Summary of Statistics and Error Analysis

PHY 575/675 Spring 2012
Tanja Horn

I. STATISTICS

A series of measurements can be taken to be a representation of a parent population that describes the distribution of events. The parent population is the distribution seen when the number of measurements goes to infinity.

In a graph of the frequency of occurances of a measurements x, the parent distribution is represented by the mean,

$$\mu = \lim_{N\to+\infty} \frac{1}{N}\sum_{i=1} x_i, \tag{1}$$

and the variance,

$$\sigma^2 = \lim_{N\to+\infty} \frac{1}{N}\sum_{i=1} (x_i - \bar{x})^2., \tag{2}$$

For a finite series of measurements, the measured distribution is represented by the sample mean,

$$\bar{x} = \frac{1}{N}\sum_{i=1} x_i, \tag{3}$$

and the sample variance,

$$s^2 = \frac{1}{N-1}\sum_{i=1} (x_i - \bar{x})^2., \tag{4}$$

We will often work under the assumption that the sample distribution properly represents the parent distribution and thus use μ and σ instead of x and s.

A. Types of Parent Distributions

1. BINOMIAL

The binomial distribution is typically used when the total number of possible outcomes is small and its probability function is,

$$P_B(x,n,p) = \frac{n!}{x!(n-x)!} p^x q^{n-1}, \tag{5}$$

where p is the probability of success, $q = 1-p$ is the probability of failure, n denotes the number of attempts, and x the number of successes. The mean and variance are given by,

$$\mu = np \tag{6}$$

and

$$\sigma^2 = \mu(1-p). \tag{7}$$

2. POISSON

The Poisson distribution applies when the total number of possible outcomes large with few successes, i.e., $p << 1$, $x << 1$, and $\mu << n$) and its probability function is,

$$P_P(x,\mu) = \frac{\mu^x}{x!}e^{-\mu}, \tag{8}$$

where x is an integer, for instance, the number of counts detected over a certain time interval. The relation between mean and variance is given by,

$$\sigma^2 = \mu. \tag{9}$$

3. GAUSSIAN

The Gaussian distribution applies when we have a very large number of trials, and p is not too small, i.e., $n \to +\infty$ and $np >> 1$) and its probability function is,

$$P_G(x,\mu) = \frac{1}{\sigma\sqrt{2\pi}}exp(-\frac{(x_i-\mu)^2}{2\sigma^2}), \tag{10}$$

where

$$\int_{-\sigma}^{\sigma} P(x)dx = 0.68 \tag{11}$$

and

$$\int_{-2\sigma}^{+2\sigma} P(x)dx = 0.95 \tag{12}$$

The Gaussian distribution is symmetric while the Poisson distribution is not necessarily symmetric. The Poisson distribution looks Gaussian when x is greater than about 10.

II. TYPES OF ERRORS

Measurements errors can be categorized into the following categories:

- Random Instrumental: limitations coming from the apparatus (resolution, reproducibility)
- Random Statistical: in counting experiments due to finite number of counts in a random process
- Systematic Instrumental: gain shifts in apparatus, overall scale uncertainties

III. PROPAGATION OF ERRORS

Let $x = f(u, v, w, ...)$, where u, v, w are the measured quantities with variances σ_u^2, σ_v^2, and σ_w^2, and x is the quantity to be determined (dependent). To determine the variance in x, assume that x is linear over a small range of $(u, v, w...)$, and perform a Taylor series expansion to first order. Also assume tht the mean value of x is defined to be at the mean values of the measured quantities. Then,

$$\sigma_x^2 = (x_i - x)^2 = \frac{1}{N}\sum_{i=1}(u_i - \bar{u})^2(\frac{\partial^2 f}{\partial u^2}) + (v_i - \bar{v})^2(\frac{\partial^2 f}{\partial v^2}) + (w_i - \bar{w})^2(\frac{\partial^2 f}{\partial w^2}) + \frac{2}{N}\sum_{i=1}(u_i - \bar{u})(v_i - \bar{v})(\frac{\partial f}{\partial u})(\frac{\partial f}{\partial v}) + ... \tag{13}$$

or

$$\sigma_x^2 = \sigma_u^2(\frac{\partial f}{\partial u})^2 + \sigma_v^2(\frac{\partial f}{\partial v})^2 + \sigma_w^2(\frac{\partial f}{\partial w})^2 + \sigma_{uv}^2(\frac{\partial f}{\partial u})(\frac{\partial f}{\partial v}) + ... \tag{14}$$

The covariance σ_{uv}^2 is zero if u and v are independent.

IV. COMBINING MEASUREMENTS

If we have a series of measurements x_i of the same observable, say the ratio $\frac{e}{m}$, where e is the electron charge and m is the electron mass, each with an error σ_i, then we assume that the measurements come from a Gaussian parent distribution with mean value μ. The probability of measuring a particular result x_i with error σ_i is

$$P(x_i, \mu) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(x_i - \mu)^2}{2\sigma_i^2}) \tag{15}$$

and the probability of seeing a given distribution of events is

$$P = \prod_{i=1}^{N} P(x_i, \mu) = \frac{1}{(2\pi)^{N/2}\sigma_1\sigma_2} exp(-\frac{1}{2}\sum_{i=1}\frac{(x_i - \mu)^2}{\sigma_i^2}) \tag{16}$$

The maximum likelihood method states that the best value of μ is when the probability is a maximum, which happens when the exponent is minimized. Let us define,

$$\chi^2 = \sum_{i=1}^{N}(\frac{x_i - \mu}{\sigma_i})^2 \tag{17}$$

which is minimized when

$$\mu = \sum_{i=1}^{N}\frac{x_i/\sigma_i^2}{1/\sigma_i^2} \tag{18}$$

and then

$$\sigma_\mu^2 = \frac{1}{\sum_{i=1}^{N} 1/\sigma_i^2} \tag{19}$$

If all the errors are the same, then $\sigma_\mu^2 = \sigma^2/N$.

V. LINEAR LEAST SQUARES FIT

If the physical quantities of interest have an linear relationship, e.g., $y = ax+b$, then using the same logic regarding the Gaussian parent distribution as above, the probability of obtaining a series of pairs of measurements (x_i, y_i) distributed in a certain way is

$$P = \prod_{i=1}^{N} P(x_i, \mu) = \frac{1}{(2\pi)^{N/2}\sigma_1\sigma_2} exp(-\frac{1}{2}\sum_{i=1}\frac{(y_i - y(x_i))^2}{\sigma_i^2}) \tag{20}$$

and minimizing the exponent again finds the best linear relationship for $y(x)$. As long as the function $y(x)$ is linear in the parameters (a, b), then the minimization can be done analytically using matrix inversion. For example, if we now minimize χ^2 simultaneously with respect to the two parameters in a linear fit, then

$$\frac{\partial\chi^2}{\partial a} = 0 \rightarrow (\sum\frac{1}{\sigma_i^2})a + (\sum_i\frac{x_i}{\sigma_i^2})b = (\sum_i\frac{y_i}{\sigma_i^2}) \tag{21}$$

$$\frac{\partial\chi^2}{\partial b} = 0 \rightarrow (\sum\frac{x_i}{\sigma_i^2})a + (\sum_i\frac{x_i^2}{\sigma_i^2})b = (\sum_i\frac{x_iy_i}{\sigma_i^2}) \tag{22}$$

These can be solved for the parameters a and b. The variance in the parameters are determined by taking the derivative in the parameters with respect to the data points.

The same procedure holds for higher order matrices as described in, for instance, Bevington chapters 6 and 7, or the program LINREG. If the function if not linear in the parameters then different techniques must be used. Several alternative fitting algorithms are available as discussed in, for instance, Bevington chapter 8. Mathematica has a nonlinear regression routine that can be used, for example, to fit peaks on top of a continous background.

If the parent distribution is not Gaussian but follows a Poisson distribution, then the exponent in the product probability will probably be different. It can, however, also be minimized. The expression for χ^2 will be different. See Bevington chapter 6.6 for a discussion of the use of Poisson statistics.

VI. USEFUL RELATIONS FOR DATA ANALYSIS

When

$$f(u, v, ...) = au \pm bv \tag{23}$$

the errors add absolutely and we have

$$\sigma_f^2 = a^2\sigma_u^2 + b^2\sigma_u^2 \pm 2ab\sigma_{uv}^2 \tag{24}$$

When

$$f(u, v, ...) = \pm auvor \pm au/v \tag{25}$$

the errors add relatively and we have

$$\frac{\sigma_f^2}{f^2} = \frac{\sigma_u^2}{u^2} + \frac{\sigma_v^2}{v^2} \pm 2\frac{\sigma_{uv}^2}{uv} \tag{26}$$

where the + sign denotes multiplication and the − sign division. Note that Bevington page 50 includes more expressions that may come in handy.