Physics 403

Parameter Estimation

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

Table of Contents

Reading

- Sivia: Ch. 2
- Cowan: Ch. 5

Last Time: The Odds Ratio

To select between two models, it is useful to calculate the ratio of the posterior probabilities of the models. This is called the odds ratio:

$$\begin{aligned} O_{ij} &= \frac{p(D|M_i, I)}{p(D|M_j, I)} \frac{p(M_i|I)}{p(M_j|I)} \\ &= B_{ij} \frac{p(M_i|I)}{p(M_j|I)} \end{aligned}$$

The first term is called the Bayes Factor [1, 2] and the second is called the prior odds ratio. Interpration:

- **Prior odds**: the amount by which you favor M_i over M_j *before taking data*. There is no analog in frequentist statistics.
- **Bayes Factor**: the amount that the data D causes you favor M_i over M_j. Frequentist analog: *likelihood ratio* (but frequentists can't marginalize nuisance parameters)

Last Time: Occam Factors

- We can express any likelihood of data D given a model M as the maximum value of its likelihood times an Occam factor:

$$p(D|M,I) = \mathcal{L}_{\max}\Omega_{\theta}$$

- The Occam factor corrects the likelihood for the statistical trials incurred by scanning the parameter space for $\hat{\theta}$.
- Occam's Razor: when selecting from among competing models, generally prefer the simpler model
- Statistical Trials: it becomes harder to reject the "null hypothesis" when the number of hypotheses in a test becomes large.

Example

You have a histogram and look for a spike in any one bin. The look-elsewhere effect: any bin could be a background fluctuation.

Table of Contents

Principle of Indifference

As a general rule, we want priors that do not inadvertently push us toward a result. We want non-informative priors. **Principle of Indifference**: given $n > 1$ mutually exclusive and exhaustive possibilities, each should be assigned a probability equal to $1/n$.

Principle of Indifference

As a general rule, we want priors that do not inadvertently push us toward a result. We want non-informative priors. **Principle of Indifference**: given $n > 1$ mutually exclusive and exhaustive possibilities, each should be assigned a probability equal to $1/n$.

Example

Drawing from a deck of cards, we apply the principle of indifference and assume the probability of selecting a given card is $1/52$.

Principle of Indifference

As a general rule, we want priors that do not inadvertently push us toward a result. We want non-informative priors. **Principle of Indifference**: given $n > 1$ mutually exclusive and exhaustive possibilities, each should be assigned a probability equal to $1/n$.

Example

Drawing from a deck of cards, we apply the principle of indifference and assume the probability of selecting a given card is $1/52$.

Example

Rolling dice with n faces, we assume the die lands on one face (exclusive possibility) with probability $1/6$.

Principle of Indifference

As a general rule, we want priors that do not inadvertently push us toward a result. We want non-informative priors. **Principle of Indifference**: given $n > 1$ mutually exclusive and exhaustive possibilities, each should be assigned a probability equal to $1/n$.

Example

Drawing from a deck of cards, we apply the principle of indifference and assume the probability of selecting a given card is $1/52$.

Example

Rolling dice with n faces, we assume the die lands on one face (exclusive possibility) with probability $1/6$.

Example

Statistical mechanics: any two microstates of a system with the same energy are equally probable at equilibrium.

Principle of Indifference

Continuous Location Parameter

- Consider an event that we locate with respect to some origin (a "location parameter"
- Example: we are interested in $p(X|I)$, where $X =$"the tallest tree in the woods is between x and $x + dx$."
- In the problem, x is measured with respect to some origin. What if we change the origin so that

$$x \to x' = x + c$$

- In the limit of complete ignorance, our choice of prior must be completely indifferent to shifts in location. This implies

$$p(X|I)\, dX = p(X'|I)\, dX' = p(X'|I)\, d(X + c) = p(X'|I) dX$$
$$\therefore p(X|I) = \text{constant}$$

Uniform Prior

Continuous Location Parameter

- If we have upper and lower bounds on x (we know the dimensions of the woods), then

$$p(X|I) = \text{constant} = \frac{1}{x_{\text{max}} - x_{\text{min}}},$$

  the uniform prior we have already used a few times.
- If the bounds x_{min} and x_{max} are not known, then technically $p(X|I)$ is not normalized. It is called an improper prior.
- **Note 1**: improper priors can be used in parameter estimation problems, as long as the posterior distribution is normalized.
- **Note 2**: improper priors cannot be used in model selection problems, because the Occam factors depend on knowing the prior range for each model parameter.

Principle of Indifference

Continuous Scale Parameter

- Consider a problem where we are interested in the mean lifetime of a particle. Lifetime is a scale parameter because it can only have positive values.
- We are interested in $p(\mathcal{T}|I)$, where $\mathcal{T}$=“the “mean lifetime is between τ and $\tau + d\tau$.”
- In the limit of complete ignorance, our prior must be indifferent to changes in scale β, e.g., if we change our time units $\tau \to \tau' = \beta\tau$:

$$p(\mathcal{T}|I)\,d\mathcal{T} = p(\mathcal{T}'|I)\,d\mathcal{T}' = p(\mathcal{T}'|I)\,d(\beta\mathcal{T}) = \beta p(\mathcal{T}'|I)\,d\mathcal{T}$$

If we represent the PDF by $g(\tau)$, then

$$g(\tau) = \beta g(\tau') = \beta g(\beta\tau) \implies g(\tau) = \text{constant}/\tau$$

Jeffreys Prior

Continuous Scale Parameter

- Since $g(\tau) \propto 1/\tau$, we must also have

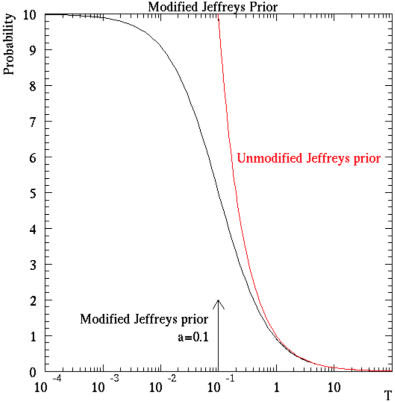
$$p(\mathcal{T}|I) \propto \frac{1}{\tau}$$

- This form of the prior is called the Jeffreys prior [1].
- If we have upper and lower bounds on τ then

$$p(\mathcal{T}|I) = \frac{1}{\tau \ln\left(\tau_{\text{max}}/\tau_{\text{min}}\right)}$$

- The Jeffreys prior is very convenient for problems in which we are ignorant about scale. It provides logarithmic uniformity via equal probability per decade. Using a uniform prior in this case would cause you to weight your PDF toward the highest decade

Modified Jeffreys Prior



- The Jeffreys prior is not normalizable if a scale parameter like τ can be zero.
- Alternative (from S. Oser): the modified Jeffreys prior, which becomes uniform for $\tau < a$:

$$p(\mathcal{T}|I) = \frac{1}{(\tau + a)\ln\left((a + \tau_{\max})/a\right)}$$

Caution: Parameterization Matters

Example from S. Oser

Two theorists predict the mass of a new particle:

1. **A**: There should be a new particle whose mass is between 0 and 1 in rationalized uints. Having no other knowledge about the mass, assume it has equal chance of being between 0 and 1: $p(m|I) = 1$.
2. **B**: There is a particle described by a free parameter $y = m^2$. The true value of y must lie between 0 and 1, but otherwise having no knowledge about it, $p(y|I) = 1$.

Both statements express ignorance about the same theory, but with different parameterizations. By the transformation rule,

$$p(y|I) = p(m|I)\left|\frac{dm}{dy}\right| \sim \frac{1}{\sqrt{y}}$$

Uh oh: transformation of variables makes a uniform prior non-uniform.

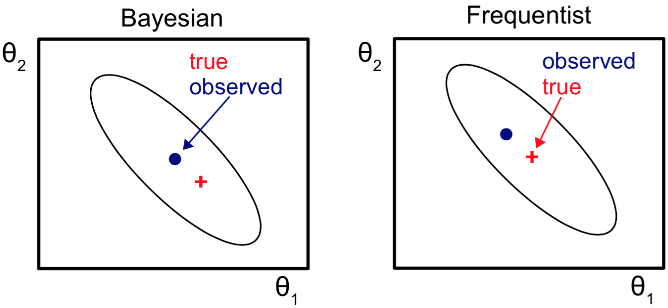
Table of Contents

Estimators

- We have seen how the PDF encodes what we want to know about a parameter given data D and relevant background information I.
- An estimator is a summary of this distribution
  - Could be a parameter of the PDF. E.g., p for a binomial distribution
  - Could be a property of the distribution, like the mean
- You have total freedom to make up any estimator you want, but you'll want to report two numbers:
  1. The best estimate itself
  2. A measure of the reliability of the estimate
- Question: what do we mean by "best" estimator?
- Question: what do we mean by the "reliability" of the estimator?

Bayesian vs. Frequentist Interpretations

- **Bayesian**: given D, the uncertainties tell us that the true value of the parameter lies within the ellipse centered on the observation with some probability
- **Frequentist**: given the true value of the parameters, the observation lies within an error ellipse centered on the true value with some probability



What is a Best Estimator?

- Let's answer the question of what defines a best estimator.
- Intuitive: it should be where the posterior PDF $p(x|D,I)$ is a maximum, meaning

$$\left.\frac{dp}{dx}\right|_{\hat{x}} = 0$$

  For this to be a maximum, we also require that

$$\left.\frac{d^2p}{dx^2}\right|_{\hat{x}} < 0$$

- If $\hat{x}$ gives the best estimator, then how do we define the reliability of the estimator?
- Look at the behavior of the PDF in a small region around the peak.

Reliability of an Estimator?

- Let's look at the Taylor expansion of p about $\hat{x}$, or better yet, $\ln p$:

$$L = \ln p = \ln p(x|D, I)$$

- We use the logarithm because p will often be a "peaky" function of x near $\hat{x}$. L varies more slowly and is a monotonic function of p.
- Taylor expanding L about $\hat{x}$, we get

$$L = L(\hat{x}) + \frac{1}{2}\frac{d^2L}{dx^2}\bigg|_{\hat{x}}(x - \hat{x})^2 + \ldots$$

- The first term is a constant. The linear term vanishes (we're at the maximum). So the quadratic term dominates, and

$$p(x|D, I) \approx A\ \exp\left[\frac{1}{2}\frac{d^2L}{dx^2}\bigg|_{\hat{x}}(x - \hat{x})^2\right]$$

Reliability of an Estimator?

- Compare the Taylor-expanded posterior PDF

$$p(x|D,I) \approx A \ \exp\left[\frac{1}{2}\frac{d^2L}{dx^2}\bigg|_{\hat{x}}(x-\hat{x})^2\right]$$

to the Gaussian

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

- We can identify the width of the Gaussian as

$$\sigma = \left(-\frac{d^2L}{dx^2}\bigg|_{\hat{x}}\right)^{-1/2}$$

with $d^2L/dx^2 < 0$ (we're at the maximum). Hence, we express the parameter as

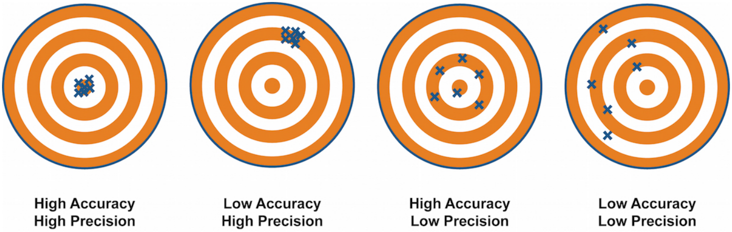
$$x = \hat{x} \pm \sigma,$$

where $\hat{x}$ is the best estimate and σ is its reliability.

Accuracy and Precision

Frequentist Aside

- It is useful to think of an estimator in terms of accuracy and precision
- **Accuracy**: how close is the estimator to true value? (Systematics)
- **Precision**: how clustered is the estimator about a central value? (Variance/Statistics)



Consistency and Bias

Caution: Frequentist Concept

- In the context of a sample of N measurements, we say that an estimator of θ, called $\hat{\theta}$, is consistent if

$$\lim_{N\to\infty} P(|\hat{\theta}-\theta| > \epsilon) = 0, \quad \forall\, \epsilon > 0$$

  I.e., $\hat{\theta}$ converges to θ in the large N limit.

- We call an estimator unbiased if the bias b

$$b(\theta) = \mathrm{E}\,(\hat{\theta}) - \theta$$

  is zero.

- An estimator can be biased even if it is consistent. If $\hat{\theta} \to \theta$ for an infinite set of measurements in one experiment, it is not necessarily true that $\hat{\theta} \to \theta$ in an infinite set of experiments with a finite number of measurements.

Mean Squared Error (or Deviation)

- It is helpful to think of bias as a systematic error which does not improve with more data
- Another popular measure of the quality of an estimator is the mean squared error, defined as

$$\begin{aligned} d = \mathrm{MSE} &= \mathrm{E}\left((\hat{\theta} - \theta)^2\right) \\ &= \mathrm{E}\left((\hat{\theta} - \mathrm{E}(\hat{\theta}))^2\right) + \left(\mathrm{E}(\hat{\theta}) - \theta\right)^2 \\ &= \mathrm{var}(\hat{\theta}) + b^2 \end{aligned}$$

- I.e., the mean squared error (MSE) is the sum of the variance and the square of the bias.
- Classical interpretation: since the variance is the square of the uncertainty in the estimator, the MSE is the quadrature sum of statistical and systematic uncertainties.
- Root mean square (RMS) is defined as $\sqrt{\mathrm{MSE}}$.

What Makes a Good Estimator?

Let's define the three properties we expect from a good estimator.

1. **Consistent**: a consistent estimator will tend to the true value as the amount of data approaches infinity:

$$\lim_{N\to\infty} \hat{\theta} = \theta$$

2. **Unbiased**: the expectation value of the estimator is equal to the true value, so its bias b vanishes:

$$b = \langle\hat{\theta}\rangle - \theta = \int d\boldsymbol{x}\, p(\boldsymbol{x}|\theta)\, \hat{\theta}(\boldsymbol{x}) - \theta = 0$$

3. **Efficient**: the variance of the estimator is as small as possible (we'll see how small when we discuss the method of maximum likelihood):

$$\text{var}\,(\hat{\theta}) = \int d\boldsymbol{x}\, p(\boldsymbol{x}|\theta)\, (\hat{\theta}(\boldsymbol{x}) - \hat{\theta})^2$$

$$\text{MSE} = \langle(\hat{\theta} - \theta)^2\rangle = \text{var}\,(\hat{\theta}) + b^2$$

It is not always possible to satisfy all three requirements.

Case Study: Efficiency Uncertainty

Example

Suppose you use simulation to determine a selection efficiency: n out of N events pass some cuts. What is the selection efficiency ϵ and its uncertainty?

This is a binomial process: fixed trials N, fixed successes n, probability of success ϵ. Therefore,

$$p(n|N,\epsilon) \propto \epsilon^n (1-\epsilon)^{N-n}$$

and

$$\begin{aligned} L = \ln p &= \text{constant} + n \ln \epsilon + (N-n) \ln (1-\epsilon) \\ \frac{dL}{d\epsilon} &= \frac{n}{\epsilon} - \frac{N-n}{1-\epsilon} \\ \frac{d^2L}{d\epsilon^2} &= -\frac{n}{\epsilon^2} - \frac{N-n}{(1-\epsilon)^2} \end{aligned}$$

Case Study: Efficiency Uncertainty

Example

For the optimal value of ϵ, $dL/d\epsilon = 0$:

$$\left.\frac{dL}{d\epsilon}\right|_{\hat{\epsilon}} = \frac{n}{\hat{\epsilon}} - \frac{N-n}{1-\hat{\epsilon}}$$
$$\therefore \hat{\epsilon} = \frac{n}{N}$$

This is a pretty intuitive result: the best estimate of the efficiency is just n/N. Mixing in a frequentist concept: is it biased?

$$b = \mathrm{E}\left(\hat{\epsilon}\right) - \epsilon = \frac{\mathrm{E}\left(n\right)}{N} - \epsilon = \frac{N\epsilon}{N} - \epsilon = 0$$

So $\hat{\epsilon}$ is an unbiased estimator.
What about its uncertainty?

Case Study: Efficiency Uncertainty

Example

The estimated variance is given by

$$\hat{\sigma}^2 = -\left(\left.\frac{d^2L}{d\epsilon^2}\right|_{\hat{\epsilon}}\right)^{-1}$$

After substituting $\hat{\epsilon} = n/N$ and combining terms, this reduces to

$$\left.\frac{d^2L}{d\epsilon^2}\right|_{\hat{\epsilon}} = -\frac{N}{\hat{\epsilon}(1-\hat{\epsilon})}$$

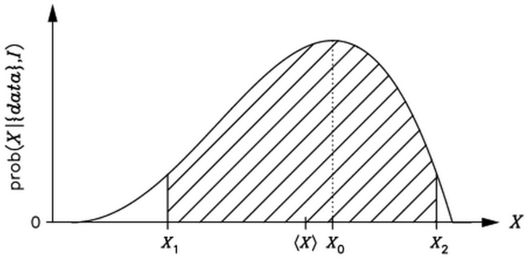
$$\therefore \hat{\sigma}^2 = \frac{\hat{\epsilon}(1-\hat{\epsilon})}{N} = \frac{n(N-n)}{N^3}$$

The expectation of $\hat{\sigma}^2$ is, after some more algebra,

$$\mathrm{E}\left(\hat{\sigma}^2\right) = \frac{N+1}{N}\sigma^2 \qquad \text{(slight bias)}$$

Asymmetric PDFs

- What happens when we have a very asymmetric PDF? In this case the expansion about the maximum may not be so reasonable.



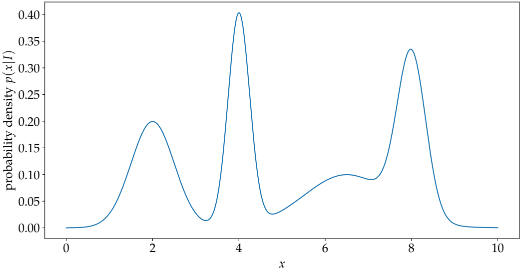
- This is where the concept of confidence intervals (or "credible regions" for a Bayesian) come in. We define

$$p(x_1 \leq x < x_2 | D, I) = \int_{x_1}^{x_2} p(x|D,I)\,dx \approx \alpha,$$

where $\alpha = 0.68$ (for example), and identify x_1 and x_2.

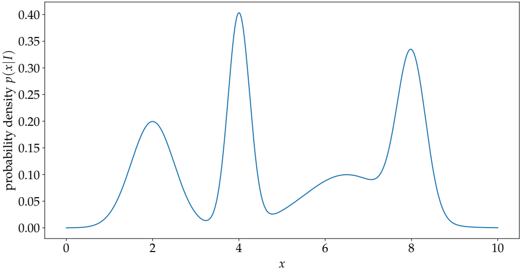
Multimodal PDFs

- What happens when we the PDF is multimodal? Can we even describe a "best parameter" and its uncertainty properly?



Multimodal PDFs

- What happens when we the PDF is multimodal? Can we even describe a “best parameter” and its uncertainty properly?



- You could try to summarize the posterior using ≥ 2 best estimates and their error bars, or some kind of disjoint confidence interval.
- Alternatively: cut your losses and just report the full posterior PDF.

Gaussian Uncertainties

- Suppose we are measuring values $\boldsymbol{x} = \{x_i\}$ drawn from a Gaussian distribution of mean μ and variance σ^2.
- For today, assume σ^2 is known but μ is not. How do we estimate μ given the data?
- Starting from Bayes' Theorem,

$$p(\mu|\boldsymbol{x},\sigma^2,I) \propto p(\boldsymbol{x}|\mu,\sigma^2,I)\, p(\mu|\sigma^2,I)$$

- **Likelihood**: If the measurements x_i are independent, then

$$p(\boldsymbol{x}|\mu,\sigma^2,I) = \prod_{i=1}^{N} p(x_i|\mu,\sigma^2,I) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\sum_i \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

- **Prior**: μ is a location parameter, so we'll use a uniform prior

$$p(\mu|\sigma^2,I) = \frac{1}{\mu_{\text{max}} - \mu_{\text{min}}}$$

which vanishes outside $x \in [\mu_{\text{min}}, \mu_{\text{max}}]$.

Gaussian Uncertainties

Estimate of the Mean

- As in the earlier examples, let's maximize the logarithm of the posterior PDF to get the best estimate for μ:

$$L = \ln p(\mu|\boldsymbol{x},\sigma^2,I) = \text{constant} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

- Differentiating, we have

$$\left.\frac{dL}{d\mu}\right|_{\hat{\mu}} = \sum_{i=1}^{N} \frac{x_i-\mu}{\sigma^2} = 0$$

$$\therefore \boxed{\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i}.$$

So the best estimate of μ is the arithmetic mean of the measurements, independent of the spread given by σ.

Gaussian Uncertainties

Uncertainty of the Mean

- The uncertainty of the mean comes from the second derivative:

$$\left.\frac{d^2L}{d\mu^2}\right|_{\hat{\mu}} = -\sum_{i=1}^{N} \frac{1}{\sigma^2} = -\frac{N}{\sigma^2}$$

- Therefore, our best estimate and uncertainty on the mean is

$$\boxed{\mu = \hat{\mu} \pm \frac{\sigma}{\sqrt{N}}}$$

- We have derived the expression often referred to as the "error on the mean," including the rule that the uncertainty decreases as $1/\sqrt{N}$.
- The only requirement is the validity of the quadratic expansion of the posterior PDF, which is exactly true for the Gaussian.
- This rule applies often thanks to the tendency of additive sources of noise to look Gaussian (Central Limit Theorem)

Different-Sized Error Bars

Weighted Mean

- What happens if the uncertainties in each x_i differ? As long as the source of uncertainties is Gaussian, then

$$p(\boldsymbol{x}|\mu,\sigma_i^2,I) = \prod_{i=1}^{N} p(x_i|\mu,\sigma_i^2,I) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}}\exp\left(-\sum_i \frac{(x_i-\mu)^2}{2\sigma_i^2}\right)$$

where $\boldsymbol{\Sigma}$ is the diagonal covariance matrix of the $\{x_i\}$.

- Taking the logarithm and differentiating gives

$$L = \ln p = \text{constant} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma_i^2}$$

$$\left.\frac{dL}{d\mu}\right|_{\hat{\mu}} = \sum_{i=0}^{N} \frac{x_i-\mu}{\sigma_i^2} = 0$$

$$\therefore \hat{\mu} = \sum_{i=1}^{N} x_i/\sigma_i^2 \Bigg/ \sum_{i=1}^{N} 1/\sigma_i^2 = \sum_{i=1}^{N} x_i w_i \Bigg/ \sum_{i=1}^{N} w_i$$

Different-Sized Error Bars

Weighted Error on the Mean

- For the uncertainty on the mean, we have

$$\left.\frac{d^2L}{d\mu^2}\right|_{\hat{\mu}} = -\sum_{i=0}^{N}\frac{1}{\sigma_i^2}$$

$$\therefore \mu = \hat{\mu} \pm \left(\sum_{i=1}^{N} w_i\right)^{-1/2}, \qquad w_i = 1/\sigma_i^2$$

- So for the case of different uncertainties on each measurement x_i, the best estimator of the mean is the arithmetic sum of the data inversely weighted by the uncertainties.
- This makes a lot of sense; we want the data points with the biggest uncertainties to contribute the least to the sum

Table of Contents

Principle of Indifference

Uniform and Jeffreys Priors

- **Principle of Indifference**: given $n > 1$ mutually exclusive and exhaustive possibilities, each should be assigned a probability equal to $1/n$.
- Matches our intuition, and we've been applying it throughout the course. We can also use it to derive PDFs.
- Uniform prior is appropriate for a location parameter:

$$p(X|I) = \text{constant} = \frac{1}{x_{\text{max}} - x_{\text{min}}},$$

- Jeffreys prior is appropriate for a scale parameter:

$$p(X|I) = \frac{1}{x \ln\left(x_{\text{max}}/x_{\text{min}}\right)}$$

It gives equal probability per decade.

Summary

- We can identify the best estimator of a PDF by maximizing it, so that
$$\left.\frac{dp}{dx}\right|_{\hat{x}} = 0$$
- We assessed the reliability of the estimator by Taylor expanding $L = \ln p$ about the best value:
$$\hat{\sigma}^2 = \left(-\left.\frac{d^2L}{dx^2}\right|_{\hat{x}}\right)^{-1}$$
- This only works when the quadratic approximation is reasonable. It may not be:
  1. **Asymmetric PDF**: better to use a confidence interval
  2. **Multimodal PDF**: no clear best estimate; report full PDF
- Frequentists: desire efficient, unbiased, and consistent estimators.