Parameter Estimation and Hypothesis Testing

Parameter estimation
Maximum likelihood
Least squares
Hypothesis tests
Goodness-of-fit

Elton S. Smith
Jefferson Lab

Con ayuda de
Eduardo Medinaceli
y Cristian Peña
<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>systematic</td>
<td>sistemático</td>
<td>spin</td>
<td>espin</td>
</tr>
<tr>
<td>background</td>
<td>trasfondo</td>
<td>scintillator</td>
<td>contador de centelleo</td>
</tr>
<tr>
<td>random</td>
<td>aleatorio (al azar)</td>
<td>scaler</td>
<td>escalar</td>
</tr>
<tr>
<td>histogram</td>
<td>histograma</td>
<td>degrees of freedom</td>
<td>grados de libertad</td>
</tr>
<tr>
<td>signal</td>
<td>senal</td>
<td>power</td>
<td>potencia</td>
</tr>
<tr>
<td>maximum likelihood</td>
<td>maxima verosimilitud</td>
<td>test statistic</td>
<td>prueba estadística</td>
</tr>
<tr>
<td>least squares</td>
<td>minimos cuadrados</td>
<td>goodness</td>
<td>bondad/fidelidad</td>
</tr>
<tr>
<td>chi-square</td>
<td>ji-cuadrado</td>
<td>estimation</td>
<td>estima</td>
</tr>
<tr>
<td>confidence intervals and limits</td>
<td>intervalos de confianza y limites</td>
<td>significance</td>
<td>significancia</td>
</tr>
<tr>
<td>frequentist</td>
<td>frecuentista?</td>
<td>conditional p</td>
<td>p condicional</td>
</tr>
<tr>
<td>nuisance parameters</td>
<td>parametros molestosos?</td>
<td>(A given B)</td>
<td>(A dado B)</td>
</tr>
<tr>
<td>outline</td>
<td>indice</td>
<td>fit</td>
<td>ajuste (encaje)</td>
</tr>
<tr>
<td>asymmetry</td>
<td>asimetria</td>
<td>parent distribution</td>
<td>distr del patron</td>
</tr>
<tr>
<td>variance</td>
<td>varianza</td>
<td>signal-to-background</td>
<td>fraction senal/trasfondo</td>
</tr>
<tr>
<td>root-mean-square</td>
<td>raiz quadrada media (rms)</td>
<td>sample</td>
<td>muestra</td>
</tr>
<tr>
<td>biased</td>
<td>sesgo</td>
<td>flip-flopping</td>
<td>indeciso</td>
</tr>
<tr>
<td>counting experiments</td>
<td>experimentos de conteo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>multiple scattering</td>
<td>dispersion multiple</td>
<td></td>
<td></td>
</tr>
<tr>
<td>correlations</td>
<td>correlaciones</td>
<td></td>
<td></td>
</tr>
<tr>
<td>weight</td>
<td>pesadas</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Parameter estimation

The parameters of a pdf are constants that characterize its shape, e.g.

$$f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}$$

r.v. \hspace{2cm} parameter

Suppose we have a sample of observed values: $\vec{x} = (x_1, \ldots, x_n)$

We want to find some function of the data to estimate the parameter(s):

$\hat{\theta}(\vec{x})$ \hspace{2cm} estimator written with a hat

Sometimes we say ‘estimator’ for the function of $x_1, \ldots, x_n$; ‘estimate’ for the value of the estimator with a particular data set.
Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:

\[ g(\hat{\theta}; \theta) \]

We want small (or zero) bias (systematic error):  
\[ b = E[\hat{\theta}] - \theta \]

→ average of repeated measurements should tend to true value.

And we want a small variance (statistical error):  
\[ V[\hat{\theta}] \]

→ small bias & variance are in general conflicting criteria
An estimator for the mean

Parameter: $\mu = E[x]$

Estimator: $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \equiv \bar{x}$ (‘sample mean’)

We find: $b = E[\hat{\mu}] - \mu = 0$

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

$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$
An estimator for the variance

Parameter: $\sigma^2 = V[x]$  

Estimator:  
$$ \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \equiv s^2 $$  

(‘sample variance’)  

We find:  

$$ b = E[\hat{\sigma}^2] - \sigma^2 = 0 \quad \text{(factor of } n-1 \text{ makes this so)} $$  

$$ V[\hat{\sigma}^2] = \frac{1}{n} \left( \mu_4 - \frac{n-3}{n-1} \mu_2 \right), \quad \text{where} $$  

$$ \mu_k = \int (x - \mu)^k f(x) \, dx $$
The Likelihood function via example

We have a data set given by N data pairs \((x_i, y_i \pm \sigma_i)\) graphically represented below.

The goal is to determine the fixed, but unknown, \(\mu = f(x)\). \(\sigma\) is known or estimated from the data.
Gaussian probabilities (least-squares)

We assume that at a fixed value of $x_i$, we have made a measurement $y_i$ and that the measurement was drawn from a Gaussian probability distribution with mean $y(x_i) = a + bx_i$ and variance $\sigma_i^2$.

$$f(y_i; a, b) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(y_i - y(x_i))^2}{2\sigma_i^2}}$$

$$y(x_i) = a + bx_i$$

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(y_i - a - bx_i)^2}{2\sigma_i^2}}$$

$$\chi^2 = -2\ln L + k = \sum_{i=1}^{N} \frac{(y_i - a - bx_i)^2}{\sigma_i^2}$$
$\chi^2$ minimization

\[ \frac{\partial \chi^2}{\partial a} = \sum_{i=1}^{N} \frac{-2}{\sigma_i^2} (y_i - a - bx_i) = 0 \]

\[ \frac{\partial \chi^2}{\partial b} = \sum_{i=1}^{N} \frac{-2x_i}{\sigma_i^2} (y_i - a - bx_i) = 0 \]

\[ a \left( \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \right) + b \left( \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \right) = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} \]

\[ a \left( \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \right) + b \left( \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \right) = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} \]
Solution for linear fit

For simplicity, assume constant $\sigma = \sigma_i$. Then solve two simultaneous equations for 2 unknowns:

$$a = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{N \sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - (\sum x_i)^2}$$

Parameter uncertainties can be estimated from the curvature of the $\chi^2$ function.

$$V[a] \approx 2 \left( \frac{\partial^2 \chi}{\partial a^2} \right)^{-1}_{\theta = \hat{\theta}} = \frac{\sigma^2 \sum x_i^2}{N \sum x_i^2 - (\sum x_i)^2}$$

$$V[b] \approx 2 \left( \frac{\partial^2 \chi}{\partial b^2} \right)^{-1}_{\theta = \hat{\theta}} = \frac{N \sigma^2}{N \sum x_i^2 - (\sum x_i)^2}$$
Parameter uncertainties

In a graphical method the uncertainty in the parameter estimator $\theta_0$ is obtained by changing $\chi^2$ by one unit.

$$\chi^2(\theta_0 \pm \sigma_{\theta}) = \chi^2(\theta_0) + 1$$

In general, using maximum likelihood

$$\ln L(\theta_0 \pm \sigma_{\theta}) = \ln L(\theta_0) - 1/2$$
The Likelihood function via example

What does the fit look like?

**Fit function** $y(x) = a + bx$
- $a = 4.97 \pm 0.34$
- $b = 1.09 \pm 0.06$

**Additional information about the fit:**
- $\chi^2$ and probability
- Were the assumptions for the fit valid?
  - We will return to this question after a discussion of hypothesis testing
More about the likelihood method

Recall likelihood for least squares:

\[ L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{(y_i-a-bx_i)^2}{2\sigma_i^2}} \]

But the probability density depends on application

\[ L = \prod_{i=1}^{N} f(y_i; \text{parameters}) \]

Proceed as before maximizing \( \ln L \) (\( \chi^2 \) has minus sign). The values of the estimated parameters might not be very different, but the uncertainties can be greatly affected.
Applications

\[ L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi \sigma_i}} e^{-\frac{(y_i - a - bx_i)^2}{2\sigma_i^2}} \]

\begin{align*}
\text{a) \quad} \sigma_i^2 &= \text{constant} \\
\text{b) \quad} \sigma_i^2 &= y_i \\
\text{c) \quad} \sigma_i^2 &= Y(x)
\end{align*}

Poisson distribution (see PDG Eq. 32.12)

\[ L = \prod_{i=1}^{N} \frac{(a - bx_i)^{n_i}}{n_i!} e^{-(a-bx_i)} \]

Stirling’s approx \( \ln(n!) \sim n \ln(n) - n \)

\[ -2\ln L = 2 \sum_{i=1}^{N} \left[ (a + bx_i) - n_i + n_i \ln\left[ n_i / (a + bx_i) \right] \right] \]
In addition to estimating parameters, one often wants to assess the validity of statements concerning the underlying distribution.

Hypothesis tests provide a rule for accepting or rejecting hypotheses depending on the outcome of an experiment. [Comparison of $H_0$ vs $H_1$]

In goodness-of-fit tests one gives the probability to obtain a level of incompatibility with a certain hypothesis that is greater than or equal to the level observed in the actual data. [How good is my assumed $H_0$?]

Formulation of the relevant question is critical to deciding how to answer it.
Selecting events

Suppose we have a data sample with two kinds of events, corresponding to hypotheses $H_0$ and $H_1$ and we want to select those of type $H_1$.

Each event is a point in $\vec{x}$ space. What ‘decision boundary’ should we use to accept/reject events as belonging to event types $H_0$ or $H_1$?

Perhaps select events with ‘cuts’:

\[ x_i < c_i \]
\[ x_j < c_j \]
Other ways to select events

Or maybe use some other sort of decision boundary:

linear

or nonlinear

How can we do this in an ‘optimal’ way?
Test statistics

Construct a ‘test statistic’ of lower dimension (e.g. scalar)

\[ t(x_1, \ldots, x_n) \]

Try to compactify data without losing ability to discriminate between hypotheses.

We can work out the pdfs \( g(t|H_0), \ g(t|H_1), \ldots \)

Decision boundary is now a single ‘cut’ on \( t \).

This effectively divides the sample space into two regions, where we accept or reject \( H_0 \).
Significance level and power of a test

Probability to reject $H_0$ if it is true (error of the 1st kind):

$$\alpha = \int_{t_{cut}}^{\infty} g(t|H_0) \, dt$$

(significance level)

Probability to accept $H_0$ if $H_1$ is true (error of the 2nd kind):

$$\beta = \int_{-\infty}^{t_{cut}} g(t|H_1) \, dt$$

$$1 - \beta = \text{power}$$
Efficiency of event selection

Probability to accept an event which is signal (signal efficiency):

$$\varepsilon_s = \int_{-\infty}^{t_{\text{cut}}} g(t|s) \, dt = 1 - \alpha$$

Probability to accept an event which is background (background efficiency):

$$\varepsilon_b = \int_{-\infty}^{t_{\text{cut}}} g(t|b) \, dt = \beta$$
Linear test statistic

Ansatz: \[ t(\vec{x}) = \sum_{i=1}^{n} a_i x_i \]

Choose the parameters \( a_1, \ldots, a_n \) so that the pdfs \( g(t|s), g(t|b) \) have maximum ‘separation’. We want:

large distance between mean values, small widths

\[ \rightarrow \text{Fisher: maximize } J(\vec{a}) = \frac{(\mu_s - \mu_b)^2}{\sigma_s^2 + \sigma_b^2} \]
Fisher discriminant

Using this definition of separation gives a Fisher discriminant.

Corresponds to a linear decision boundary.

Equivalent to Neyman-Pearson if the signal and background pdfs are multivariate Gaussian with equal covariances; otherwise not optimal, but still often a simple, practical solution.
Testing significance/goodness of fit

- Quantify the level of agreement between the data and a hypothesis without explicit reference to alternative hypotheses.
- This is done by defining a goodness-of-fit statistic, and the goodness-of-fit is quantified using the p-value.
- For the case when the $\chi^2$ is the goodness-of-fit statistic, then the p-value is given by

$$p = \int_{\chi^2_{\text{obs}}}^{\infty} f(\chi^2;n_d) d\chi^2$$

- The p-value is a function of the observed value $\chi^2_{\text{obs}}$ and is therefore itself a random variable. If the hypothesis used to compute the p-value is true, then p will be uniformly distributed between zero and one.
\( \chi^2 \) distribution

The chi-square pdf for the continuous r.v. \( z \) \( (z \geq 0) \) is defined by

\[
f(z; n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}
\]

\( n = 1, 2, \ldots \) = number of ‘degrees of freedom’ (dof)

\[ E[z] = n, \quad V[z] = 2n. \]

For independent Gaussian \( x_i, i = 1, \ldots, n \), means \( \mu_i \), variances \( \sigma_i^2 \),

\[
z = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}
\]

follows \( \chi^2 \) pdf with \( n \) dof.

Example: goodness-of-fit test variable especially in conjunction with method of least squares.
p-value for $\chi^2$ distribution

![Contour plot of $\chi^2$ distribution]

**Figure 32.1:** One minus the $\chi^2$ cumulative distribution, $1 - F(\chi^2; n)$, for $n$ degrees of freedom. This gives the $p$-value for the $\chi^2$ goodness-of-fit test as well as one minus the coverage probability for confidence regions (see Sec. 32.3.2.4).
Using the goodness-of-fit

Data generated using
\[ Y(x) = 6.0 + 0.2 \times, \sigma = 0.5 \]

Compare three different polynomial fits to the same data.
$\chi^2$/DOF vs degrees of freedom

**Figure 32.2:** The ‘reduced’ $\chi^2$, equal to $\chi^2/n$, for $n$ degrees of freedom. The curves show as a function of $n$ the $\chi^2/n$ that corresponds to a given $p$-value.
Summary of second lecture

- Parameter estimation
- Illustrated the method of maximum likelihood using the least squares assumption
- Reviewed how hypotheses are tested
- Use of goodness-of-fit statistics to determine the validity of underlying assumptions used to determine parent parameters
Constructing a test statistic

How can we select events in an ‘optimal way’?

Neyman-Pearson lemma (proof in Brandt Ch. 8) states:

To get the lowest $\varepsilon_b$ for a given $\varepsilon_s$ (highest power for a given significance level), choose acceptance region such that

$$\frac{f(x|s)}{f(x|b)} > c$$

where $c$ is a constant which determines $\varepsilon_s$.

Equivalently, optimal scalar test statistic is

$$t(x) = \frac{f(x|s)}{f(x|b)}$$