Introduction to Statistics and Error Analysis II

Physics 116C, 4/14/06

D. Pellett

References:
Data Reduction and Error Analysis for the Physical Sciences by Bevington and Robinson
Particle Data Group notes on probability and statistics, etc.—online at http://pdg.lbl.gov/2004/reviews/contents_sports.html
(Reference is S. Eidelman et al, Phys. Lett. B 592, 1 (2004))

Any original presentations copyright D. Pellett 2006
Topics

• Propagation of errors examples
  • Background subtraction
  • Error for \( x = a \ln(bu) \)
• Estimates of mean and error
  • method of maximum likelihood and error of the mean
  • Weighted mean, error of weighted mean
• Confidence intervals
  • Chi-square \((\chi^2)\) distribution and degrees of freedom \((\nu)\)
• Histogram \(\chi^2\)
• Format for lab writeup
• Next - least squares fit to a straight line and errors of parameters
Overview: Propagation of Errors

- Brief overview

- Suppose we have \( x = f(u, v) \) and \( u \) and \( v \) are uncorrelated random variables with Gaussian distributions.

- Expand \( f \) in a Taylor’s series around \( x_0 = f(u_0, v_0) \) where \( u_0, v_0 \) are the mean values of \( u \) and \( v \), keeping the lowest order terms:

\[
\Delta x \equiv x - x_0 = \left( \frac{\partial f}{\partial u} \right) \Delta u + \left( \frac{\partial f}{\partial v} \right) \Delta v
\]

- The distribution of \( \Delta x \) is a bivariate distribution in \( \Delta u \) and \( \Delta v \). Under suitable conditions (see Bevington Ch. 3) we can approximate \( \sigma_x \) (the standard deviation of \( \Delta x \)) by

\[
\sigma_x^2 \approx \left( \frac{\partial f}{\partial u} \right)^2 \sigma_u^2 + \left( \frac{\partial f}{\partial v} \right)^2 \sigma_v^2
\]
Bevington Problem 2.8 Solution

- Cars at fork in road: two choices, right or left; p=0.75 for right “on a typical day.”

- Binomial distribution 1035 cars

\[
P_B(x; 1035, 0.75) = \frac{1035!}{x!(1035-x)!} \left( \frac{3}{4} \right)^x \left( \frac{1}{4} \right)^{1035-x}
\]

- Mean = np = 1035x0.75 = 776.25, \( \sigma = \sqrt{npq} = 13.9 \)

- \( x = 752 \) cars turned right on a particular day

- Is this consistent with being a typical day?
  - Find probability for a deviation from the mean this large or larger, either positive or negative
  - Could sum \( P_B \) from 0 to 752 and from 800 to 1035
  - Can approximate with Gaussian since \( n, x \) large, \( x \) still many \( \sigma \) less than \( n \). (19\( \sigma \) to be exact.)
Bevington 2.8 Solution (continued)

- Use Gaussian approximation with same $\mu$, $\sigma$.
  - Transform interval to unit normal distribution in $z = (x - \mu)/\sigma$.
  - Limits of interval $\pm z$, $|z| = |752 - 776.25|/13.9 = 1.745$.
  - Probability for $z$ within this interval given in Table C.2 of Bevington (probability of being within $\pm 1.745\sigma$ of $\mu$):
    $P(-1.745 < z \leq 1.745) = 0.9190$
  - Probability that this is a typical day = $1 - P = 0.081$ (or 8.1%)
- Confidence interval: the above interval is a 91.9% confidence
  interval for $z$
  - $z$ should lie within this interval 91.9% of the time.
  - Larger probability interval required to exclude hypothesis
    - $\pm 3\sigma$ would correspond to a 99.73% confidence interval
  - Be careful: distribution might not be truly Gaussian
Bevington 2.8 (concluded)

- You could also use the cumulative distribution function to find $P$
- This is available in LabVIEW:

This calculates

$$P(z \leq -1.745) = \int_{-\infty}^{-1.745} \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} z^2 \right] = 0.0405$$

so the probability of this being a typical day is $2(0.0405)=0.081$
Chi-Square Test of a Distribution

- Define
  \[ \chi^2 \equiv \sum_{i=1}^{N} \left( \frac{x_i - \mu_i}{\sigma_i} \right)^2 \]

- For Gaussian random variables \( x_i \), it can be shown that this follows the chi-square distribution with \( \nu \) degrees of freedom.

- The expectation value (i.e., mean value) for \( \chi^2 \) is \( \nu \)

- If this is based on a fit where some parameters are determined from the fit, the number of degrees of freedom is reduced by the number of fitted parameters.

- One can set confidence intervals in \( \chi^2 \) to evaluate the goodness of fit.

- \( \chi^2 \) distributions can be found from tables, graphs or LabVIEW VI’s as was done for Gaussian probability. (See Table C.4 in Bevington)

- Can be approximated by Gaussian under suitable conditions.
Comments on $\chi^2$ of Histogram

$$\chi^2 \equiv \sum_{i=1}^{N} \frac{(N_i - n_i)^2}{n_i}$$

- Can model this with $N$ individual mutually independent binomials, so long as a fixed total is not required. Then normalize to the actual $N_{\text{total}}$, using 1 degree of freedom in the fit ($N = \text{number of bins}$) (see discussion in Bevington, Ch. 3 and Prob. 4.13 solution on next page)

- For small $n_i$, large $N_{\text{total}}$, large number of bins, $n_i$ is approx. Poisson

- $\chi^2 \approx \chi^2$ distribution if all $n_i >> 1$ ($n_i \gtrsim 5$ OK) or $N >> 1$

- For fixed $N_{\text{total}}$, model with multinomial distribution (see p. 12)

- But $n_i$ are not mutually independent with multinomial
Histogram Chi-Square: Bevington Prob. 4.13

4.13) I made a LabVIEW VI (see figures on next page) to plot the histogram, calculate the Gaussian comparison values and find $\chi^2$ according to Eq. 4.33 of Bevington:

$$\chi^2 = \sum_{j=1}^{n} \frac{[h(x_j) - NP(x_j)]^2}{NP(x_j)}$$

where $n$ is the number of bins, $N$ is the total number of trials, $h(x_j)$ is the contents of the $j$th bin and $NP(x_j)$ is the expected contents from the Gaussian distribution (see the text for details).

Assume the bins are small enough so we can approximate the integral of the p.d.f. over the bin with the central value of the p.d.f. times the bin width. Then

$$NP(x_j) = N \int_{\Delta x_j} p(x) dx \approx Np(x_j)\Delta x$$

where $p(x_j)$ is the Gaussian p.d.f. evaluated at the center of the $j$th bin and $\Delta x = 2$ is the bin width.

$$p(x_j) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x_j - \mu}{\sigma}\right)^2\right]$$
Histogram Chi-Square Result

The VI front panel with the results and the diagram which produced them are shown in the following figures.

Population and Sample Distributions

Bevington
Prob. 4.13

Gaussian distribution

Pop Mean
26

Pop Std Dev
5

Number of Samples
200

Sample Histogram
12

Chi-Square
8.2824

Deg of freedom
12

Probability of exceeding this Chi-Square
0.7626
Histogram Chi-Square: Comments

This analysis assumes the contents of each histogram bin \( h_j \) is an independent measurement. \( \mu \) and \( \sigma \) are given but the total number of trials is taken to agree with the experiment (200 trials). This represents one constraint and reduces the number of degrees of freedom by 1. We have 13 bins to compare with the Gaussian and \( \nu = 12 \) degrees of freedom.

The expectation value for \( \chi^2 \) equals the number of degrees of freedom, \( < \chi^2 > = 12 \). The resulting \( \chi^2 = 8.28 \) (disagrees with the answer in the book but was checked independently). The probability for exceeding this value of \( \chi^2 \) is 0.76 (calculated by LabVIEW but agrees with interpolated value from Table C.4). The reduced \( \chi^2 \equiv \frac{<\chi^2>}{\nu} = 0.69 \) (expectation value of 1). This is not a bad fit. Of course, the \( \chi^2 \) distribution is only valid for underlying Gaussians and this is not a good assumption for the bins with low occupancy.
Multinomial Distribution

Histogram with $n$ bins, $N$ total counts (partition $N$ events into $n$ bins), $x_i$ counts in $i^{th}$ bin, with probability $p_i$ to get a count in $i^{th}$ bin

$$P(x_1, x_2, \ldots, x_n) = \frac{N!}{x_1!x_2!\ldots x_n!} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n}$$

with

$$\sum_{i=1}^{n} x_i = N, \quad \sum_{i=1}^{n} p_i = 1.$$

Then

$$\mu_i = Np_i, \quad \sigma_i^2 = Np_i(1 - p_i), \quad \sigma_{ij}^2 = -Np_ip_j$$

“Complete” Lab Writeup

- Similar to research report. Outline as follows:
  - Abstract (very brief overview stating results)
  - Introduction (Overview and theory related to experiment)
  - Experimental setup and procedure
  - Analysis of data and results with errors
    - Graphs should have axes labeled with units, usually points should have error bars and the graph should have a caption explaining briefly what is plotted
  - Comparison with accepted values, discussion of results and errors; conclusions, if any.
  - References
  - Have draft/outline of paper and preliminary calculations at lab time Wednesday