

## Chapter 3

# Data, Statistics, and Parameter Estimation

When an experiment is undertaken and data are collected, it is the duty of any scientific observer to apply objectivity to the analysis and interpretation to the highest degree. An experiment may be founded upon a clear hypothesis or hypotheses, but the process of observation is plagued by uncertainties in both the underlying physical process generating the *true* distribution of the data and by the known (and sometimes unknown) limitations of the measurement processes that modify the *observed* distribution of the data. This experimental “filter” through which the data become an observed quantity is called the sensitivity function, which may include the relative efficiencies of intergalactic or interstellar scattering, atmospheric transparency, instrumental reflecting elements, filters, detectors, and electronic responses.

Ultimately, objectivity means that the observer must distill through the experimentally induced limitations to get at the true unknown underlying distribution of the data. In the end, the observer can never be 100% certain that the observed distribution is the true distribution. The best that can be done is to assume an underlying true distribution and then apply a statistical treatment to the observed data that incorporates treatment of the sensitivity function to yield a confidence level (percent probability) that the observed distribution describes the underlying true distribution.

The practice of quasar absorption lines can be broken into two classes: surveys and detailed studies. In the case of surveys, discussed further in Chapter 9, a complete and unbiased sample of quasar spectra is obtained, from which a large database of absorption lines, for example, is

compiled. This database is then studied in the context of the cosmological setting to obtain, for example, the redshift distribution, the statistical cross-section of the absorbing gas complexes, or the distribution of absorption line strengths. These distributions are obtained by assuming a functional form for the distribution and then using formal parameter estimation to obtain best estimates and errors for the assumed form of the distribution function. Because observational data do not have uniform nor infinite sensitivity, estimating the parameters for the distribution function must include the incompleteness of the data. This incompleteness is modeled by a so-called sensitivity function.

Further tests can then be employed to examine if the observed distribution function is statistically consistent or inconsistent with a distribution function derived upon physically motivated principles. Often, then, the form of the distribution function applied to model the data is motivated by the particular test for which the data were originally obtained. For example, one may be testing for evolution in the redshift path density ( $dN/dz$ , Eq. 2.176, outlined in § 2.11), in which case the chosen distribution function would include evolution parameters amended to the functional form of  $dN/dz$ .

In this chapter, we review the standard statistical descriptors of data, i.e., the mean, standard deviation, etc. We then introduce a few distribution functions commonly employed in astronomical studies, followed by discussion on the principles of establishing confidence levels on the statistical descriptors, on tests designed to determine if an assumed underlying distribution is consistent with the measure distribution, and tests to determine correlations between two independent data sets. We then, discuss the maximum likelihood method for parameter estimation and derive a general procedure for two-parameter distribution functions. We also discuss the  $\chi^2$  minimization technique for modeling data. Examples are provided for both the maximum likelihood and the  $\chi^2$  minimization methods. We include a brief introduction to some of the numerical software available to undertake  $\chi^2$  minimization.

## 3.1 Data and Underlying Distributions

### 3.1.1 Statistical Descriptors of Data

Consider a data set of  $m$  independent measurements of the quantity  $t$  with values

$$t_1, t_2, t_3, t_4, \dots, t_m,$$

where  $t_i$  could be redshifts, equivalent widths, column densities, metallicities, etc.

A set of discrete data values are statistically characterized by the mean value,  $\langle t \rangle$  the standard deviation,  $\sigma$ , which is the "dispersion" about the mean, the skew,  $s$ , and the kurtosis,  $k$ . The latter three are each successive higher moments of the sum of the residuals from the mean. The mean and  $\sigma$  have the units of the data, whereas the skew and kurtosis are dimensionless quantities that measure the asymmetry and the flatness (or peakness) of the distribution. A distribution is considered "normal", Gaussian, if it is symmetric about the mean (null skew) and has a null kurtosis.

For a collection of  $m$  data values,  $t_i$ , the mean is given by

$$\langle t \rangle = \frac{1}{m} \sum_{i=1}^m t_i, \quad (3.1)$$

the standard deviation is given by the square root of the variance,  $V = \sigma^2$ , where

$$\sigma^2 = \frac{1}{m-1} \sum_{i=1}^m (t_i - \langle t \rangle)^2, \quad (3.2)$$

where the term  $m-1$  appears because one degree of freedom is removed by invoking  $\langle t \rangle$  in the sum. The skew is

$$s = \frac{1}{m} \sum_{i=1}^m \left( \frac{t_i - \langle t \rangle}{\sigma} \right)^3, \quad (3.3)$$

where  $s > 0$  indicates a tail in the distribution toward  $t > \langle t \rangle$  and  $s < 0$  indicates a tail toward  $t < \langle t \rangle$ . The kurtosis is

$$k = \left[ \frac{1}{m} \sum_{i=1}^m \left( \frac{t_i - \langle t \rangle}{\sigma} \right)^4 \right] - 3, \quad (3.4)$$

where  $k > 0$  indicates a narrow peak and  $k < 0$  indicates a flat, broad peak (or lack of a clear peak).

The median of the data,  $\bar{t}$ , is the value at which equal numbers of data points lie at  $t < \bar{t}$  and  $t > \bar{t}$ . If the data can be sorted in ascending order, then the median is easily estimated by

$$\bar{t} = t_{(m+1)/2} \quad (3.5)$$

if  $m$  is odd and by

$$\bar{t} = \frac{1}{2} (t_{m/2} + t_{m/2+1}), \quad (3.6)$$

if  $m$  is even.

Applications where statistical probabilities and a formal confidence level, CL, is desired, require the cumulative distribution function (CDF). The CDF ranges from  $0 \leq F(t) \leq 1$  over the domain  $t_{min} \leq t \leq t_{max}$ . The definition of the CDF for a known probability density distribution,  $f(t)$ , is

$$F(t) = \frac{\int_{t_{min}}^t f(w)dw}{\int_{t_{min}}^{t_{max}} f(w)dw}, \quad (3.7)$$

where the integrals are performed over the domain of the data set. In many statistical applications (for example, when determining confidence limits),  $t_{min} = 0$  and/or  $t_{max} = \infty$  are used for the limits of integration. Computing the CDF for the observed data requires sorting the data in ascending order. Attention must be given to possible multiplicity of data values, or identical  $t_i$  in the sorted list. Of the  $m$  data points in the data set, let there be  $j = 1, 2, 3, \dots, n$  unique values in the sorted list and let the multiplicity of the  $j$ th data point be  $d_j$ . Then  $m = n + \sum_{j=1}^n (d_j - 1)$  holds. Once this book keeping is accomplished, the CDF is computed from

$$F(t_j) = \sum_{k=1}^j \frac{d_k}{m}. \quad (3.8)$$

### 3.1.2 Natural Distributions

Need an introduction. Be sure to discuss the CDF and CL connection.

#### 3.1.2.1 Normal Distribution

The normal distribution, also called the Gaussian distribution, is the most widely applied probability distribution in the physical sciences. The probability density function is

$$f(t, \langle t \rangle, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t - \langle t \rangle)^2}{2\sigma^2}\right). \quad (3.9)$$

Note that the function is symmetric about the mean. A full characterization of the statistical moments is

$$\begin{aligned}
 \text{mean} &= \langle t \rangle \\
 \text{median} &= \langle t \rangle \\
 \text{mode} &= \langle t \rangle \\
 \text{standard deviation} &= \sigma \\
 \text{skew} &= 0 \\
 \text{kurtosis} &= 0.
 \end{aligned} \tag{3.10}$$

The cumulative distribution function (CDF) is

$$F(t, \langle t \rangle, \sigma) = \frac{1}{2} \left[ 1 + \operatorname{erf} \left( \frac{t - \langle t \rangle}{\sqrt{2}\sigma} \right) \right], \tag{3.11}$$

where  $\operatorname{erf}(x)$  is the error function

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-w^2} dw. \tag{3.12}$$

Note that the integrand is of the form of the probability density function evaluated at

$$x = \frac{t - \langle t \rangle}{\sqrt{2}\sigma}. \tag{3.13}$$

The range of  $\operatorname{erf}(x)$  is  $\operatorname{erf}(0) = 0$  and  $\operatorname{erf}(\infty) = 1$ , and the asymmetry is  $\operatorname{erf}(-x) = -\operatorname{erf}(x)$ .

The probability density function of the normal distribution, Eq. 3.9, is illustrated in upper-left panel of Fig. 3.1. Three Gaussian functions are shown, for  $\sigma = 0.5$  (short dash), 1.0 (solid), and 2.0 (long-dash) for  $\langle t \rangle = 0$ . The function has  $f(t, 0, \sigma) = 0.5$  at  $t = \pm\sigma$ . The cumulative distribution functions (CDF), Eq. 3.11, are shown in the lower-left panel of Fig. 3.1. The CDF provides the area under the probability density distribution as a function of  $t$ . Note that the CDF is 0.5 for  $t = \langle t \rangle$  for all normal distributions. The percent area under the probability distribution curve over the range  $t - \langle t \rangle = \pm N \sigma$ , where  $N$  is an integer, can be obtained from the CDF integrated over the appropriate limits. We have

$$\begin{aligned}
 t - \langle t \rangle = \pm 1 \sigma &= 68.26895\% \\
 t - \langle t \rangle = \pm 2 \sigma &= 95.44997\% \\
 t - \langle t \rangle = \pm 3 \sigma &= 99.73002\% \\
 t - \langle t \rangle = \pm 4 \sigma &= 99.99367\% \\
 t - \langle t \rangle = \pm 5 \sigma &= 99.99994\%.
 \end{aligned} \tag{3.14}$$

These areas are interpreted as the percent number of data values that lie within  $\pm N \sigma$  of the mean if the data are drawn from a normal distribution.

Do not mistake the percent area under the probability density distribution over the limits  $t - \langle t \rangle = \pm N \sigma$  as a confidence level in statistical testing (see § 3.2).

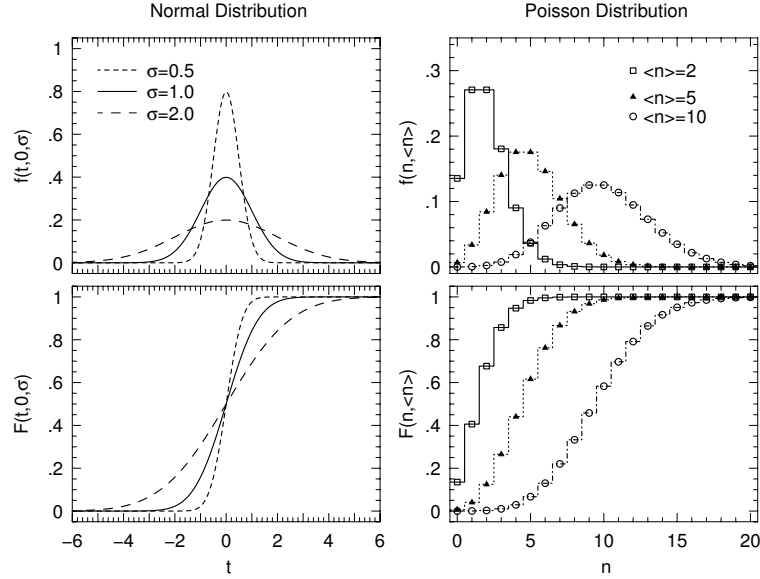


Figure 3.1: — (upper left) The continuous normal (Gaussian) probability density function,  $f(t, \langle t \rangle, \sigma)$  for  $\langle t \rangle = 0$  and  $\sigma = 0.5$  (short dash),  $1.0$  (solid), and  $2.0$ . (long-dash). — (lower left) The cumulative distribution function (CDF),  $F(t, \langle t \rangle, \sigma)$ , for the three normal distributions in the upper panel. The CDF provides the area under the probability distribution as a function of  $t$ ; Note that the CDF is  $0.5$  at  $t = \langle t \rangle$ . — (upper right) The discrete Poisson probability mass function,  $f(n, \langle n \rangle)$  for  $\langle n \rangle = 2$  (square, solid),  $5$  (triangle, dotted), and  $10$  (circle, dot-dash). — (lower right) The cumulative distribution function (CDF),  $F(n, \langle n \rangle)$ , for the three Poisson distributions in the upper panel.

### 3.1.2.2 Poisson Distribution

The Poisson distribution is a discrete probability distribution. It expresses the probability of a number of events occurring in a fixed interval if these events are independent and occur with a known average rate. An example would include the number of photons collected from an object having constant luminosity over a fixed time interval. The Poisson distribution is the discrete counterpart of the continuous normal (Gaussian) distribution. The probability of detecting  $n$  (integer) events in a fixed interval established through the nature of the experiment in which  $\langle n \rangle$  integer events are

the expected average number of events in this interval, is

$$f(n, \langle n \rangle) = \frac{\langle n \rangle^n e^{-\langle n \rangle}}{n!}. \quad (3.15)$$

Eq. 3.15 is often called the probability "mass" function (where the term "mass" accentuates the fact that it is a discrete probability function). The statistical moments are

$$\begin{aligned} \text{mean} &= \langle n \rangle \\ \text{standard deviation} &= \langle n \rangle^{1/2} \\ \text{skew} &= \langle n \rangle^{-1/2} \\ \text{kurtosis} &= \langle n \rangle^{-1}. \end{aligned} \quad (3.16)$$

In the limit of large  $\langle n \rangle$ , a Poisson distribution converges to a normal distribution. The cumulative distribution function, CDF, is

$$F(n, \langle n \rangle) = \frac{\Gamma(n+1, \langle n \rangle)}{\Gamma(n+1)} = \frac{\Gamma(n+1, \langle n \rangle)}{n!}, \quad (3.17)$$

where  $\Gamma(a, x)$  is the incomplete  $\Gamma$  function,

$$\Gamma(a, x) = \int_0^x e^{-w} w^{a-1} dw, \quad (3.18)$$

where the complete  $\Gamma$  function is

$$\Gamma(a) = \int_0^\infty e^{-w} w^{a-1} dw. \quad (3.19)$$

For the Poisson distribution,  $a = n+1$  and  $x = \langle n \rangle$ . Note that the integrands of the  $\Gamma$  functions take on the form of the probability mass function following the substitution of  $a$ . When  $a$  is an integer the  $\Gamma$  function is the factorial function offset by  $n+1$ ,

$$n\Gamma(n) = \Gamma(n+1) = n!, \quad (3.20)$$

where the latter gives the recurrence relationship (thus, the origin of  $n!$  in the denominators of Eq. 3.21 and 3.23).

The probability mass function of the Poisson distribution, Eq. 3.15, is illustrated in upper-right panel of Fig. 3.1. Three Poisson functions are shown as a function  $n$  for  $\langle n \rangle = 2$  (square), 5 (triangle), and 10 (circle). Note that as  $\langle n \rangle$  increases, the Poisson distribution approaches the normal distribution in its properties. The cumulative distribution function (CDF), Eq. 3.17, for the three distributions is shown in the lower-left panel of Fig. 3.1. The CDF provides the area under the probability mass distribution as a function of  $n$ .

### 3.1.2.3 Binomial Distribution

The binomial distribution describes the probability of outcomes in an experiment where the data are bimodal, either a “1” or a “0”, and “up” or a “down”, a “yes” or a “no”. The probability of obtaining  $n$  positive outcomes in a sample of  $m$  trials for which the probability of a positive outcome for a single trial is known to be  $p$ , is given by the probability mass function

$$f(n, m, p) = \left( \frac{m!}{n!(m-n)!} \right) p^n (1-p)^{m-n}. \quad (3.21)$$

The statistical moments are

$$\begin{aligned} \text{mean} &= mp \\ \text{mode} &= (m+1)p \\ \text{standard deviation} &= [mp(1-p)]^{1/2} \\ \text{skew} &= (1-2p)/[mp(1-p)]^{1/2} \\ \text{kurtosis} &= [1-6p(1-p)]/[mp(1-p)]. \end{aligned} \quad (3.22)$$

The cumulative distribution function (CDF) is

$$F(n, m, p) = \frac{\beta_{1-p}(m-n, n+1)}{\beta(m-n, n+1)} \quad (3.23)$$

where  $\beta_x(a, b)$  is the incomplete  $\beta$  function

$$\beta_x(a, b) = \int_0^x w^{a-1} (1-w)^{b-1} dw, \quad (3.24)$$

and  $\beta(a, b)$  is the complete  $\beta$  function

$$\beta(a, b) = \int_0^1 w^{a-1} (1-w)^{b-1} dw. \quad (3.25)$$

For the binomial distribution,  $x = 1-p$ ,  $a = m-n$ , and  $b = n+1$ . Substitution of  $a$  and  $b$  into the general expression for the incomplete  $\Gamma$  function recovers the functional form of the probability mass function for  $x = 1-p$ .

Examples of the probability mass function for the binomial distribution, Eq. 3.21, are illustrated in upper panels of Fig. 3.2. Three experiments are shown for  $m = 5$ , (square, solid) 10 (triangle, dotted), and 20 (circle, dot-dash) trials, or measurements. For each experiment, the probabilities for positive results are  $p = 0.25, 0.5$ , and  $0.75$ , respectively shown from left to right in the upper panels. An applicable example might be that  $m$  quasar spectra are obtained for which the probability detecting an absorbing system is  $p$ . For the  $m = 20$  and  $p = 0.25$  scenario, the mean is  $(0.25)(20) = 5$ ,



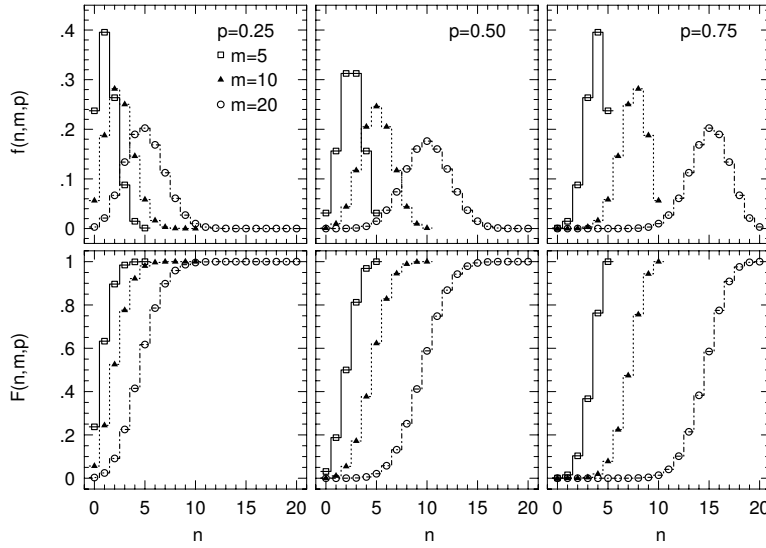


Figure 3.2: — (upper) The binomial probability mass function,  $f(n, m, p)$   $n$  successes  $n$  in  $m = 5$  (square), 10 (triangle), and 20 (circle) trials. Three probabilities  $p = 0.25$ , 0.50, and 0.75 are shown. — (lower) The cumulative distribution function (CDF),  $F(n, m, p)$ , for the distributions in the upper panels. The CDF provides the area under the probability distribution as a function of  $n$ ; Note that the CDF is 0.5 at the most probable  $n$ .

and the probability mode is  $(0.25)(21) = 5.25$ . The cumulative distribution functions (CDF), Eq. 3.23, for the various experiments are shown in the lower panels of Fig. 3.2.

#### 3.1.2.4 Exponential and Power-Law Distribution

The exponential distribution is a continuous probability that is primarily applied to describe the change in the number of data values over a finite domain over which the data have a constant average rate of change. Examples include the decay rate of nuclear isotopes, or the density scale height of atoms in atmospheres. The exponential distribution thus has a characteristic scale,  $t_*$ . The probability density distribution is

$$f(t, t_*) = \frac{1}{t_*} \exp(-t/t_*). \quad (3.26)$$

The statistical moments are

$$\begin{aligned}
 \text{mean} &= t_* \\
 \text{median} &= t_* \ln(2) \\
 \text{mode} &= 0 \\
 \text{standard deviation} &= t_* \\
 \text{skew} &= 2 \\
 \text{kurtosis} &= 6
 \end{aligned} \tag{3.27}$$

The cumulative distribution function (CDF) is

$$F(t, t_*) = 1 - \exp(-t/t_*) \tag{3.28}$$

The probability density function of the exponential distribution, Eq. 3.26, is illustrated in upper-left panel of Fig. 3.3. Three functions are shown as a function  $t$  for  $t_* = 0.5$  (solid), 1.0 (dotted), and 2.0 (dashed). The exponential function follows an e-folding scale relationship for  $t$  equal to integer multiples of  $t_*$

$$f[(n+1)t_*, t_*] = e^{-1} f(nt_*, t_*), \tag{3.29}$$

where  $n$  is an integer over the domain  $0 \leq n \leq \infty$ . For this reason, the characteristic scale parameter,  $t_*$ , is often called the e-folding scale (the distribution function decreases one power of  $e$  for each additional  $t_*$ ). The cumulative distribution function (CDF), Eq. 3.28, for the three distributions is shown in the lower left panel of Fig. 3.1.

The Power-law distribution is probably the most frequently observed scaling law. It describes the scale invariance found in many natural phenomena. It has no characteristic scale length and has no natural normalization (over the domain  $\infty \leq t \leq \infty$  the area under the distribution diverges). Thus, there is no true probability density distribution to describe the power law. The functional form is

$$f(t, a, b) = at^b, \tag{3.30}$$

where  $a$  is the constant of proportionality and  $b$  is the exponent of the power law. Both are constants. On a log-log plane, the power law is a linear relation

$$\log f(t, a, b) = \log a + b \log t. \tag{3.31}$$

When it is suspected that a data set may follow a power-law distribution, a good rule of thumb is to examine whether the data are linear on a graph of  $\log f$  versus  $\log t$  for more than three orders of magnitude.

### 3.1.2.5 Gamma Distribution

The  $\Gamma$  distribution accounts for a combination of both the exponential and power-law distributions. It describes distributions that have scale invariance for large values yet have a characteristic scale for small values. The scale at which the transition occurs is governed by the characteristic scale of the exponential. Two important applications in the astronomical sciences include the mass distribution function of gaseous structures and the luminosity distribution function of galaxies and quasars. A slightly modified version of the  $\Gamma$  distribution is better known as the Schechter function. The probability density distribution is

$$f(t, \alpha, t_*) = \frac{1}{\Gamma(\alpha)} \left( \frac{t}{t_*} \right)^{\alpha-1} \left[ \frac{1}{t_*} \exp(-t/t_*) \right], \quad (3.32)$$

which has been written in a form to accentuate that the  $\Gamma$  distribution is product of the exponential and power-law distributions with constant of proportionality  $a = [t_*^{\alpha-1} \Gamma(\alpha)]^{-1}$ . The statistical moments are

$$\begin{aligned} \text{mean} &= \alpha t_* \\ \text{mode} &= (\alpha - 1) t_* \quad (\alpha \geq 1) \\ \text{standard deviation} &= \sqrt{\alpha} t_* \\ \text{skew} &= 2/\sqrt{\alpha} \\ \text{kurtosis} &= 6/\alpha \end{aligned} \quad (3.33)$$

The cumulative distribution function (CDF) is

$$F(t, \alpha, t_*) = \frac{\gamma(\alpha, t/t_*)}{\Gamma(\alpha)} = 1 - \frac{\Gamma(\alpha, t/t_*)}{\Gamma(\alpha)}, \quad (3.34)$$

where  $\gamma(a, x)$  is the complementary incomplete  $\Gamma$  function,

$$\gamma(a, x) = \int_x^\infty e^{-w} w^{a-1} dw, \quad (3.35)$$

and where the complete  $\Gamma$  function (Eq. 3.19) is

$$\Gamma(a) = \int_0^\infty e^{-w} w^{a-1} dw. \quad (3.36)$$

The probability density function of the  $\Gamma$  distribution, Eq. 3.32, is illustrated in the upper-right panel of Fig. 3.3. Nine functions are shown as a function  $t$  for  $t_* = 0.5, 1.0$ , and  $2.0$ , for  $\alpha = 1.5$  (solid)  $3.0$  (dotted), and  $8.0$  (dashed). The functions have been presented in log-log, as is customary in the astronomical sciences. The slope for small  $t$  is governed by  $\alpha$ , which is the origin of the name “faint-end slope” for  $\Gamma$  distributions. The cumulative distribution functions (CDF), Eq. 3.34, for the nine distributions are shown in the lower right panel of Fig. 3.3.

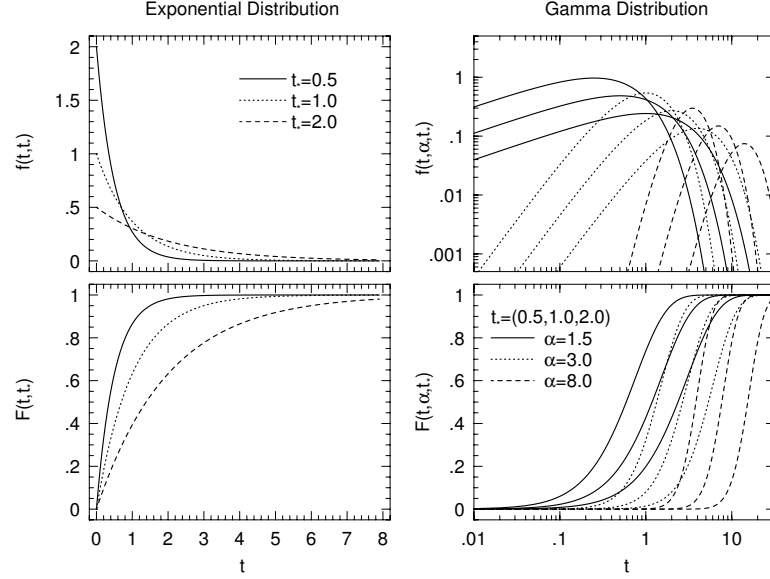


Figure 3.3: — (upper left) The exponential probability density function,  $f(t, t_*)$ , as a function  $t$  for  $t_* = 0.5$  (solid), 1.0 (dotted), and 2.0 (dash). Note that the curves follow the e-folding scale relation (Eq. 3.29) — (lower left) The cumulative distribution functions (CDF),  $F(t, t_*)$ , for the three distributions in the upper-left panel. — (upper right) The  $\Gamma$  probability density function,  $f(t, \alpha, t_*)$ , for combinations of  $\alpha = (1.5, 3.0, 8.0)$  and  $t_* = (0.5, 1.0, 2.0)$  (see legend), presented in log-log. — (lower right) The cumulative distribution function for the nine distributions. The parameter  $\alpha$  governs the slope for small  $t$ .

## 3.2 Confidence Limits

Not yet written.

## 3.3 Distribution Tests

### 3.3.1 Student's $t$ -test

The Student's  $t$ -test is applied to determine if the the mean of two distributions of discrete data are statistically different.

Consider two complete data sets, 1 and 2, of  $m_1$  and  $m_2$  data points. Under the condition that the standard deviations are similar, i.e.,  $\sigma_1 \simeq \sigma_2$ , the student's  $t$  is

$$t_s = \frac{\langle t_1 \rangle - \langle t_2 \rangle}{\sigma_{12}}, \quad (3.37)$$

where  $\langle t_1 \rangle$  and  $\langle t_2 \rangle$  are the means of the respective distributions, and where

the degrees of freedom for the combined data sets is

$$\nu = \nu_1 + \nu_2 = (m_1 - 1) + (m_2 - 1) \quad (3.38)$$

The value of  $\sigma_{12}$  is

$$\sigma_{12}^2 = \left[ \frac{m_1 + m_2}{m_1 m_2} \right] \frac{\nu_1 \sigma_1^2 + \nu_2 \sigma_2^2}{\nu} \quad (3.39)$$

Under the condition that the standard deviations are not similar (as determinable from the  $F$ -test, see below), the student's  $t$  is

$$t_s = \frac{\langle t_1 \rangle - \langle t_2 \rangle}{(S_1 + S_2)^{1/2}}, \quad (3.40)$$

where

$$S_1 = \frac{\sigma_1^2}{m_1} \quad S_2 = \frac{\sigma_2^2}{m_2} \quad (3.41)$$

and

$$\nu = \nu_1 \nu_2 \frac{(S_1 + S_2)^2}{\nu_2 S_1^2 + \nu_1 S_2^2} \quad (3.42)$$

is the “reduced” number of degrees of freedom.

For both scenarios, the probability that the distribution means are statistically consistent is obtained from the CDF of the binomial probability mass function,

$$P(x) = \beta_x(\nu/2, 1/2) \quad (3.43)$$

which is the incomplete  $\beta$  function given by Eq. 3.24, with

$$\begin{aligned} x &= \nu / (\nu + t_s^2) \\ a &= \nu/2 \\ b &= 1/2 \end{aligned} \quad (3.44)$$

The range of  $x$  is  $0 \leq x \leq 1$  for the domain  $\infty \geq t_s \geq 0$ .

The limiting values of the incomplete  $\beta$  function yield  $P(0) = 0$  and  $P(1)$ . The value of  $P(x)$  is the probability, that the measured value of  $x$  could be smaller (due to larger  $t_s$ ) than if the means of the distributions were identical, i.e.,  $t_s = 0$  and  $x = 1$ . When  $P(x) < 0.01$ , it indicates that the probability of the two means being statistically consistent is less than 1%. Thus, the confidence level that the means are *not* statistically identical is  $CL = 1 - P(x)$ .

### 3.3.2 The $F$ -test

The  $F$ -test is the well-known test to reject the null-hypothesis that two data sets have statistically consistent variances, i.e.,  $\sigma^2$ .

Consider two complete data sets, 1 and 2, of  $m_1$  and  $m_2$  data points. The  $F$  statistic is

$$F = \frac{\sigma_1^2}{\sigma_2^2}, \quad (3.45)$$

with degrees of freedom  $\nu_1 = m_1 - 1$  and  $\nu_2 = m_2 - 1$ .

As with the Student's  $t$ -test, the probability that the null-hypothesis holds true is obtained from the CDF of the binomial probability mass function. However, since the goal is to determine if very small or very large  $F$  is statistically improbable, we need to account for both directions in the distribution of  $F$ . Thus, we have

$$P(x_1, x_2) = [1 - \beta_{x_1}(\nu_1/2, \nu_2/2)] + \beta_{x_2}(\nu_2/2, \nu_1/2), \quad (3.46)$$

from the incomplete  $\beta$  function given by Eq. 3.24, with

$$\begin{aligned} x_1 &= \nu_1/(\nu_1 + \nu_2 F) \\ x_2 &= \nu_2/(\nu_2 + \nu_1 F) \\ a_1 = b_2 &= \nu_1/2 \\ b_1 = a_2 &= \nu_2/2 \end{aligned} \quad (3.47)$$

The range of each  $x$  is  $0 \leq x \leq 1$  for the domain  $\infty \geq F \geq 0$ . If  $P(x_1, x_2) < 0.01$ , it would mean that the probability that the two variances are statistically consistent is less than 1%. The confidence level that the variances are *not* statistically consistent is  $CL = 1 - P(x_1, x_2)$ .

### 3.3.3 The Kolmogorov–Smirnov Test

Known as the K–S test, this statistic examines if two distributions are consistent. One of the benefits of the K–S test is that a data set can be compared directly to an analytical distribution function, such as a power-law or exponential function. It can also be applied to examine if the distributions of two independent data sets are statistically consistent.

The test directly compares the cumulative distribution functions. The K–S statistic is simply the maximum absolute difference between the CDF obtained from the data,  $F(t_j)$  (Eq. 3.8), and the candidate analytical CDF function,  $F(t)$  (Eq. 3.7),

$$\Delta_{KS} = \max |F(t_j) - F(t)|. \quad (3.48)$$

The probability of significance is provided by the converging alternating series

$$P_{KS}(x) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2x^2 j^2), \quad (3.49)$$

which can be converged to some arbitrary accuracy for

$$x = m^{1/2} \Delta_{KS}. \quad (3.50)$$

The function  $P_{KS}(x)$  increases monotonically with  $x$ , over the range  $I_{KS}(0) = 1$  and  $I_{KS}(\infty) = 0$ . Thus, it can be seen that larger  $x$  yields a lower probability that the measured distribution is consistent with the assumed analytical probability density function.

If instead of comparing a measured distribution to an analytical distribution function, one compares the distributions of two data set, then the K-S statistic is

$$\Delta_{KS} = \max |F_1(t_j) - F_2(t_j)|. \quad (3.51)$$

for data set 1 with  $m_1$  points and data set 2 with  $m_2$  points. The test is conducted over the domain of the minimum value to the maximum value of the  $t_i$  of the two data sets. In this case, the probability that the distributions are statistically consistent is given by Eq. 3.49 with

$$x = \left[ \frac{m_1 m_2}{m_1 + m_2} \right]^{1/2} \Delta_{KS} \quad (3.52)$$

### 3.3.4 Standard Linear Correlation Test

The standard linear correlation coefficient between two data sets 1 and 2 of equal number of data points,  $m$ , is

$$r = \frac{\sum_{i=1}^m \Delta t_{1,i} \Delta t_{2,i}}{\sqrt{\sum_{i=1}^m (\Delta t_{1,i})^2} \sqrt{\sum_{i=1}^m (\Delta t_{2,i})^2}} \quad (3.53)$$

where

$$\begin{aligned} \Delta t_{1,i} &= t_{1,i} - \langle t_1 \rangle \\ \Delta t_{2,i} &= t_{2,i} - \langle t_2 \rangle \end{aligned} \quad (3.54)$$

For  $r \sim 0$ , it can be claimed that there is no strong evidence for a correlation. However, even when  $r$  approaches  $-1$  (anticorrelation) or  $+1$  (correlation), there is no robust method to obtain the probability that the

null hypothesis (no correlation) is ruled out. This is because the  $r$  statistic is an arbitrary number and is not based upon any knowledge of the underlying distribution of the either data set.

However, under the (perhaps unfounded) assumption that  $r$  follows a normal distribution with a mean  $\langle r \rangle = 0$  and standard deviation  $\sigma = 1/\sqrt{m}$ , then the probability that  $|r|$  would have its value as compared to the null hypothesis of  $r = 0$  is given by

$$P(x) = \text{erfc}(x) = 1 - \text{erf}(x) \quad (3.55)$$

where  $\text{erfc}(x)$  is the complementary error function and  $\text{erf}(x)$  is the error function, given by Eq. 3.12 and where

$$x = \frac{|r|}{\sqrt{2}\sigma}. \quad (3.56)$$

The range of  $P(x)$  is  $P(0) = 1$  and  $P(\infty) = 0$ . The significance level at which a lack of correlation is ruled out is  $\text{CL} = 1 - P(k)$ .

The Student's  $t$ -test can also be applied to test the significance, where the degrees of freedom is  $\nu = m - 2$ , and

$$t_s = r \sqrt{\frac{\nu}{1 - r^2}}. \quad (3.57)$$

The probability and confidence level is then determined from Eq. 3.43 as applied in § 3.3.1.

Because the the underlying distribution of the data is unknown, confidence level is fraught with assumption based uncertainty. It is often a wise idea to compare the two methods for statistical significance. For these reasons, the standard linear correlation coefficient,  $r$ , is sometimes abandoned for so-called non-parametric or "rank" correlation statistics. Two that are commonly employed in the literature are the Spearman and the Kendall rank order tests.

### 3.3.5 Spearman Correlation Test

The Spearman correlation test requires that the two data sets, 1 and 2, be arranged in ascending order. Thus, they must be sorted before the Spearman coefficient can be computed. Then, each datum is assigned a rank (effectively its number of appearance in the sorted list). If  $R_{1,i}$  is the rank of  $t_{1,i}$  in data set 1, and If  $R_{2,i}$  is the rank of  $t_{2,i}$  in data set 2, then



the Spearman coefficient is

$$r_S = \frac{\sum_{i=1}^m \Delta R_{1,i} \Delta R_{2,i}}{\sqrt{\sum_{i=1}^m (\Delta R_{1,i})^2} \sqrt{\sum_{i=1}^m (\Delta R_{2,i})^2}} \quad (3.58)$$

where

$$\begin{aligned} \Delta R_{1,i} &= R_{1,i} - \langle R_1 \rangle \\ \Delta R_{2,i} &= R_{2,i} - \langle R_2 \rangle \end{aligned} \quad (3.59)$$

If  $r_S < 0$  then the data tend to be anticorrelated and if  $r_S > 0$  then the data tend to be correlated.

The Student's  $t$ -test is applied to test the significance, where the degrees of freedom is  $\nu = m - 2$ , and

$$t_s = r_S \sqrt{\frac{\nu}{1 - r_S^2}}. \quad (3.60)$$

The probability is then determined from Eq. 3.43 as applied in § 3.3.1, with

$$x = \frac{\nu}{\nu + t_s^2}. \quad (3.61)$$

If  $P(x) < 0.01$ , it would mean that the probability that the data are uncorrelated (null hypothesis) is less than 1%. The confidence level that the correlation is present is  $CL = 1 - P(x)$ .

### 3.3.6 Kendall Correlation Test

An advantage of the Kendall correlation test over that of the Spearman correlation test is that it employs only the relative ranking of data pairs. Thus, the data do not need to be sorted.

The Kendall  $\tau$  is computed from

$$\tau = \frac{\sum_{j=1}^{m-1} \sum_{k=1}^m \text{sign}(\Delta t_{1,jk} \Delta t_{2,jk})}{\sqrt{\sum_{j=1}^{m-1} \sum_{k=1}^m \hat{H}(|\Delta t_{1,jk}|)} \sqrt{\sum_{j=1}^{m-1} \sum_{k=1}^m \hat{H}(|\Delta t_{2,jk}|)}}, \quad (3.62)$$

where

$$\begin{aligned} \Delta t_{1,jk} &= t_{1,j} - t_{1,k} \\ \Delta t_{2,jk} &= t_{2,j} - t_{2,k} \end{aligned} \quad (3.63)$$

and where  $\hat{H}(t)$  is a modified Heavisite step function, such that

$$\hat{H}(t) = \begin{cases} 0 & t = 0 \\ 1 & t > 0 \end{cases} \quad (3.64)$$

If  $\tau < 0$  then the data tend to be anticorrelated and if  $\tau > 0$  then the data tend to be correlated. Under the null-hypothesis of no correlation,  $\tau$  has the mean  $\langle \tau \rangle = 0$  and is normally distributed with variance

$$\sigma^2 = \frac{1}{9} \frac{2m + 5}{N_p}, \quad (3.65)$$

where  $N_p = m(m-1)/2$  is the number of unique data pairs (no self pairing). Because  $\tau$  is normally distributed, the probability of significance that the measured  $\tau$  departs from the null-hypothesis (i.e.,  $\tau = 0$ ) is provided by the complimentary error function (Eq. 3.55 as applied in § 3.3.4), with

$$x = \frac{|\tau|}{\sqrt{2}\sigma}. \quad (3.66)$$

If  $P(x) < 0.01$ , it would mean that the probability that the data are uncorrelated (null hypothesis) is less than 1%. The confidence level that the correlation is present is  $CL = 1 - P(x)$ .

### 3.4 Distribution Functions

In an experiment, it is often desirable to determine the functional form of a distribution of measured data. Consider a set of independent measurements of the quantity  $t$  with values

$$t_1, t_2, t_3, t_4, \dots, t_m,$$

where  $t_i$  could be redshifts, equivalent widths, column densities, metallicities, etc. The assumed distribution function will depend upon a set of  $n$  parameters,  $\mathbf{a}$ , where

$$\mathbf{a} = a_1, a_2, a_3, a_4, \dots, a_n,$$

where  $n < m$ . Common examples of assumed distribution functions in astronomy include the Gaussian distribution,

$$n(t; a_1, a_2, a_3)dt = a_1 \exp \left[ - \left( \frac{t - a_2}{a_3} \right)^2 \right] dt \quad (3.67)$$

the exponential distribution,

$$n(t; a_1, a_2)dt = a_1 \exp \left[ - \left( \frac{t}{a_2} \right) \right] dt \quad (3.68)$$

the power-law distribution,

$$n(t; a_1, a_2)dt = a_1 t^{-a_2} dt \quad (3.69)$$

and the modified  $\Gamma$  distribution,

$$n(t; a_1, a_2, a_3)dt = a_1 \left( \frac{t}{a_3} \right)^{-a_2} \exp \left[ - \left( \frac{t}{a_3} \right) \right] dt, \quad (3.70)$$

where the latter is often referred to as the Schechter function. Depending upon the context of the problem, the distribution functions are often normalized to either the number of data points,  $m$ , or some other related constant derived from the survey,

$$I = \int_{t_{min}}^{\infty} n(t; \mathbf{a})dt, \quad (3.71)$$

where  $t_{min}$  is the minimum of the  $t_i$  measured in the data. The expectation value of a given distribution function is defined by

$$\langle E \rangle = \frac{1}{I} \int_{t_{min}}^{\infty} E(t) n(t; \mathbf{a})dt. \quad (3.72)$$

Thus, the mean of the measure data is

$$\langle t \rangle = \frac{1}{I} \int_{t_{min}}^{\infty} t n(t; \mathbf{a})dt, \quad (3.73)$$

and the variance in the measured data is

$$\sigma^2 = \frac{1}{I} \int_{t_{min}}^{\infty} (t - \langle t \rangle)^2 n(t; \mathbf{a})dt, \quad (3.74)$$

### 3.5 Maximum Likelihood Method

The goal is to determine the “best” values of  $\mathbf{a}$  that describe the true values  $\hat{\mathbf{a}}$ . Three favorable properties of maximum likelihood estimation are (1) that as  $m \rightarrow \infty$ , the  $\mathbf{a}$  approach the true  $\hat{\mathbf{a}}$ , and (2) the uncertainties in the  $\mathbf{a}$  are normally distributed with a *minimized* variance, and (3) the data

do not need to be arbitrarily grouped or binned, which results in a loss of information.

In practice, there is not an equal probability that all  $t_i$  will be detected in the data. The probability of measuring  $t_i$  in a survey is

$$p(t_i; \mathbf{a}) = \frac{g(t_i)n(t_i; \mathbf{a})}{I(\mathbf{a})}, \quad (3.75)$$

where the “sensitivity” function,  $g(t)$ , is determined from the data set. The computation of the sensitivity function will be discussed in detail in § 9.2. The  $p(t_i; \mathbf{a})$  are the probability distribution functions describing the likelihood that  $t_i$  would have been measured assuming the  $t_i$  are distributed following  $n(t_i; \mathbf{a})$  in a survey with a sensitivity  $g(t_i)$  for detecting  $t_i$ . The normalization  $I(\mathbf{a})$  is determined by normalizing the probability distribution to unity

$$\int_{t_{min}}^{\infty} p(t; \mathbf{a}) dt = 1, \quad (3.76)$$

giving

$$I(\mathbf{a}) = \int_{t_{min}}^{\infty} g(t)n(t; \mathbf{a}) dt. \quad (3.77)$$

The best estimate of  $\mathbf{a}$  is obtained when the joint probability function is maximized. The joint probability function is the product of the  $p(t_i; \mathbf{a})$ , written

$$\mathcal{L} = p(t_1; \mathbf{a})p(t_2; \mathbf{a})p(t_3; \mathbf{a}), \dots, p(t_m; \mathbf{a}), \quad (3.78)$$

where  $m$  is the number of data points. More concisely,  $\mathcal{L}$  is called the likelihood function, and is written

$$\mathcal{L} = \prod_{i=1}^m p(t_i; \mathbf{a}). \quad (3.79)$$

For computational reasons, it is convenient to maximize the natural logarithm of  $\mathcal{L}$ . This is due to the property that the logarithm of products and quotients is equal to the sums and differences of the logarithms for each term, for example,

$$\ln \left( \prod_{i=1}^m \frac{x_i}{y_i} \right) = \sum_{i=1}^m \ln x_i - \sum_{i=1}^m \ln y_i.$$

Writing Eq. 3.79 in logarithm form, gives,

$$\ln \mathcal{L} = \sum_{i=1}^m \ln p(t_i; \mathbf{a}), \quad (3.80)$$

which is maximized for a given  $a_j$  by setting the partial derivatives with respect to  $a_j$  to zero,

$$\begin{bmatrix} \frac{\partial \ln \mathcal{L}}{\partial a_1} \\ \frac{\partial \ln \mathcal{L}}{\partial a_2} \\ \vdots \\ \frac{\partial \ln \mathcal{L}}{\partial a_n} \end{bmatrix} = 0. \quad (3.81)$$

This matrix is solved numerically. However, as we shall demonstrate below, the exponential and power law distribution functions (Eqs. 3.68 and 3.69), which involve only two parameters, can be solved without matrix inversion.

Assuming the errors in  $a_j$  are normally distributed, the variance of  $a_j$  is given by

$$\sigma_{a_j}^2 = - \left( \frac{\partial^2 \ln \mathcal{L}}{\partial a_j^2} \right)^{-1}, \quad (3.82)$$

as  $m \rightarrow \infty$ . This can be seen as follows: Assume the distribution of  $a_j$  is

$$\mathcal{L}(a_j) = C \exp \left[ -\frac{(a_j - \hat{a}_j)^2}{2\sigma_{a_j}^2} \right]$$

where  $\hat{a}_j$  is the true value. We have

$$\ln \mathcal{L}(a_j) = \ln C - \frac{(a_j - \hat{a}_j)^2}{2\sigma_{a_j}^2},$$

with

$$\frac{\partial \ln \mathcal{L}}{\partial a_j} = -\frac{(a_j - \hat{a}_j)}{\sigma_{a_j}^2},$$

and

$$\frac{\partial^2 \ln \mathcal{L}}{\partial a_j^2} = -\frac{1}{\sigma_{a_j}^2},$$

the latter of which is equivalent to Eq. 3.82.

### 3.5.1 Generalized Two Parameter Estimation

The exponential function (Eq. 3.68) and the power-law function (Eq. 3.69) are commonly invoked to describe the distribution of observed data, for example, equivalent widths, column densities, and the redshift density of absorbers. These two functions involve only two parameters, a normalization constant, and a characteristic data value in the case of the exponential function and a power index in the case of the power law. The maximum likelihood method is well suited for estimating the parameters of two-parameter distribution functions. Here, we outline generalized formulae for two-parameter estimation.

Consider the distribution function for a set of  $m$  measured data values  $t_1, t_2, t_3, \dots, t_m$ , that is the product of its normalization and its functional form

$$n(t; \mathbf{a}) = a_1 f(t, a_2), \quad (3.83)$$

where  $a_1$  is the normalization constant and  $a_2$  is a characteristic data value or a power index. For a survey sensitivity function,  $g(t)$ , the normalization is

$$I(\mathbf{a}) = \int_{t_{min}}^{\infty} g(t) n(t, \mathbf{a}) dt = \int_{t_{min}}^{\infty} a_1 g(t) f(t, a_2) dt = a_1 A_2 \quad (3.84)$$

where we introduce the unnormalized integral,

$$A_2 = \int_{t_{min}}^{\infty} g(t) f(t, a_2) dt, \quad (3.85)$$

which depends only upon the parameter  $a_2$ . From Eqs. 3.75 and 3.80, the logarithm of the joint probability function is

$$\ln \mathcal{L} = \sum_{i=1}^m \ln \left( \frac{a_1}{I(\mathbf{a})} \right) + \sum_{i=1}^m \ln g(t_i) + \sum_{i=1}^m \ln f(t_i, a_2) \quad (3.86)$$

Simplifying the first term on the right hand side yields

$$\sum_{i=1}^m \ln \left( \frac{a_1}{I(\mathbf{a})} \right) = - \sum_{i=1}^m \ln \left( \frac{I(\mathbf{a})}{a_1} \right) = - \sum_{i=1}^m \ln A_2 = -m \ln A_2, \quad (3.87)$$

giving  $\ln \mathcal{L}$  in terms of a single parameter,

$$\ln \mathcal{L} = -m \ln A_2 + \sum_{i=1}^m \ln g(t_i) + \sum_{i=1}^m \ln f(t_i, a_2). \quad (3.88)$$

The parameter  $a_2$  is determined by carrying out the differentiation

$$\frac{\partial \ln \mathcal{L}}{\partial a_2} = -\frac{m}{A_2} \frac{\partial A_2}{\partial a_2} + \frac{\partial}{\partial a_2} \sum_{i=1}^m \ln f(t_i, a_2). \quad (3.89)$$

and equating to zero, which yields

$$\frac{A'_2}{A_2} = \frac{1}{m} \frac{\partial}{\partial a_2} \sum_{i=1}^m \ln f(t_i, a_2), \quad (3.90)$$

where we have used the notation

$$A'_2 = \frac{\partial A_2}{\partial a_2} = \int_{t_{min}}^{\infty} g(t) \frac{\partial f(t, a_2)}{\partial a_2} dt, \quad (3.91)$$

which follows from Leibniz's rule for differentiation of integrals. Often, because  $g(t)$  is either a complex function or determined computationally, Eq. 3.90 must be root solved numerically. However, if  $g(t) = 1$  for all  $t_i$ , then Eq. 3.90 has a very clean analytical solution for the exponential and power-law distribution functions.

We obtain the variance in  $a_2$  by a second differentiation of Eq. 3.88

$$\frac{\partial^2 \ln \mathcal{L}}{\partial a_2^2} = m \frac{A'_2}{A_2^2} - m \frac{A''_2}{A_2} + \frac{\partial^2}{\partial a_2^2} \sum_{i=1}^m \ln f(t, a_2), \quad (3.92)$$

where

$$A''_2 = \frac{\partial^2 A_2}{\partial a_2^2} = \int_{t_{min}}^{\infty} g(t) \frac{\partial^2 f(t, a_2)}{\partial a_2^2} dt. \quad (3.93)$$

For two-parameter distribution functions typical of astronomical applications, it is very common for the last term on the right hand side of Eq. 3.92 to vanish. This is true for the exponential and power-law functions. Under this assumption, the variance can be simplified and written

$$\sigma_{a_2}^2 = \frac{A_2}{m} \left[ A''_2 - \frac{A'_2}{A_2} \right]^{-1} \quad (3.94)$$

where we have invoked Eq. 3.82. The ratio  $A'_2/A_2$  should be evaluated directly from the data using Eq. 3.90.

The normalization of the distribution function,  $a_1$ , is then determined from

$$\int_{t_{min}}^{\infty} n(t, \mathbf{a}) dt = a_1 \int_{t_{min}}^{\infty} f(t, a_2) dt = M, \quad (3.95)$$

where the constant  $M$  is taken to be the number of data points,  $m$ , or another quantity related to the survey. In certain applications,  $M$  is taken to be the measured redshift density of absorber,  $\mathcal{N}(z)$  (see § 9.3). If  $M$  is a measured quantity, then the normalization constant will have an associated uncertainty,  $\sigma_M$ . Note that the sensitivity function,  $g(t)$  does not appear in the integral of Eq. 3.95, because the value of  $a_1$  is the normalization of the distribution function and not of the probability distribution function. The variance of  $a_1$  is given by

$$\sigma_{a_1}^2 = a_1^2 \left[ \left( \frac{\sigma_M}{M} \right)^2 + \sigma_{a_2}^2 \left( \frac{B'_2}{B_2} \right)^2 \right], \quad (3.96)$$

where we introduce the notation

$$B_2 = \int_{t_{min}}^{\infty} f(t, a_2) dt \quad (3.97)$$

and

$$B'_2 = \frac{\partial B_2}{\partial a_2} = \int_{t_{min}}^{\infty} \frac{\partial f(t, a_2)}{\partial a_2} dt \quad (3.98)$$

The lower limit of integration,  $t_{min}$ , can be replaced with zero in certain applications, even if the sensitivity function,  $g(t)$ , vanishes for  $t_i < t_{min}$ . This can simplify the definite integrals to a form in which they can be evaluated analytically. Similarly, the upper limit of integration, set to  $\infty$  for this example, may be replaced with  $t_{max}$ , the maximum value of the observed data. This may eliminate improper integrals that diverge. Using  $t_{max}$  normalizes the distribution function to the observed data domain, which may be preferable in instances when an upper cutoff in the observable is physically motivated.

### 3.5.2 Example: Gaussian Distribution

Consider the Gaussian distribution function given in Eq. 3.67 to describe  $m$  observations of the quantity  $t$  in a survey with sensitivity  $g(t)$  for each  $t_i$ , for which the minimum data value is  $t_{min}$ . The likelihood function is

$$\ln \mathcal{L} = \ln \prod_{i=1}^m \frac{1}{I(\mathbf{a})} g(t_i) a_1 \exp \left[ - \left( \frac{t_i - a_2}{a_3} \right)^2 \right], \quad (3.99)$$

where  $I(\mathbf{a})$  is the normalization as given in Eq. 3.77,

$$I(\mathbf{a}) = \int_{t_{min}}^{\infty} g(t) a_1 \exp \left[ - \left( \frac{t - a_2}{a_3} \right)^2 \right] dt. \quad (3.100)$$



Note that the Gaussian is a three-parameter distribution function; only under ideal conditions can the formalism for the two-parameter estimation be employed. We will address this within the context of this example.

From Eq. 3.99, we have

$$\ln \mathcal{L} = \sum_{i=1}^m \ln \left( \frac{a_1}{I(\mathbf{a})} \right) + \sum_{i=1}^m \ln g(t_i) - \sum_{i=1}^m \left( \frac{t_i - a_2}{a_3} \right)^2. \quad (3.101)$$

Simplifying the first term on the right hand side yields

$$\sum_{i=1}^m \ln \left( \frac{a_1}{I(\mathbf{a})} \right) = -m \ln A, \quad (3.102)$$

where  $I(\mathbf{a}) = a_1 A$ , with

$$A = \int_{t_{min}}^{\infty} g(t) \exp \left[ - \left( \frac{t - a_2}{a_3} \right)^2 \right] dt. \quad (3.103)$$

The integral  $A$  is a number that depends only upon  $t_{min}$ ,  $a_2$ , and  $a_3$ . Rewriting Eq. 3.101, gives

$$\ln \mathcal{L} = -m \ln A + \sum_{i=1}^m \ln g(t_i) - \sum_{i=1}^m \left( \frac{t_i - a_2}{a_3} \right)^2. \quad (3.104)$$

The manipulation from Eq. 3.101 to 3.104 to eliminate the normalization is a standard procedure always resulting in the term  $-m \ln A$ ; it is universally invoked.

In a real survey, where  $g(t)$  is determined by the data and does not have a functional form,  $A$  would need to be determined using numerical integration. For purposes of illustration, consider a “perfect” survey with  $g(t) = 1$  for all  $t_i$ . In this survey  $t_{min} = 0$ , and each term  $\ln g(t_i)$  appearing in the sum of Eq. 3.104 vanishes, giving

$$\sum_{i=1}^m \ln g(t_i) = 0.$$

To obtain  $a_2$ , we differentiate Eq. 3.104 and equate to zero,

$$\frac{\partial \ln \mathcal{L}}{\partial a_2} = 0 = -\frac{m}{A} \frac{\partial A}{\partial a_2} - \frac{2}{a_3^2} \left[ \sum_{i=1}^m (t_i - a_2) \right]. \quad (3.105)$$

Differentiating Eq. 3.103 with respect to  $a_2$ , where we have assumed  $g(t) = 1$  for all  $t$  and  $t_{min} = 0$ , we have

$$\frac{\partial A}{\partial a_2} = \frac{2}{a_3^2} \int_0^\infty (t - a_2) \exp \left[ - \left( \frac{t - a_2}{a_3} \right)^2 \right] dt = \frac{\Gamma(1)}{2} = 0, \quad (3.106)$$

where  $\Gamma$  is the Gamma function, and we have used the property that  $\Gamma(n) = (n-1)!$ , when  $n$  is an integer. Thus, Eq. 3.105 simplifies to

$$\sum_{i=1}^m t_i - m a_2 = 0, \quad (3.107)$$

which gives

$$a_2 = \frac{1}{m} \sum_{i=1}^m t_i = \langle t \rangle. \quad (3.108)$$

The variance in  $a_2$  is given by Eq. 3.82. Carrying out the required differentiation gives

$$\sigma_{a_2}^2 = \frac{a_3^2}{2} \frac{1}{m}, \quad (3.109)$$

which is precisely the relationship expected for normally distributed data with variance  $a_3^2/2$ . Note that  $\sigma_{a_2} \propto 1/\sqrt{m}$ .

Similar mathematical manipulation of Eq. 3.104 for  $a_3$  yields

$$a_3^2 = \frac{1}{m} \sum_{i=1}^m (t_i - \langle t \rangle)^2, \quad (3.110)$$

which is also the expectation for normally distributed data.

The normalization constant,  $a_1$  is then obtained from

$$I(\mathbf{a}) = m = \int_0^\infty a_1 \exp \left[ - \left( \frac{t - a_2}{a_3} \right)^2 \right] dt, \quad (3.111)$$

where we have arbitrarily chosen the number of observed data points,  $m$ , for the normalization. Carrying out the integration yields,

$$a_1 = \frac{m}{a_3} \frac{2}{\sqrt{\pi}}, \quad (3.112)$$

with uncertainty

$$\sigma_{a_1} = \frac{m}{a_3^2} \frac{2}{\sqrt{\pi}} \sigma_{a_3}. \quad (3.113)$$

Had the above assumption of a “perfect survey” not been applied,  $\partial A/\partial a_2$  would not have vanished and would have a dependence upon  $a_3$ . Similarly,  $\partial A/\partial a_3$  would not have vanished and would have dependence upon  $a_2$ . In such cases, the derivatives must be numerically root solved simultaneously,

$$\begin{bmatrix} \frac{\partial \ln \mathcal{L}}{\partial a_2} \\ \frac{\partial \ln \mathcal{L}}{\partial a_3} \end{bmatrix} = 0. \quad (3.114)$$

### 3.6 $\chi^2$ Minimization

Consider a situation in which  $m$  data pairs,  $(x_i, y_i)$ , with  $i = 1, 2, 3, \dots, m$ , are to be modeled by a functional relationship

$$f(x) = f(x, \mathbf{a})$$

where  $\mathbf{a}$  are  $n$  adjustable parameters,  $a_j$ , with  $j = 1, 2, 3, \dots, n$ . If the uncertainties in each of the  $y_i$  is  $\sigma_i$ , then, based upon the same principles as the maximum likelihood estimation, the probability that the data set is described by the function  $f(x)$  is proportional to

$$\ln P \propto \chi^2(\mathbf{a}) = \sum_{i=1}^m \left( \frac{y_i - f(x_i; \mathbf{a})}{\sigma_i} \right)^2. \quad (3.115)$$

Minimizing  $\chi^2$  is equivalent to maximizing the joint probability function,  $P$ , which is achieved by differentiating Eq. 3.115 for each  $a_j$

$$\frac{\partial \chi^2}{\partial a_j} = -2 \sum_{i=1}^m \left( \frac{y_i - f(x_i; \mathbf{a})}{\sigma_i^2} \right) \left( \frac{\partial f(x_i; \mathbf{a})}{\partial a_j} \right), \quad (3.116)$$

and equating to zero. Obtaining Eq. 3.116 for each  $j$  yields a set of  $n$  equations, one for each unknown  $a_j$ . The matrix comprising the terms  $\partial f(x_i; \mathbf{a})/\partial a_j$ , is called the Jacobian, which must be supplied via analytical differentiation or via numerical computation during the minimization process.

The model must be started with an initial estimate of the  $\mathbf{a}$ . At each iteration toward root solving the matrix given by Eq. 3.116, the next estimate of each parameter is given by the increment  $\Delta a_k$ , which is solved for

in the matrix equation

$$\sum_{k=1}^n \alpha_{jk} \Delta a_k = -\frac{1}{2} \frac{\partial \chi^2}{\partial a_j}, \quad (3.117)$$

where the sum is over the  $n$  parameters, and the partial differentiation is given by the right hand side of Eq 3.116. The  $\alpha_{jk}$  are given by

$$\alpha_{jk} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial a_j \partial a_k} = \sum_{i=1}^m \frac{1}{\sigma_i^2} \left[ \frac{\partial f(x_i; \mathbf{a})}{\partial a_j} \frac{\partial f(x_i; \mathbf{a})}{\partial a_k} \right], \quad (3.118)$$

which is called the curvature matrix.

Once  $\chi^2$  is minimized, the variance in the  $\mathbf{a}$  are approximated by the diagonal elements,

$$\sigma_{a_j}^2 = C_{jj}, \quad (3.119)$$

of the co-variance matrix, which is computed from the inverse of the curvature matrix

$$[C] = [\alpha]^{-1}. \quad (3.120)$$

This treatment of the variances ignores the possibility of correlated errors. That is, the 68% confidence interval for a given parameter is estimated assuming all other parameters are fixed, unadjustable quantities that are true representations of the distribution from which the observed data were drawn. More sophisticated methods for estimating the confidence level of the model are expounded upon in the book *Numerical Recipes* by Press et al. (1996).

### 3.6.1 The Reduced $\chi^2$

In practice, it is the reduced  $\chi^2$  that is utilized, since it conveniently converges to unity for a “good” fit. Simply stated, the reduced  $\chi^2$  is normalized by the number of degrees of freedom characteristic of the model fit and data pairs. The degrees of freedom,  $\nu$ , of a least squares fitting procedure is

$$\nu = m - n \quad (3.121)$$

where  $m$  is the number of data points and  $n$  is the number of free parameters. For example, a single Gaussian function is characterized by three parameters, so  $n = 3N_g$ , where  $N_g$  is the number of Gaussian components used to model the data. A “good”  $\chi^2$  minimization should yield

$$\chi^2 \simeq \nu, \quad (3.122)$$

which arises when each term in the sum of Eq. 3.115 is nearly unity. The reduced  $\chi^2$  is defined as

$$\chi_\nu^2 = \frac{\chi^2}{m - n} = \frac{V_{mod}}{V_{data}}, \quad (3.123)$$

where

$$V_{mod} = \sum_{i=1}^m [y_i - f(x_i; \mathbf{a})]^2, \quad (3.124)$$

is the variance in the model and

$$V_{data} = \sum_{i=1}^m \sigma_i^2. \quad (3.125)$$

is variance in the data. For a “good” model fit to the data,

$$\chi_\nu^2 \simeq 1. \quad (3.126)$$

### 3.6.2 Example: Gaussian Distribution

Assume the model to be fitted to the data is the sum of  $N_g = n/3$  Gaussian functions

$$f(x; \mathbf{a}) = \sum_{k=1,3}^n a_k \exp \left[ - \left( \frac{x - a_{k+1}}{a_{k+3}} \right)^2 \right], \quad (3.127)$$

where the index  $k$  is incremented in intervals of 3. Most fitting routines require that the derivatives used for  $\chi^2$  minimization (in Eqs. 3.116, 3.117, and 3.118) are provided in addition to the function  $f(x; \mathbf{a})$ . Carrying out the differentiation of Eq. 3.127, the Jacobian is

$$\begin{aligned} \frac{\partial f}{\partial a_1} &= \exp \left[ - \left( \frac{x - a_2}{a_3} \right)^2 \right] \\ \frac{\partial f}{\partial a_2} &= 2 \frac{a_1}{a_3} \left( \frac{x - a_2}{a_3} \right) \exp \left[ - \left( \frac{x - a_2}{a_3} \right)^2 \right] \\ \frac{\partial f}{\partial a_3} &= 2 \frac{a_1}{a_3} \left( \frac{x - a_2}{a_3} \right)^2 \exp \left[ - \left( \frac{x - a_2}{a_3} \right)^2 \right] \end{aligned} \quad (3.128)$$

for the first Gaussian function, and

$$\frac{\partial f}{\partial a_{k-2}} = \exp \left[ - \left( \frac{x - a_{k-1}}{a_k} \right)^2 \right]$$

$$\begin{aligned}\frac{\partial f}{\partial a_{k-1}} &= 2 \frac{a_{k-2}}{a_k} \left( \frac{x - a_{k-1}}{a_k} \right) \exp \left[ - \left( \frac{x - a_{k-1}}{a_k} \right)^2 \right] \\ \frac{\partial f}{\partial a_k} &= 2 \frac{a_{k-2}}{a_k} \left( \frac{x - a_{k-1}}{a_k} \right)^2 \exp \left[ - \left( \frac{x - a_{k-1}}{a_k} \right)^2 \right]\end{aligned}\quad (3.129)$$

for the remaining Gaussian functions for  $4 \leq k \leq 3N_g$ .

In practice, the Jacobian is inserted into Eq. 3.116 from which the parameter adjustment ( $\Delta a_k$ , Eq. 3.117) is made for the next step in the iterative process. Once all  $\Delta a_k \simeq 0$  within some specified fractional tolerance, then the errors in  $a_k$  are determined from the Jacobian inserted into the curvature matrix, which is inverted to obtain the co-variance matrix.

### 3.6.3 Fitters

Not complete.

The first step of data modeling is to choose a function appropriate to the distribution of the data. Having done so, the next step is to implement a stable routine to carry out  $\chi^2$  minimization. There are many on the market, but not all are stable. Much discussion on the details of implementation can be found in Bevington & Robinson (2003)

From the author's personal experience, two recommended "canned" routines are:

1. DNLS1E, provided by the Netlib Repository
2. MRQMIN provided by *Numerical Recipes*

Both use a modified Levenberg–Marquard method and are well documented.

According to the website at [www.netlib.org](http://www.netlib.org), "the Netlib Repository contains freely available software, documents, and databases of interest to the numerical, scientific computing, and other communities. The repository is maintained by AT&T Bell Laboratories, the University of Tennessee and Oak Ridge National Laboratory, and by colleagues world-wide." Routine DNLS1E is part of the SLATEC library, which comprises several subroutines. Routine DNLS1E is somewhat more cumbersome to implement than is routine MRQMIN, but it is incredibly powerful, stable, and robust. Positive attributes include machine precision, the option to compute the Jacobian numerically or to include the Jacobian explicitly. This is particularly useful if the Jacobian is impossible to express in analytical form. One shortcoming, is that DNLS1E does not return the co-variance matrix for error estimations; the user must compute the curvature matrix and obtain its inverse following  $\chi^2$  minimization. If the Jacobian cannot be provided, then the curvature

matrix must be obtained numerically, which can be also be cumbersome. However, if you are looking for a “no-fail” routine, DNLS1E is your choice.

HERE [www.nr.com](http://www.nr.com)

On the other hand, routine MRQMIN is very easy to implement and, for most all problems, is stable. It does not handle a large data volume and a large number of parameters as well as routine DNLS1E. It does, however, return the covariance matrix.

## References

- Bevington, P. & Robinson, D. K. 2003, “Data Reduction and Error Analysis for the Physical Sciences,” (McGraw Hill : New York)
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., Metcalf, M. 1996, “Numerical Recipes,” (Cambridge University Press : Cambridge)

