + \hat{\sigma}^+_{\theta_i})$, where $\hat{\theta}_i - \hat{\sigma}^-_{\theta_i}$ and $\hat{\theta}_i + \hat{\sigma}^+_{\theta_i}$ are the two solutions of $\lambda(\theta_i) = 1$.

RooFit provides two uncertainty calculation methods. (1) `RooFit::Hesse`: directly estimates the covariance matrix via the information matrix. This is the default method; it is fast (because the Hessian matrix has already been computed during the minimization), but the uncertainty can be large when the sample size is small. (2) `RooFit::Minos`: scans the confidence interval via the profile method with high precision, but is computationally time-consuming (because after optimizing the likelihood function one must separately optimize the profile likelihood function for each parameter, compute the log-likelihood ratio statistic, and solve for the values at which the log-likelihood ratio statistic equals 1 for each parameter).

Note that what is estimated here is only the **statistical uncertainty** of the estimator, not the **systematic uncertainty**. Systematic errors usually require other methods to estimate; due to space limitations, we will not elaborate further here.

3.4 How to Report Measurement Results

When reporting measurement results, one should report not only the central value $\hat{\theta}_i$ of each parameter but also its uncertainty. This subsection introduces the conventions for reporting measurement results. Students should adhere to these conventions when reporting their measurement results.

In the previous subsection we mentioned that estimating the uncertainty of an estimator is in fact estimating the confidence interval of the estimator. The first type is a symmetric confidence interval of the form $(\hat{\theta}_i - \hat{\sigma}_{\theta_i}, \hat{\theta}_i + \hat{\sigma}_{\theta_i})$, such as the confidence interval estimated by the information matrix method. Usually, we report the 68.3% confidence interval, i.e., the "1σ" interval in the Gaussian-distribution sense. In this case, the measurement result should be reported as

$$\textbf{The measurement result of } \theta_i \textbf{ is } \hat{\theta}_i \pm \hat{\sigma}_{\theta_i}.$$

This result should be interpreted as "given the experimental sample, we believe that the true value of θ_i has a probability of about 68.3% of lying within the interval $(\hat{\theta}_i - \hat{\sigma}_{\theta_i}, \hat{\theta}_i + \hat{\sigma}_{\theta_i})$."

The second type is an asymmetric confidence interval of the form $(\hat{\theta}_i - \hat{\sigma}^-_{\theta_i}, \hat{\theta}_i + \hat{\sigma}^+_{\theta_i})$, such as the confidence interval estimated by the profile method. Likewise, the reported interval is the 68.3% confidence interval. In this case, the measurement result should be reported as

$$\textbf{The measurement result of } \theta_i \textbf{ is } \hat{\theta}_i {}^{+\hat{\sigma}^+_{\theta_i}}_{-\hat{\sigma}^-_{\theta_i}}.$$

The interpretation of this result is likewise "given the experimental sample, we believe that the true value of θ_i has a probability of about 68.3% of lying within the interval $(\hat{\theta}_i - \hat{\sigma}^-_{\theta_i}, \hat{\theta}_i + \hat{\sigma}^+_{\theta_i})$."

When reporting, attention must be paid to the number of significant figures. Usually, **the uncertainty is quoted with one or two significant figures, and the estimated value is rounded to the digit corresponding to the least significant digit of the uncertainty**. For example,

$$\begin{gathered}\text{Muon lifetime: } 2197.81 \pm 0.66 \text{ ns.} \\ J/\psi \text{ mass: } 3096.900 \pm 0.006 \text{ MeV.} \\ \text{Higgs boson width: } 3.7^{+1.9}_{-1.4} \text{ MeV.}\end{gathered}$$

It is important to note that **the operation of keeping significant figures is performed only at the very end; do not artificially round during the computation, as this introduces additional bias**. Furthermore, if, after rounding according to the convention, the two sides of an asymmetric confidence interval become numerically identical, then the result should still be reported in the symmetric confidence interval format.

When using scientific notation or percentage expressions, the power of ten or the percent sign is placed outside. For example,

$$\begin{gathered}\text{Muon anomalous magnetic moment: } (g_\mu - 2)/2 = (11659205.9 \pm 2.2) \times 10^{-10}. \\ D^+ \to K^-\pi^+\pi^+ \text{ branching fraction: } (9.38 \pm 0.16)\%. \\ \psi(2S) \to \gamma\gamma J/\psi \text{ branching fraction: } \left(3.1^{+1.0}_{-1.2}\right) \times 10^{-4}.\end{gathered}$$

In general, a rigorous data analysis will also quote the systematic uncertainty of the measured value. In such cases, the statistical and systematic uncertainties should be reported separately whenever possible. **The usual convention is that the first uncertainty term is the statistical uncertainty and the second is the systematic uncertainty; they may also be individually labeled.** For example,

The fit to these signals gives a mass of $125.3 \pm 0.4\,(\text{stat.}) \pm 0.5\,(\text{sys.})$ GeV Chatrchyan et al. (2012).

$\tau_\mu(R06) = 2196979.9 \pm 2.5 \pm 0.9$ ps Webber et al. (2011).

where "stat." stands for statistical uncertainty and "sys." stands for systematic uncertainty.

When space is limited, symmetric uncertainties may be expressed in compact notation. For instance, 2197.81(66) ns is equivalent to 2197.81 ± 0.66 ns, 3096.900(6) MeV is equivalent to 3096.900 ± 0.006 MeV, 125.3(4)(5) GeV is equivalent to $125.3 \pm 0.4 \pm 0.5$ GeV, etc.

3.5 Introduction to ROOT and RooFit

ROOT is an open-source data analysis framework developed under the leadership of CERN, specifically designed for large-scale data processing in high-energy physics experiments. Its core design philosophy is to seamlessly integrate the mathematical modeling of physical objects, numerical computation, and visualization, providing experimenters with end-to-end support from raw data to scientific conclusions. RooFit, as the statistical library of ROOT, implements functions such as probability density function modeling, fitting, and uncertainty estimation, with its theoretical framework based on rigorous statistical methods. By abstracting the statistical workflow and transforming statistical methods into programmable interfaces, RooFit allows users to focus on the physics problem without having to manually implement the details of numerical computation.

Core Components of the RooFit Library

- **Variables (`RooRealVar`)**: RooFit uses the `RooRealVar` class to define physical quantities or parameters in the model, such as the decay time t or the lifetime parameter τ. Each variable must be specified with a name, a range of allowed values, and an initial value. They correspond to random variables and the parameter space to be estimated in the statistical sense.

- **Datasets (`RooDataSet`)**: Experimental observation data are stored via `RooDataSet`, which supports univariate or multivariate analysis. Users can import data directly from files, or generate pseudo-datasets conforming to specific distributions through code for model validation. By default, RooFit performs event-by-event fits for data imported from `TTree`. To carry out binned maximum likelihood fitting, one must first construct a histogram and then build a `RooDataSet` from the histogram.

- **Probability Density Functions (`RooAbsPdf`)**: RooFit has many commonly used standard or empirical probability density functions built in, such as the Gaussian distribution (`RooGaussian`), the uniform distribution (`RooUniform`), the Breit–Wigner distribution (`RooBreitWigner`), etc. For complex distributions not predefined in RooFit, users can define their own via `RooGenericPdf`. Superimposed distributions can be automatically constructed through `RooAddPdf`; when the user introduces the numbers of signal and background events, the extended likelihood function (Eq. (32)) is automatically constructed. All probability density function classes are derived from the base class `RooAbsPdf`, which defines the common interface for probability density functions.

- **Model Fitting and Error Estimation**: Maximum likelihood estimation is performed via the `RooAbsPdf::fitTo()` method, which automatically solves for the minimum of the negative log-likelihood function using numerical optimization algorithms (such as Minuit2). Error estimation is achieved through the HESSE subroutine (computing the information matrix) or the MINOS subroutine (scanning the profile likelihood function). The concrete implementation is encapsulated by the `RooMinimizer` class, sparing users from having to manually implement complex numerical differentiation and optimization algorithms.

- **Plotting Statistical Charts**: RooFit provides plotting tools through the `RooPlot` class, which can overlay data sets and fitted curves, automatically generating publication-quality figures (see Fig. 9). Users can further draw residual plots to visually present the degree of match between the model and the data.

Based on ROOT and RooFit, the statistical analysis steps in particle physics experiments can be transformed into reproducible and verifiable workflows. This process not only improves analysis efficiency but also ensures the rigor of statistical methods and the reproducibility of results. Readers may consult the following documentation and manuals for further information:

- ROOT Primer: https://root.cern/primer/
- RooFit Quick Start Guide: https://root.cern/manual/roofit/
- Slides introducing RooFit Verkerke (2008): https://root.cern/download/roofit-strasbourg-v10.pdf
- ROOT Tutorials: https://root.cern/manual/
- RooFit Manual: https://root.cern/download/doc/RooFit_Users_Manual_2.91-33.pdf

3.6 Example: Muon Lifetime Measurement

3.6.1 Model Construction

As mentioned earlier, fitting the muon experimental data with background requires numerical methods. Now that we know RooFit can meet our needs, let us see how to use RooFit to fit the data. First, let us review the model we need to fit (Eq. (28)):

$$\rho(t;\tau,n_{\text{sig}},n_{\text{bkg}}) = \frac{n_{\text{sig}}}{\tau}e^{-t/\tau} + \frac{n_{\text{bkg}}}{T_{\text{M}}} \ . \tag{37}$$

This is a composite model, being the superposition of an exponential distribution and a uniform distribution, i.e.,

$$\rho(t;\tau,n_{\text{sig}},n_{\text{bkg}}) = n_{\text{sig}} \times \text{Expo}\left(\frac{1}{\tau}\right) + n_{\text{bkg}} \times \text{Uni}\left(0,T_{\text{M}}\right) \ . \tag{38}$$

In RooFit, the exponential distribution can be defined via `RooExponential`, and the uniform distribution via `RooUniform`; these two distributions can be superimposed using `RooAddPdf`. This is a one-dimensional probability density function with three parameters, requiring the definition of three `RooRealVar` objects. In addition, since `RooExponential` takes the form e^{-cx}, one needs to take the reciprocal of the parameter τ via `RooFormulaVar` and feed it into the exponential distribution. The test data are stored in a `TTree` (`TTree` is equivalent to a data table in which each row is a sample), so we will perform an event-by-event fit. The logic for constructing this composite model in RooFit is summarized in Fig. 8.

3.6.2 Program Code Example

Following the logic in Fig. 8, the following ROOT C++ script can be written to perform the data fitting. The code for plotting the results is placed after the model construction and fitting code. Pay close attention to the connection between the code and the theory; try making modifications and observe the results.

```
// fit_muon_lifetime.cxx - Fit the lifetime measurement experimental data
#include "RooAddPdf.h"
#include "RooDataSet.h"
#include "RooExponential.h"
#include "RooFormulaVar.h"
#include "RooPlot.h"
#include "RooRealVar.h"
#include "RooUniform.h"
#include "TCanvas.h"
#include "TLegend.h"
```

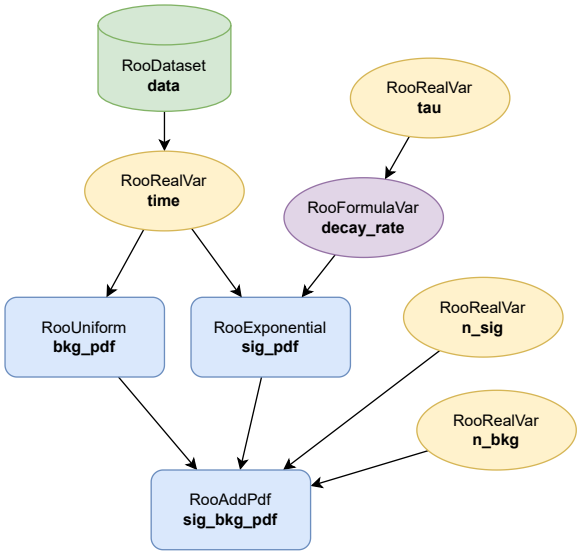


Figure 8: Composite model for fitting muon lifetime measurement experimental data with RooFit.

```
#include "TLegendEntry.h"

auto fit_muon_lifetime() -> void {
    const auto t_min{0.};     // minimum time (ns)
    const auto t_max{20000.}; // maximum time (ns)

    // Define variables and read data
    RooRealVar time{"time", "Time", t_min, t_max, "ns"};
    RooDataSet data{"data", "Dataset", RooArgSet{time},
                    RooFit::ImportFromFile("muon_lifetime_mc.root", "data")};

    // Construct the signal p.d.f.

    RooRealVar tau{"tau", "Lifetime", 2200, 1000, 3000, "ns"};
    RooFormulaVar decay_rate{"decay_rate", "Decay␣rate", "1/tau", RooArgList(tau)};
    RooExponential sig_pdf{"sig_pdf", "Signal␣PDF", time, decay_rate, true}; // the
        ↪ last true adds a minus sign to the exponent

    // Construct the background p.d.f.
    RooUniform bkg_pdf{"bkg_pdf", "Background␣PDF", RooArgSet(time)};

    // Construct the composite model
    const auto n_event{static_cast<double>(data.numEntries())};
    RooRealVar n_sig{"n_sig", "Signal␣events", n_event / 2., 0, n_event * 1.2};
    RooRealVar n_bkg{"n_bkg", "Background␣events", n_event / 2., 0, n_event * 1.2};
    RooAddPdf sig_bkg_pdf{"sig_bkg_pdf", "Signal+Background",
                          RooArgList(sig_pdf, bkg_pdf), RooArgList(n_sig, n_bkg)};

    // Perform the fit
    //   Because RooAddPdf is used, extended maximum likelihood estimation is
        ↪ performed automatically.
    sig_bkg_pdf.fitTo(data);
    // Draw the fit result
    const auto canvas(new TCanvas{"canvas", "Fit␣result", 800, 600});
    const auto frame{time.frame(RooFit::Title("Fit␣result"))};
    data.plotOn(frame, RooFit::Name("data"));              // draw data points
    sig_bkg_pdf.plotOn(frame, RooFit::Name("total"));      // total fitted curve
    sig_bkg_pdf.plotOn(frame, RooFit::Components(sig_pdf), // draw signal component
                       RooFit::LineColor(kRed),            // red
                       RooFit::LineStyle(kDashed),         // dashed
                       RooFit::Name("signal"));            // name the signal
                           ↪ component
    sig_bkg_pdf.plotOn(frame, RooFit::Components(bkg_pdf), // draw background
        ↪ component
                       RooFit::LineColor(kGreen),          // green
```

```
52                          RooFit::LineStyle(kDashed),          // dashed
53                          RooFit::Name("background"));         // name the background
                                ↪ component
54      frame->Draw();
55      // Create legend
56      const auto legend(new TLegend{0.65, 0.65, 0.85, 0.85});
57      legend->AddEntry(frame->findObject("data"), "Data", "EP");            // data
            ↪ points
58      legend->AddEntry(frame->findObject("total"), "Total␣fit", "L");       // total fit
59      legend->AddEntry(frame->findObject("signal"), "Signal", "L");         // signal
            ↪ component
60      legend->AddEntry(frame->findObject("background"), "Background", "L"); //
            ↪ background component
61      legend->SetBorderSize(0);                                             // remove
            ↪ border
62      legend->Draw();                                                       // draw
            ↪ legend
63      // Set logarithmic scale
64      canvas->SetLogy();
65      // Save as PDF
66      canvas->SaveAs("muon_lifetime_fit_result.pdf");
67  }
```

Save the file as `fit_muon_lifetime.cxx` (the filename must match the function name to be executed), and place the ROOT file containing the data to be fitted (specified on line 20 of the script; here it is `muon_lifetime_mc.root`) in the same directory as the script. Execute the following command in the terminal to run the script:

```
1  root -l fit_muon_lifetime.cxx
```

If the data volume is relatively large, the number of terms in the likelihood function will be relatively large, and execution may take some time. During execution, some log messages will be output indicating the status of the optimization algorithm. After the fit completes, the parameter fit results will be output, for example:

```
n_bkg      = 76099        +/-  551.728    (limited)
n_sig      = 5.00702e+06  +/-  2288.05    (limited)
tau        = 2196.52      +/-  1.14106    (limited)
```

Here no uncertainty estimation method was specially specified, so the default HESSE method (i.e., the information matrix method introduced earlier) was used to estimate the parameter uncertainties. If one wishes to instruct RooFit to use the MINOS method (i.e., the profile method) during the fit, line 40 should be changed to:

```
1      sig_bkg_pdf.fitTo(data, RooFit::Minos());
```

That is, add a `RooFit::Minos()` parameter when calling `fitTo`. However, it should be noted that because the number of samples involved in the fit here is very large, the errors estimated by the HESSE method and the MINOS method will not differ appreciably. But when the sample size is very small, using the MINOS method is almost essential. An example using the MINOS method is presented in Section 4.7.2.

After the fit completes, the fit result figure is drawn as specified by the plotting section of the code, as shown in Fig. 9. Students may refer to the ROOT and RooFit documentation, modify the script as needed, and use it to fit other models or analyze other experimental data.

4 Hypothesis Testing

In experiments, beyond estimating unknown parameters in a given model, one often needs to judge whether a model or hypothesis is supported by the experimental data. The final judgment is reached during the data analysis stage, and this process involves the statistical branch of **hypothesis testing**. Since the establishment of particle physics as a discipline in the early 20th century, the relationship between particle physics experiments and theory has gradually shifted from advancing side by side to theory leading experimental research. This has become even more pronounced since the 1960s, with the completion of the Standard Model. Over the past 50 years, a vast number of new physics models beyond the Standard Model have also

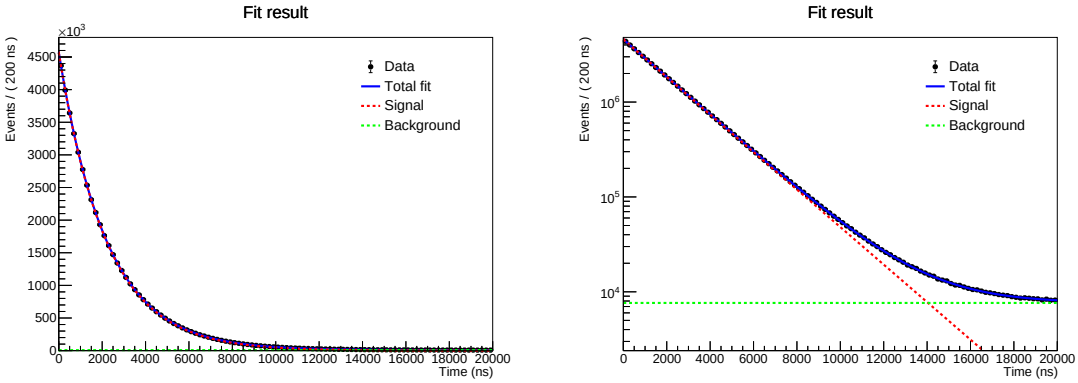


Figure 9: Fit results for the lifetime measurement experimental data. The right figure is the logarithmic-scale version of the left figure.

emerged. Theoretical progress has spawned a large number of corresponding experiments aimed at precisely testing the Standard Model or searching for new physical phenomena or particles predicted by various new physics theories. The core statistical method involved in all of these experiments is hypothesis testing. Before the experiment, we need to optimize the experimental design and evaluate the experimental timeline by simulating the final test performance; after the experiment, we need to draw conclusions about the theory or hypothesis being tested based on the experimental data through hypothesis testing methods.

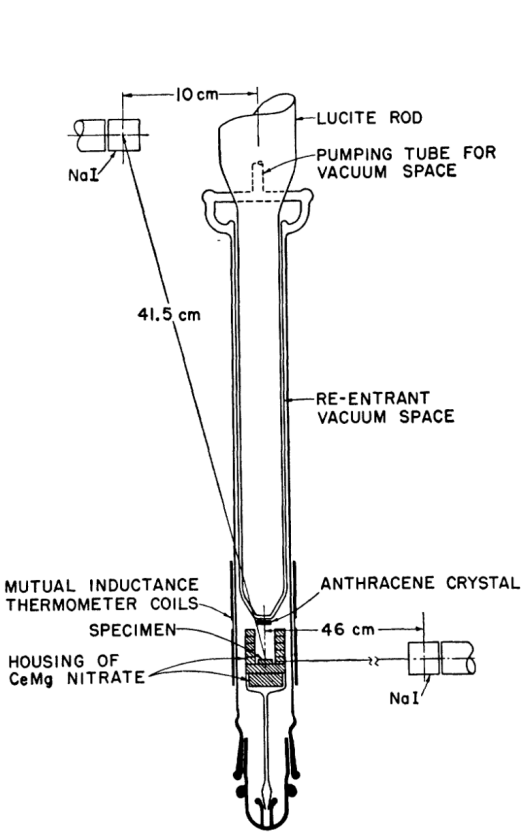
Hypothesis testing theory is quite fascinating, as it can open up new perspectives on many common questions. Is this breakfast safe? Will it rain today? Does the newly purchased phone have a quality defect? In their statistical essence, these questions are no different from scientific questions such as "Do sterile neutrinos exist?" or "Does the muonium–antimuonium conversion process occur?" When we question whether a certain drug is truly effective, when a quality inspector judges whether the acceptance rate of a batch of parts meets the standard, or when a fund manager assesses whether the risk of a quantitative strategy is controllable, the underlying logical framework is all based on the fundamental ideas of hypothesis testing. Although superficially the problems addressed by hypothesis testing are very simple, they contain rich philosophical reflections. These reflections have ultimately been distilled by philosophers and mathematical statisticians into a systematic set of statistical theories and methods, allowing us to make rational answers to these questions based on quantified formulas. A thorough understanding of the ideas behind hypothesis testing will give us new perspectives on many issues and enable us to make more rational and scientific decisions when faced with incomplete information.

The theory and methods of hypothesis testing are quite extensive; here we will focus only on the likelihood ratio test, which is a relatively general-purpose testing method.

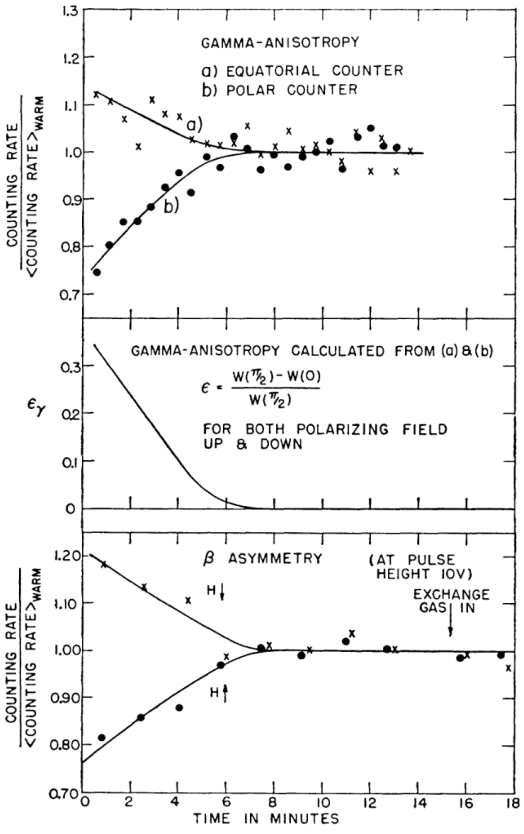
4.1 Introduction 1: How to Test Whether Parity Is Conserved?

In Section 2.2.1 it was mentioned that counting experiments hold a special status in particle physics. In many experiments, we need to compare the difference between counts or count rates. Examples include studying the cosmic-ray muon flux at different altitudes, or studying the spatial asymmetry of a certain process. Here we take the experiment that verified parity non-conservation in weak interactions as an example.

In electromagnetic interactions and gravitational interactions, parity is conserved. In other words, the laws of electromagnetic interaction and gravitational interaction are invariant under spatial reflection transformations. Until the 1950s, the vast majority of physicists believed that all fundamental physical laws were parity-conserving. In June 1956, Chen Ning Yang and Tsung-Dao Lee co-authored a theoretical article titled "Question of Parity Conservation in Weak Interactions" in the journal *Physical Review*, in which they proposed for the first time that there was no clear experimental evidence in weak interactions to prove or disprove that weak interactions are parity-conserving Lee et al. (1956). At the time, Yang and Lee were both in New York, USA; Yang was affiliated with Brookhaven National Laboratory, and Lee with Columbia University. Meanwhile, Chien-Shiung Wu, also at Columbia University, held discussions with



(a) Experimental apparatus schematic.



(b) Experimental results. The bottom panel shows the difference in electron count rates between the two detectors.

Figure 10: The experiment testing parity conservation in weak interactions carried out by Wu et al. Wu et al. (1957).

Yang and Lee and rapidly carried out the experiment. That experiment demonstrated for the first time that parity is not conserved in weak interactions, and the experimental results were published in January 1957 as a rapid communication in *Physical Review*, titled "Experimental Test of Parity Conservation in Beta Decay" Wu et al. (1957). Owing to the importance of this experiment, it is often referred to as the "Wu experiment" in the literature. Yang and Lee shared the 1957 Nobel Prize in Physics for this work, though it is regrettable that Wu was not awarded. Only less than a year elapsed between the experimental verification of parity non-conservation in weak interactions and the awarding of the Nobel Prize—a remarkably short time. This also illustrates, on another level, how sufficient and compelling the Wu experiment's proof of parity non-conservation in weak interactions was. The discovery of parity non-conservation in weak interactions ultimately led to the weak interaction being established as a $V - A$ interaction (an interaction of the vector-minus-axial-vector form), becoming an important component of the Standard Model of particle physics.

In Yang and Lee's article, they proposed that β decay, as a weak-interaction decay, might exhibit an asymmetric angular distribution of the outgoing electron Lee et al. (1956):

$$I(\theta)\mathrm{d}\theta \propto (1 + \alpha \sin\theta) \sin\theta \mathrm{d}\theta \ . \tag{39}$$

where α is the asymmetry factor; if $\alpha \neq 0$ is measured experimentally, this indicates that the angular distribution of the decay products is asymmetric, thereby demonstrating that parity is not conserved in weak interactions. The coordinate system for this flux expression takes the spin direction of the parent nucleus as the polar axis; if $\alpha > 0$, this indicates that the electrons are preferentially emitted in the direction of the spin; if $\alpha < 0$, this indicates that the electrons are preferentially emitted in the direction opposite to the spin.

Nuclear β decay is a weak-interaction process. Wu's experimental design for testing parity

non-conservation in weak interactions was based on examining the spatial symmetry of the electron emission direction in the β decay of cobalt-60 (^{60}Co) nuclei.

$$^{60}\mathrm{Co} \to {}^{60}\mathrm{Ni}^{*} + e^{-} + \nu_e \ . \tag{40}$$

In the experiment, the cobalt-60 nuclei were highly spin-polarized by a strong magnetic field and cooled to extremely low temperatures to suppress thermal motion. The electrons produced by the decay were counted by two scintillator detectors placed at different positions: one detector (the polar counter) was positioned near the direction along the polarization axis, while the other (the equatorial counter) was located in the plane perpendicular to the polarization axis (as shown in Fig. 10(a)). If parity were conserved in weak interactions, the electron fluxes entering these two detectors should not differ significantly. The experimental results showed that the count rates of the two detectors differed markedly (Fig. 10(b)), and this was not due to a bias in the experimental apparatus. Therefore, the experimental results proved that parity is not conserved in weak interactions.

In Wu's experiment, no rigorous hypothesis testing method was explicitly used to quantify whether the difference in count rates between the two detectors was significant, because the difference was too significant![14] Even without computing the significance of the difference between the two detector counts, the figure alone suffices to show that the electron fluxes in different emission directions are significantly different.

However, suppose that in the experiment we are going to carry out the sample size is relatively limited, and the obtained count rates carry certain errors. After the experiment, we need to compare the following two estimated (measured) count rates:

$$\hat{r}_1 = 4.07 \pm 0.18 \ \mathrm{s}^{-1} \ , \quad \hat{r}_2 = 4.47 \pm 0.19 \ \mathrm{s}^{-1} \ .$$

Numerically, 4.07 s^{-1} is nearly 10% smaller than 4.47 s^{-1}. **Can we conclude that "the count rate r_1 is smaller than the count rate r_2"?**

4.2 Introduction 2: How to Search for New Physics at a Collider?

It is now the year 2077. Thanks to the advent of room-temperature superconducting materials and breakthroughs in muon cooling technology, a muon collider has been built, with the center-of-mass collision energy raised to $\sqrt{s} = 10$ TeV, and the first batch of data has been collected and reconstructed. Particle physicists finally have the opportunity to directly study new physics at the TeV energy scale.[15]

Over the preceding fifty years (2025–2075), the field of particle physics experiments has witnessed a large number of important results. After careful analysis, theoretical physicists have concluded that a new particle with a mass near 4.2 TeV and a width of about 100 GeV likely exists. At that time your advisor is participating in the muon collider collaboration, responsible for coordinating part of the physics analysis work. After internal discussions, the collaboration deems it valuable to search for this new particle at the muon collider, and your advisor has secured the sub-project on physics analysis within this larger program and obtained access to the data. Your advisor assigns the data analysis task primarily to a postdoc, while you are responsible for cross-checking the analysis workflow and assisting with the statistical analysis. After several months of work, most of the event selection has been completed, and it is now time to analyze whether there is any trace of a new particle in the data.

The particle to be searched for has a mass of 4.2 TeV and a width of 100 GeV. How does one search for this particle in the data? First, we must identify the signature of this particle in the event data. Because the particle's lifetime is extremely short, it can only be found by measuring the properties of its decay products. In Section 2.1.2 we mentioned that the mass of an unstable particle is not fixed; the center-of-mass energy of its decay products is also not fixed, and follows

[14] Significance test is only significant when the results are not significant.
— A saying

[15] Although the current Large Hadron Collider (LHC) has already reached a center-of-mass energy of $\sqrt{s} =$ 14 TeV, the LHC is a proton–proton collider—a collision of composite particles—which is closer in nature to parton–parton collisions. Therefore, the effective center-of-mass energy is lower than 14 TeV. Muons are elementary particles, so in a muon collider the center-of-mass energy of a muon pair is the full collision energy.

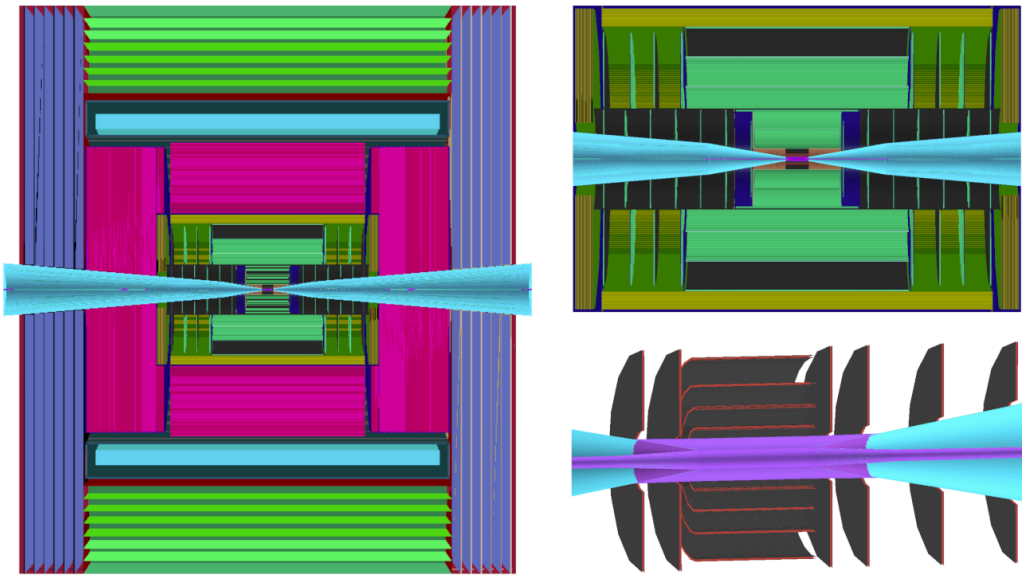


Figure 11: Conceptual design of the muon collider detector Bartosik et al. (2022). Its total length is 5.64 m and height 6.45 m, roughly the height of a two-story building.

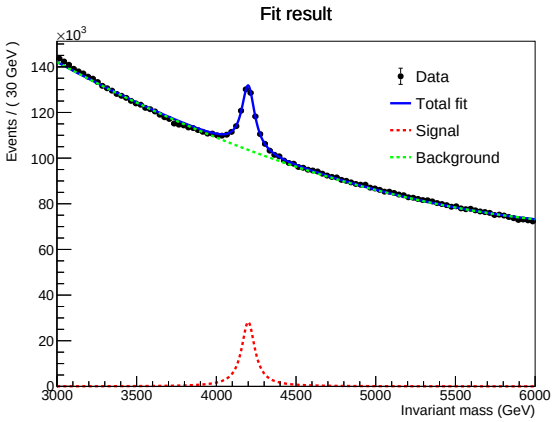


Figure 12: Distribution shapes for fitting signal and background.

a Breit–Wigner distribution, whose central value and width are the particle's mass and width, respectively. Therefore, the basic approach is to look for a mass peak with this property on the invariant mass spectrum of the observed decay final state.[16] Since the particle's width is much larger than the broadening introduced by the detector resolution, this mass peak is precisely the signal distribution we want to fit:

$$f_{\text{sig}}(m) = \frac{\Gamma}{2\pi} \frac{1}{(m - m_0)^2 + \Gamma^2/4} \ . \tag{41}$$

Of course, in addition to the potentially existing signal events, there are also inevitably existing background events. Here, the background corresponding to the decay final state you are studying happens to be very complex: there are all kinds of processes that can leak into the signal region of the final state of interest, and because there are too many components, simulating each background component individually is no longer realistic.[17] Hence, you consider modeling the background distribution with a quadratic polynomial:

$$f_{\text{bkg}}(m; a_1, a_2) = A(a_1, a_2) \left(a_2 m^2 + a_1 m + 1\right) \ . \tag{42}$$

where $A(a_1, a_2)$ is a normalization coefficient. The data may contain background as well as signal, so we apparently would want to fit a superposition distribution of signal and background:

$$\boxed{\rho_1(m; n_{\text{sig}}, n_{\text{bkg}}, a_1, a_2) = n_{\text{sig}} f_{\text{sig}}(m) + n_{\text{bkg}} f_{\text{bkg}}(m; a_1, a_2) \ .} \tag{43}$$

Based on the content introduced earlier, we know that this distribution can be fitted using extended maximum likelihood estimation. But a problem seems to appear. **Fitting such a distribution does not seem to answer a crucial preliminary question: do new physics events exist or not?** If new physics events exist, then fitting this distribution seems reasonable; but if they do not exist, should we instead be fitting a background-only distribution?

$$\boxed{\rho_0(m; n_{\text{bkg}}, a_1, a_2) = n_{\text{bkg}} f_{\text{bkg}}(m; a_1, a_2) \ .} \tag{44}$$

But fitting only the pure-background distribution does not seem right either, because the signal of interest has disappeared from the model.

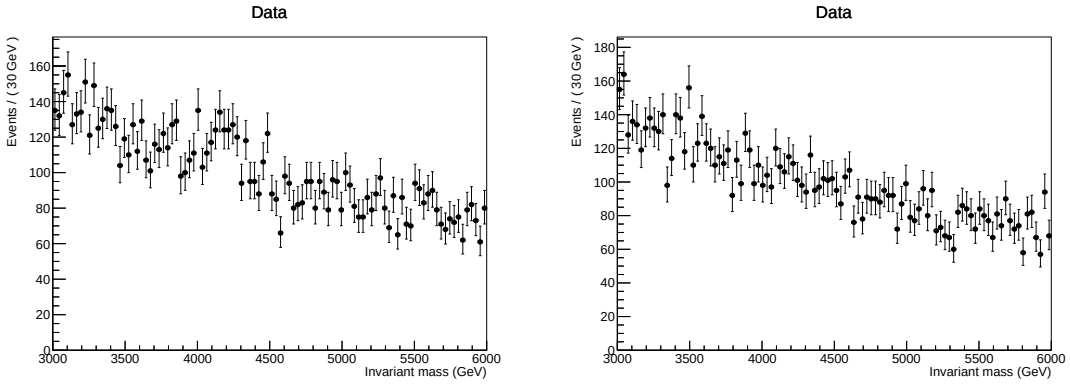


Figure 13: Two sets of Monte Carlo generated data. Without using statistical methods, can one tell at a glance whether there is a signal in the data?

At this point, an idea springs to mind: if we fit both $\rho_0(m; n_{\text{bkg}}, a_1, a_2)$ and $\rho_1(m; n_{\text{sig}}, n_{\text{bkg}}, a_1, a_2)$, and after fitting see which model agrees better with the data, would that not allow us to judge

[16] In a real experiment, one would not fix the mass and perform the analysis; rather, one would scan over a mass range and use the local significance to determine the most likely mass value. Here, for ease of understanding, we simplify the treatment.

[17] Of course, this is largely a simplification made for ease of understanding. In most real cases, especially at lepton colliders, it is possible to simulate the dominant backgrounds. But at hadron colliders (e.g., the LHC), it is sometimes indeed very difficult to analyze every type of background properly.

whether there is a signal? Put differently, we need to determine whether the case $n_{\rm sig} = 0$ (i.e., $\rho_0(m; n_{\rm bkg}, a_1, a_2)$) or the case $n_{\rm sig} > 0$ (i.e., $\rho_1(m; n_{\rm sig}, n_{\rm bkg}, a_1, a_2)$) is more consistent with the experimental data. Therefore, the problem of searching for new physics is transformed into an equivalent question: **Is $n_{\rm sig} = 0$ or $n_{\rm sig} > 0$ better supported by the experimental data?**

4.3 Hypothesis Testing

4.3.1 Null Hypothesis and Alternative Hypothesis

Before introducing the likelihood ratio test, we first present some basic concepts in hypothesis testing. **Hypothesis testing** is a statistical method for making inferences about population parameters or distributional forms based on sample data. Its essence consists in formulating mutually exclusive and exhaustive **null** (H_0) and **alternative** (H_1) hypotheses, constructing a test statistic, measuring its extremeness under the assumption that H_0 is true (the p-value and statistical significance), and ultimately deciding whether to reject H_0 based on a pre-specified significance level (α).

Let the directly observable quantity in the experiment be the random variable x, with probability density function $f(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_n\}$ are the model parameters, the parameter space is Ω, and $\boldsymbol{\theta} \in \Omega$. The **null hypothesis H_0 is usually the initial proposition that the researcher wishes to falsify**, such as "the count rates do not differ" or "no new physics signal exists." Here we concretize H_0 as imposing certain constraints on the model parameters $\boldsymbol{\theta}$ (e.g., fixing some parameters, specifying that some parameters are equal to each other, specifying that some parameters are greater than others, or more general constraints), so that the parameters are confined to a subspace ω of Ω, i.e.,

$$\boxed{H_0 : \boldsymbol{\theta} \in \omega \ .} \tag{45}$$

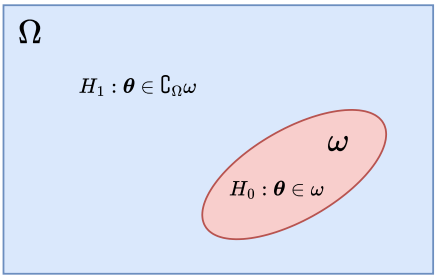


Figure 14: Schematic illustration of the null hypothesis H_0 and the alternative hypothesis H_1 in the full parameter space Ω of the model parameter $\boldsymbol{\theta}$.

The **alternative hypothesis H_1 is usually the proposition that the researcher attempts to prove**, such as "the count rates differ" or "new physics signals exist." The alternative hypothesis H_1 is mutually exclusive with H_0, and either H_0 or H_1 is always true ($H_0 \vee H_1 = \boldsymbol{\theta} \in \Omega$). H_1 is expressed as

$$\boxed{H_1 : \boldsymbol{\theta} \in \complement_\Omega \omega \ .} \tag{46}$$

The above are the definitions of the null and alternative hypotheses, which form the foundation for carrying out hypothesis testing. To aid understanding, we illustrate with specific examples. Taking the test for a difference in count rates as an example, the null and alternative hypotheses are respectively:

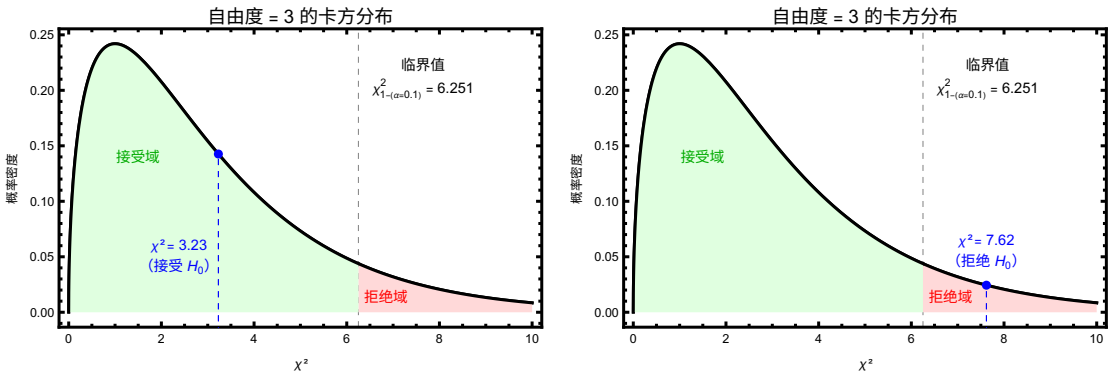
$$\begin{aligned} H_0 &: r_1 = r_2 \text{ (the count rates do not differ)}, \\ H_1 &: r_1 \neq r_2 \text{ (the count rates differ)}. \end{aligned} \tag{47}$$

Taking the search for new physics as an example, the null and alternative hypotheses are respectively:

$$\begin{aligned} H_0 &: n_{\rm sig} = 0 \text{ (new physics does not exist)}, \\ H_1 &: n_{\rm sig} > 0 \text{ (new physics exists)}. \end{aligned} \tag{48}$$

After N measurements of the random variable x, we obtain a sample $\boldsymbol{x} = \{x_1, \cdots, x_N\}$ of size N. **Hypothesis testing refers to the following procedure:**

- Construct a **test statistic**: Based on the sample $\boldsymbol{x}$ and the null hypothesis H_0, construct some test statistic λ that, when H_0 is true, follows a known probability density function $h(\lambda|H_0)$.



(a) χ^2 takes a value that is common under H_0 (falls in the acceptance region, corresponding to the interval with probability $1-\alpha$); H_0 is therefore accepted and H_1 is deemed incorrect.

(b) χ^2 takes a value that is rare under H_0 (falls in the rejection region, i.e., the interval with the remaining probability α); H_0 is therefore rejected and H_1 is deemed correct.

Figure 15: Situations where H_0 is accepted and rejected in a χ^2 test.

- Construct a **rejection region**: Specify a **significance level** α, and construct the rejection region for the statistic λ as $\{\lambda|\lambda \leq \lambda_\alpha\}$ (left-tailed test) or $\{\lambda|\lambda \geq \lambda_{1-\alpha}\}$ (right-tailed test) or $\{\lambda|\lambda \leq \lambda_{\alpha/2} \vee \lambda \geq \lambda_{1-\alpha/2}\}$ (two-tailed test). Here λ_β denotes the upper-β quantile of λ, defined as the value λ_β such that $P(\lambda \leq \lambda_\beta) = \beta$.
- Draw a conclusion: **If the test statistic λ falls in the rejection region, reject H_0 and deem H_1 to be true; otherwise, deem H_0 to be true**.

This procedure appears rather convoluted. Let us try to explain the core idea behind it in more intuitive language. The basic idea of hypothesis testing is somewhat analogous to proof by contradiction. For the test statistic λ constructed here (e.g., χ^2 in a χ^2 test, Z in a Z test, etc.), if H_0 is true, λ will follow its "original" distribution $h(\lambda|H_0)$ (e.g., χ^2 will follow a χ^2 distribution, Z will follow the standard normal distribution, etc.). In other words, as a random variable obeying $h(\lambda|H_0)$, λ should, under the premise that H_0 is correct, typically take values in the region where the probability density is relatively large (e.g., Fig. 15(a)). However, if the experimental data do not conform to H_0, then the test statistic computed from the sample will be more likely to take a value that is quite extreme with respect to H_0 (e.g., Fig. 15(b)). If this happens, we have confidence to believe that H_0 does not hold, and therefore deem H_1 to hold.

From the frequentist perspective, suppose we repeat the experiment many times under the condition that H_0 is true (referred to here as null-control experiments), obtaining a series of test statistic values; these values should typically fall within the high-probability region (the acceptance region) of $h(\lambda|H_0)$, while the low-probability region (the rejection region) of $h(\lambda|H_0)$ should contain very few test statistic samples. On this basis, in order to test whether H_1 holds, we set up a new experiment and introduce the experimental factors to be tested. After carrying out this experiment we compute the test statistic again; if the test statistic value is still a relatively common value in the null-control experiments (i.e., falls in the acceptance region), then we have no reason to say that H_0 does not hold. Conversely, if the newly obtained test statistic takes a value that is very rare in the null-control experiments (i.e., falls in the rejection region), then we can conclude that H_0 is most likely no longer true, and thus have confidence to believe that H_1 holds.

In summary, **the core idea of hypothesis testing is: low-probability events do not occur. The testing procedure first computes the test statistic under the assumption that H_0 is correct, and the test statistic should take common values under the premise that H_0 is correct. If the test statistic obtained from the experiment indeed takes a common value (falls in the acceptance region), then this indicates that H_0 is very likely true. If the obtained test statistic takes a rather extreme value (falls in the rejection region, i.e., a "low-probability event" has occurred), then this conversely indicates that H_0 is very likely not true**.

4.3.2 Significance Level and Two Types of Errors

The basic idea of hypothesis testing has been introduced above, as have the concepts of hypothesis and rejection region. One quantity that remains to be explained is the **significance level** α, whose range is $0 < \alpha < 1$. The significance level is a probability threshold that must be specified subjectively in hypothesis testing; it serves to specify how extreme the test statistic value must be in order to reject H_0.

To define the significance level concretely, we first introduce the concept of type I error in hypothesis testing. As shown in Fig. 15, the acceptance region corresponds to the interval with probability $1-\alpha$ under $h(\lambda|H_0)$, while the rejection region corresponds to the probability α. Hence, there is a probability equivalent to α of rejecting H_0 when it is in fact true. **"Rejecting H_0 when H_0 is actually true," or equivalently "accepting H_1 when H_1 is actually false," is called a type I error**, i.e., a **false positive**. For example, in a new physics search experiment, claiming the discovery of new physics when in fact no new physics phenomenon exists is to commit a type I error.

Strictly speaking, **the significance level α is defined as the probability of committing a type I error**. In other words, if the final test result is to accept H_1, then the probability of committing a type I error is α. The smaller the specified significance level, the lower the probability of a type I error, but a more significant experimental result is also required to support the validity of H_1. To reduce type I errors, the significance level is usually chosen to be a relatively small value, such as 0.1, 0.05, or even 2.8×10^{-7} (the "5σ criterion").

Logically, there is another type of error corresponding to type I error, namely, **"accepting H_0 when H_0 is actually false," or equivalently "rejecting H_1 when H_1 is actually true," which is called a type II error**, i.e., a **false negative**. The probability of this error occurring is given by the statistical power, which slightly exceeds the scope of our discussion and will not be defined here. However, an important conclusion is that **increasing the sample size can reduce the probability of a type II error**. For instance, in a new physics search experiment, the reason we ultimately claim not to have found new physics may, in addition to new physics genuinely not existing, be that the branching fraction or cross section of the new physics process is too small, so that the expected number of events is too low given the current experimental means. After increasing the sample size, it may become possible to find enough new physics events.

Table 1: The two types of errors in hypothesis testing. The colors indicate the general attitudes in scientific research toward different situations.

Reality	**Test Result**	
	Accept H_0 (reject H_1)	**Reject H_0 (accept H_1)**
H_0 true (H_1 false)	Correct inference	Type I error
H_0 false (H_1 true)	Type II error	Correct inference

It is not hard to realize that in published scientific research, out of concern for the rigor of scientific conclusions, we generally consider type I errors to be more serious, while type II errors can sometimes be tolerated. When exploring scientific laws, compared to finding a potentially false "new law," we may prefer to be more conservative and refrain from drawing such a conclusion for the time being. For example, in particle physics, not finding new physics beyond the Standard Model is always more scientifically sound than finding spurious new physics. On the other hand, if the experimental techniques and methods have not yet reached the stage where new laws can be definitively discovered, then committing a type II error in the experiment is still relatively normal. In other fields, the interpretation of the two types of errors may well differ and should be weighed according to the actual context.[18] After all, statistics only reveals patterns; the key in the real world lies in how it is applied.

In particle physics experiments, it is conventional to adopt the "5σ criterion" for determining the significance level, corresponding to $\alpha = 2.8 \times 10^{-7}$. In other words, only when the probability of committing a type I error is smaller than 2.8×10^{-7} can the particle

[18] For example: in COVID-19 nucleic acid testing, should one prefer a lower false-positive rate or a lower false-negative rate?

physics community unambiguously claim a new phenomenon. We will discuss the "5σ criterion" in more detail in Section 4.6.

4.4 Likelihood Ratio Test

We have already introduced the basic ideas of hypothesis testing above, but have not yet introduced specific testing methods. In addition to the test of independence that students have already encountered in high-school mathematics, relatively commonly used testing methods also include the t-test, Z-test, F-test, likelihood ratio test, Kolmogorov–Smirnov test, etc. Although there are already perhaps dozens of named testing methods, they are all based on the fundamental framework of hypothesis testing and differ only in their applicability and choice of test statistic. As long as one understands how the test statistic of a given method is constructed and what distribution it follows, the remaining logic is consistent with the framework of hypothesis testing. Among the many testing methods, the likelihood ratio test is a relatively simple and general-purpose method; some other testing methods can be viewed as approximations or special cases of the likelihood ratio test.

The likelihood ratio test is used to compare whether one model is significantly more consistent with the observed data than its constrained form, i.e., it can be used to compare which of two nested-parameter-space models is clearly superior, but it cannot be directly used to compare two completely different models. For example, it can be used to test whether a signal-plus-background composite model is significantly more consistent with the data than a pure-background model (i.e., the original model with the constraint $n_{\text{sig}} = 0$), but it cannot be used to compare whether a Gaussian distribution or a Breit–Wigner distribution better describes the signal peak. The likelihood ratio test is commonly used in fitting particle physics experimental data when comparing different hypotheses, owing to its advantages of simple operation, broad applicability, and good performance. Its drawback, however, is that it can only compare which of two models is relatively more consistent with the data, but cannot directly test the agreement between a single model and the observed data, nor compare two models whose parameter spaces are not nested.[19]

We now formally introduce the likelihood ratio test. For experimental data $\boldsymbol{x} = \{x_1, \cdots, x_N\}$ and a model $f(x; \boldsymbol{\theta})$ to be tested, its likelihood function is $L(\boldsymbol{\theta}; \boldsymbol{x})$. Let the null hypothesis to be tested be $H_0 : \boldsymbol{\theta} \in \omega$ and the alternative hypothesis be $H_1 : \boldsymbol{\theta} \in \complement_\Omega \omega$. Define the **likelihood ratio** as

$$\Lambda = \frac{\max_{\boldsymbol{\theta}\in\omega} L(\boldsymbol{\theta}; \boldsymbol{x})}{\max_{\boldsymbol{\theta}\in\complement_\Omega\omega} L(\boldsymbol{\theta}; \boldsymbol{x})} = \frac{L\left(\arg\max_{\boldsymbol{\theta}\in\omega} \boldsymbol{\theta}; \boldsymbol{x}\right)}{L\left(\arg\max_{\boldsymbol{\theta}\in\complement_\Omega\omega} \boldsymbol{\theta}; \boldsymbol{x}\right)} \ . \tag{49}$$

In other words, the likelihood ratio is the ratio of the maximum of the likelihood function constrained by H_0 to the maximum of the unconstrained likelihood function. The **log-likelihood ratio test statistic** is defined as minus twice the natural logarithm of the likelihood ratio,

$$\begin{aligned} \lambda = -2\log\Lambda &= -2\left(\log\max_{\boldsymbol{\theta}\in\omega} L(\boldsymbol{\theta}; \boldsymbol{x}) - \log\max_{\boldsymbol{\theta}\in\complement_\Omega\omega} L(\boldsymbol{\theta}; \boldsymbol{x})\right) \\ &= -2\left(\max_{\boldsymbol{\theta}\in\omega}\log L(\boldsymbol{\theta}; \boldsymbol{x}) - \max_{\boldsymbol{\theta}\in\complement_\Omega\omega}\log L(\boldsymbol{\theta}; \boldsymbol{x})\right) \ . \end{aligned} \tag{50}$$

Put differently, the likelihood ratio test statistic is minus twice the difference between the maximum of the H_0-constrained log-likelihood function and the maximum of the unconstrained log-likelihood function. S. S. Wilks proved in 1937 that **if the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ approximately follows a Gaussian distribution, then the log-likelihood ratio test statistic asymptotically follows a χ^2 distribution with $\dim\Omega - \dim\omega$ degrees of freedom**,

$$\lambda \sim \chi^2\left(\dim\Omega - \dim\omega\right) \ , \quad \text{s.t.} \ \ \hat{\boldsymbol{\theta}} \sim \mathcal{N}\left(\boldsymbol{\theta}_0, \boldsymbol{\mathcal{I}}^{-1}\right) \ . \tag{51}$$

where $\dim\Pi$ is the dimensionality of the parameter space Π, $\boldsymbol{\theta}_0$ is the true parameter value, and $\boldsymbol{\mathcal{I}}$ is the Fisher information matrix (Eq. (34)). This theorem is known as **Wilks' theorem**. Note that for the likelihood ratio test to be meaningful, one requires $\dim\Omega - \dim\omega \geq 1$, i.e., $\dim\Omega >$

[19] In such cases, the Kolmogorov–Smirnov test or the χ^2 test can be used Zhu (2016).

$\dim\omega$. In other words, **the likelihood ratio test requires that the null hypothesis H_0 in fact constrains the number of free parameters, or that the alternative hypothesis H_1 in fact introduces new free parameters**.

According to Wilks' theorem, provided that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ follows a Gaussian distribution reasonably well, the distribution of the log-likelihood ratio statistic λ under the null hypothesis can be taken as the χ^2 distribution, and hypothesis testing can be carried out on that basis. We know that Lyapunov's central limit theorem states that $\hat{\boldsymbol{\theta}}$ approximately follows a Gaussian distribution when the sample size is relatively large. However, even when the sample size is relatively small—say, only a few tens of events—the departure of $\hat{\boldsymbol{\theta}}$ from the Gaussian distribution is usually not very large, so one can still treat $\lambda \sim \chi^2\,(\dim\Omega - \dim\omega)$. Nevertheless, if the sample size is extremely small (e.g., fewer than ten events), one must cautiously judge whether this premise is still satisfied. If it is no longer satisfied, the distribution of λ must be estimated via Monte Carlo methods.

After computing the log-likelihood ratio statistic λ corresponding to the sample and the hypotheses, and having clarified the distribution of λ, the remaining steps of the likelihood ratio test are the same as those described in Section 4.3 and will not be repeated here.

In some simple cases, the likelihood ratio test statistic admits an analytic form. Let us take the test for a difference in count rates as an example. Consider two independent count-rate measurements lasting times T_1 and T_2, with the counts obtained being N_1 and N_2, respectively. Assuming that the time measurement errors are relatively small, only the two counts among the observables are random; the experimental observables form a two-dimensional random variable. Hence the experimental sample is (N_1, N_2), with sample size 1. Let the count rates of the two measurements be r_1 and r_2, respectively. Since the two measurements are independent and the counts follow a Poisson distribution, the probability mass function of the experimental sample is

$$P(N_1 = k_1, N_2 = k_2; r_1, r_2) = \left(\frac{(r_1T_1)^{N_1}}{N_1!}e^{-r_1T_1}\right)\left(\frac{(r_2T_2)^{N_2}}{N_2!}e^{-r_2T_2}\right) . \tag{52}$$

The sample size is 1, so the likelihood function has only one term. The log-likelihood function is

$$\begin{aligned} \log L(r_1, r_2; N_1, N_2) &= \log P(N_1 = k_1, N_2 = k_2; r_1, r_2) \\ &= N_1 \log r_1T_1 - \log N_1! - r_1T_1 + N_2 \log r_2T_2 - \log N_2! - r_2T_2 , \\ \Leftrightarrow \quad \log L(r_1, r_2; N_1, N_2) &= N_1 \log r_1T_1 - r_1T_1 + N_2 \log r_2T_2 - r_2T_2 . \end{aligned} \tag{53}$$

As stated earlier, the null and alternative hypotheses for testing a difference in count rates are

$$\begin{aligned} H_0 : r_1 = r_2 \equiv r \text{ (the count rates do not differ)}, \\ H_1 : r_1 \neq r_2 \text{ (the count rates differ)}. \end{aligned} \tag{54}$$

To compute the likelihood ratio, we must first compute the maximum likelihood estimators of the parameters under the two hypotheses separately. First, for H_0,

$$\frac{\partial}{\partial r}\log L(r, r; N_1, N_2) = \frac{N_1}{r} - T_1 + \frac{N_2}{r} - T_2 = 0 , \tag{55}$$

which yields

$$\hat{r}_1(H_0) = \hat{r}_2(H_0) = \underset{r_1=r_2\equiv r}{\arg\max} \log L(r_1, r_2; N_1, N_2) = \frac{N_1 + N_2}{T_1 + T_2} . \tag{56}$$

For H_1, the estimators are evidently

$$\hat{r}_1(H_1) = \frac{N_1}{T_1} , \quad \hat{r}_2(H_1) = \frac{N_2}{T_2} . \tag{57}$$

From these we can compute the log-likelihood ratio statistic:

$$\begin{aligned} \lambda &= -2\left(\log L(\hat{r}_1(H_0), \hat{r}_2(H_0); N_1, N_2) - \log L(\hat{r}_1(H_1), \hat{r}_2(H_1); N_1, N_2)\right) \\ &= -2\Bigg(N_1 \log\left(\frac{N_1 + N_2}{T_1 + T_2}T_1\right) - \frac{N_1 + N_2}{T_1 + T_2}T_1 + N_2 \log\left(\frac{N_1 + N_2}{T_1 + T_2}T_2\right) - \frac{N_1 + N_2}{T_1 + T_2}T_2 \\ &\qquad - (N_1 \log N_1 - N_1 + N_2 \log N_2 - N_2)\Bigg) , \end{aligned} \tag{58}$$

which simplifies to

$$\lambda = 2\left(N_1 \log\left(\frac{T_1+T_2}{T_1}\frac{N_1}{N_1+N_2}\right) + N_2 \log\left(\frac{T_1+T_2}{T_2}\frac{N_2}{N_1+N_2}\right)\right) . \tag{59}$$

This is the analytic result for the log-likelihood ratio statistic.

Now let us examine the distribution satisfied by λ. If $N_1, N_2 \gtrsim 20$, then these two observables each approximately follow a Gaussian distribution. Since $\hat{r}_{1,2} = N_{1,2}/T_{1,2}$, it follows that $\hat{r}_{1,2}$ also approximately follow a Gaussian distribution. According to Wilks' theorem, if H_0 holds, λ follows a χ^2 distribution with $\dim\{(r_1, r_2)|r_1, r_2 > 0\} - \dim\{(r_1, r_2)|r_1 = r_2 > 0\} = 2 - 1 = 1$ degree of freedom. After specifying the significance level α, if $\lambda > \chi^2_{1-\alpha}(1)$ then we reject H_0, i.e., we have confidence $1-\alpha$ that the count rates differ significantly; if $\lambda \leq \chi^2_{1-\alpha}(1)$, then we accept H_0, i.e., we deem that the count rates do not differ significantly. If N_1 and N_2 are both relatively small, the distribution of λ must be estimated separately.

If students are familiar with common testing methods, they may notice that testing the difference in count rates could be done directly with a Z-test. That is correct; it can be shown that the Z-test in this case can be regarded as the asymptotic form of the likelihood ratio test. Furthermore, when finally drawing a conclusion, in addition to directly constructing the rejection region and checking whether the statistic falls within it, one can also compute the p-value or statistical significance and compare it with the given threshold. This content will be introduced in the upcoming Section 4.5.

In the vast majority of cases, however, the likelihood ratio test statistic has no analytic form. This is to a large extent because the maximum likelihood estimator has no analytic form. In such cases a numerical solution is required. In Section 3.2 we already introduced how to perform maximum likelihood fitting using RooFit; therefore, as long as one knows the maximum value of the log-likelihood function obtained by RooFit, one can compute the log-likelihood ratio statistic and carry out the likelihood ratio test. RooFit of course provides such an interface. The specific operation will be introduced in Section 4.7.

4.5 p-value and Statistical Significance

In hypothesis testing, we need to compute the test statistic and compare it with the critical value of the test statistic at the given significance level to decide whether to accept or reject the null hypothesis. However, the test statistic differs from one testing method to another; we would like to have a more intuitive indicator to evaluate whether the test statistic is very extreme with respect to H_0. Here we introduce two commonly used quantities, the p-value and the statistical significance, which can serve this purpose.

4.5.1 p-value

Earlier we mentioned the need to judge whether the test statistic λ falls in the rejection region, i.e., one must compare λ with the quantiles of $h(\lambda|H_0)$ to reach a conclusion on the hypothesis being tested. The logic of this approach is: from the given significance level α one derives the quantile, and then compares the quantile with the test statistic. Would it not be nice if one could provide a statistic to which all other statistics can be converted, one that can be directly compared with α? In this way, the conclusion could be read off directly from the value of this new statistic.

The statistic that satisfies this requirement is called the p-value. **The p-value is defined as the probability, under the null hypothesis, of obtaining a test statistic value that is more extreme than the observed value λ,** i.e., the probability of obtaining this result "by chance." The specific expression for the p-value depends on how the rejection region is constructed, i.e., on whether a one-tailed or two-tailed test is used. For a left-tailed test (i.e., the rejection region has the form $\{\lambda|\lambda \leq \lambda_\alpha\}$), with the observed value of the test statistic λ denoted as $\hat{\lambda}$ (here and below), the corresponding p-value is

$$\boxed{p = P(\lambda \leq \hat{\lambda}|H_0) = \int_{-\infty}^{\hat{\lambda}} h(\lambda|H_0)\mathrm{d}\lambda = F_{h(\lambda|H_0)}(\hat{\lambda}) .} \tag{60}$$

where $F_{h(\lambda|H_0)}(\lambda)$ is the **cumulative distribution function** (c.d.f.) corresponding to $h(\lambda|H_0)$. For a right-tailed test (i.e., the rejection region has the form $\{\lambda|\lambda \geq \lambda_{1-\alpha}\}$), the p-value is

$$p = P(\lambda \geq \hat{\lambda}|H_0) = \int_{\hat{\lambda}}^{+\infty} h(\lambda|H_0)\mathrm{d}\lambda = 1 - F_{h(\lambda|H_0)}(\hat{\lambda}) \ . \tag{61}$$

For a two-tailed test (i.e., the rejection region has the form $\{\lambda|\lambda \leq \lambda_{\alpha/2} \vee \lambda \geq \lambda_{1-\alpha/2}\}$), the p-value is

$$p = 2\min\left\{P(\lambda \leq \hat{\lambda}|H_0), P(\lambda \geq \hat{\lambda}|H_0)\right\} \ . \tag{62}$$

From the expression for the p-value it can be seen that **judging whether the test statistic λ falls in the rejection region is equivalent to judging the ordering between the p-value and the significance level α: $p \leq \alpha$ is equivalent to λ falling in the rejection region**. In other words, if one treats the p-value as a test statistic, then the rejection region for the p-value is $\{p|p \leq \alpha\}$. For example, suppose the significance level is set to 0.05; if the p-value computed from the sample is larger than 0.05, then accept H_0 (do not reject H_0); if the p-value is smaller than 0.05, then accept H_1 and reject H_0.

Beyond being an indicator that can be directly compared with the given significance level, the p-value also serves as an intuitive indicator to characterize the rarity of the experimental sample under H_0, since its definition is precisely the probability, under H_0, of obtaining a sample even rarer than the given data sample. Put conversely, **if the data sample is credible, then the p-value to some degree represents the credibility of H_0. The smaller the p-value, the lower the credibility of H_0, i.e., the higher the credibility of H_1.**[20] Therefore, in many fields the p-value is used directly to express the statistical significance. But in particle physics and related fields, the p-value is usually further converted into a quantile of the standard normal distribution, which is another, perhaps more intuitive, indicator.

4.5.2 Statistical Significance

In particle physics, compared to the p-value, we prefer its corresponding quantile of the standard normal distribution, i.e., the "number of σ" commonly referred to in experiments. The conversion relation is:

$$Z = F^{-1}_{\mathcal{N}(0,1)}(1-p) = -\sqrt{2}\,\mathrm{erfc}^{-1}(2(1-p)) \ . \tag{63}$$

where $F^{-1}_{\mathcal{N}(0,1)}$ is the inverse cumulative distribution function (inverse c.d.f.) of the standard normal distribution, and erfc is the complementary error function. In other words, the probability corresponding to the region beyond Z standard deviations in the Gaussian distribution is p. In particle physics this quantity is commonly used to measure the significance of a result; the larger Z is, the more significant the result. Because people have a deeper intuitive grasp of the normal distribution, reporting that "the p-value of the result reaches 2.8×10^{-7}" never seems as intuitive as saying "the deviation of the result from the Standard Model reaches 5σ."

4.6 How to Report Test Results

In contrast to reporting parameter estimation results, there are not many formatting conventions for reporting test results; as long as the statistical significance of the result is clearly stated before drawing the final conclusion, it is essentially a writing issue. For example:

- At a Higgs boson mass of about 125 GeV, the 7 TeV and 8 TeV data sets show excess signals with significances of 3.2σ and 3.8σ, respectively. In the combined analysis of the full data set, the significance reaches 5.0σ at $m_H = 125.5$ GeV Chatrchyan et al. (2012).

[20] However, one should note that the p-value is neither the probability that the null hypothesis is true, nor the probability that the alternative hypothesis is false. This is a common misinterpretation.

- Under the best-fit $\nu_\mu \leftrightarrow \nu_e$ oscillation hypothesis ($A = 0.205$), the expected asymmetry of the multi-GeV electron-like events deviates from the measured result ($A = -0.036 \pm 0.067 \pm 0.02$) by 3.4σ Fukuda et al. (1998).[21]

- In a joint analysis of the dilepton and lepton-plus-jets channels, the probability that the experimental result is caused by a fluctuation of the expected background is estimated to be 0.26%. Calculated under a Gaussian distribution, this corresponds to an excess signal of 2.8σ. Due to the limited sample size, the existence of the top quark cannot yet be definitively confirmed Abe et al. (1994).[22]

Table 2: Conclusions typically drawn for results of different statistical significances in particle physics experiments.

Statistical significance	p**-value**	**Typical conclusion**
$Z < 3$	$p > 0.0013$	No evidence that H_1 holds
$3 \leq Z < 5$	$2.8 \times 10^{-7} < p \leq 0.0013$	No clear evidence supporting H_1, or there are hints that H_1 may hold
$Z \geq 5$	$p \leq 2.8 \times 10^{-7}$	The experimental data clearly support H_1

Where a clear convention exists is in the choice of significance level before drawing a conclusion. Although in most disciplines choosing $\alpha = 0.05$ is conventional, in particle physics experiments, for establishing new laws or searching for new phenomena—such as determining the neutrino mass ordering or searching for charged lepton flavor violation processes—**one needs the statistical significance to reach 5σ (i.e., $Z \geq 5$) in order to unambiguously claim a new law or a new discovery. This is the so-called 5σ criterion**. The general attitudes of the particle physics community toward results of different statistical significances are roughly as shown in Table 2.

This does not mean, however, that results cannot be published when the significance is below 5σ. When a complete round of data analysis is finished, even if the significance of the result is small and a new discovery cannot be claimed, it is still very necessary to set constraints on the model parameters. For example, in a search for new physics decays of the muon, even if the branching fraction cannot be given, an upper limit on the branching fraction can be provided. In a dark matter search experiment, even if the mass or production cross section of dark matter cannot be given, upper limits on each of them can be provided, and so forth. In the sense of interval estimation, a significant experimental result typically yields a parameter range within which the true parameter value lies; an insignificant experimental result can likewise yield a boundary that the true parameter value, at least with high probability, does not cross.

Interestingly, one can infer what conclusion an experiment has reached simply from the title of the article reporting particle physics experimental results.

- In general, when the statistical significance is below 3σ, the article title usually only carries the wording "Search for," e.g., "A search for $\mu^+ \to e^+\gamma$ with the first dataset of the MEG II experiment."

- When the significance exceeds 3σ, some articles begin to adopt titles with "Evidence for," e.g., "Evidence for top quark production in $\bar{p}p$ collisions at $\sqrt{s} = 1.8$ TeV." However, some studies still use titles of the "Search for" form.

[21] This is the experimental result of the search for atmospheric neutrino oscillations published by the Super-Kamiokande experiment in Japan in 1998. That result discovered ν_μ–ν_τ oscillations in atmospheric neutrinos. Takaaki Kajita shared the 2015 Nobel Prize in Physics with A. McDonald for this work. Prior to Super-Kamiokande's discovery of atmospheric neutrino oscillations, A. McDonald led the SNO collaboration in discovering solar neutrino oscillations.

[22] One year after this result was obtained by the CDF experiment at the Tevatron collider at Fermilab in the United States, i.e., in 1995, CDF confirmed the existence of the top quark with a statistical significance of 5σ Abe et al. (1995). Henceforth, all six flavors of quarks in the Standard Model had been discovered. On September 30, 2011, the Tevatron ceased operation.

- Only when the significance exceeds 5σ is a title with "Observation of" used, e.g., "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC." However, some studies still use titles of the "Evidence for" form.

Should students later embark on the path of particle physics research, may you also achieve results worthy of "Observation of…."

4.7 Example 1: Searching for New Physics

4.7.1 Basic Method

Now let us solve the problem posed in the introduction: judging whether new physics events exist in the experimental data. Earlier we mentioned that we can separately fit the signal-plus-background composite model and the pure-background model, i.e.,

$$\begin{aligned} \rho_1(m; n_{\text{sig}}, n_{\text{bkg}}, a_1, a_2) &= n_{\text{sig}} f_{\text{sig}}(m) + n_{\text{bkg}} f_{\text{bkg}}(m; a_1, a_2) \ , \\ \rho_0(m; n_{\text{bkg}}, a_1, a_2) &= n_{\text{bkg}} f_{\text{bkg}}(m; a_1, a_2) \ . \end{aligned} \tag{64}$$

where the signal component is the Breit–Wigner distribution and the background distribution is a second-order polynomial, the specific forms of which were already introduced in Section 4.2. In Section 3.6 we already learned how to fit both of these models using RooFit.

The key, however, is to judge whether the number of signal events in the data is significantly greater than zero. The likelihood ratio test has already been introduced above, and this testing method is well-suited to the present case. First, let us clarify the hypotheses to be tested:

$$\begin{aligned} H_0 &: n_{\text{sig}} = 0 \text{ (new physics does not exist)}, \\ H_1 &: n_{\text{sig}} > 0 \text{ (new physics exists)}. \end{aligned} \tag{65}$$

Thus, these two hypotheses correspond exactly to the pure-background model and the signal-plus-background composite model, respectively. As long as we fit both separately and obtain the maximum values of their corresponding log-likelihood functions, we can compute the log-likelihood ratio statistic. In RooFit, maximum likelihood estimation is in fact achieved by minimizing the negative log-likelihood function; therefore, one can retrieve from the RooFit fit result the minimum value of the negative log-likelihood function found by RooFit, and compute the test statistic from it. The specific method can be found in the program code example below.

Next, we must determine what distribution the test statistic follows. Wilks' theorem has been introduced above: under the null hypothesis, the log-likelihood ratio test statistic asymptotically follows a χ^2 distribution when the number of events is relatively large, with the number of degrees of freedom equal to the difference in dimensionality between the parameter spaces corresponding to the null and alternative hypotheses. The number of events in the data set we are about to fit is far greater than several tens, so applying Wilks' theorem poses no major problem. Here, under the null hypothesis the free parameters are $n_{\text{sig}}, n_{\text{bkg}}, a_1, a_2$, making 4 in total; under the alternative hypothesis the free parameters are n_{bkg}, a_1, a_2, making 3 in total. Therefore, under the null hypothesis the log-likelihood ratio test statistic follows a χ^2 distribution with 1 degree of freedom.

Finally, we need to compute the p-value and the statistical significance. The likelihood ratio test is a right-tailed test; hence, after computing the value of the log-likelihood ratio test statistic $\hat{\lambda}$, the p-value is computed as follows:

$$p = P(\lambda \geq \hat{\lambda} | H_0) = \int_{\hat{\lambda}}^{+\infty} \chi^2(\lambda; 1) \mathrm{d}\lambda = 1 - F_{\chi^2(1)}(\hat{\lambda}) \ . \tag{66}$$

How is this computed? The ROOT math library provides a ready-made function:

$$1 - F_{\chi^2(n)}(\hat{\lambda}) = \texttt{TMath::Prob(}\hat{\lambda}\texttt{,}n\texttt{)} \ . \tag{67}$$

Thus, one simply calls `TMath::Prob(`$\hat{\lambda}$`,1)`. Another question is how to compute the statistical significance, i.e., to solve for the corresponding quantile of the standard normal distribution from the p-value. This is also already provided in the ROOT math library:

$$Z = F^{-1}_{\mathcal{N}(0,1)}(1 - p) = \texttt{TMath::NormQuantile(}1 - p\texttt{)} \ . \tag{68}$$

After obtaining the statistical significance, one draws a conclusion according to the conventions described in Section 4.6.

4.7.2 Program Code Example

```
// new_physics_analysis.cxx - Fit and test for a signal
#include "RooAddPdf.h"
#include "RooBreitWigner.h"
#include "RooChebychev.h"
#include "RooDataSet.h"
#include "RooExtendPdf.h"
#include "RooFitResult.h"
#include "RooPlot.h"
#include "RooRealVar.h"
#include "TCanvas.h"
#include "TLegend.h"
#include "TMath.h"

#include <iostream>
#include <memory>

auto new_physics_analysis() -> void {
    const auto energy_min{3000.}; // minimum center-of-mass energy (GeV)
    const auto energy_max{6000.}; // maximum center-of-mass energy (GeV)

    // Define variables and read data
    RooRealVar invariant_mass{"InvMass", "Invariant mass", energy_min, energy_max,
        ↪ "GeV"};
    RooDataSet data{"data", "Dataset", RooArgSet{invariant_mass},
                    RooFit::ImportFromFile("mucol_data1.root", "data")};

    // Construct the signal p.d.f. The particle mass and width are known to be
    // near 4200 GeV and near 100 GeV.
    RooRealVar mass{"mass", "Mass", 4200, "GeV"};
    RooRealVar width{"width", "Width", 100, "GeV"};
    RooBreitWigner sig_pdf{"sig_pdf", "Signal PDF", invariant_mass, mass, width};

    // Construct a quadratic-polynomial background p.d.f. (Chebyshev polynomial)
    RooRealVar bkg_c1{"c1", "Background coefficient b", 0, -3, 3};
    RooRealVar bkg_c2{"c2", "Background coefficient a", 0, -3, 3};
    RooChebychev bkg_pdf{"bkg_pdf", "Background PDF", invariant_mass,
                         RooArgList(bkg_c1, bkg_c2)};

    // Construct the pure-background model and the background+signal composite model
    const auto n_event{static_cast<double>(data.numEntries())};
    RooRealVar n_sig{"n_sig", "Signal events", n_event / 2., 0, n_event * 1.2};
    RooRealVar n_bkg{"n_bkg", "Background events", n_event / 2., 0, n_event * 1.2};
    RooExtendPdf ex_bkg_pdf{"ex_bkg_pdf", "Background PDF", bkg_pdf, n_bkg};
    RooAddPdf sig_bkg_pdf{"sig_bkg_pdf", "Signal+Background",
                          RooArgList(sig_pdf, bkg_pdf), RooArgList(n_sig, n_bkg)};

    // There are now two hypotheses:
    // 1. Null hypothesis H0: no signal, i.e., only the background p.d.f.
    // 2. Alternative hypothesis H1: signal exists, with the p.d.f. being the
    //    superposition of background and signal.
    // To perform the likelihood ratio test, do the following in two steps:
    // 1. Find the maximum of each likelihood function (i.e., perform the fit),
    //    and compute the corresponding likelihood values L0 and L1.
    // 2. Construct the likelihood ratio statistic lambda = -2 log(L0/L1),
    //    and compute the p-value and significance.

    // Perform the H0 fit
    //   Because RooExtendPdf is used, extended maximum likelihood estimation
    //   is performed automatically.
    //   RooFit::Save() makes fitTo return the fit result.
    std::unique_ptr<RooFitResult> h0_result{
        ex_bkg_pdf.fitTo(data, RooFit::Minos(), RooFit::Save())};
    // Minimum of the negative log-likelihood (NLL, i.e., -log(L0)) under H0
    const auto h0_min_nll{h0_result->minNll()};
    // Draw the H0 fit result
    const auto h0_canvas(new TCanvas{"h0_canvas", "Fit result", 800, 600});
    const auto h0_frame{invariant_mass.frame(RooFit::Title("Fit result"))};
    data.plotOn(h0_frame, RooFit::Name("data"));          // name the data points
    ex_bkg_pdf.plotOn(h0_frame, RooFit::Name("bkg_fit")); // name the background fit
    h0_frame->Draw();
    // Add H0 legend
```

```
71      const auto h0_legend{(new TLegend{0.6, 0.7, 0.85, 0.85})};
72      h0_legend->AddEntry(h0_frame->findObject("data"), "Data", "PE");
73      h0_legend->AddEntry(h0_frame->findObject("bkg_fit"), "Background␣fit", "L");
74      h0_legend->SetBorderSize(0);
75      h0_legend->Draw();
76      h0_canvas->SaveAs("h0_fit_result.pdf");
77
78      // Perform the H1 fit
79      //   Because RooAddPdf is used, extended maximum likelihood estimation
80      //   is performed automatically.
81      //   RooFit::Save() makes fitTo return the fit result.
82      std::unique_ptr<RooFitResult> h1_result{
83          sig_bkg_pdf.fitTo(data, RooFit::Minos(), RooFit::Save())};
84      // Minimum of the negative log-likelihood (NLL, i.e., -log(L1)) under H1
85      const auto h1_min_nll{h1_result->minNll()};
86      // Draw the H1 fit result
87      const auto h1_canvas(new TCanvas{"h1_canvas", "Fit␣result", 800, 600});
88      const auto h1_frame{invariant_mass.frame(RooFit::Title("Fit␣result"))};
89      data.plotOn(h1_frame, RooFit::Name("data"));              // name the data points
90      sig_bkg_pdf.plotOn(h1_frame, RooFit::Name("total"));      // total fitted curve
91      sig_bkg_pdf.plotOn(h1_frame, RooFit::Components(sig_pdf), // draw signal component
92                         RooFit::LineColor(kRed),               // red
93                         RooFit::LineStyle(kDashed),            // dashed
94                         RooFit::Name("signal"));               // name the signal
                               ↪ component
95      sig_bkg_pdf.plotOn(h1_frame, RooFit::Components(bkg_pdf), // draw background
            ↪ component
96                         RooFit::LineColor(kGreen),             // green
97                         RooFit::LineStyle(kDashed),            // dashed
98                         RooFit::Name("background"));           // name the background
                               ↪ component
99      h1_frame->Draw();
100     // Add H1 legend
101     const auto h1_legend{(new TLegend{0.65, 0.6, 0.85, 0.85})};
102     h1_legend->AddEntry(h1_frame->findObject("data"), "Data", "EP");
103     h1_legend->AddEntry(h1_frame->findObject("total"), "Total␣fit", "L");
104     h1_legend->AddEntry(h1_frame->findObject("signal"), "Signal", "L");
105     h1_legend->AddEntry(h1_frame->findObject("background"), "Background", "L");
106     h1_legend->SetBorderSize(0);
107     h1_legend->Draw();
108     h1_canvas->SaveAs("h1_fit_result.pdf");
109
110     // Compute the statistical significance of the result. Output p-value and
            ↪ quantile.
111     // Construct test statistic: lambda = -2(log L0 - log L1) = 2(NLL0 - NLL1)
112     const auto lambda{2 * (h0_min_nll - h1_min_nll)};
113     // p-value for a right-tailed test with the chi^2 distribution (1 d.o.f.)
114     const auto p{TMath::Prob(lambda, 1)};
115     // Significance
116     const auto z{TMath::NormQuantile(1 - p)};
117     // Output results
118     std::cout << "p-value:␣␣␣␣␣␣" << p << '\n'
119               << "Significance:␣" << z << std::endl;
120 }
```

Save the file as `new_physics_analysis.cxx` (the filename must match the function name to be executed), and place the ROOT file containing the data to be fitted (specified on line 24 of the script; here it is `mucol_data1.root`) in the same directory as this script. Execute the following command in the terminal to run the script:

```
1 root -l new_physics_analysis.cxx
```

The script performs two fits, so the parameters from the two fits will be output successively. Because the data volume here is relatively small, both fits use the MINOS algorithm to estimate parameter uncertainties (lines 56 and 77 of the script) as a demonstration. The log output from MINOS is much longer than that from HESSE, but the key is to find the final estimated central values and confidence intervals. Among these, the output of the central value is the same as without MINOS, but the errors are those given by the HESSE method. One should look for output in this form, where the important part is the central value of the parameter:

```
c1        = -0.440951    +/-  0.234625  (limited)
```

```
c2          = -0.0171495   +/-  0.281086   (limited)
n_bkg       = 50.0155      +/-  7.95257    (limited)
n_sig       = 13.984       +/-  5.2841     (limited)
```

In addition, for each fitted parameter MINOS outputs an asymmetric uncertainty. One should look for output in this form, which gives the two sides of the parameter confidence interval:

```
Minos: Lower error for parameter n_sig  :  -4.904
Minos: Upper error for parameter n_sig  :  5.73122
```

For instance, here the final fit result for n_{sig} is $n_{\text{sig}} = 14.0^{+5.7}_{-4.9}$. Be careful: **one must ignore output of this form**:

```
  Pos |     Name     |  type    |       Value        |     Error +/-
    0 |          c1 | limited  |    -0.4219982009   | 0.2351117177
    1 |          c2 | limited  |   0.05578524545    | 0.2834626212
    2 |      n_bkg  | limited  |      47.4884445    |  7.455669603
    3 |      n_sig  |  fixed   |      19.71629515   |   5.28308836
```

This is intermediate output produced by the MINOS algorithm while optimizing the profile likelihood function; it has no significant meaning for the final result, and one should not mistake it for the final fit result.

The fit result figures are also output, as shown in Fig. 16. Finally, the script computes and outputs the p-value and significance, for example:

```
p-value:      0.000411275
Significance: 3.34509
```

After obtaining the results, one draws a conclusion according to the conventions described in Section 4.6.

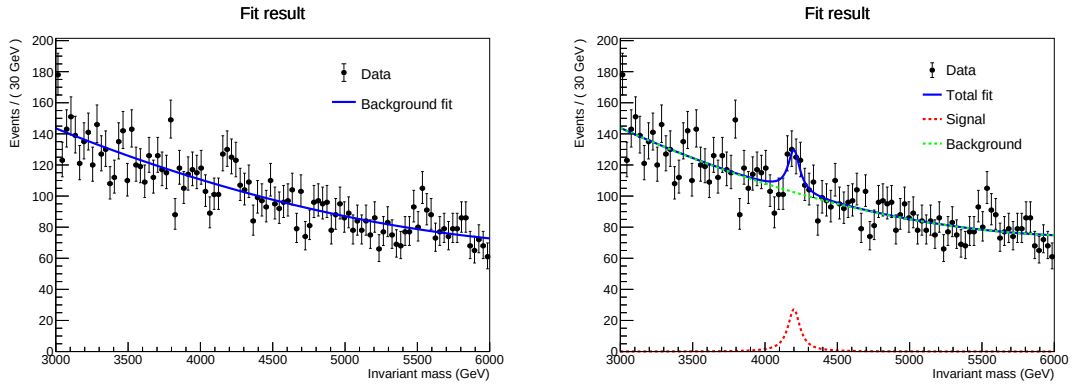


Figure 16: Fit results for the null hypothesis and the alternative hypothesis on the invariant mass spectrum.

4.8 Example 2: Testing a Difference in Count Rates

Finally, we use an example to answer the other question posed at the start: in two independent counting experiments, how does one test whether one count rate differs from another? From the perspective of hypothesis testing, this is a test of which of the following two hypotheses about the count rates r_1 and r_2 is more consistent with the observations:

$$\begin{aligned} H_0 &: r_1 = r_2 \text{ (the count rates do not differ)}, \\ H_1 &: r_1 \neq r_2 \text{ (the count rates differ)}. \end{aligned} \tag{69}$$

In the example provided in Section 4.4, we have already derived the analytic form of the log-likelihood ratio test statistic for this problem. Considering two independent count-rate measurements lasting times T_1 and T_2, with the counts obtained being N_1 and N_2, the log-likelihood ratio test statistic is

$$\lambda = 2\left(N_1 \log\left(\frac{T_1+T_2}{T_1}\frac{N_1}{N_1+N_2}\right) + N_2 \log\left(\frac{T_1+T_2}{T_2}\frac{N_2}{N_1+N_2}\right)\right) . \tag{70}$$

Table 3: Counting experiment data.

Group	Total count	Acquisition time (s)	Count rate (s^{-1})
0	1243	3612	0.3441 ± 0.0098
1	567	1803	0.314 ± 0.013
2	743	1730	0.429 ± 0.016

Suppose we have now carried out three experiments, with the experimental data shown in Table 3. Among them there is one control experiment (number 0) and two measurement experiments (numbers 1 and 2). We wish to know whether the count rates in experiments 1 and 2 differ significantly from that of the control experiment. To this end, we must compute the log-likelihood ratio test statistic λ for group 0 vs. group 1 and for group 0 vs. group 2 separately, and then compute the corresponding p-values and statistical significances Z. Here, under the null hypothesis H_0, the constraint $r_1 = r_2$ restricts the number of free parameters to 1, whereas under the alternative hypothesis H_1, both r_1 and r_2 are free. Because the counts here are all far larger than a few tens, the estimators of the count rates approximately follow a Gaussian distribution, so Wilks' theorem can be applied: i.e., under the premise that H_0 holds, λ follows a χ^2 distribution with $2 - 1 = 1$ degree of freedom. Finally, the two test statistics need to be converted into p-values and statistical significances Z; the functions provided by ROOT can accomplish this directly, as already introduced in Section 4.7. In fact, one can complete the computation of the p-value and Z here using software or libraries such as GNU Octave, SciPy, Excel, Mathematica, MATLAB, etc.; students may explore the specific methods themselves. The results are listed in Table 4.

Table 4: Results of the count rate difference test.

Group	Log-likelihood ratio test statistic	p-value	Statistical significance
$0 \leftrightarrow 1$	3.20	0.074	1.4
$0 \leftrightarrow 2$	22.4	2.3×10^{-6}	4.6

According to the conventions in Section 4.6, based on the results in Table 4 we can conclude that the count rates of group 0 and group 1 do not differ significantly, whereas the count rates of group 0 and group 2 very likely differ.

References

BRUN R, RADEMAKERS F. ROOT: An object oriented data analysis framework[J/OL]. Nucl Instrum Meth A, 1997, 389: 81-86. DOI:10.1016/S0168-9002(97)00048-X.

CERN Higgs gallery[EB/OL]. https://home.cern/resources/image/physics/higgs-collection-images-gallery.

CHATRCHYAN S et al. Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC[J/OL]. Phys Lett B, 2012, 716: 30-61. DOI:10.1016/j.physletb.2012.08.021.

AAD G et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC[J/OL]. Phys Lett B, 2012, 716: 1-29. DOI:10.1016/j.physletb.2012.08.020.

NAVAS S, AMSLER C, GUTSCHE T, et al. Review of particle physics[J/OL]. Phys Rev D, 2024, 110(3): 030001. DOI:10.1103/PhysRevD.110.030001.

YU Z H. Lecture Notes on Quantum Field Theory[EB/OL]. https://yzhxxzxy.github.io/teaching/1807_QFT.pdf.

SKBKEKAS. Poisson_pmf.svg[EB/OL]. https://upload.wikimedia.org/wikipedia/commons/1/16/Poisson_pmf.svg?download.

PSI secondary beamlines[EB/OL]. https://www.psi.ch/en/sbl.

WEBBER D M et al. Measurement of the Positive Muon Lifetime and Determination of the Fermi Constant to Part-per-Million Precision[J/OL]. Phys Rev Lett, 2011, 106: 041803. DOI:10.1103/PhysRevLett.106.079901.

TISHCHENKO V et al. Detailed Report of the MuLan Measurement of the Positive Muon Lifetime and Determination of the Fermi Constant[J/OL]. Phys Rev D, 2013, 87(5): 052003. DOI:10.1103/PhysRevD.87.052003.

COWAN G. Statistical Data Analysis[M]. Oxford: Oxford University Press, 1997.

ZHU Y S. Statistical Analysis of High Energy Physics Experiments[M]. Beijing: Science Press, 2016.

VALLVERDÚ J. Bayesians Versus Frequentists: A Philosophical Debate on Statistical Reasoning[M]. Springer Nature, 2016.

MA W A. Computational Physics[M]. Beijing: Science Press, 2001.

NOCEDAL J, WRIGHT S J. Numerical Optimization[M]. Springer Nature, 2006.

VERKERKE W, KIRKBY D P. The RooFit toolkit for data modeling[J]. eConf, 2003, C0303241: MOLT007.

VERKERKE W. RooFit (lecture)[EB/OL]. https://root.cern/download/roofit-strasbourg-v10.pdf.

LEE T D, YANG C N. Question of Parity Conservation in Weak Interactions[J/OL]. Phys Rev, 1956, 104: 254-258. DOI:10.1103/PhysRev.104.254.

WU C S, AMBLER E, HAYWARD R W, et al. Experimental Test of Parity Conservation in β Decay[J/OL]. Phys Rev, 1957, 105: 1413-1414. DOI:10.1103/PhysRev.105.1413.

BARTOSIK N et al. Simulated Detector Performance at the Muon Collider[A]. arXiv.

FUKUDA Y et al. Evidence for oscillation of atmospheric neutrinos[J/OL]. Phys Rev Lett, 1998, 81: 1562-1567. DOI:10.1103/PhysRevLett.81.1562.

ABE F et al. Evidence for top quark production in $\bar{p}p$ collisions at $\sqrt{s} = 1.8$ TeV[J/OL]. Phys Rev D, 1994, 50: 2966-3026. DOI:10.1103/PhysRevD.50.2966.

ABE F et al. Observation of top quark production in $\bar{p}p$ collisions[J/OL]. Phys Rev Lett, 1995, 74: 2626-2631. DOI:10.1103/PhysRevLett.74.2626.