Likelihood

Physics 252C - Lecture 9

Prof. John Conway

warmup example: "error on the error"

- this has little to do with likelihoods, but it's interesting...
- what is the "error on the estimate of the error"?
- equivalently, what is the variance of the estimate of the variance?
- we have a sample $\{x_i, i = 1, 2, ..., N\}$

$$\hat{\mu} = \frac{1}{N}\sum_{I=1}^{N} x_i \qquad \hat{\sigma} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

- first, what is the error on the estimate of the mean?

$$V[\hat{\mu}] = E[\hat{\mu}^2] - E[\hat{\mu}]^2 = ... = \frac{\hat{\sigma}^2}{N}$$

warmup example: "error on the error"

- can apply this to the estimator for the standard deviation; we get

$$V[\hat{\sigma}^2] = \frac{1}{N}\left(m_4 - \frac{N-3}{N-1}\hat{\sigma}^4\right)$$

- for gaussian distributed numbers

$$V[\hat{\sigma}^2] = \frac{\hat{\sigma}^2}{2N}$$

- therefore the "error on the error" is $\sigma/\sqrt{2N}$

what is a likelihood?

- simply put, a likelihood is a number proportional to a probability
- a likelihood as a function of a parameter α could, for example, be set to the value of the probability density for some observation x given α :

$$\mathcal{L}(\alpha) \equiv \mathcal{P}(x; \alpha)$$

- note that integrating the likelihood with respect to α is not a probability! (wrong dimensions, for starters)

uses of likelihoods

- parameter estimation
  - can use likelihood as a means to derive estimates of parameters given observations; could be the basis for example of a Neyman construction for a frequentist approach
- Bayesian posterior densities
  - idea is to take likelihood and prior in some parameter, and derive a posterior density in the parameter using Bayes' Theorem
- hypothesis testing (LR)
  - use likelihood ratios to decide between competing hypotheses

the likelihood "principle"

- "The likelihood function contains all of the information about a sample."
- this is controversial!
- read Edwards book <u>Likelihood</u> (missing from UC Davis library, alas)
- tons of literature on the subject...
- not to be confused with the maximum likelihood principle !

maximum likelihood estimators

- maximum likelihood principle

  "The values of a set of parameters which maximize the likelihood for a given set of observations is the best estimate of the parameters." - Fisher, 1912

- such parameter estimates are called maximum likelihood estimators; they are
  - unbiased
  - efficient
  - asymptotically normal
  - invariant under transformation

R.A. Fisher

example: maximum likelihood mean

- we again have a sample $\{x_i, i = 1, 2, ..., N\}$
- we want to write the likelihood for the mean
- need hypothesis for pdf !

$$\mathcal{L}(\mu) = \prod_{i=1}^{N} f(x_i; \mu)$$

- suppose f(x;μ) is a gaussian; then need σ also?

$$\mathcal{L}(\mu) = \prod_{i=1}^{N} e^{-(x_i-\mu)^2/2\sigma^2}$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

calculating likelihoods

- if we have large N, then the product can get very small:

$$\mathcal{L}(\mu) = \prod_{i=1}^{N} e^{-(x_i-\mu)^2/2\sigma^2}$$

- numerically it is almost always necessary to deal with the log of the likelihood; in this case

$$\log \mathcal{L}(\mu) = -\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

- this turns the likelihood calculation into a sum
- maximizing the likelihood is equivalent to minimizing the negative of the log likelihood!

likelihood and χ^2

- for gaussian data, we see that there is a connection between -ln$\mathcal{L}$ and χ^2:

$$-2\log\mathcal{L}(\mu) = -\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^2} = \chi^2$$

- another way to say it is that

$$\mathcal{L}(\mu) = e^{-\chi^2/2}$$

- minimize $\chi^2 \Rightarrow$ maximize likelihood

- but don't get fooled into thinking you can use likelihood for goodness of fit (though, maybe...)

joint likelihoods

- we needn't restrict ourselves to such simple examples
- suppose we have several measurements which depend on the same parameter
- then we can write

$$\mathcal{L}(\mu) = \mathcal{L}_1(x_1; \mu) \times \mathcal{L}_2(x_2; \mu) \times \ldots$$

- product of likelihoods is a joint likelihood
- do need to worry about correlations among measurements, however!

likelihoods for spectra

- in general, we shall refer to an ordered set of measurements like this as a spectrum:

$$\{y_i(x_i), i = 1, 2, ..., n\}$$

- as in the case of χ^2, we can write a functional form to describe the data

$$\tilde{y}(x; \bar{\alpha})$$

- if we know the applicable probability (density) we can write

$$\mathcal{L}(\bar{\alpha}) = \prod_{i=1}^{N} \mathcal{P}(y_i(x_i); \tilde{y}(\bar{\alpha}))$$

- "likelihood fit": maximize likelihood w.r.t. the α

likelihoods for Poisson-distributed spectra

- most common example: likelihood fit to observed spectrum, with data in bins of x
- in this case we know the number of events we observe in each bin is described by a Poisson distribution:

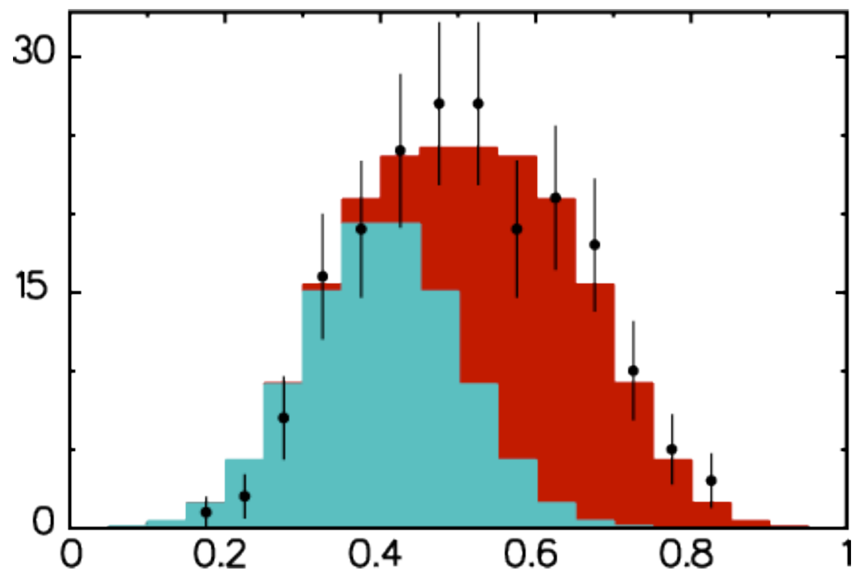
$$\mathcal{L}(\bar{\alpha}) = \prod_{i=1}^{N} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

- here the μ_i depend in some way on the α parameters
- one way to write the unknown parameters is simply as the cross sections:

$$\mu_i = \sigma_1 L \epsilon_{1i} + \sigma_2 L \epsilon_{2i} + ...$$

likelihoods for Poisson-distributed spectra

- example: two overlapping Gaussians



- the y_i are the data points with $\sqrt{n}$ error bars
- the predicted number of events in each bin is given by

$$\mu_i = \alpha_1 G(x_i; \mu_1, \sigma_1) + \alpha_2 G(x_i; \mu_2, \sigma_2)$$

- fit can be from one to six parameters...

likelihoods for Poisson-distributed spectra

- for the Poisson spectrum likelihood the log is

$$\log \mathcal{L} = \sum_{i=1}^{N} y_i \log \mu_i - \mu_i - \log y_i!$$

- the last term is usually dropped since it is a constant; we only care about minimizing -logL with respect to changes in the parameters
- note that empty bins contribute to the likelihood, but not bins where nothing is expected

combining results using likelihoods

- clearly we can use the multiplicative property of likelihoods to combine quite different measurements of a parameter, even from different experiments:

$$\mathcal{L}(\mu) = \mathcal{L}_1(x_1;\mu) \times \mathcal{L}_2(x_2;\mu) \times ...$$

- again: the hardest part is that there may be correlations between the experiments
- typically these correlations can be captured in additional parameters that co-vary
- in practice, people from different experiments must sit together, swap code/data, etc. LEPEWWG, etc.

delta-log-likelihood

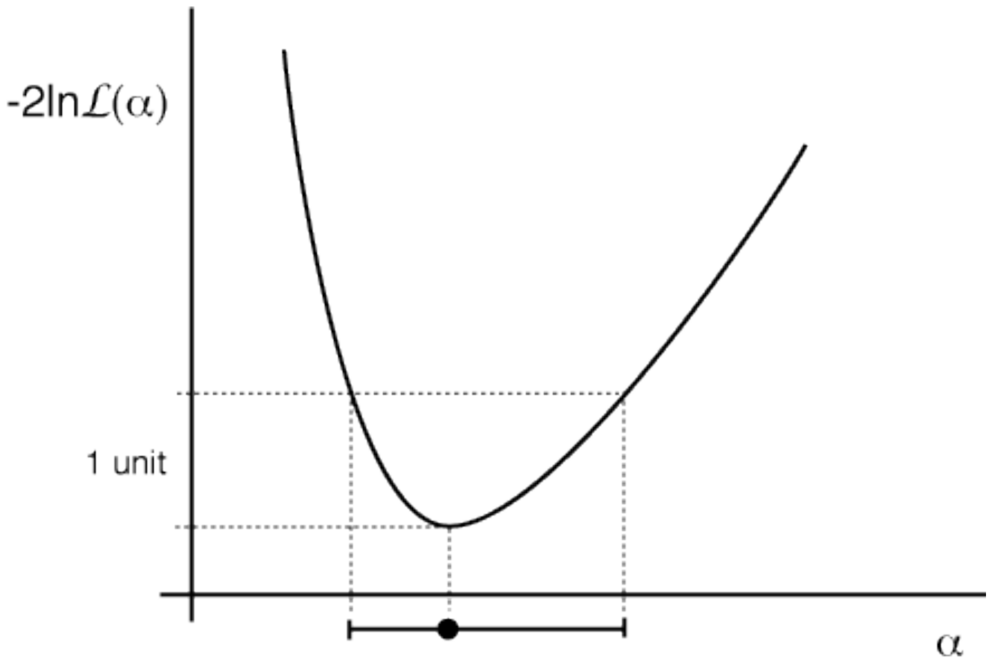
- as we saw, the change in χ^2 by one unit corresponds to a 1-standard-deviation shift in the parameter
- we use different $\Delta\chi^2$ for different numbers of parameters varying simultaneously
- $\Delta(2\ln L)$ behaves very much like $\Delta\chi^2$

Table 32.2: $\Delta\chi^2$ or $2\Delta\ln L$ corresponding to a coverage probability $1-\alpha$ in the large data sample limit, for joint estimation of m parameters.

$(1-\alpha)$ (%)	$m=1$	$m=2$	$m=3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

delta log-likelihood intervals

- can use this property of $\Delta(2\ln L)$ to determine intervals for measurements, just as with $\Delta\chi^2$



very much like highest posterior density Bayes intervals!

- coverage properties are remarkably well behaved
- generalizes to multiple dimensions: likelihood “contours”

when do we use χ^2 versus likelihood fits?

- in general, if your problem is definitely gaussian, and, even better, you have parameters which enter linearly only, use a chi square fit (LLS)
- LLS fit is as fast as matrix inversion; speed counts!
- likelihood fits will allow you to account for non-gaussian behavior (Poisson) and nonlinear functions of the parameters
- multi-bin Poisson spectrum gives very gaussian results
- likelihood fit is limited by the minimization technique
- try to minimize analytically!
- otherwise we are stuck with MINUIT, FUMILI, etc.

likelihoods and Bayesian posteriors

- if we have a likelihood function we can use the Bayesian treatment to convert it into a posterior pdf in the parameter if interest

$$\mathcal{P}(\alpha;\bar{x}) = \frac{\mathcal{L}(\bar{x};\alpha)\mathcal{P}(\alpha)}{\int \mathcal{L}(\bar{x};\alpha')\mathcal{P}(\alpha')d\alpha'}$$

- denominator ensures that the pdf is normalized properly regardless of the prior
- jargon in the field "we integrated the likelihood to set our limit" $\Rightarrow$ they used a Bayesian treatment
- all the same techniques with intervals and limits apply here

unbinned likelihood (a.k.a. extended ML)

- why bin the data, since that just takes away information about the sample?
- if predicted distributions are from Monte Carlo, then it is quite natural to bin the data to get the ε_i
- can define an unbinned likelihood, which is a product over events, not bins
- must know the functional forms for the event distributions, from each of the event sources; then

$$\mathcal{L}(\alpha; \bar{x}) = \frac{\nu^n e^{-\nu}}{n!} \prod_{i=1}^{n} p(x_i; \alpha)$$

- here ν is the mean number of events, and n the observed

unbinned likelihoods (a.k.a. extended ML)

$$\log \mathcal{L}(\alpha; \bar{x}) = n \log \nu(\bar{\alpha}) - \nu(\bar{\alpha}) + \sum_{i=1}^{n} \log p(x_i; \bar{\alpha})$$

- here we explicitly show that ν is a function of the unknown parameters
- often the $p(x_i;\alpha)$ are superpositions of m different sources (backgrounds + signal for example) and the unknown parameters are the fractions in each source
- my preference is to recast this using cross sections:

$$\log \mathcal{L}(\bar{\sigma}; \bar{x}) = -\sum_{j=1}^{m} \mu_j + \sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} \mu_j f_j(x_i) \right)$$

$$\mu_j = \sigma_j L \epsilon_j, \quad j = 1, 2, ..., m$$

unbinned likelihoods

- use of unbinned likelihoods is surging
- this squeezes the most information possible out of a sample
- have to do work to figure out the analytic/ numerical form of pdfs!
- lots of examples in high energy physics...