B
Basics of Probability

The key concept discussed here is that of the *discrete random variable.* Such a variable X is defined by considering the results of a sequence of random events E_i. We associate a real number x_i with each event E_i, and define X as the set of all the different x_is (a random variable can also be continuous, but this case is not discussed here).

A simple example is the set of results of throwing a die. The results are numbers between 1 and 6, obtained with equal probability. A more interesting example is the result of throwing two dice. The result is a number between 2 and 12, but these are obtained with different probabilities. A result of 2 can be obtained only if each die falls on 1; thus, it has low probability. A result of 12, similarly, is also obtained with low probability. A result of 6, however, is obtained when the two dice fall on (1,5), (2,4), or (3,3), so it is much more common. A practical example is the pixels of an image. In an image with 8-bit pixels, e.g., the value of a pixel is between 0 and 255, but not all values occur with the same probability. Imagine two images, **A**, with a lot of red and **B**, with mostly blue. The pixel distributions of those images form different random variables, because a pixel with, say, value 112 may occur in **A** with a probability 0.01 and in **B**, with probability 0.02.

◇ **Exercise B.1:** Consider the event of throwing three coins, where a coin can fall on a head (H) or a tail (T). Let our random variable X be the number of coins that fall on heads. List all the values of X and their probabilities.

It is clear that we can write $X = (x_1, p_1, x_2, p_2 \ldots, x_n, p_n)$ where x_i are the different values that X takes and p_i is the probability of x_i. The average (or *mean*) of X is $\overline{X} = (1/n)\sum x_i$.

The expected value, or *expectation*, of the variable X is denoted $E(X)$ and is defined as $E(X) = \sum p_i x_i = \sum x_i P(X = x_i)$. This is the sum of the possible values x_i of X, each weighted by its probability p_i. The notation $P(X = x_i)$ should be read "the probability that X will take the value x_i." If all the probabilities p_i are the same ($p_i = 1/n$), then $E(X) = \overline{X}$. In the case of three coins, the expected number of heads is $E(X) = \overline{X} = (3+2+2+1+2+1+1+0)/8 = 3/2$. We don't expect, of course, any individual event (throwing) to produce 1.5 heads, but this will be the number obtained on average.

◇ **Exercise B.2:** What is the expected value of the random variable X that takes just one value v?

Next we consider the two random variables $X = (0, 20)$ and $Y = (10, 10)$. In either case, the expected value of the variable is 10. However, variable Y always equals 10, whereas X never takes this value. This indicates that the expectation of a random variable does not give enough information about it, and we should consider how far the individual values x_i are from the expectation $E(X)$ (how far the individual values deviate from the expected value). We thus want to derive an expression for the *variance* of a random variable. Intuitively it seems that Y should have a zero variance, while the variance of X should be nonzero.

The first method that comes to mind is to calculate the sum of the differences $[x_i - E(X)]$. This definition of the variance is simple, but also counterintuitive, since it produces a zero in the case of X. Better results are obtained when the absolute differences $|x_i - E(X)|$ are used, instead of the differences.

To understand why even this definition is not satisfactory, consider the case of $n-1$ values x_i, perhaps the results of $n-1$ measurements of lengths, that are all in the range $[0, 10]$. The expectation is also in this range. We now make the next measurement and get the value $x_n = 1000$. It makes sense to require that this single value, which is so far from the other ones and from the expectation, should be given more weight in the calculation of the variance. This is why the squares of the differences, instead of the differences themselves, are used in the definition of the variance. If the expectation is 10, then a value $x_i = 5$ contributes $(5-10)^2 = 25$, but a value $x_i = 110$ makes the much bigger contribution of $(110-10)^2 = 10000$.

This is why the variance of the random variable X is defined as (see also page 370)

$$\mathrm{Var}(X) = \sum p_i [x_i - E(X)]^2.$$

If all the p_i are identical, then $p_i = 1/n$ and the variance becomes

$$\mathrm{Var}(X) = \frac{1}{n} \sum [x_i - E(X)]^2.$$

The standard deviation is another useful statistical measure. It is defined as the square root of the variance.

The variance is a natural unit for measuring how a variable varies, but the standard deviation is also useful because a random variable X has a dimension such as height or weight. The dimension of the variance is the square of the dimension of the variable (height squared or weight squared), so the variance and the random variable cannot be used together in calculations (for example, they cannot be added or subtracted). The standard deviation, however, has the dimension of the variable.

> The computer informed here that three spaces accounted for eighty-one percent of variance.
>
> — Michael Crichton, *The Terminal Man*

The standard deviation is a measure of the "radius" of the variable. If X has a Gaussian distribution (Section B.3.1), 68% of its values will be within one standard deviation of the mean, and 95% will be within two standard deviations.

◇ **Exercise B.3:** The definition above implies $\mathrm{Var}(X) = E[(X - E(X))^2]$. Show that $\mathrm{Var}(X) = E(X^2) - [E(X)]^2$.

Bivariate observations are those in which values of two random variables, X and Y, are taken. Such observations are referred to as "paired," and pairing can occur in a number of situations. The following are some examples:

1. When there are two different variables for each case (e.g., age and shoe size, height and weight, sex and IQ, country's infant mortality and average education).

2. When the same variable is measured for each case at two different times (e.g., reading level before training and reading level after, IQ at age 3 and IQ at age 6).

3. When the same or different variables are measured from related cases (e.g., father's and son's educational attainment, husband's height and wife's height; mother's anxiety and child's security).

4. When the same or different variables are measured in unrelated cases at the same time (for example, unemployment rate in city A and in city B in a given month).

The two variables can be independent or dependent. Imagine two persons living in different areas, whose phone numbers (minus the area codes) are the same. We intuitively feel that these persons' heights are independent. On the other hand, the incomes of two persons having similar phone numbers within the same area code may be dependent (although perhaps just to a small degree), because such persons may live close to each other, so there is a chance that they interact or even work together.

The question is how to measure the relation between two random variables. It is easy to define independence. Two random variables X and Y are independent if

$$P(X = a, Y = b) = P(X = a)P(Y = b), \quad \text{for any } a,\, b.$$

The *covariance* of two random variables is a measure of the linear relation between them. The covariance indicates only the direction of the linear relation, not its strength. It is defined as

$$COV_{xy} = \frac{\sum(x_i - \overline{X})(y_i - \overline{Y})}{n-1}.$$

The following simple Matlab code calculates the covariance matrix of a square matrix where each column is a variable:

```
function xy=covarmat(x)
[m,n]=size(x);
xc=x-repmat(sum(x)/m,m,1); % subtract average
xy=xc' * xc/(m-1);
```

A positive covariance indicates that the values of one variable increase as the values of the other increase. A negative covariance indicates that the values of one variable decrease as the values of the other increase. A zero covariance indicates that there is no linear relation between the two variables; they are *decorrelated.* The precise value of the covariance normally does not have any meaning because it depends on the units of measurement of X and Y and on their variances. To actually measure the amount of correlation between two variables we use the statistical measure of *correlation.* It is defined as

$$R = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}}. \tag{B.1}$$

It measures the linear relation between two paired variables. The values of R range from -1 (perfect negative relation), to 0 (no relation), to $+1$ (perfect positive relation).

Assume that we have two arrays x_i and y_i of numbers, perhaps measurements of the height and weight of a group of people. The correlation coefficient of the two arrays is a measure of how well the variation of one of them (say, y_i) can be predicted from the variation of the other (x_i). We can also consider x_i and y_i variables instead of arrays. If we look at x_i as the independent variable and at y_i as the dependent variable, then the correlation coefficient of the two variables is a measure of how well the variation of the dependent variable can be predicted from the variation of the independent variable.

If $y_i = ax_i$ for every i (for some real a), then the variation of y_i can easily and accurately be predicted from that of x_i and the correlation coefficient is 1 (or -1 if a is negative). If y follows x (i.e., if $x_{i+1} < x_i$ implies that $y_{i+1} < y_i$, and $x_{i+1} > x_i$ implies that $y_{i+1} > y_i$), then the correlation coefficient is a positive number. In the opposite case (where $x_{i+1} < x_i$ implies $y_{i+1} > y_i$, and $x_{i+1} > x_i$ implies $y_{i+1} < y_i$), the correlation coefficient is a negative number.

> Variance is what any two statisticians are at.
>
> — Unknown

B.1 Joint and Union of Events

We first discuss the probabilities of *independent events.* Those are events that do not affect each other's probabilities. An example is the throw of a die or of a coin. Each outcome of a throw is independent of the preceding (and also of the following) throws. Even when two dice are thrown together, the two outcomes are independent. We denote the probability of outcome x by $P(x)$. The *joint probability* $P(x \cdot y)$ of the independent events x and y is the probability that one event will result in x and another event, in y. This probability is simply the product of the individual probabilities. For example, if the probability of having snow today (event A) is 25% and that of having snow tomorrow (event B) is 40%, then the probability of having snow today *and* tomorrow, $P(A\cdot B)$, is the product $0.25{\times}0.4 = 0.1$, i.e., 10% (it is, of course, smaller than each of the individual probabilities). The probability of the *complement* $\bar{x}$ of an event x equals one minus the probability of the event, $P(\bar{x}) = 1-P(x)$. Thus, the probability of *not* having snow today is $1-0.25 = 0.75$.

Another important concept is the probability of the *union* of events. For example, what is the probability of having snow today *or* tomorrow? To understand how this is computed, consider the opposite case. The opposite of it snowing today or snowing tomorrow is the case where it does not snow today (event $\bar{A}$) and it does not snow tomorrow (event $\bar{B}$). The probability of this event is, of course, the joint probability $P(\bar{A}\cdot\bar{B})$, i.e., $P(\bar{A})P(\bar{B})$, so we conclude that this joint event is the complement of the union that we are looking for. Thus, we end up with $P(A+B) = 1-P(\bar{A}\cdot\bar{B}) = 1-P(\bar{A})P(\bar{B})$. In the snowing example the result is $P(A+B) = 1-P(\bar{A})P(\bar{B}) = 1-(1-0.25)(1-0.4) = 1-0.45 = 0.55$. This result is greater than 25% and 40%, but it is still a probability, i.e., in the range $[0,1]$.

◇ **Exercise B.4:** (For gamblers.) Calculate the probability of winning in the following two games. In game A, the player rolls two dice up to 24 times and wins if he rolls a double-six. In game B the player rolls a single die and wins if he gets a six in four rolls. These winning probabilities were first calculated by Blaise Pascal, one of the founders of modern probability theory, in 1654.

B.2 Conditional Probability

Not all events are independent. When dealing with dependent events, we have to calculate *conditional probabilities*. We usually ask the question: What is the probability of event A given that event B has occurred? This is the conditional probability of A (more precisely, the probability of A conditioned on B) and it is denoted by $P(A|B)$. The field of conditional probability is sometimes called *Bayesian statistics*, since it was first developed by the Reverend Thomas Bayes [1702–1761], who came up with the basic formula

$$P(A|B) = \frac{P(A\cdot B)}{P(B)}. \tag{B.2}$$

Example: Three cards are given. One is red on both sides, another is black on both sides, and the third is red on one side and black on the other side. You pick up a card without looking at it, place it on the table, then look at it. If you see a red side, what is the probability of the other side also being red? This probability is conditional and is denoted by $P(\mathit{Rup}|\mathit{Rdown})$. Equation (B.2) becomes

$$P(\mathit{Rup}|\mathit{Rdown}) = \frac{P(\mathit{Rup}\cdot\mathit{Rdown})}{P(\mathit{Rdown})}.$$

The probability $P(\mathit{Rup}\cdot\mathit{Rdown})$ is $1/3$, since only one of the three cards has two red sides. The probability $P(\mathit{Rdown})$ can be calculated as follows: We have three cards. The probability of any card being picked up is 1/3. The probability that the first card will be picked up *and* will have red down is $(1/3)\times 1$, since both its sides are red. The probability that the second card will be picked up *and* will have red down is $(1/3)\times 0$, since none of its sides is red. The probability that the third card will be picked up *and* will have red down is $(1/3)(1/2)$, since only one of its sides is red. Thus, the total probability of picking a card and having red down is $(1/3)(1+0+1/2)=1/2$.

The same result can be obtained by considering the events points in some *event space*. The first card will have red down no matter what, so it contributes a point (point 1, with probability 1/3) to the event space. The second card will never have red down, no matter what. It contributes another point, point 2, also with probability 1/3, to the event space. The third card will have red down if picked up in a certain way, and red on top if picked up another way. It therefore contributes two points (points 3 and 4, with probability 1/6 each) to the event space. We are interested in points 1 and 3, whose probabilities are 1/3 and 1/6, that add up to 1/2. Equation (B.2) therefore yields the conditional probability

$$P(Rup|Rdown)=\frac{P(Rup\cdot Rdown)}{P(Rdown)}=\frac{1/3}{1/2}=\frac{2}{3};$$

a nonintuitive result.

◇ **Exercise B.5:** Three companies A, B, and C are competing for a NASA contract to develop a new space vehicle. Experts agree that the chances of A, B, and C to win the contract are 2/5, 2/5, and 1/5, respectively. At a certain point company B withdraws from the competition. What are the winning chances of A and C now?

Example: A family has two children, where a child can be a boy or a girl. If nothing is known a priori about the children, then there are the four possibilities (BB, GG, BG, GB), and the probability of having two boys is 1/4. What is the conditional probability of the family having two boys, if it is known that one of the children is a boy? Intuitively this knowledge eliminates the GG case and reduces the possibilities to (BB, BG, GB), of which BB is what we are after. The conditional probability is therefore

$$P(\text{two boys}|\text{one boy})=\frac{P(BB)}{P(BB)+P(BG)+P(GB)}=\frac{\frac{1}{4}}{\frac{1}{4}+\frac{1}{4}+\frac{1}{4}}=\frac{1}{3}.$$

Using Bayesian statistics we have to calculate $P(A|B)$ where A is the event of having two boys and B is the event of having a boy. The quantity $P(A\cdot B)$ is the probability of having two boys *and* having a boy. Well, if a family has two boys, then it has a boy, so looking at the three possibilities (BB, GG, BG), we find that $P(A\cdot B)=P(A)=1/3$. The quantity $P(B)$ is the probability of having a boy. It equals 2/3 because this event makes up two of the three possibilities. Equation (B.2) yields

$$P(A|B)=\frac{P(A\cdot B)}{P(B)}=\frac{1/3}{2/3}=\frac{1}{2}.$$

An important practical case is where an event A has certain alternatives A_i. In such a case, Equation (B.2) can be generalized to

$$\begin{aligned}P(A_i|B) &= \frac{P(A_i \cdot B)}{P(A_1 \cdot B) + P(A_2 \cdot B) + \cdots + P(A_n \cdot B)}\\ &= \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_n)P(B|A_n).}\end{aligned}$$

This is the well-known Bayes' theorem, published posthumously by Thomas Bayes in 1763. It requires knowledge of the probabilities $P(A_i)$ of the alternatives and of the conditional probabilities $P(B|A_i)$ of B relative to the alternatives. This theorem is used by the QM coder to construct the probability estimation tables used by JPEG, JBIG, and JBIG2 (see, for example, Table 2.68).

Example: A student has to select one science course among mathematics, physics, chemistry, and biology. Based on his personal knowledge of the student, his advisor figures that the probabilities that the student will select one of the four classes are 0.4, 0.3, 0.2, and 0.1, respectively (these probabilities add up to 1). The advisor does not know what course was actually selected by the student, but at the end of the semester, the advisor hears that the student got an A in the selected course. Based on the difficulties of these courses, the advisor estimates the probabilities of the student's getting an A in mathematics, physics, chemistry, and biology to be 0.1, 0.2, 0.3, and 0.9, respectively (these don't have to add up to 1). Using Bayes' theorem, the advisor can now revise the original probabilities that the student will select one of the four courses. The probability that the student has actually selected, for example, the mathematics course is

$$\begin{aligned}P(\text{selected math}|\text{got an A}) &= \frac{0.4 \cdot 0.1}{0.4 \cdot 0.1 + 0.3 \cdot 0.2 + 0.2 \cdot 0.3 + 0.1 \cdot 0.9}\\ &= \frac{4}{25} = 0.16.\end{aligned}$$

◇ **Exercise B.6:** Based on the fact that the student got an A, calculate the new probabilities that the student has actually selected physics, chemistry, and biology.

Example: This is the well-known *three prisoners* problem. Tom, Dick, and Harry are in prison. The next morning one of them is going to be executed and the other two will be freed. They don't know who the unlucky one is, but their guard knows. Close to midnight, Tom asks the guard, "I don't know who is going to be executed, but I do know that either Dick or Harry is going to be freed. Please tell me which one because, after all, it does not make any difference in my probability of being executed." The guard (an amateur statistician who never heard of Bayes) answers, "You now estimate the probability of your being executed to be 1/3. If I told you that, say, Dick is going to be freed, you would know that only you and Harry remain candidates for execution, so the probability of your being executed would go up to 1/2 and you won't sleep well tonight." The guard's answer is simple but wrong, since it does not consider conditional probability. The guard could have given Tom a name, such as Dick, but this knowledge would not have changed Tom's conditional probability of being executed. Here is the analysis.

Without any prior knowledge, each prisoner has probability 1/3 of being executed and probability 2/3 of being freed. We denote by A the event that Tom will be executed. Clearly $P(A) = 1/3$. We want to calculate the conditional probability that Tom will be executed, given that he was told by the guard that Dick would be freed. Event B is, therefore, the guard's telling Tom that Dick is going to be released. The subtle point is that event B is not necessarily the release of Dick; it's the guard's telling Tom that Dick is to be freed. The difference is easy to see by considering all the possible events points in an event space. There are four such points, as follows:

Point 1: Harry is to be executed. In this case the guard tells Tom that Dick is to be freed. Since each prisoner has probability 1/3 of being executed, the probability of this point is 1/3.

Point 2: Dick is to be executed. In this case the guard tells Tom that Harry is to be freed. The probability of this point is, again, 1/3.

Points 3 and 4: Tom is to be executed (with probability 1/3). In this case the guard tells Tom that either Harry or Dick are going to go free. These are points 3 and 4, respectively, in the event space, each with a probability of 1/6.

The joint probability $P(A \cdot B)$ that Tom will be executed *and* the guard says that Dick will be freed is, therefore, the probability of point 4, or 1/6. The probability of event B (the guard will say that Dick is to be released) is the sum of points 1 and 4, or $1/3+1/6$. The conditional probability that Tom will be executed, if the guard says that Dick is going to be freed is, therefore,

$$P(A|B) = \frac{P(A \cdot B)}{P(B)} = \frac{1/6}{1/3+1/6} = 1/3,$$

the same probability of Tom's being executed without the guard's saying anything.

(The author is indebted to J. Robert Henderson for pointing out this problem and its solution.)

The conclusion is: Stay out of trouble, but if you are in it (trouble, i.e.) place your trust in old Reverend Bayes.

B.3 Probability Distributions

Imagine an input stream where symbols have equal probabilities of occurrence. An example may be an image consisting of N 8-bit pixels where each of the 256 possible colors appears exactly $N/256$ times. We say that the colors in this image have a *flat distribution* of values. The graph whose x-axis shows the 256 colors and whose y-axis shows the number of times each color appears will be a horizontal line. Now imagine a similar image, but with mostly green tones. There are many pixels with different shades of green and few with any other colors. The same graph will be high for x values that represent shades of green and low elsewhere. The distribution of colors in this image is not flat.

> The theory of symbiogenesis assumes that the most probable explanation for improbably complex structures (living or otherwise) lies in the association of less complicated parts. Sentences are easier to construct by combining words than by combining letters. Sentences then combine into paragraphs, paragraphs combine into chapters, and, eventually, chapters combine to form a book—highly improbable, but vastly more probable than the chance of arriving at a book by searching the space of possible combinations at the level of letters or words.
>
> — George B. Dyson, *Darwin Among The Machines.*

B.3.1 Gaussian Distribution

The Gaussian (also known as the Normal) distribution is an important statistical tool used in many branches of science. It provides a good model for continuous distributions that occur in many everyday situations. Examples are the following:

1. The distribution of peoples' heights. Most people are of medium height. Few are tall or short. Even fewer are very tall or very short. Practically no one is a giant or a dwarf. Imagine a sample of people whose height is known. If the sample is large enough and is not biased, the graph describing the number of people of height h as a function of h will look very similar to Figure B.1.

2. The speed of gas molecules. The molecules of a gas are in constant motion. They move randomly, collide with each other and with objects around them, and change their velocities all the time. However, most molecules in a given volume of gas move at about the same speed, and only a few move much faster or much slower than this speed. This speed is related to the temperature of the gas. The higher this average speed, the hotter the gas feels to us. (This example is asymmetric, since the minimum speed is zero, but the maximum speed can be very high.)

3. Château Chambord in the Loire valley of France has a magnificent staircase, designed by Leonardo da Vinci in the form of a double ramp spiral. Worn out by the innumerable footsteps of generations of tourists, the marble tread of this staircase now looks like an inverted normal distribution curve. It is worn mostly in the middle, were most people tend to step, and the wear tapers off to either side from the center. This staircase, and others like it, are physical embodiments of the abstract mathematical concept of probability distribution.

4. Prime numbers are familiar to most people. They are attractive and important to mathematicians because any positive integer can be expressed as a product of prime numbers (its *prime factors*) in one way only. The prime numbers are thus the building blocks from which all other integers can be constructed. It turns out that the number of distinct prime factors is distributed normally. Few integers have just one or two distinct prime factors, few integers have many distinct prime factors, while most integers have a small number of distinct prime factors. This is known as the Erdős-Kac theorem.

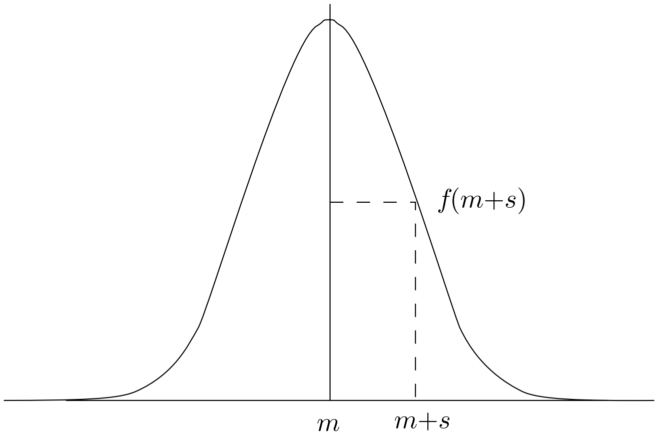

Figure B.1: Gaussian (Normal) Distribution.

> The general features of the bell curve make intuitive sense, but why it should have precisely the shape it does, which I can confidently predict with the help of a few lines of grade school arithmetic, and not some other shape—a little wider in the hip, say, or more pointed, like a witch's hat—that remains a mystery.
>
> — Hans Christian von Baeyer, *Maxwell's Demon*, 1998.

The Gaussian distribution with mean m and standard deviation s is defined as

$$f(x) = \frac{1}{s\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-m}{s}\right)^2\right\}.$$

This function has a maximum for $x = m$ (i.e., at the mean), where its value is $f(m) = 1/(s\sqrt{2\pi})$. It is also symmetric about $x = m$, since it depends on x according to $(x-m)^2$. It has the general "bell" shape of Figure B.1. At $x = m+s$ and $x = m-s$ its value is

$$f(m \pm s) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{1}{2}} \approx \frac{0.6065}{s\sqrt{2\pi}},$$

which means that at one standard deviation from the mean it has dropped to about 60% of its maximum. At two standard deviations it drops to about 0.1353 of its maximum value.

The total area under the normal curve is one unit. The area one standard deviation from the mean equals ≈ 0.682. At two standard deviations it equals ≈ 0.9545 and at three standard deviations it is ≈ 0.9973.

(The Gaussian distribution and the concept of standard deviation were discovered and discussed by Abraham de Moivre in 1733, when he was 18.)

The precise shape of the curve depends on s. As s increases, the curve gets wider. For small values of s the curve approaches a narrow spike of height $1/(s\sqrt{2\pi})$ placed at $x = m$.

When we talk about random numbers we usually mean numbers that are uniformly distributed over a certain interval $[a, b]$. If we divide interval $[a, b]$ into equal-size subintervals, any of them would contain the same amount of random numbers. It is possible to draw random numbers that have different distributions, such as Gaussian. When we compute many random numbers that are normally (i.e., Gaussian) distributed with mean m and standard deviation s, then count the amount y of these numbers in a small subinterval $[x, x + \epsilon]$, plot (x, y) as a point, and repeat for many subintervals, we get the normal distribution with mean m and standard deviation s.

> Because of the very nature of the tables, it did not seem necessary to proofread every page of the final manuscript in order to catch random errors.
>
> — *A Million Random Digits With 100,000 Normal Deviates*, RAND Corp., 1955.

Here are two ways to compute random numbers that are normally distributed with mean 0 and standard deviation 1.

1. Draw n uniformly-distributed random numbers R_i in the range $[-a, +a]$ for any real a and calculate their average $(1/n)\sum R_i$. The average is the first of the normally-distributed random numbers N_i. Repeat this process to get N_2, N_3, and so on. The larger n, the closer to normal will be the distribution of these numbers. The reason that the N_i are normally distributed is that it is rare for the average of R_i to be $-a$ or $+a$ or close to these values, but it is common for it to be around 0. This is an aspect of the *law of large numbers* that says: if R_i are random numbers of any distribution, then the averages $(1/n)\sum R_i$ are normally distributed.

2. Method 1 above is simple but slow, since n should be large. The *Polar method* (see [Knuth 81] Vol. 2, Sec. 3.4.1) is more efficient. Let U_1 and U_2 be two uniformly-distributed random numbers in the range $[0, 1]$. We calculate two normally-distributed random numbers X_1, X_2 by the following two simple steps:

Step 1. Compute $V_1 := 2U_1 - 1$, $V_2 := 2U_2 - 1$, and $S := V_1^2 + V_2^2$.

Step 2. If $S \geq 1$ go to step 1; else compute $X_1 := V_1\sqrt{\frac{-2\ln S}{S}}$, $X_2 := V_2\sqrt{\frac{-2\ln S}{S}}$.

Once a sequence N_i is obtained of normally-distributed random numbers with mean 0 and standard deviation 1, it is easy to convert them to normally-distributed random numbers with mean m and standard deviation s. Just transform each N_i to $m + N_i \times s$.

Assuming that a function `Rnd()`, which returns uniformly-distributed random numbers in the range $[0, 1]$, is given. Gaussian random numbers with zero mean and a standard deviation of one can be obtained by the following:

```
x:=0.0;
for i:=1 to 12 do x:=x+Rnd();
Gauss:=x-6.0;
```

B.3.2 Laplace Distribution

The Laplace probability distribution is similar to the normal (Gaussian) distribution, but is narrower and sharply peaked. The general Laplace distribution with variance V and mean m is given by

$$L(V,x)=\frac{1}{\sqrt{2V}}\exp\left(-\sqrt{\frac{2}{V}}|x-m|\right).$$

Table B.2 shows some values for the Laplace distributions with $m=0$ and $V=3,4,5$, and 1,000.

	x					
V	0	2	4	6	8	10
3:	0.408248	0.0797489	0.015578	0.00304316	0.00059446	0.000116125
4:	0.353553	0.0859547	0.020897	0.00508042	0.00123513	0.000300282
5:	0.316228	0.0892598	0.025194	0.00711162	0.00200736	0.000566605
1,000:	0.022360	0.0204475	0.018698	0.0170982	0.0156353	0.0142976

Table B.2: Some Values of the Laplace Distribution with $V=3,4,5$, and 1,000.

The factor $1/\sqrt{2V}$ is included in the definition of the Laplace distribution in order to scale the area under the curve of the distribution to 1.

◇ **Exercise B.7:** What is the indefinite integral of the Laplace distribution?

The Laplace distribution is used by the MLP image compression method (Section 4.19).

B.3.3 Discrete Distributions

Imagine n independent events being performed such that in each event a certain result (that we call a "success") occurs with the same probability p. A simple example is a coin throw, where a success is defined as the coin falling on head. The probability is, of course, 0.5. The number of successes in the n events is a random variable X that has a *binomial distribution.* The probability of X taking a value x is

$$P(X=x)=\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x},\quad \text{for } x=0,1,\ldots,n,$$

and the binomial distribution function is

$$P(X\le x)=\sum_{i=0}^{n}\frac{n!}{i!(n-i)!}p^i(1-p)^{n-i}.$$

> There is a story about two friends, who were classmates in high school, talking about their jobs. One of them became a statistician and was working on population trends. He showed a reprint to his former classmate. The reprint started, as usual, with the Gaussian distribution, and the statistician explained to his former classmate the meaning of the symbols for the actual population, for the average population, and so on. His classmate was a bit incredulous and was not quite sure whether the statistician was pulling his leg.
>
> "How can you know that?" was his query. "And what is this symbol here?"
>
> "Oh," said the statistician, "this is π."
>
> "What is that?"
>
> "The ratio of the circumference of the circle to its diameter."
>
> "Well now you are pushing your joke too far," said the classmate, "surely the population has nothing to do with the circumference of the circle."
>
> — Eugene P. Wigner, 1960

The mean of X is np and its variance is $np(1-p)$.

If the number of occurrences of an event x per unit (unit of time, length, area, volume, or whatever) is, on average, a real positive number λ, then the actual number of such occurrences is a discrete random variable X with a Poisson distribution. The probability that X will take a value x is

$$P(X=x)=\frac{\lambda^x e^{-\lambda}}{x!}, \quad \text{for } x=0,1,\ldots,\infty,$$

and the distribution function is

$$P(X\le x)=\sum_{i=0}^{x}\frac{\lambda^i e^{-\lambda}}{i!}.$$

The mean and the variance of the Poisson distribution both equal λ.

> The theory of probability is at bottom nothing
> but common sense reduced to calculus.
>
> Laplace, Pierre Simon de [1749–1827]

> Socrates took poisson
>
> Unknown

Siméon Denis Poisson [1781–1840]

Originally intended by his parents to study medicine, Poisson instead shifted to mathematics. He published his first book, on finite differences, at age 18.

Poisson taught at the École Polytechnique from 1802. In 1808 he became an astronomer at the Bureau des Longitudes. In 1809 he was appointed chair of pure mathematics in the newly established Faculté des Sciences.

His most important works are on definite integrals, Fourier series, probability theory, and mechanics.

In 1837 he published an important work on probability where he introduced the Poisson distribution. This distribution describes the probability that a random event will occur in a time or space interval under the conditions that the probability of the event occurring is very small, but the number of trials is very large, so that the event actually occurs a few times.

His important text *Traité de mécanique* was published in 1811 and again in 1833. This book was the standard work on mechanics for many years.

The mathematician Guglielmo Libri said of him: His only passion has been science: he lived and is dead for it.

One of Poisson's best known quotations is: Life is good for only two things, discovering mathematics and teaching mathematics.