6 Classical Inference III: Confidence Intervals and Statistical Tests

In previous chapters, we presented several methods for the estimation of values of unknown model parameters $\vec{\theta}$ based on measured data. We also discussed how experimental errors on directly measured quantities impact the errors of estimated model parameter values. However, we did not discuss in detail the interpretation or significance of these error estimates. We shall remedy this situation with comprehensive discussions, in this and the next chapter, of the notions of confidence intervals, confidence levels, and how these concepts can be used toward the formulation of scientific hypotheses and hypothesis testing. The discussions presented in this chapter are primarily frequentist in nature, while those of Chapter 7 rely on the Bayesian interpretation of probability.

We begin with a detailed discussion of the notion of error intervals in §6.1 and present practical considerations and examples in §§6.2 and 6.3. We next introduce, in §6.4, the notions of hypothesis, hypothesis testing, and significance level. Key properties of test power, consistency, and test bias are introduced in §6.5. We complete the chapter with a presentation, in §6.6, of statistical tests commonly used in nuclear and particle physics, as well as a discussion, in §6.7, on test optimization and most particularly the Neyman–Pearson test.

6.1 Error Interval Estimation

6.1.1 Basic Goals and Strategy

The physical and measurement processes involved in the observation of physical observables are intrinsically stochastic in nature. Whether a single or several measurements of an observable X are carried out, fluctuations inherent to the measurement process yield observed values that randomly deviate from the actual or true value of an observable. Either explicitly or implicitly, it is generally assumed that these fluctuations may be described by a specific (but possibly partly unknown) probability model $p(x|x_{\text{true}})$. In the frequentist paradigm, this model describes the probability (density) of observing values x in the range $[x, x + dx]$, expected from an arbitrarily large number of observations, and it is the purpose of statistical estimators, introduced in Chapters 4 and 5, to provide estimates of the true value x_{true} based on measured values. Estimates evidently deviate from the true value by some unknown amount[1] but one wishes to use the probability model $p(x|x_{\text{true}})$ to make a

[1] If the actual deviation was known, it could obviously be corrected for and the true value would be revealed. . .

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

statement about the possible size of such deviations. Clearly, if $p(x|x_{\text{true}})$ is very broad, the dispersion of measured values will be quite large and any particular measured value x may deviate substantially from the actual value x_{true}. On the other hand, if $p(x|x_{\text{true}})$ is narrowly distributed, observed values should be rather close to the true value. One may then use the model $p(x|x_{\text{true}})$ to characterize the error or uncertainty of a measurement. However, the problem arises, typically, that the model $p(x|x_{\text{true}})$ is not perfectly known and that it may feature finite probability density, albeit very small, for arbitrarily large deviates $|x - x_{\text{true}}|$. It is common practice, for instance, to approximate the probability model $p(x|x_{\text{true}})$ by a Gaussian distribution

$$p(x|x_{\text{true}}) \approx \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x - x_{\text{true}})^2}{2\sigma^2}\right], \tag{6.1}$$

which, evidently, features low and high side tails that extended to infinity. For this and similarly behaved models, it is consequently not possible to obtain error bounds that completely enclose the distribution $p(x|x_{\text{true}})$. One must then characterize the uncertainty of an estimate based on a finite interval $[a, b]$ with a specific probability β of containing the true value, given a particular estimate.

To this end, and for simplicity's sake, let us first consider an experiment yielding a single parameter estimate $\hat{\theta}$. We wish to determine the probability, $\beta \equiv P(\hat{\theta}|\theta, \theta_a, \theta_b)$, of finding this estimate within the range $\theta_a \leq \hat{\theta} \leq \theta_b$. In essence, we wish to establish a quantitative method to characterize the errors on the estimate $\hat{\theta}$ based on the probability content β of the interval. Intuitively, it is clear that the larger this probability is, the wider will the range be. Conversely, the wider this range is, the larger should the probability of finding the value $\hat{\theta}$ within it. Evidently, if the probability model of $p(\hat{\theta}|\theta)$ is fully known, there should exist a one-to-one relation between the probability content β and the width of the interval $[\theta_a, \theta_b]$. Unfortunately, there is no unique way of choosing or defining what constitutes a desirable probability value β. Indeed, the choice of a particular value of β is in fact rather arbitrary.

Two approaches are commonly used in the scientific community. For the first, one arbitrarily chooses β to be a convenient and easy-to-remember "round" value, such as 90% or 99%. For the second, one relies on the central limit theorem and assume data tend to follow a Gaussian distribution, with standard deviation σ, and one then typically quotes error intervals in terms of integer multiples of σ.

Recall from §3.9 that for a Gaussian distribution, the probability of obtaining values in the range $\mu \pm n\sigma$ amounts to 68.3%, 95.5%, and 99% for $n = 1$, 2, and 3, respectively. One can then specify error ranges in terms of ± 1, ± 2, or ± 3 standard deviations, with probabilities $\beta = 68.3\%$, 95.5%, and 99%, respectively. However, these probabilities are valid exclusively for Gaussian distributions: random variables obeying a different probability model should in general bear a rather different relation between the width of the range expressed in some multiple of their standard deviation σ and the probability content of the range.

It is worth reiterating that the discussion of errors presented in this chapter is based on the frequentist interpretation of probability. In this context, the use of Bayes' theorem is not strictly possible and one must find a roundabout way to make a statement about θ based on an observed estimate $\hat{\theta}$. We will see in Chapter 7 that posterior probabilities

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

based on Bayes' theorem enable a far more direct reasoning and a much simpler technique to deduce confidence intervals but it is nonetheless convenient and useful to begin with a detailed presentation of frequentist techniques in this and following sections.

6.1.2 Confidence Interval and Confidence Level of Continuous Variables

Let us examine the notion of error quantitatively by considering a set of observations of a *continuous* observable X determined by a PDF $p(x|\theta)$. By definition, $p(x|\theta)\,dx$ is the probability of measuring x in a given interval $[x, x+dx]$ given a parameter θ, and, as such, encapsulates the uncertainty of the measurement. The variable θ may correspond to the actual value of X or some other physical parameter that determines its value. At times fundamentally unknown, $p(x|\theta)$ is often assumed to be Gaussian. In some cases, it can be determined from performance analysis and simulations of the measurement procedure and apparatus. Alternatively, it can also be estimated from a large number of repeated measurements of the same observable X.

Let us assume the function $p(x|\theta)$ represents the density of probability of measurement outcomes of a physical observable X. The probability β of observing (measuring) a value x in the range $[a, b]$ is thus in principle given by

$$\beta \equiv P(a \leq x \leq b|\theta) = \int_a^b p(x|\theta)\,dx. \tag{6.2}$$

If both the functional form of $p(x|\theta)$ and the parameter θ are known, the integral in Eq. (6.2) completely specified and the probability β easy to compute numerically, if not analytically. In practice, however, the value of θ may not be known a priori and it might in fact be the very purpose of the measurement to determine its value. Without a value for θ, calculation of Eq. (6.2) seems impossible, as well as an assessment of the range $[a, b]$ with probability content β. How, then, can one obtain any meaningful estimate of the error on the parameter θ and the observable X?

It turns out that knowledge of the variable θ is not absolutely required so long as the functional form of $p(x|\theta)$ is known, and, provided it is possible to formulate a function $z(x, \theta)$ of the observation of x and the unknown parameter θ whose PDF is **independent** of θ. Indeed, as we shall describe in the following, if such a function exists, it may become possible to reformulate Eq. (6.2) as a problem of **interval estimation**.

In this context, and given a value β, the problem then consists in finding an (optimal) range $[\theta_a, \theta_b]$ such that

$$P(\theta_a \leq \hat{\theta} \leq \theta_b) = \beta, \tag{6.3}$$

where $\hat{\theta}$ is an estimate of the true value of the parameter θ. The interval $[\theta_a, \theta_b]$ is then known as a **confidence interval** of $\hat{\theta}$ with a probability β. In other words, the probability of obtaining an estimate within that interval amounts to β. The interval is then said to be a $100 \times \beta\%$ **confidence interval**. For instance, an interval whose content has a probability of $\beta = 0.9$ is said to be a 90% confidence interval.

Formally, a method which produces an interval $[\theta_a, \theta_b]$ satisfying Eq. (6.3) is said to have the property of **coverage**, and, as it happens, an interval that does not possess this property

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

cannot be considered a confidence interval [115]. This notion will become important and much clearer, later in this chapter, when we consider confidence intervals near a physical boundary.

The process of finding a function $z(x, \theta)$ that enables the determination of a confidence interval is best illustrated with examples. We begin our discussion with the determination of the mean of Gaussian deviates.

6.1.3 Confidence Interval for the Mean of Gaussian Deviates

Let the function $p(x|\theta)$, considered in Eq. (6.2), be a Gaussian PDF $p_{\mathrm{G}}(x|\mu, \sigma)$, where the mean, μ, has been substituted for θ and is now the parameter of primary interest. If the mean μ and standard deviation σ are known, the probability β may be evaluated for any arbitrary interval $[a, b]$:

$$\beta \equiv P(a \leq x \leq b|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^b \exp\left[-\frac{(x'-\mu)^2}{2\sigma^2}\right] dx', \tag{6.4}$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right), \tag{6.5}$$

where $\Phi(z)$ is the **Cumulative Distribution Function** of the normal PDF, given by Eq. (3.131).

The preceding calculation is well defined if μ and σ are known. But what if μ is in fact unknown and the measured value of x intended to obtain an estimate of μ? One would then be interested in calculating the probability β that the value x lies within some interval $[\mu + \Delta_1, \mu + \Delta_2]$. This may be written

$$\beta = P(\mu + \Delta_1 \leq x < \mu + \Delta_2|\sigma) \tag{6.6}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{\mu+\Delta_1}^{\mu+\Delta_2} \exp\left[-\frac{(x'-\mu)^2}{2\sigma^2}\right] dx'. \tag{6.7}$$

On first inspection, this integral seems impossible because the mean μ is unknown. However, consider a change of variable $z = (x' - \mu)/\sigma$. The preceding integral then becomes

$$\beta = \frac{1}{\sqrt{2\pi}} \int_{\Delta_1/\sigma}^{\Delta_2/\sigma} \exp\left(-\frac{1}{2}z^2\right) dz = \Phi\left(\frac{\Delta_2}{\sigma}\right) - \Phi\left(\frac{\Delta_1}{\sigma}\right), \tag{6.8}$$

which is clearly independent of μ and calculable for specific bounds Δ_1 and Δ_2 provided σ is known.

By construction, the integral (6.8) corresponds to the probability of finding x in the interval $\mu + \Delta_1 \leq x \leq \mu + \Delta_2$. One may, however, rearrange the terms of the two inequalities and get

$$\beta = P(x - \Delta_2 \leq \mu \leq x - \Delta_1|\sigma). \tag{6.9}$$

The value β thus corresponds to the probability of finding the unknown value μ in the range $x - \Delta_2 \leq \mu \leq x - \Delta_1$ given the measured value x, and for a known value of σ. Although, strictly speaking, the integral (6.8) is really a statement about the probability of x given

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

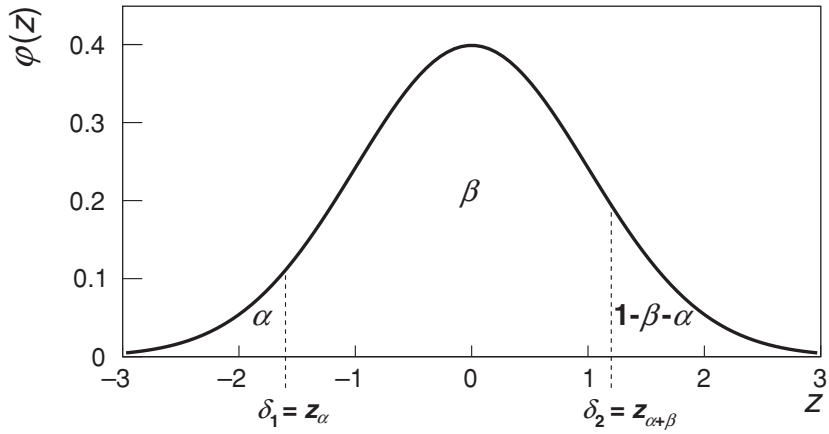


Fig. 6.1 1D confidence interval $[x - \delta_2\sigma, x - \delta_1\sigma]$ for the mean μ of a Gaussian distribution with $\delta_1 = z_\alpha$ and $\delta_2 = z_{\alpha+\beta}$. The probabilities of the three regions delimited by vertical lines are α, β, and $1 - \alpha - \beta$, respectively.

μ, the capacity to reformulate its integrant in terms of $z = (x' - \mu)/\sigma$ enables us to use x to find some bounds on μ with probability β. The interval $x - \Delta_2 \leq \mu \leq x - \Delta_1$ is thus indeed a **confidence interval** for μ with probability β given by Eq. (6.8).

If neither σ nor μ are known, one may instead calculate $P(\delta_1 \leq z \leq \delta_2)$ according to

$$\beta \equiv P(\delta_1 \leq z \leq \delta_2) = \frac{1}{\sqrt{2\pi}} \int_{\delta_1}^{\delta_2} \exp\left(-\frac{1}{2}z^2\right) dz, \tag{6.10}$$

$$= \Phi(\delta_2) - \Phi(\delta_1). \tag{6.11}$$

Converting the inequalities $\delta_1 \leq z \leq \delta_2$ in terms of μ, one gets

$$P(x - \delta_2\sigma \leq \mu \leq x - \delta_1\sigma) = \Phi(\delta_2) - \Phi(\delta_1), \tag{6.12}$$

which provides a generic confidence interval $[x - \delta_2\sigma, x - \delta_1\sigma]$ of probability $\beta = \Phi(\delta_2) - \Phi(\delta_1)$ expressed in terms of some multiples (not necessarily integers) δ_1 and δ_2 of the standard deviation of the PDF of x.

There is much latitude in the choice of δ_1 and δ_2 for a given value of β. For instance, choosing $\delta_1 = -1$ and $\delta_2 = 1$, one obtains a symmetric plus/minus one sigma confidence level with $P(x - \sigma \leq \mu \leq x + \sigma) = 68.3\%$. But arbitrarily many other interval choices also have the same probability. Indeed, recalling the definition of the α–point from Eq. (2.63), one notes that the difference $\Phi(\delta_2) - \Phi(\delta_1)$ in the preceding integral may be written $\Phi(z_{\alpha+\beta}) - \Phi(z_\alpha) = \beta$, where z_α and $z_{\alpha+\beta}$ are the α–point and the $(\alpha + \beta)$−point of the distribution, respectively. The confidence level of probability β can consequently be expressed in any interval $[z_\alpha, z_{\alpha+\beta}]$, as schematically illustrated in Figure 6.1. There are obviously infinitely many such intervals. It is customary, however, to use symmetrical intervals, known as **central intervals**, consisting of integer multiples of the standard deviation or some convenient values of β, as shown in Table 6.1.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

Table 6.1 Central confidence intervals commonly used with a standard normal distribution

$\beta = 1 - 2\alpha$	z_α	$z_{\alpha+\beta}$
0.6827	−1.00	1.00
0.9000	−1.65	1.65
0.9500	−1.96	1.96
0.9545	−2.00	2.00
0.9900	−2.58	2.58
0.9973	−3.00	3.00

6.1.4 Central vs. Minimal Error Intervals

For a strongly peaked, symmetric, but not necessarily Gaussian PDF, the notion of central interval is easy to visualize: it suffices to find a symmetric range $\hat{\theta} \pm \Delta$ with a specific probability content β. If the distribution is not symmetric about its peak, however, one must define the interval based on the probability content outside of the interval and require the probability of observing values below and above the central range to be equal. A **central interval** may then be defined by

$$\begin{aligned} \tfrac{1-\beta}{2} &= P(\hat{\theta} < \theta_a|\theta) = \textstyle\int_{-\infty}^{\theta_a} p(x|\theta)\, dx, \\ \tfrac{1-\beta}{2} &= P(\hat{\theta} > \theta_b|\theta) = \textstyle\int_{\theta_b}^{\infty} p(x|\theta)\, dx, \end{aligned} \tag{6.13}$$

which, in general, shall yield bounds θ_a and θ_b with unequal probability density, that is, such that $p(\theta_a) \neq p(\theta_b)$. We show below that such an interval does not correspond to the shortest interval with probability content β. So, rather than a central interval, one may alternatively seek the **shortest interval** consistent with probability content β.

In order to identify the shortest interval with probability content β, consider the intervals $[\theta_a, \theta_b]$ and $[\theta'_a, \theta'_b]$ displayed in Figure 6.2, and defined such that they both have a probability content β, that is, such that they satisfy

$$\begin{aligned} \textstyle\int_{\theta_a}^{\theta_b} p(\hat{\theta}|\theta)\, d\hat{\theta} &= \beta \\ \textstyle\int_{\theta'_a}^{\theta'_b} p(\hat{\theta}|\theta)\, d\hat{\theta} &= \beta, \end{aligned} \tag{6.14}$$

where the bounds θ_a and θ_b were chosen in order to satisfy the condition

$$p(\theta_a|\theta) = p(\theta_b|\theta), \tag{6.15}$$

illustrated by the horizontal dashed line in Figure 6.2. We proceed to show that these bounds define the most compact (i.e., shortest) interval compatible with Eq. (6.14). To this end, we calculate the integral of the PDF $p(\hat{\theta}|\theta)$ in the range $[\theta_a, \theta'_b]$. By construction, we may write

$$\int_{\theta_a}^{\theta'_b} p(\hat{\theta}|\theta)\, d\hat{\theta} = \int_{\theta_a}^{\theta_b} p(\hat{\theta}|\theta)\, d\hat{\theta} + \int_{\theta_b}^{\theta'_b} p(\hat{\theta}|\theta)\, d\hat{\theta} = \beta + \int_{\theta_b}^{\theta'_b} p(\hat{\theta}|\theta)\, d\hat{\theta}, \tag{6.16}$$

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

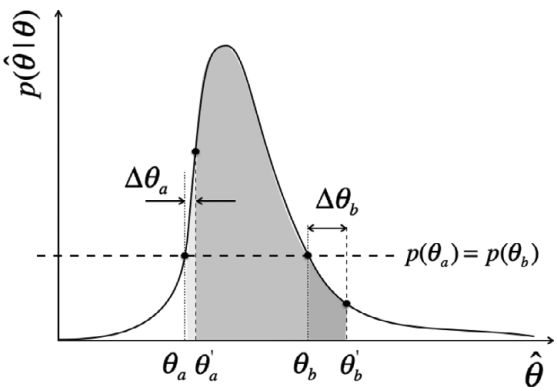


Fig. 6.2 Given a specific probability content β and an asymmetric PDF $p(\hat{\theta}|\theta)$, a minimal interval $[\theta_a, \theta_b]$ is obtained for values θ_a and θ_b that simultaneously satisfy Eqs. (6.14, 6.15).

or alternatively,

$$\int_{\theta_a}^{\theta_b'} p(\hat{\theta}|\theta)\, d\hat{\theta} = \int_{\theta_a}^{\theta_a'} p(\hat{\theta}|\theta)\, d\hat{\theta} + \int_{\theta_a'}^{\theta_b'} p(\hat{\theta}|\theta)\, d\hat{\theta} = \int_{\theta_a}^{\theta_a'} p(\hat{\theta}|\theta)\, d\hat{\theta} + \beta. \tag{6.17}$$

We thus find that

$$\int_{\theta_a}^{\theta_a'} p(\hat{\theta}|\theta)\, d\hat{\theta} = \int_{\theta_b}^{\theta_b'} p(\hat{\theta}|\theta)\, d\hat{\theta}. \tag{6.18}$$

Invoking the mean value theorem, the preceding equality may be written

$$p(\hat{\theta}_{a,0}|\theta)\Delta\theta_a = p(\hat{\theta}_{b,0}|\theta)\Delta\theta_b, \tag{6.19}$$

where we defined $\Delta\theta_a = \theta_a' - \theta_a$ and $\Delta\theta_b = \theta_b' - \theta_b$, while $\hat{\theta}_{a,0}$ and $\hat{\theta}_{b,0}$ are values found in the intervals $[\theta_a, \theta_a']$ and $[\theta_b, \theta_b']$, respectively. Based on the PDF construction shown in Figure 6.2, one knows that the ratio of the densities at $\theta_{a,0}$ and $\theta_{b,0}$ satisfies

$$\frac{p(\theta_{a,0})}{p(\theta_{b,0})} > 1. \tag{6.20}$$

In order to maintain the equality Eq. (6.19), one must then have $\Delta\theta_b > \Delta\theta_a$ which implies that

$$\theta_b' - \theta_a' > \theta_b - \theta_a. \tag{6.21}$$

We thus conclude that the interval $[\theta_a, \theta_b]$ is indeed the most compact interval with a probability content β, and as such provides a unique and consequently meaningful choice of central interval. For PDFs symmetric about their mean μ, such as Gaussian distributions, Eq. (6.15) defines a symmetric interval of width 2Δ such that

$$p(\mu - \Delta) = p(\mu + \Delta), \tag{6.22}$$

which, evidently, amounts to a central and most compact interval.

It is important to acknowledge that the notion of shortest interval is not strictly well defined for continuous variables because it is not invariant under a reparameterization

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

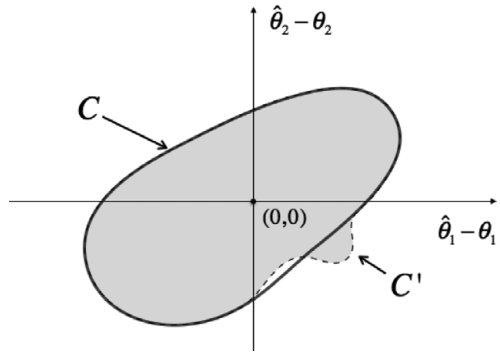


Fig. 6.3 Given a specific probability content β and an asymmetric PDF $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2)$, a minimal confidence region $A_{\min}$ is obtained for values $\hat{\theta}_1$ and $\hat{\theta}_2$ bound within an iso-contour C of probability density $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2) = k$ shown as a solid line. The dashed line illustrates a nonisometric contour C' (i.e., a contour with nonuniform probability density), which, by construction, features an area $A > A_{\min}$.

$\theta \to \xi(\theta)$. Central intervals, defined by Eq. (6.13), are invariant under such transformation, however, and are thus properly defined.

6.1.5 Multiparametric Confidence Intervals: Confidence Regions

The concepts and techniques presented in the previous two subsections are readily extended to multiparametric systems, that is, optimization problems involving $n \geq 2$ unknown parameters. For instance, let us consider an experiment producing n estimators, $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n)$ of model parameters $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$. Assuming the probability model of the data is known, one can write

$$\beta \equiv P(\hat{\theta} \in V) = \int_V p(\hat{\theta}|\vec{\theta}) \prod \hat{\theta}_i, \tag{6.23}$$

where V is an n dimensions hyper-volume bounded by an $n-1$ dimensions hypersurface C defining a **confidence region** for the parameters $\hat{\theta}$ given the true values $\vec{\theta}$. For $n = 2$, V reduces to a two-dimensional bounded surface, and C amounts to a closed contour surrounding V, as illustrated in Figure 6.3. For $n = 3$, C is a closed surface, while for $m > 3$, it amounts to a closed hypersurface.

In two or more dimensions, $n \geq 2$, just as for a single dimension, there is much latitude in selecting boundaries with a specific probability content β. It is legitimate, however, to seek a bounding hypersurface (contour in 2D, surface in 3D, etc.) that circumscribes the smallest "volume" with probability content β. The task of selecting such an enclosing surface is schematically illustrated in Figure 6.3, which for the sake of simplicity is limited to $n = 2$ parameters. The solid curve here represents an iso-contour line corresponding to a locus of constant probability density, that is, where all points $(\hat{\theta}_1, \hat{\theta}_2)$ have the same probability density $p(\hat{\theta}_1, \hat{\theta}_2|\vec{\theta}) = k$, with k chosen such that the bounded area (i.e., the confidence region) has a probability β.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

Without repeating in detail the reasoning carried out for a single parameter, it is clear that any contour (with probability content β) deviating from the iso-contour would have a larger area than that of the iso-contour. The condition $p(\hat{\theta}_1, \hat{\theta}_2|\vec{\theta}) = k$ thus defines, for a given β, the bounded region with the least area. The density $p(\hat{\theta}_1, \hat{\theta}_2|\vec{\theta})$ being continuous, the contour is well defined and unique, it thus constitutes a legitimate and sensible choice of bounding region. The reasoning is readily extended in $n \geq 3$ dimensions: bounding surfaces or hypersurfaces may be chosen as iso-surfaces, that is, hypersurfaces corresponding to loci of equal probability density, $p(\hat{\theta}|\vec{\theta}) = k$, where the constant k is determined by the required value of β.

6.1.6 Confidence Region for a Multiparametric Gaussian Distribution

Let us consider a system determined by n parameters $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ for which one has obtained n estimates $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n)$. For illustrative purposes, let us assume the PDF of the estimators may be formulated as a multidimensional Gaussian distribution

$$p(\hat{\theta}|\vec{\theta}) = \frac{1}{(2\pi)^{n/2}|\mathbf{V}|^{1/2}} \exp\left[-\frac{1}{2}(\hat{\theta} - \vec{\theta})^T \mathbf{V}^{-1}(\hat{\theta} - \vec{\theta})\right], \tag{6.24}$$

with covariance matrix $\mathbf{V}$. We are interested in using the estimates $\hat{\theta}$ to establish bounds on the model parameters $\vec{\theta}$. In general, the covariance matrix $\mathbf{V}$ shall be nondiagonal, thereby revealing correlations between the measured estimates $\hat{\theta}_i$. The bounds on each of the model parameters θ_i, $i = 1, \ldots, n$ will then also be correlated. Errors and confidence intervals thus cannot be defined independently for each parameter. Rather than calculating disjoint confidence intervals for all parameters, one must then define and calculate **confidence regions**. Recall from §3.10 that a multiparametric Gaussian distribution can always be transformed in terms of orthogonal coordinates corresponding to the principal axes of the covariance matrix $\mathbf{V}$. The distribution is thus symmetric relative to these principal axes, and based on the discussion on central regions presented in the previous section, one would like to define central confidence regions in terms of iso-contours with probability content β.

Iso-contours with probability content β are readily defined by recalling, from §2.10.4, that the **quadratic form**

$$Q(\hat{\theta}|\vec{\theta}) = (\hat{\theta} - \vec{\theta})^T \mathbf{V}^{-1}(\hat{\theta} - \vec{\theta}) \tag{6.25}$$

has a χ^2 distribution with n degrees of freedom. This implies Q is in fact independent of $\vec{\theta}$. We can then determine the probability β that Q will be found within a domain κ_β^2 based on

$$P\left(Q(\hat{\theta}|\vec{\theta}) \leq \kappa_\beta^2\right) = \beta, \tag{6.26}$$

where the bound κ_β^2 corresponds to the β-point of a χ^2 distribution with n degrees of freedom. By construction, κ_β^2 delimits a hyper-ellipsoidal region in the n-dimensional space of the estimators $\hat{\theta}$ and model parameters $\vec{\theta}$ with probability β:

$$Q(\hat{\theta}|\vec{\theta}) \leq \kappa_\beta^2. \tag{6.27}$$

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

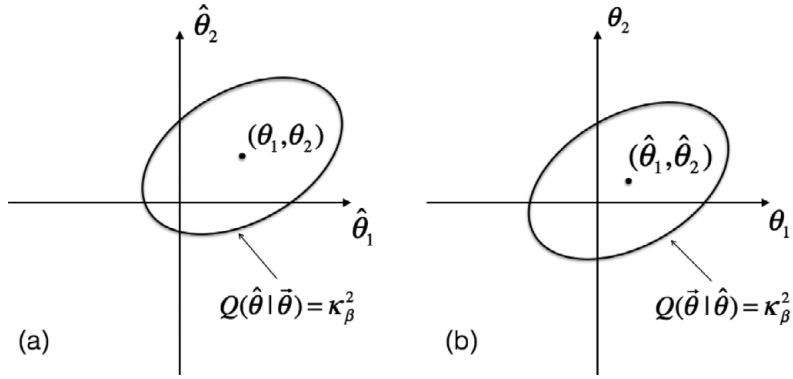


Fig. 6.4 Confidence region of probability β in (a) the space of estimators $\hat{\theta}$ and (b) the model parameter space $\vec{\theta}$ for a two-dimensional Gaussian PDF.

If the model parameter values $\vec{\theta}$ are known a priori, the foregoing condition delimits a hyper-ellipsoidal region in the n-dimensional space of the estimators $\hat{\theta}$, illustrated schematically in Figure 6.4a for a case involving $n = 2$ model parameters. The probability of measuring $\hat{\theta}$ within the ellipsoidal boundary is β, by construction. In the context of an experiment, $\hat{\theta}$ is measured, and then one seeks to invert the above bound to define a bounding region for the model parameters $\vec{\theta}$. By virtue of the symmetry of Q in terms of variables $\hat{\theta}$ and $\vec{\theta}$, this leads to an ellipsoidal domain surrounding the measured values $\hat{\theta}$, as shown schematically in Figure 6.4b.

We proceed to illustrate the quantitative determination of a confidence region in two dimensions, but the technique is readily extended to problems involving more than two parameters. We express the covariance matrix $\mathbf{V}$ as

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \tag{6.28}$$

where σ_1^2 and σ_2^2 are the variances of $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively, while ρ expresses the Pearson correlation coefficient between $\hat{\theta}_1$ and $\hat{\theta}_2$. Calculation of the inverse of $\mathbf{V}$ yields

$$\mathbf{V}^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}. \tag{6.29}$$

Defining $\Delta\theta_i = \hat{\theta}_i - \theta_i$, the quadratic form (6.25) may then be expressed as

$$\sigma_2^2\Delta\theta_1^2 - 2\rho\sigma_1\sigma_2\Delta\theta_1\Delta\theta_2 + \sigma_1^2\Delta\theta_2^2 = \kappa_\beta^2\sigma_1^2\sigma_2^2(1-\rho^2), \tag{6.30}$$

which is indeed the equation of an ellipse, as illustrated in Figure 6.5.

Figure 6.5 actually displays three distinct regions. The ellipse is defined by the quadratic form (6.25) and has a probability content of β_1, which means that the probability of measuring a pair of values $(\hat{\theta}_1, \hat{\theta}_2)$ within the ellipse is β_1. The rectangular region is defined by the condition

$$P(|\hat{\theta}_1 - \theta_1| < \kappa_\beta\sigma_1 \text{ and } |\hat{\theta}_2 - \theta_2| < \kappa_\beta\sigma_2) = \beta_2, \tag{6.31}$$

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

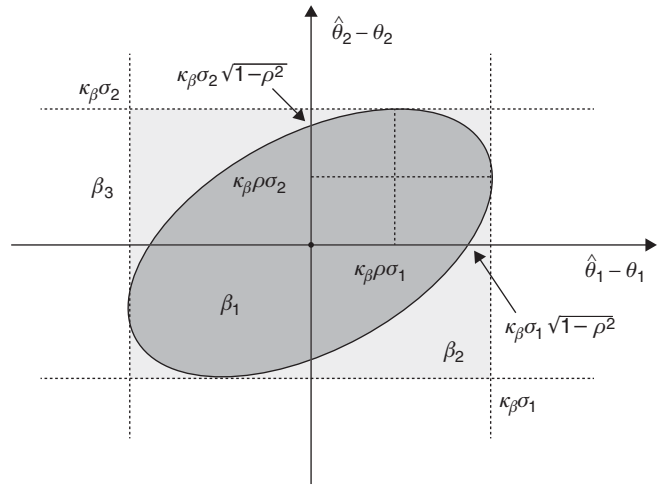


Fig. 6.5 Examples of confidence regions for specific values of σ_1, σ_2, and ρ plotted for arbitrary choices of probability content β_1, β_2, and β_3. See text for definitions of these probabilities.

and has a probability content $\beta_2 > \beta_1$. The third region is delimited by the long horizontal lines and has probability content

$$P(|\hat{\theta}_2 - \theta_2| < \kappa_\beta\sigma_2) = \beta_3, \tag{6.32}$$

which is by construction larger than β_2. One must consequently be careful when quoting error intervals of observables obtained in joint measurements. If the error ellipse can be explicitly reported, it is then possible to quote the region it delimits as a $100 \times \beta_1$ % confidence region. If instead the errors on θ_1 and θ_2 are reported as independent confidence intervals defined by the bounds $\kappa_\beta\sigma_1$ and $\kappa_\beta\sigma_2$ independently, the confidence region then has probability content β_2 and should thus be quoted as such. Finally, if the value of $\hat{\theta}_1$ is ignored and the error on $\hat{\theta}_2$ is reported based on $\kappa_\beta\sigma_2$, than it constitutes a $100 \times \beta_3$ % confidence region.

A practical example of application of these concepts is discussed in Problem 6.1.

6.1.7 Confidence Interval for Discrete Distributions

The technique described in the previous sections applies for the determination of confidence intervals of Gaussian continuous deviates and can be readily extended to any continuous variable distributions using, for instance, the Neyman construction presented in §6.1.8. It must, however, be slightly modified for distributions of discrete random variables.

Let us consider a measurement whose outcome is an integer value n_o determined by a discrete probability distribution $p(n|\theta)$ expressing the probability of measuring a positive definite integer value n given some model parameter θ. One wishes to use the measured value n_o to establish a confidence interval $\theta_\alpha \leq \theta \leq \theta_\beta$ with a specific confidence level CL.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

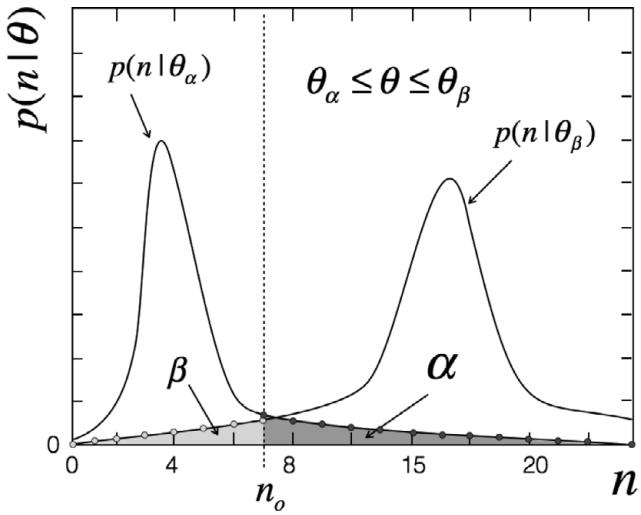


Fig. 6.6 Determination of confidence intervals for discrete PDFs: one seeks parameters θ_α and θ_β such that the probability of observing n with values equal and larger (smaller) to n_o is equal to α and β, respectively.

As for continuous intervals, much flexibility exists in defining such an interval. However, in contrast to continuous variables, one must define the interval on the basis of probabilities of observing values of n smaller and larger than n_o, as illustrated in Figure 6.6. Specifically, one must define an upper bound θ_β by calculating the probability β that n be equal or smaller than the observed value n_o and a lower bound θ_α by calculating the probability α that n be equal or greater than the observed value n_o and such that $\mathrm{CL} = 1 - \alpha - \beta$ according to

$$\alpha = P(n \geq n_o|\theta_\alpha), \tag{6.33}$$

$$\beta = P(n \leq n_o|\theta_\beta). \tag{6.34}$$

Assuming n is positive definite, the probability α may be calculated according to

$$\alpha = \sum_{n=n_o}^{\infty} p(n|\theta_\alpha) = 1 - \sum_{n=0}^{n_o-1} p(n|\theta_\alpha), \tag{6.35}$$

while β is given by

$$\beta = \sum_{n=0}^{n_o} p(n|\theta_\beta). \tag{6.36}$$

We present an application of these rules for Poisson distribution in the next section. Intervals for binomial distributions are discussed in §6.1.7.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

Confidence Interval for a Poisson Distribution

As a first example of confidence interval for a discrete distribution, let us consider the estimation of confidence intervals for the mean of a Poisson distribution:

$$p_P(n|\lambda) = \frac{\lambda^n}{n!}e^{-\lambda}. \tag{6.37}$$

We saw in §3.3 that the Poisson distribution has equal mean and variance, that is, $\mathrm{E}[n] = \mathrm{Var}[n] = \lambda$. It is also possible to verify that for a given single measurement of n noted n_o, the $\mathbb{ML}$ estimator of λ is equal to n_o. Indeed, an extremum of $p_P(n|\lambda)$ given $n = n_\mathrm{o}$ is obtained for

$$0 = \left.\frac{\partial p_P(n|\lambda)}{\partial \lambda}\right|_{n=n_\mathrm{o}} = \left.\frac{e^{-\lambda}}{n!}\left(n\lambda^{n-1} - \lambda^n\right)\right|_{n=n_\mathrm{o}}, \tag{6.38}$$

which yields the stated result $\hat{\lambda}_\mathrm{o} = n_\mathrm{o}$. Although the actual value of λ may be any real number ($\lambda \in \mathbf{R}$), the $\mathbb{ML}$ estimate $\hat{\lambda}_\mathrm{o}$ obtained from a single measurement is by construction an integer. One is thus in a situation in which one is trying to establish a confidence interval $\lambda_\alpha \leq \lambda \leq \lambda_\beta$ for a continuous variable based exclusively on an integer value n_o. We thus apply Eqs. (6.35, 6.36) and write

$$\alpha = 1 - \sum_{n=0}^{n_\mathrm{o}-1} \frac{\lambda_\alpha^n}{n!}e^{-\lambda_\alpha}, \tag{6.39}$$

$$\beta = \sum_{n=0}^{n_\mathrm{o}} \frac{\lambda_\beta^n}{n!}e^{-\lambda_\beta}, \tag{6.40}$$

which, given values α and β, can be solved for both λ_α and λ_β. As noted in ref. [67] §9.4, this can be accomplished using the identity

$$\sum_{k=0}^{n} \frac{\lambda^k}{k!}e^{-\lambda} = \int_{2\lambda}^{\infty} p_{\chi^2}(z|n_d)\,dz = 1 - P\chi^2(2\lambda|n_d), \tag{6.41}$$

where $p_{\chi^2}(z|n_d)$ represents the χ^2 PDF, given by Eq. (3.163), for $n_d = 2(n+1)$ degrees of freedom and $P_{\chi^2}(2\lambda|n_d)$ is its cumulative integral evaluated at $z = 2\lambda$. The values λ_α and λ_β sought for may then be calculated according

$$\lambda_\alpha = \tfrac{1}{2}P^{-1}_{\chi^2}(\alpha|n_d), \tag{6.42}$$

$$\lambda_\beta = \tfrac{1}{2}P^{-1}_{\chi^2}(1-\beta|n_d), \tag{6.43}$$

with $n_d = 2(n_\mathrm{o}+1)$. Given an observed value $\hat{\lambda}_\mathrm{o} = n_\mathrm{o}$, the probabilities of observing specific ranges of λ can then be characterized according to

$$P(\lambda \geq \lambda_\alpha) \geq 1 - \alpha, \tag{6.44}$$

$$P(\lambda \leq \lambda_\beta) \geq 1 - \beta, \tag{6.45}$$

$$P(\lambda_\alpha \leq \lambda \leq \lambda_\beta) \geq 1 - \alpha - \beta. \tag{6.46}$$

Obviously, a lower limit λ_α cannot be meaningfully defined if $n_\mathrm{o} = 0$, and one can only establish an upper limit λ_β. One can then seek the largest λ_β consistent with the observed

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

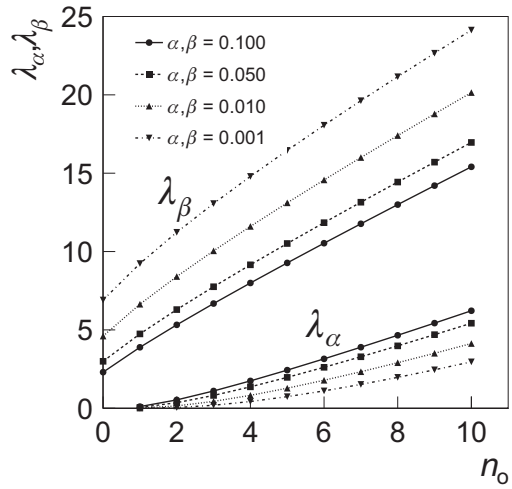


Fig. 6.7 Examples of lower λ_α and upper λ_β Poisson limits for values α, $\beta = 0.01, 0.05, 0.01, 0.001$ as a function of the observed number of counts.

null value for a specific probability β and write

$$\beta = P(n_o = 0|\lambda_\beta) = \sum_{k=0}^{0} \frac{\lambda_\beta^k}{k!} e^{-\lambda_\beta} = e^{-\lambda_\beta}, \tag{6.47}$$

which yields $\lambda_\beta = -\ln\beta$. For example, requiring the probability of $n_o = 0$ to be 1% one obtains a 99% confidence interval with an upper limit of $\lambda_\beta = -\ln(0.01) \approx 4.605$, thereby setting an upper limit of approximately $\lambda_\beta = 4.6$ given the observed null value. Choosing instead a 99.9% confidence level, one would get an upper limit of $\lambda_\beta = 6.9$. This means that the value of λ could be as large as 6.9 and there would be a probability of 0.001 of observing $n = 0$ counts (of whatever entity is being sought for). Limits λ_α and λ_β are shown in Figure 6.7 for $\alpha, \beta = 0.01, 0.05, 0.01, 0.001$ as a function of the observed number of counts in the range $0 \leq n_o \leq 10$. Note that significant difficulties may arise in the determination of an upper limit if the measurement of a signal n_o involves subtraction of a background. Since the background may fluctuate, n_o may be negative, and one then deals with an empty confidence interval. Evidently, as for continuous variables in the presence of a physical boundary, one also runs the risk of flip-flopping and incomplete coverage: use of the Neyman construction, discussed in §6.1.8, should thus be considered whenever practical.

Confidence Interval for a Binomial Distribution

For large sample sizes N, the determination of error intervals for the fraction ϵ of the binomial distribution, Eq. (3.1), may be obtained with a normal approximation. Recall that the mean and variance of the binomial distribution are respectively $N\epsilon$ and $N\epsilon(1-\epsilon)$ for a sample size N determined by a binomial fraction ϵ. One can then use an observed number

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

of successes n, given N trials, to determine an estimate $\hat{\epsilon} = n/N$ of the binomial fraction. And for a reasonably large sample size N, an $100 \times \alpha\%$ confidence interval can then be approximated by

$$\hat{\epsilon} \pm z_{1-\alpha/2}\sqrt{\hat{\epsilon}(1-\hat{\epsilon})/N}, \tag{6.48}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution. For instance, for a 95% confidence level, one has $\alpha = 0.05$ and $1-\alpha/2 = 0.975$, which yields $z = 1.96$.

The normal approximation relies on the central limit theorem, valid for large values of sample size N, and thus yields poor results for small N. The approximation also fails if the estimate $\hat{\epsilon}$ is identically zero or unity since it then yields a null interval. Brown et al. [58] discussed a simple rule of thumb which states that the normal approximation is reasonable when $\epsilon N > 5$ and $N(1-\epsilon) > 5$. Several other approximations are discussed in the literature for such situations: they include the Wilson score interval [196], the Wilson score interval with continuity correction [148], the Jeffreys interval based on the Bayesian credible interval obtained when using a noninformative Jeffreys' prior (see §7.3.2), the Clopper–Pearson interval [65], and various other approximations. Arguably, binomial intervals are best obtained with the unified approach discussed in §§6.1.8 and 6.2.2.

6.1.8 Confidence Interval: The General Case

The Neyman Construction

In previous sections, we described basic techniques for the determination of confidence intervals and confidence regions when measured variables are Gaussian distributed with expectation values equal to the model parameters. While the central limit theorem implies that a vast majority of practical applications involve physics parameters that fluctuate with Gaussian distributions, cases are nonetheless encountered in which parameters fluctuate according to other types of distributions that cannot be approximated by Gaussian PDFs. It is thus desirable to have a general technique, applicable to any PDFs, to determine confidence intervals and confidence regions of model parameters. Such a technique was first developed by Jersey Neyman [149] and is now commonly known as the **Neyman construction**.

In order to introduce the Neyman construction, let us consider a single parameter estimator defined as a function $\hat{\theta} \equiv \hat{\theta}(\vec{x})$ of measured data $\vec{x}$ with a PDF $p(\hat{\theta}|\theta)$, which may be non-Gaussian. For a known value of θ, one may calculate the probability β that a specific estimate $\hat{\theta}$ be found in some interval $[\hat{\theta}_1, \hat{\theta}_2]$ according to

$$\beta = P(\hat{\theta}_1 \leq \hat{\theta} \leq \hat{\theta}_2|\theta) = P(\hat{\theta}_1(\theta) \leq \hat{\theta} \leq \hat{\theta}_2(\theta)) \tag{6.49}$$

$$= \int_{\hat{\theta}_1}^{\hat{\theta}_2} p(\hat{\theta}|\theta)\, d\hat{\theta}. \tag{6.50}$$

Obviously, as for the Gaussian case discussed in §6.1.3, there are arbitrarily many values $\hat{\theta}_1(\theta)$ and $\hat{\theta}_2(\theta)$ that satisfy this condition. However, it is convenient and most common to

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

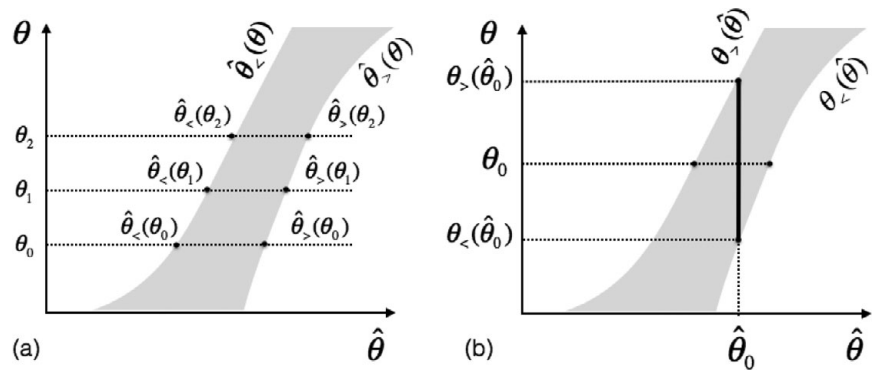


Fig. 6.8 (a) A confidence belt is constructed horizontally by finding the values $\theta_<(\theta)$ and $\theta_>(\theta)$ that satisfy Eq. 6.49. (b) The confidence level is obtained vertically as the range $[\theta_<(\hat{\theta}_0), \theta_>(\hat{\theta}_0)]$ corresponding to a measured value $\hat{\theta}_0$.

define a **central interval** according to

$$\frac{1-\beta}{2} = \int_{-\infty}^{\hat{\theta}_1} p(\hat{\theta}|\theta)\, d\hat{\theta}, \tag{6.51}$$

$$\frac{1-\beta}{2} = \int_{\hat{\theta}_2}^{\infty} p(\hat{\theta}|\theta)\, d\hat{\theta}, \tag{6.52}$$

but one might also use **upper limits**, **lower limits**, or the so-called **unified approach** discussed in §6.1.8.

Let us assume we have values $\hat{\theta}_<(\theta)$ and $\hat{\theta}_>(\theta)$ that satisfy Eq. (6.49). By construction, these are functions of the model parameter value θ and can be plotted as such, as illustrated in Figure 6.8a. The region (shaded area) between the functions $\hat{\theta}_<(\theta)$ and $\hat{\theta}_>(\theta)$ is called a **confidence belt**. It is defined such that for any specific value θ (vertical axis), the interval $[\hat{\theta}_<(\theta), \hat{\theta}_>(\theta)]$ has a probability content β defined by

$$P(\hat{\theta}_<(\theta_0) \le \hat{\theta} \le \theta_>(\theta_0)) = \beta. \tag{6.53}$$

The width of the belt along the $\hat{\theta}$ axis is a function of the value θ, which depends on the behavior of the function $p(\hat{\theta}|\theta)$. Our goal is to identify an interval $[\theta_<, \theta_>]$ for θ that has the same probability content β given a measured value of $\hat{\theta}$, as illustrated in Figure 6.8b. If the true but unknown value of the model parameter θ were θ_0, one would expect that a large ensemble of measurements should yield values in the range $\hat{\theta}_<(\theta_0) \le \hat{\theta} \le \theta_>(\theta_0)$ with probability β. Assuming that $\hat{\theta}_<(\theta)$ and $\hat{\theta}_>(\theta)$ are both monotonically increasing functions of the parameter θ, as they should if $\hat{\theta}$ is a good estimator, one can then consider their inverse functions and write

$$\theta_< \equiv \theta_<(\hat{\theta}_0) = \hat{\theta}_>^{-1}(\hat{\theta}_0) \tag{6.54}$$

$$\theta_> \equiv \theta_>(\hat{\theta}_0) = \hat{\theta}_<^{-1}(\hat{\theta}_0) \tag{6.55}$$

We can then invert the inequalities in Eq. (6.53):

$$P(\theta_<(\hat{\theta}_0) \le \theta_0 \le \theta_>(\hat{\theta}_0)) = \beta, \tag{6.56}$$

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

and conclude that, given a measured value $\hat{\theta}_0$, the interval $[\theta_<(\hat{\theta}_0), \theta_>(\hat{\theta}_0)]$ is guaranteed, on average, to cover (i.e., include or contain) the true value θ_0 and thus constitutes a confidence interval of the parameter θ_0 with probability β.

It is important to first note that given the estimate $\hat{\theta}_0$ is a random value, the bounds $\theta_<(\hat{\theta}_0)$ and $\theta_>(\hat{\theta}_0)$ are by construction also random values. The bound inversion, Eq. (6.54), indeed yields an estimate of the confidence interval that varies measurement by measurement: just as $\hat{\theta}_0$ fluctuate measurement by measurement, so do the bounds defining the confidence interval. However, the Neyman construction and Eq. (6.56) guarantee that the interval $[\theta_<(\hat{\theta}_0), \theta_>(\hat{\theta}_0)]$ has a probability content β. This implies that if the measurement is repeated a large number of times N, the N confidence intervals obtained from these measurements should cover the actual value of the parameter with probability β (on average or in the large N limit).

One should also note that if the functions $\theta_<(\hat{\theta}_0)$ and $\theta_>(\hat{\theta}_0)$ are not monotonically increasing with $\hat{\theta}_0$, the construction is still usable but one may end up with confidence intervals that are not *simply connected*, that is, which consist of several disconnected ranges.

One-Sided Confidence Intervals

One-sided confidence intervals are of interest when a measured physical quantity lies near a boundary it cannot exceed and lead to lower and upper limit confidence intervals. For instance, the mass of a particle cannot be negative but could be vanishingly small or even null. Defining a central interval in this context might become problematic because part of the confidence interval might extend across the physically forbidden region. Indeed, if the mass is small, or possibly null, it would be preferable to define a confidence level ranging from zero up to some upper limit $m_{\max}$. Upper limits are also of interest when searching for rare phenomena, including, for instance, rare decays of known particles and upper limits on the production cross section of predicted particles whose existence is yet to be established.

Figure 6.9 illustrates the difference between the confidence belt associated with a central interval and a one-sided interval, in this case an upper limit. The intervals are based on a Gaussian distribution with unit standard deviation. The solid lines represent the lower and upper boundaries of the Neyman construction for a 90% confidence belt assuming the model parameter θ cannot be smaller than zero. The dashed line shows the 90% lower limit on the measured value $\hat{\theta}$ given any value of θ. This line translates into a 90% **upper limit** confidence interval on θ. For instance, for a measured value $\hat{\theta} = 1$, one would conclude that θ has a 90% probability of being smaller than ~ 2.2, whereas the central limit would imply that there is a finite probability for the mean to lie in the negative nonphysical region. Clearly, the use of an upper limit is more appropriate in this case. However, if the measured value is $\hat{\theta} = 4$, the upper limit could only tell us that the mean has an upper bound of ~ 5.2 with probability of 90% and no statement about a minimal value. It thus appears appropriate to switch methods for the definition of the confidence interval based on the observed value. It is in fact common practice to use an upper-limit interval if the measured value is within 3σ of the origin (or nonphysical region) and a central (symmetric) interval, otherwise. Unfortunately, this switching of methods, sometimes called **flip-flopping**, leads

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

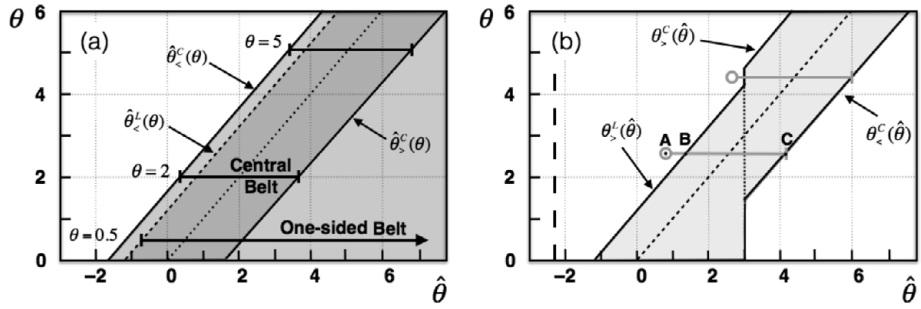


Fig. 6.9 (a) The solid lines $\hat{\theta}^C_<$ and $\hat{\theta}^C_>$ represent the boundaries of a 90% central interval and the dashed lines $\hat{\theta}^L_<$ the lower limit of a 90% one-sided interval, for a standard normal distribution of the parameter θ. (b) The shaded gray region represents the effective confidence belt obtained when flip-flopping between a central limit and an upper limit **after** a measurement $\hat{\theta}$ has been taken. The effective belt lacks coverage for model values in the range $1.35 \leq \mu \leq 4.65$, as illustrated with the light gray segments. The vertical long dashed line represents a measurement $\hat{\theta} = -2.2$ with no allowed physical value and resulting in an empty confidence interval.

to loss of coverage and inconsistent confidence interval definitions, which we describe in the next section.

Flip-flopping, Unphysical Values, and Empty Intervals

The loss of coverage occurring with flip-flopping of confidence belts is illustrated in Figure 6.9(b). The gray area represents the confidence belt obtained when using an **upper-limit method** for a measurement $\hat{\theta}$ within 3σ of the origin or nonphysical region, and switching to a **central interval method** for values larger than 3σ. The flip-flop belt has full coverage for values of $\theta \leq 1.35$ and $\theta > 4.65$ but lacks coverage in the intermediate range. The horizontal gray segments provide examples where the parameter θ lacks coverage. For $\theta = 2.6$, the intervals AC and $B\infty$ each contain 90% of the probability but the interval BC contains only 85% of the probability. There is consequently a lack of coverage for this and other values of θ in the range [1.35, 4.65]. Confidence intervals thus cannot be quoted consistently with the flip-flop method.

Figure 6.9b illustrates an additional problematic issue encountered with confidence belts near the boundary of a physical region. While the model parameter θ is assumed to be non-negative, a measurement $\hat{\theta}$ of θ may nonetheless yield a negative value. For $\theta = 0$, the 90% central confidence interval amounts to $-1.65 \leq \hat{\theta} \leq 1.65$. If the measurement is repeated a large number of times, one expects that 5% of the measured values will fall below -1.65. Although such values have a smaller probability, they are nonetheless possible, even if the model parameter θ cannot be negative. The problem arises that a vertical line drawn anywhere in this range, say at $\hat{\theta} = -2$, would not intersect the chosen confidence belt in the physical region of θ. Extrapolation into a nonphysical confidence interval for θ obviously does not make sense. One is then dealing with an empty confidence interval for that value, that is, an interval that contains no physically allowed values of θ. So, although the 90% upper limit method has exact coverage, it should nonetheless be avoided since it cannot be used to produce a meaningful confidence interval for θ whenever measurements of $\hat{\theta}$ fall below the 90% confidence interval.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

Fortunately, issues of lack of coverage encountered with flip-flopping and empty intervals obtained with one-sided limits may both be remedied with the unified approach discussed in the next section.

Unified Approach for the Definition of Confidence Intervals

The Neyman construction requires that the integral Eq. (6.49) yield the probability β but it does not stipulate which elements should be part of the interval. It is in fact possible to choose which elements of probability should be accepted in the interval. This leads to the notion of an ordering principle.

The purpose of this ordering principle is to produce an optimal confidence belt, with full coverage, no issue associated with flip-flopping, and no empty intervals. Feldman and Cousins found an ordering principle that satisfies these requirements and has now become the basis for the unified approach mentioned earlier [80]. Known as the **likelihood ratio ordering principle**, the ordering is based on the Neyman–Pearson test (discussed in §6.7) and sorts elements of the measurable parameter space on the basis of a likelihood ratio. The construction of intervals for each value of the model parameter θ proceeds by including the elements of probability $P(\hat{\theta}|\theta)$ that maximize the likelihood ratio

$$R(\hat{\theta}) = \frac{P(\hat{\theta}|\theta_0)}{P(\hat{\theta}|\theta_{\text{Best}})}, \tag{6.57}$$

where θ_{Best} corresponds to the value of θ with maximum likelihood $P(\hat{\theta}|\theta)$ for the measured $\hat{\theta}$, while θ_0 is the values one wishes to test for. For a Gaussian distribution, the Feldman and Cousins method yields a confidence belt as illustrated schematically by the dark solid lines in Figure 6.10 (see §6.2.3 for a formal derivation of this result). This unified approach belt is identical to the 90% central interval belt for values of $\theta > 1.65$ but significantly deviates from it below that value. For small values of θ, the most likely θ_{Best} is at zero. The different ordering then produces a belt that extends toward negative values of $\hat{\theta}$. This long negative tail prevents the occurrence of empty intervals that would otherwise occur for $\hat{\theta} < -1.65$. Consistent 90% coverage entails that the high side of the belt must shift to the left relative to the central interval. This leads to the small deviation seen in Figure 6.10 on the high side of the belt. The method then not only eliminates empty intervals but also avoids flip-flopping and provides proper and consistent coverage.

We present calculations of the Felman and Cousins ordering principle applied to Poisson, binomial, and Gaussian processes in §6.2.

6.2 Confidence Intervals with the Unified Approach

6.2.1 Poisson Intervals

Measurements of rare or predicted but as of yet unobserved decay processes constitute a perfect exemplar for the calculation of Poisson intervals: a measurement is designed

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

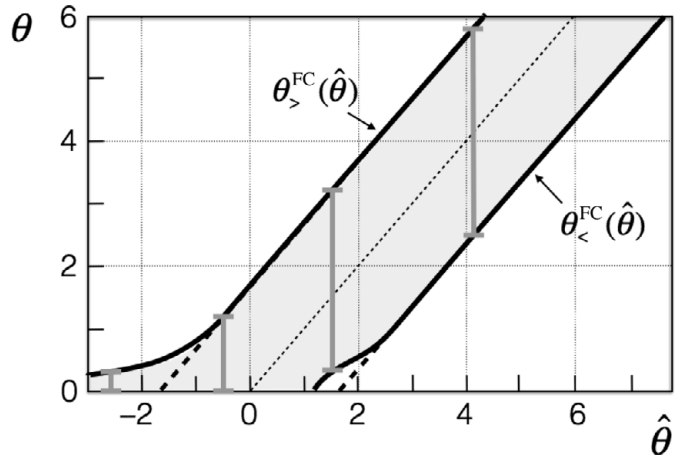


Fig. 6.10 Schematic representation of the 90% confidence belt (light gray) obtained with the Feldman–Cousins unified approach for a Gaussian distribution of unit standard deviation compared to a 90% central interval confidence belt (dashed lines). Dark gray vertical segments represent confidence intervals corresponding to measured values $\hat{\theta} = -2.6, -0.5, 1.5,$ and 4.1.

to observe and count the number of decays of a certain particle. Given instrumental and physical process ambiguities, it is possible that the observed decays might originate from other improperly measured physical processes, or decays, as well as random combination of particles mistakenly interpreted as decays. Observed counts of decays thus amount to a combination of a true signal (of unknown strength) and one or several background signals. We here proceed to determine the confidence belt for such Poisson processes with known background using the Neyman construction and the ordering principle of Feldman and Cousins [80].

For a Poisson process with a known background, one can write the probability of observing n events (or counts) as

$$P(n|\mu, b) = \frac{(\mu + b)^n \exp(-(\mu + b))}{n!}, \tag{6.58}$$

where b is the expected (known) background and μ is the unknown mean yield of the Poisson process under consideration. We will assume, for illustrative purposes, that the mean of the background is known to have a specific value.

We build confidence intervals for each value of μ by ordering probabilities of different values of n based on the ratio of likelihoods

$$R(n, \mu) = \frac{P(n|\mu, b)}{P(n|\mu_{\text{best}}, b)}, \tag{6.59}$$

where $P(n|\mu, b)$ is the probability of observing n counts given an actual mean μ and background b, whereas $P(n|\mu_{\text{best}}, b)$ corresponds to the probability of n for the value of $\mu + b$ that yields the largest probability of observing n counts. The construction proceeds as follows. First, one selects ranges of interest for n and μ for a given value of b. Both ranges are divided in finitely many bins with a unit bin width for n (since it is an integer variable)

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

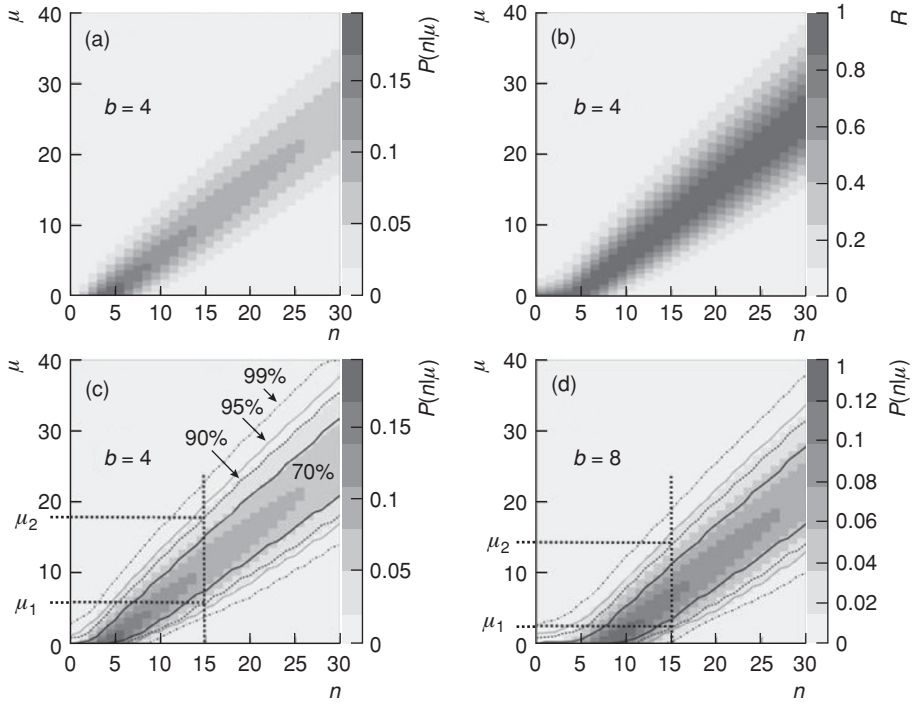


Fig. 6.11 Determination of confidence intervals for Poisson processes with known finite background b. (a) Probability $P(n|\mu, b = 4)$. (b) Ordering likelihood ratio $R(n, \mu|b = 4)$. (c) Confidence intervals calculated with $b = 4$ for CL 70%, 90%, 95%, and 99%. (d) The same confidence intervals calculated with $b = 8$. The 90% confidence intervals for $n = 15$ amount to $[\mu_1, \mu_2] = [5.6, 18.0]$ with $b = 4$ and $[2.2, 14.1]$ for $b = 8$.

and a bin width of order 0.005 for μ. One calculates and stores the probability $P(n|\mu, b)$ in each bin (using the low side of the bin for μ) into an array, or a two-dimensional histogram as illustrated in Figure 6.11a. For each value of n, one scans probabilities vs. μ to find the maximum $P(n|\mu_{\text{best}}, b)$ which can then be used to compute the ratio $R(n, \mu)$ in each bin n, μ. This yields a two-dimensional array, shown as a histogram in Figure 6.11b, which can be used to carry out the ordering of probabilities and the construction of the confidence interval for each value of μ. This is achieved by sorting values of the ratio R in descending order, for each bin in μ. Starting with the most probable bin in n for a given value of μ, one sums probabilities until the desired level of confidence is met or exceeded. The bins in n that were added to reach the desired confidence level constitute the confidence interval for the given value of μ. The sorting procedure is repeated iteratively for each value of μ. The boundaries of the confidence belt can be easily identified by searching the lowest and largest values of μ bins that have been assigned to confidence intervals. Confidence belts for selected confidence levels are shown in Figure 6.11c,d for values $b = 4$ and $b = 8$, respectively.

With an expected background $b = 4$, and no actual signal, observation of $n = 4$ counts would lead to an upper limit of $\mu = 5.1$ at the 90% confidence level ($\mu = 8.4$ at CL 99%).

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

Observation of fewer counts, consistent with downward fluctuations of the background, would yield reduced upper limits as shown in Figure 6.11c. Upper limits are further reduced for $b = 8$ as shown in panel (d) of the figure.

It should be noted that the confidence intervals obtained with the procedure outlined in the preceding produces slight overcoverage, that is, the probability of values n included in confidence intervals slightly exceeds, typically, the required confidence level. Overcoverage is not a consequence of the method, however, but merely the result of the integer nature of the variable n and thus cannot be avoided.

6.2.2 Binomial Intervals: Detection Efficiency

The unified approach can also be applied to binomial distributions, and most particularly, to calculations of error intervals on detection efficiency. Let us consider, as an example, the determination of the track reconstruction efficiency of a large detector. This can be accomplished, for instance, with a technique known as embedding (§12.4.6). The embedding technique involves the insertion of hits of simulated tracks into actual detector events. One then reconstructs the embedded events as if they were normal events. The ratio of found embedded tracks to the number actually embedded constitutes a measure of the detection efficiency of such tracks. This determination can be carried out as a function of transverse momentum, rapidity, and so on. Other approaches are also available to determine detection efficiencies that depend on the specificities of the measurement of interest. The following discussion also applies to a host of situations in which binomial probabilities determine the outcome of an observation, including the determination of the mortality associated with a disease, the efficacy of a medical treatment or test, or even the success of a marksman in hitting a target.

The efficiency is defined as the ratio of the number of reconstructed tracks, n, by the number of true tracks N:

$$\varepsilon = \frac{n}{N}. \tag{6.60}$$

Since the n reconstructed tracks are a subset of the N embedded tracks, it would be incorrect to use the "usual" error propagation technique. Instead, one can assume N as given and based the evaluation of the error and confidence interval solely on n. One must, however, be cognizant of the fact that the efficiency cannot be larger than unity. So, for instance, if 1000 tracks are embedded and 980 tracks are reconstructed, one might be tempted to assess the error based on $\sqrt{980} \approx 30$. But this would suggest the efficiency might be larger than unity and is thus nonsensical. One might then resort to using a lower limit on the efficiency only for large values of n and a two-sided interval for smaller values. This is technically known as flip-flopping and should be avoided, as per our discussion in §6.1.8. The unified approach provides, thankfully, a better approach to the estimation of the error of the efficiency. As for the calculation of the mean μ in the previous section, one calculates the probability $P(n|N, \varepsilon)$ for a range of multiplicity n, given N and efficiency ε of interest.

$$P(n|N, \varepsilon) = \frac{N!}{(N-n)!n!}\varepsilon^n(1-\varepsilon)^{N-n}. \tag{6.61}$$

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

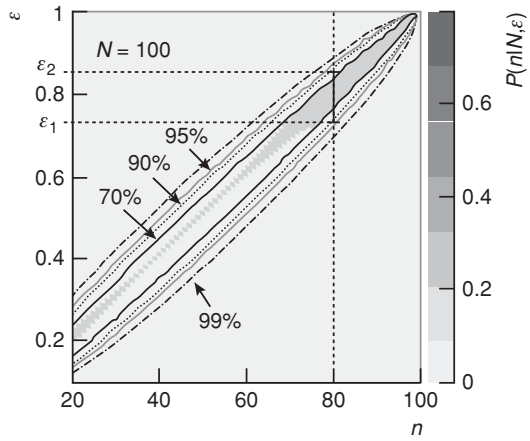


Fig. 6.12 Confidence intervals for the detection efficiency based on n detected objects (e.g., particles) based on $N = 100$ produced objects. Shown is the 90% confidence interval [$\varepsilon_1 = 0.73$, $\varepsilon_2 = 0.86$] for $n = 80$ reconstructed objects.

Given a specific value N, one can then proceed to calculate the ratio R defined as

$$R(n, \varepsilon|N) = \frac{P(n|N, \varepsilon)}{P(n|N, \varepsilon_{\max})}, \tag{6.62}$$

where $\varepsilon_{\max}$ corresponds to the most likely efficiency for the given n and N. R values are then used to rank, for given values of ε, the probabilities of n, and establish confidence intervals at the desired confidence level, as per the procedure described in the previous section. Figure 6.11 provides an example of 70%, 90%, 95%, and 99% confidence intervals for a relatively modest number of tracks $N = 100$, plotted as a function of n and the detection efficiency ε.

6.2.3 Gaussian Intervals

As we discussed in §6.1.3, the determination of confidence intervals for Gaussian distributed variables is rather straightforward, given the intrinsic symmetry $\mu \rightleftharpoons x$ of the argument of the exponential defining the Gaussian distribution. The symmetry breaks down, however, in the presence of physical boundaries. For instance, the mass of a particle cannot be negative, but measurements of mass can and often do yield negative values, when the mass being measured is either very small or null. We saw in §6.1.8 that the decision to report a double-sided interval or a one-sided interval (e.g., an upper limit) leads to flip-flopping, inconsistent coverage, or worse empty intervals. The unified approach avoids such problems and is of rather simple application for Gaussian distributions.

Let us consider a measurement involving a physical boundary at $\mu = 0$, that is, such that only positive values are allowed for the physical variable, while measurements can produce both positive and negative values. Let us also assume the standard deviation of the measurement σ is known, so one can rescale measurements by σ to obtain measured values of x with unit standard deviation.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

The unified approach requires that one finds, for a given measured value x, the value of the mean μ_{best} which is most probable. Assume x is given. If $x \geq 0$, the most probable value of the mean, that is, the value of μ that has the largest probability is obviously $\mu = x$, while for $x < 0$, the most probable value of μ is zero, since it cannot be negative. One thus writes

$$P(x|\mu_{\text{best}}) = \begin{cases} \frac{1}{\sqrt{2\pi}} & \text{for } x \geqslant 0, \\ \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) & \text{for } x < 0. \end{cases} \tag{6.63}$$

The ratio of likelihoods required for ordering probabilities is thus simply

$$R(x) = \frac{P(x|\mu)}{P(x|\mu_{\text{best}})} = \begin{cases} \exp\left(-\frac{1}{2}(x-\mu)^2\right) & \text{for } x \geqslant 0, \\ \exp\left(x\mu - \frac{1}{2}\mu^2\right) & \text{for } x < 0. \end{cases} \tag{6.64}$$

The ratio R determines the order in which values of x are added to the acceptance region of a given value μ. Starting with the most probable value of x, one adds values to the left and to the right until

$$\alpha = \int_{x_1}^{x_2} P(x|\mu)\, dx = \frac{1}{2}\left(\text{erf}(x_2/\sqrt{2}) - \text{erf}(x_1/\sqrt{2})\right), \tag{6.65}$$

where α is the desired confidence level. By construction, x_1 and x_2 are given equal ordering priority, so that one has

$$R(x_1) = R(x_2). \tag{6.66}$$

One must thus simultaneously solve Eqs. (6.65) and (6.66). Two cases need to be considered. First, if both $x_1, x_2 \geq 0$, the equality $R(x_1) = R(x_2)$ implies

$$x_1 = 2\mu - x_2, \tag{6.67}$$

whereas if $x_1 < 0, x_2 \geq 0$, one has

$$x_1 = \frac{x_2\,(2\mu - x_2)}{2\mu}. \tag{6.68}$$

Since neither x_2 nor x_1 are known, a priori, one can write $x_2 = \mu + \Delta x$ and seek a numerical solution for Δx iteratively. Starting with a small value Δx, one determines x_2 and x_1 according to (6.67). If x_1 is negative, one recalculates it with (6.68) before calculating the integral (6.65). One iteratively increases Δx until a value is reached that matches the required confidence level α. One plots the values x_1 and x_2 as a function of μ that define the required confidence intervals. This is illustrated in Figure 6.13 for confidence levels of 70%, 90%, 95%, and 95%. Given a measured value x, one finds the desired confidence interval $[\mu_1, \mu_2]$ by drawing a vertical line through the relevant confidence belt. For instance, for $x = 2$ one concludes the observable has a value $\mu = 2$ with a 90% confidence interval [0.58, 3.65], while for a measured value $x = -2$, one would conclude $\mu = 0$ with an upper limit of 0.4 at the 90% confidence level. Note how for smaller values of x still, upper limits are defined for all confidence levels.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

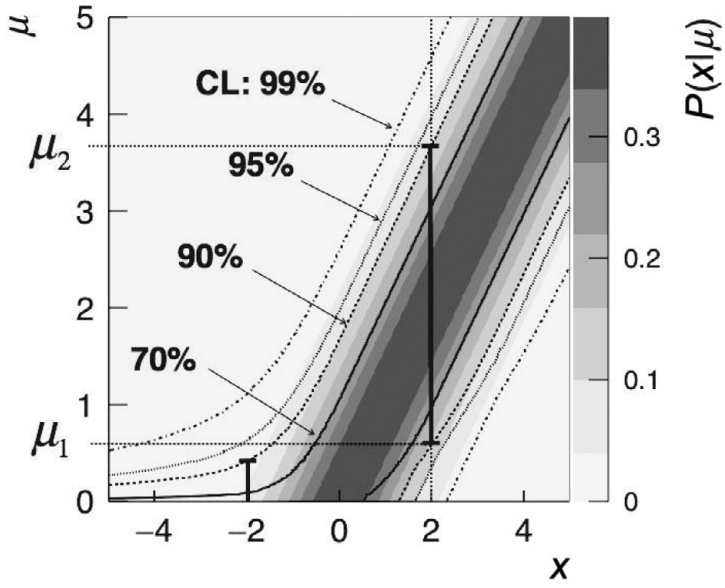


Fig. 6.13 Selected Gaussian confidence belts near a physical boundary, $\mu \geq 0$. Note how for $x \geq 0$, confidence levels are two-sided, whereas for $x < 0$, confidence intervals become one-sided and yield upper limits. Shown is the 90% confidence interval $[\mu_1 = 0.58, \mu_2 = 3.65]$ for $x = 2$, as well as the upper limit $\mu_2 = 0.42$ for $x = -2$.

6.3 Confidence Intervals from ML and LS Fits

In §§6.1 and 6.2, we discussed the notions of confidence intervals and confidence region with the implicit assumption that the joint probability density, or likelihood, of measured estimators, $p(\hat{\theta}|\vec{\theta})$, is known ab initio, and could thus be used, for instance, to generate a Neyman construction, thereby enabling the determination of confidence belts from which confidence intervals can be calculated on a case-by-case basis for estimates $\hat{\theta}$. However, the density $p(\hat{\theta}|\vec{\theta})$ may often be a rather complicated multiparameter function of measured data $\vec{x}$ and as such too difficult to model in detail. This is particularly the case in problems of ML and LS optimization involving several parameters. In such situations, obtaining an explicit model for $p(\hat{\theta}|\vec{\theta})$ might indeed be technically difficult, tedious, or too fastidious. Fortunately, as we saw in our discussions of the ML and LS methods, in §§5.1 and 5.2, the function $p(\hat{\theta}|\vec{\theta})$ may be approximated by a multiparametric Gaussian distribution,

$$p(\hat{\theta}|\vec{\theta}) \approx \frac{1}{(2\pi)^{m/2}\,|\mathbf{V}|^{1/2}} \exp\left[-(\hat{\theta}-\vec{\theta})^T \mathbf{V}^{-1} (\hat{\theta}-\vec{\theta})\right] \tag{6.69}$$

where m is the number of model (fit) parameters. One can show that, by virtue of the central limit theorem, this approximation becomes exact in the large sample limit.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

For illustrative purposes, let us consider a fit involving a single parameter θ. The likelihood function $L(\theta)$ may then be written

$$L(\theta) \equiv p(\hat{\theta}|\theta) = L_{\max} \exp\left[-\frac{(\hat{\theta}-\theta)^2}{2\sigma_{\hat{\theta}}^2}\right], \tag{6.70}$$

where $\hat{\theta}$ is the value of the estimator with maximum likelihood $L_{\max}$, and $\sigma_{\hat{\theta}}$ corresponds to the standard deviation of the parameter $\hat{\theta}$. One may readily obtain $n\sigma$ central confidence intervals by computing the values of $\theta_{n\sigma} = \hat{\theta} \pm n\sigma_{\hat{\sigma}}$ satisfying

$$\ln L(\theta_{n\sigma}) = \ln(L_{\max}) - \frac{n^2}{2}, \tag{6.71}$$

or equivalently, when using the $\mathbb{LS}$ method with Gaussian errors,

$$\chi^2(\theta_{n\sigma}) = \chi^2_{\min} + n^2. \tag{6.72}$$

For finite datasets, the likelihood function $L(\theta)$ may appreciably deviate from a Gaussian distribution; central confidence intervals $[\theta_<, \theta_>] = [\hat{\theta} - \delta_L, \hat{\theta} + \delta_H]$ may nonetheless be approximated according to

$$\ln L\left(\hat{\theta}^{-\delta_H}_{+\delta_L}\right) = \ln(L_{\max}) - \frac{n^2}{2}, \tag{6.73}$$

where $n = \Phi^{-1}(1-\alpha/2)$ is the quantile of standard normal distribution and the desired confidence level is $CL = 1-\alpha$. Clearly, for $\mathbb{LS}$ fits with Gaussian errors, this translates into

$$\chi^2\left(\hat{\theta}^{-\delta_H}_{+\delta_L}\right) = \chi^2_{\min} + n^2. \tag{6.74}$$

The foregoing confidence interval determination method is readily extended to fits involving a multiparametric model. Iso-contours or iso-surfaces are defined according to the discussion presented in §6.1.6. For a multiparameteric Gaussian model, one requires the quadratic function $Q(\hat{\theta}|\theta)$, given by Eq. (6.25), to satisfy

$$P\left(Q(\hat{\theta}, \vec{\theta}) \leq k^2_\beta\right) = \int_0^{\kappa^2_\beta} p_{\chi^2}(z|n)\, dz = \beta, \tag{6.75}$$

where β is the required confidence level, and k^2_β is the quantile of order β of the χ^2 distribution

$$k^2_\beta = P^{-1}_{\chi^2}(\beta|n). \tag{6.76}$$

Operationally, for a likelihood multiparametric Gaussian function, the m-dimension confidence region may thus be constructed by finding the locus of parameter values $\vec{\theta}$ that satisfy

$$\ln L(\vec{\theta}) = \ln L_{\max} - \kappa^2_\beta, \tag{6.77}$$

where $L_{\max}$ is the likelihood corresponding to the $\mathbb{ML}$ solution $\hat{\theta}$.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

The determination of multiparametric confidence regions may become technically challenging when there are several parameters or if the $\mathbb{ML}$ or $\mathbb{LS}$ solution $\hat{\theta}$ is found near a physical boundary of one or several of the model parameters. A variety of methods and approximation techniques have been developed to deal with such issues. A detailed discussion of these techniques is beyond the scope of this textbook but may be found in the book by F. James [116].

6.4 Hypothesis Testing, Errors, Significance Level, and Power

Progress in the physical sciences, and for that matter in all sciences, is based on a cycle of measurements and hypotheses. A specific model provides a prediction of a certain observable. An experiment is built and carried out to determine the value of the observable. Given the outcome of one or several observations, scientists must then determine whether the predicted and measured values are mutually compatible. If the values are compatible, the model is considered provisionally acceptable and might be used as the lead model until another measurement invalidates its predictions. On the other hand, if the values are not compatible, the model may be abandoned or require substantial revisions before it is considered acceptable. This raises the question of how one can assess whether two values, a measurement and a prediction, or even two distinct measurements, can be considered compatible. Because most observables of interest are continuous variables, one cannot sensibly expect a measured value to exactly equal the predicted value. All experiments are also obviously limited in statistical accuracy and might additionally entail sizable systematic errors. At best, one can hope that the difference between measured and predicted values is small. But how small should the difference be given the measurement errors? Indeed, one needs to quantify how large the difference can be, considering uncertainties in the measurement as well as, possibly, in the formulation of the prediction. At the end of the day, one is interested in assessing whether the model has a large likelihood of being valid given the measured value(s) and their estimated errors. In essence, one wishes to make a decision: Should the model be accepted (provisionally) or rejected? What is required is a statistical test to determine whether the measured and predicted values are compatible. Methods for the formulation of such tests and their properties are within the realm of a branch of statistics called **hypothesis testing**.

Hypothesis testing and decision making are of interest in a wide range of scientific and generic human activities. In car manufacturing, for instance, a manufacturer obviously wishes to reduce the likelihood of car defects and recalls. Parts included in the assembly of cars must then be tested (measured) to determine whether they fall within acceptable limits, or **tolerance limits**, before inclusion in a vehicle. In medicine, a doctor might use a low-cost test to determine the cholesterol level of her patient and decide to prescribe a medication if the measured level is outside a specific range. In particle physics, analyses often require a test to determine the species or type of a specific particle observed in a collision. Is the particle a charged pion, a kaon, or perhaps a proton? The possibilities are as varied as the activities undertaken by humans. It is thus useful to define the notion of hypothesis testing in a broad context based on the methods of statistics.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

The notions of hypothesis, hypothesis testing, and test errors are introduced in §6.4.1 while properties of significance, power, consistency, and bias are discussed in §6.5. Methods for the construction of tests, and most particularly the Neyman–Pearson test, are presented in §6.7. Selected examples of applications of statistical tests are presented in the subsections that follow.

6.4.1 Hypothesis Testing

The formulation of a (scientific) test begins with a statement about the expected outcome of a measurement based on a specific model. This statement constitutes the hypothesis to be tested and is commonly called the **null hypothesis**. It is usually represented as H_0.

The notion of null hypothesis is often perceived as counterintuitive by nonscientists, and its interpretation thus requires some elaboration. The point is that it is not possible to scientifically prove that a theory is correct; one can only demonstrate that it is incorrect. This is a principle of **falsifiability**. One can demonstrate that a model is wrong or incomplete under some or all circumstances, but one cannot prove that it is correct because it is not possible to examine all circumstances in which the model might be applied. It is also impossible to test the model with infinite precision. A null hypothesis typically adopts the point of view that an existing model (or theory) is correct but that measurements can be used to falsify the model, that is, to demonstrate the model is in fact incorrect. Consequently, to prove that a new model is required to explain measured data, one must first demonstrate that the data is **incompatible** with the null hypothesis.

As an example, consider a search for a new short-lived particle using the invariant mass technique (§8.5.1). Energy-momentum conservation dictates that the invariant mass of a pair of particles produced by the decay of the purported particle should equal its mass. An invariant mass spectrum obtained from measured particle pairs is thus expected to exhibit a peak at the mass of the decaying particle, as well as some broad-range background. But to prove the existence of the new particle, one must first show that the measured invariant mass spectrum is incompatible with the null hypothesis, which states that there is no new particle, or in other words, that the data can be satisfactorily explained without requiring the existence of a new particle. Discovery of a new particle is achieved when a peak is observed that cannot be explained by a statistical fluctuation. Nuclear and particle physicists typically require the probability a peak might be explainable by statistical fluctuations with a probability smaller than 3×10^{-7}, corresponding to a 5σ deviation from the "background" or null hypothesis.[2]

The null hypothesis H_0 may specify the expected value of an observable or the probability density $f(x)$ of measured values according to the model. It may also represent the status quo (e.g., no need for new physics). If the function $f(x)$ has no free parameters and is unique, the hypothesis H_0 is said to be a **simple hypothesis**. If the functional form of $f(x)$ is known or given but includes parameters to be determined by the measurement, the hypothesis is considered a **composite hypothesis**. We will here focus on the definition and properties of simple hypotheses.

[2] See §6.6.2 for a more in-depth discussion of this notion.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

It is often of interest to consider one of several alternative hypotheses $H_1, H_2, \ldots, H_m$. The purpose of the test shall then be to determine whether the null hypothesis H_0 can be rejected in favor of one of these alternative H_i, $i = 1, \ldots, m$.

Testing the null and alternative hypotheses requires making specific predictions about the outcome of one of several measurements of a specific observable x. These predictions are usually formulated in terms of probability densities $f(x|H_i)$, $i = 0, \ldots, m$, the observable x should follow if the hypotheses H_i, $i = 0, \ldots, m$, are considered valid (i.e., true). The PDFs $f(x|H_i)$ would ideally have small or no overlap for a given measurement in order to unambiguously discriminate against the various hypotheses. In practice, these PDFs may have considerable overlap, and one must consequently use boundaries or cuts to establish ranges that might be used to accept or reject particular hypotheses.

A measurement may involve n repeated observations of the same random variable or a single observation of an n-dimensional variable. Testing the agreement between the hypothesis H_0 (and its alternatives, if any) and the measurements $\vec{x} = \{x_i\}$, $i = 1, 2, \ldots, n$ may be investigated using a function of the measured variables called a **test statistic**, denoted $t(x)$. The test statistic is meant to encapsulate all (or selected) measured values and enable a straightforward comparison with model predictions. Given the PDFs $f(x|H_i)$, $i = 0, \ldots, m$, and a specific test statistic $t(\vec{x})$, the m hypotheses determine PDFs, denoted $p(t|H_i)$, which predict the distribution of values of t. Obviously, the test statistic $t(\vec{x})$ could specify a vector of values $\hat{\theta} = \{\theta_1, \theta_2, \ldots, \theta_k\}$, and even the vector $\vec{x}$ itself, or one of its subsets. Test statistics with $k < n$ or with a single value ($k = 1$) are usually preferred, however, because they enable a reduction of the data without, hopefully, a loss of discriminating power, that is, the capacity to keep and reject competing hypotheses.

For the purpose of our discussion, let us assume $t(\vec{x})$ is a scalar function (i.e., that it yields a single number). Let us also limit our considerations, without loss of generality, to a null hypothesis H_0 and one alternative hypothesis H_1 only. In this context, one expects the probability of t to be given by the PDF $p(t|H_0)$ if H_0 is true, and the PDF $p(t|H_1)$ if H_1 is true. The basis for our statistical test is illustrated in Figure 6.14. Let us assume the statistic t spans a domain D. Since the densities $p(t|H_0)$ and $p(t|H_1)$ cover different ranges of t, one selects a value t_{cut} that defines a zone of acceptance D_a for H_0 and a critical region $D_c = D - D_a$. Observations that yield a value t in D_c are considered an indication that the null hypothesis is false, even though there might be a finite probability of observing values of t in that range when H_0 is true. The formulation of a test for H_0 consequently amounts to a choice of a test statistic $t(x)$ and a critical region D_c. In the example presented in Figure 6.14, the region D_c consists of a single range of t, but one can in principle choose zones involving multiple disconnected regions or subsets of D.

The size of the region D_c and the shape of $p(t|H_0)$ determine the probability of finding t in that region if H_0 is true. As such, it determines the **level of significance** α of the test:

$$P(t \in D_c|H_0) = \alpha. \tag{6.78}$$

The value of α corresponds to the probability of rejecting H_0 based on the measurements $\vec{x}$, even if it is in fact true. Obviously, a test can be considered useful insofar as it can

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

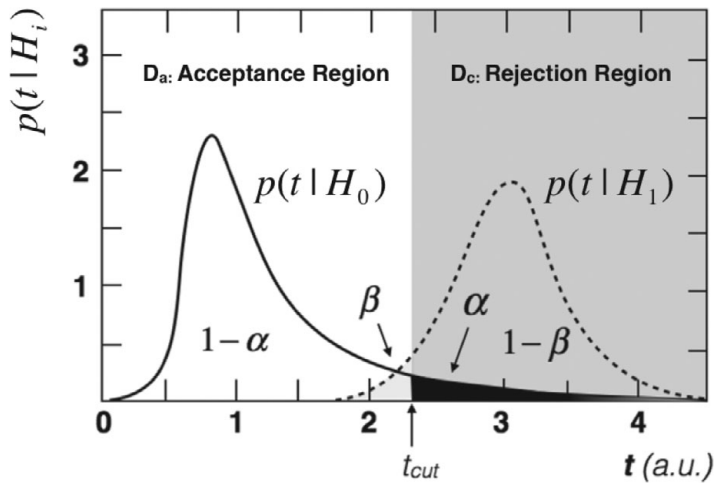


Fig. 6.14 Illustrative examples of test statistics $p(t|H_0)$ and $p(t|H_1)$ determined by PDFs $f(x|H_0)$ and $f(x|H_1)$, respectively. The variable θ_{cut} delineates the zone of acceptance D_a for H_0 and the critical region $D_c = D - D_a$ under which H_0 is rejected with a significance level α. The probability $1 - \beta$ determines the power of the test and the probability that the alternative hypothesis H_1 might be accepted.

discriminate against one or several alternative hypotheses, H_i, $i = 1, 2, \ldots$. The usefulness of the test is commonly expressed in terms of its **power**, defined as the probability $1 - \beta$ of t falling in the critical region D_c if H_1 is true. One thus has

$$P(t \in D_a|H_1) = \beta. \tag{6.79}$$

Given a null hypothesis H_0, alternative hypothesis H_1, and specific choices for the regions D_a and D_c, four distinct situations may be encountered:

1. t falls within D_a and H_0 is true. The hypothesis H_0 is *correctly* accepted with probability $1 - \alpha$.
2. t falls within D_c and H_0 is true. The hypothesis H_0 is *wrongly* rejected with a probability α. This is commonly known as an error of the **1st kind**, although when a test is used for classification purposes it may be more useful to consider this situation as a **loss** of efficiency.
3. t falls within D_a and H_0 is false. The hypothesis H_0 is *wrongly* accepted with a probability β. This is commonly known as an error of the **2nd kind**.
4. t falls within D_c and H_0 is false. The hypothesis H_0 is correctly rejected with a probability $1 - \beta$, known as the **power** of the test.

These four situations and their respective probabilities are summarized in Table 6.2.

Statistical tests may be used in a variety of contexts and various purposes. We consider specific examples of applications in §6.6. But first, let us examine the key properties of statistical tests.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

Table 6.2 Definition of errors of the 1st and 2nd kind. Shown are probabilities of a test statistic t to be found in the critical region D_c and used to correctly or falsely reject the null hypothesis H_0 when either H_0 or the alternative hypothesis H_1 is actually correct.

	Probability & Error Type	
True Hypothesis	$t \in D_c$	$t \notin D_c$
	H_0 Rejected	H_0 Accepted
H_0	α	$1-\alpha$
	Error 1st kind	OK
H_1	$1-\beta$	β
	OK	Error 2nd kind

6.5 Test Properties

The capacity of a test to discriminate between a null hypothesis H_0 and an alternative hypothesis H_1 depends on a variety of factors including the measurement protocol and instrumentation performance, the number of measurements n, the choice of test statistic $t(x)$, and the choice of critical region D_c. The extent to which the PDFs $p(t|H_0)$ and $p(t|H_1)$ overlap is key in discriminating between the two hypotheses. The shape and overlap of these two functions are largely determined by the measurement performance embodied in the PDFs $f(x|H_0)$ and $f(x|H_1)$, but also depend on the number of measurements and the specific choice of test statistic $t(x)$ which altogether determine the functions $p(t|H_0)$ and $p(t|H_1)$. The discriminating power of the test and the likelihood of encountering errors of the 1st and 2nd kind obviously also depend on the specific choices of cut(s) used to select the critical region D_c and its complement D_a. All facets of a measurement and the statistical test used to test a hypothesis thus play a role in one's capacity to validate or reject a null hypothesis.

Overall, it is typically the measurement method and protocol that have the most influence on the discriminating power of a test. If the measurement resolution and efficiency are excellent, one surely can expect the measurement to provide a better test of a theory than a measurement with poor resolution and efficiency. Given a particular measurement technique, however, the choice of statistic $t(x)$ and the critical region D_c shall also influence the discriminating capacity of the test. It is thus of interest to examine the properties of a statistical test from a formal point of view and consider how or whether these properties may be optimized.

The most important property of a statistical test is its **power** $1-\beta$, but one should also be concerned about the **test consistency** and the possibility of a **test bias**. These three

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

properties are briefly discussed in §§6.5.1–6.5.3. A more extensive discussion of these properties is presented by James [116].

6.5.1 Test Power

Definition

Let us consider the determination of a parameter θ based on some measurements $\vec{x} = (x_1, x_2, \ldots, x_n)$ and a statistic $t(\vec{x})$. Let us assume that the null hypothesis stipulates that the parameter's expectation value is $\theta = \theta_0$, whereas the alternative hypothesis implies the value should be $\theta = \theta_1$. We wish to optimize the choice of the statistic $t(\vec{x})$ in order to maximize the power of the test, and minimize the chance of incurring an error of the first kind.

$$\begin{aligned} H_0 &: \theta = \theta_0 \qquad \text{Null hypothesis} \\ H_1 &: \theta = \theta_1 \qquad \text{Alternative hypothesis} \end{aligned}$$

The test statistic t shall have PDFs

$$\begin{aligned} &p(t|H_0(\theta_0)) \quad \text{if } H_0 \text{ is true,} \\ &p(t|H_1(\theta_1)) \quad \text{if } H_1 \text{ is true.} \end{aligned} \tag{6.80}$$

The power of a statistical test is formally defined as the probability the test statistic t might fall in the critical region D_c defined by a cut value t_{cut}, known as a **decision boundary**, as illustrated in Figure 6.14.

$$p = 1 - \beta = \int_{t_{\text{cut}}}^{\infty} p(t|H_1(\theta_1))\, dt. \tag{6.81}$$

The cut value t_{cut} may in principle be selected to achieve a specific significance level α according to

$$\alpha = \int_{t_{\text{cut}}}^{\infty} p(t|H_0(\theta_0))\, dt. \tag{6.82}$$

But for a test to be useful, it should also have as large a power $p = 1 - \beta$ as possible in order to minimize errors of second kind (i.e., the probability of wrongly accepting H_0) and to maximize the probability of identifying H_1 as the correct hypothesis. The choice of t_{cut} is thus often predicated by the discriminating power one wishes to achieve and the losses one can tolerate as determined by the significance level.

A Simple Example with Gaussian Deviates

The notion of power is illustrated with a concrete example in Figure 6.15a, where one considers a measurement of an observable X based on a procedure involving Gaussians fluctuations characterized by a standard deviation σ. One assumes there exist two hypotheses or models predicting the value of X: the null hypothesis H_0 stipulating it has a value $\mu_0 = 0$ and an alternative hypothesis stating it has a value $\mu_1 = \mu_0 + \Delta$, with $\Delta = 2\sigma$. One assumes that it is legitimate to accept or reject the null hypothesis with a 1%

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

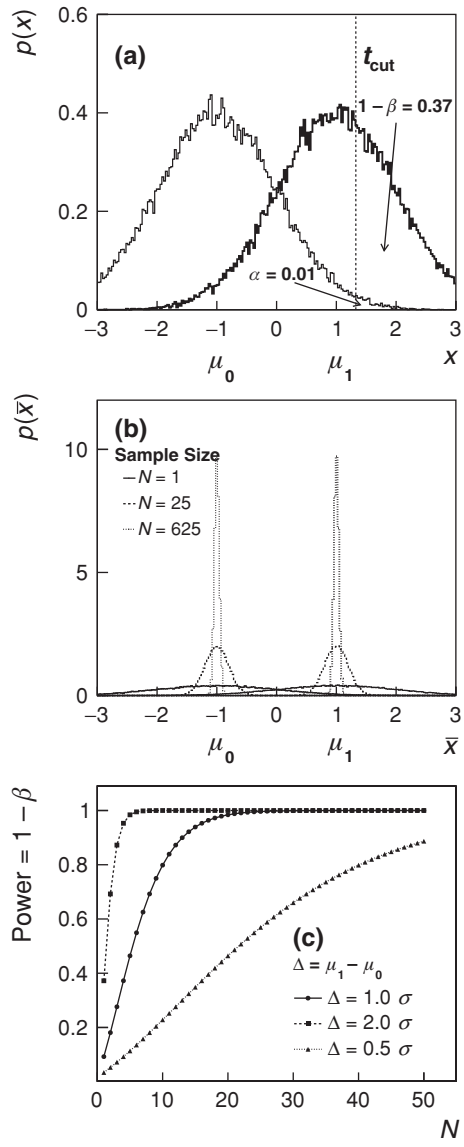


Fig. 6.15 Illustration of the notion of power based on a test involving two hypotheses $x = \mu_0$ and $x = \mu_1$ for a measurement with Gaussian deviates of standard deviation σ. In (a), the test is performed based on a single measurement of x while in (b) it is carried out based on the arithmetic mean of N sampled values. (c) Power of the test as a function of the sample size for three values of relative separation Δ/σ between the two hypotheses.

significance level based on a single measurement of X. The critical region is defined by a cut x_{cut} selected to yield the desired significance level, that is, such that

$$\alpha = 0.01 = \int_{x_{\text{cut}}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu_0)^2}{2\sigma^2}\right] dx = 1 - \Phi(x_{\text{cut}}). \tag{6.83}$$

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

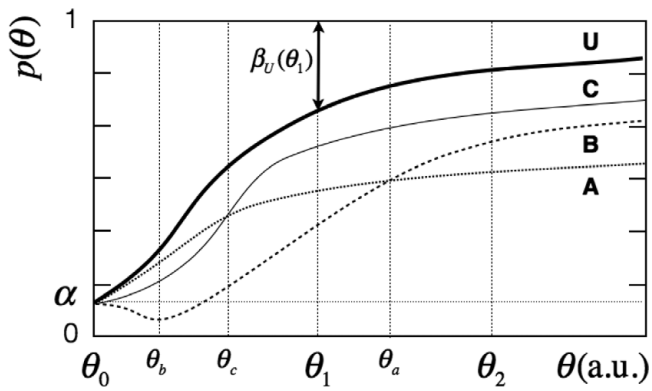


Fig. 6.16 Illustration of the concepts of power function, most powerful test, uniformly most powerful test, and test bias.

Inverting the CDF Φ, one gets $x_{\text{cut}} = \Phi^{-1}(0.99) = 2.326$ shown as a vertical dash line in Figure 6.15a. The power $1 - \beta$ of the test is calculated on the basis of the probability distribution of X assuming $x = \mu_1$ is the correct hypothesis. One finds

$$1 - \beta = \int_{x_{\text{cut}}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma^2}\right] dx = 0.37. \tag{6.84}$$

A power of 0.37 is relatively weak and implies that there is a large probability, 63%, that H_0 might be accepted even it is wrong (error of the second kind). Evidently, the power would be larger if the separation $\Delta = \mu_1 - \mu_0$ was larger or if the standard deviation σ of the measurements was smaller, or both.

Power as a Function of the Alternative Hypothesis

The power achievable, for a given significance level α, depends on the value of θ_1 predicted by the alternative hypothesis H_1:

$$p(\theta_1) = 1 - \beta(\theta_1) = \int_{t_{\text{cut}}}^{\infty} p(t'|H_1(\theta_1))\, dt'. \tag{6.85}$$

It is thus useful to regard the power as function of θ:

$$p(\theta) = 1 - \beta(\theta) = \int_{\theta_{\text{cut}}}^{\infty} p(t'|\theta)\, dt'. \tag{6.86}$$

By construction, the power function should yield α for $\theta = \theta_0$:

$$p(\theta_0) = 1 - \beta(\theta_0) = \alpha. \tag{6.87}$$

One can thus characterize a test in terms of its **power function**, which consists of a plot of the test's power as a function of θ. Examples of power functions are displayed in Figure 6.16 for several hypothetical tests. In general, the power of a test increases monotonically

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

(tests A, C, U) when the separation between the null hypothesis, θ_0, and the alternative θ_1 increases, but there can be pathological cases (e.g., test B) where the power decreases in some range of θ (see the discussion of test bias in the next section).

The power function provides a useful tool to compare different tests. If the alternative hypothesis is simple (i.e., θ_1 is specified a priori), the best test of H_0 against H_1 at a given significance level α shall be the test with the maximum power at $\theta = \theta_1$. As an illustrative example of this notion, Figure 6.16 compares the power functions of three hypothetical tests noted **A**, **B**, and **C**. Test **C** exhibits a larger power than **A** and **B** at all values $\theta > \theta_c$, and particularly at $\theta = \theta_1$. It is thus considered the best test to discriminate against that particular hypothesis. However, test **A** would be best for testing against hypotheses with $\theta < \theta_c$, because it has the largest power in that region.

A test which is at least as powerful as any other possible test at a specific value of t is called **most powerful test** at that particular value of t. A test which is most powerful for all values of θ under consideration is known as **Uniformly Most Powerful** (UMP). Figure 6.16 includes an example of a UMP test, noted **U**. A UMP test is obviously the best choice to discriminate against an alternative hypothesis, but the use of a non-UMP test might be preferred for reasons of simplicity and/or robustness (e.g., a test with low sensitivity or dependence on unimportant changes in H_0).

6.5.2 Test Consistency

Definition

Experimentally, one wishes for a test that provides an increasingly better discrimination between a null hypothesis H_0 and an alternative hypothesis H_1 as the number of observations/measurements increases. For instance, the relative error $\delta\theta/\theta$ on the mean of an observable θ typically decreases inversely as the square root of the number N of sampled values. One would thus expect that the PDFs $p(t|H_0)$ and $p(t|H_1)$ become progressively distinct as N increases, thereby leading to an improved power at a fixed significance level.

A test is considered consistent if its power tends to unity as the number of observations N goes to infinity, even though this cannot be achieved in practice. This can be written formally as

$$\lim_{N\to\infty} p(t \in D_c|H_1) = 1, \tag{6.88}$$

where D_c is the critical region with probability α under H_0.

Consistent tests feature a power function that tend to a step function for $N \to \infty$, as illustrated schematically in Figure 6.17.

A Simple Example with Gaussian Deviates (cont'd)

In order to illustrate the notion of consistency, we revisit the test example presented in §6.5.1 and replace a single measurement of the observable X by the mean $\bar{x}$ of a sample consisting of N measurements of X. The error on $\bar{x}$ is $\sigma_N = \sigma/\sqrt{N}$. The relative distance Δ/σ_N between the two hypotheses then scales in proportion to $\sqrt{N}$. The PDFs of $\bar{x}$ for both

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

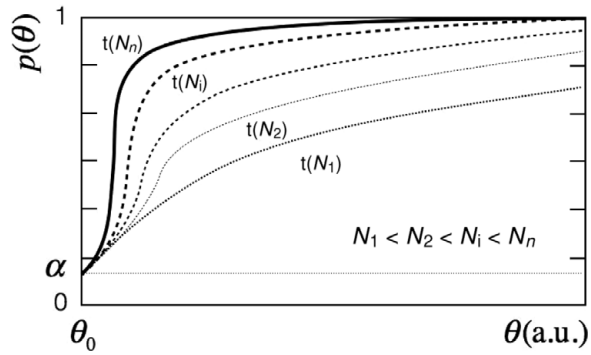


Fig. 6.17 A test is considered consistent if its power tends to unity as the number of observations N increases. Its power function thus approaches a step function as $N \to \infty$.

hypotheses thus narrow proportionally to $\sqrt{N}$ and for large N, distinguishing between the two hypotheses becomes quite straightforward, as illustrated in Figure 6.15b. The power of the test, based on a significance level $\alpha = 0.01$, is easily calculated. One finds

$$1 - \beta = \int_{\mu_0 + 2.326\sigma/\sqrt{N}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma^2/N}\right] dx, \tag{6.89}$$

$$= 1 - \Phi\left(-\sqrt{N}\Delta/\sigma + 2.326\right), \tag{6.90}$$

which is plotted in Figure 6.15c for three values of $\Delta = \mu_1 - \mu_0$. One finds, as anticipated, that the test power increases with $\sqrt{N}\Delta/\sigma$. Distinguishing between hypotheses H_0 and H_1 is thus possible even if Δ/σ is small, provided sufficiently many repeated measurements of the observable X can be carried out. Indeed, the probability $\Phi(-\sqrt{N}\Delta/\sigma + 2.326)$ vanishes in the large N limit, and one concludes the test can be considered consistent. In fact, it should nominally be possible to carry the test for an arbitrary large power $1 - \beta$, for any significance level α, provided measurements of X be carried out for sufficiently large N. In practice, however, limitations also arise from systematic errors inherent to the measurement of X. Unlike the statistical error on the sample mean, systematic errors cannot be reduced or suppressed, usually, by increasing the size N of the data sample. They thus impose a minimum limit on the difference $\Delta = \mu_1 - \mu_0$ which can be resolved in practice. Consequently, although a statistical test may be consistent, its power shall remain smaller than unity, in practice, because of systematic errors involved in the measurement procedure.

6.5.3 Test Bias

A test is generally considered biased if the power function $p(\theta)$ does not rise monotonically with θ. An example of a biased test is provided by test **B** in Figure 6.16. The power of this test decreases below the confidence level α in the range $\theta_0 < \theta < \theta_c$ and reaches a minimum near $\theta = \theta_b$. This implies that H_0 is more likely to be accepted when $\theta = \theta_b$ than

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

for $\theta = \theta_0$, that is, the hypothesis H_0 is more likely to be accepted when it is wrong than when it is actually right, a clearly undesirable situation.

Although undesirable, a biased test might in principle be usable if, for instance, it provides a better power than other tests in a specific region of interest of the test statistic t. This is illustrated in Figure 6.16 by the fact that test **B** has better power than test **A** in the region $\theta = \theta_2$. Test **B** might then be usable in spite of its bias if it is particularly important to discriminate against the hypothesis $\theta = \theta_2$.

6.6 Commonly Used Tests

A simple test can be formulated on the basis of the probability (density) distribution that governs the fluctuations of a variable x of interest in a specific situation. Indeed, given a PDF $f(x|\vec{\theta})$ determined by parameters $\vec{\theta}$, assumed known a priori, one may define a null hypothesis H_0 as stating that observed values of x are consistent with the PDF $f(x|\vec{\theta})$ and parameters $\vec{\theta}$. The desired significance α of the test determines the bounds $x_{\min}$ and $x_{\max}$ beyond which H_0 is rejected. For a two-sided interval, one has

$$\alpha = 1 - \int_{x_{\min}}^{x_{\max}} f(x|\vec{\theta})\, dx. \tag{6.91}$$

The integral is replaced by a sum for discrete probability distributions.

We consider several commonly used tests in the remainder of this section.

6.6.1 Binomial Test

A binomial test is of interest whenever one has reasons to suspect changing conditions in a fixed probability sampling experiment or process. One can, for instance, rely on binomial sampling to monitor the sanity and good performance of the operation of a particle detector. An analysis stream might be set up to identify a specific type of collision or process (e.g., a very high charged particle multiplicity), hereafter called diagnostic events, which happen with a probability p. Once the experimental conditions are fixed, one would expect this particular process to "fire" a certain fraction of the time p and yield, on average, $\langle n \rangle = pN$ diagnostic events in a set of N collected events. If the number n of observed diagnostic events is too small or too large, one might then conclude that the experimental conditions have changed, that is, that the response of the detector has shifted and now yields incorrect results. Periodic measurements of event samples of size N could then be used to determine whether the experimental response is improperly shifting.

Defining a central acceptance interval as $[pN - \Gamma, pN + \Gamma]$, one writes

$$\alpha = 1 - \sum_{n=pN-\Gamma}^{pN-\Gamma} \frac{N!}{n!(N-n)!} p^N (1-p)^{N-n}, \tag{6.92}$$

which (approximately) determines the desired significance level α. For large values n and N, one may use a Gaussian approximation with mean $\mu = pN$ and standard deviation

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

$\sigma = \sqrt{p(1-p)N}$:

$$\alpha = 1 - \frac{1}{\sqrt{2\pi}\sigma}\int_{\mu-\Gamma}^{\mu-\Gamma} dx \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{6.93}$$

$$= 1 - \Phi(-\Gamma/\sigma) + \Phi(\Gamma/\sigma). \tag{6.94}$$

Suppose, for instance, that the probability of the diagnostic events is $p = 0.10$. A sample of $N = 1000$ events should thus, on average, yield $\langle n \rangle = 100$ such events. Fluctuations well below or far above this mean could be used to identify a shift in the detector performance. Since one wishes to be on the safe side and identify potential problems early, one might set the significance of the test to 10%. The standard deviation of a binomial distribution is $\sigma = \sqrt{p(1-p)N}$. For $N = 1000$, this yields $\sigma = 9.5$. Considering n and N are both quite large, one can approximate the binomial distribution by a Gaussian distribution and conclude that for a significance level $\alpha = 10\%$, degradation of the performance could be identified if samples featured deviations by more than $1.65 \times 9.5 = 15.7$ from the mean $\langle n \rangle = 100$. One could of course reduce the number of inquiries by reducing the significance level of the test to 5%, but it would be probably be unwise to use a smaller value given that changes in performance might be slow or small and not necessarily yield large shifts in the value of the probability p.

6.6.2 Poisson Test

Whether by design or serendipitous happenstance, the discovery of new entities such as elementary particles, stars, or galaxies, often occurs with the observation of a peak in a probability distribution, spectrum, or histogram. However, such a peak is commonly observed atop a more or less flat background produced by known processes. What peak strength or amplitude shall constitute grounds for a claim of discovery then seemingly stands as the key question. Yet, proper scientific procedure requires that one first formulate a null hypothesis H_0 stipulating that the peak might be simply the result of background fluctuations. The null hypothesis shall indeed be that the observed spectrum is consistent with known processes or backgrounds, while the alternative hypothesis would be that the observed peak corresponds to a new process or entity (e.g., a new particle). The task at hand, the test, shall be to establish whether the observed peak is consistent with the null hypothesis, that is, whether it might be caused by an upward fluctuation of background processes.

As a practical example, consider a search for a new particle based on the invariant mass technique (described in §§8.2.1 and 8.5.1). The technique assumes that the purported particle decays in some specific channel involving two (or three) known particles. Having measured the momenta of the particles, and accessorily identified their species, one can determine the mass of the parent particle based on the invariant mass of the pair:

$$m_p = \sqrt{m_1^2 + m_2^2 + 2|\vec{p}_1||\vec{p}_2|\left(\sqrt{1+\frac{m_1^2}{|\vec{p}_1|^2}}\sqrt{1+\frac{m_2^2}{|\vec{p}_2|^2}} - \cos\theta_{12}\right)},$$

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

where θ_{12} represents the angle between the momenta $\vec{p}_1$ and $\vec{p}_2$ of the detected particles. Several (known) processes and combinatorial background could yield particle pairs with a mass in the range of interest. Suppose that based on some prior knowledge, one expects the background in a specific mass range will consist of $n_b = 52$ pairs, but that $n_{\text{tot}} = 90$ pairs are actually measured. Can one conclude that the observed yield is incompatible with the expected average, and that there is evidence for a new particle?

Clearly, one must decide, based on some general principles, what constitutes grounds for inconsistency with the null hypothesis and thus a claim of discovery. While a significance level of 10% or 5% is often considered sufficient to identify anomalous behaviors, it does not seem justified to single out evidence for a new particle. Particle and nuclear physicists alike typically require the level of significance to be of order of 3×10^{-7} or less to reject the null hypothesis and announce the discovery of a new particle. In our example, based on Poisson fluctuations one expects the standard deviation of the background yield to be $\sigma = \sqrt{52} = 7.2$. Excess above the expected background would thus be considered significant if it amounts to (approximately) five times the standard deviation, that is, $n = 52 + 5 \times 7.2 = 88$. Observation of 90 counts in the mass window of interest might then be considered grounds for a claim of discovery, provided of course that one can unambiguously demonstrate that the expected background level is correct. Observation of a smaller number of counts (e.g, $n_{\text{tot}} = 73$), though nearly 3σ in excess of the expected number of counts, would not be considered grounds for a claim of discovery. Deviations of this magnitude, though rare, have in fact been observed experimentally with relatively small datasets and later shown, with increased statistics, to indeed be only due to fluctuations.

Obvious complications arise in practice because the expected background may not be known with absolute certainty, and thus, it has an estimated error of its own. The observed counts should then exceed fluctuations from the expected background while accounting for the fact that it might be underestimated. If the uncertainty on the background is $\sigma_{\text{b}} = 3$, one might require the observed number of counts to exceed

$$n = n_b + 3 \times \sigma_{\text{B}} + 5 \times \sqrt{n_b + 3 \times \sigma_{\text{b}}} = 100 \tag{6.95}$$

in order to conclude a new particle has been observed in our example, and then the observed 90 counts would not match this challenge. Note that we here used a factor of 3 toward the estimation of an upper limit of the expected background. Even more conservative factors (i.e., larger) can be used in practice.

6.6.3 Gaussian Test: Comparison of Sample Means with Known Standard Deviations

A common situation encountered in statistical analysis involves the comparison of the sample means $\hat{\mu}_1$ and $\hat{\mu}_2$ obtained from two distinct sets of measurements of the same quantity. One might, for instance, be interested in finding out whether values produced by different experiments or techniques are compatible with one another. Alternatively, one might be interested in assessing whether a change or improvement applied to a process has an impact on the outcome of that process. In general, two samples taken from the same parent distribution will have a different arithmetic mean. One is thus interested in distinguishing

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

samples that are actually different from those that originate from the same parent distribution. The null hypothesis shall be that the two samples and their means are consistent with one another.

Let us first consider the comparison of samples with means $\hat{\mu}_i$, $i = 1, 2$, and known standard deviation σ_i. The two values $\hat{\mu}_i$ are known to be randomly distributed about their true value(s) according to Gaussian distributions with standard deviations σ_1 and σ_2, each determined by the actual standard deviation of the samples divided by the number of measurements in each sample. If the two means $\hat{\mu}_i$, $i = 1, 2$ are representative of the same parent distribution, their difference $z = \hat{\mu}_1 - \hat{\mu}_2$ should be compatible with zero. Actually, the difference z should be distributed according to a Gaussian distribution with a null mean and a variance equal to $V[z] = \sigma_1^2 + \sigma_2^2$. We formulate a null hypothesis stipulating that z should have fluctuations determined by $\sigma_z = \sqrt{V[z]}$ with a null expectation value. One can then test whether the means $\hat{\mu}_i$, $i = 1, 2$ are compatible with one another by assessing whether their difference z exceeds some threshold z_0 considered to identify anomalous differences. The threshold z_0 determines the significance level α of the test (and conversely a choice of α determines z_0) according to

$$\alpha = 1 - [\Phi(z_0/\sigma_z) - \Phi(-z_0/\sigma_z)]\,. \tag{6.96}$$

Choosing, for instance, the confidence level α to correspond to a difference in excess of 3σ, one would reject the null hypothesis with a probability of 1% if the observed difference $|z|$ exceeds 3σ. Changes in the outcome of a measurement procedure could thus be readily identified if the means differ by more than 3σ. Of course, one can require the test to be less or more stringent based on the circumstances and needs of the situation. If the selected value of α is large, the difference z required to flag an event or difference as anomalous is small. Conversely, if α is chosen to be very small, the acceptance interval will be very large, and very few normal events (or differences) will be considered as anomalous, and if differences in excess of $|z_0|$ are observed, they will be more likely due to genuine differences in the two samples.

6.6.4 Student's *t*-Test: Comparison of Sample Means with Unknown Standard Deviations

While it might be reasonable to assume that two samples have Gaussian distributions, their standard deviations may not be known a priori. A comparison of the means of the two samples thus requires that one estimates their standard deviations according to

$$\hat{\sigma}_i = \sqrt{\frac{1}{n_i - 1} \sum_{k=1}^{n_i} \left(x_k^{(i)} - \hat{\mu}_i\right)^2}, \tag{6.97}$$

where $\hat{\mu}_i$ represents the arithmetic means of the two samples, which consists of n_i, $i = 1, 2$, observed values $x_k^{(i)}$.

Because estimates $\hat{\sigma}_i$ rather than actual values σ of the standard values are available, one cannot make use of the standard normal distribution to determine a confidence interval but

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

must instead use Student's t-distribution:

$$p_t(t|n) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(n+1/2)}{\Gamma(n/2)} \frac{1}{(1+t^2/n)^{(n+1)/2}}, \tag{6.98}$$

for n degrees of freedom. A derivation of this expression is provided in §6.8. The variable t is nominally defined as

$$t = \frac{(x-\mu)}{\sigma} \frac{\sigma}{\hat{\sigma}} = \frac{z}{\sqrt{\chi^2/n}}. \tag{6.99}$$

In the context of the comparison of two sample means, one has $x = \hat{\mu}_1 - \hat{\mu}_2$, and $\mu = 0$ if the null hypothesis is correct. The variable z is the ratio of x by its standard deviation

$$z = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}, \tag{6.100}$$

while the χ^2 can be written

$$\chi^2 = \frac{(n_1-1)\hat{\sigma}_1^2}{\sigma_1^2} + \frac{(n_2-1)\hat{\sigma}_2^2}{\sigma_2^2} \tag{6.101}$$

for $n_1 + n_2 - 2$ degrees of freedom. By virtue of the null hypothesis, it is legitimate to assume that $\sigma_1 = \sigma_2$ and the variable t simplifies to

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{S\sqrt{1/n_1 + 1/n_2}}, \tag{6.102}$$

with

$$S^2 = \frac{(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}, \tag{6.103}$$

which is a properly weighted estimate of the two samples' standard deviation, known as a "pooled estimate."

Use of Student's t-distribution to establish a confidence interval and/or significance level proceeds as for other distributions. The probability that the difference $|x|$ exceeds a specific value x_0 is given by

$$\alpha = \int_{-t_0}^{t_0} p_t(t|n)\, dt, \tag{6.104}$$

where p_t is computed according to Eq. (6.98) for $n = n_1 + n_2 - 2$ degrees of freedom and t_0 given by

$$t_0 = \frac{x_0}{S\sqrt{1/n_1 + 1/n_2}}. \tag{6.105}$$

Integrals of p_t and values of t_0 for specific significance levels α must nominally be computed for each value of n. Such integrals are tabulated in the literature dating back to times when computing resources were scarce [5]. However, tools to evaluate Student's t-distribution and the foregoing integral are now also available in most statistical software packages, including ROOT [59].

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

6.6.5 The χ^2 Test: Goodness-of-Fit

Modeling and fitting data is the bread and butter of physicists and scientists alike. Yet, the question remains for every fit: does the model used in the fit provide a suitable representation of the data, or is there evidence of incompatibilities between the data and the model? To answer this question, we formulate a null hypothesis that posits that the model does in fact provide a good description of the data, but that fluctuations and errors may introduce apparent deviations between the two. One then needs to consider the probability that such fluctuations arise and establish, for a given significance level, a measure of what might constitute a meaningful deviation between the data and the model (fit). This measure is formulated in the form of a statistics based on the data sample and the results of the fit. If the statistics deviates "too much" from values expected to be representative of a good fit, one should reject the fit and/or the mathematical model used in the description of the data.

The χ^2 test, by far the most commonly used goodness-of-fit test, is based on the statistics χ^2 defined as

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(x_i|\vec{\theta}))^2}{\sigma_i^2}. \tag{6.106}$$

The sum is taken over N relevant data points (x_i, y_i) where x_i and y_i are considered independent and dependent variables, respectively. The function $f(x_i|\vec{\theta})$ is a mathematical model meant to describe the relation between the observables x and y and is determined by m parameters θ_k. The values σ_i^2 represent the variance of the measurements y_i, that is, they represent the variance that (infinitely many) repeated measurements of the values y_i would yield if such repeated measurements could be performed.

If the measurements errors of the N data points y_i are correlated, the χ^2 is written

$$\chi^2 = (\vec{y} - \vec{f})^T \mathbf{V}^{-1} (\vec{y} - \vec{f}), \tag{6.107}$$

where $\mathbf{V}^{-1}$ is the inverse of the covariance matrix of the measurements y_i.

The χ^2 test can be used to compare data with a fixed-parameters model or with the outcome of a fit. Fitting techniques to find model parameters that best match the data (i.e., minimize the χ^2) are discussed in §§5.2 and 7.6.

The null hypothesis is that the function $f(x_i|\vec{\theta})$ does describe the data. If that is truly the case, one would expect that each of the differences $y_i - f(x_i|\vec{\theta})$ would be, on average, of the order of σ_i owing to statistical fluctuations. One would then expect the χ^2 to be roughly equal to N on average. If χ^2 is much larger than N, it provides an indication that the fit is bad or that the model function $f(x_i|\vec{\theta})$ does not provide an appropriate description of the data.

We showed in §3.13.3 that if the values y_i fluctuates according to Gaussian distributions with standard deviations σ_i, the χ^2 function is expected to fluctuate according to a χ^2-distribution for n degrees of freedom:

$$p_{\chi^2}(\chi^2|n) = \frac{1}{2^{n/2}\Gamma(n/2)} \chi^{n-2} e^{-\chi^2/2}. \tag{6.108}$$

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

The number of degrees of freedom n here corresponds to the total number of points if the model has no free parameter, or to the number of fitted points N minus the number of free parameters m used in the fit. The probability $\alpha(\chi_0^2, N)$ of observing a χ^2 in excess of a certain value χ_0^2 with n degrees of freedom is then

$$\alpha(\chi_0^2, N) = \int_{\chi_0^2}^{\infty} p_{\chi^2}(\chi^2|n)\, d\chi^2. \tag{6.109}$$

At this point, two paths are possible. One may decide on a desired level of significance and determine the minimal values χ_0^2 beyond which all observed χ^2 would be flagged as inconsistent with the null hypothesis. Unfortunately, this requires the values χ_0^2 be calculated for each possible values of the number of degrees of freedom n, a rather tedious process. A second path is thus often used in practice, which consists in calculating the integral in Eq. (6.109) with a lower boundary χ_0^2 replaced by the χ^2 obtained in specific fits. If the integral yields a value smaller than a preselected significance level (e.g., $\alpha = 0.01$), one may then reject the null hypothesis and consider the function to be an invalid description of the measured data.

However, it should be stressed that the probability $\alpha(\chi_0^2, N)$ is strictly accurate only if the fluctuations of the variables y_i are actually Gaussian and that the errors σ_i are properly evaluated. If the errors are inflated (i.e., overestimated), the measured χ^2 will be too small and the function might not be appropriately rejected. Conversely, if the errors are underestimated, the measured χ^2 will be too large, and the function might then be improperly rejected. It is thus rather important to calibrate measurement errors if a χ^2 test is to be used to accept/reject fits performed on data sets. A very low χ^2 might alternatively result from "overfitting," that is, the use of a function with too many parameters given the number of data points. For instance, the fit of a distribution involving three data points with a polynomial of order 3 or more would obviously yield a perfect fit and a null χ^2, but it is clear that three points alone do not provide sufficient information to uniquely determine the coefficients of such a polynomial. A small or null χ^2 thus does not always provide a reliable indicator of the goodness of a fit. In this context, indeed, the χ^2 is essentially meaningless because the fit function simply has too many parameters and the data can seemingly be fitted to any degree of accuracy. While it is readily obvious that fitting three points with a polynomial of order 2 (or higher order) shall yield a perfect fit, there are situations in which it might not be readily obvious whether a fit function (i.e., a model) might have too many parameters. In such cases, the χ^2 shall be very low and the uncertainty on some the model parameters very large, but there is no simple recipe to evaluate whether the data quality warrants the given model complexity. Simply put, the frequentist paradigm does not provide a statistic to readily evaluate whether a model might feature too many or too few parameters. In contrast, we will see in Chapter 7 that the Bayesian paradigm provides a natural Occam's razor mechanism to identify unnecessary model parameters.

Figure 6.18 illustrates how the discriminating power of a fit (§6.4.1) and associated χ^2 increases in inverse proportion to the size of measurements errors. For illustrative purposes, data (x_i, f_i) were generated with Gaussian deviates to obey a parabolic dependence $f(x) = b_0 + b_1 x + b_2 x^2$ with relative measurement errors, σ/f, ranging between 0.05% and 50%, as shown in panels (a) through (e). We then proceeded as if the true functional form of

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

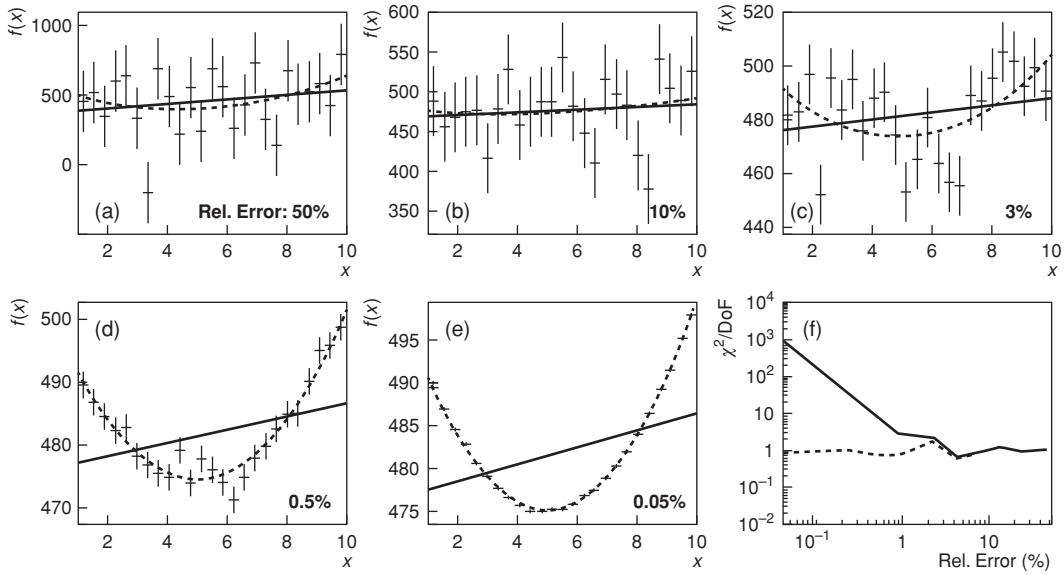


Fig. 6.18 Illustration of the discriminating power of the χ^2-test with decreasing errors. All simulated data were generated with the same quadratic function but such that measurement errors ranged from 50% in panel (a) to 0.05% in panel (e). Panel (f) displays the χ^2/DoF of the linear and quadratic fits to the data. The quadratic fit consistently yields $\chi^2/\text{DoF} \approx 1$ but the χ^2/DoF of the linear fits rises to very large values for relative errors below 1%, and the null hypothesis stipulating that a linear dependency is consistent with the data is readily rejected for such small errors.

the data was unknown and fitted the data with a linear model $f_1(x) = a_0 + a_1 x$ as a null hypothesis, and a quadratic form $f_2(x) = a_0 + a_1 x + a_2 x^2$ as an alternative hypothesis. The χ^2 of both fits were plotted as a function of the relative error in panel (e). One finds that as the relative errors decrease the quality of the linear fit (solid line) progressively worsens and yields large χ^2/DoF, whereas, in contrast, the quadratic fit consistently improves and yields more or less constant values $\chi^2/\text{DoF} \approx 1$. The rejection of the linear model is not warranted for large measurement errors but can be readily accomplished with small errors. The power of the test $1 - \beta$ to reject the linear fit as the alternative hypothesis is plotted in Figure 6.20 as a function of the relative measurement errors. One finds that for relative measurement errors smaller than 0.5%, the χ^2 distributions of the linear and quadratic fits are very well separated, thereby yielding (within the precision of the calculation shown) a power $1 - \beta$ essentially equal to unity. In situations where the variables y_i represents a yield N_i (e.g., number of counts, a number of events, etc.) and the function $f(x_i|\vec{\theta})$ amounts to the expected rate of occurrences $N p_i(x_i)$, with N being the total number of observed events, the test becomes a Pearson's χ^2 test with

$$\chi^2 = \sum_{i=1}^{N} \frac{(N_i - N(x_i|\vec{\theta}))^2}{N(x_i|\vec{\theta})}, \tag{6.110}$$

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

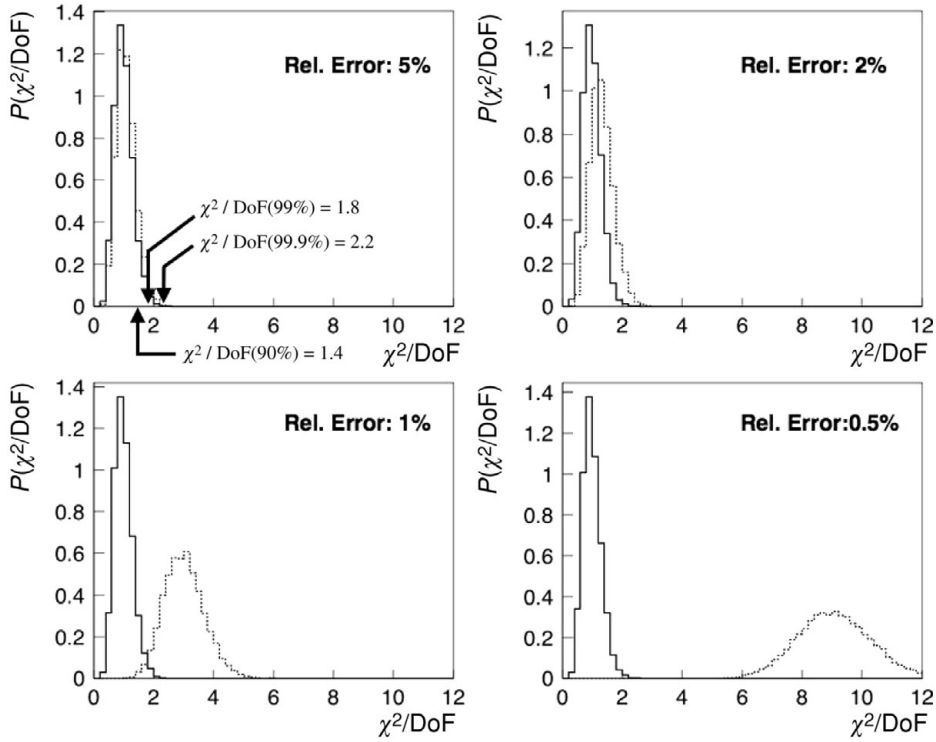


Fig. 6.19 χ^2-Distributions obtained with the linear and quadratic fits shown in Figure 6.18 for 10,000 repeated simulated measurements with relative errors ranging from 0.5% to 5% as shown. The confidence level shown in the top left panel are for 22 degrees of freedom. The rejection power of the linear hypothesis (dotted line) clearly increases when measurement errors are reduced.

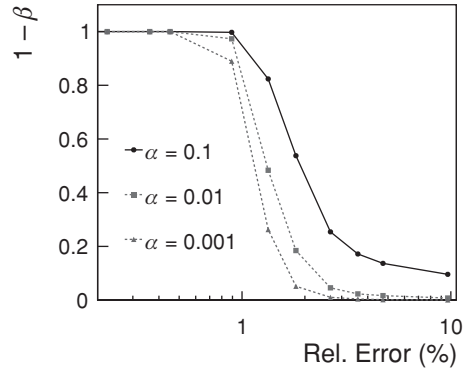


Fig. 6.20 Power of the χ^2 test shown in Figure 6.19 for three significance levels as a function of the relative errors used in the simulations.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

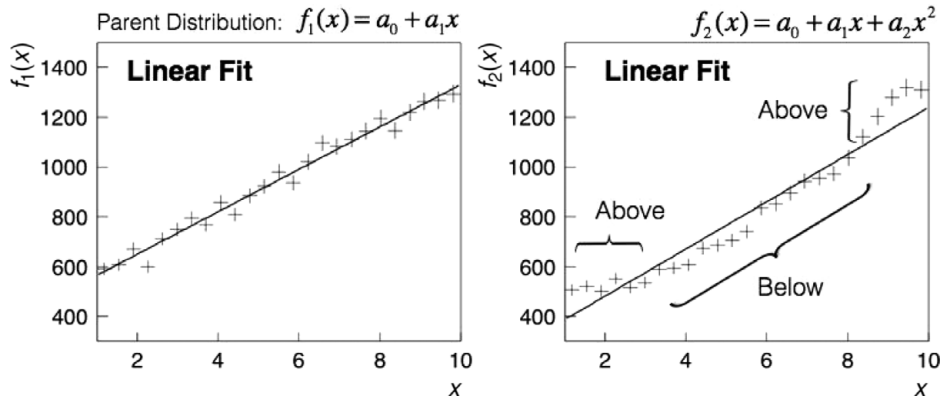


Fig. 6.21 The principle of the **Run Test** is here illustrated with a linear fit of two distinct parent distributions. One expects that statistical fluctuations should yield points randomly distributed below and above the fit if the parameterization used is a reasonable model of the parent distribution, as shown on the left for f_1. This basic expectation is clearly not met with the linear fit of the quadratic function f_2.

where one has replaced the variances σ_i^2 by $N(x_i|\vec{\theta})$. This modified χ^2 functions fluctuates according to a χ^2 distribution provided the values N_i fluctuate according to Gaussian distributions or Poisson fluctuations in the large N limit.

The χ^2-test is often not sufficient to convincingly accept or reject the null hypothesis. For instance, structures in residuals (i.e., difference between the observed and values predicted by a fit) may serve as an indicator of poor fit quality. Indeed, residuals considered as a function of the independent variable x should show a random (Gaussian if the measured values have Gaussian deviates) distribution around zero to be trustworthy. Significant structures or orderings of the residues thus provide an indicator that the null hypothesis might need to be rejected. An additional goodness-of-fit test known as a **run test** may then be used to discriminate situations where the χ^2 is relatively small but the model visually incompatible with the data, as illustrated in Figure 6.21. The run test, discussed for instance by Barlow [26], determines the probability that such seemingly unlikely situation (i.e., large nonrandom residue structures) might occur and thus enables additional power toward the rejection of a fit and the model on which it is based.

6.6.6 Histograms' Compatibility

Just as it is commonly of interest to compare predictions of a model with data, it is often useful to compare two or more histograms against one another. One may, for instance, compare data distributions produced by different experiments and examine their degree of compatibility. Alternatively, one may wish to compare results produced by a single experiment but across the life of the experiment. This is particularly useful in identifying whether the conditions or performance of the experiment evolve over time.

Let us assume we wish to compare two histograms, a control distribution $\vec{H}^{(0)}$ and a sample $\vec{H}^{(1)}$, in order to establish whether the two samples (distributions) have the same

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

parent distribution. Let us assume the two histograms were obtained with the same number of bins m and are both normalized to represent average number of entries per event. If the two histograms have the same parent distribution, their difference calculated bin-by-bin, $\Delta_i = H_i^{(0)} - H_i^{(1)}$, should average to zero with a variance $\sigma_i^2 = (\sigma_i^{(0)})^2 + (\sigma_i^{(1)})^2$ where $(\sigma_i^{(0)})^2$ and $(\sigma_i^{(1)})^2$ are the variances of the respective bins. Let us further assume that the variances are either known or can be estimated with reasonably high precision. The sum $\sum_i (\Delta_i/\sigma_i)^2$ then constitutes a χ^2 variable distributed according to a χ^2-distribution. One can thus use a χ^2-test with m degrees of freedom to evaluate the compatibility of the two histograms to any desired level of significance.

Few remarks are in order. First, note that if the variances $(\sigma_i^{(1)})^2$ of the m bins of sample $\vec{H}^{(1)}$ are not very well known, one can set $\sigma_i^2 = 2(\sigma_i^{(0)})^2$ since the null hypothesis stipulates that the two histograms have same parent distribution. Additionally, note that if the two histograms are normalized to have the same integral rather than being normalized *per event*, the same procedure may be applied but the number of degrees of freedom is then reduced by one unit. If the two distributions have unequal numbers of bins or points, it might be possible to instead use the Kolmogorov test, presented in the next section.

6.6.7 Kolmogorov–Smirnov Test

The **Kolmogorov–Smirnov test** (KS test), also often called the **Kolmogorov test**, is a nonparametric test of the equality of continuous, one-dimensional probability distributions. It can be used to test the compatibility of a sample distribution with a reference probability distribution (one-sample KS test), or to compare two samples (two-sample KS test). Unlike the histogram compatibility test presented in the previous section, the KS test makes no assumption on the number of points (bins) or the variances of the samples.

The test is based on the Kolmogorov–Smirnov statistic, which expresses the distance or difference between the empirical distribution function (EDF) of the sample (defined below) and the cumulative distribution function (CDF) of the reference distribution (or alternatively, the sample and control empirical distributions).

The null hypothesis is that the sample being tested is determined by the reference distribution (one-sample KS test) or that two samples being compared have the same parent distribution (two-sample case). The distributions considered under the null hypothesis are continuous but otherwise unrestricted. Formally, the EDF $F_n(x)$ of a sample is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x) \tag{6.111}$$

where n is the size of the sample and $I(x_i < x)$ is called the **indicator function**, defined such that

$$I(x_i < x) = 1 \qquad \text{if } x_i \leq x, \tag{6.112}$$

$$0 \qquad \text{otherwise.} \tag{6.113}$$

This is equivalent to sorting the measured values of a sample in ascending order and summing the number of sampled values smaller than x. Alternatively, if the sample is provided

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

Table 6.3 Critical values d_α of the KS test listed for selected significance levels

α	20%	10%	5%	2.5%	1%	0.5%	0.1%
d_α	1.07	1.22	1.36	1.48	1.63	1.73	1.95

in the form of an m bins histogram $\vec{H} = (H_1, \ldots, H_m)$, where the values of x are implicitly sorted in ascending order, rather than a list of values as earlier, the function $F_n(x)$ becomes

$$F_k = \frac{1}{m} \sum_{i=1}^{k} H_i, \tag{6.114}$$

where $k = 1, \ldots, m$. For a given cumulative distribution function, $F(x)$, the Kolmogorov–Smirnov statistic is defined as

$$D_n = \sup_x |F_n(x) - F(x)|\,, \tag{6.115}$$

which corresponds to the maximum deviation between the cumulative distributions. One finds that $d_n = \sqrt{n} D_n$ converges to the Kolmogorov distribution [129], and one may reject the null hypothesis at a level of significance α for

$$d_n \equiv \sqrt{n} D_n > d_\alpha. \tag{6.116}$$

Several critical values d_α and their respective levels of significance are listed in Table 6.3.

When the KS test is used to compare to measured samples, the Kolmogorov–Smirnov statistic becomes

$$D_{n,n'} = \sup_x \left|F_{1,n}(x) - F_{2,n'}(x)\right|. \tag{6.117}$$

The null hypothesis is then rejected at level α if

$$D_{n,n'} > d_\alpha \sqrt{\frac{n + n'}{nn'}} \tag{6.118}$$

with values of the coefficient d_α, listed in Table 6.3.

Example of the KS Test

Two samples of 1,000 events each, shown in Figure 6.22b, were generated according to parent distributions

$$f_1(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{(dashed line)}, \tag{6.119}$$

$$f_2(x) = [1 + 0.2(x - \mu)] \times f_1(x) \quad \text{(dotted line)}, \tag{6.120}$$

with $\mu = 3$ and $\sigma = 0.7$, shown in Figure 6.22a. While it is straightforward to distinguish the two (parent) distributions (or with infinite statistics plotted on a log scale), it is rather difficult to determine whether the two samples shown in Figure 6.22b originate from the

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

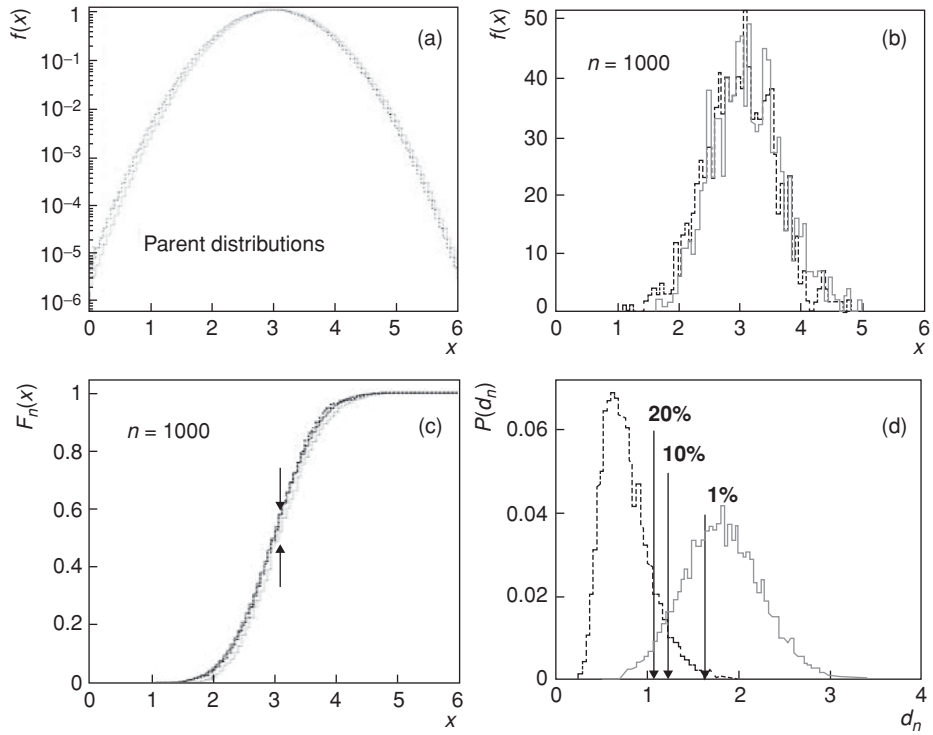


Fig. 6.22 Example of application of the Kolmogorov–Smirnov test. (a) Parent distributions f_1 and f_2 listed in the text. (b) 1,000 events samples generated with distributions f_1 and f_2. (c) Cumulative distributions. (d) Relative probability of the maximum difference d_n for both types of samples.

same or distinct parent distributions. We thus carry out a KS test to establish whether either sample is compatible with the parent f_1. Cumulative distributions of the two samples are compared to the cumulative distribution F_1 of the first parent distribution f_1 in Figure 6.22c. One observes that the cumulative distribution of the first sample essentially overlaps with F_1, whereas the cumulative distribution (dotted line) of the second sample exhibit significant differences with the parent cumulative distribution F_1. The largest difference, of order 0.016, multiplied by $\sqrt{1{,}000}$, is found to be 1.19, which is in excess of the 20% confidence level of 1.07 and thus constitutes grounds for excluding the null hypothesis that the second sample might originate from the same parent distribution as the first sample.

The "experiment" was then repeated 100,000 times: for each experiment, we generated a sample of 1,000 values and created their cumulative distributions. The maximum differences between these and the cumulative distribution F_1 are plotted in Figure 6.22d where arrows indicate the 20%, 10%, and 1%. One finds the distribution of d_n obtained with the first sample (dashed line) covers values typically smaller than unity whereas the second distribution lies mostly above this value. The test thus indeed enables a distinction between parent distributions to be made even when the difference may not be readily obvious to the human eye.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

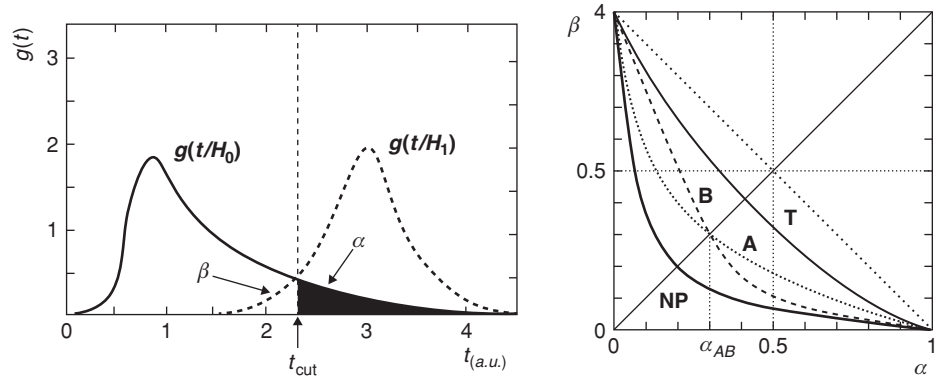


Fig. 6.23 (Left) A symmetry exists between parameters α and β. If t_{cut} is raised β increases while α decreases, and conversely. Right: Tests viewed as functions of both α and β.

6.7 Test Optimization and the Neyman–Pearson Test

6.7.1 Test Optimization

The choice between tests usually relies on a comparison of their power curves for a given value of the significance level α. While the value of α determines the risk of an error of the first kind (false rejection of H_0), the value of $\beta = 1 - p$ controls the risk of an error of the second kind (contamination or incorrect acceptance of H_0). It is common practice to first choose α and thence find which test provides the best power. However, there is clearly a certain form of symmetry between α and β, as shown in the left panel of Figure 6.23, and one obviously wishes to minimize both types of errors. It is thus useful to regard tests as simultaneous functions of both α and β, as illustrated in the right panel of the figure. The function $\beta_\alpha = 1 - \alpha$ corresponds to the dotted diagonal line below which, for a fixed value of α, all other tests have a large power than α, and can thus be considered consistent. By extension, all tests described by curves below this straight line are therefore strictly unbiased (for all values of α). Actual tests tend to follow a curve such as **T** (typical). For specific values of θ_0 and θ_1, decreasing α usually requires raising the value t_{cut}, used to define the critical boundary, and this then simultaneously increases β, thereby reducing the power of the test. Conversely, efforts to reduce β for a specific test will end up producing a larger value of α. The question thus arises as to whether there might exists some tests **A** or **B**, as illustrated in Figure 6.23, that are more acceptable than **T** because they reduce the values of both α and β simultaneously. In 1933, Jerzy Neyman and Egon S. Pearson established that given two hypotheses H_0 and H_1, there exists a most powerful test, now known as the Neyman–Pearson (NP) test, which is at least as good as any other tests for all values of α and β. This test is based on a ratio of likelihood functions and is described below. Because the implementation of the Neyman–Pearson test is often cumbersome or does not exist if hypotheses are composite, then it is commonly acceptable to resort to lower

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

performing tests, such as tests **A** and **B** in Figure 6.23. Obviously, test **A** should be chosen, in this example, if $\alpha < \alpha_{AB}$ is desired given that it enables a higher power in that range of α. Conversely, **B** should be used if a value $\alpha > \alpha_{AB}$ is deemed acceptable. Comparing tests according to their simultaneous performance as a function of α and β permits the selection of the best test for a given job or analysis. In general, deciding on a specific value of α (and consequently β) requires knowing the cost or consequences of a wrong decision. Errors of the first kind might be more acceptable in certain experiments, while others might better tolerate errors of the second kind.

6.7.2 Formulation of the Neyman–Pearson Test

In order to formulate the Neyman–Pearson test, let us consider a measured data sample $\vec{x} = (x_1, \ldots, x_n)$ in its entirety, and let us assume it is determined by a PDF $f_n(\vec{x}|\theta)$. For simplicity's sake, we will also assume that the parameter θ takes a value $\theta = \theta_0$ if the null hypothesis H_0 is correct and a value $\theta = \theta_1$ according to the alternative hypothesis H_1. The goal of the test is to find out whether the data $\vec{x}$ are consistent with the null hypothesis H_0, or if H_0 should be rejected in favor of H_1. We will additionally assume that $f_n(\vec{x}|\theta)$ is a continuous and smooth function of θ between θ_0 and θ_1.

Given a desired significance level α, one expects that there is a portion of the domain of $\vec{x}$ which can be identified as a critical (rejection) region D_c such that

$$\int_{D_c} f_n(\vec{x}|\theta_0)\, d\vec{x} = \alpha, \tag{6.121}$$

$$\int_{D_c} f_n(\vec{x}|\theta_1)\, d\vec{x} = 1 - \beta. \tag{6.122}$$

Our goal is to find the region D_c, given α, which maximizes the power $1 - \beta$. To this end, it is convenient to rewrite Eq. (6.122) as

$$1 - \beta = \int_{D_c} \frac{f_n(\vec{x}|\theta_1)}{f_n(\vec{x}|\theta_0)} f_n(\vec{x}|\theta_0)\, d\vec{x}, \tag{6.123}$$

$$= \int_{D_c} r_n(\vec{x}|\theta_0, \theta_1) f_n(\vec{x}|\theta_0)\, d\vec{x}, \tag{6.124}$$

where we have introduced the ratio $r_n(\vec{x}|\theta_0, \theta_1)$ as

$$r_n(\vec{x}|\theta_0, \theta_1) = \frac{f_n(\vec{x}|\theta_1)}{f_n(\vec{x}|\theta_0)}. \tag{6.125}$$

Inspection of Eq. (6.124) reveals that it corresponds to the expectation value of the function r_n over the domain D_c according to the null hypothesis $H_0(\theta_0)$:

$$1 - \beta = \mathrm{E}_{D_c}\left[r_n(\vec{x}|\theta_0, \theta_1)\right], \tag{6.126}$$

$$= \mathrm{E}_{D_c}\left[\left.\frac{f_n(\vec{x}|\theta_1)}{f_n(\vec{x}|\theta_0)}\right|_{\theta=\theta_0}\right]. \tag{6.127}$$

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

The power is consequently optimal if and only if the D_c is a fraction of the domain of $\vec{x}$ which contains the maximum value of the ratio r_n. This is easily guaranteed by selecting points such that

$$r_n(\vec{x}|\theta_0, \theta_1) = \frac{f_n(\vec{x}|\theta_1)}{f_n(\vec{x}|\theta_0)} \geq r_\alpha, \tag{6.128}$$

where the constant r_α must be chosen to correspond to the desired level of significance α. The Neyman–Pearson test is thus formulated as follows:

- If $r_n(\vec{x}|\theta_0, \theta_1) > r_\alpha$, reject H_0, and choose H_1.
- If $r_n(\vec{x}|\theta_0, \theta_1) \leq r_\alpha$, choose H_0.

It is important to note that the function $r_n(\vec{x}|\theta_0, \theta_1)$ is a statistic evaluated for specific points $\vec{x}$ and corresponds, in fact, to the ratio of the likelihood of $\vec{x}$ according to the hypotheses H_1 and H_0. Clearly, for the test to be usable it should be possible to calculate the likelihood function $f_n(\vec{x}|\theta)$ for both $\theta = \theta_0$ and $\theta = \theta_1$ at all points of the space of interest. This, in turn, implies that both H_1 and H_0 must be simple hypotheses. If these conditions are verified, the Neyman–Pearson test provides by construction the best power $1 - \beta$ against the chosen value of the significance level α and the given data $\vec{x}$. Extensions of the Neyman–Pearson test to composite hypotheses are possible but not guaranteed to yield an optimal test [116].

6.7.3 Comparison of Sample Means with the Neyman–Pearson Test

Let us reexamine the situation, already discussed in §6.6.3, where a change is made to a process that yields entities with a certain property x. The purpose of the test is to determine whether the change has in fact affected the property value x. We once again assume that the mean value of x prior to the change is known to be μ, and that measurement errors are Gaussian deviates with a known standard deviation σ. The test is conducted by measuring a sample of N values x_i and taking their mean $\bar{x}$. The null hypothesis H_0 shall be that the change in process has not altered the mean, that is, the measured mean $\bar{x}$ shall have an expectation value $\mu_0 = \mu$. The alternative hypothesis shall then be that the mean has changed and now has a value μ_1.

Let us proceed to build a Neyman–Pearson test. Given that measurements are assumed to yield Gaussian fluctuations with standard deviation σ, the test may then be formulated based on the ratio r of the likelihood functions of $\vec{x}$ given hypotheses H_1 and H_0:

$$r = \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu_1)^2\right]}{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu_0)^2\right]}. \tag{6.129}$$

Introducing the sample mean $\bar{x} = \frac{1}{N} \sum_i^N x_i$, Eq. (6.129) simplifies to

$$r = \exp\left[-\frac{N}{2\sigma^2} (\mu_1 - \mu_0)(\mu_0 + \mu_1 - 2\bar{x})\right] \tag{6.130}$$

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

Taking the log on both sides, one writes

$$\frac{2\sigma^2}{N}\ln r = -(\mu_1 - \mu_0)(\mu_0 + \mu_1 - 2\bar{x}). \tag{6.131}$$

The expectation value of $\bar{x}$ depends on which of the two hypotheses is correct. If H_0 is correct, one gets $\mathrm{E}[\bar{x}] = \mu_0$, whereas if H_1 is correct, one has $\mathrm{E}[\bar{x}] = \mu_1$. Either way, the variance of $\bar{x}$ is σ^2/N. It is then useful to define a deviate z according to

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}}, \tag{6.132}$$

which, by construction, is Gaussian distributed with unit standard deviation and expectation values

$$E[z] = \begin{cases} 0 & \text{for } H_0 \\ \frac{\mu_1 - \mu_0}{\sigma/\sqrt{N}} & \text{for } H_1 \end{cases} \tag{6.133}$$

Substituting in Eq. (6.131), one gets, after some simple algebra

$$r' \equiv \frac{2\sigma^2 \ln r}{N(\mu_1 - \mu_0)^2} = -\left(1 - 2\frac{z\sigma}{\sqrt{N}(\mu_1 - \mu_0)}\right), \tag{6.134}$$

where, for notational convenience, we introduced the statistics r' which clearly has expectations values

$$\mathrm{E}[r'] = \mathrm{E}\left[\frac{2\sigma^2 \ln r}{N(\mu_1 - \mu_0)^2}\right] = \begin{cases} -1 & \text{for } H_0, \\ +1 & \text{for } H_1 \end{cases} \tag{6.135}$$

Since z is Gaussian distributed with unit standard deviation, the statistics r' is also Gaussian distributed but its standard deviation decreases proportionally to $\sqrt{N}$. For a given σ, the PDFs $p(r'|H_i)$, $i = 0, 1$, are thus increasingly separated for increasing N, as illustrated in Figure 6.24. For sufficiently large N, it then becomes possible to distinguishes which of the two hypotheses is the best. The Neyman–Pearson test requires the ratio of likelihood functions for hypotheses H_1 and H_0 to exceed some value r_α, for a given significance level α. Given Eq. (6.134), this translates in a cut on z. The critical region is then defined for values $z > z_\alpha$ (assuming $\mu_1 > \mu_0$) where z_α is the α-point of the standard normal distribution. The power of the test is then

$$1 - \beta = \int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z - \Delta)^2}{2}\right) dz, \tag{6.136}$$

where

$$\Delta = \frac{\sqrt{N}}{\sigma}(\mu_1 - \mu_0). \tag{6.137}$$

One thus find that the power increases with $\mu_1 - \mu_0$ as well as N, as expected from Figure 6.24.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

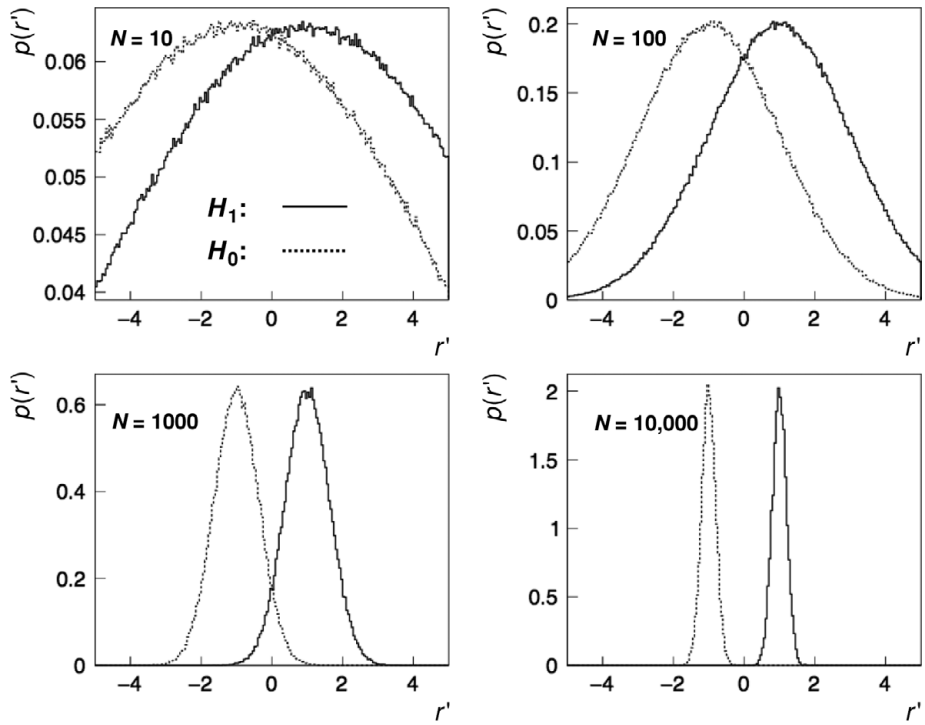


Fig. 6.24 Probability density of r' corresponding to the scaled ratio of likelihood of hypotheses H_1 and H_0, where $r' \equiv 2\sigma^2 \ln r/N(\mu_1 - \mu_0)^2$.

Other tests of the equivalence of two samples may be applied, but one can show that the aforementioned Neyman–Pearson test indeed yields the most powerful test and discrimination power between the null and alternative hypotheses [116].

6.8 Appendix 1: Derivation of Student's *t*-Distribution

Student's t-distribution, Eq. (3.161), expresses the probability (density) of observing a certain value of t defined as

$$t = \frac{x - \mu}{\hat{\sigma}}, \tag{6.138}$$

where x is a continuous random variable Gaussian distributed with a mean μ and a standard deviation σ, while $\hat{\sigma}$ constitutes an estimates of σ obtained from a sample $\{x_i\}$ according to

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \tag{6.139}$$

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

if the mean μ is known and

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2 \tag{6.140}$$

if calculated in terms of an estimate $\hat{\mu}$ of the mean.

In order to derive Student's t-distribution, it is convenient to rewrite t as

$$t = \frac{(x-\mu)}{\sigma}\frac{\sigma}{\hat{\sigma}} = \frac{(x-\mu)/\sigma}{\sqrt{(n\hat{\sigma}^2/\sigma^2)/n}} = \frac{z}{\sqrt{\chi^2/n}}, \tag{6.141}$$

where, in the second equality, we introduced $z = (x-\mu)/\sigma$ and $\chi^2 = n\hat{\sigma}^2/\sigma^2$, with n equal to N or $N-1$ as appropriate. By construction, z and χ^2 should be distributed as a standard normal (i.e., $\sigma = 1$) and χ^2-distribution for n degrees of freedom, respectively. The joint probability of observing values z and χ^2 is

$$p(z, \chi^2|n) = \frac{e^{-z^2/2}}{\sqrt{2\pi}} \times \frac{\left(\chi^2\right)^{n/2-1} e^{-\chi^2/2}}{2^{n/2}\Gamma(n/2)}. \tag{6.142}$$

One obtains a distribution for t by integrating all values of z and χ^2 meeting the condition $t = z/\sqrt{\chi^2/n}$:

$$p_t(t|n) = \int d\chi^2 \int dz\, \delta(t - z/\sqrt{\chi^2/n}) \frac{e^{-z^2/2}}{\sqrt{2\pi}} \times \frac{\left(\chi^2\right)^{n/2-1} e^{-\chi^2/2}}{2^{n/2}\Gamma(n/2)}, \tag{6.143}$$

$$= \frac{1}{\sqrt{n\pi}} \frac{\Gamma(n+1/2)}{\Gamma(n/2)} \frac{1}{(1+t^2/n)^{(n+1)/2}}. \tag{6.144}$$

Exercises

6.1 Consider an experiment measuring the mass and width of a short-lived resonance based on an invariant mass spectrum. Assume the estimates are normally distributed with a covariance matrix

$$V = \begin{pmatrix} \sigma_M^2 & \rho\sigma_M\sigma_\Gamma \\ \rho\sigma_M\sigma_\Gamma & \sigma_\Gamma^2 \end{pmatrix}.$$

Show that one can establish a symmetric β confidence interval for the mass of the resonance M_0 conditional on its width according to

$$\hat{M} - \frac{\rho\sigma_M}{\sigma_\Gamma}\left(\hat{\Gamma} - \Gamma_0\right) - Z_{(1-\beta)/2}\sigma_M\sqrt{1-\rho^2} \leq M_0$$
$$\leq \hat{M} - \frac{\rho\sigma_M}{\sigma_\Gamma}\left(\hat{\Gamma} - \Gamma_0\right) - Z_{(1+\beta)/2}\sigma_M\sqrt{1-\rho^2} \tag{6.145}$$

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

where $\hat{M}$ and $\hat{\Gamma}$ are estimates obtained by the experiment for the mass and width of the resonance, respectively; $Z_{(1\pm\beta)/2}$ correspond to the α and β points of the probability of the mass M given by

$$P(Z_\alpha \leq Z \leq Z_{\alpha+\beta}|\Gamma) = \beta,$$

where

$$Z = \frac{M - \left[M_0 + \frac{\rho\sigma_M}{\sigma_\Gamma}(\Gamma = \Gamma_0)\right]}{\sigma_M\sqrt{1-\rho^2}}.$$

Discuss how this result can be used to constrain the mass M_0 if a prior estimate Γ^* for the width Γ_0 is known from a previous experiment.

6.2 The melting point of two material specimens are determined by actually melting a fraction of each of the specimens. One finds that the first specimen has a melting point of 303 ± 5 K, while the second specimen is determined to have a melting point of 307 ± 4 K. Assuming measurements of temperatures yield Gaussian distributions with the quoted standard deviations, establish if the two samples are likely to be the same substance based on a Gaussian test using a confidence level of 5%.

6.3 A steel company changes the composition and shape of the steel beams it manufactures in order to reduce production costs. However, the company needs to demonstrate that the new beams feature the same quality standards as the older version. If their measurements show that the original beams can withstand a tensile stress of 412 ± 20 MPa[3], and the new beams have a mean strength of 393 ± 18 MPa, would you conclude that the quality of the new beams is necessarily inferior to that of the original beams?

6.4 An experiment searching for proton decays based on a sample of 2,000,000 kg of hydrogen for one year reports the observation of five background events. Calculate the 99% confidence level for the number of proton decays, and determine an upper limit for the half-life of the proton.

6.5 A very high-precision magnetic spectrometer is used to measure the momentum of electrons produced by beta decay of tritium nuclei. Researchers use the end-point of the electrons' momentum spectrum to determine the mass of the neutrino and find a raw mass value of -0.1 eV. Assuming the spectrometer has a Gaussian response with a standard deviation $\sigma = 0.05$ eV, determine a 90% upper limit on the mass of the neutrino using (a) a basic upper limit belt and (b) the unified approach. Also determine the minimum mass the neutrino must have for this experiment to claim a discovery with a 5σ significance level. Compare your estimates to those of the KATRIN experiment [142].

6.6 Consider an experiment for which you must determine the particle detection efficiency as a function of momentum. Assume it is possible to determine the efficient

[3] MPa = Mega-Pascal.

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press

based on the embedded track technique. Use the table below to estimate the detection efficiency and its error using (a) a simple Gaussian approximation and (b) the Feldman–Cousins ordering technique.

$p_\perp$ range	Embedded	Found
0.5–1.0	10,000	9,945
1.0–2.0	9,010	8,456
2.0–3.0	7,890	7,333
3.0–5.0	5,550	5,211
5.0–10.0	4,442	4,012

6.7 An experiment produces 100 data points, which you proceed to fit with linear and quadratic models. You obtain a $\chi^2/\text{DoF} = 2.5$ with the linear fit whereas the quadratic fit yields $\chi^2/\text{DoF} = 2.1$. Can you reject the linear hypothesis? Can you accept the quadratic hypothesis?

https://doi.org/10.1017/9781108241922.008 Published online by Cambridge University Press