STATISTICS

To reduce the size of this section's PostScript file, we have divided it into two PostScript files. We present the following index:

PART 1

PART 2

29.6. Exact confidence intervals

29.6.1. *Two methodologies*:

There are two different approaches to statistical inference, which we may call Frequentist and Bayesian. For the cases considered up to now, both approaches give the same numerical answers, even though they are based on fundamentally different assumptions. However, for exact results for small samples and for measurements near a physical boundary, the different approaches may yield very different confidence limits, so we are forced to make a choice. There is an enormous amount of literature devoted to the question of Bayesian vs non-Bayesian methods, most of it written by people who are fervent advocates of one or the other methodology, which often leads to exaggerated conclusions. For a reasonably balanced discussion, we recommend the following articles: by a statistician [9], and by a physicist [6].

29.6.2. *Bayesian*: The Bayesian concept of probability is not based on limiting frequencies, but is more general and includes *degrees of belief.* It can therefore be used for experiments which cannot be repeated, where a frequency definition of probability would not be applicable (for example, one can consider the probability that it will rain tomorrow). Bayesian methods also allow for a natural way to input additional information such as physical boundaries and subjective information; in fact they *require* as input the *prior distribution* for any parameter to be estimated.

The Bayesian methodology, while well adapted to decision-making situations, is not in general appropriate for the objective presentation of experimental data. This can be seen from the following example.

An experiment sets out to measure the value of a parameter whose true value cannot be negative (such as the neutrino mass squared), but let us assume that the true value is in fact zero. We should then expect that about half of the time, an unbiased experimental measurement should yield a negative (unphysical) result. Now if our experiment produces a negative result, the question arises what value to report. If we wish to make a decision concerning the most likely value of this parameter, we would use a Bayesian approach which would assure that the reported value is positive, since it would be nonsense to assert that the most likely value is one which cannot be true. On the other hand, if we wish to report an unbiased result which can be combined with other measurements, it is better to report the unphysical result. Everyone understands what it means to quote a result of, for example, $m^2 = -1.2 \pm 2.0\ \mathrm{eV}^2$. This result could then be averaged with other results, half of which would be positive, and the average would eventually converge toward zero, the true value. If Bayesian estimates are averaged, they do not converge to the true value, since they have all been forced to be positive.

29.6.3. *Frequentist, or classical confidence intervals*: As the name implies, the Frequentist concept of probability is based entirely on the limiting frequency, so it only makes sense in situations where experiments are repeatable, at least in principle. This is clearly the case for the kind of data we are concerned with, and the methods we present here are based on the Frequentist point of view.

The classical construction of exact confidence intervals which we describe here was first proposed by Neyman [10].

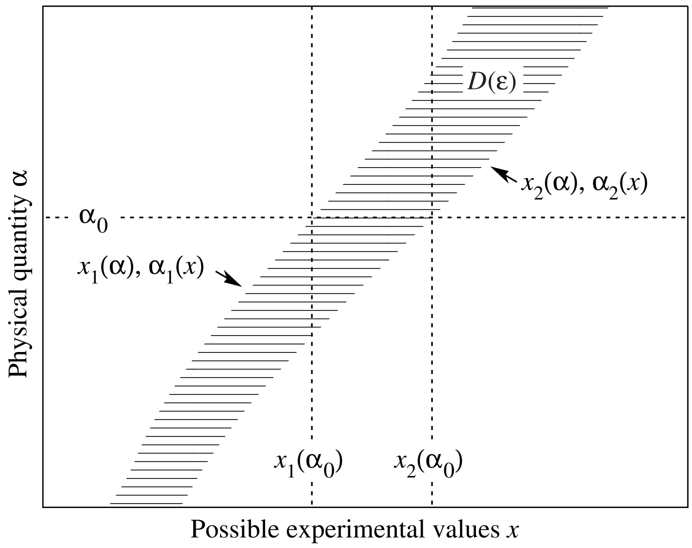

Figure 29.1: Confidence intervals for a single unknown parameter α. One might think of the p.d.f. $f(x;\alpha)$ as being plotted out of the paper as a function of x along each horizontal line of constant α. The domain $D(\varepsilon)$ contains a fraction $1-\varepsilon$ of the area under each of these functions.

We wish to set limits on the parameter α whose true value is fixed but unknown. The properties of our experimental apparatus are expressed in the function $f(x;\alpha)$ which gives the probability of observing data x if the true value of the parameter is α. This function must be known, otherwise it is impossible to interpret the results of an experiment. For a large complex experiment, this function is usually determined numerically using Monte Carlo simulation.

Given the function $f(x;\alpha)$, we can find for every value of α, two values $x_1(\alpha,\varepsilon)$ and $x_2(\alpha,\varepsilon)$ such that repeated experiments would produce results x in the interval $x_1 < x < x_2$ a fraction $1-\varepsilon$ of the time, where

$$P(x_1 < x < x_2) \;=\; 1-\varepsilon \;=\; \int_{x_1}^{x_2} f(x;\alpha)\,dx \;. \tag{29.33}$$

This situation is shown in Fig. 29.1, where the region between the curves $x_1(\alpha,\varepsilon)$ and $x_2(\alpha,\varepsilon)$ is indicated by the domain $D(\varepsilon)$. We require that the curves $x_1(\alpha,\varepsilon)$ and

$x_2(\alpha, \varepsilon)$ be monotonic functions of α, so they can be labeled either as functions of x or of α. Dropping the argument ε for simplicity, we may then label the curve $x_1(\alpha)$ as $\alpha_1(x)$ and $x_2(\alpha)$ as $\alpha_2(x)$. Now consider some arbitrary particular value of α, say α_0, as indicated in the figure. We notice from the figure that for all values of x between $x_1(\alpha_0)$ and $x_2(\alpha_0)$, it happens that α_0 lies between $\alpha_1(x)$ and $\alpha_2(x)$. Thus we can write:

$$P\big[x_1(\alpha_0) < x < x_2(\alpha_0)\big] \;=\; 1-\varepsilon \;=\; P\big[\alpha_2(x) < \alpha_0 < \alpha_1(x)\big]. \tag{29.34}$$

And since, by construction, this is true for any value α_0, we can drop the subscript 0 and obtain the relationship we wanted to establish for the probability that the confidence limits will contain the true value of α:

$$P[\alpha_2(x) < \alpha < \alpha_1(x)] \;=\; 1-\varepsilon\;. \tag{29.35}$$

In this probability statement, α_1 and α_2 are the random variables (not α), and we can verify that the statement is true, as a limiting ratio of frequencies in random experiments, for any assumed value of α. In a particular real experiment, the numerical values α_1 and α_2 are determined by applying the algorithm to the real data, and the probability statement appears to be a statement about the true value α since this is the only unknown remaining in the equation. It should however be understood that it gives only the probability of obtaining values α_1 and α_2 which include the true value of α, in an ensemble of identical experiments. Any method which gives confidence intervals that contain the true value with probability $1-\varepsilon$ (no matter what the true value of α is) is said to have *coverage*. The frequentist intervals as constructed above have *coverage* by construction. Coverage is considered the most important property of confidence intervals [6].

The condition of coverage Eq. (29.33) does not determine x_1 and x_2 completely, since any range which gives the desired value of the integral would give the same coverage. Additional criteria are needed to determine the intervals uniquely. The most common criterion is to choose *central intervals* such that the area of the excluded tail on either side is $\varepsilon/2$. This criterion is sufficient in most cases, but there is a more general *ordering principle* which reduces to centrality in the usual cases and produces confidence intervals with better properties when in the neighborhood of a physical limit. This ordering principle, which consists of taking the interval which includes the largest values of a likelihood ratio, is described by Feldman and Cousins [11].

29.6.4. *Gaussian errors*:

If the data are such that the distribution of the estimator(s) satisfies the central limit theorem discussed in Sec. 28.3.3, the function $f(x;\alpha)$ is the Gaussian distribution. If there is more than one parameter being estimated, the multivariate Gaussian is used. For the univariate case with known σ,

$$1-\varepsilon = \int_{\mu-\delta}^{\mu+\delta} e^{\dfrac{-(x-\mu)^2}{2\sigma^2}}\, dx = \mathrm{erf}\left(\frac{\delta}{\sqrt{2}\,\sigma}\right) \tag{29.36}$$

Table 29.1: Area of the tails ε outside $\pm\delta$ from the mean of a Gaussian distribution.

ε (%)	δ	ε (%)	δ
31.73	1σ	20	1.28σ
4.55	2σ	10	1.64σ
0.27	3σ	5	1.96σ
6.3×10^{-3}	4σ	1	2.58σ
5.7×10^{-5}	5σ	0.1	3.29σ
2.0×10^{-7}	6σ	0.01	3.89σ

is the probability that the measured value x will fall within $\pm\delta$ of the true value μ. From the symmetry of the Gaussian with respect to x and μ, this is also the probability that the true value will be within $\pm\delta$ of the measured value. Fig. 29.2 shows a $\delta = 1.64\sigma$ confidence interval unshaded. The choice $\delta = \sqrt{\mathrm{Var}(\,\mu\,)} \equiv \sigma$ gives an interval called the *standard error* which has $1 - \varepsilon = 68.27\%$ if σ is known. Confidence coefficients ε for other frequently used choices of δ are given in Table 29.1.

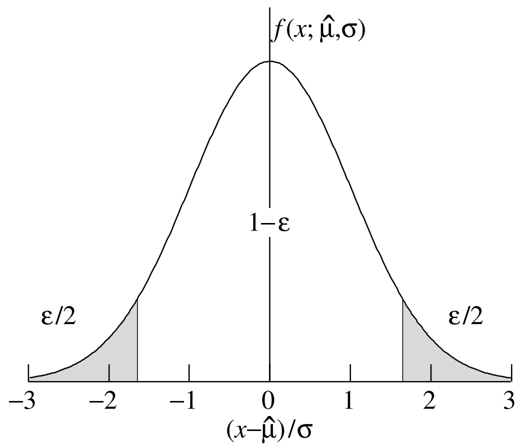

Figure 29.2: Illustration of a symmetric 90% confidence interval (unshaded) for a measurement of a single quantity with Gaussian errors. Integrated probabilities, defined by ε, are as shown.

For other δ, find ε as the ordinate of Fig. 28.1 on the $n = 1$ curve at $\chi^2 = (\delta/\sigma)^2$. We can set a one-sided (upper or lower) limit by excluding above $\mu + \delta$ (or below $\mu - \delta$); ε's for such limits are 1/2 the values in Table 29.1.

For multivariate $\boldsymbol{\alpha}$ the scalar $\mathrm{Var}(\mu)$ becomes a full variance-covariance matrix. Assuming a multivariate Gaussian, Eq. (28.22), and subsequent discussion the standard error ellipse for the pair $(\,\widehat{\alpha}_m, \widehat{\alpha}_n)$ may be drawn as in Fig. 29.3.

The minimum χ^2 or maximum likelihood solution is at $(\,\widehat{\alpha}_m, \widehat{\alpha}_n)$. The standard errors σ_m and σ_n are defined as shown, where the ellipse is at a constant value of $\chi^2 = \chi^2_{\rm min} + 1$

or $\ln \mathscr{L} = \ln \mathscr{L}_{\max} - 1/2$. The angle of the major axis of the ellipse is given by

$$\tan 2\phi = \frac{2\rho_{mn}\, \sigma_m\, \sigma_n}{\sigma_m^2 - \sigma_n^2} \; . \tag{29.37}$$

For non-Gaussian or nonlinear cases, one may construct an analogous contour from the same χ^2 or $\ln \mathscr{L}$ relations. Any other parameters $\widehat{\alpha}_\ell, \ell \neq m, n$ must be allowed freely to find their optimum values for every trial point.

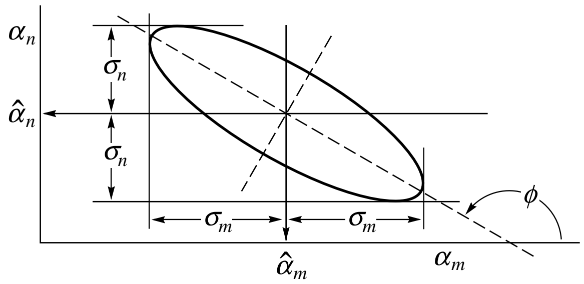

Figure 29.3: Standard error ellipse for the estimators $\widehat{\alpha}_m$ and $\widehat{\alpha}_n$. In this case the correlation is negative.

For any unbiased procedure (*e.g.*, least squares or maximum likelihood) being used to estimate k parameters α_i, $i = 1, \ldots, k$, the probability $1 - \varepsilon$ that the true values of all k lie within the s-standard deviation ellipsoid may be found from Fig. 28.1. Read the ordinate as ε; the correct value of ε occurs on the $n = k$ curve at $\chi^2 = s^2$. For example, for $k = 2$, the probability that the true values of α_1 and α_2 simultaneously lie within the one-standard-deviation error ellipse ($s = 1$), centered on $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$, is 39%. This probability only assumes Gaussian errors, unbiased estimators, and that the model describing the data in terms of the α_i is correct.

29.6.5. *Upper limits and two-sided intervals*:

When a measured value is close to a physical boundary, it is natural to report a one-sided confidence interval (often an upper limit). It is straightforward to force the procedure of Sec. 29.6.3 to produce only an upper limit, by setting $x_2 = \infty$ in Eq. (29.33). Then x_1 is uniquely determined. Clearly this procedure will have the desired coverage, but *only if we always choose to set an upper limit.* In practice one might decide after seeing the data whether to set an upper limit or a two-sided limit. In this case the upper limits calculated by Eq. (29.33) will not give exact coverage, as has been noted in Ref. 11.

In order to correct this problem and assure coverage in all circumstances, it is necessary to adopt a *unified procedure*, that is, a single ordering principle which will provide coverage globally. Then it is the *ordering principle* which decides whether a one-sided or two-sided interval will be reported for any given set of data. The appropriate unified procedure and ordering principle are given in Ref. 11. We reproduce below the main results.

29.6.6. *Gaussian data close to a boundary*:

One of the most controversial statistical questions in physics is how to report a measurement which is close to the edge or even outside of the allowed physical region. This is because there are several admissible possibilities depending on how the result is to be used or interpreted. Normally one or more of the following should be reported:

(a) The actual measurement should be reported, even if it is outside the physical region. As with any other measurement, it is best to report the value of a quantity which is nearly Gaussian distributed if possible. Thus one may choose to report mass squared rather than mass, or $\cos\theta$ rather than θ. For a complex quantity z close to zero, report $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$ rather than amplitude and phase of z. Data carefully reported in this way can be unbiased, objective, easily interpreted and combined (averaged) with other data in a straightforward way, even if they lie partly or wholly outside the physical region. The reported error is a direct measure of the intrinsic accuracy of the result, which cannot always be inferred from the upper limits proposed below.

(b) If the data are to be used to make a decision, for example to determine the dimensions of a new experimental apparatus for an improved measurement, it may be appropriate to report a Bayesian upper limit, which must necessarily contain subjective feelings about the possible values of the parameter, as well as containing information about the physical boundary. Its interpretation requires knowledge of the prior distribution which was necessarily used to obtain it.

(c) If it is desired to report an upper limit in an objective way such that it has a well-defined statistical meaning in terms of a limiting frequency, then report the Frequentist confidence bound(s) as given by the unified Feldman-Cousins approach. This algorithm always gives a non-null interval (that is, the confidence limits are always inside the physical region, even for a measurement well outside the physical region), and still has correct global coverage. These confidence limits for a Gaussian measurement close to a non-physical boundary are summarized in Fig. 29.4. Additional tables are given in Ref. 11.

29.6.7. *Poisson data for small samples*:

When the observable is restricted to integer values (as in the case of Poisson and binomial distributions), it is not generally possible to construct confidence intervals with exact coverage for all values of α. In these cases the integral in Eq. (29.33) becomes a sum of finite contributions and it is no longer possible (in general) to find consecutive terms which add up exactly to the required confidence level $1-\varepsilon$ for all values of α. Thus one constructs intervals which happen to have exact coverage for a few values of α, and unavoidable over-coverage for all other values. This is the best that can be done and still guarantee coverage for any true value.

In addition to the problem posed by the discreteness of the data, we usually have to contend with possible background whose expectation must be evaluated separately and may not be known precisely. For these reasons, the reporting of this kind of data is even more controversial than the Gaussian data near a boundary as discussed above. This is especially true when the number of observed counts is greater than the expected

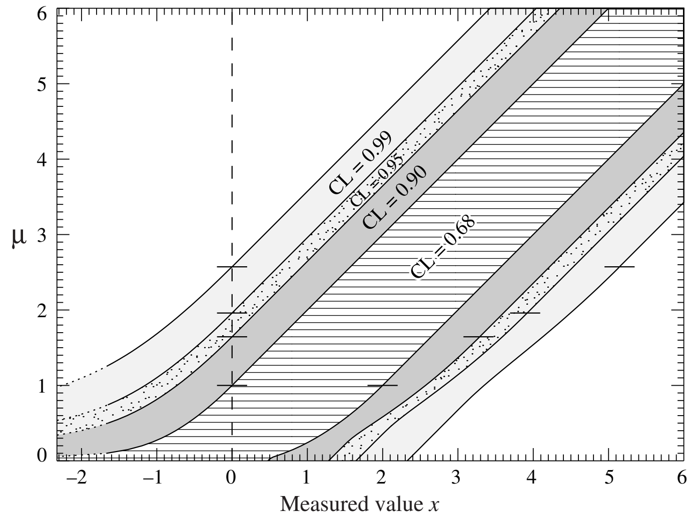

Figure 29.4: Plot of 99%, 95%, 90%, and 68.27% ("one σ") confidence intervals for a physical quantity μ based on a Gaussian measurement x (in units of standard deviations), for the case where the true value of μ cannot be negative. The curves become straight lines above the horizontal tick marks. The probability of obtaining an experimental value at least as negative as the left edge of the graph ($x = -2.33$) is less than 1%. Values of x more negative than -1.64 (dotted segments) are less than 5% probable, no matter what the true value of μ.

background. As for the Gaussian case, there are at least three possibilities for reporting such results depending on how the result is to be used:

(a) The actual measurements should be reported, which means (1) the number of recorded counts, (2) the expected background, possibly with its error, and (3) normalization factor which turns the number of counts into a cross section, decay rate, *etc.* As with Gaussian data, these data can be combined with that of other experiments, to make improved upper limits for example.

(b) A Bayesian upper limit may be reported. This has the advantages and disadvantages of any Bayesian result as discussed above. It is especially difficult to find an acceptable prior probability distribution for this case.

(c) An upper limit (or confidence region) with optimal coverage can be reported using the unified approach of Ref. 11. At the moment these confidence limits have been calculated only for the case of exactly known background expectation. The main results can be read from Fig. 29.5 or from Table 29.2; more extensive tables can be found in Ref. 11.

None of the above gives a single number which quantifies the quality or sensitivity of the experiment. This is a serious shortcoming of most upper limits including those of method (c), since it is impossible to distinguish, from the upper limit alone, between a clean experiment with no background and a lucky experiment with fewer observed counts than expected background. For this reason, we suggest that in addition to (a) and (c)

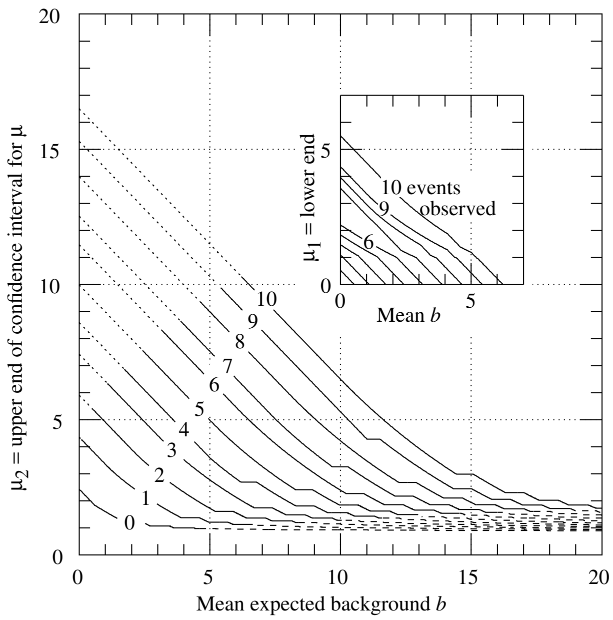

Figure 29.5: 90% confidence intervals $[\mu_1, \mu_2]$on the number of signal events as a function of the expected number of background events b. For example, if the expected background is 8 events and 5 events are observed, then the signal is 2.60 or less with 90% confidence. Dotted portions of the μ_2 curves on the upper left indicate regions where μ_1 is non-zero (as shown by the inset). Dashed portions in the lower right indicate regions where the probability of obtaining the number of events observed or fewer is less than 1%, even if $\mu = 0$. Horizontal curve sections occur because of discrete number statistics. Tables showing these data as well as the CL = 68.27%, 95%, and 99% results are given in Ref. 11.

above, a measure of the sensitivity should be reported whenever expected background is larger or comparable to the number of observed counts. The best such measure we know of is that proposed and tabulated in Ref. 11, defined as the average upper limit that would be attained by an ensemble of experiments with the expected background and no true signal.

References:

1. A. Stuart and A. K. Ord, *Kendall's Advanced Theory of Statistics*, Vol. 2 *Classical Inference and Relationship* 5th Ed., (Oxford Univ. Press, 1991), and earlier editions by Kendall and Stuart.

2. W.T. Eadie, D. Drijard, F.E. James, M. Roos, and B. Sadoulet, *Statistical Methods in Experimental Physics* (North Holland, Amsterdam and London, 1971).

3. H. Cramér, *Mathematical Methods of Statistics,* Princeton Univ. Press, New Jersey (1958).

Table 29.2: Poisson limits $[\mu_1, \mu_2]$ for n_0 observed events in the absence of background.

	CI = 90%		CI = 95%	
n_0	μ_1	μ_2	μ_1	μ_2
0	0.00	2.44	0.00	3.09
1	0.11	4.36	0.05	5.14
2	0.53	5.91	0.36	6.72
3	1.10	7.42	0.82	8.25
4	1.47	8.60	1.37	9.76
5	1.84	9.99	1.84	11.26
6	2.21	11.47	2.21	12.75
7	3.56	12.53	2.58	13.81
8	3.96	13.99	2.94	15.29
9	4.36	15.30	4.36	16.77
10	5.50	16.50	4.75	17.82

4. B.P. Roe, *Probability and Statistics in Experimental Physics,* (Springer-Verlag, New York, 208 pp., 1992).
5. S. Baker and R. Cousins, Nucl. Instrum. Methods **221**, 437 (1984).
6. R.D. Cousins, Am. J. Phys. **63**, 398 (1995).
7. W.H. Press *et al.*, *Numerical Recipes* (Cambridge University Press, New York, 1986).
8. F. James and M. Roos, "MINUIT, Function Minimization and Error Analysis," CERN D506 (Long Writeup). Available from the CERN Program Library Office, CERN-IT Division, CERN, CH-1211, Geneva 21, Switzerland.
9. B. Efron, Am. Stat. **40**, 11 (1986).
10. J. Neyman, Phil. Trans. Royal Soc. London, Series A, **236**, 333 (1937), reprinted in *A Selection of Early Statistical Papers on J. Neyman* (University of California Press, Berkeley, 1967).
11. G.J. Feldman and R.D. Cousins, Phys. Rev. **D57**, 3873 (1998).
12. F. James and M. Roos, Phys. Rev. **D44**, 299 (1991).