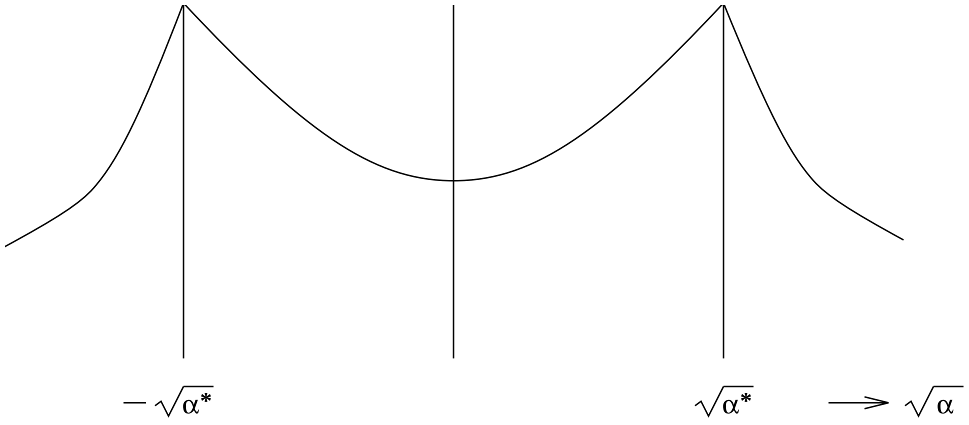
General Maximum Likelihood Methods—I

- The above is glib in many ways—especially in its treatment of errors.
  - We will repair this later with real live theorems.
  - First, we state the obvious generalization.
- The general maximum likelihood method is to put $P(\alpha|T) = 1$ in Equation (14)
- and form $L(\alpha) = \Pi_{i=1}^{n} p\left(x_i, \alpha\right)$
- for any measurement x_i whose probability $p\left(x_i, \alpha\right)$ depends on value x_i and theoretical parameter(s) α.
- The maximum likelihood estimate of α is the value α^* where $L(\alpha^*)$ is a maximum.
- The error in estimate is the standard deviation of the distribution of L as a function of α, i.e.,

$$L(\alpha) = \exp\left\{\text{const.} - \frac{1}{2\Delta\alpha^2}\left(\alpha - \alpha^*\right)^2\right\}$$

General Maximum Likelihood Methods—II

- One must issue some obvious warnings.
- For instance, L could have several maxima



- For instance, this would happen if in earthquake problem you were still drunk in morning and decided to find $\sqrt{\alpha}$.
  - Then $\sqrt{\alpha}$ would peak at $\pm\sqrt{\alpha^*}$.
- Less obviously one should remember that theorems (to come) only say that maximum likelihood method works for large n.
- Then it is a universally good method.
- However, for small n it is much more likely to give nonsense than many other methods. Thus, ...

Warning Example of Misuse of Likelihood Method

- Consider probability distribution appropriate for polarization experiment:

$$p(\varphi, \alpha) = \frac{(1 + \alpha \cos \varphi)}{2\pi} \quad : -1 \leq \alpha \leq 1.$$

- If we had but one event (a very small experiment!)

$$L(\alpha) = \frac{(1 + \alpha \cos \varphi_1)}{2\pi}$$

  - if $\cos \varphi_1 > 0$: $\alpha = 1$ (i.e., end of range) maximizes L
  - or $\cos \varphi_1 < 0$: $\alpha = -1$ (other end of range) maximizes L

- If you have two events:

$$L(\alpha) = \frac{(1 + \alpha(\cos \varphi_1 + \cos \varphi_2) + \alpha^2 \cos \varphi_1 \cos \varphi_2)}{2\pi}$$

- This has a maximum at end of range, again, unless $\cos \varphi_1 \cos \varphi_2 < 0$.

Practical Advice on Use of Maximum Likelihood—I

- So usually people warn you by saying:
- "If distribution of α **not** Gaussian, then do not simply quote maximum and standard deviation, rather plot the likelihood."
- Unfortunately, this is dubious, because for **small** n, nobody has proved any theorems and $P(\alpha|T)$ philosophy important
- while for large n, theorems are proven on the relevance of the likelihood method
  - but they also prove L is Gaussian!

Practical Advice on Use of Maximum Likelihood—II

- So correct warning is: "If distribution not Gaussian—everything is OK if either:
  - (α) Just many sharp humps as in $\sqrt{\alpha}$ example, data can be fitted with > 1 value of parameter.
  - (β) If several parameters, can be Gaussian in some and flat in others if depends weakly (or not at all!) on them.
  - (γ) Gaussians maybe distorted by effects due to end of range.
- If (α), (β), or (γ) don't apply and distribution of α **isn't** Gaussian, then one can only be unhappy!
- You should apply for a new grant and do a better experiment!

Determination of Lifetime Example of Maximum Likelihood—I

- All books give the example of determining the lifetime τ_0 of a particle from the observation of decay times $t_1 \ldots t_n$ of n decays.
- The probability distribution for decay problem is Poisson:
$$p(t)dt = \frac{1}{\tau} \exp\left(-\frac{t}{\tau}\right) dt$$
- Now, dt is meaningless—it is independent of τ and only affects normalization of likelihood.
  - The unknown $P(\alpha|T)$, $P(x_i|T)$ terms in Equation (14) made this arbitrary anyway.
- We must, however, keep terms in normalization $\left(\frac{1}{\tau}\text{ here}\right)$ that depend on parameters to be determined.
- The likelihood of n decays is:
$$\begin{aligned} L(\tau) \; &\propto \; \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \; \ldots \; \frac{1}{\tau} e^{-\frac{t_n}{\tau}}, \quad \text{or} \\ L(\tau) \; &= \; \exp\left[-\frac{1}{\tau} \sum_{i=1}^{n} t_i - n \ln \tau\right] \end{aligned}$$
where we lazily convert $\propto$ into an $=$ sign.

Determination of Lifetime Example of Maximum Likelihood—II

- The maximum of L is given by:

$$\begin{aligned}\frac{dL}{d\tau} &= 0 \quad \text{at } \tau = \tau^* \quad \text{or} \\ \frac{1}{{\tau^*}^2}\sum_{i=1}^{n} t_i - \frac{n}{\tau^*} &= 0 \quad \text{or} \\ \tau^* &= \frac{1}{n}\sum_{i=1}^{n} t_i\end{aligned}$$

- or the maximum likelihood estimate is just the mean:
- such a simple result is only true because of particular Poisson distribution used in the example.

Central Limit Estimate of Error

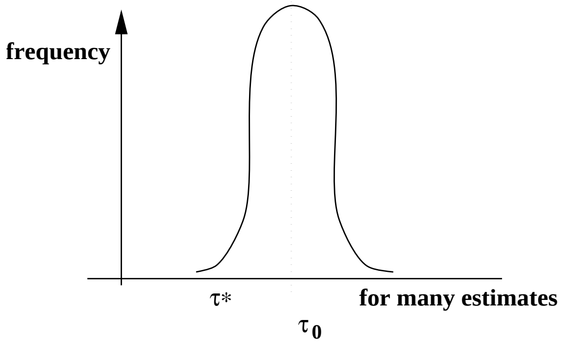
- We can use this example to illustrate an interesting theoretical point about the error, and now compare two methods of finding an error.
- (a) According to the central limit theorem, the variable $y = \frac{1}{n}\sum_{i=1}^{n} t_i$ is a sum over independent random variables and has mean $\langle y \rangle = \tau_0$ and standard deviation σ.

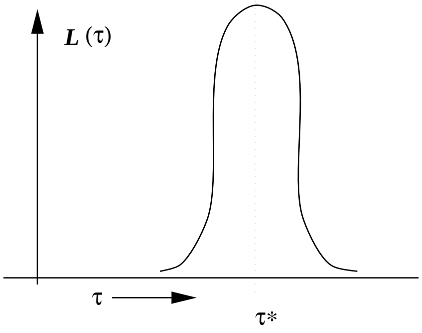
$$\begin{aligned} \sigma^2 &= \frac{1}{n}\left\{\langle t_i^2 \rangle - \langle t_i \rangle^2\right\} \\ &= \frac{1}{n}\left\{\frac{1}{\tau_0}\int t^2 e^{-\frac{t}{\tau_0}}\, dt - \tau_0^2\right\} \quad \text{or} \\ \sigma &= \frac{\tau_0}{\sqrt{n}} \end{aligned} \tag{15}$$

- The mean and standard deviation of y refer to distribution of values of y in **many** experiments—**each** of n observations.

Maximum Likelihood Estimate of Error in Lifetime—I

- (b) However, maximum likelihood method of finding error is, off hand, quite different.
- Rather than looking at distribution of y over infinitely many experiments each of n observations,
- It considers τ dependence of L for one given experiment of n observations.





Maximum Likelihood Estimate of Error in Lifetime—II

- Then we found L became Gaussian in τ, and calling maximum likelihood standard deviation $\tilde{\sigma}$

$$L(\tau) \propto \exp\left\{-\frac{1}{2\tilde{\sigma}^2}(\tau - \tau^*)^2\right\}$$

- where we can find width $\tilde{\sigma}$—width of Gaussian by:

$$\begin{aligned} \frac{1}{\tilde{\sigma}^2} &= -\frac{d}{d\tau^2} \ln L(\tau)\big|_{\tau=\tau^*} \\ &= \frac{2}{\tau^3}\sum_{i=1}^{n} t_i - \frac{n}{\tau^2}\big|_{\tau=\tau^*} \quad \text{or} \\ \tilde{\sigma} &= \frac{\tau^*}{\sqrt{n}} \end{aligned} \tag{16}$$

- as $\tau^* = \tau_0 + O\left(\frac{1}{n}\right)$ we see that Equation (15) and (16) are in fact the same asymptotically (large n, that is).

Comparison of Two Error Estimates in Maximum Likelihood—I

- Note that physics books (Solmitz is an exception) don't mention method (a) of finding error in maximum likelihood method.
- On the other hand, mathematics books never mention (b) as a serious method.
- They do, however, prove (a) and (b) are the same for large n (as we indicated in *example* above) for general distribution.
- As far as I can see, method (a) is only correct definition of error.
  - It follows that method (b) is wrong when it differs from method (a).
  - (b) does differ from (a) for small n when, in particular, likelihood isn't Gaussian.
  - So again warnings for small n ("plot likelihood") are very dubious.

Comparison of Two Error Estimates in Maximum Likelihood—II

- The correct statement is that your results are simply unreliable for small n, and error is not calculable from likelihood function alone.
- The error is defined for small n, but must be calculated as in (a)—this is general as not possible to do with just the single experiment with only a few (n) observations.
- Note that when math books discuss maximum likelihood estimate for small n—it is only the value α^* they discuss and not the shape of L and the method (b) of finding error.

Theorem on Asymptotic Validity of Maximum Likelihood Method—Conditions and Statement of Theorem

- This is described in Cramer Chapters 22 and 23
- **Theorem:**
  Let $P(x_i, \alpha)$ be a normalized probability for a result x_i given theoretical parameter α. We have n such observations x_i:
  - $L = \Pi_{i=1}^{n} P(x_i, \alpha)$ is usual likelihood function.
  - Let α^* be maximum of L with respect to α.
  - There are three basic results that can be proven.

Theorem on Asymptotic Validity of Maximum Likelihood Method Conditions—First Assertion

- (a) If we fix n and conduct many such series of n observations, the mean of α^* satisfies:

$$\langle \alpha^* \rangle = \alpha_0 + O\left(\frac{1}{n}\right)$$

- Note $\langle \alpha^* \rangle$ is not $= \alpha_0$ exactly (α_0 is true theoretical parameter).
- If for *any* estimate $\langle \alpha_{\text{estimate}} \rangle = \alpha_0$, then we say estimate is unbiased:
- if $\langle \alpha_{\text{estimate}} \rangle = \alpha_0 + b(\alpha_0)$ then we have a biased estimate with bias $b(\alpha_0)$.
- So maximum likelihood estimate is biased for finite n and is only asymptotically unbiased.
- However, bias is smaller than Standard Deviation and so it is of no real importance.
- Bias is discussed in detail by Cramer.

Theorem on Asymptotic Validity of Maximum Likelihood Method Conditions—Second and Third Assertions

- (b) We can regard α^* as a random variable: then its standard deviation is $O\left(\frac{1}{\sqrt{n}}\right)$ and to this order we can correctly calculate standard deviation in the "physics way" already discussed (i.e., from α dependence of $L(\alpha)$).
- (c) There is *no* other estimate of α_0 which has a smaller standard deviation asymptotically in n, i.e., the maximum likelihood method is the best as long as n is large.

Proof of Maximum Likelihood Theorem Part (a)—I

- Can be found in Cramer, p. 479 or Hoel p. 379.
- For (a) and (b), write

$$\ln L(\alpha) = \sum_{i=1}^{n} \ln P(x_i, \alpha)$$

- Expand L about true theoretical value.

$$\ln L(\alpha) = L_0 + (\alpha - \alpha_0)L_1 + \frac{(\alpha-\alpha_0)^2}{2}L_2 + \frac{(\alpha-\alpha_0)^3}{6}L_3 + \ldots$$

$$\begin{aligned} L_0 &= \sum_{i=1}^{n} \ln P(x_i, \alpha_0) \\ L_1 &= \sum_{i=1}^{n} \frac{\partial}{\partial \alpha} \ln P(x_i, \alpha_0) \\ L_2 &= \sum_{i=1}^{n} \frac{\partial^2}{\partial \alpha^2} \ln P(x_i, \alpha_0), \quad \text{etc.} \end{aligned}$$

- We get maximum likelihood estimate, by calculating $\partial \ln L / \partial \alpha^* = 0$ or

$$\begin{aligned} L_1 \;+\; (\alpha^* - \alpha_0)L_2 = 0 \quad \text{i.e.,} \\ \alpha^* = \alpha_0 - \frac{L_1}{L_2} \end{aligned} \qquad (17)$$

Proof of Maximum Likelihood Theorem Part (a)—II

- The error in this is $O\left\{(\alpha^* - \alpha_0)^2 \frac{L_3}{L_2}\right\}$ which is $O\left(\frac{1}{n}\right)$ as we will soon show that $(\alpha^* - \alpha_0)^2$ is $O\left(\frac{1}{n}\right)$.
- Now L_1 and L_2 are random variables and Equation (17) defines α^* as a random variable.
- We can calculate from (17):

$$\begin{aligned} \langle \alpha^* \rangle &= \alpha_0 - \langle L_1/L_2 \rangle \\ &= \alpha_0 - \frac{\left\langle \frac{\partial}{\partial\alpha} \ln P(x, \alpha_0) \right\rangle}{\left\langle \frac{\partial^2}{\partial\alpha^2} \ln P(x, \alpha_0) \right\rangle} \end{aligned}$$

where we use theorem on F/G, proved at end of Section 2.

Proof of Maximum Likelihood Theorem Part (a)—III

- Note, this form for $\langle\alpha^*\rangle$ is only valid to $O(1/n)$.

$$\langle\alpha^*\rangle = \alpha_0 - \frac{\left\langle \frac{\partial}{\partial\alpha} \ln P(x,\alpha_0)\right\rangle}{\left\langle \frac{\partial^2}{\partial\alpha^2} \ln P(x,\alpha_0)\right\rangle}$$

- Now, the numerator of formula for $\langle\alpha^*\rangle$ is $\left\langle \frac{\partial}{\partial\alpha} \ln P(x,\alpha_0)\right\rangle$ which is by definition

$$\begin{aligned}
\int dx P(x,\alpha_0)\frac{\partial}{\partial\alpha}\ln P(x,\alpha_0) &= \int dx \frac{\partial}{\partial\alpha} P(x,\alpha)|_{\alpha=\alpha_0} \\
&= \frac{\partial}{\partial\alpha}\left[\int dx P(x,\alpha)\right]|_{\alpha=\alpha_0} \\
&= \frac{\partial}{\partial\alpha}\cdot 1 \\
&= 0
\end{aligned}$$

as P is truely normalized.

- This proves part (a) that $\langle\alpha^*\rangle = \alpha_0$ with corrections from both of order L_3/L_2 and $L_1/L_2 \neq$ quite 0 terms which are both of order $1/n$.
- These corrections have been calculated by Yellin.

Proof of Maximum Likelihood

Part (b): Formula for Error—I

- Now to prove part (b).
- First consider "physics" method.
- This takes standard deviation of α^* as $\sqrt{-\frac{1}{L_2}}$ or rather this evaluated at $\alpha = \alpha^*$ not $\alpha = \alpha_0$ but these are same asymptotically.

$$\sigma^2_{\text{physics}} = \frac{-1}{\sum_{i=1}^{n} \left[\frac{\partial^2}{\partial\alpha^2} \ln P(x_i, \alpha_0)\right]}$$

$$\left\langle \sigma^2_{\text{physics}} \right\rangle = \frac{-1}{n \left\langle \frac{\partial^2}{\partial\alpha^2} \ln P(x_i, \alpha_0) \right\rangle} \qquad \text{where}$$

$$\left\langle \frac{\partial^2}{\partial\alpha^2} \ln P(x, \alpha) \right\rangle = \int dx P \left\{ \frac{-(\partial P/\partial\alpha)^2}{P^2} + \frac{\partial^2 P/\partial\alpha^2}{P} \right\}$$

- Now second term is just $\frac{\partial^2}{\partial\alpha^2} \int P$ and is zero again because of normalization condition. So,

$$\left\langle \sigma^2_{\text{physics}} \right\rangle = \frac{1}{\left\{ n \int dx \left(\frac{\partial P}{\partial\alpha}\right)^2 / P \right\}} \tag{18}$$

Proof of Maximum Likelihood

Part (b): Formula for Error—II

- We can calculate the correct *statistical* standard deviation in α^* by taking Equation (17) and using our F/G theorem with $f_i = \frac{\partial}{\partial\alpha} \ln P(x_i, \alpha)$, $g_i = \frac{\partial^2}{\partial\alpha^2} \ln P(x_i, \alpha)$. Generally

$$\sigma^2 = \frac{1}{n} \left\langle \left(\frac{f}{\langle g \rangle} - \frac{g\langle f \rangle}{\langle g \rangle^2} \right)^2 \right\rangle$$

- but $\langle f \rangle = 0$—proved above, so: $\sigma^2 = \frac{1}{n} \frac{\langle f^2 \rangle}{\langle g \rangle^2}$.
- It is easy to see that

$$\left\langle f^2 \right\rangle = \int \frac{\left(\frac{\partial P}{\partial \alpha} \right)^2}{P^2} \cdot P dx$$

and this is $\langle g \rangle$.

Proof of Maximum Likelihood

Part (b): Formula for Error—III

- So $\sigma^2_{\text{statistical}}$ is just $\frac{1}{n\langle g\rangle}$ which is exactly equal to $\langle\sigma^2_{\text{physics}}\rangle$ from Equation (18).
- The difference between σ^2_{physis} and $\langle\sigma^2_{\text{physics}}\rangle$ is no sweat:
- Thus, from central limit theorem $\langle\sigma^2_{\text{physics}}\rangle$ and $\sigma^2_{\text{physics}}$ differ only by relative error $O\left(\frac{1}{\sqrt{n}}\right)$ which tends to 0 as $n \to \infty$.
- One can normally ignore such "errors" in the error estimate.

Multiplication of Experiments

- If we have two experiments that are independent but depend on the same theoretical parameter α. Then

$$L_1 = \prod_{i=1}^{n_1} p(x_i, \alpha) \quad \text{is likelihood of experiment \# 1}$$
$$L_2 = \prod_{i=1}^{n_2} q(y_i, \alpha) \quad \text{is likelihood of experiment \# 2}$$

- and clearly the combined likelihood is

$$L = L_1 L_2$$

- This final L is correct expression of results of both experiments.
- i.e., one should **multiply** likelihoods when combining experiments.
- We will now apply this to case when each L_i is Gaussian in theoretical parameter α.

3.2 Overview of χ^2 Method

- This is standard and (often) fully correct (best) method for combining results from several experiments.
- It is, in fact, used (because it is easy conceptually and for computer) even when it is inapplicable,
  - e.g., if there are correlations between supposed "independent" random variables
  - or if the basic experiments have too few events to be Gaussian.
- The problem to be solved is: given N experimental measurements x_i and errors σ_i plus estimates ξ_i of the x_i, which are functions of n_T theoretical parameters.
  - Find best values of theoretical parameters.

Derivation of χ^2 from Maximum Likelihood

- Then the probability of each observation is

$$p_i = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{\frac{-(x_i - \xi_i)^2}{2\sigma^2}\right\}$$

- The likelihood

$$L = \prod_{i=1}^{N} p_i \propto \exp\left\{- \sum_{i=1}^{N} \frac{(x_i - \xi_i)^2}{2\sigma_i^2}\right\}$$

- so, the principle of maximizing the likelihood is equivalent to minimizing χ^2 where

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - \xi_i)^2}{\sigma_i^2} \tag{19}$$

- with respect to n_T theoretical parameters in ξ_i.

Mean and Standard Deviation of χ^2

- Notice that χ^2 is the sum $\sum_{i=1}^{N} v_i^2$ where v_i are independent Gaussian random variables of unit standard deviation and zero mean.
- We can calculate the χ^2 distribution exactly (Mathews and Walker pages 393–395) but in practice this is rarely necessary.
- Thus, we can easily show that—for fixed ξ_i—that
$$\langle \chi^2 \rangle = N \qquad \sigma^2_{\chi^2} = 2N \tag{20}$$
and this is sufficient for large N as central limit theorem says χ^2 will be Gaussian with above mean and standard deviation.
- Now Equation (20) is not quite correct because if we determine the theoretical parameters to minimize χ^2, then implicitly ξ_i are also random variables that are functions of x_j.
- It is not too hard to show that in this case one should just replace N by $N - n_T$ (number of degrees of freedom) in (20).
- In any case, (20) is very important because it allows an easy criterion for the goodness of fit.

Interpretation of Value of χ^2

- Thus, if value of χ^2 is many (say > 3) standard deviations away from its mean, then the theoretical form is in doubt.
- The usual maximum likelihood method does not have this advantage: remember L was unnormalized and so its value had no significance for the goodness of fit.
- This is quite a difficult problem and an elementary discussion is given in MacDonald notes, page 14 onwards.
- The only way to get a feeling for χ^2 is to use it: so I won't dwell here on the important techniques used to minimize χ^2 to get good theoretical parameters and estimate errors in the same. We will come back to this later in the notes.
- Mathews and Walker solve the simplest problem, i.e., ξ_i are linear in theoretical parameters.
- We do an example of this below.
- Bevington or similar book should be read before tackling a real problem.

Correlated Observations in χ^2

- Note that if x_i are calculated from the same population, e.g.,

$$x_1 = \frac{1}{n} \sum_{k=1}^{m} y_k$$

$$x_2 = \frac{1}{n} \sum_{k=1}^{m} y_k^2, \quad \text{etc.}$$

- Then they are not independent,
- but we can use the theorem proved at the end of Section 2 to show that x_i will (for large m) be Gaussian.
- However, there is a correlation matrix, i.e., $L \propto \exp\left\{-\frac{1}{2}(\mathrm{x} - \xi)^T M (\mathrm{x} - \xi)\right\}$ as discussed in Solmitz.
- By diagonalizing M we can (in principle) reduce this to the usual χ^2 method with x_i replaced by eigenvectors of M.

Example of Multiple Moments

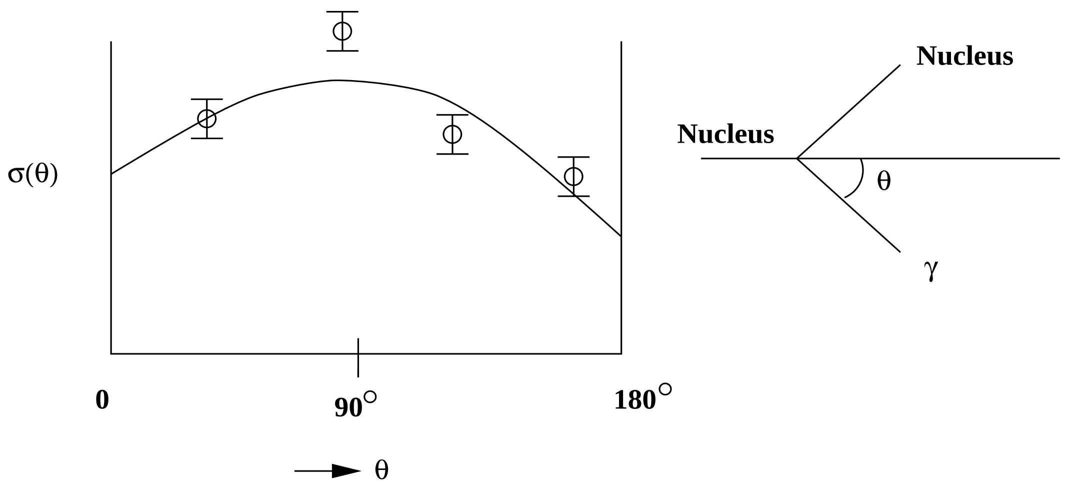
- This idea is important in the method of moments when one calculates several moments from the same distribution of events.
- In fact, one usually drops (forgets) off diagonal terms in $\mathbf{M}$ and forms χ^2 from non-independent random variables.
- This is done in high energy physics when fitting the moments ρ_{00}, ρ_{1-1}, ρ_{10} of a spin 1^- particle decay.
- Of course, in this case, the off-diagonal elements are never given by the experimentalist recording his or her data: so one is forced to set them zero.

Linear fits to Cross Section Data

Example of Use of χ^2—Overview

- Consider the example in Mathews and Walker, page 390.

$$\sigma(\theta) = a_0 + a_1 \cos\theta + a_2 \cos^2\theta + a_3 \cos^3\theta$$



- The measurements x_i's are cross sections $\sigma(\theta_i)$ at particular scattering angles θ_i, and the theoretical predictions ξ_i are linear in unknown parameters a_m.

$$\xi_i = \sum_{m=1}^{n_T} C_{im} a_m$$

χ^2 Formulation of Linear Fit for Cross Section Data—I

- Minimize:

$$\sum_{i=1}^{N} \frac{(x_i - \xi_i)^2}{\sigma_i^2}$$

$$\begin{aligned} x_i &= \text{const.} \;\; N_i \\ \sigma_i &= \text{const.} \;\; \sqrt{N_i} \\ &= \frac{x_i}{\sqrt{N_i}} \end{aligned}$$

where i'th bin (i.e., range of θ observed, which are centered at θ_i) has N_i events.

- Minimizing χ^2 with respect to unknown a_m gives

$$\sum_{i=1}^{N} \frac{(x_i - \xi_i^*)}{\sigma_i^2} \frac{\partial \xi_i^*}{\partial a_m} = 0 \;\; \text{or}$$

$$\sum_{l} M_{ml}\, a_l^* = b_m$$

- The superscript * denotes χ^2 or maximum likelihood estimate.

χ^2 Formulation of Linear Fit for Cross Section Data—II

- The "χ^2" matrix M_{ml} is given by:

$$M_{ml} = \sum_i \frac{C_{im} C_{il}}{\sigma_i^2}$$

- where M_{ml} is a positive definite symmetric matrix

$$= \sum_i \frac{(\partial \xi_i / \partial a_m)(\partial \xi_i / \partial a_l)}{\sigma_i^2}$$

- The right hand side

$$b_m = \sum_i \frac{x_i C_{im}}{\sigma_i^2} = \sum_i \frac{x_i}{\sigma_i^2} \frac{\partial \xi_i}{\partial a_m}$$

- Remember:
  - $\xi_i = \sigma_{\text{theoretical}}(\theta_i) =$
    $a_0 + a_1 \cos\theta_i + a_2 \cos^2\theta_i + a_3 \cos^3\theta_i$
  - $C_{im} = \cos^m \theta_i \quad \text{for} \quad m = 1, 2, 3$

Solution and Errors from Linear χ^2

- We can write the result in matrix form

$$\mathbf{a}^* = \mathbf{M}^{-1}\mathbf{b}$$

- We can find errors in either of the two ways
  - a) regard $\mathbf{a}^*$ as a function of x_i and calculate standard deviation from central limit theorem.
  - b) Use the shape of likelihood, $L = \exp\left(-\frac{1}{2}\chi^2\right)$.
- We have—as far as shape of L is concerned:

  $L = \exp$ (independent of $\mathbf{a}+$ no terms linear in $(\mathbf{a} - \mathbf{a}^*)$ due to "maximum" condition $-\frac{1}{2}(\mathbf{a} - \mathbf{a}^*)^T\mathbf{M}(\mathbf{a} - \mathbf{a}^*))$,
- i.e., χ^2 matrix $\mathbf{M}$ is indeed error matrix for a.

Estimate of Errors to be Used in χ^2 Fits

- One should also mention that there are two versions of χ^2 depending on whether error σ_i taken from "experiment" or "estimated theoretically."
- In the latter case, it will be a function of the parameters to be varied.
- For example, suppose x_i is number of events in some counting experiment.
  - The σ_i (experimental) $= \sqrt{x_i}$
  - but σ_i (theory) $= \sqrt{\xi_i}$.
- Now x_i is only Gaussian for large x_i, but then $\sqrt{x_i}$ and $\sqrt{\xi_i}$ are essentially same.
- If $x_i = 0$, then theoretical σ is to be preferred!—although χ^2 method itself is in difficulties as x_i not Gaussian.
- It is technically best to go back to maximum likelihood and put in true Poisson distribution for x_i.

3.3 Counter Physics—Introduction and Poisson Process—I

- For historical reasons, I am forced to take my examples from high energy physics.

- Take M happy smiling counters—all arranged in a row: Suppose they count random events that fall in their province.
- Then each of them has a parameter λ_i associated with them such that distribution of number n_i of events in counter is a Poisson distribution with parameter $\lambda_i t$ (t is time).

Counter Physics—Introduction and Poisson Process—II

- This follows from our theory of Birth processes. In high energy physics, λ_i is proportional to $\int_{\mathrm{BIN}} dt \frac{d\sigma}{dt}$ where BIN is t-bite covered by counter (t the momentum transfer is a particular kinematic range—not time!).
- The analysis is a straightforward application of maximum likelihood method:

$$p_i = e^{-\lambda_i t} \frac{(\lambda_i t)^{n_i}}{n_i!} \tag{21}$$

- (n_i observed counts, λ_i theory: t—now time!—may be known or not).

$$L = \prod_{i=1}^{M} p_i \qquad \text{or}$$

$$\ln L = -\sum_{i=1}^{M} \ln n_i! - t \sum_{i=1}^{M} \lambda_i + \sum_{i=1}^{M} n_i \ln(\lambda_i t) \tag{22}$$

Unnormalized Counter Experiment—I

- Suppose the time t—which includes beam intensity, etc.—is unknown.
- Then we use Equation (22) to determine it (regard t as a theoretical parameter)

$$\frac{\partial(\ln L)}{\partial t} = 0 \qquad \text{or}$$

$$t = \frac{\sum_{i=1}^{M} n_i}{\sum_{i=1}^{M} \lambda_i}$$

$$= \frac{N}{\Lambda}$$

where N = total Poisson parameter summed over all counters, and $\Lambda = \sum_{i=1}^{M} \lambda_i$.

- Now if we eliminate t, we get

$$\ln L = -\sum_{i=1}^{M} \ln n_i! - N + N \ln N + \sum_{i=1}^{M} n_i \ln\left(\frac{\lambda_i}{\Lambda}\right)$$

Unnormalized Counter Experiment—II

- or dropping terms independent of λ_i :

$$\begin{aligned}\ln L &= \sum_{i=1}^{M} n_i \ln\left(\frac{\lambda_i}{\Lambda}\right) \\ &= \sum_{j=1}^{N} \ln \tilde{p}(C_j) \end{aligned} \tag{23}$$

  where C_j is counter in which j'th event arrived and $\tilde{p}(k) = \frac{\lambda_k}{\Lambda}$ is relative probability of landing in k'th counter.

- Note $\tilde{p}$ is normalized $\sum_{k=1}^{M} \tilde{p}(k) = 1$.
- This is the discrete maximum likelihood method:
  - we have N observations:
  - each has only M possible outcomes
- and the $L = \prod_{j=1}^{N} \tilde{p}(C_j)$ is product of probabilities—normalized properly.

Unnormalized Counter Experiment—III

- If λ_i are functions of some theoretical parameters, we find them by maximizing Equation (23).
- In particular, if all values of $\mu_i = \frac{\lambda_i}{\Lambda}$ are free parameters, we maximize (23) subject to constant $\sum_{i=1}^{M} \mu_i = 1$:
- this gives of course:

$$\mu_i = \frac{n_i}{N}$$

Normalized Counter Experiment

- Now, we suppose t is no longer an unknown theoretical parameter, but a known number.

$$\ln L = \text{const.} + \underbrace{\{N\ln(t\Lambda) - t\Lambda\}}_{(a)} + \underbrace{\sum_{i=1}^{M} n_i \ln(\lambda/\Lambda)}_{(b)}$$

- (b) is the form we found before.
- (a) can be regarded as $\ln p_N$ where p_N is probability of getting N (which should now be regarded as an experimental observable) total counts in time t
- ($p_N = e^{-t\Lambda}(t\Lambda)^N$ which is up to a constant, the usual Poisson distribution).
- This requires some care to apply previous theorems as probabilities must be normalized in proofs in Equation (14). However, theorems do hold.
- Note Orear calls this normalized case as the extended maximum likelihood method.

Gaussian Limit of Counter Physics

- Of course, if each n_i is large

$$p_i \propto \exp\left[\frac{-\{n_i - \lambda_i t\}^2}{\lambda_i t}\right]$$

  and we get back χ^2 method.

- Note: σ_i is derived as theoretical value of error! (replace $\lambda_i t$ in denominator by n_i to get experimental value).

- So we only get something different by using the full maximum likelihood method, if some n_i are small and Poisson distribution is not in regime where the distribution is essentially Gaussian.

3.4 Bubble Chamber or Spectrometer Physics—I

- This is like Equation (16) except that every event (not every counter) has its own probability p_i.
- Typically, the events are labelled by continuous parameter(s), and so p_i is a continuous probability density.
- So far, we haven't discussed how to apply maximum likelihood method to a counting problem with continuous probability distribution
  - (lifetime example was continuous but not a counting experiment).

Bubble Chamber or Spectrometer Physics—II

- What we do is divide range of continuous parameters {such as $(0 \le \theta \le 2\pi)$ in $(1 + p\cos\theta)/2\pi$ polarization distribution} into M small bins.
- M should be so large that
  - (a) probability of two events in a given experiment falling in same bin is zero.
  - (b) the theoretical p_i is constant over each bin.
- then we have M observations—each with one of two results:
  - success—1 event in bin
  - failure—0 events in bin

Bubble/Spectrometer—Physics

Individual Events with Continuous Distributions—III

- Let N be total number of events: then likelihood is

$$\begin{aligned} L &= \prod_{\text{successes}} p(\text{success}) \prod_{\text{failures}} p(\text{failure}) \\ &= \prod_{i=1}^{N} \exp(-p_i\, t\, dx_i)(p_i\, dx_i\, t) \\ &\otimes \prod_{j=1}^{M-N} \exp(-p_j\, dx_j\, t) \end{aligned}$$

- writing out L in terms of Poisson probabilities for 0 or 1 events.
- Now $M >> N$ and M bins dx_j cover space. Thus

$$L = \exp\left[-\int p(x)\, dx\, t\right] \prod_{i=1}^{N} p_i\, dx_i\, t$$

Bubble/Spectrometer—Physics

Individual Events with Continuous Distributions—IV

- Now, fortunately, arbitrary dx_i, $i = 1 \ldots N$ are independent of theoretical parameters and so can be dropped (at further cost to the significance of normalization of L).
- So our fundamental formula for L in a continuous distribution is

$$L = \exp\left[-\int p(x)\,dx\,t\right] \prod_{i=1}^{N} p(x_i, \alpha)t$$

Bubble/Spectrometer—Physics

Individual Events with Continuous Distributions—V

- Note that just as in Section 3.3—if time t is unknown we can find it from

$$\frac{\partial L}{\partial t} = 0 \qquad \text{which gives}$$
$$t = \frac{N}{\int p(x)\, dx}$$

- when we get, with respect to remaining parameters α,

$$L(\alpha) \propto \prod_{i=1}^{N} \left\{ \frac{p(x_i, \alpha)}{\int p(x, \alpha) dx} \right\} \tag{24}$$

  —the product of normalized probability distributions.

- If t is known, the discussion is again similar to this case in Section 3.3.

Individual Events Become Binned Data—I

- If we have so many events that I can divide them into K groups of n_k $(k = 1 \ldots K)$ events where p has the same value for each member of the group, then:

$$\begin{aligned} L &= \exp\{-\textstyle\int p\, dx\, t\} \prod_{k=1}^{K} (pt)^{n_k} \\ &\propto \exp\left\{-\sum_{k=1}^{K} \lambda_k t\right\} \prod_{k=1}^{K} (\lambda_k t)^{n_k} \end{aligned} \tag{25}$$

where

$$\lambda_k = \int_{\text{bin for k'th group}} p\, dx,$$

and we return exactly to formulae in Section 3.3.

- Further, if n_k is large for each k, our Poisson distributions become Gaussian and we can use χ^2 method to analyze data.
  - This is the only case where χ^2 is correct in the sense of giving the best answer.

Individual Events Become Binned Data—II

- However, even if N is not so large that p is constant in each group, then one can still bin data and *throw away exact* values of x_i for each event—just record what bin it is in.
- Obviously, if p is not constant over each bin, this throws away information.
- However, Equation (25) is now exact for L and one chooses bins so that n_k is large and χ^2 applicable.
- This is used in practice because χ^2 is so much easier and cheaper to use.
- Often one can bin data and minimize χ^2: this gives values for theoretical parameters which can be used as initial values for expensive full likelihood fit which will then need only a few iterations.
- One can carry piggy back further—and use a super low grade method (e.g., method of moments—see Section 3.5) to initialize parameters before χ^2 fit.
- This is discussed by Solmitz ("histogram method") and Eadie, et al.

Method of Moments—I

- This is a simple and, in general, non-optimal method.
- Even here—for more than one parameter—there are two versions
  - (a) ignore correlations (used in practice but wrong) and
  - (b) include correlations (best of a bad lot!). Of course, simplicity can lead to huge savings in computer time, and so it should not be sneered at.
- The basic idea of method of moments is simple:
  - Consider a set of observations x_i, then central limit theorem implies

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i) \rightarrow \mu$$

  - where

$$\mu(\alpha) = \int p(x,\alpha) f(x) dx$$

- Here, μ is $\mu(\alpha)$ where α is some unknown parameter. So putting

$$\mu(\alpha) = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

- gives one displaymath for α with solution $\alpha = \alpha^*$:
  - α^* is the moment estimate of α.

Method of Moments—II

- Note:
  - (a) One gets a different moment method for each choice of f:
  - from maximum likelihood theorem, all of these are asymptotically worse or perhaps just not better than maximum likelihood method.
  - One must choose moment which is simple and has small expected error.
- (b) If we have to estimate n theoretical parameters α, then we just get n equations by taking n moments (i.e., n different values of f).

Method of Moments—III

- (c) Error in μ is theoretically

$$\frac{1}{\sqrt{N}}\sqrt{\int f^2 p(x,\alpha)dx - \mu_0^2}$$

  - and experimentally

$$\frac{1}{\sqrt{N}}\sqrt{\frac{1}{N}\sum_{i=1}^{N} f^2(x_i) - \left(\frac{1}{N}\sum_{i=1}^{N} f(x_i)\right)^2}$$

  - using μ_0 to be true value of mean and
  - $\mu = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$ to be estimate.

- By central limit theorem, for large N, experimental estimate will give correct error.
- However, if all x_i are the same (always true if $N = 1!$), then experimental error = 0, which is simply wrong.
- The moral is: use theoretical estimate if you only have a few events.

Example of Moment Method

- Consider

$$p(x) = \frac{1}{2}(1 + ax) \qquad -1 \leq x \leq 1$$

- Use Moment Method for any power m of x

$$\begin{aligned} \langle x^m \rangle &= \frac{a}{2} \int_{-1}^{+1} x^{m+1} dx \\ &= \frac{1}{m+2} \qquad : m \;\; \text{odd} \end{aligned}$$

- So if M_m = moment method using $\langle x^m \rangle$ and let M_m produce estimate a^*_m.
- Then

$$a^*_m = (m+2) \sum_{i=1}^{N} \frac{x_i^m}{N}$$

- The error^2 in estimate a^*_m is

$$\begin{aligned} \sigma^2|_m &= \frac{(m+2)^2}{N} \left\{ \int_{-1}^{+1} x^{2m}(1 + ax) dx - \frac{a^2}{(m+2)^2} \right\} \\ &= \frac{1}{N} \left\{ \frac{(m+2)^2}{2m+1} - a^2 \right\} \end{aligned}$$

- which is a minimum for $m = 1$ when

$$\sigma^2|_1 = \frac{1}{N}(3 - a^2)$$

Comparison of Moment Method with Maximum Likelihood—I

- Now, Maximum Likelihood Method gives result with error estimate given by

$$\frac{1}{[N\sigma^2]}\Big|_{m.l.m.} = \int_{-1}^{+1} \frac{dx}{p}\left(\frac{\partial p}{\partial a}\right)^2$$

- and from our general theory, this general formula gives in our general case:

$$\begin{aligned}
\frac{1}{[N\sigma^2]}\Big|_{m.l.m.} &= \frac{1}{2}\int_{-1}^{+1} dx \frac{x^2}{(1+ax)} \\
&= \frac{1}{2a^2}\left\{\ln\left[\frac{1+a}{1-a}\right] - 2a\right\} \\
&= \frac{1}{3} + \frac{a^2}{5} + \frac{a^3}{7} + \ldots \\
&= \frac{1}{3} \text{ when } a = 0
\end{aligned}$$

Comparison of Moment Method with Maximum Likelihood—II

- The value of $\frac{1}{N\sigma^2}|_{m.l.m.}$ is equal to $\frac{1}{N\sigma^2}|_1$ value when $a = 0$
- but
$$\frac{1}{[N\sigma^2]}|_{m.l.m.} = \infty \text{ when } a = \pm 1.$$
- So
$$\begin{aligned}\sigma^2|_{m.l.m.} &= \sigma^2|_1 \quad a = 0 \\ &< \sigma^2|_1 \quad a \neq 0.\end{aligned}$$
- i.e., maximum likelihood is always better, or as good as, moment method—in agreement with general theorem.

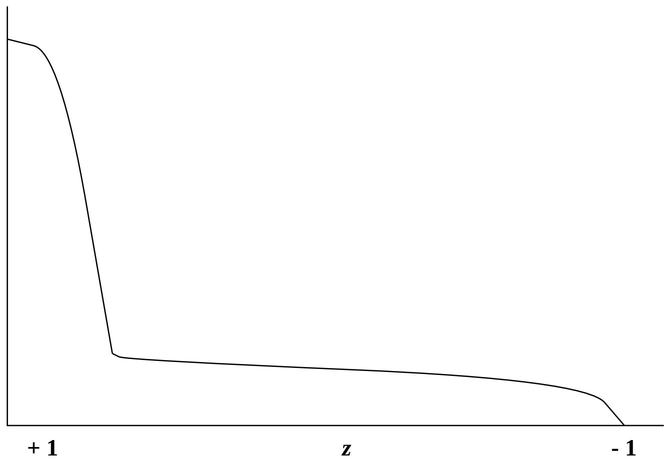
Several Parameters in Method of Moments—I

- As we have already mentioned, the case of several parameters is treated by taking several moments.
- Now our moment errors are described by a full second moment matrix for off diagonal terms come from $\int (f(x) - \langle f \rangle)\,(g(x) - \langle g \rangle)\,p(x)\,dx$ which is nonzero in general.
- It is conventional to ignore correlations:
  - The expansion in terms of $Y^l_m(\theta, \varphi)$ of $d\sigma/dt$ or an angular decay distribution is typical example.
  - The expansion functions Y^l_m are orthonormal, so

$$W(\theta,\varphi) = \sum_{lm} a_{lm} Y_{lm}$$
$$a_{lm} \propto \langle Y^*_{lm} \rangle$$

Several Parameters in Method of Moments—II

- but still a_{lm} have correlated errors as $\langle Y_{lm} Y_{l'm'} \rangle \neq \langle Y_{lm} \rangle \langle Y_{l'm'} \rangle$.
- For mean is weighted with probability distribution in θ and φ, and Y_{lm} are only orthogonal with respect to unit weight.
- These correlations in high energy physics example of $d\sigma/dt$ express fact that backward peak (<< forward) comes from correlated cancellations.



Remarks on Counting Errors in Moments—I

- In formula

$$\langle f \rangle = \frac{\sum_{i=1}^{N} f(x_i)}{N} \tag{26}$$

- I used to worry if you should include counting error ($\sqrt{N}$) in N in error estimate which otherwise only involves the Standard Deviation of f.
- If N is a priori fixed, clearly there is no $\sqrt{N}$ error but suppose the events used in Equation (26) are those that fall in a particular subregion (counter) of a bigger sample space (and $f(x)$ is defined only over this subregion).
- Then the estimate of probability of falling into counter certainly involves $\sqrt{N}$ error! What about the mean of a quantity within subregion—does this involve this error?
- The problem is: what is error in a moment: should it include counting error?
- The answer is no—one uses identical formulae in case where number of events prescribed and not subject to counting error.

Remarks on Counting Errors in Moments—II

- We show this in a model problem in two variables:
- x is variable on which moment taken and y is binned.
- Choose:

$$\begin{aligned} P(x,y) &= p(y)\frac{[1+\alpha(y)x]}{2} \\ \text{Take a } y\text{-bin} \quad & \quad y_1 \le y \le y_2 \\ \text{set } S(y) = 1: \quad & \quad y_1 \le y \le y_2 \\ = 0: \quad & \quad \text{otherwise.} \end{aligned}$$

- Assume $\alpha(y)$ is constant at α over bin and then method of moments is

$$\alpha = 3 \ \frac{\sum_{i=1}^{N_T} f(x_i, y_i)}{\sum_{i=1}^{N_T} g(x_i, y_i)} = \frac{3F}{G}$$

- where

$$\begin{aligned} f(x,y) &= xS(y) \\ g(x,y) &= S(y) \end{aligned}$$

Remarks on Counting Errors in Moments—III

- There are N_T total events and N fall into bin $y_1 \leq y \leq y_2$.
- Now we can apply F/G theorem from end of Section 2,
- and it is not hard to show that f and g separately have counting errors $\sqrt{N}$,
- but they are correlated and correlation cancels counting errors, so that α has identical error formula to standard moment method.
- (The estimate of $p(y)$ will involve counting error, of course).

4: MINIMIZATION

4 Minimization

4.1 Statement of Problem

- The basic problem is to minimize $f(\alpha_1 \ldots \alpha_n)$—a function of n variables $\alpha_1 \ldots \alpha_n$.
- Equivalently, we must find the zero of an arbitrary function $f'(x_1 \ldots x_n)$ of n variables.
- This arises in previous sections when we wish either to minimize χ^2 or maximize the likelihood L (equivalently to minimize—$\ln L$).
- The methods divide according to

  (i) Can calulate f only

  (ii) Can calculate f, f', and f''.
- (i) is extremely time consuming (i.e., it takes a lot of function evaluations to converge to desired minimum).
- In all problems I have faced, it has been simple to calculate derivatives and so I will only discuss (ii).

4.2 Derivative Formalism for χ^2—I

- One great advantage of the χ^2 or maximum likelihood problem is that to calculate the second derivative of the function to be minimized only requires the *first* derivative of the theoretical functions.
- First consider this in the case of χ^2:

$$\chi^2 = \sum_{i=1}^{N} \frac{(t_i - e_i)^2}{\sigma_i^2} \tag{27}$$

- where e_i is experimental measurement, σ_i is error, and t_i is theory—a function of n parameters.
- We take, without loss of generality, $n = 1$ and one parameter α in the following:

$$\begin{aligned}
\frac{\partial \chi^2}{\partial \alpha} &= \sum_{i=1}^{N} 2(t_i - e_i)\frac{\partial t_i}{\partial \alpha}/\sigma_i^2 \\
\frac{\partial^2 \chi^2}{\partial \alpha^2} &= S_1 + S_2 \\
S_1 &= \sum_{i=1}^{N} 2(t_1 - e_i)\frac{\partial^2 t_i}{\partial \alpha^2}/\sigma_i^2 \\
S_2 &= 2\sum_{i=1}^{N} \frac{\left(\frac{\partial t_i}{\partial \alpha}\right)^2}{\sigma_i^2}
\end{aligned}$$

Derivative Formalism for χ^2—II

- Now, S_1 is $O(\sqrt{N})$ and asymptotically negligible.
- To show this, regard it as a random variable which is a function of the N random variables $e_1 \ldots e_N$.
- By the Central Limit Theorem, S_1 has mean

$$\mu = \sum_{i=1}^{N} \mu_i$$

- and Standard Deviation σ.

$$\sigma^2 = \sum_{i=1}^{N} \omega_i^2$$

- where

$$\frac{2(t_i - e_i)}{\sigma_i^2} \frac{\partial^2 t_i}{\partial \alpha^2}$$

  has mean μ_i and Standard Deviation ω_i.

Derivative Formalism for χ^2—III

- Now if t_i was evaluated with true theoretical parameters $\mu_i = 0$ and ω_i is some number.
- So σ^2 is $O(N)$ and S_1 is $O(\sqrt{N})$ as desired.
- Meanwhile, S_2 is $O(N)$ as it is sum of N positive numbers.
- Whence—near solution

$$\frac{\partial^2 \chi^2}{\partial \alpha^2} = S_2 = 2 \sum_{i=1}^{N} \left(\frac{\partial t_i}{\partial \alpha} \right)^2 / \sigma_i^2$$

- As claimed: second derivatives of χ^2 only require evaluation of first derivatives of theory (only true when N large).

Derivative Calculation for Maximum Likelihood—I

- Let's do the same thing for maximum likelihood method.

$$-\ln L = -\sum_{i=1}^{N} \ln q_i$$

- where q_i are probability of our $i = 1 \ldots N$ events.
- Assume q_i are normalized $\int q(\alpha)dx = 1$ as in Equation (25).
- This normalization is crucial but even when it does not seem to be true as in our subtle modifications in Section 4.3 with the "generalized" maximum likelihood method, one can easily modify the analysis, and show it is still valid.
- We get in normalized q_i case,

$$\frac{-\partial^2 \ln L}{\partial \alpha^2} = S_1 + S_2$$

- with

$$S_1 = \sum_{i=1}^{N} \frac{1}{q_i} \frac{\partial^2 q_i}{\partial \alpha^2}$$

$$S_2 = \sum_{i=1}^{N} \frac{1}{q_i^2} \left(\frac{\partial q_i}{\partial \alpha}\right)^2$$

Derivative Calculation for Maximum Likelihood—II

- Now, $S_{1,2}$ are directly of form suitable to apply central limit theorem.
- Thus, they are sums over independent random variables taken from same distribution.

$$
\begin{aligned}
\langle S_1 \rangle &= N \left\langle \frac{1}{q} \frac{\partial^2 q}{\partial \alpha^2} \right\rangle \\
&= N \int q \cdot \frac{1}{q} \frac{\partial^2 q}{\partial \alpha^2} d\alpha \\
&= N \frac{\partial^2}{\partial \alpha^2} \left\{ \int q \, d\alpha \right\} \\
&= N \frac{\partial^2}{\partial \alpha^2} (1) = 0
\end{aligned}
$$

- So again S_1 only has standard deviation term and is $O(\sqrt{N})$.
- Similarly S_2 is $O(N)$ and dominates.
- It follows that exactly same methods can be applied to χ^2 and maximum likelihood fits.
- In the future we will only mention the former.

Derivative Calculation for Maximum Likelihood—III

- As far as your computer program is concerned, one can handle χ^2 and likelihood with same program.
- They just differ in calculational complexity N.
  - In χ^2, N is the number of "bins."
  - In likelihood, N is number of events.
- My old fitting programs
  - Manxcat—Marquardt's Method
  - "Feefoefum's Well Nigh Perfect Fit" using eigenvalue method
  - see later for methods
- were designed to interface with either χ^2 or maximum likelihood.

4.3 The Derivative Method for χ^2—I

- Let's first generalize the previous discussion to n parameters—getting a vector F_j of first and matrix M_{jk} of second derivatives

$$F_j = \frac{\partial \chi^2}{\partial \alpha_j} = \sum_{i=1}^{N} \frac{(t_i - e_i)}{\sigma_i^2} \frac{\partial t_i}{\partial \alpha_j}$$
$$M_{jk} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial \alpha_j \partial \alpha_k} = \sum_{j,k=1}^{N} \frac{1}{\sigma_i^2} \frac{\partial t_i}{\partial \alpha_j} \frac{\partial t_i}{\partial \alpha_k} \qquad (28)$$

- Note that—in our approximation—M is positive semi-definite—which is also a technical advantage.

The Derivative Method for χ^2—II

- Now given any guess $\alpha = \alpha_0$, expand χ^2.

$$\chi^2(\alpha) = \chi^2(\alpha_0) + \sum_j (\alpha - \alpha_0)_j \, F_j + \sum_{jk} (\alpha - \alpha_0)_j \, (\alpha - \alpha_0)_k \, M_{jk}$$

- Adopt a vector notation with $\Delta\alpha_j = \alpha_j - \alpha_{0j}$.

$$\chi^2(\underline{\alpha}) = \chi^2(\underline{\alpha}_0) + \Delta\underline{\alpha \text{ F}} + \Delta\underline{\alpha}^T M \Delta\underline{\alpha} \qquad (29)$$

- Minimum condition is

$$\frac{\partial\chi^2(\underline{\alpha})}{\partial\alpha_j} = 0 \qquad \text{or}$$

$$M\Delta\underline{\alpha} = -\underline{F} \qquad (30)$$

- So we have new guess

$$\underline{\alpha} = \Delta\underline{\alpha} + \underline{\alpha}_0 = \underline{\alpha}_0 - M^{-1}\underline{F}$$

  for true minimum.

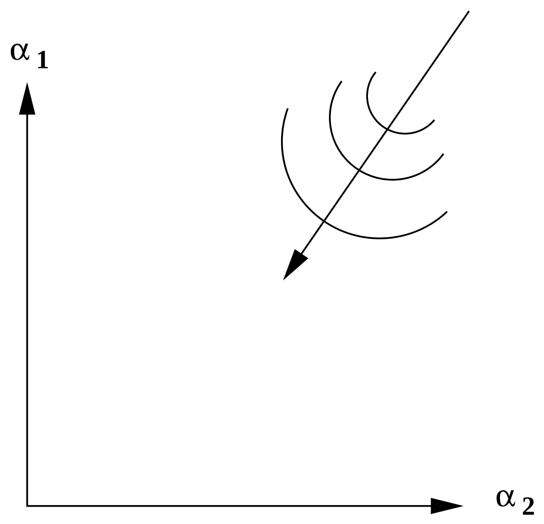
- Note that as **M** positive definite in Equation (30), this new value of $\underline{\alpha}$ indeed gives a minimum (not a maximum or saddle point), **as long as second order Taylor expansion valid**.

Error Calculation from χ^2 Expansion

- When we are at a minimum, $\underline{\mathrm{F}} = 0$ and Equation (29) becomes
$$\chi^2(\underline{\alpha}) = \chi^2(\underline{\alpha}_0) + \Delta\underline{\alpha}^T M \Delta\underline{\alpha}$$
- or remembering that the likelihood
$$L = \exp\left(-\frac{1}{2}\chi^2\right)$$
$$L = L(\alpha_0) \otimes \exp\left(-\frac{1}{2}\Delta\underline{\alpha}^T M \Delta\underline{\alpha}\right) \qquad (31)$$
- Equation (31) is expected Gaussian form for L.
- According to Section 3, M is just inverse of error matrix, i.e.,
$$\langle(\alpha^* - \tilde{\alpha})_j(\alpha^* - \tilde{\alpha})_k\rangle = M^{-1}_{jk}$$
  - where $\underline{\alpha}^*$ = estimate from minimizing χ^2,
  - $\underline{\tilde{\alpha}}$ = true parameters.
- In particular, "error" σ_j in α_j is
$$\sigma_j = \sqrt{(M^{-1})_{jj}} \qquad (32)$$
and the other elements in M^{-1} are the much forgotten correlation parameters.
- Note that $\Delta\alpha_j = \sigma_j$, all other $\Delta\alpha_j = 0$, produces a χ^2 change of 1.

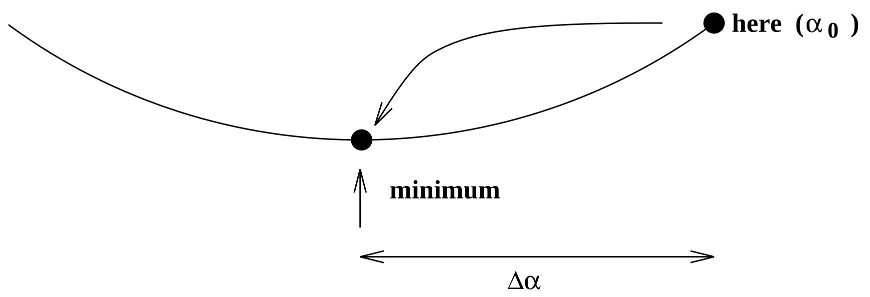
4.4 The Real World and χ^2 Minimization—I

- For a real problem, the iterative method described in Section 4.3 always diverges: one must invent various artistic devices to ensure procedure converges.
- The simplest reliable algorithm is called *Marquardt's* method and this is intuitively easy to understand.
- Consider χ^2 as a function of several parameters.



The Real World and χ^2 Minimization—II

- χ^2 will have a bunch of hills and valleys and we are trying to discover the location of the lowest point in the valley.
- If our assumptions are correct, then χ^2 is just a neat quadratic function, and we have found a step that exactly moves to the minimum.



- Unfortunately, this is not usually correct, as the step $\underline{\alpha}_0 \to \underline{\alpha}_0 + \Delta\underline{\alpha}$ *maybe too large* to support the assumption, that terms in $\frac{\partial^k(\chi^2)}{\partial\alpha^k}$, $k \geq 3$ are zero, to be correct.

The Real World and χ^2 Minimization

Example of Two Variables—I

- Consider the case when there are two variables.

$$\begin{aligned}\chi^2 &= \chi_0^2 + \text{linear terms} \\ &+ A_1\left(\alpha_1 - \alpha_1^0\right)^2 + A_2\left(\alpha_2 - \alpha_2^0\right)^2 \\ &+ B_1\left(\alpha_1 - \alpha_1^0\right)^3 + B_2\left(\alpha_1 - \alpha_1^0\right)^2\left(\alpha_2 - \alpha_2^0\right) + \ldots \\ &+ \ldots\end{aligned}$$

- where we have taken the special case where there is no term proportional to $(\alpha_1 - \alpha_1^0)(\alpha_2 - \alpha_2^0)$;
- such a term only confuses the current discussion, and we will soon see how to put it back in.
- The difficulty (lack of convergence of iterative scheme) occurs when

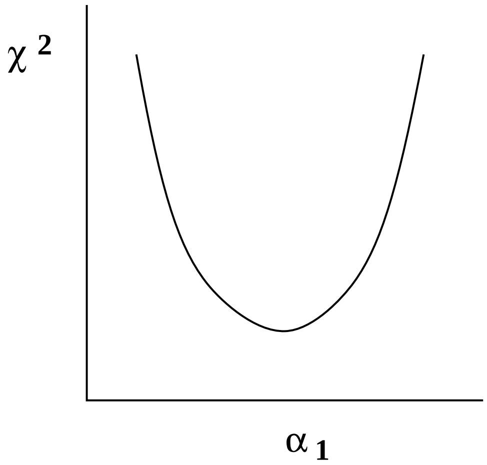
$$A_1 >> A_2$$

The Real World and χ^2 Minimization

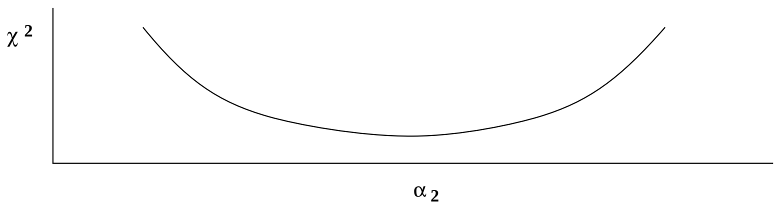
Example of Two Variables—II

Pictures of Two Variable Case

- so that as a function of α_1, we get



- and as a function of α_2, we get



  with a valley with steep walls in α_1 and gentle walls in α_2 direction.

- The predicted shift from Equation (30), $\Delta\alpha_2 \sim \frac{1}{A_2}$ and is typically large compared to $\Delta\alpha_1 \sim \frac{1}{A_1}$.

The Real World and χ^2 Minimization

Example of Two Variables—III

- Unfortunately, there is no reason to believe that this large shift is well estimated as assumptions that lead to expansion break down.
  - Terms like $B_2(\Delta\alpha_1)^2\Delta\alpha_2$ are quite likely to be big.
- We see that estimate in Equation (30) is false and one will typically jump out of valley and find

$$\chi^2\left(\underline{\alpha}_0 + \Delta\underline{\alpha}\right) >> \chi^2\left(\underline{\alpha}_0\right)$$

- The method will diverge!
- In the above example, the estimate for $\Delta\alpha_1$ was valid and not spoiled by $\Delta\alpha_2$ estimate.
- In a real example, these will be mixing between α_1 and α_2 due to nonzero $\frac{\partial^2\chi}{\partial\alpha_1\partial\alpha_2}$.
- Both shifts $\Delta\alpha_1$, $\Delta\alpha_2$ will be large and specious.

Eigenfunction Analysis

- There is an intuitively obvious solution of this.
- Consider the second order expansion term

$$(\underline{\alpha} - \underline{\alpha}_0)^T \, \mathbf{M} \, (\underline{\alpha} - \underline{\alpha}_0)$$

- $\mathbf{M}$ is a symmetric positive semi-definite matrix.
- $\mathbf{M}$ can be diagonalized

n x n matrix

$$M \longrightarrow P^T \; D P$$

$$D = \begin{bmatrix} \lambda_1 & & & & \bigcirc \\ & \cdot & & & \\ & & \cdot & & \\ & & & \cdot & \\ \bigcirc & & & & \lambda_n \end{bmatrix}$$

- λ_j are eigenvalues of M.
- D is a diagonal matrix with diagonal entries λ_j.
- $\mathbf{P}$ is $n \times n$ matrix formed by eigenfunctions of M.

Transformation of M to Eigendirections

- Put $\underline{\beta} = \mathbf{P}\underline{\alpha}$.
- Then $\chi^2 =$
  $\chi_0^2 +$ linear terms $+ \lambda_1 (\beta_1 - \beta_{01})^2 + \lambda_2 (\beta_2 - \beta_{01})^2$
  $+ \ldots$.
- Now the shifts $\Delta\beta_j$ can be calculated independently and do not affect each other.
- Assume that we order the eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$$

- Then clearly the first new shifts $\Delta\beta_j$ corresponding to $\lambda_1 \ldots \lambda_L$, say are reliable and the remainder $j = L + 1 \ldots n$ unreliable.
- Suppose we ignore the predictions for these last and set

$$\begin{aligned} \Delta\beta_j &= \text{Predicted value } j \leq L \\ &= 0 \quad j > L \end{aligned}$$

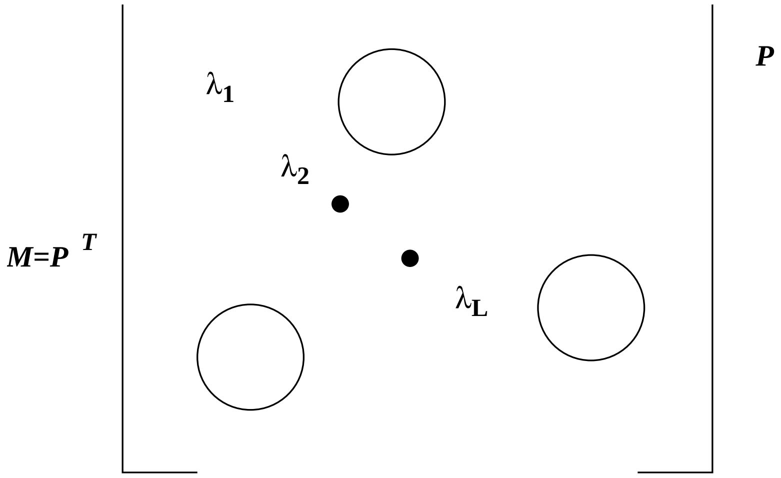
- We need a criterion for choosing L but subject to this, this method is intuitively reasonable.

What I Used to Do!

- In a program that I used for many years successfully, I chose L by the largest value such that

$$\lambda_L \geq r\lambda_1 \quad \text{and} \quad \lambda_{L+1} < r\lambda_1$$

- where initially the ratio $r = 0.1$ and r was increased or decreased based on experience,
- i.e., if the resultant $\chi^2\left(\underline{\beta}_0 + \Delta\beta\right)$ was indeed smaller (as predicted) or larger than $\chi^2\left(\underline{\beta}_0\right)$.
- This method is equivalent to replacing M by the regularized version.



Why is M Poorly Conditioned?

- I note that in using this method, I got to look at values of the λ_j. It was not uncommon for
$$\frac{\lambda_{\max}}{\lambda_{\min}} = \frac{\lambda_1}{\lambda_n} \sim 10^7 \text{ for } n \sim 30$$
- You can ask why one gets such large ratios.
- Intuitively, the eigenvectors of M with small eigenvalue correspond to linear combinations of α_j that are badly determined.
- So very small λ's say that your theoretical formalism is redundant or alternatively the data inadequate.
- For instance, in fitting a cross section $\frac{d\sigma}{d(\cos\theta)}$ to $\sum_{j=1}^{n} \alpha_j P_{j-1}(\cos\theta)$
  - then often the coefficients of the higher order Legendre functions are poorly determined, and
  - if you choose n to be "too big," then small eigenvalues will develop.
- In this case, one can cure the problem by reducing the top index n but in general no such simple solution exists.

Marquardt's Method

- Marquardt's method is a simpler technique but with similar goals to the eigenvalue method.
- One replaces

$$\mathrm{M} \rightarrow \mathrm{M} + QI$$

- where **I** is the identity matrix.
- So this means we replace M by

$$M = P^T \begin{bmatrix} \lambda_1 + Q & & & & & \\ & \lambda_2 + Q & & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & \lambda_1 + Q \end{bmatrix} P$$

- Clearly, if $\lambda_L \sim Q$, then the two formalism's are roughly equivalent.
- Eigenvectors with eigenvalues $> Q$, are shifted by naive prediction while those with eigenvalue $<< Q$, have their shift drastically reduced.

Estimate of Marquardt

Parameter Q

- One needs an artistic method to estimate Q which again is varied based on experience. Q is increased when predicted

$$\chi^2\left(\underline{\beta}_0 + \Delta\beta\right) > \chi^2\left(\underline{\beta}_0\right)$$

and is otherwise decreased or left unchanged.

- The natural "scale" for Q is given by

$$\frac{1}{n} Tr\,\mathrm{M} = \frac{1}{n} \sum_{j=1}^{n} \lambda_j$$

the average eigenvalue.

- Marquardt's method is less precise than the diagonalization method but much faster;
- finding eigenvalues and vectors is time consuming.

Experience with Marquardt's Method

- I have used both successfully.
- I used Marquardt's method in the analysis of some 10^6 Fermilab events where on each event, I fitted known shower slopes to measured photon energy measurements in a segmented lead-scintillator sandwich.
- Here $n \sim 10$ and the large number of fits (10^6) required a fast method.
- In these million fits, I saw no problem with divergence.
- This is case where it was necessary to find a fast reliable method.
- I used the eigenvalue method in analysis of experimental data in terms of theoretical models.
- Here, one is typically only doing a few (maybe one) fit and so speed of minimizer is irrelevant.

5: GOODNESS OF FIT

5. Goodness of Fit

5.1 The χ^2 Distribution

- When the maximum likelihood method reduces to χ^2, it is relatively easy to judge goodness of fit.

$$\chi^2 = \sum_{i=1}^{n} \frac{(e_i - t_i)^2}{\sigma_i^2} \tag{33}$$

where t_i depend on m theoretical parameters.

- Then the result of minimizing Equation (33) with respect to m theoretical parameters should be distributed according to χ^2 distribution with $n - m$ degrees of freedom.
- This function is plotted on p. 64 of EDJRS.
- If k is number of degrees of freedom.

$$p(\chi^2) = \frac{\left(\frac{\chi^2}{2}\right)^{\frac{k}{2}-1} \exp\left(-\frac{\chi^2}{2}\right)}{2\Gamma\left(\frac{k}{2}\right)} \tag{34}$$

- $\langle\chi^2\rangle = k$ and Standard Deviation is $\sqrt{2k}$.

χ^2 Distribution for One Degree of Freedom

- Consider χ^2 with one degree of freedom

$$\begin{aligned}
\chi^2 &= \frac{(x-t)^2}{\sigma^2} \text{ where} \\
p(x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-t)^2}{2\sigma^2}\right) \\
p(\chi^2) &= p(x)\frac{dx}{d\chi^2} = \frac{p(x)\sigma^2}{2(x-t)} \cdot 2 \\
p(\chi^2) &= \frac{\exp\left(-\frac{\chi^2}{2}\right)}{\sqrt{2\pi}\ 2\ \sqrt{\chi^2}} \cdot 2
\end{aligned}$$

- which is value in Equation (34) using $\Gamma(1/2) = \sqrt{\pi}$
- (Here, factor of two comes as $\pm(x-t)$ given same χ^2 value).

General χ^2 Distribution

- One can derive the χ^2 distribution for k degrees of freedom in two ways.
- 1.) Mathews and Walker
- Direct calculation as for $k = 1$ noting that $\sum_{i=1}^{k} x_i^2$ with x_i of mean 0, Standard Deviation 1 is r^2 where r is radial distance in k-dimensional space.
- Then factors multiplying $\exp\left(-\frac{\chi^2}{2}\right)$ in Equation (34) are just volume element in k dimensions.
- 2.) Eadie, et al. show characteristic function of χ^2 with one degree of freedom is

$$\int_{-\infty}^{+\infty} \exp(i\,t\,\chi^2)\,p(\chi^2)\,d(\chi^2) = \frac{1}{\sqrt{1 - 2\,i\,t}}$$

- Therefore, characteristic function for χ^2 with k degrees of freedom is $(1 - 2\,i\,t)^{-\frac{k}{2}}$
- and this can be inverse fourier transformed to give Equation (34).

χ^2 Characteristic Function

- Introduce the Characteristic Function $F(\chi^2)$ by

$$
\begin{aligned}
F(\chi^2) &= \int_0^{\chi^2} p(\chi^2) d(\chi^2) \\
&= \frac{1}{2\Gamma\left(\frac{k}{2}\right)} \int_0^{\chi^2} \left(\frac{u}{2}\right)^{\left(\frac{k}{2}\right)-1} e^{\frac{-u}{2}} du
\end{aligned}
$$

- F has probability distribution $w(F)$

$$
w(F) = p(\chi^2) \frac{d(\chi^2)}{dF} = 1
$$

so F is uniformly distributed between 0 and 1.

5.2 Confidence Level

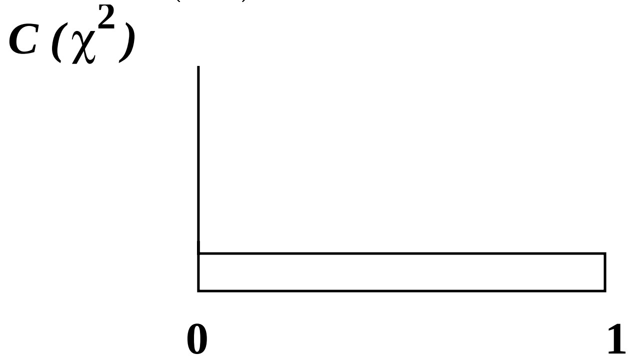
- We prefer to define $C(\chi^2) = 1 - F(\chi^2)$ which is called the confidence level

  $$\begin{aligned} C(\chi^2) &= 1 \ \text{ is } \ \chi^2 = 0 \quad \text{“good”} \\ &= 0 \ \text{ is } \ \chi^2 = \infty \ \ \text{“bad” or “low confidence”} \end{aligned}$$

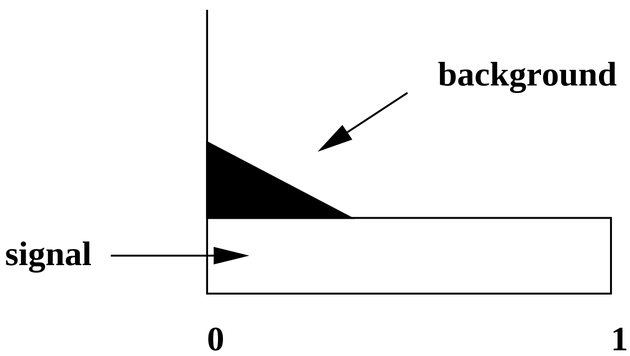
- Suppose you conduct a series of experiments which gives you a sequence of χ^2 values, $\chi^2|_{1,2\ldots}$.
  - An example in high energy physics would come from analyzing a set of events each containing particle tracks—each track would have an individual χ^2.
  - A given experiment would generate many million such χ^2 values!

Typical Confidence Level Distributions

- If one can do the "errors" right, the resultant distribution in $C(\chi^2)$ should be flat



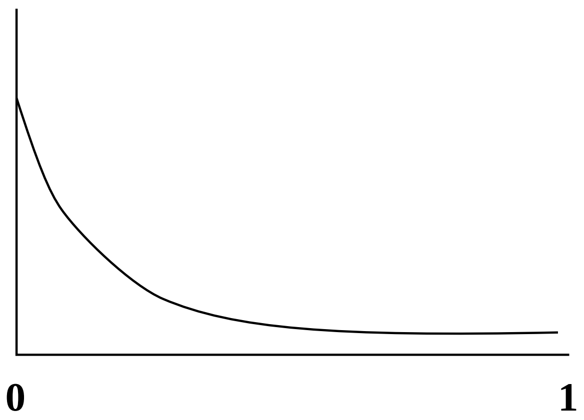
- But if you have "background," i.e., events that do not satisfy claimed hypothesis, then one will have an excess of events at large $\chi^2 >> (n - m)$ or small $C(\chi^2)$.
- $C(\chi^2)$ will look like



- The plot of $C(\chi^2)$ is a good way of estimating background.

Real World Confidence Levels

- Unfortunately, one often gets $C(\chi^2)$ plots that look like



- with no clear "flat" piece.
- This maybe because your "signal" is very small.
- It could be because you got your errors wrong (!), e.g., maybe the true σ_i are all larger than ones used in calculating χ^2.
- In this case, it makes sense to scale χ^2 $\left(\chi^2_{\text{new}} = \text{const.}\ \chi^2_{\text{old}}\right)$ until we get a nice flat distribution for confidence level C near 1 and analyze as above.

Error Estimate from χ^2 Value

- If we expand χ^2 near its minimum with—for convenience—just one theoretical parameter, λ.

$$\chi^2 = \chi_0^2 + \alpha\,(\lambda - \lambda_0)^2 \tag{35}$$

- where λ_0 is our estimate of λ.
- Then probability distribution of λ is just

$$\alpha \exp\left(-\frac{\chi^2}{2}\right)$$
$$\alpha \exp\left(-\frac{\alpha(\lambda - \lambda_0)^2}{2}\right)$$

  i.e., $\frac{1}{\sqrt{\alpha}}$ is Standard Deviation of estimate.

- Returning to Equation (35), we find that if we put $\lambda = \lambda_0 \pm \frac{1}{\sqrt{\alpha}}$ (i.e., at its one standard deviation value),
- χ^2 is increased by 1.
- This χ^2 increase "rule" is common method of finding "error" in estimate λ even when the Taylor expansion equation is not a good approximation.

5.3 Bartlett's Function—Overview

- The previous analysis can be generalized to Bartlett's S function which is defined as

$$S(\lambda) = \frac{1}{\sqrt{I(\lambda)}} \frac{\partial \ln L}{\partial \lambda}$$

- $I(\lambda)$ is the information $\left\langle \left(\frac{\partial}{\partial\lambda} \ln L\right)^2 \right\rangle$

$$I(\lambda) = N \int \left[\frac{\partial}{\partial\lambda} \ln P(x,\lambda)\right]^2 P(x,\lambda) dx$$

- where N = Number of events and $\int \left[\frac{\partial}{\partial\lambda} \ln P(x,\lambda)\right]^2 P(x,\lambda) dx$ can be calculated from theoretical forms.

Bartlett's Function for χ^2

- From Equation (35) we have

$$\frac{\partial}{\partial\lambda}\ln L = -\frac{2\alpha(\lambda-\lambda_0)}{2} \qquad \text{for} \quad L = \exp\left(-\frac{\chi^2}{2}\right)$$

- and using our earlier asymptotic result

$$\begin{aligned} \alpha &= I(\lambda) \quad \text{we find} \\ S(\lambda) &= \sqrt{I(\lambda)}(\lambda_0-\lambda) \\ &= \sqrt{\alpha}(\lambda_0-\lambda) \end{aligned}$$

- Therefore, $\lambda = \lambda_0 \pm \mathit{one}$. Standard Deviation is just $S(\lambda) = \mp 1$.
- In fact, for any number of parameters n,
- $S(\lambda)$ has mean 0 and variance 1, and so the rule $S(\lambda) = \pm 1$ ($\pm l$ for l Standard Deviations) is an appropriate generalization of the rule that one increases χ^2 by $1(l)$.

5.4 Hypothesis Testing—I

- We wish to test a specific hypothesis H_0 (referred to as null hypothesis) against a set of possible other hypotheses.
- In goodness of fit analysis, we test H_0 (= hypothesis that data described by claimed theory with parameters found from maximizing L) against "all" other hypotheses.
- If we are trying to find type of a particle then
  - H_0 might be particle is an electron,
  - H_1, that is a pion and these could be only possibilities.

Hypothesis Testing—II

- We are given our Standard N observations $(x_1 \ldots x_N)$ on which to make our decision.
- We do this by constructing yet another statistic Y which is some function of $x_1 \ldots x_N$.
- We assume that Y has distribution ${P_Y}^{(i)}$ if hypothesis i is true ($i = 0$ or 1 if two hypotheses).
- Y could be vector valued but we will take scalar case here.
- In the best of all possible worlds, one can find a Y such that the values taken by it under two hypotheses are disjoint, i.e.,

$$\int {P_Y}^{(0)}(y){P_Y}^{(1)}(y)dy = 0 \tag{36}$$

- Unfortunately, Equation (36) is not always attainable
  - but one should always try to minimize the probability overlap integral.

Hypothesis Testing—III

- We define r as region in Y space where we will deem hypothesis (1) to be true and (0) to be false.
- In the remaining region of Y space, $R - r$, (R is total Y range) we will take (0) to be true, (1) to be false.
- This scheme is characterized by two constants:
- 1. α = level of significance or size of test $= \int_r P_Y^{(0)}(y)dy$.
  - α is chance that (0) will be declared false when it is in fact true.
- 2. $\beta = \int_{R-r} P_Y^{(1)}(y)dy$ is the contamination, i.e.,
  - β is chance that (0) will be declared true when it is in fact false.
- Clearly, it depends on the application for the balance between efficiency $1 - \alpha$ and contamination β.

5.5 Student's t Distribution for Hypothesis Testing—I

- Suppose that the x_i are drawn from a normal (Gaussian) distribution of unknown mean and Standard Deviation.

$$H^{(0)} \text{ is hypothesis that mean } = \mu^{(0)}$$
$$H^{(1)} \text{ is hypothesis that mean } = \mu^{(1)}$$

- Define

$$t = \frac{\left(\overline{x} - \mu^{(0)}\right)\sqrt{n}}{\overline{\sigma}} \text{ where}$$
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\overline{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

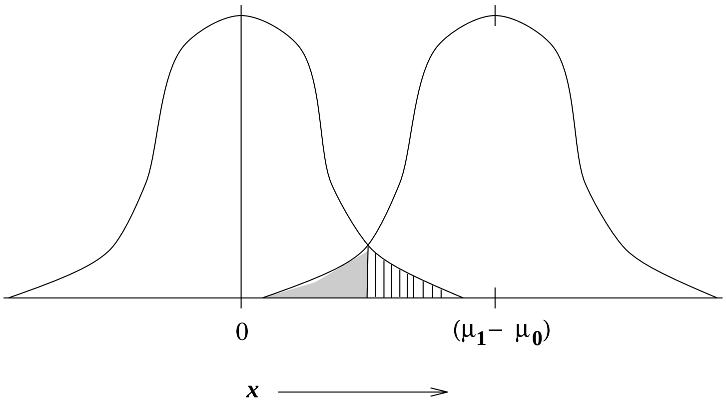
- Under hypothesis (0), t follows a so-called t-distribution with $n-1$ degrees of freedom. $(f = n-1)$.
- t would be Gaussian with unit standard deviation if $\overline{\sigma}$ taken from theory—not from experiment.

Student's t Distribution for Hypothesis Testing—II

- The probability distribution is

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

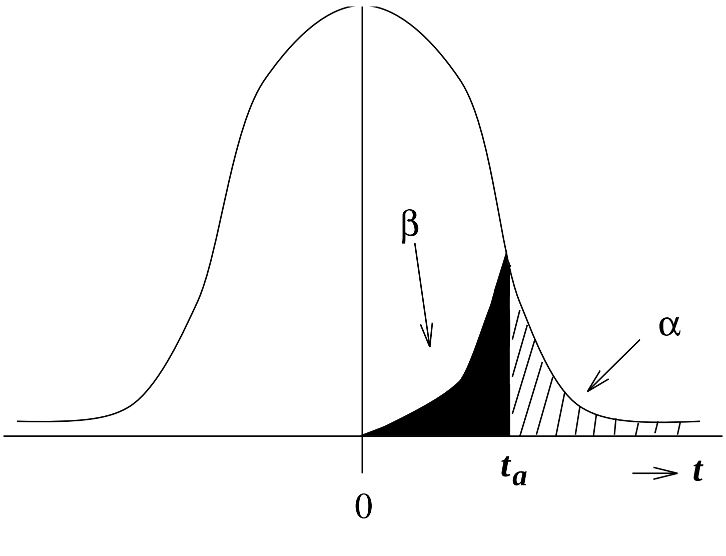
- $f = 1$ is Cauchy distribution.
- As $f \to \infty$, the t distribution becomes Gaussian with mean 0 and Standard Deviation 1.
- This can be expected for as $f \to \infty$, using data to find $\overline{\sigma}$, should be accurate.



- If $\mu_1 > \mu_0$, choose $0 < t_a < (\mu_1 - \mu_0)$, and
- one could say that 0 is true if $t < t_a$ and 1 is true if $t > t_a$.

Student's t Distribution for Hypothesis Testing—III

- The corresponding efficiency/contamination are trivially calculable from cumulative probability functions of the student's t distribution.
- As discussed on page 168 of Brandt, this method can be adapted (use average $\overline{\sigma}$) to test if two sets of measurements have the same mean
  - (e.g., Set 1: Smokers, Set 2: Non-Smokers)



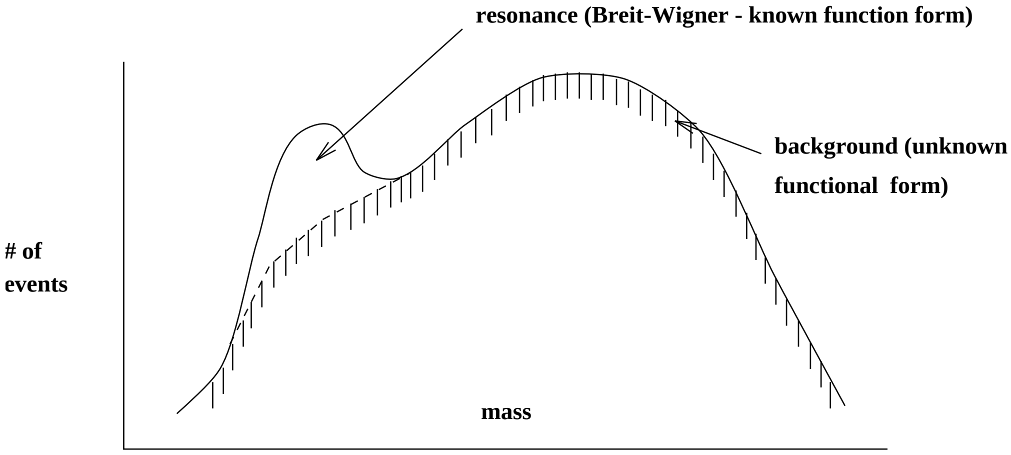
- Above, we show the regions with α (level of significance) and β (contamination), defined earlier.

5.6 Robust Estimation–Motivation–I

- See Andrews, et al. "Robust Estimates of Location," Princeton University Press;
- "Robustness in Statistics," edited by R. L. Laurer, G. N. Wilkinson, Academic Press (1979).
- In the maximum likelihood method we found the method that had the smallest possible Standard Deviation (asymptotically at least).
- However, as we emphasized in the study of non-parametric estimates, it is quite possible (and indeed probable) that systematic error from *incorrect* theoretical form of probability is large.
- In many cases, one knows reasonably well the analytic form of the "interesting" effect but what one observes has other totally unknown contributions.
- One must choose between parameterizing the unknown or ignoring known form of known effect.

Robust Estimation—Motivation—II

- Consider an example from physics where one observes mass distribution with a Breit Wigner for a resonance, say, plus "small" (usually large!) contributions to the probability distribution (e.g., background which must be estimated as polynomial/phase space) which must be parameterized in an ad hoc fashion.



- It is hard to evaluate distortion due to an incorrect parameterization of these unknown contributions.
- Obviously, the correct choice of method will depend on relative size of signal (with known functional form), and background.

Robust Estimation—Motivation—III

- There is thus a great deal of interest in the study of so called *robust* techniques which are insensitive to errors in our knowledge about the underlying distribution.
- Statisticians are very interested in such techniques but (apart from non-parametric methods we have already discussed) there has been little use/experience of them in the type of hard problems that physicists are want to find.
- Further, the derivation of robust methods is rather intuitive and not as mathematically precise as those of classical statistics.

Robust Determination of Mean—I

- The first book cited (Andrews) has an instructive study of a very simple problem.
- Given a univariate sample $\{x^{(i)} \quad i = 1, \quad N\}$ from some distribution,
- Find an estimate of the mean μ of the underlying probability distribution.
- The natural method is to estimate μ from the arithmetic mean

$$\mu^*_{Am} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$$

- As long as the underlying distribution has a finite Standard Deviation, this will have an error $\propto \frac{1}{\sqrt{N}}$ which $\rightarrow 0$ as $N \rightarrow \infty$.

Robust Determination of Mean—II

- Further, the method is unbiased for any N, i.e.,

$$\langle \mu^*_{Am} \rangle = \mu$$

- If the $x^{(i)}$ are distributed according to either the
- Gaussian

$$p(x) \propto \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

- or exponential form

$$p(x) \propto \exp\left(-\frac{x}{\mu} \right)$$

- then we know from maximum likelihood method that the arithmetic mean is the best method.
- However, for other distributions, the arithmetic mean is unlikely to be best.

Robust Determination of Mean—III

- For a trivial example, suppose I measured $y = x^2$ where x had an exponential distribution.
- Then the best estimate of the mean of y is obviously

$$y^*_{\text{clever}} = \left(\frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} \right)^{1/2} \right)^2$$

and *not*

$$y^*_{Am} = \frac{1}{N} \sum_{i=1}^{N} y^{(i)}$$

- The problem of estimating μ is hardest for distributions $p(x)$ that have long tails.
- Then with a small probability, we will generate events from the tail, but these will be way off the means,
  - and if included in calculation of mean, will distort results in a finite sample.

Robust Determination of Mean

Trimmed Mean

- One method that is more robust than the arithmetic mean, is the trimmed mean

$$\mu^*_{\text{trim}} = \frac{\sum_{i=pN+1}^{N(1-p)} x^{(i)}}{N(1-2p)}$$

where pN events are left off each end.

- For the special case of distributions symmetric about their mean, the trimmed μ^* will also have mean μ and error $\alpha \frac{1}{\sqrt{N}}$.
- The variance of μ^*_{trim} will be larger than μ^*_{Am} for Gaussianly distributed $x^{(i)}$.
- But as you investigate distributions broader than Gaussians you will find the variance of μ^*_{trim} becoming smaller than that of μ^*_{Am}.
- A special case of μ^*_{trim} is the median (i.e., middle value of $x^{(i)}$ when they are numerically ordered) gotten with $p = 0.5$.

Likelihood Based Robust Estimation of Means—I

- An interesting class of robust estimators are based on the modification of the maximum likelihood equations.
- Thus, let
$$z^{(i)} = \frac{\left(x^{(i)} - \mu\right)}{\sigma}$$
and suppose that $p(x, \mu)$ is just a function of z. (True for Gaussian case!).
- The likelihood equation for μ is
$$\frac{\partial}{\partial \mu} \ln L = 0$$
or
$$\sum_{i=1}^{N} \frac{\partial}{\partial \mu} \ln p \left(\frac{\left(x^{(i)} - \mu\right)}{\sigma} \right) = 0$$

Likelihood Based Robust Estimation of Means—II

- Now, as we are assuming p to be an even function of z, i.e., symmetric about the mean, we write the above as

$$\sum_{i=1}^{N} \Psi\left(\frac{\left(x^{(i)}-\mu\right)}{\sigma}\right)=0 \tag{37}$$

- where Ψ is an *odd* function of its argument.
- For the case of Gaussian p,
- Ψ is just $(x^{(i)}-\mu)/\sigma^2$ and Equation (37) is

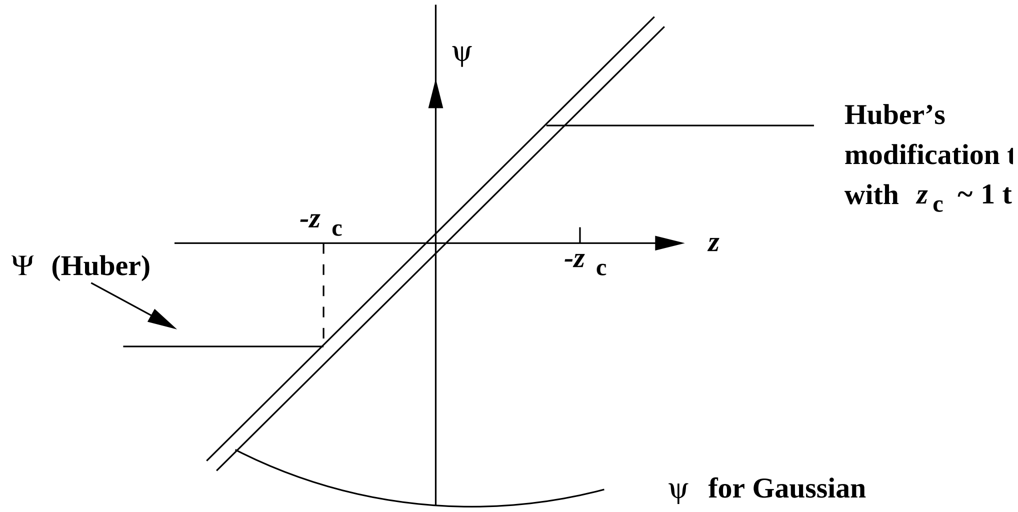
$$\frac{1}{\sigma^2}\sum_{i=1}^{N}\left(x^{(i)}-\mu_L^*\right)=0$$

- which gives the arithmetic mean as the maximum likelihood estimate of the mean.

Particular Robust Estimates for Likelihood—Huber

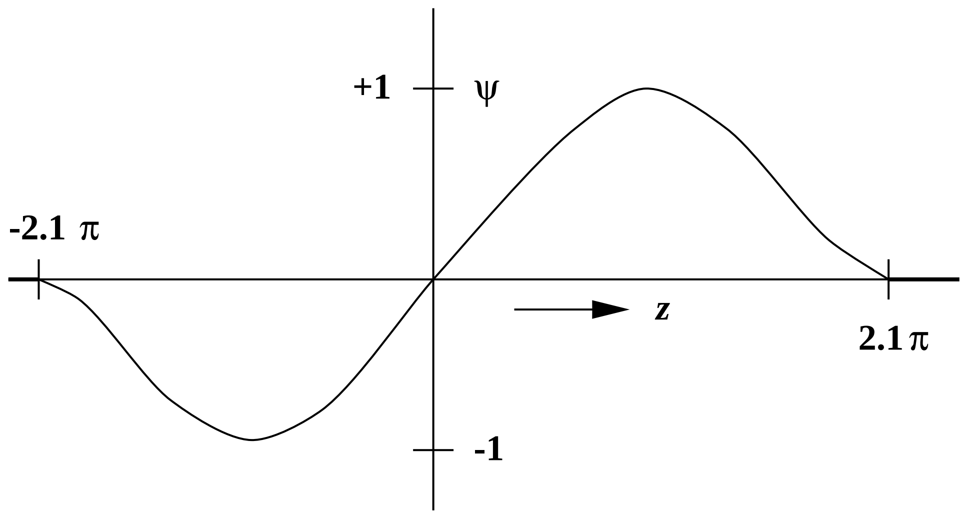
- Now we invent a function Ψ which is less sensitive to $x^{(i)}$'s far from the mean than $(x^{(i)} - \mu)$.
- In Huber's method, one sets

$$\begin{aligned} \Psi(z) &= z && \text{for } |z| \le z_c \\ \text{and } \Psi(z) &= z_c && \text{for } z > z_c \\ &= -z_c && \text{for } z > -z_c \end{aligned}$$



Particular Robust Estimates for Likelihood—Andrews

- Many forms for Ψ have been proposed.
- The most successful was that due to Andrews with



$$\begin{aligned} \Psi(z) &= \sin(z/2.1) \quad & |z| < 2.1\pi \\ &= 0 & |z| \geq 2.1\pi \end{aligned}$$

- Note z is $\frac{(x-\mu)}{\sigma}$ where σ supplies natural scale for problem.
- We do not need a very good estimate for σ but it should be robust!
- A simple choice is $\sigma = \text{mean}|x^{(i)} - \mu|$ which formula is applied iteratively as value of μ is refined.

Comparison of Methods—I

- A large number of tests are described in first reference to this section. Choose three distributions.
  - Let N = Gaussian Distribution with $\sigma = 1$.
  - C = Cauchy.
  - M = Gaussian $\div$ uniform random variable between 0 and 1.
- Here is a sample of their results for 10 observations

Method	N	C	M	$\longleftarrow$ Distribution
Mean	1	27314	5255	Relative Value of Variances
Median	1.37	3.66	7.33	
Andrews	1.08	4.66	9.29	

- The above table gives variances in units of which I am not certain; but the relative values within a column are independent of this.

Comparison of Methods—II

- We see very clearly that the arithmetic mean is hopeless for broad distributions.
- The median and Andrews methods are very good in these cases with the Andrews being better in Gaussian case (σ only 4% more than best estimate) but less robust in other examples.
- Note that all the methods *except* the arithmetic mean require all the $x^{(i)}$ being in memory so that we can order them.
- So the computer complexity is much higher for the robust estimates.
- The book of Andrews gives computer programs for all the robust estimators.

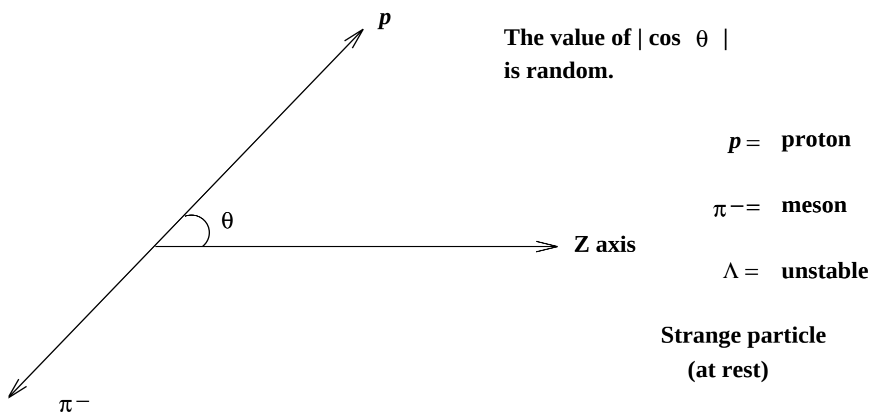
6: GENERATION OF RANDOM NUMBERS

6 Overview of Generation of Random Numbers

- Knuth is a reasonable pedagogical description although it is a little out of date.
  - Chiu (C3P-526) describes newer algorithms which are important in some cases.
- Press, et al. (Numerical Recipes) implements algorithms at about the level of Knuth.
- The National High Performance Software Exchange has a review of parallel random number generators.

Generation of Random Numbers—Genuinely Random

- One needs specialized hardware to generate true random numbers,
- e.g.,



for successive decays ($\Lambda \rightarrow p\pi^-$) of a bunch of Λ's.

- There are a lot of difficulties here.
  - Not only must you measure θ in an unbiased fashion (impossible),
  - but also to find a random number with 32 bits of precision would require several separate θ values.
  - Maybe each gives four bits of information and so eight measurements of θ are needed to get one random number.

Congruential Pseudorandom Numbers

- Rather, one usually employs a deterministic method which gives "essentially" random numbers. These are called pseudorandom.
- Most "standard" library random number generators use the so called linear congruential generators which generate a sequence of integers I_j by

$$I_{j+1} = [aI_j + c] \bmod(m)$$

- Here, m is called the modulus, a is called the multiplier, and c is called the increment.
- This is deterministic—given any seed I_0, one generates an identical sequence I_0, I_1, $I_2 \ldots$
- There are most m possible random numbers, but there may be less. One could generate a cycle if

$$I_{j_1} = I_{j_2} \qquad j_1 > j_2$$

- Then $I_{j_1+1} = I_{j_2+1}$, etc. Also, $I_{j_1-1} = I_{j_2-1}$ and in particular $I_{j_1-j_2} = I_0$.

Congruential Random Numbers in Monte Carlo

- If the cycle has a length m, then imagine our calculation of π

$$\pi = \int_{-1}^{+1} \int_{-1}^{+1} dx\, dy\, \theta\, [1 - x^2 - y^2]$$

- Then random numbers after the m'th are no longer independent and so our use of central limit theorem breaks down.
- Our error is $\propto \max\left[\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{m}}\right]$ where $N =$ number of events and m is cycle length.
- The references give various formal criteria to guide the choice of a, c, and m.
- Unfortunately, I do not see any of these references as being very good as they do not test and compare some key choices.

A Practical Congruential Pseudorandom Number Generator

- In my group, we have used in the past

$$\begin{aligned} m &= 2^{31} \\ a &= 1103515245 \\ c &= 12345 \end{aligned}$$

- with C code which is essentially
- newran = [a * oldran + c]& MASK,
- where MASK has lower 31 bits = 1. All higher bits = 0.
- We get a floating point number between 0 and 1 by:

$$\text{floatran} = \text{newran}/2147483648.0$$

- This is very simple to code, especially in C which allows you to use machine capabilities effectively.

Shift Register Pseudorandom Numbers—Overview

- The shift register methods are not discussed (correctly) in Knuth, and may well be better, as they are as easy to calculate and have longer period.
- Chiu has implemented a particularly simple method.

  $$I_j = I_{j-q} \oplus I_{j-p}$$

  where $\oplus$ = Exclusive OR

  $$\begin{aligned} 1 \oplus 1 &= 0 \\ 1 \oplus 0 &= 0 \oplus 1 = 1 \\ 0 \oplus 0 &= 0 \end{aligned}$$

  This $\oplus$ operates independently on each bit of the numbers I_{j-q}, I_{j-p}.
- The period of this could be as large as $2^p - 1$, which is big for Chiu's choice $p = 250$, $q = 147$.