

## 3. Probability Distribution Functions

- *Gaussian distribution*
- *Binomial distribution*
- *Poisson distribution*
- *Student's  $t$  distribution*
- *Confidence intervals*

### 3. Probability Distribution Functions

#### *a. Introductory comment*

The following sections discuss some probability distribution functions that apply to common measurement situations. The concept of a probability distribution function was discussed briefly in Chapter 2, as a normalized function whose integral gives the probability corresponding to the space of integration. Although the material of this section is readily available elsewhere, it is included here for review and reference. The probability function also determines the expectation value of a function:

$$\langle f(x) \rangle = \int f(x)\phi(x)dx. \quad (3.1)$$

The following are particularly important forms of the probability distribution function.

*b. Gaussian or normal distribution*

This distribution occurs frequently and has great generality. For large numbers of events, it is the limiting form for many other distribution functions, and by virtue of the central limit theorem it is the appropriate form for the sum of many variables even if those variables individually follow other distributions. It is

$$\Phi_G(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}. \quad (3.2)$$

The Gaussian distribution provides a realistic approximation to the distribution of deviations in many experimental situations, especially for the “central” portion of the deviations. The distribution function is plotted in Fig. 3.1. The width of the distribution is characterized by the standard deviation  $\sigma$ , or sometimes by the full-width-at-half-maximum,  $\Gamma = 2.354\sigma$ . See Fig. 3.2 for examples with various widths.

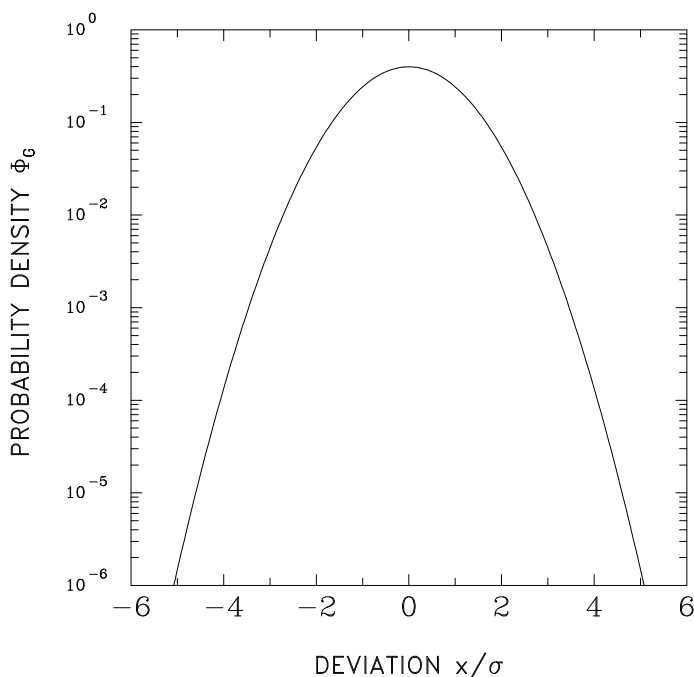


Fig. 3.1: Frequency distribution for the Gaussian distribution function  $\Phi_G$  as a function of the normalized deviation  $x/\sigma$ , for the case with zero mean value.

*c. The binomial distribution*

Suppose that there is a probability  $p$  that a particular event will occur, and therefore a probability  $(1 - p)$  that the event will not occur, in a given trial. In a set of  $N$  trials, what is the probability that there will be  $n$  events? For example, the events might be coin tosses where the event in question is “heads” with probability  $p=1/2$ . A first guess might be  $\phi(n) = p^n(1 - p)^{N-n}$ , and this would be correct if the order of events were specified. However, if the order is not specified, there may be many different sequences that lead

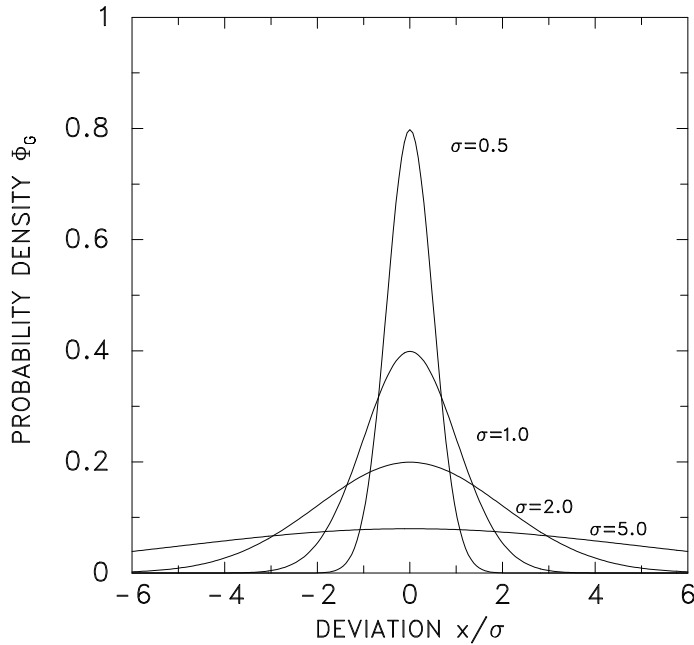


Fig. 3.2: Gaussian probability distribution, as a function of the unnormalized deviation, for cases where  $\sigma$  assumes the values 0.5, 1.0, 2.0, and 5.0.

to the same final number of events, and the probability must correct for the multiple ways that the final number can be reached. The correction factor is called the binomial coefficient, and the resulting probability distribution is the binomial distribution function:

$$\Phi_B(n; p, N) = \binom{N}{n} p^n (1-p)^{N-n} \quad (3.3)$$

where

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} \quad (3.4)$$

Figure 3.3 shows an example for 30 events and a probability  $p = 0.4$ .

The mean of the distribution is given by  $\mu = pN$ , as can be demonstrated by integration of the probability distribution function. The variance is given by

$$\sigma^2 = Np(1-p). \quad (3.5)$$

Figure 3.3 also shows a comparison between Gaussian and binomial distribution functions having the same mean and standard deviation. For these conditions, the distribution functions are almost indistinguishable.

The binomial distribution characterizes the probability of discrete events, while the Gaussian distribution describes the probability of a continuously varying result. Both distributions describe events that are independent. Violation of this assumption is a common source of error. For example, if in 100 days rain is observed on 40 days, one might erroneously estimate that the standard deviation in the number of rain events is  $\sqrt{(100)(0.4)(0.6)} \approx 5$ . Because rain events in most locations are highly correlated from day to day, rain events cannot be treated as independent, and this use of the binomial distribution would usually underestimate the true variability.

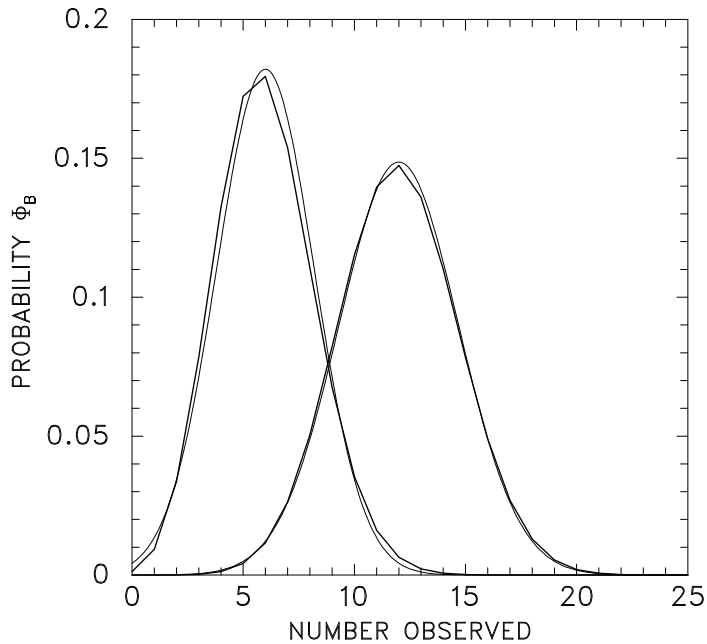


Fig. 3.3: The binomial distribution functions (3.3) [heavier line] for  $p = 0.4$  and  $p = 0.2$ , both with  $N=30$ . For comparison, the Gaussian distribution functions having the same mean and standard deviation are also plotted as the thinner smooth lines. Values for the binomial distribution function are shown only for integer numbers of events, with adjacent values connected by straight lines.

#### d. Poisson distribution

The Poisson distribution applies to the number of randomly occurring and countable events that occur in an interval. For example, the expected average rate  $R$  at which cloud droplets are detected by an airborne counter is given by

$$R = AVc \quad (3.6)$$

where  $A$  is the sample area within which passing droplets are counted,  $V$  the airspeed, and  $c$  the droplet concentration. However, the droplets counted during any particular interval will differ from  $R$  because only discrete and not fractional events can be counted, and because statistical fluctuations will cause the number counted to vary from the true population mean. The binomial distribution appears to be applicable, because it gives the probability of seeing a given number if the probability  $p$  is known. However, in this case, the number of locations at which a droplet could be observed is infinitely large, and the probability of an observation at each location infinitesimally small. The appropriate distribution is therefore the limit of the binomial distribution as the number of possible events approaches infinity while the probability of any specific event approaches zero, maintaining the correct average number of events in a specific interval. This limit is the Poisson distribution function.

The Poisson distribution function thus gives the probability of observing  $n$  discrete events if the true mean is  $\mu$ . It has the form

$$\Phi_P(n; \mu) = \frac{\mu^n}{n!} e^{-\mu}. \quad (3.7)$$

The mean of this distribution is  $\mu$ , and the variance is also  $\mu$ . This is the source of the common estimate that the standard deviation in a counted number of events is the square root of the number counted. Figure 3.4 shows that the distribution function, even for small numbers of events, is similar to a Gaussian distribution.

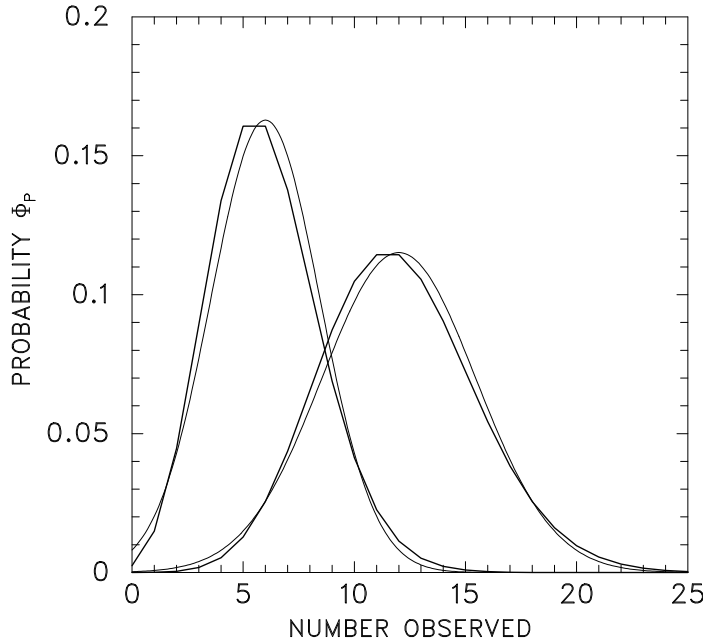


Fig. 3.4: Poisson distribution functions (3.7), shown as heavier lines, for cases having means 6 and 12. The Gaussian distribution functions having the same means and standard deviations are also shown as thinner smooth lines. Values for the Poisson distribution function are plotted only at integer values, and adjacent values are connected by straight lines.

*e. Student's  $t$  distribution*

Suppose that a set of observations has mean  $\bar{x}$ . To test the hypothesis that this sample came from a population with mean  $\mu$  obeying a Gaussian distribution, we might try the test statistic

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - \mu}{\sigma_\mu} \quad (3.8)$$

where  $\sigma$  is the true standard deviation in  $x$  and  $\sigma_\mu = \sigma/\sqrt{n}$  is the standard deviation in the mean. However, usually the true standard deviation  $\sigma$  for the population from which the sample was collected is unknown. An estimator for  $\sigma$ , calculated from the observations, is

$$s = \left[ \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 \right]^{1/2}. \quad (3.9)$$

A candidate test statistic is thus

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\bar{x} - \mu}{s_\mu} \quad (3.10)$$

where  $s_\mu = s/\sqrt{n}$ .

Although  $z$  would obey a Gaussian distribution if the individual measurements entering  $\bar{x}$  do,  $t$  will not be Gaussian distributed. Instead, the distribution in  $t$  is determined by the ratio of the Gaussian distribution to the square root of the chisquare distribution which characterizes deviations of the sample estimate of the standard deviation from the true standard deviation:<sup>1</sup>

$$t = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \left[ \frac{\sigma^2}{s^2} \right]^{1/2} \quad (3.11)$$

where  $s^2/\sigma^2$  is distributed as  $\chi^2(n-1)/(n-1)$ .

The distribution in  $t$  can be derived from this ratio (as shown, e.g., in Brownlee 1965):

$$\Phi_t(t, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad (3.12)$$

where  $\nu$  is the number of degrees of freedom, which will be  $(n-1)$  in the case where  $\bar{x}$  is the average of  $n$  independent measurements. The variance of the  $t$  distribution is given by

$$V_{tt} = \frac{\nu}{\nu - 2}, \quad (3.13)$$

and for large numbers of degrees of freedom the  $t$  distribution approaches a Gaussian distribution with this variance.

The  $t$  statistic supports a test of the hypothesis that the mean is  $\mu$  without requiring that the true standard deviation  $\sigma$  be known. Only the sample standard deviation is needed. The change from a Gaussian distribution function is most significant for small numbers of degrees of freedom. Figure 2.1 showed the cumulative form of this distribution function for various degrees of freedom and compared its shape to that of the Gaussian distribution.

**EXAMPLE 3.1:** *Two instruments measuring the concentration of droplet concentration in clouds collect measurements that, by linear regression, are related by a slope parameter  $b$  and a standard deviation  $s_b$ . We want to test if these results are consistent with a slope of 1, as would be expected if the two instruments were calibrated and operating consistently. The regression fit has  $N - 2$  degrees of freedom if  $N$  measurements are used in the comparison, so the appropriate test statistic is*

$$t = \frac{\bar{b} - 1}{\sigma_b} \quad (3.14)$$

*with  $N - 2$  degrees of freedom. This can be used to test the probability that a value as large as  $t$  will be obtained if the true mean is 1.*

<sup>1</sup>The chisquare distribution is discussed in Chapter 7.

**EXAMPLE 3.2:** Test if two experimental results are in conflict or are consistent with expected deviations in repeated experiments. Suppose that two different experimenters obtain, respectively,  $R_1 \pm s_1$  and  $R_2 \pm s_2$ . To test for consistency of these results, use the difference  $(R_2 - R_1)$ . The standard deviation in this difference can be estimated from

$$s = \sqrt{s_1^2 + s_2^2}. \quad (3.15)$$

Then

$$t = \frac{R_2 - R_1}{s}$$

is distributed according to the  $t$  distribution with  $(n_1 + n_2 - 1)$  degrees of freedom, where  $n_1$  and  $n_2$  are the respective degrees of freedom in the two experiments. A common mistake is to assume that the degrees of freedom should be 1 for this case because there are two measurements and one difference. Recall that the reason for introducing the  $t$  distribution was to account for lack of knowledge of the true standard deviation in the phenomenon being observed. If there is a large set of measurements, this standard deviation is known much better than if there are only a few, and this improved knowledge must be reflected in the number of degrees of freedom in the comparison.

### f. Confidence intervals

The distributions of the preceding sections are often used for inference. If the statistical character of the observations requires a particular distribution function, that function can be used to estimate the probability that the true mean is greater than some limit  $\mu_2$  if the observed mean is  $\bar{x}$ .

First, consider the *direct* probability question: If the true mean is  $\mu$  and the true standard deviation is  $\sigma$ , what is the probability that an observation will lie between  $x_1$  and  $x_2$ ? The answer is

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} \phi(x) dx. \quad (3.16)$$

Often, the integrals of probability functions do not have analytical forms. Tables of cumulative values for the Gaussian distribution function and other standard forms are available in statistical handbooks and other reference books (e.g., Abramowitz and Stegun 1972), and values are also available on many computer systems through mathematical libraries. For example, Figure 2.1 showed the probability of exceeding various deviations for the Gaussian and Student- $t$  distribution functions.

The *inverse* problem follows a similar procedure. If the experimentally determined estimate of the mean is  $\bar{x}$  and the estimate of the standard deviation is  $s$ , the procedure of the preceding paragraphs can be used to calculate the probability that a set of observations will give a mean  $\bar{x}$  when the true mean is  $\mu$ . If the value of  $\mu$  is determined for which the observed mean  $\bar{x}$  would be a deviation exceeded only with probability  $f$ , it is sometimes said that there is only a probability  $f$  that the true mean exceeds the limit  $\mu$ . The weakness in this argument is that the true standard deviation  $\sigma$  is also unknown, and the estimate of probability depends on knowledge of this true standard deviation. If the

experimental estimate  $s$  is used in place of  $\sigma$ , this is not a true inverse procedure and can lead to erroneous indications of confidence limits.

#### *SOURCES AND FURTHER READING*

- Abramowitz, M. and I. A. Stegun, 1972: Handbook of Mathematical Functions. Dover Publications, New York, 1046 pp.
- Bevington, P. R., 1969: Data Reduction and Error Analysis for the Physical Sciences. McGraw-Hill, New York, 336 pp.
- Box, G. E. P., W. G. Hunter, and J. S. Hunter, 1978: Statistics for Experimenters. John Wiley and Sons, New York, 653 pp.
- Brownlee, K. A., 1965: Statistical Theory and Methodology in Science and Engineering. John Wiley and Sons, New York, 590 pp.
- Feller, William, 1950: *An Introduction to Probability Theory and its Applications*. John Wiley and Sons, New York, 461 pp. *Statistics to Meteorology*. Pennsylvania State University, 224 pp.
- 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.