Rejecting H_0 if it is true is called an error of the first kind. The probability for this to occur is called the *significance level* of the test, α, which is often chosen to be equal to some pre-specified value. It can also happen that H_0 is false and the true hypothesis is given by some alternative, H_1. If H_0 is accepted in such a case, this is called an error of the second kind. The probability for this to occur, β, depends on the alternative hypothesis, say, H_1, and $1 - \beta$ is called the *power* of the test to reject H_1.

In High Energy Physics the components of $\boldsymbol{x}$ might represent the measured properties of candidate events, and the acceptance region is defined by the cuts that one imposes in order to select events of a certain desired type. That is, H_0 could represent the signal hypothesis, and various alternatives, H_1, H_2, etc., could represent background processes.

Often rather than using the full data sample $\boldsymbol{x}$ it is convenient to define a *test statistic*, t, which can be a single number or in any case a vector with fewer components than $\boldsymbol{x}$. Each hypothesis for the distribution of $\boldsymbol{x}$ will determine a distribution for t, and the acceptance region in $\boldsymbol{x}$-space will correspond to a specific range of values of t. In constructing t one attempts to reduce the volume of data without losing the ability to discriminate between different hypotheses.

In particle physics terminology, the probability to accept the signal hypothesis, H_0, is the selection efficiency, *i.e.*, one minus the significance level. The efficiencies for the various background processes are given by one minus the power. Often one tries to construct a test to minimize the background efficiency for a given signal efficiency. The *Neyman–Pearson lemma* states that this is done by defining the acceptance region such that, for $\boldsymbol{x}$ in that region, the ratio of p.d.f.s for the hypotheses H_0 and H_1,

$$\lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|H_0)}{f(\boldsymbol{x}|H_1)} , \tag{32.26}$$

is greater than a given constant, the value of which is chosen to give the desired signal efficiency. This is equivalent to the statement that (32.26) represents the test statistic with which one may obtain the highest purity sample for a given signal efficiency. It can be difficult in practice, however, to determine $\lambda(\boldsymbol{x})$, since this requires knowledge of the joint p.d.f.s $f(\boldsymbol{x}|H_0)$ and $f(\boldsymbol{x}|H_1)$. Instead, test statistics based on *neural networks* or *Fisher discriminants* are often used (see [10]).

32.2.2. *Goodness-of-fit tests* :

Often one wants to quantify the level of agreement between the data and a hypothesis without explicit reference to alternative hypotheses. This can be done by defining a *goodness-of-fit statistic*, t, which is a function of the data whose value reflects in some way the level of agreement between the data and the hypothesis. The user must decide what values of the statistic correspond to better or worse levels of agreement with the hypothesis in question; for many goodness-of-fit statistics there is an obvious choice.

The hypothesis in question, say, H_0, will determine the p.d.f. $g(t|H_0)$ for the statistic. The goodness-of-fit is quantified by giving the p-value, defined as the probability to find t in the region of equal or lesser compatibility with H_0 than the level of compatibility observed with the actual data. For example, if t is defined such that large values correspond to poor agreement with the hypothesis, then the p-value would be

$$p = \int_{t_{\rm obs}}^{\infty} g(t|H_0)\, dt \; , \tag{32.27}$$

where $t_{\rm obs}$ is the value of the statistic obtained in the actual experiment. The p-value should not be confused with the significance level of a test or the confidence level of a confidence interval (Section 32.3), both of which are pre-specified constants.

The p-value is a function of the data and is therefore itself a random variable. If the hypothesis used to compute the p-value is true, then for continuous data, p will be uniformly distributed between zero and one. Note that the p-value is not the probability for the hypothesis; in frequentist statistics this is not defined. Rather, the p-value is the probability, under the assumption of a hypothesis H_0, of obtaining data at least as incompatible with H_0 as the data actually observed.

When estimating parameters using the method of least squares, one obtains the minimum value of the quantity χ^2 (32.13), which can be used as a goodness-of-fit statistic. It may also happen that no parameters are estimated from the data, but that one simply wants to compare a histogram, *e.g.*, a vector of Poisson distributed numbers $\boldsymbol{n} = (n_1, \ldots, n_N)$, with a hypothesis for their expectation values $\nu_i = E[n_i]$. As the distribution is Poisson with variances $\sigma_i^2 = \nu_i$, the χ^2 (32.13) becomes *Pearson's* χ^2 *statistic*,

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} \; . \tag{32.28}$$

If the hypothesis $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$ is correct and if the measured values n_i in (32.28) are sufficiently large (in practice, this will be a good approximation if all $n_i > 5$), then the χ^2 statistic will follow the χ^2 p.d.f. with the number of degrees of freedom equal to the number of measurements N minus the number of fitted parameters. The same holds for the minimized χ^2 from Eq. (32.13) if the y_i are Gaussian.

Alternatively one may fit parameters and evaluate goodness-of-fit by minimizing $-2 \ln \lambda$ from Eq. (32.12). One finds that the distribution of this statistic approaches the asymptotic limit faster than does Pearson's χ^2 and thus computing the p-value with the χ^2 p.d.f. will in general be better justified (see [9] and references therein).

Assuming the goodness-of-fit statistic follows a χ^2 p.d.f., the p-value for the hypothesis is then

$$p = \int_{\chi^2}^{\infty} f(z; n_{\rm d})\, dz \; , \tag{32.29}$$

where $f(z; n_{\rm d})$ is the χ^2 p.d.f. and $n_{\rm d}$ is the appropriate number of degrees of freedom. Values can be obtained from Fig. 32.1 or from the CERNLIB routine PROB. If the

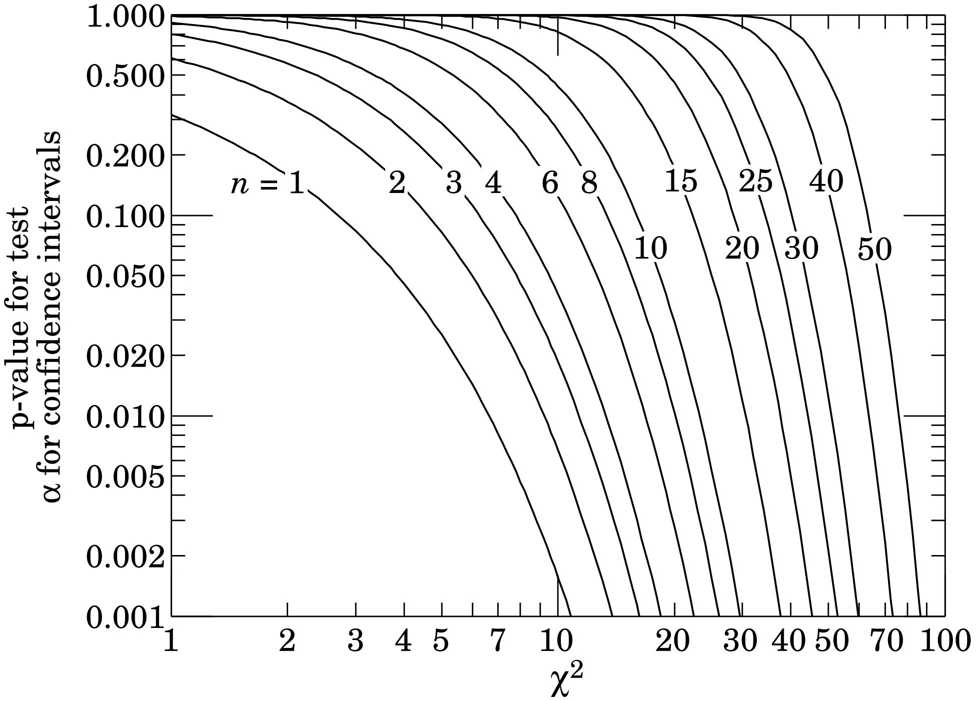

Figure 32.1: One minus the χ^2 cumulative distribution, $1-F(\chi^2;n)$, for n degrees of freedom. This gives the p-value for the χ^2 goodness-of-fit test as well as one minus the coverage probability for confidence regions (see Sec. 32.3.2.3).

conditions for using the χ^2 p.d.f. do not hold, the statistic can still be defined as before, but its p.d.f. must be determined by other means in order to obtain the p-value, *e.g.*, using a Monte Carlo calculation.

If one finds a χ^2 value much greater than n_{d} and a correspondingly small p-value, one may be tempted to expect a high degree of uncertainty for any fitted parameters. Although this may be true for systematic errors in the parameters, it is not in general the case for statistical uncertainties. If, for example, the error bars (or covariance matrix) used in constructing the χ^2 are underestimated, then this will lead to underestimated statistical errors for the fitted parameters. But in such a case an estimate $\hat{\theta}$ can differ from the true value θ by an amount much greater than its estimated statistical error. The standard deviations of estimators that one finds from, say, equation (32.11) reflect how widely the estimates would be distributed if one were to repeat the measurement many times, assuming that the measurement errors used in the χ^2 are also correct. They do not include the systematic error which may result from an incorrect hypothesis or incorrectly estimated measurement errors in the χ^2.

Since the mean of the χ^2 distribution is equal to n_{d}, one expects in a “reasonable” experiment to obtain $\chi^2 \approx n_{\mathrm{d}}$. Hence the quantity χ^2/n_{d} is sometimes reported. Since the p.d.f. of χ^2/n_{d} depends on n_{d}, however, one must report n_{d} as well in order to make a meaningful statement. The p-values obtained for different values of χ^2/n_{d} are shown in Fig. 32.2.