

## Lecture 6

### Chi Square Distribution ( $\chi^2$ ) and Least Squares Fitting

#### 1) Chi Square Distribution ( $\chi^2$ ):

(See Taylor P. 181-197, App. B or Barlow P. 106-8 and 150-3)

Suppose we have a set of measurements  $(x_1, x_2, \dots, x_n)$  and suppose we know the true value of each  $x_i$  ( $x_{t1}, x_{t2}, \dots, x_{tm}$ ).

We would like some way to measure how good these measurements really are. Obviously the closer the  $(x_1, x_2, \dots, x_n)$ 's are to the  $(x_{t1}, x_{t2}, \dots, x_{tm})$ 's the better (or more accurate) the measurements. But can we get more specific?

Let's further assume that the measurements are independent of each other and that they come from a Gaussian distribution. Let  $(\sigma_1, \sigma_2, \dots, \sigma_n)$  be the standard deviation associated with each measurement.

Consider the following two possible measures of the quality of the data:

$$R \equiv \sum_{i=1}^n \frac{x_i - x_{ti}}{\sigma_i}$$

$$\chi^2 \equiv \sum_{i=1}^n \frac{(x_i - x_{ti})^2}{\sigma_i^2}$$

Which of the above gives more information on the quality of the data?

Both  $R$  and  $\chi^2$  are zero if the measurements agree with the true value. In addition, the measure  $R$  looks good because via the Central Limit Theorem as  $n \rightarrow \infty$  the sum  $\rightarrow$  Gaussian. However,  $\chi^2$  is better! One can show that the probability distribution for  $\chi^2$  is exactly:

$$p(\chi^2, n) = \frac{1}{2^{n/2} \Gamma(n/2)} [\chi^2]^{n/2-1} e^{-\chi^2/2} \quad 0 \leq \chi^2 \leq \infty$$

This is a continuous probability distribution that is a function of two variables,  $\chi^2$  and  $n =$  number of degrees of freedom. Here  $\Gamma$  is the "Gamma Function". We call this the "Chi Square" ( $\chi^2$ ) distribution.

A few words about the number of degrees of freedom  $n$ :

$$n = \text{\#data points} - \text{\# of parameters calculated from the data points}$$

For example, you collected  $N$  events in an experiment and you histogram your data in  $n$  bins before performing the fit, then you have  $n$  data points!

#### EXAMPLE:

Assume we have 10 data points and let  $\mu$  and  $\sigma$  be the mean and standard deviation of the data set.

If we calculate  $\mu$  and  $\sigma$  from the 10 data point then  $n = 8$

If we know  $\mu$  and calculate  $\sigma$  then  $n = 9$

If we know  $\sigma$  and calculate  $\mu$  then  $n = 9$

If we know  $\mu$  and  $\sigma$  then  $n = 10$

Like the situation with the Gaussian probability distribution the probability integral cannot be done in closed form and we must resort to a table to find out the probability of exceeding a given  $\chi^2$ .

$$P(\chi^2 > a) = \int_a^{\infty} p(\chi^2, n) d\chi^2 = \int_a^{\infty} \frac{1}{2^{n/2} \Gamma(n/2)} [\chi^2]^{n/2-1} e^{-\chi^2/2} d\chi^2$$

EXAMPLE: What's the probability to have  $\chi^2 > 10$  with the number of degrees of freedom  $n = 4$ ?

Using the table we find  $P(\chi^2 > 10, n = 4) = 0.04$ .

Thus we say that the probability of getting a  $\chi^2 > 10$  with 4 degrees of freedom by chance is 4%.

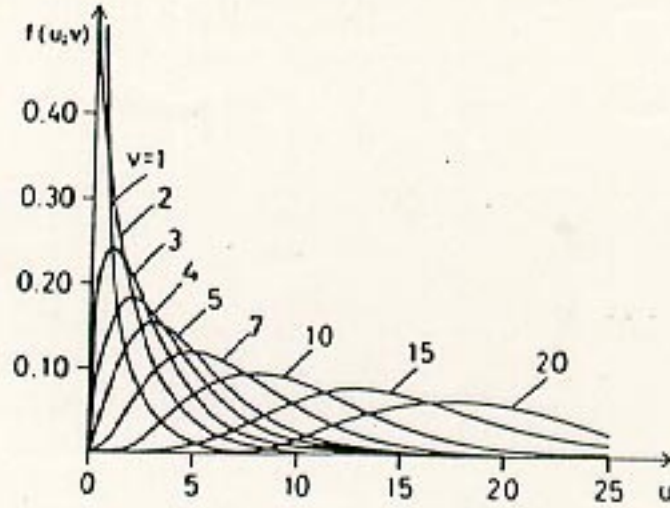
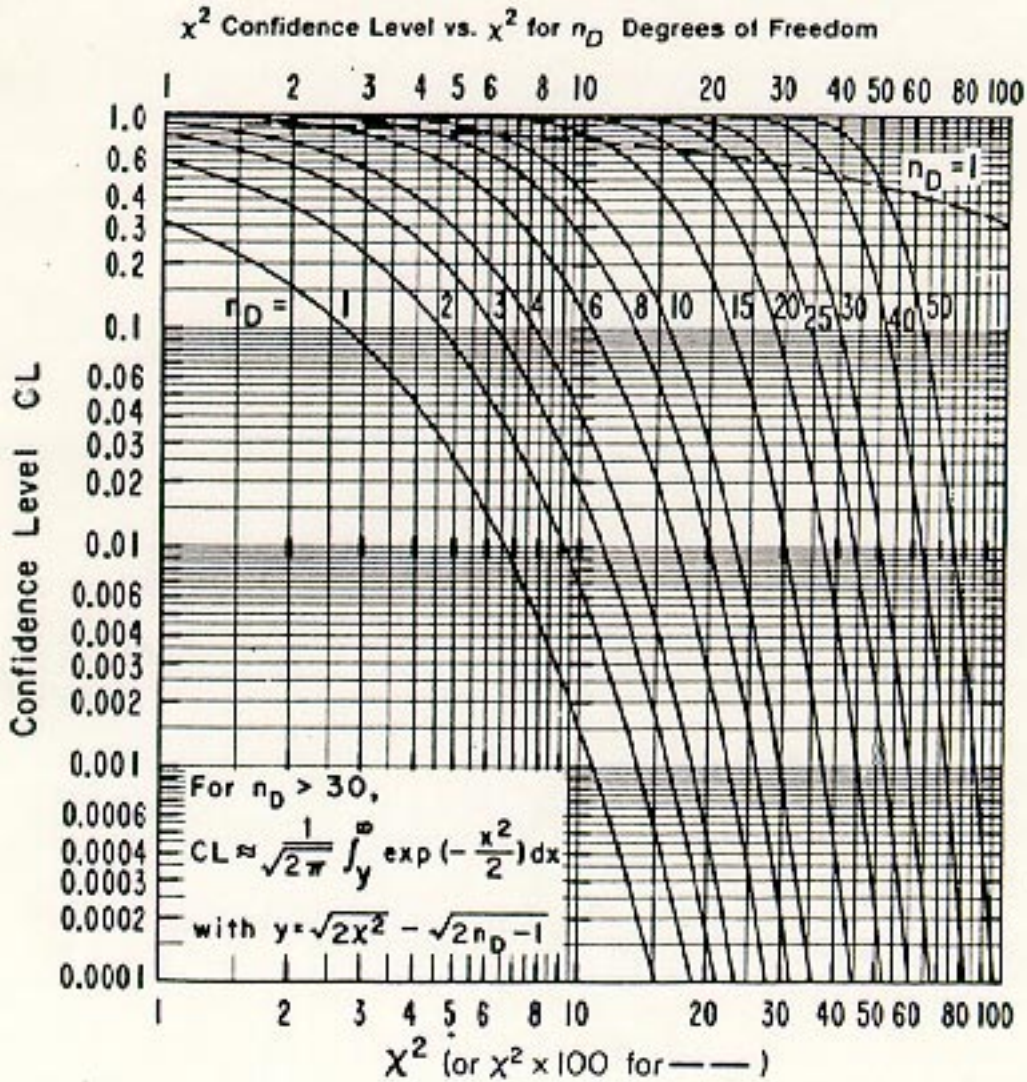


Fig. 5.1. The chi-square distribution for different degrees of freedom  $v$ .



Some not so nice things about the  $\chi^2$  distribution:

- a) Given a set of data points two different functions can have the same value of  $\chi^2$ .  
 b) The distribution does not look at the order of the data points, ignores trends in the data points, and ignores the sign of differences between the data points and “true” values.

There are other distributions/statistical test that do use the order of the points:  
 “run tests” and “Kolmogorov test”

## 2) Least Squares Fitting:

(See Taylor Chapter 8 or Barlow Chapter 6)

Suppose we have  $n$  data points  $(y_i, \sigma_i)$  and we believe that we know a functional relationship between the points, i.e.

$$y = f(x, a, b \dots).$$

Assume that for each  $y_i$  we know  $x_i$  exactly. The parameters  $a, b \dots$  are constants that we wish to determine. A procedure to obtain  $a$  and  $b$  is to minimize the following  $\chi^2$  with respect to  $a$  and  $b$ . Note this is very similar to the Maximum Likelihood Method. For the Gaussian case MLM and LS are identical.

$$\chi^2 = \sum_{i=1}^n \frac{[y_i - f(x_i, a, b)]^2}{\sigma_i^2}$$

Technically this is a  $\chi^2$  distribution only if the  $y$ 's are from a Gaussian distribution. Since most of the times the  $y$ 's are not from a Gaussian distribution we speak of “least squares” rather than  $\chi^2$ .

EXAMPLE:  $f(x, b) = 1 + bx$

We have a function with one unknown parameter. Find  $b$  using the least squares technique.

We need to minimize the following:

$$\chi^2 = \sum_{i=1}^n \frac{[y_i - f(x_i, a, b)]^2}{\sigma_i^2} = \sum_{i=1}^n \frac{[y_i - 1 - bx_i]^2}{\sigma_i^2}$$

To find the  $b$  that minimizes the above function, we do the following:

$$\frac{\partial \chi^2}{\partial b} = 0 = \frac{\partial}{\partial b} \sum_{i=1}^n \frac{[y_i - 1 - bx_i]^2}{\sigma_i^2} = \sum_{i=1}^n \frac{-2[y_i - 1 - bx_i]x_i}{\sigma_i^2}$$

$$\frac{\partial \chi^2}{\partial b} = 0 = \sum_{i=1}^n \frac{y_i x_i}{\sigma_i^2} - \sum_{i=1}^n \frac{x_i}{\sigma_i^2} - \sum_{i=1}^n \frac{bx_i^2}{\sigma_i^2}$$

solving for  $b$  we obtain:

$$b = \frac{\sum_{i=1}^n \frac{y_i x_i}{\sigma_i^2} - \sum_{i=1}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{x_i^2}{\sigma_i^2}}$$

Note: Here each measured data point ( $y_i$ ) is allowed to have a different standard deviation ( $\sigma_i$ ).

The LS technique can be generalized to two or more parameters for simple and complicated (e.g. non-linear) functions. One especially nice case is a polynomial function that is linear in the unknowns, e.g.:

$$f(x, a_1 \dots a_n) = a_1 + a_2 x + a_3 x^2 + a_n x^{n-1}.$$

In this case we can always recast problem in terms of solving  $n$  simultaneous linear equations. Thus we use the techniques from linear algebra and wind up inverting an  $n \times n$  matrix to solve this problem!

Example: Given the following data perform a least squares fit to find the value of  $b$ .

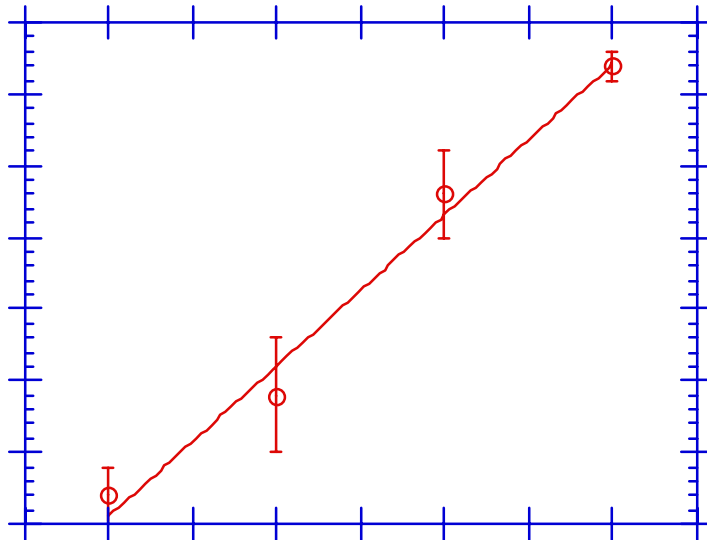
$$f(x, b) = 1 + bx$$

$x$	$y$	$\sigma$
1.0	2.2	0.2
2.0	2.9	0.4
3.0	4.3	0.3
4.0	5.2	0.1

Using the above expression for  $b$  we calculate:

$$b = 1.05$$

Below is a plot of the data points and the line from the least squares fit.



If we assume that the data points are from a Gaussian distribution then we can calculate a  $\chi^2$  and the probability associated with the fit.

For this example there are 3 degrees of freedom and  $\chi^2$  is:

$$\chi^2 = \sum_{i=1}^n \frac{[y_i - 1 - 1.05x_i]^2}{\sigma_i^2} = \left(\frac{2.2 - 2.05}{0.2}\right)^2 + \left(\frac{2.9 - 3.1}{0.4}\right)^2 + \left(\frac{4.3 - 4.16}{0.3}\right)^2 + \left(\frac{5.2 - 5.2}{0.1}\right)^2 = 1.04$$

The probability to get  $\chi^2 > 1.04$  for 3 degrees of freedom is 80% (read it off the enclosed chart or get a table of  $\chi^2$  values, the book by Bevington contains such a table). For this example the probability is close to 100% so we would call this a "good" fit. If, however the  $\chi^2$  was large (e.g. 15) then the corresponding probability would be small (< 0.2% for 3 dof) and we'd say this was a "bad" fit.

## RULE OF THUMB

A "good" fit has  $\chi^2 / \text{dof} \approx 1$