Michael E. Peskin
July 12, 2011

Basic Training in Statistics - 2

Building on the concepts developed in the previous lecture, I will explain here how to define the most probable regions for parameters as these are determined by experimental measurements. I will also discuss procedures for placing upper limits on the parameters of theories, for example, on the masses of hypothetical particles. First, though, I will discuss one more set of issues associated with parameter determination.

Goodness-of-fit tests

Many of the methods presented in the previous lecture depended on the assumption that we can accurately write down the correct functional form of the likelihood function. Thus, the question arises: Given a proposal for the form of the likelihood, or any other probability distribution, how do we test this form from data? If the proposed distribution is a Gaussian, this can be tested rather straightforwardly. Even if the distribution is of a very general form, interesting methods are available. There are many such statistical tests adapted to a wide variety of purposes. I will illustrate this variety here by presenting three examples.

First, assume that the likelihood function is described by a Gaussian with known variance (or covariance matrix). In the previous lecture, we considered the situation in which N experimental measurements x_i determine M parameters θ_a, with $N > M$. Then, in general, the quadratic form in the exponent of the likelihood function will have a nonzero minimum value. I argued that this value

$$z = -2\mathcal{L} = \sum_{i,j=1}^{N} (x_i - g_i(\hat{\theta}))\, V_{ij}^{-1}\, (x_j - g_j(\hat{\theta}))$$

varies from realization to realization according to a χ^2 distribution with $(N - M)$ degrees of freedom. We can then test the shape of the probability distribution by applying a χ^2 hypothesis test. The typical value of χ^2 will be 1 for each degree of freedom, or

$$\chi^2 = z \sim (N-M)$$

If $z \gg (N - M)$, we can exclude the hypothesis that the Gaussian with stated covariance matrix V is a good description of the data by using the p-values for the χ^2 distribution quoted in the first lecture.

A weaker assumption is that the distribution of measured values x_i is Gaussian but with an unknown variance. For simplicity, return to the case of one parameter. We can test that the mean of the distribution is given correctly, independently of the value of the variance, using a goodness-of-fit test based on a new variable t.

Let $\{x_i\}$, $i = 1, \ldots, N$, be independent Gaussian random variables with mean 0 and variance 1. Then

$$z = \sum_{i}^{N} x_i^2$$

follows a χ^2 distribution. Let x be an additional independent Gaussian random variable with mean 0 and variance 1. Then one can show that the variable

$$t = \frac{x}{\sqrt{z/N}}$$

distributed according to the *Student's t-distribution* with N degrees of freedom,

$$p(t,n) = \frac{\Gamma(\frac{N+1}{2})}{\sqrt{N\pi}\,\Gamma(\frac{N}{2})} \left(1 + \frac{t^2}{N}\right)^{-(N+1)/2}$$

The proof of this assertion is straightforward and is given in Appendix B. The t-distribution has a mean of zero and variance

$$\langle t^2 \rangle = \frac{N}{N-2}$$

The test of the mean of a Gaussian distribution then proceeds as follows: Consider an experiment that produces N values expected to be distributed according to a Gaussian distribution with mean μ and unknown variance σ^2. Let

$$\bar{x} = \frac{1}{N}\sum_{1}^{N} x_i \qquad s^2 = \frac{1}{N-1}\sum_{i} (x_i - \bar{x})^2$$

be the estimators of mean and variance from the data. The variable $(\overline{x} - \mu)$ follows a Gaussian distribution with mean 0, for the correct value of μ, and variance equal to

$$\sigma^2/N$$

Then

$$X = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/N}}$$

follows a Gaussian distribution with mean 0 and variance 1. Each term in the numerator of s^2 has variance 1 (except for the common value subtracted in the determination of $\overline{x}$), so the quantity

$$z = \frac{(N-1)\, s^2}{\sigma^2}$$

follows a χ^2 distribution with $(N-1)$ degrees of freedom. Then

$$t = \frac{X}{\sqrt{z/N-1}}$$

follows a Student's t-distribution with $(N-1)$ degrees of freedom. We can simplify the calculation of t,

$$t = \frac{(\bar{x}-\mu)/\sqrt{\sigma^2/N}}{\sqrt{(N-1)s^2/\sigma^2(N-1)}}$$

so that

$$t = \frac{\bar{x}-\mu}{\sqrt{s^2/N}}$$

The parameter σ^2 cancels out, and so we can compute p-values for deviations of t from zero without any knowledge of the true variance of the distribution. The price of this is that the t-distribution has a much longer tail than a Gaussian, so the test is not as sharp and one that invokes the correct σ^2.

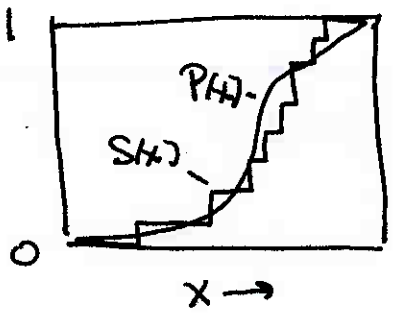
The name associated with the test deserves comment. The author "Student" was William Gosset, a mathematician hired out of Oxford in 1899 by the Guinness brewing company. Guinness regarded their use of statistics as a trade secret, so the company required Gosset to publish under a pseudonym.

There are goodness-of-fit tests for even weaker hypotheses. It is useful, for example, to test whether a set of independently generated values $\{x_i\}$ are consistent with a given pdf $p(x)$ without making any assumptions on the form $p(x)$. A similar problem is that of testing whether two data sets $\{x_{1i}\}$ and $\{x_{2i}\}$ could be draw from the same pdf without any assumption about the form of that pdf. One test that answers these questions is that of Kolmogorov and Smirnov.

Consider first the problem of testing whether $\{x_i\}$, $i = 1, \ldots, N$, could be drawn from $p(x)$. Let $P(x)$ be the lower cumulative distribution of $p(x)$. Let $S(x)$ be the cumulative distribution of the data, that is,

$$S(x) = \frac{k}{N} \qquad k = \#\text{ of points in } \{x_i\} \text{ s.t. } x_i < x$$

Both distributions start at 0 at the smallest possible value of x and approach 1 as x becomes large. Compare these distributions as a function of x.



Let

$$D = \max_x |S(x) - P(x)|$$

Kolmogorov was able to compute the cumulative distribution of D with no further assumptions. The result, for large N, is

$$P(D) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2 z^2}$$

where

$$z \cong \sqrt{N}\, D$$

That is, the p-value for $D_* \gg 1/\sqrt{N}$ is

$$Q(D) \sim 2 e^{-2ND^2} + \cdots$$

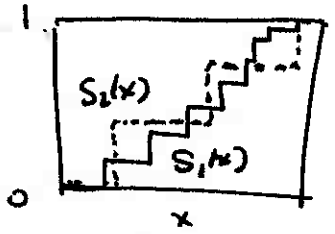
The factor of 2 in the exponent is helpful in defining a quite strong test. Stephens showed that these formula can be used for finite N; with the modification

$$Z = (\sqrt{N} + 0.12 + 0.11/\sqrt{N})\, D$$

the formula for $P(D)$ is accurate down to $N = 4$. For further discussion, and help in computing $P(D)$ and $Q(D)$, see the discussion in *Numerical Recipes*.

Smirnov showed that the same method can be used to test whether two data sets $\{x_{1i}\}$, $i = 1, \ldots, N_1$ and $\{x_{2i}\}$, $i = 1, \ldots, N_2$ are drawn from the same distribution. Smirnov showed that, if $S_1(x)$ and $S_2(x)$ are the (step-function) cumulative distributions of the x_{1i} and x_{2i}, and

$$D = \max_x |S_1(x) - S_2(x)|$$



the cumulative distribution for D is again given by the above formulae, with

$$N^{-1} = N_1^{-1} + N_2^{-1}$$

The proofs of these assertions are relatively straightforward, at least for the case of asymptotic N. I give these proofs in Appendix C.

In the literature, you will many other interesting statistical tests that address these and other questions about a data set and its description by an underlying pdf.

Confidence intervals

Finally, we turn to the following question related to parameter determination: Once we have obtained our estimate of a parameter θ, we should assign an error to this estimate. More precisely, we should define a *confidence interval* $[\theta_-, \theta_+]$ with

$$\theta_- < \hat{\theta} < \theta_+$$

and quantify the probability that the value of θ lies in this interval.

Consider first the case in which the likelihood function is well approximated by a Gaussian. I will also specialize here to the simple situation in which the relation between the measured value x and the parameter θ is linear. Then the likelihood function is

$$p(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(x-a\theta)^2}{\sigma^2}\right]$$

If the value $x = x_0$ is measured, the maximum likelihood estimate of θ is

$$\hat{\theta} = \frac{x_0}{a}$$

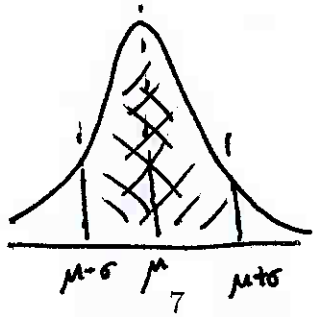
In terms of $\hat{\theta}$, the likelihood function takes the form

$$p(x_0|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\theta-\hat{\theta})^2}{2\Sigma^2}\right] \qquad \Sigma^2 = \frac{\sigma^2}{a^2}$$

For this case, I will formally define the *$n\sigma$ confidence interval* for θ to be the interval $[\theta - n\Sigma, \theta + n\Sigma]$. In particle physics, we typically quote 1σ intervals as the standard error, writing

$$\theta = \hat{\theta} \pm \Sigma$$

For a Gaussian distribution, we recall that the interval $[\mu - \sigma, \mu + \sigma]$ contains 68.3% of the area under the Gaussian.



Thus, the interval I have just defined is also called a *68% confidence interval.* In some fields, it is conventional to quote 2σ errors, corresponding to a 95% confidence interval.

In a moment, I will discuss the interpretation of the confidence level in this construction as a probability. First, though, I will give the generalization to higher-dimensional Gaussian likelihood functions.

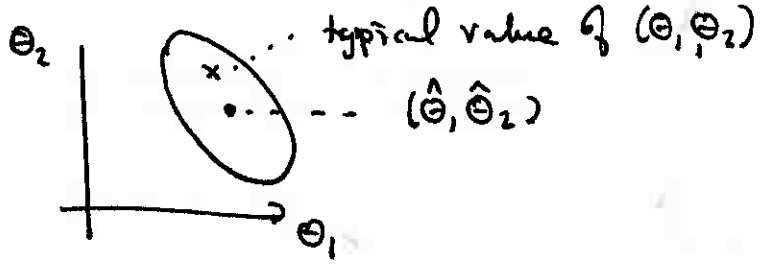
At the end of the previous lecture, I wrote the likelihood function for parameters θ_a, $a = 1, \ldots, M$, as

$$\text{likelihood} \sim \exp\left[-\tfrac{1}{2} \sum_{ab} (\theta_a - \hat{\theta}_a)\, U^{-1}_{ab} (\theta_b - \hat{\theta}_b)\right]$$

For such a form, we can define the standard error contour by the condition

$$z = \sum_{ab} (\theta_a - \hat{\theta}_a)\, U^{-1}_{ab} (\theta_b - \hat{\theta}_b) = 1$$

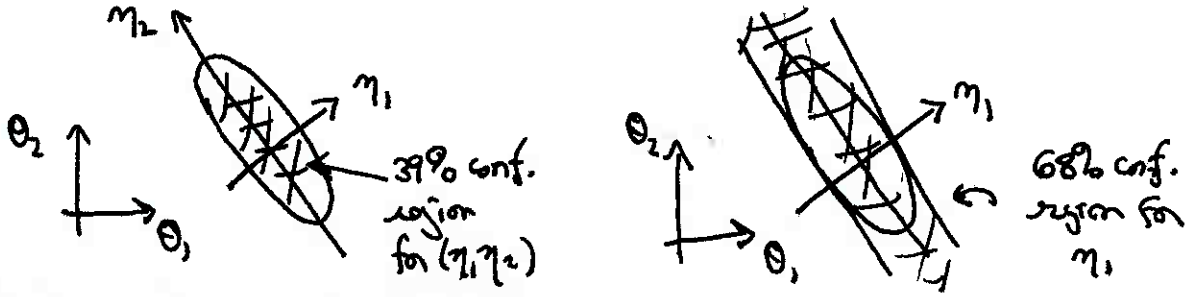
This directly generalizes the definition above. For $M = 2$, this contour is called the *error ellipse.* For $M > 2$, the contour is an ellipsoid in the higher-dimensional space. A typical outcome for $M = 2$ would be an ellipse of the form



It is important to note that this contour contains a probability that decreases steeply with the number of variables M:

M	area enclosed
1	68.3%
2	39.3%
3	19.9%
4	9.0%

Often, one of the determined variables is more important than the others, and we would like to quote a separate overally error for that parameter. The simplest way to work out that error is to integrate the Gaussian form of the likelihood function over the other variables. A rule of thumb follows from the fact that the Gaussian separates in the coordinate system of principal axes of the ellipsoid. Then the 1-dimensional 68% confidence interval for a variable in this frame is given by the volume between the two planes tangent to the error ellipsoid perpendicular to the corresponding axis.



More often, the variable we are interested in is not a principle axis; then we must go back to the full form of the likelihood function. In general, the standard error on the variable θ_k with all other variables integrated over is

$$\Sigma^2_{(k)} = U_{kk} \qquad (\text{not } [(U^{-1})_{kk}]^{-1}!)$$

Now it is time to discuss the probability interpretation of the confidence interval. Bayesians and frequentists approach this question differently. For Bayesians, the interpretation is simple. We can speak of a probability that the parameter θ has a certain value. The probability distribution of θ is informed by the measurement of x. In the Gaussian case that we are discussing, the posterior probability distribution for θ is given by

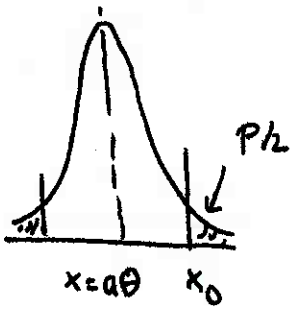
$$p(\theta) = \text{const.} \cdot e^{-\frac{(\theta-\hat{\theta})^2}{2\Sigma^2}} \cdot p_0(\theta)$$

where $p_0(\theta)$ is the prior probability. If $p_0(\theta)$ is very broad compared to the Gaussian factor in this equation, then the posterior probability of θ will be given accurately by the properly normalized Gaussian

$$P(\theta) = \frac{1}{\sqrt{2\pi\Sigma^2}} e^{-\frac{(\theta-\hat{\theta})^2}{2\Sigma^2}}$$

Then the 68% confidence interval that I have defined above is exactly the symmetrical region about $\hat{\theta}$ that contains 68% of the total area. Bayesians would then be comfortable in saying that θ lies in the interval $[\hat{\theta} - \Sigma, \hat{\theta} + \Sigma]$ with probability 68%.

Frequentists do not accept the idea that there is a well-defined probability distribution for θ. Instead, they would argue as follows: If $\theta = \hat{\theta}$, the value $x = x_0$ obtained in the experiment would be a likely outcome of the measurement. On the other hand, for values of θ on either side of $\hat{\theta}$, the measurement of x_0 would be less likely. Moving to higher values of θ, we eventually reach a point at which the probability of measuring x_0 or a point less consistent with that value of θ is a small probability p. We can define this value of θ to be a boundary of the probability $(1-p)$ confidence interval.

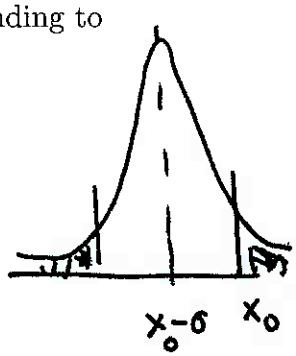


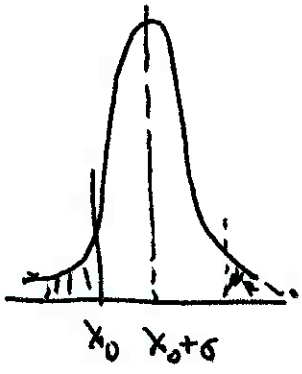
To put it another way, the 68% confidence interval in θ is the set of values of θ for which, given that value of θ, the measured value x_0 lies within a region that encloses 68% of the probability for the values of x.

For this simple case, the boundary values of θ are

$$\hat{\theta} - \frac{\sigma}{a} \qquad \hat{\theta} + \frac{\sigma}{a}$$

corresponding to





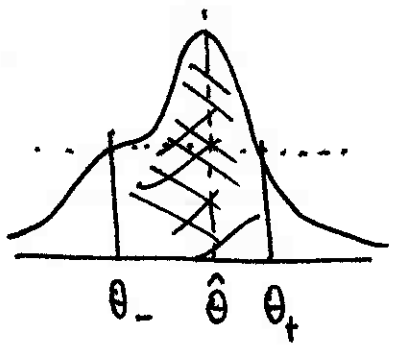
Then we obtain the same standard error interval

$$\theta \in [\hat{\theta} - \Sigma, \hat{\theta} + \Sigma] \qquad \Sigma = \frac{\sigma}{a}$$

as we found from the Bayesian argument.

If the likelihood function is not a Gaussian, we can still apply the Bayesian or frequentist arguments given above. We will obtain similar, but somewhat different, confidence intervals.

The Bayesian analysis is quite straightforward. We construct the likelihood function $p(x|\theta)$, evaluate this at x_0, and normalize it as a function of θ. We then find the values θ_-, θ_+—at equal values of the likelihood—such that the interval $[\theta_-, \theta_+]$ contains 68% of the area under the likelihood function



Writing

$$\theta_- = \hat{\theta} - \Sigma_- \qquad \theta_+ = \hat{\theta} + \Sigma_+$$

we would quote the error interval for θ as

$$\theta = \hat{\theta}^{+\Sigma_-}_{-\Sigma_-}$$

Note that the two ends of the error interval will typically be asymmetrically separated from $\hat{\theta}$.

Another convention that is often used is to determine $\Sigma_\pm$ as the two solutions to the equation

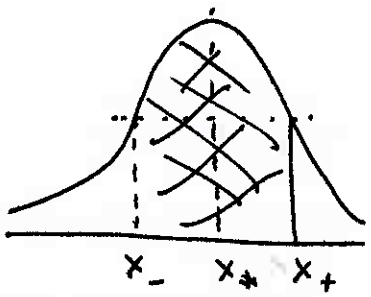
$$-2\log(\text{likelihood}) = -2\mathcal{L}(\theta) = 1$$

This gives an approximate 68% confidence interval for likelihood functions that are approximately Gaussian

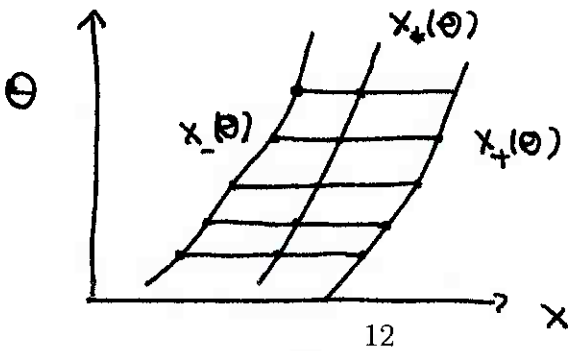
To define a confidence interval in frequentist terms, we need a more elaborate generalization. Here is the construction, due to Jerzy Neyman: Begin, again, from

$$p(x|\theta)$$

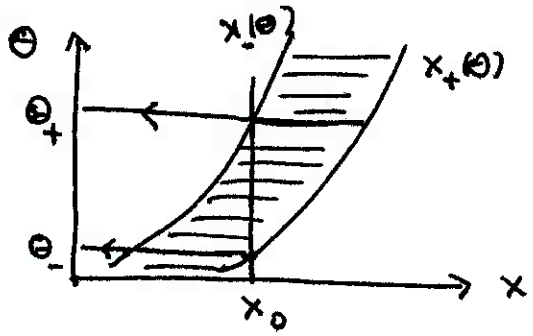
We first study this function for each fixed θ as a function of x, so that it has meaning as a probability. For each θ, there is a most probable value of x, which I will call $x_*(\theta)$. About this point, we can construct an interval $[x_-(\theta), x_+(\theta)]$ such that the likelihood function takes equal values at x_+ and x_- and the interval contains, for the standard error, 68% of the total probability.



(More generally, we could consider intervals that contain a fraction $(1-p)$ of the total probability.) Plot these intervals on the plane of x versus θ:



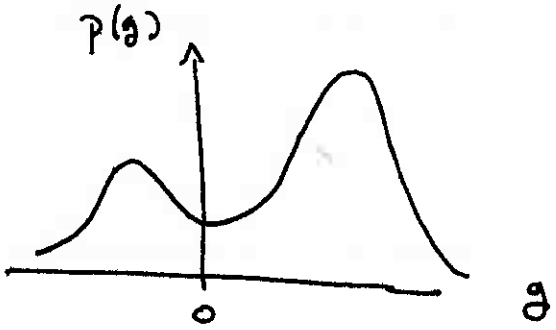
The two-dimension region covered is called the *confidence belt.* Now apply the information that the measured value is x_0. The line $x = x_0$ intersects the confidence belt in an interval $[\theta_-, \theta_+]$:



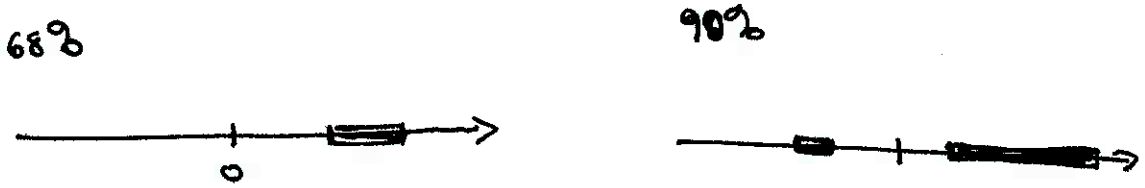
This interval $[\theta_-, \theta_+] = [\hat{\theta} - \Sigma_-, \hat{\theta} + \Sigma_+]$ is the *Neyman 68% confidence interval.* This is the set of values of θ for which the measured value x_0 lies within an interval about the most likely value of x that contains 68% of the total probability.

In the case of a simple Gaussian likelihood function, the Neyman construction reduces to the construction that we made earlier and will agree with the error interval obtained from Bayesian arguments. For more general distributions, the Bayesian and Neyman constructions will disagree. Both methods are used in the literature, so it is important, in using a result from a paper, to understand carefully which (or what other) method was used.

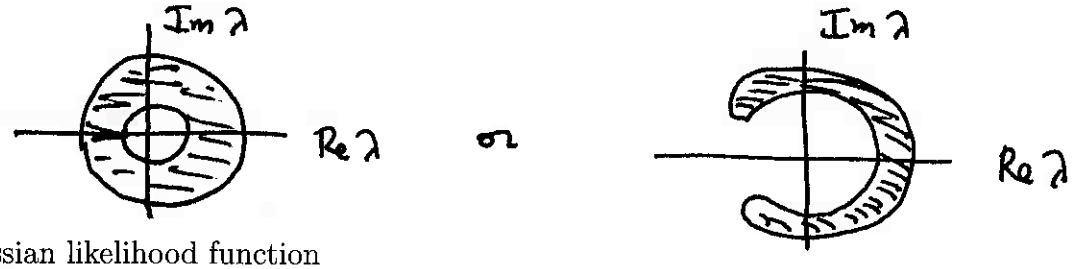
It is important to note that confidence intervals need not be connected, or, in higher dimensions, simply connected. An example that often arises is that of a new interaction parametrized by a coupling λ. Depending on the sign of λ, this interaction might interfere constructive or destructively with the Standard Model, and either case might allow a fit to experimental data. Then the (Bayesian) probability for g would have the form



giving confidence intervals



depending on the significance level chosen. Another possibility is that λ might be complex, with the signal strength depending weakly on its phase. Then the two-dimensional confidence ellipse for (Re λ, Im Λ) might have the form



A Gaussian likelihood function

$$p(x|\lambda) \sim \exp\left[-\frac{1}{2\sigma^2}(x - g(\lambda))^2 \right]$$

in which the function $g(\lambda)$ is nonlinear can easily give rise to these and other odd behaviors.

Confidence intervals near a boundary

The methods for determining confidence intervals that I have just described run into difficulties for situations in which the signals that we are trying to extract from data are marginal. You might think that this is a special case, but, actually, it applies very often. Many experiments concern signals that have not yet been discovered but will—we hope—appear in the current data set. To discuss this situation, we must address the definition of confidence intervals when the space of parameters θ has natural boundaries.

Consider, for example, the case in which θ is required to be positive on physical grounds. For example, if we are trying to measure the mass of a light particle from kinematics, the particle mass will enter the kinematic expressions as m^2. For example, experiments that attempt to measure the neutrino mass from the endpoint energy E_e of a β-decay electron analyze a likelihood function

$$p(E_e|m^2)$$

If there is a nonzero energy resolution for E_e, then the form of this function will be a Gaussian about the predicted value $E_e(m^2)$. If the true value of m^2 is zero, we will find values of E_e below $E_e(0)$, corresponding to a *negative* inferred value of m^2, half

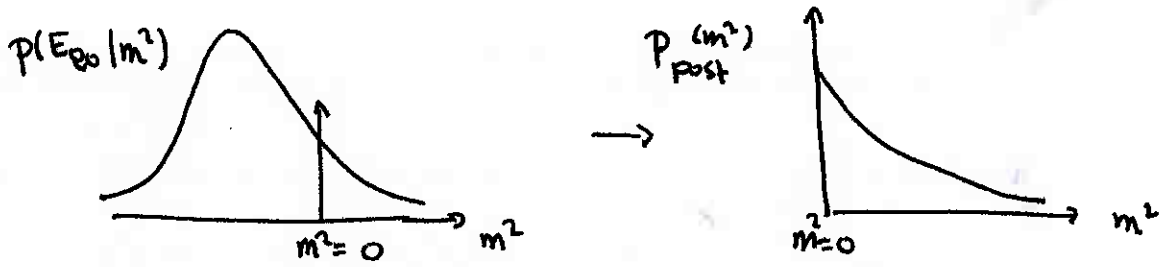
of the time. And, 15% of the time, we will find values of E_e that are below $E_e(0)$ by more than 1σ, so that the entire standard confidence interval for m^2 lies at negative values. How do we quote the result of the experiment in this case? If the preferred value of m^2 is negative, that is clearly evidence that the mass of the neutrino is small. But, what is the precise upper limit on m^2 that is implied?

Frequentists traditionally dealt with this problem by constructing the Neyman 68% confidence interval and then arbitrarily translating the central value to $m^2 = 0$, keeping the size of the interval fixed. Alternatively, they constructed the Neyman interval at a higher level of confidence so that the region would include some positive values. I will discuss a better frequentist solution to this problem below.

Some textbooks say that, to solve this problem, one must become a Bayesian. Indeed, Bayesians have a simple solution at hand. In the Bayesian philosophy, a measurement improves our prior knowledge of the parameter being determined. We can claim that, *a priori*, we are ignorant of the value of m^2, but we do know from basic principles that it is non-negative. So, we can reasonably take the prior probability $p_0(m^2)$ to be constant for $m^2 \geq 0$ but *zero* for $m^2 < 0$. Then the resulting posterior probability distribution for m^2 would be

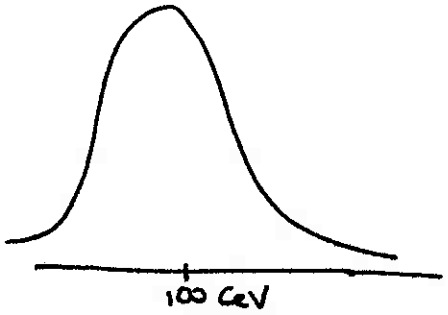
$$p(m^2 | E_{e0}) = \frac{p(E_{e0} | m^2)\ \Theta(m^2)}{\int_0^\infty dm^2\ p(E_{e0} | m^2)}$$

Graphically, we slice the likelihood function at $m^2 = 0$ and renormalize the tail.

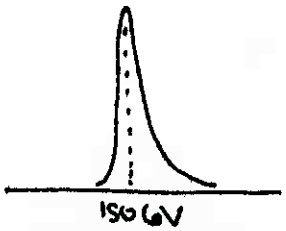


Then we can use this normalized distributions to set upper limits. For example, if m_*^2 is such that values $m^2 > m_*^2$ contain only 5% of the area under the renormalized $p(m^2|E_{e0})$, we would say that $m^2 < m_*^2$ at 95% confidence.

A difficulty with this technique is that is works too well, creating a situation that we might call the "Strumia problem". Imagine that you have published a theory of supersymmetric particle masses that predict low values, say, about 100 GeV. The likelihood function predicted by your theory might have the form



Now an experiment excludes masses below 150 GeV. It is tempting to keep the same theory but include this new information in the likelihood function. This gives a prediction for the next experiment. Typically, the prediction will be sharper, because the tails of the original likelihood function fall off rapidly.



Proceeding in this way, we obtain a set of predictions that the supersymmetry particle masses lie just above the current limits, with smaller and smaller error intervals.

The real problem here is that, at a certain point, one should decide that the theory is wrong and needs to be improved. This evolution actually took place in experiments that measured the electron neutrino mass from the β-decay endpoint measured with films of solid Tritium. The best experiments of the 1990's, at Livermore, Mainz, and Troitsk, measured negative values of m^2 at increasing levels of confidence. Many arguments were given about the correct statistical treatment to define upper limits on m^2. Finally, the Mainz group realized that they were working above the temperature of the roughening transition of solid Tritium surfaces. That is, they were correcting the electron spectrum for energy loss in the solid Tritium assuming a uniform flat surface, while in reality the surface thickness was fluctuating substantially from point to point. By running the experiment at a lower temperature of 2°K, they removed this effect and found estimates of m^2 scattered more reasonably about $m^2 = 0$.

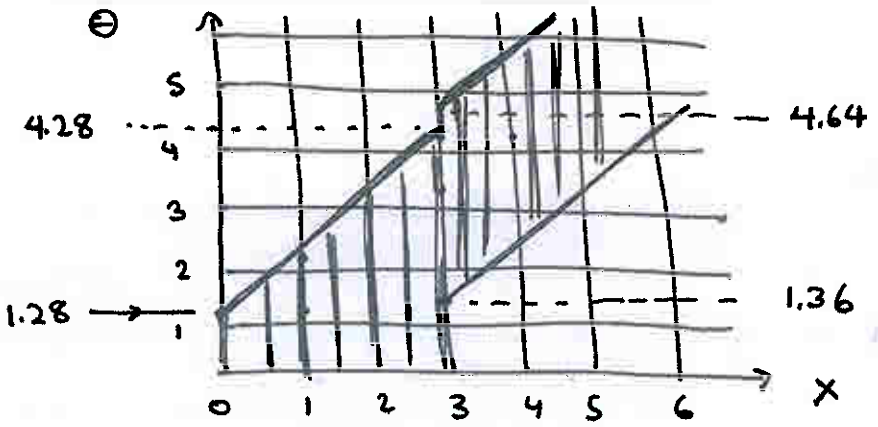
Let's now return to the frequentist approach to confidence intervals and limits for a parameter θ with a natural boundary at $\theta = 0$. When the preferred value $\hat{\theta}$ is greater than 0, it would seem reasonable to quote a 95% confidence limit on θ at the value $\hat{\theta} + 1.64\sigma$. However, another problem can appear with this approach, as pointed out by Gary Feldman and Robert Cousins in a paper that is worth a careful reading (Phys. Rev. D57, 3873 (1998)). Feldman and Cousins analyze experiments that search for small or marginal signals. It is reasonable, they say, to decide before the data is taken whether to quote an upper limit on θ or a measurement of θ with a confidence interval. However, they point out, making this choice on the basis of the data and then following standard practices can lead to errors.

Imagine an experiment that is searching for an effect parametrized by θ. It is very tempting to follow a practice that Feldman and Cousins call "flip-flopping": If the experiment finds a value $\hat{\theta}$ that is less than 3σ from zero, quote an upper limit as described above, but if the value $\hat{\theta}$ is greater than 3σ, announce observation of the effect and quote a nonzero measurement. Unfortunately, the choice of strategy is can be influenced by fluctuations of the measurement. It is instructive to analyze this carefully for the case of the simple Gaussian likelihood function

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$$

with maximum at $x = \theta$ and $\sigma = 1$. I will consider 90% confidence upper limits and 90% confidence intervals.

First consider the case in which the measurement obtained is less then 3σ. Then we quote an upper limit. For $x_0 = 0$, the 90% confidence upper limit lies at $\theta_* = 1.28\sigma$; for higher values of x_0, $\theta_* = x_0 + 1.28$. This situation applies up to $x_0 = 3$. Above that point, we quote a 90% confidence interval $\theta = x_0 + 1.64\sigma$. The resulting confidence belt, the union of these confidence intervals plotted vertically in θ as a function of x_0, is



In a correct frequentist interpretation, the horizontal slices through this figure should be intervals of x that contain 90% of the total probability. However, as you can see, in the region $1.36 < \theta < 4.28$, these regions are too narrow and in fact contain only 85% of the total probability. Then the results quoted for 90% confidence are stronger than the experimental measurement actually allows.

Feldman and Cousins gave a solution to this problem, and I will discuss it below. First, though, there is one more basic introductory topic that we should discuss.

Poisson distribution

The Poisson distribution describes the distribution of the number of events in a sample for a process that generates individual events indepedently with some probability. So far in these lectures, all of the pdf's that we have studied have been functions of continuous variables. The Poisson distribution is a function of a discrete variable, the number of generated events. Nevertheless, the same basic concepts apply to the analysis of this distribution.

The pdf of the Poisson distribution is

$$p(k|\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

where k is the generated number of events. The mean and variance of the distribution are

$$\langle k \rangle = \lambda \qquad \mathrm{Var}(k) = \lambda$$

For an experiment in which we expect n events (n need not be an integer), $\lambda = n$ and the typical fluctuation in k about this value is $\sqrt{n}$. As $\lambda \to \infty$, $p(k|\lambda)$ converges to a Gaussian distribution with $\mu = \lambda$ and $\sigma^2 = \lambda$; this is a consequence of the Central Limit Theorem.

Because k is discrete, we cannot define exact confidence intervals for λ at any chosen level from a measurement of k. Instead, we define *conservative* confidence intervals, that is intervals that are at least as large as required. A measurment k_0 defines a conservative upper limit according to

$$\lambda < \lambda_+ \text{ with confidence } (1-p) \text{ if } Q(k_0, \lambda_+) \leq p$$

and a conservative lower limit according to

$\lambda > \lambda_-$ with confidence $(1-p)$ if $P(k_0, \lambda_-) \leq p$

We could then define $[\lambda_-, \lambda_+]$ to be a standard error interval (68% confidence) if the regions $\lambda > \lambda_+$ and $\lambda < \lambda_-$ are each excluded at a confidence level at most 31%.

An important special case is that in which *zero* events are observed. Since

$$P(k=0, \lambda) = e^{-\lambda}$$

the value $\lambda = 3$ defines the 95% confidence limit. That is, *the observation of 0 events excludes models in which 3 or more events are expected at the 95% confidence level.*

Here is a table of the upper and lower 95% confidence limits (defining a 90% confidence interval) for higher numbers of observed events:

observed	lower	upper
0	—	3
1	0.05	4.7
2	0.36	6.3
3	0.82	7.75

If the process we are studying has known backgrounds, the values quoted are for the expected number of signal plus background processes. If the signal rate is small, it will often happen that the observed number of events will be less than that expected number of background events. In that case, there are difficulties in quoting upper limits on the rate of signal events similar to the problems discussed in the previous section.

In this context, I will now explain the proposal for "unified" confidence intervals suggested by Feldman and Cousins. In the frequentist construction of the confidence belt, we begin by constructing, for each value of θ, an interval in x that contains a fraction $(1-p)$ of the total probability (or, for a discrete distribution, a set of points k that contains at least a fraction $(1-p)$ of the total probability). The boundaries of

this interval are to some extent arbitrary. We may choose different conventions that shift the interval to the left or the right, as long as it contains the preferred value $x_*(\theta)$ that maximizes the likelihood. Feldman and Cousins proposed that we include those points with the highest values of the quantity

$$R = \frac{p(x_0|\theta)}{p(x_0|\theta_{best})}$$

where θ_{best} is the value of θ in the physically allowed region that maximizes the likelihood function $p(x|\theta)$ for the value of x being considered. If the maximum of the function $p(x|\theta)$ lies in a unphysical region, the point θ_{best} would typically lie at the boundary. For a process governed by a Poisson distribution, the physical region for the expected number of events λ would exclude values that are less than the expectation from known background processes.

The prescription gives results that are shifted somewhat from those described above. Applying the prescription to a Poisson process with zero background, it gives the 90% confidence intervals

observed	lower	upper
0	0.0	2.4
1	0.11	4.4
2	0.53	5.9
3	1.1	7.4

For a Gaussian distribution with mean at zero, it gives the same upper limits as the Bayesian argument given earlier,

90% : 1.64σ 95% : 1.96σ

This frequentist method for defining confidence intervals near a boundary is now often used in particle physics papers dealing with marginal signals.

This concludes my introduction to statistics and their use in elementary particle physics. I hope that these lectures will prepare you to understand the language that is used in the presentation of experimental results in this field.