

## 11. Special Topics

- *transforming distribution functions*
- *moments and associated uncertainties*
- *nomograms*

### 11. Special topics

#### a. Introduction

This chapter contains some miscellaneous topics that did not fit well elsewhere. Most of this chapter can be considered supplementary material, although the sections on transforming distribution functions, the log-normal distribution, and moments include material that is often encountered in atmospheric science.

#### b. Using units in equations

At one time, particularly in engineering usage, most equations were expressed in forms that required a particular set of units. An example might be the following form for the equivalent potential temperature:

$$\Theta_e = (T + 273.15) \left( \frac{1000}{(p - e)} \right)^{2/7} \exp \left\{ \frac{2.5r_v}{(T + 273.15)} \right\} . \quad (11.1)$$

This form gives the equivalent potential temperature in Kelvin if the temperature  $T$  is in °C, the pressure  $p$  and water vapor pressure  $e$  are in mb, and the mixing ratio  $r_v$  is expressed in g/kg. This form of equation ties the expression to a particular set of units, and often obscures the underlying physical relationships. For example, in this case the factor

2.5 arises from the dependence on the latent heat of vaporization, the specific heat of air at constant pressure, and conversion of units. Because the latent heat varies significantly with temperature, (11.1) only applies to a single temperature and it is not obvious without further explanation how this must be modified for other temperatures. Because of these problems, and the lack of flexibility in units, this form of equation should be avoided.

Instead, current usage favors expressing equations in dimensionless form. For example, the above equation in dimensionless form is

$$\Theta_e = T \left( \frac{p_0}{p - e} \right)^{R_d/C_p} \exp \left\{ \frac{L_v r_v}{C_p T} \right\} \quad (11.2)$$

where  $R_d$  is the gas constant for dry air,  $L_v$  is the latent heat of vaporization,  $C_p$  the specific heat of air at constant pressure, and  $T$  is the absolute temperature. Where constants occur, like  $p_0$  in this equation, they are represented by symbols and the value is then given in some system of units, in this case  $p_0=1000$  mb. This form of the equation is dimensionally consistent in any system of units. In this form, arguments like that of the exponential in (11.2) will be dimensionless, and the units used should balance in any correct evaluation of the expression.

A common error, or at least a common imprecision in usage, is to use quantities with dimensions in logarithmic, exponential, or trigonometric functions. For example, the units to be associated with an expression like  $\ln(r)$  are unclear, because the value of the expression changes when the units used to express  $r$  change. It is preferable to use expressions like  $\ln(r/r_0)$  where  $r_0$  is a specified reference radius, because then the expression is independent of the system of units used even when the value of  $r_0$  is expressed in a particular system (e.g.,  $r_0=1\mu\text{m}$ ). A dimensionally consistent expression can always be found for equations involving logarithms or exponential or trigonometric functions. For precise and clear usage, these forms are preferred, although they are not always adhered to in common usage. If this convention were applied strictly, an expression like

$$\ln(\Theta) = \ln(T) - \frac{R_d}{C_p} \ln(p) + \text{constant} \quad (11.3)$$

would be changed to

$$\ln \left( \frac{\Theta}{T} \right) = \frac{R_d}{C_p} \ln \left( \frac{p}{p_0} \right), \quad (11.4)$$

because in this form the dimensionally ambiguous terms like  $\ln(\Theta)$  are avoided. All arguments in logarithmic, exponential, trigonometric, etc., functions should be dimensionless.<sup>1</sup>

Balancing dimensions in dimensionless equations is a powerful check on the consistency of units, and provides a convenient basis for conversion among units.

---

<sup>1</sup> For an exception, see Example 11.2.

**EXAMPLE 11.1:**

The pattern in the following example, if followed consistently with dimensionless equations, handles all unit conversion problems and in addition provides a valuable check on the consistency of the equation and the values used.<sup>2</sup>

The problem is to find the equivalent potential temperature using (11.2), for the following conditions: temperature 20°C, pressure 27 inches of mercury, and dew point 18°C. In addition, the following values are known:  $C_p=1005 \text{ J kg}^{-1} \text{ K}^{-1}$ ,  $L_v = 586 \text{ cal/g}$ , and the gas constant  $R_d$  is  $287 \text{ J kg}^{-1} \text{ K}^{-1}$ . At the dewpoint temperature, the saturation vapor pressure is  $e=20.630 \text{ mb}$ . The molecular weights of water ( $M_w$ ) and of air ( $M_a$ ) are  $18.015 \text{ g mol}^{-1}$  and  $28.964 \text{ g mol}^{-1}$ , respectively.

Units can be converted as needed by multiplying parts of the equation by ratios of units that are unity, such as  $(100 \text{ cm})/(1 \text{ m})$ , so as to cancel all units. In this example, that is accomplished as follows:

Step 1: Find the mixing ratio from the vapor pressure and the pressure:

$$\begin{aligned} r_v &= \frac{M_w}{M_a} \frac{e}{(p - e)} \\ &= \frac{18.015 \text{ g mol}^{-1}}{28.964 \text{ g mol}^{-1}} \times \frac{20.630 \text{ mb}}{27 \text{ in. Hg} \times \frac{1013.25 \text{ mb}}{29.921 \text{ in. Hg}} - 20.630 \text{ mb}} \\ &= 0.622 \times \frac{20.630 \text{ mb}}{914.33 \text{ mb}} \\ &= 0.01436 \text{ [dimensionless]} \end{aligned} \quad (11.5)$$

Note that the pressure ratio  $1013.25 \text{ mb} : 29.921 \text{ in. Hg}$  has been used to convert the pressure to mb; both numerator and denominator correspond to one atmosphere, so this ratio is easier to remember and find than the resulting conversion factor of  $33.86 \text{ mb} / (\text{in. Hg})$ .

Step 2: Evaluate the exponential factor. This should be dimensionless.

$$\begin{aligned} \frac{L_v r_v}{C_p T} &= \frac{586 \frac{\text{cal}}{\text{g}} \times 0.01436}{1005 \frac{\text{J}}{\text{kgK}} + 293.15 \text{ K}} \\ &= \frac{586 \frac{\text{cal}}{\text{g}} \times \frac{4.184 \text{ J}}{\text{cal}} \times 0.01436}{1005 \frac{\text{J}}{\text{kgK}} \times \frac{1 \text{ kg}}{1000 \text{ g}} \times 293.15 \text{ K}} \\ &= 0.1195 \end{aligned} \quad (11.6)$$

<sup>2</sup>This example may appear straightforward and standard, but it has been the author's experience that even at the graduate level a surprising number of students benefit from this systematic approach.

Step 3: Evaluate the expression (11.2):

$$\Theta_e = 293.15K \left( \frac{1000mb}{27in. Hg \times \frac{1013.25mb}{29.92in. Hg} - 20.630mb} \right)^{\frac{287Jkg^{-1}K^{-1}}{1005Jkg^{-1}K^{-1}}} e^{0.1195} \quad (11.7)$$

$$= 338.9K$$

In these examples, the units are carried in the equations and evaluated to obtain the units of the answer. Whenever units fail to cancel, this indicates that a conversion factor is needed (or that incorrect units have been used for one of the quantities in the equation).

Atmospheric science journals and most scientific journals now require use of International System (SI) units. If that system is used consistently, it is not necessary to carry the units in dimensionless equations; units will automatically cancel if SI units are used in a correct nondimensional equation. It is still useful to carry units and check for consistency as in the preceding example.

### c. Nomograms

With the current availability of computers, there has been a decline in reliance on graphic methods for the solution of equations, and the material in this section is mostly unknown to the current generation of scientists. In engineering, graphic presentations have long been important tools, and they still serve important roles in atmospheric science (e.g., in thermodynamic diagrams). This optional section on nomograms is included here partly because this material is seldom seen and yet is a useful technique for the rapid, although approximate, solution of many equations. The author often uses nomograms in the course of airborne research, where it is difficult to solve equations with the speed needed to make flight decisions.

A nomogram is a graphic representation of the solution of an equation, usually involving more than 2 variables, where the solution can be read by using a straightedge to connect points representing two variables and then determine the value of a third. The equation  $A+B=C$  can be used to illustrate the geometrical basis for nomograms. Because  $(A+B)/2=C/2$ , the average of  $A$  and  $B$  is  $(1/2)C$ , and the solution can be represented by three lines as in Fig. 11.1 where the scale relationship used for the variable  $C$  is half that used for  $A$  and  $B$ .

The diagram becomes more useful and less trivial if  $A$ ,  $B$ , and  $C$  are functions of other variables, e.g., as in the relationship  $A(x) + B(y) = C(z)$ , in which case the axes can be labeled according to the values of  $x$ ,  $y$ , and  $z$  that give the corresponding values of  $A$ ,  $B$ , and  $C$ . Product forms like  $A(x)B(y) = C(z)$  can be constructed by using logarithmic scales, because  $\log(A) + \log(B) = \log(C)$ .

It is also possible to construct nomograms representing relationships among more than three variables. For example,  $A+B+C+D = E$  can be represented by an initial nomogram to give the sum  $A+B$ , followed by another that uses the resulting point from the first as input to give  $(A+B)+C$ , followed by a fourth to complete the sum. Another technique often used in such sequences is to incorporate an ordinary graph as a way of representing one relationship or to change units or scales, as in the example that follows.

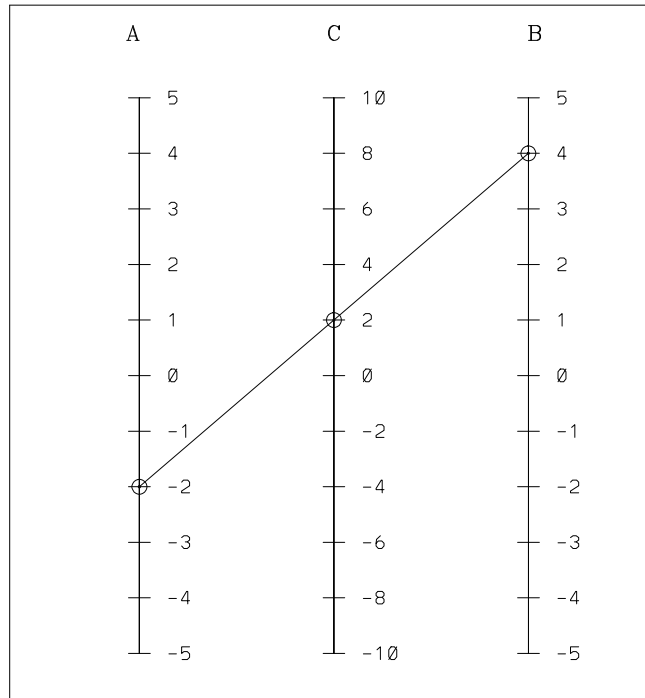


Fig. 11.1: Nomogram representing the solution of the equation  $A + B = C$ .

Figure 11.2 shows an example. In this case, the nomogram was constructed to represent the wet-bulb equation (cf. Example 9.1):

$$T_{wb} = T + \frac{L_v}{C_p}(r - r_s(T_{wb})) \quad (11.8)$$

where  $L_v$  is the latent heat of vaporization,  $C_p$  the specific heat of air at constant pressure,  $r$  the water vapor mixing ratio, and  $r_s(T_{wb})$  the saturation mixing ratio at the wet-bulb temperature. The mixing ratio is a function of the dewpoint temperature  $T_d$  and the pressure  $p$ :

$$r = \epsilon \frac{e_s(T_d)}{(p - e_s(T_d))} \quad (11.9)$$

where  $\epsilon$  is the ratio of molecular weights of water and air and  $e_s(T_d)$  is the saturation vapor pressure at the temperature  $T_d$ . The wet-bulb equation can then be rewritten as

$$T_{wb} + \frac{L_v}{C_p}r_s(T_{wb}) = T + \frac{L_v}{C_p}r_s(T_d) . \quad (11.10)$$

The left side of this equation is the quantity plotted on the left side of the nomogram, and the right side is the last term in this equation; these two terms are labeled by the appropriate values of  $T_{wb}$  and  $T_d$ , respectively, and in both cases a set of lines is plotted to show the pressure dependence of the terms.

For a more extensive discussion of nomograms, including those with more complicated geometry, see Menzel (1960), pp. 141..

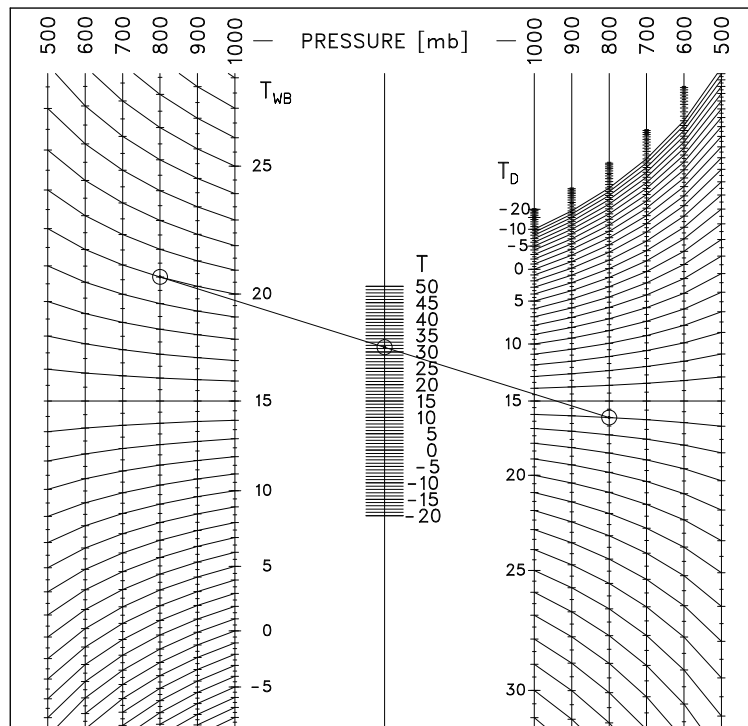


Fig. 11.2: Nomogram representing the relationship among temperature ( $T$ ), dewpoint ( $T_D$ ), and isobaric wet-bulb temperature ( $T_{WB}$ ) as a function of pressure. The circled points connected by the straight line show corresponding values for a pressure of 800 mb.

d. *Transforming distribution functions*

A differential probability distribution function  $f(x)$ , to describe the probability associated with events having (possibly multidimensional) coordinate  $x$ , must have these properties:

- It must be normalized so that the integral over the defined space gives unity:

$$\int_{-\infty}^{\infty} f(x) dx = 1 . \quad (11.11)$$

- It must be positive definite:

$$f(x) \geq 0 \quad \text{for all } x. \quad (11.12)$$

- The integral between specified limits must give the probability that events will occur in that interval:

$$\int_a^b f(x) ds = P(a < x < b) . \quad (11.13)$$

One often needs to transform such distribution functions to find probability distribution functions in terms of new variables. Consider the transformation from variable  $x$  to variable  $y$  where the relationship between  $y$  and  $x$  is known ( $y(x)$ ). Then, if the probability

distribution function  $g(y)$  represents the same probability distribution function in terms of the variable  $y$ , to preserve the relationship of integrals to probability it must be true that

$$\int_{x_1}^{x_2} f(x) dx = \int_{y(x_1)}^{y(x_2)} g(y) dy . \quad (11.14)$$

Differentiating gives:

$$\begin{aligned} \frac{d}{dx_2} \left[ \int_{x_1}^{x_2} f(x) dx \right] &= f(x_2) = \frac{d}{dx_2} \left[ \int_{y(x_1)}^{y(x_2)} g(y) dy \right] \\ &= \frac{d}{dy_2} \left[ \int_{y(x_1)}^{y(x_2)} g(y) dy \right] \frac{dy_2}{dx} \Big|_{x_2} \\ &= g(y_2) \frac{dy_2(x_2)}{dx_2} = g(y) \frac{dy}{dx} \Big|_{x_2} . \end{aligned} \quad (11.15)$$

A shorthand way of representing this relationship is

$$f(x) dx = g(y) dy \quad (11.16)$$

where  $dx$  and  $dy$  are understood to be corresponding infinitesimal intervals.<sup>3</sup>

The transformation between variables is then, from (11.16),

$$f(x) = g(y) \frac{dy}{dx} . \quad (11.17)$$

**EXAMPLE 11.2:** From a droplet distribution function in terms of radius,  $n(r)$ , find the distribution functions  $f(x)$  in terms of droplet mass  $x$  and  $g(\log_{10}(r))$  in terms of the base-10 logarithm of the radius.

Because

$$x = \frac{4}{3} \pi r^3 \rho_w,$$

$$dx = 4\pi r^2 \rho_w dr$$

so

$$f(x) = \frac{n(r)}{4\pi r^2 \rho_w} . \quad (11.18)$$

The second part of the problem might be considered poorly stated. As explained in section (11.b), for dimensional consistency it might appear that the distribution function should be presented in terms of a dimensionless argument. The appropriate distribution

<sup>3</sup> A convenient way of representing this relationship is with the cumulative distribution function  $F(x^*)$ , giving the probability for all values of  $x$  smaller than  $x^*$ . If the corresponding cumulative distribution function in terms of  $y$  is  $G(y)$ , then  $G(y(x^*)) = F(x^*)$ . But  $dF(x)/dx = f(x)$  and  $dG(y)/dy = g(y)$  so, for corresponding intervals such that  $dF(x) = dG(y)$  it must be true that  $f(x)dx = g(y)dy$ .

function would then have the form  $g(\log_{10}(r/r_0))$  where  $r_0$  is a constant reference radius. Then

$$\begin{aligned} \frac{d(\log_{10}(r/r_0))}{dr} &= \frac{d(\log_{10}(r/r_0))}{d(\ln_e(r/r_0))} \times \frac{d(\ln_e(r/r_0))}{dr} \\ &= \frac{1}{\ln_e(10)} \frac{1}{r} \end{aligned} \quad (11.19)$$

so

$$g(\log_{10}(r/r_0)) = \ln_e(10) r n(r) \approx 2.30 r n(r) . \quad (11.20)$$

Because the resulting distribution function does not depend on the value selected for  $r_0$ , the reference dimension  $r_0$  is almost universally omitted from logarithmic distribution functions like this. There is no dimensional inconsistency in doing so because the logarithm is only offset by a fixed amount depending on the choice of units and hence the change in the logarithm remains the same for a given interval regardless of the choice of units. For example, if  $r$  is expressed in  $\mu\text{m}$  and the units of  $n(r)$  are  $\text{cm}^{-3} \mu\text{m}^{-1}$ , the units of  $g(\log_{10}(r/r_0))$  are  $\text{cm}^{-3}$  per logarithmic interval in radius, and this is the same for any other choice of unit for  $r$ .

A cumulative distribution function can be defined as the integral of the differential distribution function, so that

$$N(r) = \int_{-\infty}^r n(r') dr' \quad (11.21)$$

is the number of events with radius (for example) smaller than  $r$ . Then the derivative of this cumulative distribution function with respect to any variable is the differential probability distribution function for that variable. For example,

$$\frac{dN(\log_{10}(r))}{d\log_{10}(r)} = g(\log_{10}(r)) . \quad (11.22)$$

For clarity, plots of differential distribution functions are often labeled with the form on the left side, rather than the right side, of this equation.

It is sometimes useful to display moments of a distribution function. For example, if  $x$  is the droplet mass as before, the distribution of mass in a droplet size distribution can be displayed as

$$h(x) = x f(x) = \frac{4}{3} \pi r^3 \rho_w \frac{n(r)}{r \pi r^2 \rho_w} = \frac{r n(r)}{3}$$

to produce a plot for which area is proportional to mass.

**EXERCISE 11.1:** Show that, if a logarithmic scale is needed, the appropriate function to show the distribution ( $x dN/d\log_{10}(r)$ ) in mass is  $(4/3)\pi r^4 \ln_e(10) \rho_w n(r)$ .

## e. Moments and associated uncertainties

Moments of observed distribution functions are often needed because they represent physical quantities like mass. For example, the liquid water content  $x$  in a raindrop size distribution is obtained using the third moment of the distribution:

$$x = \left( \sum_i n_i \frac{4}{3} \pi r_i^3 \rho_w \right) / S = \frac{N}{S} \frac{4}{3} \pi \rho_w \langle r^3 \rangle \quad (11.23)$$

where  $n_i$  is the number in class  $i$ , having radius  $r_i$ ,  $S$  is the sample volume from which the sample was measured, and  $N = \sum_i n_i$  is the total number. In this expression, the angle brackets denote the average value for the sample. Other expressions involving moments of the distribution function are the radar reflectivity  $Z$ , the geometric cross-sectional area  $A$ , and the “effective radius”  $r_e$  sometimes used to represent radiation effects in terms of the liquid water content:

$$Z = \frac{N}{S} \langle d^6 \rangle \quad (11.24)$$

$$A = \frac{N}{S} \pi \langle r^2 \rangle \quad (11.25)$$

$$r_e = \frac{\langle r^3 \rangle}{\langle r^2 \rangle} \quad (11.26)$$

where  $d = 2r$  is the diameter. Another related quantity is the rainrate  $R$ :

$$R = \frac{N}{S} \frac{4}{3} \pi \rho_w \langle r^3 v_t(r) \rangle \quad (11.27)$$

where  $v_t(r)$  is the (size-dependent) fall speed of the raindrops.

The measurement errors for quantities like these are sometimes highly correlated. If the scatter in the measurements is caused primarily by counting statistics and not by natural variability, there still may appear to be a strong correlation among these moments just because the fluctuations are correlated. For example, some investigators have used dropsize distributions to determine correlations between radar reflectivity  $Z$  and rainfall  $R$  without taking into account the correlation that arises just from the correlation between measurement errors. Relationships so determined cannot be used for their intended purpose (the measurement of rainfall by radar) because the physical relationship they sought to determine is masked by the correlated measurement errors.

Error limits for moments can be determined as follows. [Cf. Bury, 1975.] Consider the  $n$ -th moment:

$$\langle r^n \rangle_S = \frac{1}{N} \sum_i r_i^n \quad (11.28)$$

where the subscript  $S$  serves as a reminder that this moment is determined using the finite sample from region  $S$ , and therefore is an inexact representation of the larger sample. If that larger or “ensemble” (true) moment is  $\alpha^n$ , the variance in the estimated moment is

$$\begin{aligned} \text{Var}(\langle r^n \rangle_S) &= \langle (\langle r^n \rangle_S - \alpha^n)^2 \rangle \\ &= \langle \langle r^n \rangle_S^2 - (\alpha^n)^2 \rangle \\ &= \frac{1}{N^2} \langle \sum_i \sum_j r_i^n r_j^n \rangle - (\alpha^n)^2 \end{aligned} \quad (11.29)$$

If two variables  $x$  and  $y$  are uncorrelated,  $\langle xy \rangle = \langle x \rangle \langle y \rangle$ .<sup>4</sup> Therefore, if the measurements are uncorrelated, the variance in the estimate of the moment is

$$\begin{aligned} \text{Var}(\langle r^n \rangle_S) &= \frac{1}{N^2} [N \langle r^{2n} \rangle + N(N-1) \langle r^n \rangle^2] - (\alpha^n)^2 \\ &= \frac{1}{N} [\langle r^{2n} \rangle - \langle r^n \rangle^2] \end{aligned} \quad (11.30)$$

**EXAMPLE 11.3:** Find the uncertainty in radar reflectivity factor  $Z$  associated with a set of  $N$  measurements of raindrop diameters  $d_i$ . The measurement uncertainty in the radar reflectivity factor  $Z = (N/S) \langle d^6 \rangle$  has contributions from uncertainty in both  $N$  and in  $\langle d^6 \rangle$ . The basic formula for error propagation (2.16) leads to

$$\begin{aligned} \text{Var}(Z) &= \left(\frac{\partial Z}{\partial N}\right)^2 (\delta N)^2 + \left(\frac{\partial Z}{\partial \langle d^6 \rangle}\right)^2 (\delta \langle d^6 \rangle)^2 \\ &= \frac{\langle d^6 \rangle^2}{S^2} N + \frac{N^2}{S^2} \left[ \frac{1}{N} (\langle d^{12} \rangle - \langle d^6 \rangle^2) \right] \\ &= \frac{N}{S^2} \langle d^{12} \rangle = \frac{1}{S} \sum_i d_i^{12} \end{aligned} \quad (11.31)$$

**EXERCISE 11.2:** Show that the variance in an estimate of rainrate is

$$\text{Var}(R) = \left(\frac{\pi}{6S}\right)^2 N \langle d^6 v_t^2 \rangle \quad (11.32)$$

and that the covariance in the errors in  $Z$  and  $R$  is

$$\text{Cov}(Z, R) = \frac{\pi N}{6S^2} \langle d^9 v_t \rangle \quad (11.33)$$

Moments of distributions are often used to summarize characteristics of those distributions. If moments are calculated relative to the means, the standard deviation  $\sigma_x$  in  $x$  is  $\langle x^2 \rangle^{1/2}$ , the skewness  $s$  is  $\langle x^3 \rangle / \sigma_x^3$ , and the kurtosis is  $[\langle x^4 \rangle / \sigma_x^4 - 3]$ . With these definitions, the kurtosis of a Gaussian distribution is zero. Other definitions for these quantities are sometimes used, including the above definitions divided by 2 or, in the case of kurtosis, the preceding formula without the term  $(-3)$ .

<sup>4</sup> To show this, replace  $x$  and  $y$  by  $\langle x \rangle + x'$  and  $\langle y \rangle + y'$ , then take the product and average.

## f. Log-normal distributions

Distributions encountered in meteorology are notoriously non-Gaussian, usually because they have more extreme events than expected for a Gaussian distribution. However, a distribution that is normal in the logarithm of a parameter often provides a much improved fit to observations. Another advantage of the log-normal distribution is that it is positive-definite, so it is often useful for representing quantities that cannot have negative values. Log-normal distributions have proven useful as distributions for rainfall amounts, for the size distributions of aerosol particles or droplets, and for many other cases.

The general form for a log-normal distribution is

$$P\left(\ln\left(\frac{x}{x_0}\right)\right) = \frac{1}{\sigma_g \sqrt{2\pi}} e^{-\left(\frac{(\ln(x) - \ln(x_0))^2}{2\sigma_g^2}\right)} = \frac{1}{\sigma_g \sqrt{2\pi}} e^{-\left(\frac{\ln^2(x/x_0)}{2\sigma_g^2}\right)} \quad (11.34)$$

where  $\ln(x_0)$  is the central value for the distribution and  $\sigma_g$  is the geometric standard deviation. This will have a Gaussian shape like that shown in Fig. 3.1 when plotted with a logarithmic abscissa. Figure 11.3 shows some examples of this distribution function, for different values of the geometric standard deviation, plotted with a linear abscissa. The shape differs if the plot is constructed using base-10 logarithms, because the meaning of the standard deviation changes and the normalization factor changes. The value of  $\sigma'_g$  applicable to a distribution in terms of base-10 logarithms is obtained from  $\sigma_g$  for Napierian logarithms by the translation

$$\sigma'_g = \frac{\sigma_g}{\ln(10)}. \quad (11.35)$$

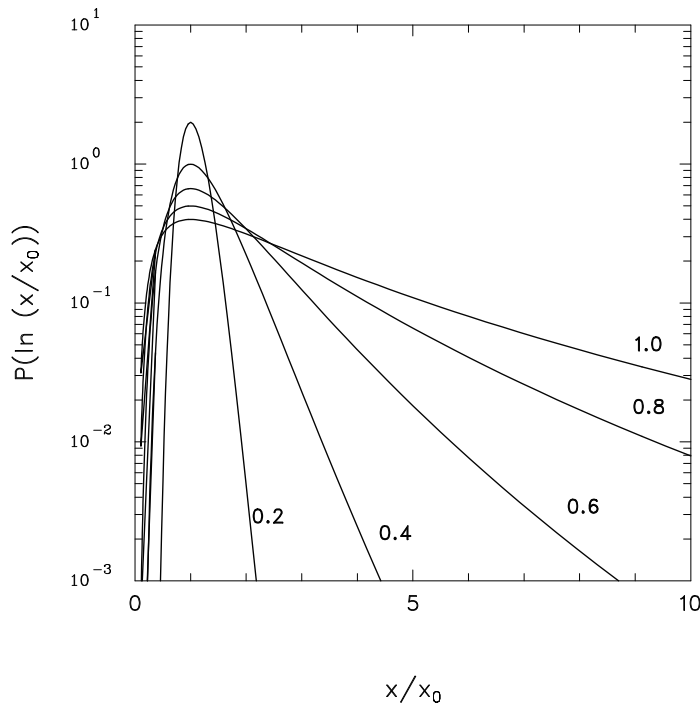


Fig. 11.3: The log-normal distribution function, for geometric standard deviations as labeled.

Because the distribution is skewed toward larger values, the mean value of  $x$  is larger than  $x_0$ . For the log-normal distribution constructed using Napierian logarithms,

$$\bar{x} = e^{\sigma_g^2/2} . \quad (11.36)$$

The variance in  $x$  is given in terms of  $x_0$  and  $\sigma_g$  by

$$\sigma_x^2 = x_0^2 e^{\sigma_g^2} (e^{\sigma_g^2} - 1) = (\bar{x})^2 (e^{\sigma_g^2} - 1) . \quad (11.37)$$

### *g. The Maximum-Entropy Approach to Data Analysis*

#### 1). Fundamentals

Chapter 3 outlined the maximum-likelihood approach and advocated that as a fundamental basis for data analysis. There it was argued that the best value of a parameter is that which gives the maximum probability that the set of observed values would be obtained in an experiment. Here we explore the logical foundation for that claim, and then extend the approach by incorporating powerful additional information. The resulting Maximum-Entropy Method (MEM) produces results that are sometimes far superior to those that don't incorporate this additional information, particularly in cases where the observations are a large set of numbers (as in an image or a variance spectrum).

The analysis of Chapter 3 considered a way of finding a best value of a parameter with the assumption that the distribution has a particular form. Here that analysis is expanded by including, as an element in the probability statement, a measure of our confidence that the functional form really is as we have assumed. For this, we can use the standard relationship applicable to conditional probabilities:

$$P(A, B) = P(A) \times P(B|A) = P(B) \times P(A|B) . \quad (11.38)$$

That is, the probability of results  $A$  and  $B$  occurring simultaneously is the probability of  $A$  occurring multiplied by the conditional probability that  $B$  occurs when  $A$  occurs. For example, if  $H$  represents a hypothesis,  $I$  the state of our knowledge before we do an experiment, and  $y$  a set of observations, then the set of observations contribute to the joint probability that  $H$  is true and  $y$  are measured, as follows:

$$P(H, y|I) = P(H|I) \times P(y|H, I) = P(y|I) \times P(H|y, I) . \quad (11.39)$$

Here  $P(H|I)$  represents our prior knowledge of  $H$ , based on prior information  $I$ . This may be our expectation that reasonable measurements of temperature should be in a particular range, for example, or any other information that could form the basis for rejecting a measurement as unreasonable. It may also represent our expectation that events will obey a particular probability distribution, or will vary smoothly with space or time. In all these cases, this prior knowledge conditions our estimate of the probability  $P(H, y|I)$ , representing the probability that we will observe the set of observations  $y$  and that  $H$  is true. The factor  $P(y|H, I)$  in (11.39) is the probability of obtaining the observations  $y$  under the assumption that  $H$  is true, and so represents the probability that is maximized in the maximum-likelihood solution. However, the factor we want to determine from evaluation of the experimental data is  $P(H|y, I)$ , the probability that  $H$

is true, given our observations and prior information. In terms that involve the maximum-likelihood solution, this is

$$P(H|y, I) = p(H|I) \frac{P(y|H, I)}{P(y|I)}, \quad (11.40)$$

which is Bayes' theorem. In the maximum-likelihood approach, we maximized  $P(y|H, I)$  and identified the result with the maximum-likelihood solution. Equation (11.40) shows that this selection of a best value should depend on our prior knowledge  $P(H|I)$  as well. (It does not depend on  $P(y|I)$  because this does not depend on  $H$ .  $P(y|I)$  is sometimes written as  $\sum_{H'} P(y|H', I)$  where the sum extends over all possible alternatives for  $H$ .)

Bayes' theorem thus describes the way in which newly acquired information ( $y$ ) modifies our prior estimate of the probability of a conclusion: the modification is proportional to the probability with which we would observe  $y$ , given that  $H$  is true and given our prior knowledge  $I$ . This is a mathematical statement describing calculations related to probability, but when applied to the analysis of experimental data Bayes' theorem introduces an explicit dependence of the answer on the prior information and hence, potentially, on the analyst.<sup>5</sup> This is often cited as a defect of the Bayesian approach: The result depends on the analyst and can vary depending on prior information, while we like to think that we strive for objective results that are independent of who does the analyses. This is of course never completely the case, but the maximum-entropy method introduces a degree of standardization into this choice by advocating a particular basis for the introduction of prior information.

The result (11.40) may seem not particularly useful in the case where a single parameter is to be determined. For example, if we are measuring a temperature and have no knowledge of what it should be, other than perhaps that it should be within some "reasonable" range, we might assign the prior probability  $P(H|I) = \text{constant}$  throughout that reasonable range, in which case Bayes' theorem states that the maximum-likelihood solution is also the most probable result for  $H$ . However, the method becomes nontrivial when  $H$  represents a more complicated result, such as an image or a variance spectrum or a droplet size distribution. We do have prior knowledge that is relevant in many of these cases. We may expect only a certain subset of possible images, or a smoothly varying distribution without sharp peaks, for example, or there may be constraints that the set of observations must satisfy. In image reconstruction, the many images that are consistent with the data may differ only in unresolved fine structure, so we may wish to incorporate a preference for a smooth image into the term  $P(y|I)$ . Bayes' theorem describes how to incorporate these expectations into the result, and the maximum-entropy method assigns a particular choice (described below) to the term representing prior information. MEM is particularly valuable in cases where the dimensionality of the result is large. Indeed, the method becomes very powerful when the number of values needed to characterize the result exceeds the number of observations, because there is extra information from  $P(y|I)$  that can be used to determine values for the extra parameters.

If we have no prior knowledge of a process, the reasonable approach would be to incorporate this lack of knowledge into Bayes' theorem. However, this is not equivalent

---

<sup>5</sup> This is the reason that much of this section is written in more personalized form, using "we" and "our", than other sections of this book.

to assuming that all possible sets of observations  $y$  are equally probable. For example, a set of observations of coin tosses in which all are heads is very improbable compared to results in which there are about equal numbers of heads and tails, not because the probability of any specified sequence differs, but rather because there are many alternate ways to obtain almost-equal numbers of events and only one way to obtain all one result. Therefore a weighting according to probability would assign the result where the number of events in the two classes is nearly equal a higher a priori probability than the result where all events are in the same class. In statistical mechanics, the entropy  $S$  of a particular macroscopic state (characterized by macroscopic variables) is related to the number of different microscopic states  $W$  that are consistent with that macroscopic state, via the relationship

$$S = k \ln W \quad \text{or} \quad W = e^{S/k} \quad (11.41)$$

where  $k$  is Boltzmann's constant. If the state is characterized by  $N$  results, each occurring for a fraction  $p_n$  of all events, then the entropy becomes

$$S = -k \sum_{n=1}^N p_n \ln p_n . \quad (11.42)$$

In an analysis that forms a cornerstone of modern information theory, Shannon (1948) showed that the same formula (apart from the constant  $k$ ) provides a measure of the uncertainty in a priori knowledge of the correct answer in the case where there are  $N$  mutually exclusive possibilities known to have probabilities  $p_n$ . The natural tendency of thermodynamic systems to increase in entropy is thus associated with a tendency to move toward disordered states having high multiplicity or more microscopic states corresponding to the macroscopic variables. Because information content in a message represents a departure from uniformity or the result of ordering, the same formula (without the factor  $-k$ ) provides a measure of the information content in a message, according to Shannon's analysis. Indeed, then the negative of the right side of this equation, expressed with logarithms having base 2, gives the number of binary operations providing the same amount of information, in the sense of the number of such results needed to specify the result with complete certainty.

A lesson from this is that, if we want to make as few assumptions as possible when assigning a priori probabilities for events, we should assign a probability having an exponential form where the exponent is proportional to the entropy. In the case where the result is the number of events falling into a discrete set of mutually exclusive possibilities, this leads to

$$P(H|I) = e^{-\sum_n p_n \ln p_n} . \quad (11.43)$$

Recall, from section 3a, that the maximum-likelihood solution was obtained by maximizing the probability of obtaining a set of observations  $\{x\}$  from a population with probability distribution function  $\phi(x, \{a\})$  via (3.2):

$$\mathcal{L}(a) = \prod_i \phi(x_i; \{a\}) . \quad (11.44)$$

Also, in the case of Gaussian behavior, the likelihood function was equivalent to

$$\mathcal{L}(\{a\}) = e^{-\chi^2/2} . \quad (11.45)$$

A maximum-entropy solution then can be obtained by maximizing the product of the likelihood function with  $P(H|I)$ , or in the case of Gaussian errors by maximizing the exponential term

$$E = - \sum (p_n \ln p_n) - \chi^2/2 . \quad (11.46)$$

Maximizing  $E$  amounts to selecting the most probable distribution consistent with the constraints of the observations. This is not in general the maximum likelihood solution, but a departure from that solution such that the gain in probability attributed to prior knowledge exceeds the decrease in probability as the solution is perturbed from its conventional maximum-likelihood value. In some cases when the number of parameters exceeds the number of observations only solutions matching the observations exactly (and hence giving  $\chi^2 = 0$  are considered, and then the solution corresponds to the state having the maximum entropy while satisfying the constraints of the observations.

Jaynes (1985) gives a brief summary of the history and evolution of this method and its ties to statistical mechanics. That article also should be read for its exposition of the underlying approach and the philosophical foundations of the method.

## 2). Application to spectral analysis

The application of these ideas to spectral analysis was developed by Burg (1967). Part of the problem he addressed was the effect on the spectrum of contributions with wavelength or period beyond the size of the record. In the method outlined in Chapter 8, such contributions are not included when the variance spectrum is calculated. Because of the connection (8.26) between the variance spectrum and the autocovariance function, the estimated spectral density at any frequency is in truth affected by values of the autocovariance function at all lags. The equivalence between the FFT-based spectral method of Chapter 8 and the spectral approach via the autocovariance function shows that neglecting the long-wavelength or long-period components is equivalent to assuming that the autocovariance function is zero for all lags longer than the time series. Unless it is known that this is true, this is an assumption that can introduce error into the result. In fact, if the experimental values for the observed portion of the autocovariance function do not taper smoothly to zero at the maximum lag, the jump in value assumed at that point introduces unwanted side lobes into the result. Furthermore, in an effort to treat these problems, window functions are introduced that have the effect of tapering the autocovariance spectrum to zero at the longest lag. This distorts the data still further from their actual values, and reduces the resolution possible (by in effect smoothing together nearby estimates of the spectral density), to gain the benefits of smoothing the resulting spectrum and reducing sidelobe contributions. Burg pointed out that this approach then proceeds on the basis of a spectrum that could not have produced the observations, when a better approach is to select a “best” spectrum from among the many spectra that could have produced the observed spectrum.

The ambiguity among these possible spectra arises because they have nonzero contributions from wavelengths or periods outside the range of observations. The Burg approach is to select from among these the result that conforms to the maximum-entropy solution; in effect, this selects a solution that matches the data exactly but extrapolates the

autocovariance function beyond the range of observations in a way that, in the maximum-entropy sense, makes the least restrictive assumption about the form of that extrapolation.<sup>6</sup>

Press et al. (1992) give the details of the maximum-entropy solution and provide a computer routine that can be used to find that solution. They also provide useful guidance for an analyst wanting to use that routine.

### 3). Image reconstruction

Another important application of the maximum-entropy method has been in image reconstruction. Here the quantity

$$s = - \sum p_n \ln p_n . \quad (11.47)$$

is maximized, subject to the constraints imposed by the data. The result is to impose a strong preference for uniform images, so that any features in the reconstructed image are required by the data and do not represent noise. A description of the method can be found in Skilling (1989), and for this case also Press et al. (1992) provide a useful computer routine to implement image reconstruction.

## SOURCES AND FURTHER READING

- Burg, J. P., 1975: Maximum Entropy Spectral Analysis. Ph.D. thesis, Stanford University.
- Bury, K.V., 1975: *Statistical Models in Applied Science*. Wiley and Sons, New York, 625 pp.
- Jaynes, E. T., 1985: Where do we go from here? In *Maximum-Entropy and Bayesian Methods in Inverse Problems*, C. R. Smith and W. T. Grandy, Jr., eds., D. Reidel Publishing Co., Dordrecht, Holland, pp. 21–58.
- Levine, R. D., and M. Tribus, 1981: *The Maximum Entropy Formalism*. The MIT Press, Cambridge, Massachusetts, 498 pp.
- Menzel, D. H., 1960: *Fundamental Formulas of Physics, Volume One*. Dover Publications, New York, 364 pp.
- Shannon, 1948:
- Skilling, J., 1991: Fundamentals of MaxEnt in data analysis. In *Maximum Entropy in Action*, B. Buck and V. A. Macaulay, eds., Oxford University Press, Oxford, pp. 19–40.
- Press, W. H., Brian P. Flannery, S. A. Teukolsky, and W. T. Vetterling, 1992: *Numerical Recipes in C*. Second Edition, Cambridge University Press, Cambridge, 735 pp. Cf. pp. XXX-XXX.
- Ulrych, T. J., 1985: Spectral analysis and time series models. In *Maximum-Entropy and Bayesian Methods in Inverse Problems*, C. R. Smith and W. T. Grandy, Jr., editors, D. Reidel Publishing Company, Dordrecht, pp. 243-272.

---

<sup>6</sup> Note that this solution does not permit the best estimates of the autocovariance to differ from the observations, as would be obtained by maximizing (11.46). Instead, the observations are treated as constraints and the maximum-entropy solution found subject to those constraints. This solution thus is appropriate in the case of data without noise.