Statistics I

Frank Porter

September 18, 2013

Plan for the Statistics Lectures

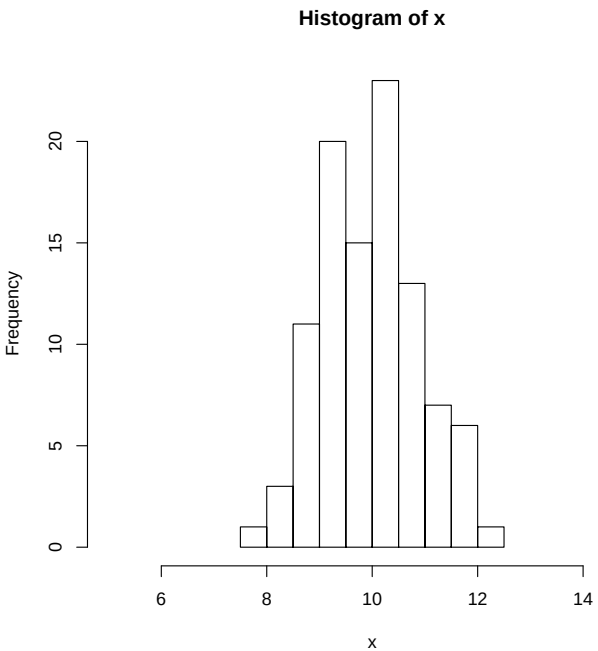
- Lecture I (Wednesday, September 18, 11:45-12:30)
  1. Important probability concepts
  2. Point estimation
- Lecture II (Thursday, September 19, 10:45-12:30)
  1. Frequency and Bayes interpretations
  2. Interval estimation
  3. Systematic uncertainties
- Lecture III (Friday, September 20, 10:45-12:30)
  1. Hypothesis tests
  2. Resampling methods
- Lecture IV (Saturday, September 21, 10:45-12:30)
  1. Density estimation

Lecture boundaries won't be so crisp!!

A couple of toolkits for statistical problems. . .

R (free, open source)

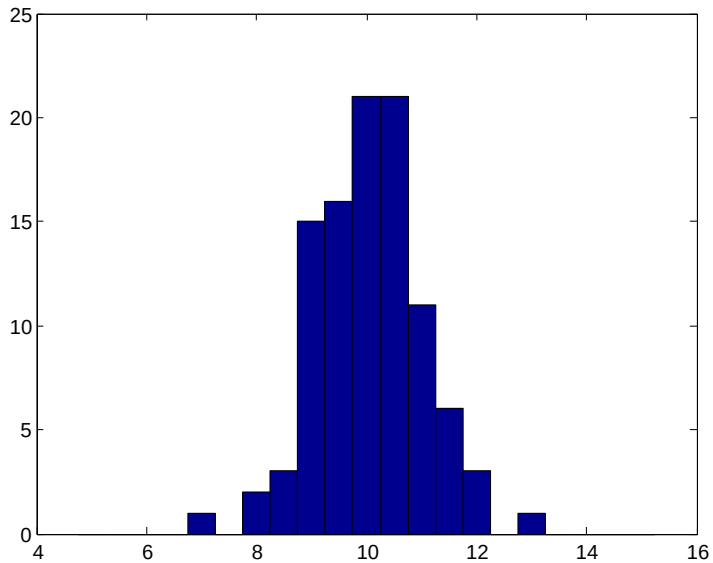
```
> x <- rnorm(100,10,1)
> hist(x,xlim=range(5,15))
>
```



http://cran.cnr.berkeley.edu

MATLAB (commercial)

```
>> x = normrnd(10, 1, 1, 100);
>> hist(x,5:.5:15)
```



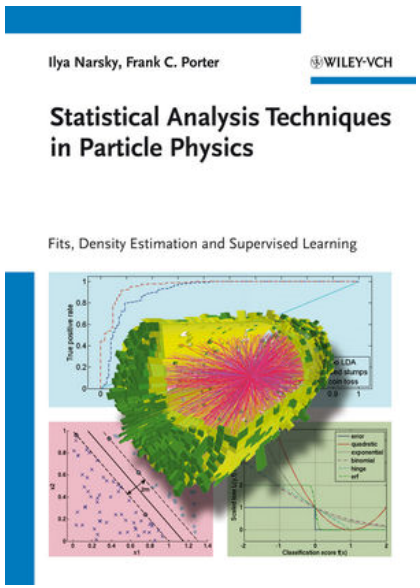
www.mathworks.com/products/matlab/

There are others. . .

Advertisement

WILEY

Home / Physics & Astronomy / Nuclear & High Energy Physics



Statistical Analysis Techniques in Particle Physics

Ilya Narsky, Frank C. Porter

ISBN: 978-3-527-41086-6

Paperback
459 pages
December 2013
US $99.95

Will refer to here as "NarskyPorter(2014), Wiley"

These lectures roughly correspond to chapters 1–5

This price is valid for United States. Change location to view local pricing and availability.

Description

Modern analysis of HEP data needs advanced statistical tools to separate signal from background. This is the first book which focuses on machine learning techniques. It will be of interest to almost every high energy physicist, and, due to its coverage, suitable for students.

Other references

1. B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman&Hall 1994.
2. E. Lehmann and J. Romano, *Testing Statistical Hypotheses*, 3rd Ed., Springer 2005.
3. D. Scott, *Multivariate Density Estimation*, Wiley 1992. A standard treatment of density estimation.
4. J. Shao, *Mathematical Statistics*, 2nd Ed., Springer 2003. A good, fairly recent general mathematical statistics textbook.

There are also many books written by particle physicists, such as Barlow, Bohm and Zech, Cowan, James, Lyons, . . .
There is much good and very relevant material in these, but be alert to implicit and explicit philosophical influences and physicist idiosyncracies.

Important probability concepts

1. The probability measure
2. Random variables
3. Some terminology
4. Central limit theorem
5. The exponential family
6. RCF theorem
7. Transformations and propagation of errors

The probability measure

- Consider a space S, called a sample space.
- A probability P defined on sets E in S is a measure on S such that $P(S) = 1$.
- We assume we deal with measurable sets. Then:
  - $P(E) \geq 0$ for $E \subset S$
  - If $E \cap F = \varnothing$, then $P(E \cup F) = P(E) + P(F)$.
  - If S is an infinite sample space and $E_1, E_2, \ldots$ is an infinite sequence of disjoint sets in S then

$$P(\cup_{i=1}^{\infty}) = \sum_{i=1}^{\infty} P(E_i).$$

Random variables (RVs)

- Assume we have a correspondence between elements of S and a set of real numbers R_S (could be vectors)
- The probability measure on S then defines a measure on $R \supset R_S$, we'll still call it P
- Mapping $X \in R_S$ is a "random variable"
- Define cumulative distribution function (cdf) $F_X(x)$:

$$F_X(x) = P(X \leq x)$$

- Define probability density function (pdf) $f_X(x)$ as the differential of $F_X(x)$

$$f_X(x)dx = dF_X(x)$$

- n.b., if distribution is "discrete" then measure is given by Dirac δ functions, and integrals are sums

Some terminology

- Expectation value of a function $U(X)$ of random variable X:

$$E_F(U) \equiv \int_R U(x) dF_X(x)$$

  Shorten to $E(U)$ where there is no confusioon
- Mean of X is $E(X)$
- Variance of X is $\mathrm{var}(X) = E[(X - E(X))^2]$

  Standard deviation is square root of variance
- In multivariate case, covariance is

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j) \equiv E\{[X_i - E(X_i)][X_j - E(X_j)\}$$

  Linear correlation coefficient is: $\rho = \frac{\mathrm{cov}(X_i, X_j)}{\sqrt{\mathrm{var}(X_i)\mathrm{var}(X_j)}}$
- Two RVs are (statistically) independent if the joint pdf factorizes:

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

  If we have a sequence of independent RVs that are identically distributed, we call them "iid"

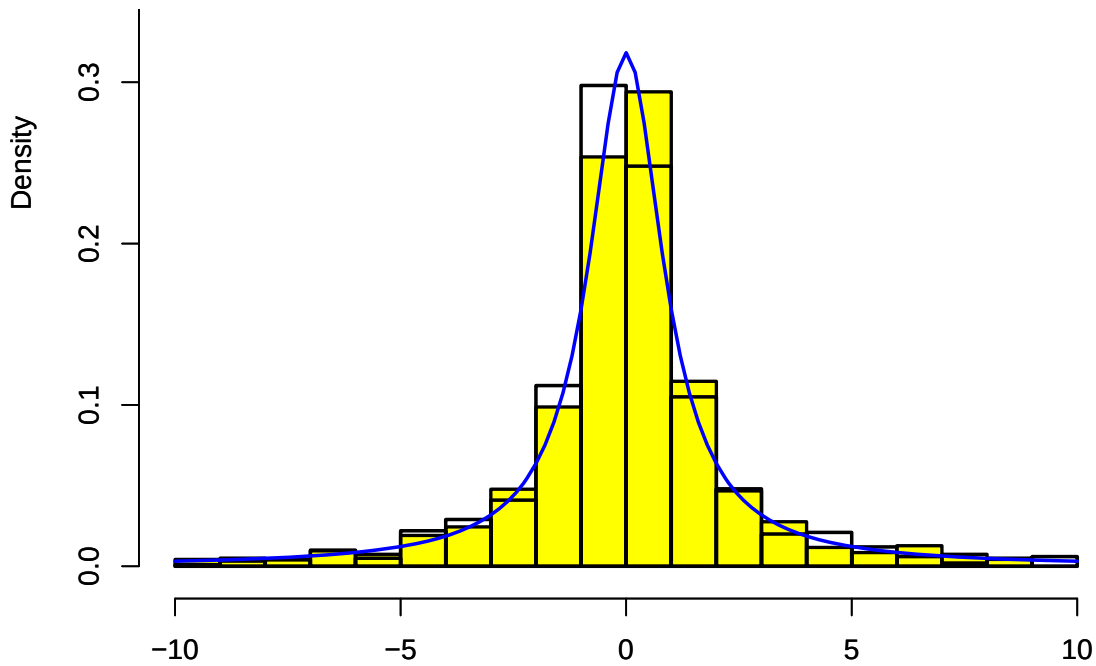
Central limit theorem

Let $X_1, X_2, \ldots, X_N$ be a sequence of iid RVs from distribution F, having finite mean μ and finite variance σ^2. Let $m = \frac{1}{N}\sum_{n=1}^{N} X_n$ be the sample mean. Then the distribution of m approaches $\mathrm{N}(\mu, \sigma^2/N)$

- We'll forego the proof, with associated discussion of convergence and characteristic functions
- Beware the proofs you find on the web, which typically make stronger than stated assumptions!
- There are variants, including generalization to non-iid sampling, with additional regularity conditions

Central limit theorem and the Breit-Wigner (Cauchy, Lortentz)

- CLT doesn't apply to Breit-Wigner (BW)
- Open histogram is sampling from BW with center at 0 and $\Gamma = 2$
- Curve is sampling distribution
- Yellow is sample mean for $N = 10000$



Sample mean has same distribution as a single sample!

Parametric and non-parametric statistics

- A distribution may depend on one or more (unknown) parameters θ; i.e.,

$$F_X(x) = F_X(x;\theta)$$

- Estimation of θ is the domain of "parametric statistics"
- We'll start with this problem
- Typically we assume a "model" for F and wish to find those parameter values that best "fit" the available data
- Later we'll talk about non-parametric methods, which can often be viewed as "model-independent"

The exponential family

Let $F_X(x;\theta)$ be a parametric cdf depending on paremeter(s) θ, on a sample space Ω with measure μ. The set of possible distributions $\{F : \theta \in \Theta\}$, where Θ is the parameter space, is called a family of distributions.

The family $\{F : \theta \in \Theta\}$ is an exponential family iff, $\forall x \in \Omega$:

$$\frac{dF_X(x;\theta)}{d\mu} = h(x) \exp\left[q(\theta)^{\mathsf{T}} g(x) - r(\theta)\right]$$

g maps X to $\mathcal{R}^k$; q maps θ to $\mathcal{R}^k$

We'll see that the exponential family distributions have nice properties.

Example of exponential family distributions: Binomial, Gaussian, Poisson

However, the BW is not in the exponential family

Transformations, propagation of errors

Consider continuous pdf $f(x)$, were $x = (x_1, \ldots x_N)$. Suppose we have a mapping $y = h(x)$ from x to $y = (y_1, \ldots y_N)$. If the y's are linearly independent, we can derive the pdf for y (call it g):

$$\begin{aligned} g(y)d^N(y) &= g[h(x)] \left| \frac{\partial h}{\partial x} \right| d^N(x) \\ &= f(x)d^N(x) \end{aligned}$$

Hence,

$$g(y) = \frac{f[h^{-1}(y)]}{\left| \frac{\partial h}{\partial x} \right|_{h^{-1}(y)}}$$

Propagation of errors

Often, content to learn the new covariance matrix, instead of the whole distribution. If $y = (y_1, \ldots y_M)$ depends linearly on $x = (x_1, \ldots x_N)$: $y = Tx + a$, where T is $M \times N$, then

$$\Sigma_y = T \Sigma_x T^{\mathsf{T}}$$

Even if y is not linearly dependent on x we can try a linear approximation:

$$T_{ij} = \left. \frac{\partial y_i}{\partial x_j} \right|_{x \sim E(X)}$$

If $M = 1$ and the x_n's are statistically independent (diagonal covariance matrix), then we obtain the most commonly-used propagation of errors formula:

$$\sigma_y^2 = \sum_{n=1}^{N} \left(\sigma_{x_n} \left. \frac{\partial y}{\partial x_n} \right|_{x \sim E(X)} \right)^2$$

Point Estimation

Point estimation deals with the estimation of parameter values. Given some parameter(s) θ we wish to determine, we use data X to form estimator $\hat{\theta}(X)$. Our estimator $\hat{\theta}(X)$ is a RV

- Desirable properties
- Least squares
- Maximum likelihood
- Substitution method

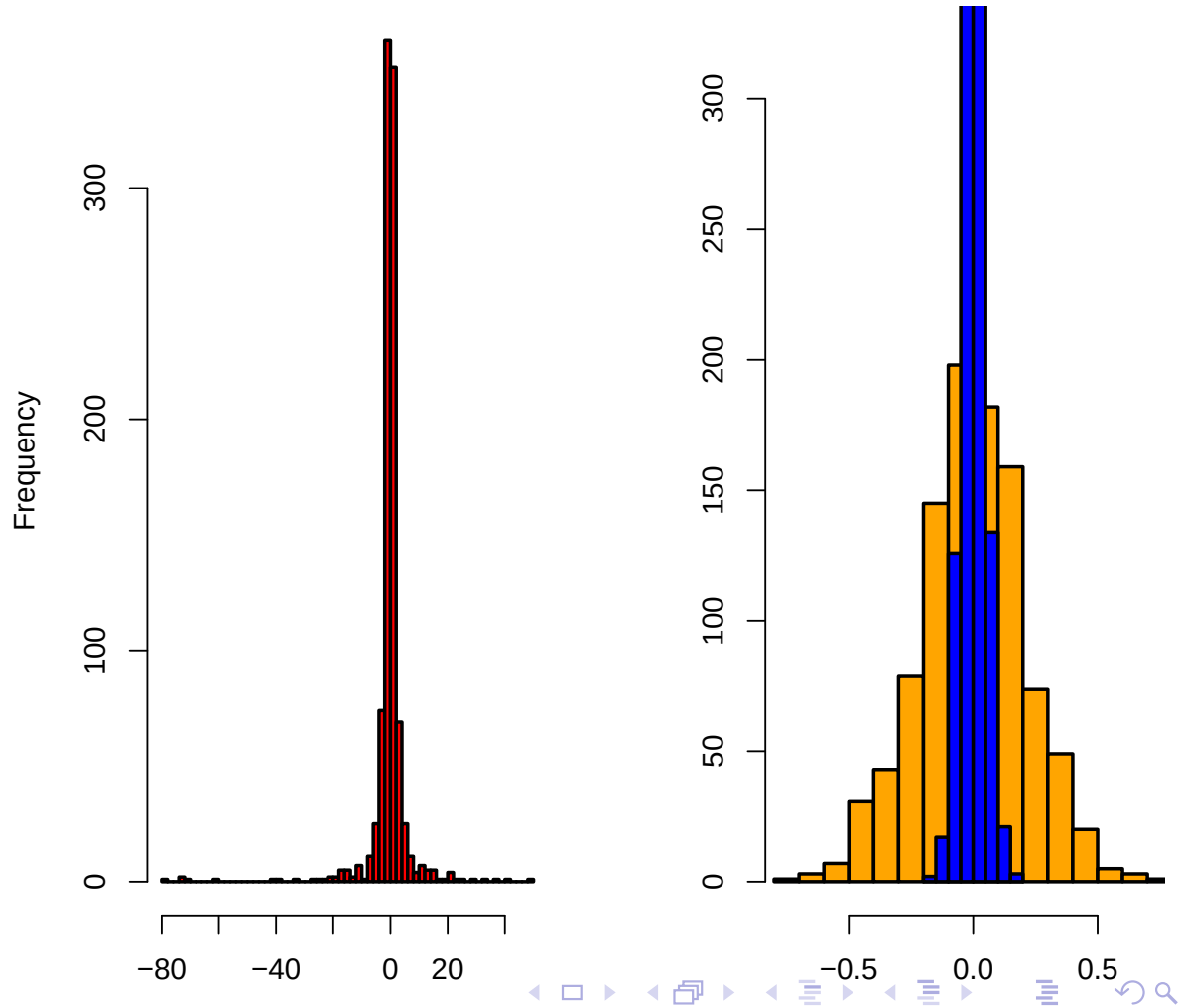
Point Estimation – Desirable properties

- Robust
- Unbiased
- Consistent
- Efficient
- Sufficient
- Practical

Point estimation – Desirable properties – Robustness

A robust estimator is an estimator that is insensitive to details of sampling distribution (i.e., to model errors).
For example, the sample mean is not in general a robust estimator of location.

- Estimation of center of a Breit-Wigner (1000 experiments with $N = 1000$)
- Sample mean has large variance
- Truncated mean does better (delete highest and lowest 1%)
- Median is even better



Point estimation – Desirable properties – Unbiased

First, a word about error: Our estimator $\hat{\theta}(X)$ is a random variable. The error in the estimator is the deviation from the true value:

$$\text{error} = \hat{\theta}(X) - \theta$$

This isn't very practical. However, we can imagine forming an average that we can use. Most common is the mean squared error:

$$\begin{aligned} \text{MSE}(\hat{\theta};\theta) &\equiv E[(\hat{\theta}(X) - \theta)^2] \\ &= \text{var}[\hat{\theta}(X)] + \left\{ E[\hat{\theta}] - \theta \right\}^2 \end{aligned}$$

The second term is the square of the bias,

$$b(\theta) \equiv E[\hat{\theta}] - \theta$$

For a given variance, we minimize the MSE by minimizing the bias.

Example of bias

The (maximum likelihood) estimator for the variance, σ^2, of X based on an iid sampling $x_1, \ldots, x_N$,

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}[x_n - \frac{1}{N}\sum_{m=1}^{N} x_m]^2$$

is biased:

$$b(\sigma^2) = -\frac{\sigma^2}{N}$$

Fortunately, this bias is easily eliminated by multiplying our estimator by $N/(N-1)$

Point estimation – Desirable properties – Consistency

An estimator is consistent if it is asymptotically unbiased.
The biased estimator in the example on the previous slide is consistent:

$$\lim_{N\to\infty}\left(-\frac{\sigma^2}{N}\right) = 0$$

Point estimation – Desirable properties – Efficiency

An estimator is efficient if it has small variance. Only makes sense to consider as function of bias (e.g., considering ubiased estimators)

To discuss efficiency, we introduce:

- The likelihood function, a function of θ for given sampling x

$$L(\theta; x) = f(x; \theta)$$

- The Fisher information number is

$$I(\theta) \equiv E\left\{\left[\frac{\partial \log L(\theta; X)}{\partial \theta}\right]^2\right\}$$

  This generalizes to the $R \times R$ Fisher information matrix in the case of a multidimensional parameter space:

$$I(\theta) = E\left\{\frac{\partial \log L(\theta; X)}{\partial \theta}\left[\frac{\partial \log L(\theta; X)}{\partial \theta}\right]^{\mathsf{T}}\right\}$$

- The quantity $S(\theta; X) = \partial_\theta \log L$ is known as the score function

The RCF bound

Answers question: How efficient can we be?

Theorem

Rao-Cramér-Frechet *(RCF) Assume:*

1. *The sample space of X is independent of θ.*
2. *The variance of $\hat{\theta}$ is finite, for any θ.*
3. $\partial_\theta \int_{-\infty}^{\infty} g(X) L(\theta; X) dX = \int_{-\infty}^{\infty} g(X) \partial_\theta L(\theta; X) dX$, *where $g(X)$ is any statistic of finite variance.*

Then the variance, $\sigma^2_{\hat{\theta}}$ of estimator $\hat{\theta}$ obeys the inequality:

$$\sigma^2_{\hat{\theta}} \geq \frac{[1 + \partial_\theta b(\theta)]^2}{I(\theta)}.$$

Proof: show that the information number is the variance of the score and consider the correlation coefficient between score and $\hat{\theta}$

The RCF bound – Case study (I)

Consider measuring CP violation via $B^0\bar{B}^0$ mixing at the $\Upsilon(4S)$. We measure the time difference, t, between the two B decays in an $\Upsilon \to B^0\bar{B}^0$ event. The sign of t is determined relative to the decay of a flavor "tag" B, e.g., a B decaying semileptonically. The pdf for this RV is:

$$f(t; A) = \frac{1}{2} e^{-|t|}(1 + A \sin rt),$$

where $t \in (-\infty, \infty)$, $r = \Delta m/\Gamma$ is a known quantity, and A is the CP asymmetry parameter of interest. In the early days there was some contorversy concerning the efficiency of a simple estimation method:

The simple analysis counts the number of times $t < 0$, N_-, and the number of times $t > 0$, N_+. The expectation value of the difference, for a total sample size $N = N_- + N_+$, is:

$$E(N_+ - N_-) = N \frac{rA}{1 + r^2}$$

The RCF bound – Case study (II)

- In the substitution method we replace the expectation value by the observed difference, and invert to obtain an estimator for the asymmetry parameter:

$$\hat{A}_{\rm subs} = d^{-1}\frac{N_+ - N_-}{N},$$

  where $d = r/(1+r^2)$ is known as the "dilution factor"
- $\hat{A}$ is by definition an unbiased estimator for A. The question is, how efficient is it? We are throwing away detailed time information – does that matter, assuming our time resolution isn't too bad?
- First, what is the variance of $\hat{A}$? For a given N, we may treat the sampling of $N_\pm$ as a binomial process, giving:

$$\mathrm{var}(\hat{A}_{\rm subs}) = (1 - d^2A^2)/Nd^2.$$

The RCF bound – Case study (III)

Second, how well can we do, at least in principle, if we do our best? Let's use the RCF bound to estimate this (and argue that, at least asymptotically, we can achieve this bound, e.g., with maximum likelihood estimation:
For N independent samplings, the RCF bound on the variance of any unbiased estimator for A is:

$$\mathrm{var}(\hat{A})_{\mathrm{RCF}} \geq \frac{1}{E\left\{\left[\frac{\partial}{\partial A}\sum_{i=1}^{N}\log f(t_i;A)\right]^2\right\}}$$

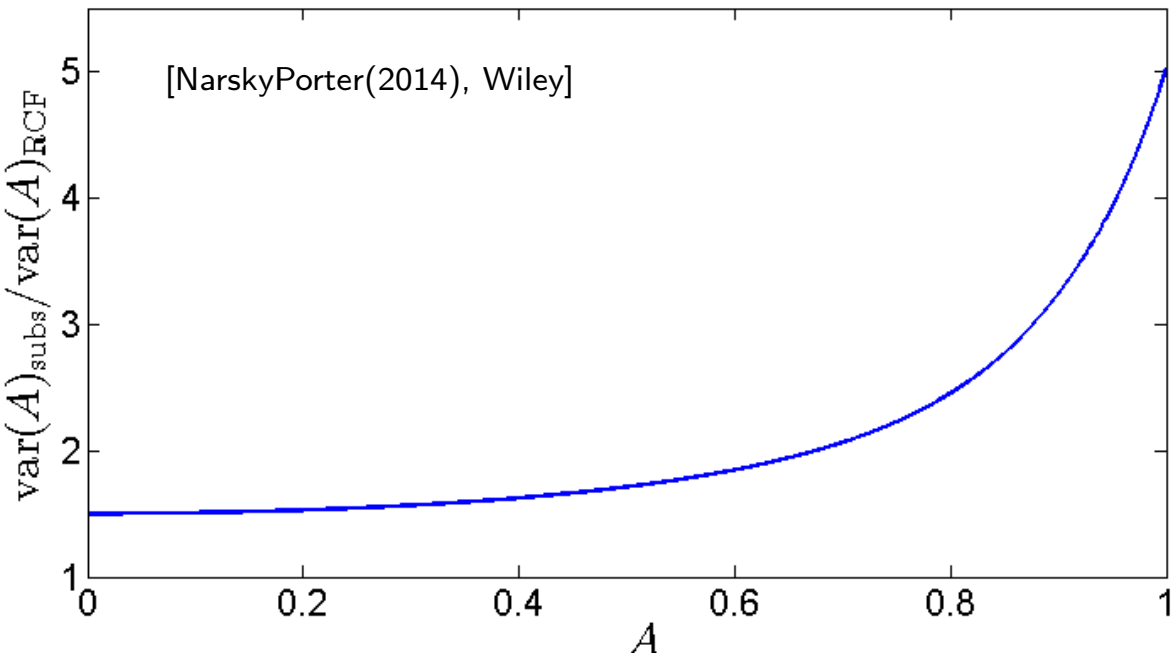
$$\geq \frac{1}{NE\left[\left(\frac{\sin rt}{1+A\sin rt}\right)^2\right]}.$$

Performing the integral gives:

$$\mathrm{var}(\hat{A})_{\mathrm{RCF}} = \frac{1}{N}\left\{\sum_{k=1}^{\infty} A^{2(k-1)}\frac{r^{2k}(2k)!}{[1+(2r)^2][1+(4r)^2]\cdots[1+(2kr)^2]}\right\}^{-1}.$$

The RCF bound – Case study (IV)

- Compare this bound with the variance from the substitution method (for $r = 0.7$)
- See that significant gains can be obtained using detailed time information



Point estimation – Desirable properties – Sufficiency

A statistic $\hat{\theta}$ is sufficient for θ if the sampling distribution for X conditioned on $\hat{\theta}$ is independent of θ
That is, $\hat{\theta}$ contains all of the information in the data with any relevance to θ
If you are not using a sufficient statistic, you may be able to do better by using some of the additional information

We saw that the median is a robust estimator of location.
However, it is usually far from sufficient.

Sufficiency – example

Consider sample size N from a $\mathrm{N}(\theta, 1)$ distribution:

$$f(x;\theta) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_n-\theta)^2}$$

If $N > 1$, then $\hat{\theta} = x_1$ is not sufficient for θ, since specifying x_1 still leaves the pdf for $x_2, \ldots, x_N$ depending on θ.
However, the sample mean is sufficient for θ. Let $m = \frac{1}{N}\sum_{n=1}^{N} x_n$ be the sample mean. Then we can write:

$$\sum_{n=1}^{N}(x_n - \theta)^2 = \sum_{n=1}^{N}(x_n - m)^2 + N(m-\theta)^2$$

Then

$$f(x|m;\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^{N-1} \exp\left[-\frac{1}{2}\sum_{n=1}^{N}(x_n - m)^2\right]$$

This is independent of θ, hence the sample mean is sufficient for θ

Point estimation – Desirable properties – Practical

Practical concerns generally arise in interval estimation and density estimation, but even in point estimation there may be an easy method that is "good enough"

Example:
Least-squares fitting of track parameters for very large data samples is best done with specialized code that knows a lot (in particular, derivatives wrt parameters) about tracks. This can be orders of magnitude more practical in computing time than just giving the problem to a general-purpose minimizer (such as MINUIT).

Point estimation – Substitution method

Have already encountered the substitution method in example of the use of the RCF bound. Let us now formalize it:
Consider pdf $f(X;\theta)$ for RV X, depending on unknown parameter θ. A statistic $U(X)$ has expectation value:

$$E(U) = \int u(x) f(x;\theta) dx = \phi_U(\theta)$$

Suppose we can invert ϕ_U:

$$\theta = \phi_U^{-1}\left[E(U)\right]$$

We use this to invent a plausible estimator for θ: Assume we have an iid sampling $(x_1, \ldots, x_N)$ from f. Compute the sample average

$$m_u = \frac{1}{N}\sum_{n=1}^{N} u(x_i)$$

Then form the estimator

$$\hat{\theta} = \phi_U^{-1}(m_u)$$

Substitution method – Comments

- The substitution method enjoys wide use as an easy method to apply
- As we have seen in the RCF example, it may not be efficient
- We can usually do better with least squares and maximum likelihood discussed below

The substitution method has been commonly used in obtaining estimates for parameters of angular distributions (may be called the "moment method" because we take the emprical moments of the distribution)
We'll see again when we discuss methods of density estimation

Point estimation – Least squares method

Consider a set of RVs $(X_1, \ldots, X_D)$ with expectation values

$$E(X_d) = g_d(\theta), \quad d = 1, \ldots, D$$

where $\theta = (\theta_1, \ldots, \theta_R)$ is an R-dimensional parameter vector. Suppose the covariance matrix, Σ, for X is known. Given a sampling $X = x$, the set of parameter values $\hat{\theta}$ which minimize the quantity

$$S = (x - g)^{\mathsf{T}} \Sigma^{-1} (x - g)$$

is called the Least Squares Estimate (LSE) for θ.

Sometimes we substitute an estimator for Σ, more on that later

Least squares method – Gauss Markov theorem

Theorem

Gauss-Markov *Consider the linear model for our observations:*

$$x_n = \sum_{r=1}^{R} \theta_r s_{rn} + \epsilon_n,$$

where s_{rn} is given, and the error ϵ_n is sampled from some distribution, not necessarily normal. If $E(\epsilon_n) = 0$ and $var(\epsilon_n) < \infty$, then the LSE for θ is unbiased and of minimum variances among all linear unbiased estimators.

- Proof left as an exercise
- This property of the LSE is sometimes denoted "BLUE", for "Best Linear Unbiased Estimator"

*Least squares method – Linear problem

Suppose the expectation values g_d for x_d are D linear functions of the R parameters θ:

$$E(x) = g = g_0 + F\theta,$$

where F is a matrix with D rows and R columns. It is convenient to translate the measurement vector by the constant vector g_0:

$$y = x - g_0$$

Then

$$S = (y - F\theta)^{\mathsf{T}} \Sigma^{-1} (y - F\theta).$$

We obtain $\hat{\theta}$, the values that minimize S by:

$$\left.\frac{\partial S}{\partial \theta_i}\right|_{\hat{\theta}} = 0.$$

Let $H \equiv F^T \Sigma^{-1} F$; this is an $R \times R$ matrix. Assuming H is non-singular, we obtain LSE estimator $\hat{\theta}$:

$$\hat{\theta} = H^{-1} F^T \Sigma^{-1} y$$

*Least squares method – Linear problem

As an exercise, show that:

- $E(\hat{\theta}) = \theta$, that is, our estimator is unbiased
- $\text{var}(\hat{\theta}) = H^{-1}$
- $S = (y - F\hat{\theta})^{\mathsf{T}} \Sigma^{-1} (y - F\hat{\theta}) + (\hat{\theta} - \theta)^{\mathsf{T}} H (\hat{\theta} - \theta)$

Suppose that our sampling distribution is multivariate normal:

$$f(y;\theta) = A \exp\left[-\frac{1}{2}(y - F\theta)^{\mathsf{T}} \Sigma^{-1} (y - F\theta)\right],$$

The pdf is thus of the form:

$$f(\hat{\theta}(y), y; \theta) = A \exp\left[-\frac{1}{2}(y - F\hat{\theta})^{\mathsf{T}} \Sigma^{-1} (y - F\hat{\theta})\right] \exp\left[-\frac{1}{2}(\hat{\theta} - \theta)^{\mathsf{T}} H (\hat{\theta} - \theta)\right]$$

- First part is original pdf, with θ replaced by the estimators $\hat{\theta}$,
- Second part is "correction term", taking into account that $\hat{\theta}$ may differ from θ.
- Have split the pdf into the probability that we observe $\hat{\theta}$, given θ, times the probability to observe y given a pdf with parameters $\hat{\theta}$

*Least squares method – Linear problem, χ^2

Let the quantity in the first exponential be denoted

$$\chi^2(y) = (y - F\hat{\theta})^{\mathsf{T}} \Sigma^{-1} (y - F\hat{\theta})$$

This is distributed according to the χ^2 distribution with $k = D - R$ degrees of freedom:

$$P(\chi^2; k) d\chi^2 = \frac{1}{\Gamma(k/2) 2^{k/2}} x^{k/2-1} e^{-\chi^2/2} d\chi^2$$

Useful in testing whether the data are consistent with the model

*Least squares method – Non-linear problem

In general, we are not lucky enough to have a linear problem. In this case:

1. See whether it is equivalent to a linear problem
2. If you don't need to do it often, plug it into a general-purpose minimizer
3. If you need to do it many times (e.g., track fitting or kinematic fitting), linearize the problem via a Taylor series expansion about some starting value for the parameters. The process is iterated until convergence is (hopefully) attained

*Non-linear problem – Linearized

The procedure in the third option is as follows: Expand in a Taylor series giving the expectation values about an initial guess for the parameter values:

$$g_d(\theta) = g_d(\theta^0) + \sum_{r=1}^{R} (\theta_r - \theta_r^0) \frac{\partial g_d}{\partial \theta_r}\Big|_{\theta^0} + \ldots$$

Try to pick a starting θ^0 near the value that minimizes S. Neglecting higher order terms, we have a problem of the form:

$$g(\theta) = g_0 + F\theta,$$

where

$$\begin{aligned} g_0 &= g(\theta^0) - F\theta^0, \\ F_{ij} &= \frac{\partial g_i}{\partial \theta_j}\Big|_{\theta^0}. \end{aligned}$$

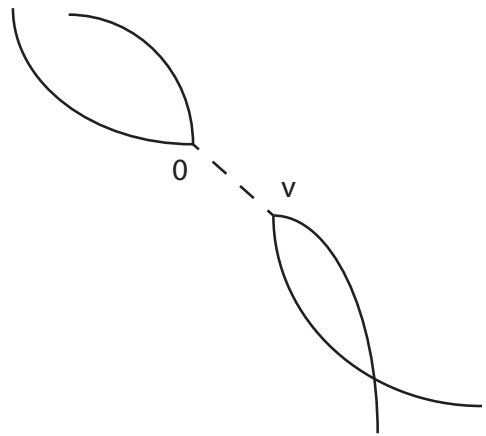
Solve this linear problem as already discussed. The solution may not be close enough to the minimum. In this case, re-expand about the new estimate and iterate. Iterate until convergence is achieved

*Least squares method – Constraints

Consider performing a kinematic fit to a set of tracks.

- Each track has measured values for the curvature, the tangent of the dip angle, and the azimuth angle
- The parameters θ are the four-momenta of the particles
- Four-momentum conservation may be imposed by eliminating some parameters
- Alternatively, impose conservation with constraint equations and Lagrange multipliers

- There may be additional features, such as detached vertices.
- Then we have additional relevant measurements on the trajectories
- We have additional parameters for the verticies as well
- Have additional constraints, eg, tracks sharing a common vertex and coplanarity



*Least squares method – Constraint formalism

Thus, write:

$$S = (x - g)^{\mathsf{T}} \Sigma^{-1} (x - g) + 2\lambda^{\mathsf{T}} c(g, u),$$

where c are equations of constraint, $c_k = 0, k = 1, \ldots, K$. The constraint equations could depend not only on g, but also on some M additional unknowns u. λ is a vector of K Lagrange multipliers. LSE is obtained by minimizing S with respect to g, u, and the Lagrange multipliers

If c is not linear in g and u, perform a linear approximation and iterate. Thus, assume:

$$c(g, u) = c_0 + G^T g + U^T u,$$

where G is a $K \times N$ matrix:

$$G_{ij} = \left. \frac{\partial c_j}{\partial g_i} \right|_{g^0, u^0}$$

and U is a $K \times M$ matrix:

$$U_{ij} = \left. \frac{\partial c_j}{\partial u_i} \right|_{g^0, u^0}$$

*Least squares method – Constraint formalism

Then

$$S = (x-g)^{\mathsf{T}}\Sigma^{-1}(x-g) + 2\lambda^{T}c_0 + 2\lambda^{\mathsf{T}}G^{\mathsf{T}}g + 2\lambda^{\mathsf{T}}U^{\mathsf{T}}u.$$

Setting the derivatives equal to zero with respect to g, u, and λ yields the equations:

$$\begin{aligned} 0 &= -\Sigma^{-1}(x-\hat{g}) + G\hat{\lambda} \\ 0 &= U\hat{\lambda} \\ 0 &= c_0 + G^{\mathsf{T}}\hat{g} + U^{\mathsf{T}}\hat{u}. \end{aligned}$$

Letting $E \equiv \Sigma G$ and $J \equiv UH^{-1}$, we solve these equations and express our estimators as:

$$\begin{aligned} \hat{u} &= -K^{-1}J(c_0 + G^{\mathsf{T}}x) \\ \hat{\lambda} &= H^{-1}(c_0 + G^{\mathsf{T}}x + U^{\mathsf{T}}\hat{u}) \\ \hat{g} &= x - E\hat{\lambda} \end{aligned}$$

*Least squares method – Checking the model

The χ^2 provides a single-number test for goodness of fit (more later).

Further information is available in the "pulls":
The "pulls" (or normalized residuals), are a handy way to tell whether the fit assumptions (e.g., M) are reasonable:

$$\text{pull}_d = \frac{x_d - g_d(\hat{\theta})}{\sqrt{\Sigma_{dd} - (FH^{-1}F^T)_{dd}}}.$$

If X is sampled from a normal distribution with mean $g(\theta)$ and covariance Σ, the pulls should be $\mathrm{N}(0,1)$ distributed (Exercise)

Point estimation – Maximum likelihood method

The maximum likelihood method consists of finding those values of θ for which the likelihood is maximized:

$$L(\hat{\theta}; x) = \max_{\theta} L(\theta; x)$$

Analytically, solve likelihood equations:

$$\left.\frac{\partial L(\theta; x)}{\partial \theta_r}\right|_{\theta=\hat{\theta}} = 0; \quad r = 1, \ldots, R$$

- Often intractable, and a numerical search is used
- Possible that likelihood equation has no solution within the domain of θ; then find the θ for which the likelihood achieves its maximum within its domain
- Value of θ for which the likelihood is maximized is the maximum likelihood estimator (MLE) for θ. Since $\hat{\theta} = \hat{\theta}(X)$, the MLE is a RV
- May be multiple solutions to the likelihood equation; called roots of the likelihood equation (RLE)
- Usually convenient to work with the logarithm of the likelihood, especially in numerical work

Point estimation – MLE and LSE

A connection can be made with the least squares method of parameter estimation, using the logarithm of the likelihood. If f is a (multivariate) normal distribution, the likelihood function for a single observation is of the form:

$$L(\theta; x) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp\left\{-\frac{1}{2}\left[x - g(\theta)\right]^T \Sigma^{-1} \left[x - g(\theta)\right]\right\},$$

where D is the dimension of X and $\Sigma = \mathrm{cov}(X)$. Take the logarithm and drop the constant (independent of θ) terms:

$$\log L(\theta; x) = -\frac{1}{2}\left[x - g(\theta)\right]^T \Sigma^{-1} \left[x - g(\theta)\right]$$

Thus, $-2 \log L$ is precisely the χ^2 expression in LSE. This connection is well-known. However, assumption of normal sampling is often forgotten. For non-normal distributions, the ML and LS procedures are distinct methods, yielding different estimators

Why is MLE good?

The popularity of the maximum likelihood method is based on several nice properties, as well as reasonably simple computation. In particular,

1. The maximum likelihood estimator is consistent and asymptotically ($N \to \infty$) efficient
2. If a sufficient statistic exists, the MLE is a function of the sufficient statistic
3. If an efficient unbiased estimator exists, the maximum likelihood algorithm will find it

See the textbooks for the fine print and proof

Why is MLE not great?

The nice asymptotic properties of the MLE may not hold for small statistics. For example, with $(x_1, \ldots, x_N)$ an iid sampling from a $N(\mu, \sigma^2)$ distribution, the MLEs for μ and σ^2 are

$$\hat{\mu} = \bar{x} \equiv \frac{1}{N} \sum_{n=1}^{N} x_n$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})^2.$$

In the first case, $\hat{\mu}$ is an unbiased estimator for μ for all values of $N > 0$. However, $\hat{\sigma}^2$ has a bias $b(\sigma^2) \equiv \langle \hat{\sigma}^2 \rangle - \sigma^2 = -\frac{1}{N}\sigma^2$. For small N, this bias can be very large. Fortunately, in this case it is easy to correct for, to obtain the familiar unbiased estimator $\frac{N}{N-1}\hat{\sigma}^2$.

MLE example

Sample N times from pdf: $\frac{1}{\theta}\exp(-X/\theta)$
The likelihood function is:

$$\mathcal{L}(\theta; x) = \prod_{n=1}^{N} \frac{1}{\theta}\exp(-x_n/\theta)$$

It is convenient to work with the negative logarithm,

$$-\ln \mathcal{L}(\theta; x) = N \ln \theta + \sum_{n=1}^{N} x_n/\theta$$

We minimize this with respect to θ:

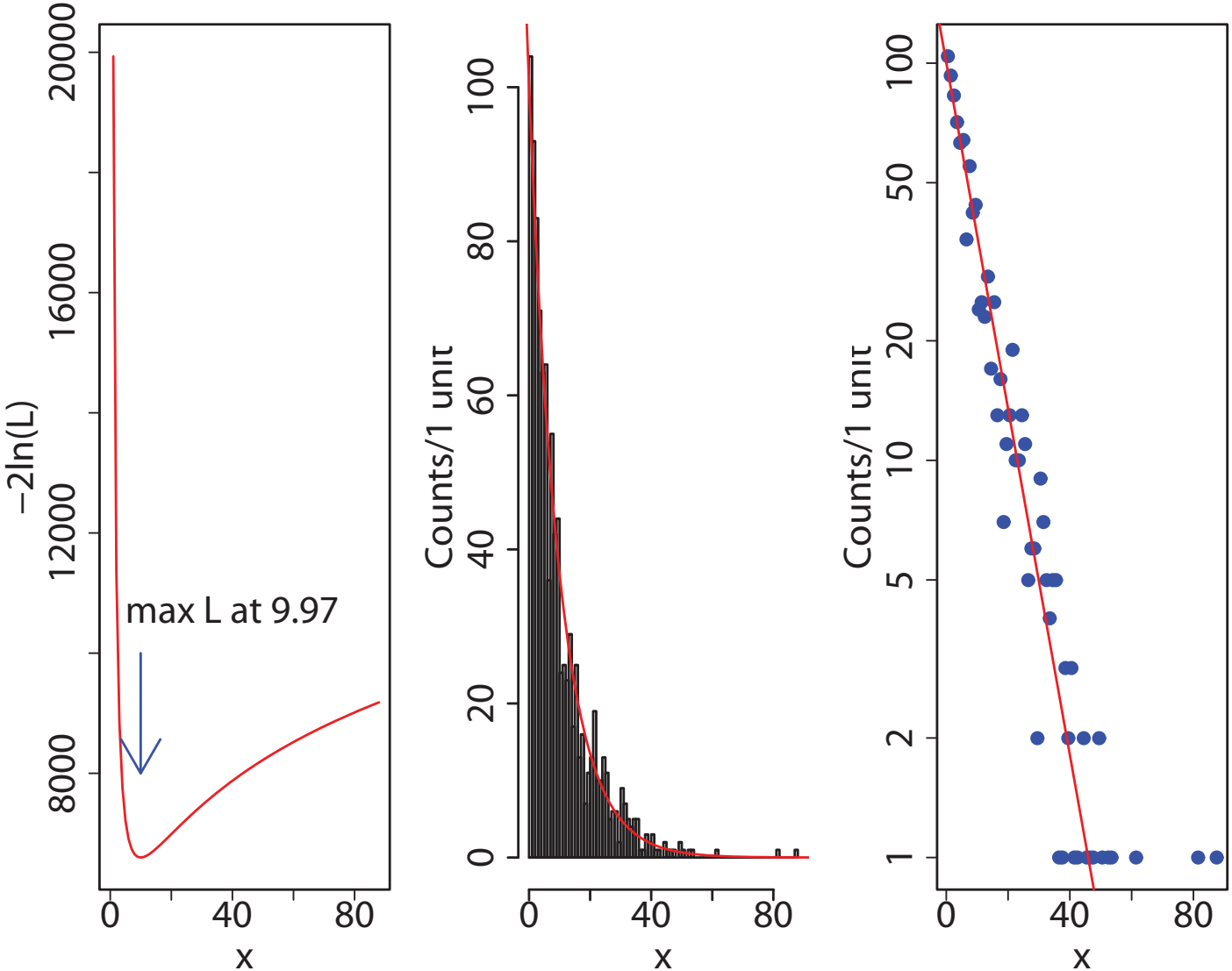
$$\left.\frac{d(-\ln \mathcal{L})}{d\theta}\right|_{\theta=\hat{\theta}} = 0$$

with the result,

$$\hat{\theta} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Numerical Maximum Likelihood Estimation

We may also do this numerically. Example for $N = 1000$ and $\theta = 10$:



The R code for the previous slide

```
# Example of exponential distribution - generate, then fit
n = 1000                    # Size of sample
theta = 10.                 # True mean
     # graphics
par(mfrow = c(1,3))        # For multiple plots
par(ps=28)                  # Font size
par(lwd=1.5)                # Line width
     # Generate samplings from exponential
x = -theta*log(runif(n))
sumx = sum(x)
     # - log-Likelihood function
lnlik <- function(par) {
     n*log(par) + sumx/par
}
     # Now do a fit
fit <- optimize(lnlik, interval=c(0,1000))
mltheta <- fit[[1]]        # ML estimator value
```

continued...

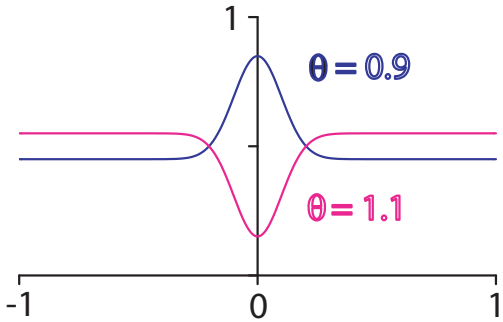
R code, continued

```
     # Now make some pictures; first plot the likelihood
plot(1:max(x+2)-1,2*lnlik(1:max(x+2)-1),ylab="-2ln(L)",
     type="l",xlab="x",main=NULL,col="red")
top = 10000
arrows(mltheta,top,mltheta,8000,col="blue")
text(mltheta-5,top+500,labels=paste("max L at",
     format(mltheta,digits=3)),adj=c(0,0))
     # Now a histogram of the data with a curve for the fit
hist(x,breaks=1:max(x+2)-1,main=NULL,
     ylab="Counts/1 unit",lwd=1.5)
curve((n/mltheta)*exp(-x/mltheta),add=TRUE,col="red")
     # Now show it on a log scale
histx <- hist(x,breaks=1:max(x+2)-1,plot=FALSE)
plot(histx[[5]],histx[[2]],log="y",main=NULL,xlab="x",
     ylab="Counts/1 unit",pch=19,cex=1.5,col="blue")
curve((n/mltheta)*exp(-x/mltheta),add=TRUE,col="red")
```

Hitting a Math Boundary
Technical issue, encountered at low statistics

Consider a ML fit of a set of events to some distribution, depending on parameters of interest

Example: $p(x;\theta) = \frac{\theta}{2} + \frac{1-\theta}{A\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, \; x \in (-1,1);$



NarskyPorter(2014), Wiley

$$\mathcal{L}(\theta; \{x_i, i = 1\ldots, N\}) = \prod_{i=1}^{N} p(x_i;\theta).$$

- Maximum wrt θ may be outside of region where pdf is defined
  - The function $p(x;\theta)$ may become negative in some regions of x
  - If there are no events in these regions, the likelihood is still "well-behaved"
  - The resulting fit, as a description of the data, will typically look poor even in region of positive pdf
  - Unacceptable (must stay within domain of θ)

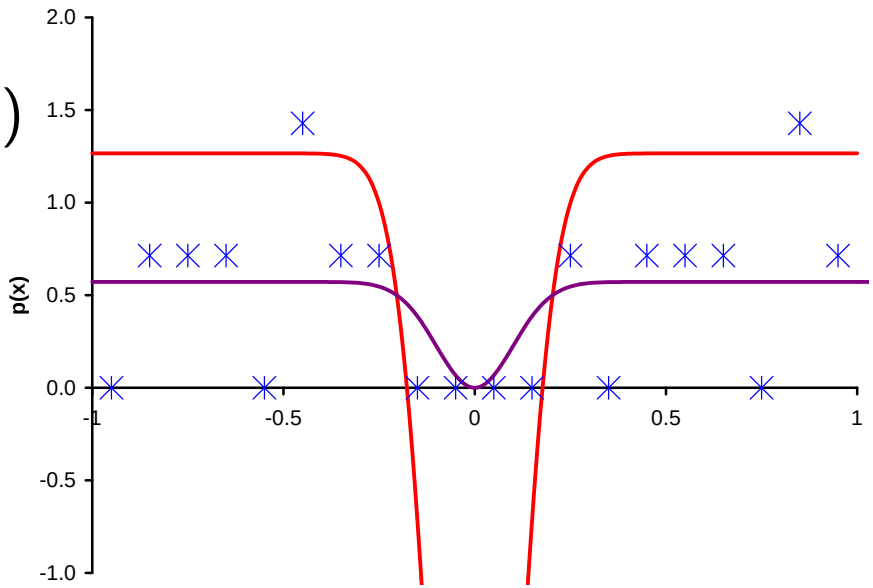
Hitting a Math Boundary (continued)

$$p(x;\theta) = \frac{\theta}{2} + \frac{1-\theta}{A\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, \; x \in (-1,1)$$

Points: Histogrammed "data".

Red: $p(x;\theta)$ allowed to go negative in ML fit

Purple: $p(x;\theta)$ constrained non-negative



NarskyPorter(2014), Wiley

- Resolution: Constrain fit to remain within bounds such that pdf is everywhere legitimate
  - n.b., parameters may still be "unphysical"
  - This gives fits which "look" like the data
  - Applies in interval evaluation also (but check coverage, as always)

Next: Interval Estimation