

### **Hypothesis Testing, Likelihood Functions and Parameter Estimation:**

We consider **estimation** of (one or more) parameters to be the experimental determination (aka “measurement”) of those parameters (which are assumed to have **fixed**, but **apriori unknown** values), and which is based on a **limited/finite number** of experimental observations.

We have already encountered the ***sample mean***,  $\bar{x} \equiv \frac{1}{N} \sum_{i=1}^N x_i$ , as an **estimator** of  $\hat{x}$ .

Now we will be more general... But before we get into a full-scale study of **estimation**, we begin by looking at **Hypothesis Testing** using **Likelihood Ratios**.

Suppose it is known that ***either*** Hypothesis ***A*** ***or*** Hypothesis ***B*** is true. And suppose further that if ***A*** is true, then the ***random variable***  $x$  is ***apriori known*** to have a P.D.F.  $f_A(x)$ , while if ***B*** is true, then the ***random variable***  $x$  is ***apriori known*** to have a different P.D.F.  $f_B(x)$ .

Suppose that we carry out  $N$  ***independent*** measurements of a ***random variable***  $x$ :  $x_1, x_2, \dots, x_N$ :

If ***A*** is true, the probability that the results are  $x_1, x_2, \dots, x_N$  is:

$$dP_A(x_1, x_2, \dots, x_N) = f_A(x_1)dx_1 \cdot f_A(x_2)dx_2 \cdot \dots \cdot f_A(x_N)dx_N = \prod_{i=1}^N f_A(x_i)dx_i$$

On the other hand, if ***B*** had instead been true, the probability of the same string of results would have been:

$$dP_B(x_1, x_2, \dots, x_N) = f_B(x_1)dx_1 \cdot f_B(x_2)dx_2 \cdot \dots \cdot f_B(x_N)dx_N = \prod_{i=1}^N f_B(x_i)dx_i$$

The ***Likelihood Ratio***  $\mathcal{R}$  is defined as:

$$\mathcal{R} \equiv \frac{dP_A(x_1, x_2, \dots, x_N)}{dP_B(x_1, x_2, \dots, x_N)} = \frac{\prod_{i=1}^N f_A(x_i)dx_i}{\prod_{i=1}^N f_B(x_i)dx_i}$$

In other words, the ***Likelihood Ratio***  $\mathcal{R}$  is:

{The probability that the particular experimental result of  $N$  measurements turned out the way that it did, assuming ***A*** is true}  $\div$  {The probability that the particular experimental result of  $N$  measurements turned out the way that it did, assuming ***B*** is true}.

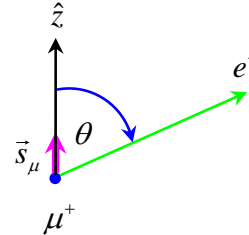
In effect, the ***Likelihood Ratio***  $\mathcal{R}$  is the “betting odds” of ***A*** against ***B***, *i.e.* we **assign** probabilities to ***A*** and ***B*** proportional to their “***Likelihoods***”:

$$\mathcal{L}_A \equiv dP_A(x_1, x_2, \dots, x_N) \equiv \prod_{i=1}^N f_A(x_i)dx_i \quad \text{and} \quad \mathcal{L}_B \equiv dP_B(x_1, x_2, \dots, x_N) \equiv \prod_{i=1}^N f_B(x_i)dx_i$$

Thus, the ***Likelihood Ratio*** is:  $\mathcal{R} \equiv \frac{\mathcal{L}_A}{\mathcal{L}_B} \equiv \frac{dP_A(x_1, x_2, \dots, x_N)}{dP_B(x_1, x_2, \dots, x_N)} = \frac{\prod_{i=1}^N f_A(x_i)dx_i}{\prod_{i=1}^N f_B(x_i)dx_i}$

*n.b.*  $\mathcal{L}_A$  and  $\mathcal{L}_B$  are **numbers** – which **will** change (slightly), *e.g.* if the **entire** experiment is repeated – hence they **are** **random variables** – but they are **not** random **distributions** of any kind.

Now in physics, it is common to have an *infinite* number (e.g. a continuum) of *hypotheses*! For example, in the weak decays of +ve muons ( $\mu^+ \rightarrow e^+ + \nu_e + \bar{\nu}_\mu$ ) whose spins  $\vec{s}_\mu$  are fully (i.e. 100%) polarized (i.e. aligned) along the  $\hat{z}$  axis {the flight direction of the  $\mu^+$  in the lab frame}, the decay positrons are emitted with a (normalized) polar angle distribution in the  $\mu^+$  center of mass frame:

$$\frac{dP(\cos \theta)}{d \cos \theta} = \frac{1}{2}(1 + \alpha \cos \theta)$$


Here,  $dP(\cos \theta) = \frac{1}{2}(1 + \alpha \cos \theta) d \cos \theta$  is the (infinitesimal) probability that the decay  $e^+$  is emitted at angle  $\theta$  whose cosine is within the (infinitesimal) range  $\cos \theta \rightarrow \cos \theta + d \cos \theta$ .

If the muons are 100% spin-polarized, then the asymmetry parameter  $\alpha$  is a number that depends on how **Parity** (space inversion symmetry, {here,  $\theta \rightarrow -\theta$ }) is *violated* (along with **Charge Conjugation**, i.e.  $\mu^+ \rightarrow \mu^-$ ) in the weak decays of the muon. The value of  $\alpha$  is physically constrained to lie between  $-1$  and  $+1$ . The Standard Model electroweak (V-A) prediction is  $\hat{\alpha} = +1.0$ . The experimentally measured *world-average* value is  $\bar{\alpha} = +1.0009^{+0.0016}_{-0.0007}$ .

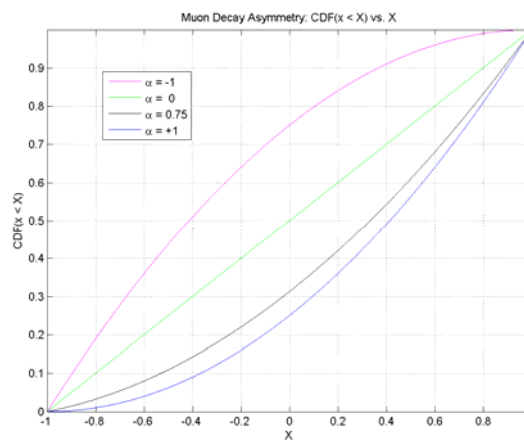
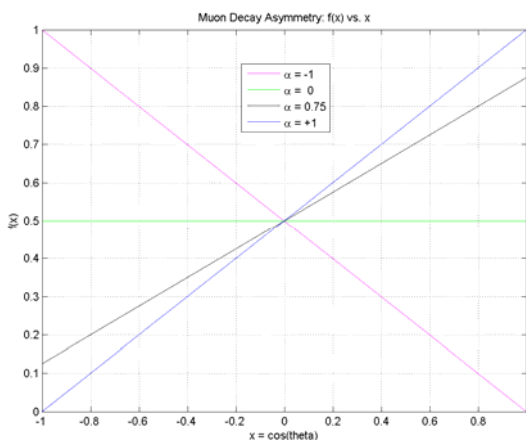
We define  $x \equiv \cos \theta$ , which here in this situation is seen as a *random variable* ranging from  $-1 \leq x = \cos \theta < +1$ , since  $0 \leq \theta < \pi$ . The Probability Density Function (PDF) for spin-polarized  $\mu^+$  decay is  $f(x, \alpha) = \frac{dP(x)}{dx} = \frac{1}{2}(1 + \alpha x) = \frac{1}{2}(1 + \alpha \cos \theta) = \frac{dP(\cos \theta)}{d \cos \theta}$ , with normalization condition:

$$\int_{-1}^{+1} f(x, \alpha) dx = \int_{-1}^{+1} \frac{1}{2}(1 + \alpha x) dx = \frac{1}{2} \left[ x + \frac{1}{2} \alpha x^2 \right]_{-1}^{+1} = \frac{1}{2} \left[ (1+1) + \frac{1}{2} \alpha (1-1) \right] = 1 \checkmark.$$

The Cumulative Distribution Function (CDF) for fully spin-polarized  $\mu^+$  decays is:

$$CDF(x < X, \alpha) = \int_{-1}^X f(x, \alpha) dx = \frac{1}{2} \left[ (X+1) + \frac{1}{2} \alpha (X^2 - 1) \right].$$

Plots of the PDF and CDF for spin-polarized  $\mu^+$  decay are shown in the two figures below, for four different physically allowed values of the asymmetry parameter,  $\alpha$ :



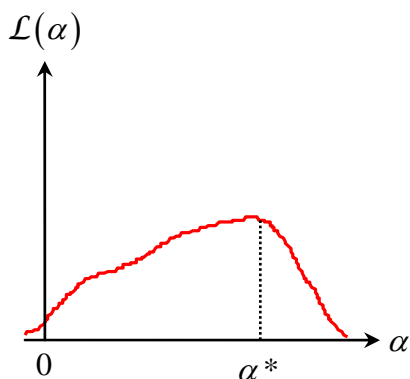
An experimentalist wants to measure the  $\alpha$  parameter, *i.e.* so as to be able to choose among **all** possible **hypotheses** – here, a **continuum**:  $\alpha = -1.00000$ ,  $\alpha = -0.99999$ , ...,  $\alpha = -0.00001$ ,  $\alpha = 0.00000$ ,  $\alpha = 0.00001$ , ...,  $\alpha = 0.99999$  ...,  $\alpha = 1.00000$ .

$N$  muon decays are measured and yield the (independent random variable) set of angles  $\theta_i$ .

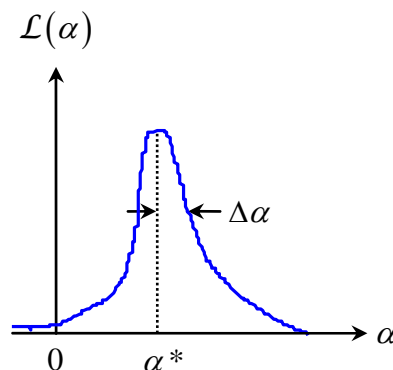
If we define  $x_i \equiv \cos \theta_i$ , then the P.D.F. becomes  $f(x, \alpha) = \frac{1}{2}(1 + \alpha x)$  and the **Likelihood** is:

$$\mathcal{L}_A \Rightarrow \mathcal{L}(\alpha) = \prod_{i=1}^N f(x_i, \alpha).$$

As before, we will think in terms of the **Likelihood Ratio**  $\mathcal{R} = \mathcal{L}(\alpha_1)/\mathcal{L}(\alpha_2)$  and assign a **Likelihood** { = “probability that  $\alpha_m$  is true” } of  $\mathcal{L}(\alpha_m)$ . Next, we plot  $\mathcal{L}(\alpha)$  vs.  $\alpha$ :



-OR-



The **Most Probable Value** of  $\alpha$ ,  $\alpha^*$  is called **the Maximum Likelihood Solution**.

The **Root Mean Square (RMS) Spread** – the square root of the **variance** of  $\alpha$  around  $\alpha^*$  is a measure of the accuracy with which  $\alpha^*$  is determined, call it  $\Delta\alpha$  ( $\equiv \sqrt{\sigma_\alpha^2}$ ).

If  $N$  is **large**, then  $\mathcal{L}(\alpha)$  will be a Gaussian (due to/because of the **Central Limit Theorem**).

But if  $N$  is **small**, then we **may** have a situation like the one depicted on the LHS of the above figure. In that case,  $\Delta\alpha$  has no real meaning and should **not** be quoted. Instead, the plot should be shown.

Now let us be very careful... If we interpret  $\mathcal{L}(\alpha)$  as a measure of the “probability of  $\alpha$ ”, then we **must** make certain that it is ***properly normalized***.

Thus, we calculate  $\int \mathcal{L}(\alpha) d\alpha$  and replace  $\mathcal{L}(\alpha)$  by  $\frac{\mathcal{L}(\alpha)}{\int \mathcal{L}(\alpha) d\alpha}$ . Then we **know** that:

$$(\Delta\alpha)^2 \equiv \sigma_\alpha^2 \equiv E[(\alpha - \alpha^*)^2] = \frac{\int (\alpha - \alpha^*)^2 \mathcal{L}(\alpha) d\alpha}{\int \mathcal{L}(\alpha) d\alpha} \quad \text{or:} \quad \Delta\alpha = \sigma_\alpha = \sqrt{\frac{\int (\alpha - \alpha^*)^2 \mathcal{L}(\alpha) d\alpha}{\int \mathcal{L}(\alpha) d\alpha}}$$

This approach to determining parameters (such as  $\alpha$ ) and their uncertainties  $\sigma_\alpha$  is called the ***Maximum Likelihood Method*** (M.L.M.)

### **A Detailed Example of the Use of the M.L.M. :**

Suppose we are trying to “directly” measure a physical parameter  $\alpha_0$ . Let each measurement be called  $x_i$ , and let  $\sigma_i$  be the standard deviation associated with each individual measurement  $x_i$ . Let us further assume that the  $N$  independent individual measurements ( $x_i$ ) are Gaussian-distributed, with  $\hat{x} \equiv \alpha_0$  as their expectation value.

Then for any individual measurement:  $f(x_i, \alpha_0) = \frac{1}{\sqrt{2\pi} \sigma_i} e^{-(x_i - \alpha_0)^2 / 2\sigma_i^2}$

Thus, for  $N$  independent measurements:  $\mathcal{L}(\alpha) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi} \sigma_i} e^{-(x_i - \alpha)^2 / 2\sigma_i^2}$

**n.b.** Here, we have used the **parameter**  $\alpha$  instead of  $\alpha_0$ , since we are trying to ***find/determine***  $\alpha_0$  (which is ***apriori unknown***). It is by **varying**  $\alpha$  (as a “free” parameter) that we find / determine / “measure” the ***particular*** value,  $\alpha^*$  (which we identify with  $\alpha_0$ ) that ***maximizes*** the ***likelihood function***  $\mathcal{L}(\alpha)$ .

Note further that we also use the “shorthand”  $\mathcal{L}(\alpha)$  instead of the more correct  $\mathcal{L}(\alpha; x_1, \dots, x_N)$ .

We will carry out the maximization of the ***Likelihood*** function explicitly, *i.e.* we will find  $\partial \mathcal{L}(\alpha) / \partial \alpha$  and then look for a ***zero*** corresponding to a ***maximum*** in  $\mathcal{L}(\alpha)$ .

Note also that in practice, we ***maximize***  $\ln \mathcal{L}(\alpha)$ , the “***log likelihood***” instead of  $\mathcal{L}(\alpha)$ . Define:

$$\begin{aligned} \ell(\alpha) \equiv \ln \mathcal{L}(\alpha) &= \sum_{i=1}^N \ln \left\{ \frac{1}{\sqrt{2\pi} \sigma_i} e^{-(x_i - \alpha)^2 / 2\sigma_i^2} \right\} \\ &= \sum_{i=1}^N \ln \left\{ \frac{1}{\sqrt{2\pi} \sigma_i} \right\} - \sum_{i=1}^N \left\{ \frac{(x_i - \alpha)^2}{2\sigma_i^2} \right\} \end{aligned}$$

Note that if  $\mathcal{L}(\alpha)$  has a maximum at some  $\alpha = \alpha^*$ , then  $\ell(\alpha) \equiv \ln \mathcal{L}(\alpha)$  will *also* have a maximum at the *same* value of  $\alpha$ . Thus:

$$\ell(\alpha) \equiv \ln \mathcal{L}(\alpha) = -\frac{1}{2} \sum_{i=1}^N \frac{(x_i - \alpha)^2}{\sigma_i^2} + \text{constant}$$

Then: 
$$\ell'(\alpha) \equiv \frac{d\ell(\alpha)}{d\alpha} = + \sum_{i=1}^N \frac{x_i - \alpha}{\sigma_i^2}$$

And: 
$$\ell''(\alpha) \equiv \frac{d^2\ell(\alpha)}{d\alpha^2} = - \sum_{i=1}^N \frac{1}{\sigma_i^2}$$
 Since this is  $< 0$ , the extremum is a *maximum*

Finally, set:  $\ell'(\alpha) \equiv \frac{d\ell(\alpha)}{d\alpha} \Big|_{\alpha^*} = 0$  and solve:

$$\sum_{i=1}^N \frac{x_i - \alpha^*}{\sigma_i^2} = 0 \Rightarrow \sum_{i=1}^N \frac{x_i}{\sigma_i^2} - \sum_{i=1}^N \frac{\alpha^*}{\sigma_i^2} = 0 \Rightarrow \sum_{i=1}^N \frac{x_i}{\sigma_i^2} = \sum_{i=1}^N \frac{\alpha^*}{\sigma_i^2} = \alpha^* \sum_{i=1}^N \frac{1}{\sigma_i^2} \Rightarrow \alpha^* = \frac{\sum_{i=1}^N \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^N \frac{1}{\sigma_i^2}}$$

Thus we see that the **Maximum Likelihood** value of the parameter  $\alpha^*$  is just a **Weighted Mean**.

Note that if *all* of the *individual* measurements  $x_i$  had the *same* variances  $\sigma_i^2 = \sigma^2$ , the above expression would reduce to the *simple/arithmetic/unweighted sample mean*:  $\alpha^* = \frac{1}{N} \sum_{i=1}^N x_i$

So we recover a result that we should have anticipated, and see how to use M.L.M. in the simplest case. But has this elegant procedure given us anything new? In order to understand this, we now “step back” and look in general at estimation.

For simplicity, let us consider “experiments” where we perform a single measurement  $x_k$ , in each, and where we are trying to “determine” (i.e. estimate) a *common* single parameter  $\lambda$ .

Later, when we look at practical schemes, we will discuss the case(s) of several measurements per experiment, and also the simultaneous determination of several parameters.

After we have made  $N$  independent measurements of a random variable  $x$ :  $x_1, \dots, x_N$ , we construct a **function**  $S(x_1, \dots, x_N)$  whose **numerical value** is the *estimate* of the *apriori unknown* parameter of interest  $\lambda$  (e.g.  $\hat{x}, \sigma_x^2, \dots$ ). Thus the *estimator*  $S$  cannot depend on  $\lambda$ . The **numerical value** of the **function**  $S$  (“the estimator”) is itself a random variable.

For example, if we wanted to *estimate* the *expectation value* of  $x$ ,  $E[x] = \hat{x}$  for a set of Gaussian-distributed measurements (whose P.D.F. is  $\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\hat{x})^2/2\sigma^2}$ , all with the same

standard deviation  $\sigma$ ) we *could* use e.g. the *sample mean*:  $S(x_1, \dots, x_N) = \bar{x} \equiv \frac{1}{N} \sum_{i=1}^N x_i$

This estimator is “**unbiased**”, since  $E[S] = \hat{x}$ . This is nice!

In general, the *estimator*  $S$  that we choose may in fact be “**biased**”.

That is,  $E[S]$  may in fact not be  $\equiv \lambda$ , the “**true**” value of the parameter.

We define the **Bias**  $B(\lambda)$  as the **difference** between  $E[S]$  and  $\lambda$ :  $B(\lambda) \equiv E[S] - \lambda$

A “**good estimator**”  $S$  will be unbiased, i.e. have  $B(\lambda) \equiv E[S] - \lambda = 0$ .

Another important property of the *estimator*  $S$  is its own **variance**,  $\sigma_S^2$ . It is obviously highly desirable to invent *estimators* with  $\sigma_S^2$  as small as possible. Can  $\sigma_S^2$  be arbitrarily small? (ans: No!)

It is not obvious (and in general not true!) that one can find an *estimator*  $S$  with both **minimum bias** and **minimum variance**.

For now, we will deal only with unbiased estimators, but later on, we will look at the biased kind.

For now, we focus on/concern ourselves with possible bias issues associated with variance:

Let  $f(x; \lambda)$  be the P.D.F. associated with the measurement of  $x$ .  $\lambda$  is the parameter we wish to determine by carrying out a series of **independent** experimental measurements  $x_1, \dots, x_N$ .

The joint P.D.F. of this series is  $f(x_1, x_2, \dots, x_N; \lambda) = f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda)$  provided that the  $N$  individual measurements of  $x_i$  are **independent**.

Let  $S(x_1, \dots, x_N)$  be an unbiased estimator of  $\lambda$ , i.e.  $E[S] = \lambda$ .

Then:  $\lambda = E[S] = \int S(x_1, x_2, \dots, x_N) f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) dx_1 dx_2 \dots dx_N$

We now go through a series of manipulations in order to arrive at a result concerning the minimum variance associated with the *estimator*  $S$ .

Along the way, we will also define a quantity known as “**information**”.

n.b.

In all that follows, we assume **all** integrals are defined, integration and differentiation commute, etc...

So:  $\lambda = E[S] = \int S(x_1, x_2, \dots, x_N) f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) dx_1 dx_2 \dots dx_N$

Differentiate both sides of the above relation with respect to the parameter  $\lambda$ :

$$1 = \int S(x_1, x_2, \dots, x_N) \frac{d}{d\lambda} \{f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda)\} dx_1 dx_2 \dots dx_N$$

Then let us

Define:  $\mathcal{L}(x_1, x_2, \dots, x_N; \lambda) \equiv \prod_{i=1}^N f(x_i; \lambda) = f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda)$

Define:  $\ell(x_1, x_2, \dots, x_N; \lambda) \equiv \ln \mathcal{L}(x_1, x_2, \dots, x_N; \lambda) = \sum_{i=1}^N \ln f(x_i; \lambda)$

Define:  $\ell'(x_1, x_2, \dots, x_N; \lambda) \equiv \frac{d\ell(x_1, x_2, \dots, x_N; \lambda)}{d\lambda} \equiv \frac{d}{d\lambda} \ln \mathcal{L}(x_1, x_2, \dots, x_N; \lambda) = \frac{d}{d\lambda} \left\{ \sum_{i=1}^N \ln f(x_i; \lambda) \right\}$

Define:

$$\begin{aligned} \ell''(x_1, x_2, \dots, x_N; \lambda) &\equiv \frac{d\ell'(x_1, x_2, \dots, x_N; \lambda)}{d\lambda} \equiv \frac{d^2 \ell(x_1, x_2, \dots, x_N; \lambda)}{d\lambda^2} \equiv \frac{d^2}{d\lambda^2} \ln \mathcal{L}(x_1, x_2, \dots, x_N; \lambda) \\ &= \frac{d^2}{d\lambda^2} \left\{ \sum_{i=1}^N \ln f(x_i; \lambda) \right\} \end{aligned}$$

Define:  $f'(x_i; \lambda) \equiv \frac{df(x_i; \lambda)}{d\lambda}$  and:  $f''(x_i; \lambda) \equiv \frac{df'(x_i; \lambda)}{d\lambda} \equiv \frac{d^2 f(x_i; \lambda)}{d\lambda^2}$

Then:

$$\begin{aligned} \ell'(x_1, x_2, \dots, x_N; \lambda) &\equiv \frac{d\ell(x_1, x_2, \dots, x_N; \lambda)}{d\lambda} \equiv \frac{d}{d\lambda} \ln \mathcal{L}(x_1, x_2, \dots, x_N; \lambda) = \frac{d}{d\lambda} \left\{ \sum_{i=1}^N \ln f(x_i; \lambda) \right\} \\ &= \sum_{i=1}^N \frac{d}{d\lambda} \ln f(x_i; \lambda) = \sum_{i=1}^N \frac{df(x_i; \lambda)/d\lambda}{f(x_i; \lambda)} = \sum_{i=1}^N \frac{f'(x_i; \lambda)}{f(x_i; \lambda)} \equiv \sum_{i=1}^N \phi(x_i; \lambda) \end{aligned}$$

Next:

$$\begin{aligned} &\frac{d}{d\lambda} \{f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda)\} \\ &= \left\{ \frac{f'(x_1; \lambda)}{f(x_1; \lambda)} + \frac{f'(x_2; \lambda)}{f(x_2; \lambda)} + \dots + \frac{f'(x_N; \lambda)}{f(x_N; \lambda)} \right\} f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) \\ &= \underbrace{\left\{ \sum_{i=1}^N \frac{f'(x_i; \lambda)}{f(x_i; \lambda)} \right\}}_{\equiv \ell'(x_1, x_2, \dots, x_N; \lambda)} f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) \\ &= \ell'(x_1, x_2, \dots, x_N; \lambda) f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) \end{aligned}$$

Plug this result into:  $1 = \int S(x_1, x_2, \dots, x_N) \frac{d}{d\lambda} \{f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda)\} dx_1 dx_2 \dots dx_N$  to get:

$$1 = \int S(x_1, x_2, \dots, x_N) \ell'(x_1, x_2, \dots, x_N; \lambda) f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) dx_1 dx_2 \dots dx_N = E[S\ell'] \quad \text{i.e. } E[S\ell'] = 1$$

Now:  $E[\ell'] = 0$ , which follows from taking  $\frac{d}{d\lambda}$  of both sides of the P.D.F. normalization condition:

$$\int f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) dx_1 dx_2 \dots dx_N = 1$$

Proof:

$$\begin{aligned}
 \frac{d}{d\lambda} \int f(x_1; \lambda) \cdot f(x_2; \lambda) \cdots f(x_N; \lambda) dx_1 dx_2 \cdots dx_N &= 0 \\
 &= \int \underbrace{\left\{ \frac{f'(x_1; \lambda)}{f(x_1; \lambda)} + \frac{f'(x_2; \lambda)}{f(x_2; \lambda)} + \cdots + \frac{f'(x_N; \lambda)}{f(x_N; \lambda)} \right\}}_{\equiv \ell'(x_1, x_2, \dots, x_N; \lambda)} f(x_1; \lambda) \cdot f(x_2; \lambda) \cdots f(x_N; \lambda) dx_1 dx_2 \cdots dx_N = 0 \\
 &= \underbrace{\int \ell'(x_1, x_2, \dots, x_N; \lambda) f(x_1; \lambda) \cdot f(x_2; \lambda) \cdots f(x_N; \lambda) dx_1 dx_2 \cdots dx_N}_{\equiv E[\ell']} = E[\ell'] = 0 \quad Q.E.D.
 \end{aligned}$$

Then:  $E[\ell'] \cdot E[S] = 0$  (assuming  $E[S] = \lambda$  is finite)

Since:  $E[S\ell'] = 1$  Then:  $E[S\ell'] - E[S] \cdot E[\ell'] = 1$

But:  $E[S\ell'] - E[S] \cdot E[\ell'] \equiv \text{cov}(S, \ell') \Rightarrow \text{cov}(S, \ell') = 1 \Rightarrow S$  and  $\ell'$  are positively correlated!

Recall that the **correlation coefficient**  $\rho(x, y) \equiv \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$  has magnitude  $|\rho(x, y)| = \left| \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \right| \leq 1$

Thus, here:  $\rho^2(S, \ell') = \frac{\overbrace{\text{cov}(S, \ell')^2}^{\equiv 1}}{\sigma_S^2 \sigma_{\ell'}^2} = \frac{1}{\sigma_S^2 \sigma_{\ell'}^2} \leq 1 \Rightarrow$  **variance** of the **estimator**  $S$ :  $\sigma_S^2 \geq \frac{1}{\sigma_{\ell'}^2}$

However, the **variance** of  $\ell' \equiv \frac{d}{d\lambda} \ln \mathcal{L}(\lambda)$  is defined as:

$$\sigma_{\ell'}^2 \equiv E[(\ell' - E[\ell'])^2] = E[\ell'^2] - \underbrace{(E[\ell'])^2}_{=0} = E[\ell'^2]$$

Thus:  $\sigma_S^2 \geq \frac{1}{\sigma_{\ell'}^2} = \frac{1}{E[\ell'^2]} = \frac{1}{E\left[\left\{\frac{d}{d\lambda} \ln \mathcal{L}(\lambda)\right\}^2\right]}$

We define the so-called “**information**”  $I(\lambda)$  of the sample (with respect to the parameter  $\lambda$ ) as:

$$I(\lambda) \equiv E[\ell'^2] = E\left[\left\{\frac{d}{d\lambda} \ln \mathcal{L}(\lambda)\right\}^2\right] = \sigma_{\ell'}^2 \quad \text{Then: } \sigma_S^2 \geq \frac{1}{I(\lambda)}.$$

Thus we see that **large information**  $I(\lambda) \Rightarrow$  **small variance**  $\sigma_S^2$  of the **estimator**  $S$  and vice versa.



Had the *estimator*  $S$  been ***biased***, then with  $B(\lambda) \equiv E[S] - \lambda \neq 0$  we would have instead arrived at the ***general form*** of this inequality:

$$\sigma_S^2 \geq \frac{\left\{1 + \frac{dB(\lambda)}{d\lambda}\right\}^2}{I(\lambda)} = \frac{\left\{1 + \frac{dB(\lambda)}{d\lambda}\right\}^2}{E\left[\left\{\frac{d}{d\lambda} \ln \mathcal{L}(\lambda)\right\}^2\right]}$$

which is known as the ***Rao-Cramér-Frechet (RCF) Inequality***, aka the “***Information Inequality***”. It gives a rigorous ***lower bound*** on the ***variance***  $\sigma_S^2$  associated with a ***biased estimator***  $S$  of the parameter  $\lambda$ .

We’re not quite done here... an interesting relationship exists between  $E[\ell'^2]$  and  $E[\ell'']$  or equivalently, exists between  $E\left[\left\{\frac{d}{d\lambda} \ln \mathcal{L}(\lambda)\right\}^2\right]$  and  $E\left[\frac{d^2}{d\lambda^2} \ln \mathcal{L}(\lambda)\right]$ , namely that:

$$E[\ell'^2] = -E[\ell''], \text{ or equivalently, that: } E\left[\left\{\frac{d}{d\lambda} \ln \mathcal{L}(\lambda)\right\}^2\right] = -E\left[\frac{d^2}{d\lambda^2} \ln \mathcal{L}(\lambda)\right].$$

We showed above that the ***expectation value*** of  $\ell'(x_1, x_2, \dots, x_N; \lambda)$  was zero, i.e.  $E[\ell'] = 0$ .

What is the ***expectation value*** of  $\ell''(x_1, x_2, \dots, x_N; \lambda)$ ? i.e. what is:

$$E[\ell''] = E\left[\frac{d\ell'}{d\lambda}\right] = E\left[\frac{d^2\ell}{d\lambda^2}\right] = E\left[\frac{d^2}{d\lambda^2} \ln \mathcal{L}(\lambda)\right] = ?$$

Repeating what we did above in the determination of  $E[\ell']$ , this time we take  $\frac{d^2}{d\lambda^2}$  of both sides of the P.D.F. normalization condition:

$$\begin{aligned} \int f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) dx_1 dx_2 \dots dx_N &= 1 \\ \frac{d^2}{d\lambda^2} \int f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) dx_1 dx_2 \dots dx_N &= 0 \\ = \frac{d}{d\lambda} \int \underbrace{\left\{ \frac{f'(x_1; \lambda)}{f(x_1; \lambda)} + \frac{f'(x_2; \lambda)}{f(x_2; \lambda)} + \dots + \frac{f'(x_N; \lambda)}{f(x_N; \lambda)} \right\}}_{\equiv \ell'(x_1, x_2, \dots, x_N; \lambda)} f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) dx_1 dx_2 \dots dx_N &= 0 \\ = \frac{d}{d\lambda} \int \ell'(x_1, x_2, \dots, x_N; \lambda) f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) dx_1 dx_2 \dots dx_N &= 0 \end{aligned}$$

$$\begin{aligned}
&= \int \frac{d}{d\lambda} \{ \ell'(x_1, x_2, \dots, x_N; \lambda) f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) \} dx_1 dx_2 \dots dx_N = 0 \\
&= \int \ell''(x_1, x_2, \dots, x_N; \lambda) f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) dx_1 dx_2 \dots dx_N \\
&+ \int \ell'(x_1, x_2, \dots, x_N; \lambda) \frac{d}{d\lambda} \{ f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) \} dx_1 dx_2 \dots dx_N = 0 \\
&= \int \ell''(x_1, x_2, \dots, x_N; \lambda) f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) dx_1 dx_2 \dots dx_N \\
&+ \int \ell'(x_1, x_2, \dots, x_N; \lambda) \underbrace{\left\{ \frac{f'(x_1; \lambda)}{f(x_1; \lambda)} + \frac{f'(x_2; \lambda)}{f(x_2; \lambda)} + \dots + \frac{f'(x_N; \lambda)}{f(x_N; \lambda)} \right\}}_{\equiv \ell'(x_1, x_2, \dots, x_N; \lambda)} \\
&\quad \times f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) dx_1 dx_2 \dots dx_N = 0 \\
&= \underbrace{\int \ell''(x_1, x_2, \dots, x_N; \lambda) f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) dx_1 dx_2 \dots dx_N}_{\equiv E[\ell'']} \\
&+ \underbrace{\int \ell'^2(x_1, x_2, \dots, x_N; \lambda) f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \dots \cdot f(x_N; \lambda) dx_1 dx_2 \dots dx_N}_{\equiv E[\ell'^2]} = 0 \\
&= E[\ell''] + E[\ell'^2] = 0 \quad \text{or:} \quad E[\ell'^2] = -E[\ell''] \quad Q.E.D.
\end{aligned}$$

Thus, we see that:

$$\begin{aligned}
E[\ell''] &= E\left[\frac{d\ell'}{d\lambda}\right] = E\left[\frac{d^2\ell}{d\lambda^2}\right] = E\left[\frac{d^2}{d\lambda^2} \ln \mathcal{L}(\lambda)\right] \\
&= -E[\ell'^2] = -E\left[\left(\frac{d\ell}{d\lambda}\right)^2\right] = -E\left[\left(\frac{d}{d\lambda} \ln \mathcal{L}(\lambda)\right)^2\right]
\end{aligned}$$

The relation between these two expectation values,  $E[\ell'^2] = -E[\ell'']$  is **quite** an amazing, and very general result – since it was derived without reference to a specific form of P.D.F. – it is therefore valid/holds for any P.D.F!

The relation  $E[\ell'^2] = -E[\ell'']$  says that the negative of the **expectation value** of the (negative!) curvature – the 2<sup>nd</sup> derivative – of the log of the **likelihood function** (which is an  $N$ -dimensional integral convolving  $\ell''(x_1, x_2, \dots, x_N; \lambda)$  with the product factor of  $N$  P.D.F.'s, integrated over all of the  $x_i$ 's, over their entire allowed physical ranges) is equal to the **expectation value** of the square of the slope – the 1<sup>st</sup> derivative – of the log of the **likelihood function** (which is another  $N$ -dimensional integral convolving  $\ell'^2(x_1, x_2, \dots, x_N; \lambda)$  with the product factor of  $N$  P.D.F.'s, integrated over all of the  $x_i$ 's, over their entire allowed physical ranges).

Restating this somewhat less rigorously, but in a more physical sense:  $E[\ell'^2] = -E[\ell'']$  tells us physically that the **negative** of the integrated-over “average” value of the (negative!) **curvature** of the log of the **likelihood function** is equal to the integrated-over “average” value of the **square** of the **slope** of the log of the **likelihood function**! Again,  $E[\ell'^2] = -E[\ell'']$  is a **very general** result – it is valid/holds for **any** P.D.F! One can also physically understand now why  $E[\ell'] = 0$ .

The above **Information Inequalities** can therefore additionally be expressed in terms of  $E[\ell'']$ , the **Information Inequality** for **unbiased estimators**,  $S$  is:

$$\sigma_s^2 \geq \frac{1}{\sigma_{\ell'}^2} = \frac{1}{E[\ell'^2]} = -\frac{1}{E[\ell'']} = \frac{1}{E\left[\left\{\frac{d}{d\lambda} \ln \mathcal{L}(\lambda)\right\}^2\right]} = -\frac{1}{E\left[\frac{d^2}{d\lambda^2} \ln \mathcal{L}(\lambda)\right]} = \frac{1}{I(\lambda)}$$

The **Rao-Cramér-Frechet (RCF) Inequality** (aka **Information Inequality**) for **biased estimators**,  $S$  with  $B(\lambda) \equiv E[S] - \lambda \neq 0$  is:

$$\begin{aligned} \sigma_s^2 &\geq \frac{\left\{1 + \frac{dB(\lambda)}{d\lambda}\right\}^2}{\sigma_{\ell'}^2} = \frac{\left\{1 + \frac{dB(\lambda)}{d\lambda}\right\}^2}{E[\ell'^2]} = -\frac{\left\{1 + \frac{dB(\lambda)}{d\lambda}\right\}^2}{E[\ell'']} \\ &= \frac{\left\{1 + \frac{dB(\lambda)}{d\lambda}\right\}^2}{E\left[\left\{\frac{d}{d\lambda} \ln \mathcal{L}(\lambda)\right\}^2\right]} = -\frac{\left\{1 + \frac{dB(\lambda)}{d\lambda}\right\}^2}{E\left[\frac{d^2}{d\lambda^2} \ln \mathcal{L}(\lambda)\right]} = \frac{\left\{1 + \frac{dB(\lambda)}{d\lambda}\right\}^2}{I(\lambda)} \end{aligned}$$

These **Information Inequalities** give rigorous **lower bounds** on the **variance**  $\sigma_s^2$  associated with the **estimator**  $S$  of the parameter  $\lambda$ .

The physical meaning of the relation  $E[\ell'^2(\lambda)] = -E[\ell''(\lambda)]$  also enables us to understand/realize some amazing physical/mathematical properties of the (log of the) likelihood function  $\ln \mathcal{L}(\lambda)$ , because physically,  $\ell'(\lambda) = \frac{d}{d\lambda} \ln \mathcal{L}(\lambda)$  is the local **slope** (i.e. 1<sup>st</sup> derivative) of the  $\ln \mathcal{L}(\lambda)$  vs.  $\lambda$  curve at the point  $\lambda$ , and  $\ell''(\lambda) = \frac{d^2}{d\lambda^2} \ln \mathcal{L}(\lambda)$  is the local (negative!) **curvature** (i.e. 2<sup>nd</sup> derivative) of the  $\ln \mathcal{L}(\lambda)$  vs.  $\lambda$  curve at the point  $\lambda$ .

In the limit of **very large**  $N$  (i.e.  $N \rightarrow \infty$ ) the 1-parameter likelihood function  $\mathcal{L}(\lambda)$  is Gaussian/normal in that parameter:  $\mathcal{L}(\lambda) = C e^{-(\lambda - \lambda^*)^2 / 2\sigma_{\lambda^*}^2}$ . Then:  $\ell(\lambda) = \ln \mathcal{L}(\lambda) = \ln C - (\lambda - \lambda^*)^2 / 2\sigma_{\lambda^*}^2$ .

At the maximum of the likelihood:  $\lambda = \lambda^*$  then:  $\ell(\lambda^*) = \ln \mathcal{L}(\lambda^*) = \ln C$ , thus:  $C = \mathcal{L}(\lambda^*)$ , thus we can write:

$$\mathcal{L}(\lambda) = \mathcal{L}(\lambda^*) e^{-(\lambda - \lambda^*)^2 / 2\sigma_{\lambda^*}^2} \quad \text{or:} \quad \ell(\lambda) = \ell(\lambda^*) - (\lambda - \lambda^*)^2 / 2\sigma_{\lambda^*}^2$$

Then for:  $\lambda_{\pm 1\sigma}^* \equiv \lambda^* \pm 1\sigma_{\lambda^*}$  or:  $(\lambda_{\pm 1\sigma}^* - \lambda^*) = \pm 1\sigma_{\lambda^*}$  then:  $\mathcal{L}(\lambda_{\pm 1\sigma}^*) = \mathcal{L}(\lambda^*) e^{-1/2}$

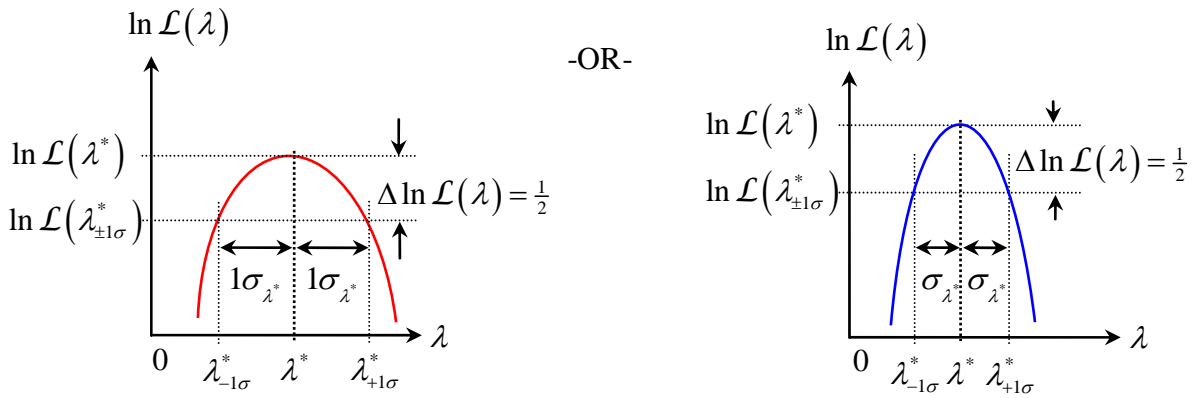
and:  $\ell(\lambda_{\pm 1\sigma}^*) = \ell(\lambda^*) - \frac{1}{2}$  or:  $\ln \mathcal{L}(\lambda_{\pm 1\sigma}^*) = \ln \mathcal{L}(\lambda^*) - \frac{1}{2}$  or:  $\Delta \ln \mathcal{L}(\lambda) \equiv \ln \mathcal{L}(\lambda^*) - \ln \mathcal{L}(\lambda_{\pm 1\sigma}^*) = \frac{1}{2}$

Then, we have the “conventional wisdom” that 68.3% of the time the “true value”  $\hat{\lambda}$  will be within  $\pm 1\sigma_{\lambda^*}$  of  $\lambda^*$ , etc.

In the **very large**  $N$  limit, the  $\ln \mathcal{L}(\lambda)$  vs.  $\lambda$  curve is an (inverted) parabola of the general form:

$$\ln \mathcal{L}(\lambda) = -\left[(\lambda - \lambda^*)^2 / 2\sigma_{\lambda^*}^2\right] - \text{constant} \quad \Leftrightarrow \quad y(x) = -A(x - x_0)^2 - B$$

This relation is shown graphically in the figure below – the RHS (LHS) plot has a wide (narrow) parabola, respectively:



When  $\ell''(\lambda) = \frac{d^2}{d\lambda^2} \ln \mathcal{L}(\lambda) \Big|_{\lambda=\lambda^*}$  is evaluated at the **local maximum**  $\lambda = \lambda^*$  of the  $\ln \mathcal{L}(\lambda)$  function:

$$\ell''(\lambda) \Big|_{\lambda=\lambda^*} = \frac{d^2}{d\lambda^2} \ln \mathcal{L}(\lambda) \Big|_{\lambda=\lambda^*} = -\frac{1}{\sigma_{\lambda^*}^2} \quad \leftarrow \quad \text{In the vicinity of the local maximum } \lambda = \lambda^*, \text{ the curvature is } \underline{\text{negative}}.$$

When  $\ell'(\lambda) = \frac{d}{d\lambda} \ln \mathcal{L}(\lambda)$  is evaluated at the  $(\pm 1\sigma_{\lambda})$   $\lambda = \lambda_{\pm 1\sigma}^*$  points of the  $\ln \mathcal{L}(\lambda)$  function on either side of the **local maximum**  $\lambda = \lambda^*$ :

$$\ell'(\lambda) \Big|_{\lambda=\lambda_{\pm 1\sigma}^*} = \frac{d}{d\lambda} \ln \mathcal{L}(\lambda) \Big|_{\lambda=\lambda_{\pm 1\sigma}^*} = -\frac{(\lambda - \lambda^*)}{\sigma_{\lambda^*}^2} \Big|_{\lambda=\lambda_{\pm 1\sigma}^*} = \pm \frac{\sigma_{\lambda^*}}{\sigma_{\lambda^*}^2} = \pm \frac{1}{\sigma_{\lambda^*}}.$$

$$\text{Thus: } \ell'^2(\lambda) \Big|_{\lambda=\lambda_{\pm 1\sigma}^*} = \left( \frac{d}{d\lambda} \ln \mathcal{L}(\lambda) \right)^2 \Big|_{\lambda=\lambda_{\pm 1\sigma}^*} = \left( -\frac{(\lambda - \lambda^*)}{\sigma_{\lambda}^2} \right)^2 \Big|_{\lambda=\lambda_{\pm 1\sigma}^*} = \left( \pm \frac{\sigma_{\lambda^*}}{\sigma_{\lambda^*}^2} \right)^2 = \frac{1}{\sigma_{\lambda^*}^2}.$$

In the limit of **very large**  $N$  (*i.e.*  $N \rightarrow \infty$ ), if we find/determine the  $\lambda_{\pm 1\sigma}^*$  points associated with reducing the value of  $\ln \mathcal{L}(\lambda)$  function by a factor of 1/2 from the value of the  $\ln \mathcal{L}(\lambda)$  function at its **local maximum**,  $\lambda = \lambda^*$ , this corresponds to the  $\pm 1 \sigma_{\lambda^*}$ , 68.3% **central/double-sided confidence level** – *i.e.*  $\lambda_{-1\sigma}^*$  and  $\lambda_{+1\sigma}^*$  are respectively the  $-1 \sigma_{\lambda^*}$  (low) and  $+1 \sigma_{\lambda^*}$  (high) points on either side of the **local maximum**,  $\lambda = \lambda^*$ .

In the limit of **very large**  $N$  (*i.e.*  $N \rightarrow \infty$ ), we have **four** equivalent methods of determining  $\sigma_{\lambda}$ :

- 1.) Compute the **Root Mean Square Deviation** – the square root of the **variance**  $\sigma_{\lambda}^2$  of the **Likelihood Distribution**,  $\mathcal{L}(\lambda)$ .
- 2.) Determine the  $(\pm 1\sigma_{\lambda})$   $\lambda = \lambda_{\pm 1\sigma}^*$  points on either side of the **local maximum**  $\lambda = \lambda^*$  of the  $\ln \mathcal{L}(\lambda)$  function from:  $\Delta \ln \mathcal{L}(\lambda) = \ln \mathcal{L}(\lambda^*) - \ln \mathcal{L}(\lambda_{\pm 1\sigma}^*) = \frac{1}{2}$ .
- 3.) Compute the (negative!) **curvature** at the **local maximum**  $\lambda = \lambda^*$  of the  $\ln \mathcal{L}(\lambda)$  function:

$$-\frac{d^2}{d\lambda^2} \ln \mathcal{L}(\lambda) \Big|_{\lambda=\lambda^*} = \frac{1}{\sigma_{\lambda}^2}$$

- 4.) Compute the **square** of the **local slope(s)** of the  $\ln \mathcal{L}(\lambda)$  function at the  $(\pm 1\sigma_{\lambda})$   $\lambda = \lambda_{\text{low}}, \lambda_{\text{high}}$  points:

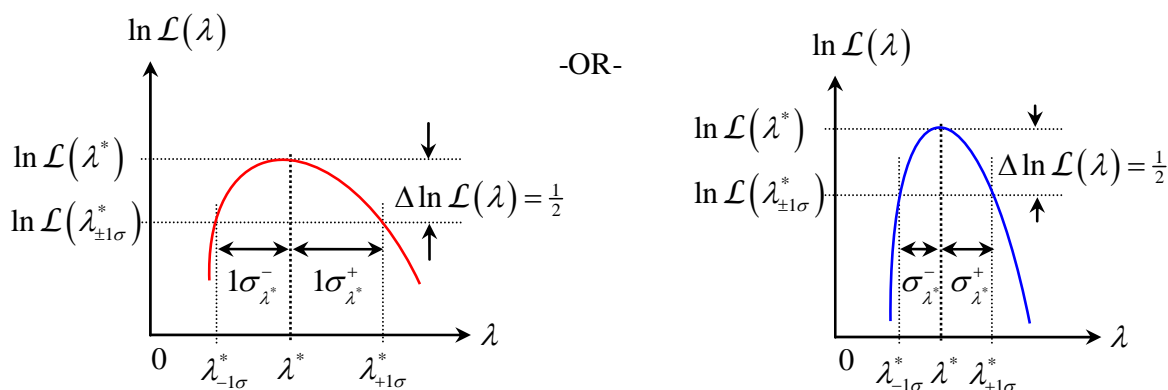
$$\left( \frac{d}{d\lambda} \ln \mathcal{L}(\lambda) \right)^2 \Big|_{\lambda=\lambda_{\pm 1\sigma}^*} = \frac{1}{\sigma_{\lambda}^2}.$$

Method 2.) above can also be used for obtaining **more** than just the  $\pm 1\sigma_{\lambda^*}$  (68.3% **double-sided/central Confidence Level limits** on the  $\lambda^*$  parameter! From the table on **central/double-sided Confidence Level Limits for Gaussian Distributions** in P598AEM Lect. Notes 8 (p. 6), we re-write it for use here, in terms of the  $\Delta \ln \mathcal{L}(\lambda) = \ln \mathcal{L}(\lambda^*) - \ln \mathcal{L}(\lambda_{\pm 1\sigma}^*)$  value associated with  $n_{\sigma} \sigma_{\lambda}$ . As can be seen from the table below, the relation is very simple:

$$\Delta \ln \mathcal{L}(\lambda) = \ln \mathcal{L}(\lambda^*) - \ln \mathcal{L}(\lambda_{\pm n_{\sigma}}^*) = \frac{n_{\sigma}}{2}.$$

$n_{\sigma} = \# \sigma_{\lambda}$	C.L. <sub>ds</sub> (%)	$\Delta \ln \mathcal{L}(\lambda)$
1.0	68.2689	0.5
2.0	95.4500	1.0
3.0	99.7300	1.5
4.0	99.9937	2.0
5.0	99.9999	2.5
6.0	100.0000	3.0

What does the  $\ln \mathcal{L}(\lambda)$  vs.  $\lambda$  curve look like when  $N$  is **finite**? This depends on the nature of the PDF associated with the **random variable**  $x$ . In general, for **finite** statistics, the  $\ln \mathcal{L}(\lambda)$  vs.  $\lambda$  can become “noisy” – *i.e.* it may not be a **perfectly** smooth curve – having increased statistical fluctuations in it as  $N$  decreases from  $\infty$ . Depending on the nature of the PDF associated with the **random variable**  $x$ , it may also become increasingly **asymmetrical** as  $N$  decreases from  $\infty$ , and may appear something like the curves shown in the figure below:



Clearly, in this asymmetrical situation, the individual results from using the 4 above methods of determining  $\sigma_{\lambda}$  can begin to **diverge** from each other as  $N$  decreases from  $\infty$ . Of the four, method # 2 ( $\Delta \ln \mathcal{L}(\lambda) = n_{\sigma}/2$ ) is the most “robust”/most widely accepted.

We see *e.g.* for the  $\pm 1\sigma$  **double-sided/central** 68.3% **Confidence Interval**, with  $\Delta \ln \mathcal{L}(\lambda) = \frac{1}{2}$ , that  $(\lambda^* - \lambda_{-1\sigma}^* = 1\sigma_{\lambda^*}^-) \neq (\lambda_{+1\sigma}^* - \lambda^* = 1\sigma_{\lambda^*}^+)$ , hence the (asymmetrical)  $\pm 1\sigma$  low-side/high-side uncertainties associated with the M.L.M.'s **most probable value** result are quoted as:  $\lambda^* \begin{smallmatrix} +\sigma_{\lambda^*}^+ \\ -\sigma_{\lambda^*}^- \end{smallmatrix}$ , *e.g.*  $5.04 \begin{smallmatrix} +0.10 \\ -0.04 \end{smallmatrix} \text{ Volts}$ .

Clearly, in this asymmetrical situation, for method # 4, the (local) **slopes** of the  $\ln \mathcal{L}(\lambda)$  vs.  $\lambda$  curve at  $\lambda = \lambda_{\text{low}}$  vs.  $\lambda = \lambda_{\text{high}}$  are **not** equal, but that is simply because they are anti-correlated with the numerical values of their respective sigmas, since they are inversely related to each other:

$$\left( \frac{d}{d\lambda} \ln \mathcal{L}(\lambda) \right)^2 \bigg|_{\lambda=\lambda_{-1\sigma}^*} = \frac{1}{(\sigma_{\lambda^*}^-)^2} \quad \text{and:} \quad \left( \frac{d}{d\lambda} \ln \mathcal{L}(\lambda) \right)^2 \bigg|_{\lambda=\lambda_{+1\sigma}^*} = \frac{1}{(\sigma_{\lambda^*}^+)^2}$$

Or simply:

$$\left| \frac{d}{d\lambda} \ln \mathcal{L}(\lambda) \right|_{\lambda=\lambda_{-1\sigma}^*} = \frac{1}{\sigma_{\lambda^*}^-} \quad \text{and:} \quad \left| \frac{d}{d\lambda} \ln \mathcal{L}(\lambda) \right|_{\lambda=\lambda_{+1\sigma}^*} = \frac{1}{\sigma_{\lambda^*}^+}$$

For method # 3, determining the **curvature** (*i.e.* 2<sup>nd</sup> derivative) of the  $\ln \mathcal{L}(\lambda)$  curve at  $\lambda = \lambda^*$  yields only a **single** number for  $\sigma_{\lambda}$ , which is actually a **weighted average** of  $\sigma_{\lambda}^-$  and  $\sigma_{\lambda}^+$ . This **could** be unfolded/unweighted, *e.g.* using the local slope information of the  $\ln \mathcal{L}(\lambda)$  vs.  $\lambda$  curve at  $\lambda = \lambda_{\text{low}}$  &  $\lambda = \lambda_{\text{high}}$ , but if one goes to all **that** effort, why not just use method # 4 instead?

Method # 1 suffers from the **same** problems as method # 3. Hence why method # 2 ( $\Delta \ln \mathcal{L}(\lambda) = n_{\sigma}/2$ ) is the most popular/most widely used method – it's very easy to carry out, it works with **any** likelihood function, and is also easily understood by others...

Please see/read **Muon Decay Asymmetry MLM Fit Example**, posted on the Physics 598AEM Software webpage, for a detailed discussion of the use of the MLM to obtain an **estimate** of the asymmetry parameter,  $\alpha^*$  and corresponding  $\pm 1\sigma_{\alpha^*}$  statistical uncertainties, from a large sample  $N$  of fully-polarized  $\mu^+ \rightarrow e^+ + \nu_e + \bar{\nu}_{\mu}$  decays.