Chapter 2
Random Variables and Their Distributions

Abstract The purpose of performing experiments and collecting data is to gain information on certain quantities of interest called random variables. The exact value of these quantities cannot be known with absolute precision, but rather we can constrain the variable to a given range of values, narrower or wider according to the nature of the variable itself and the type of experiment performed. Random variables are described by a distribution function, which is the theoretical expectation for the outcome of experiments aimed to measure it. Other measures of the random variable are the mean, variance, and higher-order moments.

2.1 Random Variables

A random variable is a quantity of interest whose true value is unknown. To gain information on a random variable we design and conduct experiments. It is inherent to any experiment that the random variable of interest will never be known exactly. Instead, the variable will be characterized by a *probability distribution function*, which determines what is the probability that a given value of the random variable occurs. Repeating the measurement typically increases the knowledge we gain of the distribution of the variable. This is the reason for wanting to measure the quantity as many times as possible.

As an example of random variable, consider the gravitational constant G. Despite the label of "constant", we only know it to have a range of possible values in the approximate interval $G = 6.67428 \pm 0.00067$ (in the standard S.I. units). This means that we don't know the true value of G, but we estimate the range of possible values by means of experiments. The random nature of virtually all quantities lies primarily in the fact that no quantity is known exactly to us without performing an experiment and that any experiment is never perfect because of practical or even theoretical limitations. Among the practical reasons are, for example, limitations in the precision of the measuring apparatus. Theoretical reasons depend on the nature of the variable. For example, the measurement of the position and velocity of a subatomic particle is limited by the Heisenberg uncertainty principle, which forbids an exact knowledge even in the presence of a perfect measuring apparatus.

The general method for gaining information on a random variable X starts with set of measurements x_i, ensuring that measurements are performed under the same

M. Bonamente, *Statistics and Analysis of Scientific Data*, Graduate Texts
in Physics, DOI 10.1007/978-1-4939-6572-4_2

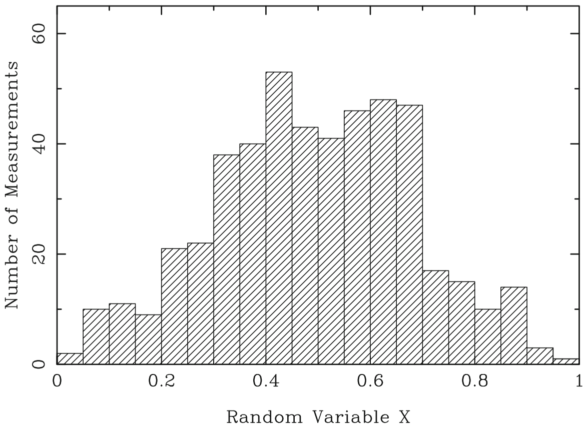

Fig. 2.1 Example of data collected to measure a random variable X. The 500 measurements were binned according to their value to construct the sample distribution. The shape of the distribution depends on the nature of the experiment and of the number of measurements

experimental conditions. Throughout the book we will reserve uppercase letters for the name of the variable itself and lowercase letters for the actual measurements. From these measurements, one obtains a histogram corresponding to the frequency of occurrence of all values of X (Fig. 2.1). The measurements x_i form the *sample distribution* of the quantity, which describes the empirical distribution of values collected in the experiment. On the other hand, random variables are typically expected to have a theoretical distribution, e.g., Gaussian, Poisson, etc., known as the *parent distribution*. The parent distribution represents the belief that there is an ideal description of a random variable and its form depends on the nature of the variable itself and the method of measurement. The sample distribution is expected to become the parent distribution if an infinite number of measurements are performed, in such a way that the randomness associated with a small number of measurements is eliminated.

Example 2.1 In Sect. 3.3 we will show that a discrete variable (e.g., one that can only take integer values) that describes a counting experiment follows a Poisson function,

$$P(n) = \frac{\mu^n}{n!} e^{-\mu}$$

in which μ is the mean value of the random variable (for short, its true-yet-unknown value) and n is the actual value measured for the variable. $P(n)$ indicates the probability of measuring the value n, given that the true value is μ. Consider the experiment of counting the number of photons reaching Earth from a given star; due to a number of factors, the count may not always be the same every time the experiment is performed, and if only one experiment is performed, one would obtain a sample distribution that has a single "bar" at the location of the measured value and this sample distribution would not match well a Poisson function. After a small number of measurements, the distribution may appear similar to that in Fig. 2.1

and the distribution will then become smoother and closer to the parent distribution as the number of measurements increases. Repeating the experiment therefore will help in the effort to estimate as precisely as possible the parameter μ that determines the Poisson distribution. ◇

2.2 Probability Distribution Functions

It is convenient to describe random variables with an analytic function that determines the probability of the random variable to have a given value. Discrete random variables are described by a *probability mass function* $f(x_i)$, where $f(x_i)$ represents the probability of the variable to have an exact value of x_i. Continuous variables are described by a *probability distribution function* $f(x)$, such that $f(x)dx$ is the probability of the variable to have values in the interval $[x, x+dx]$. For simplicity we will refer to both types of distributions as probability distribution functions throughout the book.

Probability distribution functions have the following properties:

1. They are normalized to 1. For continuous variables this means

$$\int_{-\infty}^{+\infty} f(x)dx = 1. \tag{2.1}$$

   For variables that are defined in a subset of the real numbers, e.g., only values $x \geq 0$ or in a finite interval, $f(x)$ is set to zero outside the domain of definition of the function. For discrete variables, hereafter the integrals are replaced by a sum over all values that the function of integration can have.
2. The probability distribution can never be negative, $f(x) \geq 0$. This is a consequence of the Kolmogorov axiom that requires a probability to be non-negative.
3. The function $F(x)$, called the *(cumulative) distribution function*,

$$F(x) = \int_{-\infty}^{x} f(\tau)d\tau, \tag{2.2}$$

   represents the probability that the variable has any value less or equal than x. $F(x)$ is a non-decreasing function of x that starts at zero and has its highest value of one.

Example 2.2 The *exponential random variable* follows the probability distribution function defined by

$$f(x) = \lambda e^{-\lambda x},\ x \geq 0, \tag{2.3}$$

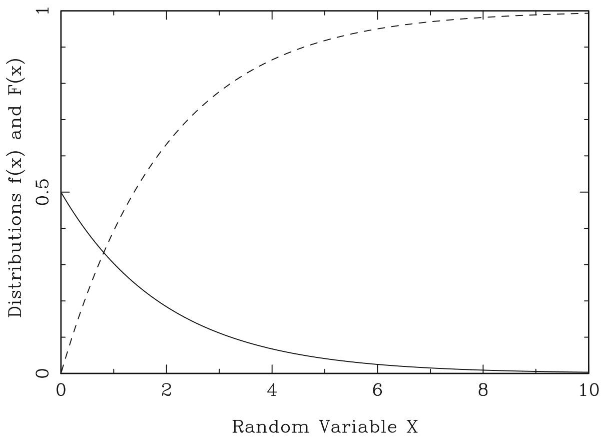

Fig. 2.2 The distribution function $f(x)$ (*solid line*) and the cumulative distribution function $F(x)$ (*dashed line*) for an exponential variable with $\lambda = 0.5$

where λ is an adjustable parameter that must be positive. The probability distribution function is therefore $f(x) = 0$ for negative values of the variable. The cumulative distribution function is given by

$$F(x) = 1 - e^{-\lambda x}. \tag{2.4}$$

In Fig. 2.2 are drawn the probability distribution function $f(x)$ and the cumulative distribution function $F(x)$ for an exponential variable with $\lambda = 0.5$. ◇

2.3 Moments of a Distribution Function

The probability distribution function $f(x)$ provides a complete description of the random variable. It is convenient to find a few quantities that describe the salient features of the distribution. The *moment* of order n, μ_n, is defined as

$$\mu_n = E[X^n] \equiv \int f(x)x^n dx. \tag{2.5}$$

The moment μ_n is also represented as $E[X^n]$, the *expectation* of the function X^n. It is possible to demonstrate, although mathematically beyond the scope of this book, that the knowledge of moments of all orders is sufficient to determine uniquely the distribution function [42]. This is an important fact, since it shifts the problem of determining the distribution function to that of determining at least some of its moments. Moreover, a number of distribution functions only have a few non-zero moments, and this renders the task even more manageable.

The moments or expectations of a distribution are theoretical quantities that can be calculated from the probability distribution $f(x)$. They are *parent* quantities that we wish to estimate via measurements. In the following we describe the two main expectations, the mean and the variance, and the *sample* quantities that approximate them, the sample mean and the sample variance. Chapter 5 describes a method to justify the estimates of parent quantities via sample quantities.

2.3.1 The Mean and the Sample Mean

The moment of the first order is also known as the *mean* or *expectation* of the random variable,

$$\mu = E[X] = \int_{-\infty}^{+\infty} xf(x)dx. \tag{2.6}$$

The expectation is a linear operation and therefore satisfies the property that, e.g.,

$$E[aX + bY] = aE[X] + bE[Y], \tag{2.7}$$

where a and b are constants. This is a convenient property to keep in mind when evaluating expectations of complex functions of a random variable X.

To estimate the mean of a random variable, consider N measurements x_i and define the *sample mean* as

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{2.8}$$

To illustrate that the sample mean $\overline{x}$ defined by (2.8) is equivalent to the mean μ, consider a discrete variable, for which

$$E[X] = \sum_{j=1}^{M} f(x_j)x_j, \tag{2.9}$$

where $f(x_j)$ is the probability distribution function and we have assumed that the variable can only have M possible values. According to the classical interpretation of the probability, the distribution function is given by

$$f(x_j) = \lim_{N\to\infty} \frac{N(x_j)}{N},$$

in which $N(x_j)$ is the number of occurrence of the value x_j. Since $\Sigma N(x_j)x_j$ is the value obtained in N measurements, it is equivalent to Σx_i. Therefore the sample mean will be identical to the parent mean in the limit of an infinite number of measurements,

$$\lim_{N\to\infty} \overline{x} = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} x_i = \lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{M} N(x_j)x_j = \sum_{j=1}^{M} f(x_j)x_j = E[X].$$

A proof that the sample mean provides an unbiased estimate of the mean will be given in Chap. 5 for Gaussian and Poisson variables.

The sample mean is therefore a representative value of the random variable that estimates the parent mean using a finite number of measurements. Other measures of a random variable include the *mode*, defined as the value of maximum probability, and the *median*, defined as the value that separates the lower 50 % and the upper 50 % of the distribution function. For distributions that are symmetric with respect to the peak value, as is the case for the Gaussian distribution defined below in Sect. 3.2, the peak value coincides with the mean, median, and mode. A more detailed analysis of the various measures of the "average" value of a variable is described in Chap. 6.

2.3.2 The Variance and the Sample Variance

The *variance* is the expectation of the square of the deviation of X from its mean:

$$Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx = \sigma^2. \tag{2.10}$$

The square root of the variance is referred to as the *standard deviation* or *standard error* σ and it is a common measure of the average difference of a given measurement x_i from the mean of the random variable. Notice that from the point of view of physical dimensions of the moments defined by (2.5), moments of the n-th order have the dimensions of the random variable to the n-th power. For example, if X is measured in meters, the variance is measured in meters square (m^2), thus the need to use the square root of the variance as a measure of the standard deviation of the variable from its mean.

The main reason for defining the average difference of a measurement from its mean in terms of a moment of the second order is that the expectation of the *deviation* $X-\mu$ is always zero, as can be immediately seen using the linearity property of the expectation. The deviation of a random variable is therefore not of common use in statistics, since its expectation is null.

The *sample variance* is defined as

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{2.11}$$

and a proof that this quantity is an unbiased estimate of the parent variance will be provided in Chap. 5. The presence of a factor of $N - 1$, and not just N, in the denominator of the sample variance, is caused by the fact that the sample variance requires also an estimate of the sample mean, since the exact value of the parent mean is unknown. This result will be explained further in Sect. 5.1.2.

Using the linear property of the expectation, it is straightforward to show that the following property applies:

$$Var(X) = E[X^2] - \mu^2. \tag{2.12}$$

This relationship is very convenient to calculate the variance from the moments of the first and second order. The deviation and the variance are moments calculated with respect to the mean, also referred to as *central moments*.

Another useful property of the variance, which follows from the fact that the variance is a moment of the second order, is

$$Var(aX) = a^2 Var(X) \tag{2.13}$$

where a is a constant.

2.4 A Classic Experiment: J.J. Thomson's Discovery of the Electron

A set of experiments by J.J. Thomson in the late nineteenth century were aimed at the measurement of the ratio between the mass and charge of a new lightweight particle, which was later named *electron*. The experiment was truly groundbreaking not just for the method used, but also because it revolutionized our understanding of physics and natural sciences by proving that the new particle was considerably lighter than the previously known charge carrier, the proton.

The experiment described in this book was reported by Thomson in [39]. It consists of measuring the deflection of negatively charged cathode rays by a magnetic field H in a tube. Thomson wanted to measure the mass m of the charged particles that constituted these cathode rays. The experiment is based on the measurement of the following quantities: W is the kinetic energy of the

(continued)

particles, $Q = Ne$ is the amount of electricity carried by the particles (N is the number of particles and e the charge of each particle) and $I = HR$, where R is the radius of curvature of the path of these rays in a magnetic field H. The measurements performed by Thomson were used to infer the ratio m/e and the speed v of the new lightweight particle according to

$$\begin{aligned} v &= \frac{2W}{QI}; \\ \frac{m}{e} &= \frac{I^2 Q}{2W}. \end{aligned} \tag{2.14}$$

For the purpose of the data analysis of this experiment, it is only necessary to know that W/Q and I are the primary quantities being measured, and inferences on the secondary quantities of interest are based on (2.14). For the proton, the mass-to-charge ratio was known to be approximately 1×10^{-4} g per electromagnetic (EMU) charge unit, where the EMU charge unit is equivalent to 10^{-10} electrostatic charge units, or ESU (a more common unit of measure for charge). In Thomson's units, the accepted value of the mass to charge ratio of the electron is now 5.7×10^{-8}. Some of the experimental data collected by Thomson are reported in Tables 2.1 and 2.2, in which "gas" refers to the gas used in the tubes he used for the experiment.

Some of Thomson's conclusions are reported here:

(a) *"It will be seen from these tables that the value of m/e is independent of the nature of the gas"*;
(b) *"the values of m/e were, however, the same in the two tubes."*;
(c) *"for the first tube, the mean for air is* 0.40×10^{-7}*, for hydrogen* 0.42×10^{-7} *and for carbonic acid* 0.4×10^{-7}*"*;
(d) *"for the second tube, the mean for air is* 0.52×10^{-7}*, for hydrogen* 0.50×10^{-7} *and for carbonic acid* 0.54×10^{-7}*"*.

Using the equations for sample mean and variance explained in Sect. 2.3, we are already in a position to measure the sample means and variances in air as $\overline{m/e}_1 = 0.42$ and $s_1^2 = 0.005$ for Tube 1, $\overline{x}_2 = 0.52$ and $s_2^2 = 0.003$ for Tube 2. These statistics can be reported as a measurement of 0.42 ± 0.07 for Tube 1 and 0.52 ± 0.06 for Tube 2. To make more quantitative statements on the statistical agreement between the two measurements, we need to know what is the probability distribution function of the sample mean. The test to determine whether the two measurements are consistent with each other will be explained in Sect. 7.5. For now, we simply point out that the fact that the range of the two measurements overlap, is an indication of the statistical agreement of the two measurements.

(continued)

Note: The three measurements marked with a star appear to have value of v or m/e that are inconsistent with the formulas to calculate them from W/Q and I. They may be typographical errors in the original publication. The first appears to be a typo in W/Q (6×10^{12} should be 6×10^{11}), the corrected value is assumed throughout this book. The second has an inconsistent value for v (should be 6.5×10^9, not 7.5×10^9), the third has inconsistent values for both v and m/e, but no correction was applied in these cases to the data in the tables.

Table 2.1 Data from Thomson's measurements of Tube 1

Gas	W/Q	I	m/e	v
Tube 1				
Air	4.6×10^{11}	230	0.57×10^{-7}	4×10^9
Air	1.8×10^{12}	350	0.34×10^{-7}	1×10^{10}
Air	6.1×10^{11}	230	0.43×10^{-7}	5.4×10^9
Air	2.5×10^{12}	400	0.32×10^{-7}	1.2×10^{10}
Air	5.5×10^{11}	230	0.48×10^{-7}	4.8×10^9
Air	1×10^{12}	285	0.4×10^{-7}	7×10^9
Air	1×10^{12}	285	0.4×10^{-7}	7×10^9
Hydrogen* .	6×10^{12}	205	0.35×10^{-7}	6×10^9
Hydrogen . .	2.1×10^{12}	460	0.5×10^{-7}	9.2×10^9
Carbonic acid*	8.4×10^{11}	260	0.4×10^{-7}	7.5×10^9
Carbonic acid	1.47×10^{12}	340	0.4×10^{-7}	8.5×10^9
Carbonic acid	3.0×10^{12}	480	0.39×10^{-7}	1.3×10^{10}

See Note for meaning of $\star$

Table 2.2 Data from Thomson's measurements of Tube 2

Gas	W/Q	I	m/e	v
Tube 2				
Air	2.8×10^{11}	175	0.53×10^{-7}	3.3×10^9
Air*	2.8×10^{11}	175	0.47×10^{-7}	4.1×10^9
Air	3.5×10^{11}	181	0.47×10^{-7}	3.8×10^9
Hydrogen .	2.8×10^{11}	175	0.53×10^{-7}	3.3×10^9
Air	2.5×10^{11}	160	0.51×10^{-7}	3.1×10^9
Carbonic acid	2.0×10^{11}	148	0.54×10^{-7}	2.5×10^9
Air	1.8×10^{11}	151	0.63×10^{-7}	2.3×10^9
Hydrogen .	2.8×10^{11}	175	0.53×10^{-7}	3.3×10^9
Hydrogen .	4.4×10^{11}	201	0.46×10^{-7}	4.4×10^9
Air	2.5×10^{11}	176	0.61×10^{-7}	2.8×10^9
Air	4.2×10^{11}	200	0.48×10^{-7}	4.1×10^9

See Note for meaning of $\star$

2.5 Covariance and Correlation Between Random Variables

It is common to measure more than one random variable in a given experiment. The variables are often related to one another and it is therefore necessary to define a measure of how one variable affects the measurement of the others. Consider the case in which we wish to measure both the length of one side of a square and the area; it is clear that the two quantities are related in a way that the change of one quantity affects the other in the same manner, i.e., a positive change of the length of the side results in a positive change of the area. In this case, the length and the area will be said to have a positive correlation. In this section we introduce the mathematical definition of the degree of correlation between variables.

2.5.1 Joint Distribution and Moments of Two Random Variables

When two (or more) variables are measured at the same time via a given experiment, we are interested in knowing what is the probability of a given pair of measurements for the two variables. This information is provided by the *joint probability distribution function*, indicated as $h(x,y)$, with the meaning that $h(x,y)dxdy$ is the probability that the two variables X and Y are in a two-dimensional interval of size $dxdy$ around the value (x,y). This two-dimensional function can be determined experimentally via its sample distribution, in the same way as one-dimensional distributions.

It is usually convenient to describe one variable at a time, even if the experiment features more than just one variable. In this case, the expectation of each variable (for example, X) is defined as

$$E[X] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xh(x,y)dxdy = \mu_x \tag{2.15}$$

and the variance is similarly defined as

$$E[(X-\mu_x)^2] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x-\mu_x)^2 h(x,y)dxdy = \sigma_x^2. \tag{2.16}$$

These equations recognize the fact that the other variable, in this case Y, is indeed part of the experiment, but is considered *uninteresting* for the calculation at hand. Therefore the uninteresting variable is integrated over, weighted by its probability distribution function.

The *covariance* of two random variables is defined as

$$Cov(X,Y) \equiv E[(X-\mu_x)(Y-\mu_y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_x)(y-\mu_y)h(x,y)dxdy = \sigma_{xy}^2. \tag{2.17}$$

The covariance is the expectation of the product of the deviations of the two variables. Unlike the deviation of a single variable, whose expectation is always zero, this quantity will be positive if, on average, a positive deviation of X is accompanied by a positive deviation of Y, or if two negative deviations are likely to occur simultaneously, so that the integrand is a positive quantity. If, on the other hand, the two variables tend to have deviations of opposite sign, the covariance will be negative. The covariance, like the mean and variance, is a parent quantity that can be calculated from the theoretical distribution of the random variables.

The *sample covariance* for a collection of N pairs of measurements is calculated as

$$s_{xy}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y}), \tag{2.18}$$

using a similar equation to the sample variance.

The *correlation coefficient* ρ is simply a normalized version of the covariance,

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_x\sigma_y}. \tag{2.19}$$

The correlation coefficient is a number between -1 and $+1$. When the correlation is zero, the two variables are said to be *uncorrelated*. The fact that the correlation coefficient is normalized to within the values ± 1 derives from (2.10) and the properties of the joint distribution function.

The *sample correlation coefficient* is naturally defined as

$$r = \frac{s_{xy}^2}{s_x s_y} \tag{2.20}$$

in which s_x^2 and s_y^2 are the sample variances of the two variables.

The covariance between two random variables is very important in evaluating the variance in the sum (or any other function) of two random variables, as explained in detail in Chap. 4. The following examples illustrate the calculation of the covariance and the sample covariance.

Example 2.3 (Variance of Sum of Variables) Consider the random variables X, Y and the sum $Z = X + Y$: the variance is given by

$$Var(Z) = \int\int (x + y - (\mu_x + \mu_y))^2 h(x,y) dxdy =$$

$$Var(X) + Var(Y) + 2Cov(X,Y)$$

which can also be written in the compact form $\sigma_z^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}^2$. This shows that variances add linearly only if the two random variables are uncorrelated. Failure to check for correlation will result in errors in the calculation of the variance of the sum of two random variables. ◇

Example 2.4 Consider the measurement of the following pairs of variables: (0, 2), (2, 5), (1, 4), (−3, −1). We can calculate the *sample* covariance by means of the following equation:

$$s_{xy}^2 = \frac{1}{3} \sum_{i=1}^{4} (x_i - \overline{x})(y_i - \overline{y}) = \frac{17}{3}$$

where $\overline{x} = 0$ and $\overline{y} = 2.5$. Also, the individual variances are calculated as

$$s_x^2 = \frac{1}{3} \sum_{i=1}^{4} (x_i - \overline{x})^2 = \frac{14}{3}$$

$$s_y^2 = \frac{1}{3} \sum (y_i - \overline{y})^2 = \frac{21}{3}$$

which results in the sample correlation coefficient between the two random variables of

$$r = \frac{17}{\sqrt{14 \times 21}} = 0.99.$$

This is in fact an example of nearly perfect correlation between the two variables. In fact, positive deviations of one variable from the sample mean are accompanied by positive deviations of the other by nearly the same amount. ◇

2.5.2 Statistical Independence of Random Variables

The independence between events was described and quantified in Chap. 1, where it was shown that two events are independent only when the probability of their intersection is the product of the individual probabilities. The concept is extended here to random variables by defining two random variables as *independent* if and only if the joint probability distribution function can be factored in the following form:

$$h(x, y) = f(x) \cdot g(y), \tag{2.21}$$

where $f(x)$ and $g(y)$ are the probability distribution functions of the two random variables. When two variables are independent, the individual probability distribution function of each variable is obtained via *marginalization* of the joint distribution with respect to the other variable, e.g.,

$$f(x) = \int_{-\infty}^{+\infty} h(x, y)dy. \tag{2.22}$$

It is important to remark that independence between random variables and uncorrelation are not equivalent properties. Independence, which is a property of the distribution functions, is a much stronger property than uncorrelation, which is based on a statement that involves only moments. It can be proven that independence implies uncorrelation, but not vice versa.

Proof The fact that independence implies uncorrelation is shown by calculating the covariance of two independent random variables of joint distribution function $h(x, y)$. The covariance is

$$\sigma_{xy}^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_x)(y-\mu_y)h(x,y)dxdy =$$
$$\int_{-\infty}^{+\infty}(x-\mu_x)f(x)dx\int_{-\infty}^{+\infty}(y-\mu_y)g(y)dy = 0,$$

since each integral vanishes as the expectation of the deviation of a random variable. □

As a counter-example of the fact that dependent variables can have non-zero correlation factor, consider the case of a random variable X with a distribution $f(x)$ that is symmetric around the origin, and another variable $Y = X^2$. They cannot be independent since they are functionally related, but it will be shown that their covariance is zero. Symmetry about zero implies $\mu_x = 0$. The mean of Y is $E[Y] = E[X^2] = \sigma_x^2$ since the mean of X is null. From this, the covariance is given by

$$Cov(X, Y) = E[X(Y-\sigma_X^2)] = E[X^3 - X\sigma_x^2] = E[X^3] = 0$$

due to the symmetry of $f(x)$. Therefore the two variables X and X^2 are uncorrelated, yet they are not independent.

Example 2.5 (Photon Counting Experiment) A photon-counting experiment consists of measuring the total number of photons in a given time interval and the number of background events detected by the receiver in the same time interval. The experiment is repeated six times, by measuring simultaneously the total number of counts T as $(10, 13, 11, 8, 10, 14)$ and the number of background counts B as $(2, 3, 2, 1, 1, 3)$. We want to estimate the mean number of source photons and its standard error.

The random variable we seek to measure is $S = T - B$ and the mean and variance of this random variable can be easily shown to be

$$\mu_S = \mu_T - \mu_B$$
$$\sigma_S^2 = \sigma_T^2 + \sigma_B^2 - 2\sigma_{TB}^2$$

(the derivation is similar to that of Example 2.3). From the data, we measure the sample means and variances as $\overline{T} = 11.0$, $\overline{B} = 2.0$, $s_T^2 = 4.8$, $s_B^2 = 0.8$ and the sample covariance as $s_{TB}^2 = +1.6$.

Notice that the correlation coefficient between T and S, as estimated via the measurements, is then given by $corr(T, B) = 1.6/\sqrt{4.8 \times 0.8} = 0.92$, indicating a strong degree of correlation between the two measurements. The measurements can be summarized as

$$\mu_S = 11.0 - 2.0 = 9.0$$
$$\sigma_S^2 = 4.8 + 0.8 - 2 \times 1.6 = 2.4$$

and be reported as $S = 9.00 \pm 1.55$ counts (per time interval). Notice that if the correlation between the two measurements had been neglected, then one would (erroneously) report $S = 9.00 \pm 2.37$, e.g., the standard deviation would be largely overestimated. The correlation between total counts and background counts in this example has a significant impact in the calculation of the variance of S and needs to be taken into account. ◇

2.6 A Classic Experiment: Pearson's Collection of Data on Biometric Characteristics

In 1903 K. Pearson published the analysis of a collection of biometric data on more than 1000 families in the United Kingdom, with the goal of establishing how certain characters, such as height, are correlated and inherited [33]. Prof. Pearson is also the inventor of the χ^2 test and a central figure in the development of the modern science of statistics.

Pearson asked a number of families, composed of at least the father, mother, and one son or daughter, to perform measurements of height, span of arms and length of left forearm. This collection of data resulted in a number

(continued)

of tables, including some for which Pearson provides the distribution of two measurements at a time. One such table is that reporting the mother's height versus the father's height, Table 2.3.

The data reported in Table 2.3 represent the joint probability distribution of the two physical characters, binned in one-inch intervals. When a non-integer count is reported (e.g., a value of 0.25, 0.5 or 0.75), we interpret it as meaning that the original measurement fell exactly at the boundary between two cells, although Pearson does not provide an explanation for non-integer values.

For every column and row it is also reported the sum of all counts. The bottom row in the table is therefore the distribution of the father's height, irrespective of the mother's height, likewise the rightmost column is the distribution of the mother's height, regardless of the father's height. The process of obtaining a one-dimensional distribution from a multi-dimensional illustrates the *marginalization* over certain variables that are not of interest. In the case of the bottom column, the marginalization of the distribution was done over the mother's height, to obtain the distribution of father's height.

From Table 2.3 it is not possible to determine whether there is a correlation between father's and mother's heights. In fact, according to (2.18), we would need all 1079 pairs of height measurements originally collected by Pearson to calculate the covariance. Since Pearson did not report these *raw* (i.e, unprocessed) data, we cannot calculate either the covariance or the correlation coefficient. The measurements reported by Pearson are in a format that goes under the name of *contingency table*, consisting of a table with measurements that are binned into suitable two-dimensional intervals.

Summary of Key Concepts for this Chapter

□ *Random variable:* A quantity that is not known exactly and is described by a probability distribution function $f(x)$.

□ *Moments of a distribution:* Expectations for the random variable or functions of the random variable, such as the mean $\mu = E[X]$ and the variance $\sigma^2 = E[(X-\mu)^2]$.

□ *Sample mean and sample variance*: Quantities calculated from the measurements that are intended to approximate the corresponding parent quantities (mean and variance).

□ *Joint distribution function*: The distribution of probabilities for a pair of variables.

□ *Covariance:* A measure of the tendency of two variables to follow one another, $Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$.

□ *Correlation coefficient:* A normalized version of the covariance that takes values between -1 (perfect anti-correlation) and +1 (perfect correlation).

□ *Statistically independent variables:* Two variables whose joint probability distribution function can be factored as $h(x,y) = f(x)g(y)$.

Table 2.3 Joint distribution of father's height (columns) and mother's height (rows) from Pearson's experiment,in inches

		Father's height																	
		58–	59–	60–	61–	62–	63–	64–	65–	66–	67–	68–	69–	70–	71–	72–	73–	74–	75–	
Mother's height	52–53	0	0	0	0	0	0	0	1	0.5	0	0	0	0	0	0	0	0	0	1.5
	53–54	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0.5
	54–55	0	0	0	0	0	0	0	0.25	0.25	0	0.5	0	0	0	0	0	0	0	1
	55–56	0	0	0	0.5	1	0	0	0.25	0.25	0	0.5	0	0	0	0	0	0	0	2.5
	56–57	0	0	0	0	0.75	1.25	0	1	1.75	1.75	0	0	0	0	0	0	0	0	6.5
	57–58	0	0	0	0.25	1	1.25	1.5	4	3.25	2.5	3	1.25	0.5	0	0	0	0	0	18.5
	58–59	0	0.25	0.75	1.25	1.25	2.75	4	7	5.75	4.5	3.75	1.25	2	0	0	0	0	0	34.5
	59–60	0	1.25	1.25	1	4	4.5	7.75	10	15	16.75	9	5.5	3.25	1.25	1	0.5	0	0	82
	60–61	0.25	0.25	0.5	2	4.25	4.5	18	16	24	14.75	23.25	12.75	7.25	5.75	4.25	0.75	0	0	138.5
	61–62	0.25	0.25	0	0	8	8.25	15	17.25	25	20.75	24	14.25	14.25	10	4	0.75	0.5	0	162.5
	62–63	0	0.5	0.5	1.25	4.75	7.75	10	26	21.25	28	28	23	14.25	10.75	4.5	2	1	0.5	184
	63–64	0	0	0.25	2	3.5	4.5	9	21	15.75	20.75	19.5	24	22.5	10.75	4	2.25	2.25	0.5	162.5
	64–65	0	0	1.25	0.75	2	6	6.5	9.75	16	18.25	23	16.75	13.75	6.75	4.75	2.25	0.25	1.5	129.5
	65–66	0	0	0	0.25	1.5	1.5	3.25	5.5	9.75	7	15.5	12.75	10.5	6.25	4.25	1.75	0.25	0	80
	66–67	0	0	0	0.25	1	0.75	0.5	3.5	5	3	7.25	7.75	7	3.5	2.75	1.5	0.25	0	44
	67–68	0	0	0	0	0	0	0	1	2.5	1.5	2.75	3.25	2.75	1.5	1	0.5	0.25	0	17
	68–69	0	0	0	0	0	0	0	0	0	1	2.5	1.25	1.25	0.5	1	0.25	0.25	0	8
	69–70	0	0	0	0	0	0	0	0	0	0	0	0.25	2.25	0	2	0	0	0	4.5
	70–71	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.5	0	0	0	1.5
		0.5	2.5	4.5	9.5	33	43	75.5	123.5	146.5	140.5	162.5	124	102.5	57	34	12.5	5	2.5	1079

Problems

2.1 Consider the exponential distribution

$$f(x) = \lambda e^{-\lambda x}$$

where $\lambda \geq 0$ and $x \geq 0$. Show that the distribution is properly normalized, and calculate the mean, variance and cumulative distribution $F(x)$.

2.2 Consider the sample mean as a random variable defined by

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2.23}$$

where x_i are identical independent random variables with mean μ and variance σ^2. Show that the variance of $\overline{x}$ is equal to σ^2/N.

2.3 J.J. Thomson's experiment aimed at the measurement of the ratio between the mass and charge of the electron is presented on page 23. Using the datasets for Tube 1 and Tube 2 separately, calculate the mean and variance of the random variables W/Q and I, and the covariance and correlation coefficient between W/Q and I.

2.4 Using J.J. Thomson's experiment (page 23), verify the statement that *"It will be seen from these tables that the value of m/e is independent of the nature of the gas"* used in the experiment. You may do so by calculating the mean and standard deviation for the measurements in each gas (air, hydrogen, and carbonic acid) and testing whether the three measurements agree with each other within their standard deviations.

2.5 Calculate the sample covariance and correlation coefficient for the following set of data: $(0, 2), (2, 5), (1, 4), (3, 1)$.

2.6 Prove that the following relationship holds,

$$Var(X) = E[X^2] - \mu^2$$

where μ is the mean of the random variable X.