

## Lecture Notes - Class 8

November 18, 2010.

# Random Variables and Probability Distributions

When we perform an experiment we are often interested not in the particular outcome that occurs, but rather in some *number* associated with that outcome.

For example, in the game of “craps” a player is interested not in the particular numbers on the two dice, but in their *sum*. In tossing a coin 50 times, we may be interested only in the *number* of heads obtained, and not in the particular sequence of heads and tails that constitute the result of 50 tosses.

In both examples, we have a rule which assigns to each outcome of the experiment a single real number. Hence, we can say that a *function* is defined.

You guys are already familiar with the function concept. Now we are going to look at some functions that are particularly useful to study probabilistic/statistical problems.

## Random Variables

In probability theory, certain functions of special interest are given special names:

**Definition 1** *A function whose domain is a sample space and whose range is some set of real numbers is called a random variable. If the random variable is denoted by  $X$  and has the sample space  $\Omega = \{o_1, o_2, \dots, o_n\}$  as domain, then we write  $X(o_k)$  for the value of  $X$  at element  $o_k$ . Thus  $X(o_k)$  is the real number that the function rule assigns to the element  $o_k$  of  $\Omega$ .*

Lets look at some examples of random variables:

**Example 1** *Let  $\Omega = \{1, 2, 3, 4, 5, 6\}$  and define  $X$  as follows:*

$$X(1) = X(2) = X(3) = 1, X(4) = X(5) = X(6) = -1.$$

Then  $X$  is a random variable whose domain is the sample space  $\Omega$  and whose range is the set  $\{1, -1\}$ .  $X$  can be interpreted as the gain of a player in a game in which a die is rolled, the player winning \$1 if the outcome is 1,2,or 3 and losing \$1 if the outcome is 4,5,6.

**Example 2** *Two dice are rolled and we define the familiar sample space*

$$\Omega = \{(1, 1), (1, 2), \dots, (6, 6)\}$$

*containing 36 elements. Let  $X$  denote the random variable whose value for any element of  $\Omega$  is the sum of the numbers on the two dice.*

Then the range of  $X$  is the set containing the 11 values of  $X$ :

$$2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12.$$

Each ordered pair of  $\Omega$  has associated with it exactly one element of the range as required by Definition 1. But, in general, the same value of  $X$  arises from many different outcomes.

For example  $X(o_k) = 5$  is any one of the four elements of the event

$$\{(1, 4), (2, 3), (3, 2), (4, 1)\}.$$

**Example 3** *A coin is tossed, and then tossed again. We define the sample space*

$$\Omega = \{HH, HT, TH, TT\}.$$

If  $X$  is the random variable whose value for any element of  $\Omega$  is the number of heads obtained, then

$$X(HH) = 2, \quad X(HT) = X(TH) = 1, \quad X(TT) = 0.$$

Notice that more than one random variable can be defined on the same sample space. For example, let  $Y$  denote the random variable whose value for any element of  $\Omega$  is the number of heads minus the number of tails. Then

$$X(HH) = 2, \quad X(HT) = X(TH) = 0, \quad X(TT) = -2.$$

Suppose now that a sample space

$$\Omega = \{o_1, o_2, \dots, o_n\}$$

is given, and that some acceptable assignment of probabilities has been made to the sample points in  $\Omega$ . Then if  $X$  is a random variable defined on  $\Omega$ , we can ask for the probability that the value of  $X$  is some number, say  $x$ .

The event that  $X$  has the value  $x$  is the subset of  $\Omega$  containing those elements  $o_k$  for which  $X(o_k) = x$ . If we denote by  $f(x)$  the probability of this event, then

$$f(x) = P(\{o_k \in \Omega | X(o_k) = x\}). \tag{1}$$

Because this notation is cumbersome, we shall write

$$f(x) = P(X = x), \tag{2}$$

adopting the shorthand “ $X = x$ ” to denote the event written out in (1).

**Definition 2** The function  $f$  whose value for each real number  $x$  is given by (2), or equivalently by (1), is called the **probability function** of the random variable  $X$ .

In other words, the probability function of  $X$  has the set of all real numbers as its domain, and the function assigns to each real number  $x$  the probability that  $X$  has the value  $x$ .

**Example 4** Continuing Example 1, if the die is fair, then

$$f(1) = P(X = 1) = \frac{1}{2}, f(-1) = P(X = -1) = \frac{1}{2},$$

and  $f(x) = 0$  if  $x$  is different from 1 or -1.

**Example 5** If both dice in Example 2 are fair and the rolls are independent, so that each sample point in  $\Omega$  has probability  $\frac{1}{36}$ , then we compute the value of the probability function at  $x = 5$  as follows:

$$f(5) = P(X = 5) = P(\{(1, 4), (2, 3), (3, 2), (4, 1)\}) = \frac{4}{36}.$$

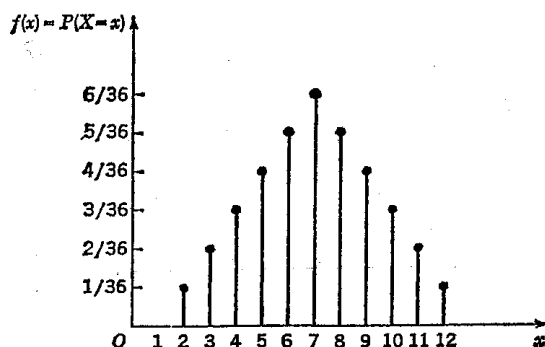
This is the probability that the sum of the numbers on the dice is 5. We can compute the probabilities  $f(2), f(3), \dots, f(12)$  in an analogous manner.

These values are summarized in the following table:

|        |                |                |                |                |                |                |                |                |                |                |                |
|--------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| $x$    | 2              | 3              | 4              | 5              | 6              | 7              | 8              | 9              | 10             | 11             | 12             |
| $f(x)$ | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

The table only includes those numbers  $x$  for which  $f(x) > 0$ . And since we include *all* such numbers, the probabilities  $f(x)$  in the table add to 1.

From the probability table of a random variable  $X$ , we can tell at a glance not only the various values of  $X$ , but also the probability with which each value occurs. This information can also be presented graphically, as in the following figure.



This is called the *probability chart* of the random variable  $X$ . The various values of  $X$  are indicated on the horizontal  $x$ -axis, and the length of the vertical line drawn from the  $x$ -axis to the point with coordinates  $(x, f(x))$  is the probability of the event that  $X$  has the value  $x$ .

Now, we are often interested not in the probability that the value of a random variable  $X$  is a particular number, but rather in the probability that  $X$  has some value *less than or equal to* some number.

In general, if  $X$  is defined on the sample space  $\Omega$ , then the event that  $X$  is less than or equal to some number, say  $x$ , is the subset of  $\Omega$  containing those elements  $o_k$  for which  $X(o_k) \leq x$ . If we denote by  $F(x)$  the probability of this event (assuming an acceptable assignment of probabilities has been made to the sample points  $\Omega$ ), then

$$F(x) = P(\{o_k \in \Omega | X(o_k) \leq x\}). \quad (3)$$

In analogy with our argument in (2), we adopt the shorthand " $X \leq x$ " to denote the event written out in (3), and then we can write

$$F(x) = P(X \leq x). \quad (4)$$

**Definition 3** *The function  $F$  whose value for each real number  $x$  is given by (4), or equivalently by (3), is called the **distribution function** of the random variable  $X$ .*

In other words, the distribution function of  $X$  has the set of all real numbers as its domain, and the function assigns to each real number  $x$  the probability that  $X$  has a value less than or equal to (i.e., at most) the number  $x$ .

It is an easy matter to calculate the values of  $F$ , the distribution function of a random variable  $X$ , when one knows  $f$ , the probability function of  $X$ .

**Example 6** *Lets continue with the dice experiment of Example 5.*

The event symbolized by  $X \leq 1$  is the null event of the sample space  $\Omega$ , since the sum of the numbers on the dice cannot be at most 1. Hence

$$F(1) = P(X \leq 1) = 0.$$

The event  $X \leq 2$  is the subset  $\{(1, 1)\}$ , which is the same as the event  $X = 2$ . Thus,

$$F(2) = P(X \leq 2) = f(2) = \frac{1}{36}.$$

The event  $X \leq 3$  is the subset  $\{(1, 1), (1, 2), (2, 1)\}$ , which is seen to be the union of the events  $X = 2$  and  $X = 3$ . Hence,

$$\begin{aligned} F(3) &= P(X \leq 3) = P(X = 2) + P(X = 3) \\ &= f(2) + f(3) \\ &= \frac{1}{36} + \frac{2}{36} = \frac{3}{36}. \end{aligned}$$

Similarly, the event  $X \leq 4$  is the union of the events  $X = 2$ ,  $X = 3$ , and  $X = 4$ , so that

$$\frac{1}{36} + \frac{2}{36} + \frac{3}{36} = \frac{6}{36}.$$

Continuing this way, we obtain the entries in the following *distribution table* for the random variable  $X$ :

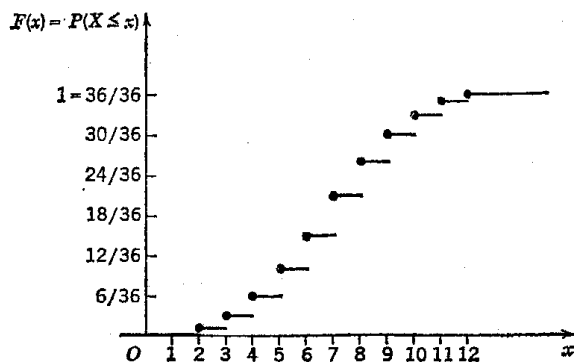
|        |                |                |                |                 |                 |                 |                 |                 |                 |                 |                 |
|--------|----------------|----------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| $x$    | 2              | 3              | 4              | 5               | 6               | 7               | 8               | 9               | 10              | 11              | 12              |
| $F(x)$ | $\frac{1}{36}$ | $\frac{3}{36}$ | $\frac{6}{36}$ | $\frac{10}{36}$ | $\frac{15}{36}$ | $\frac{21}{36}$ | $\frac{26}{36}$ | $\frac{30}{36}$ | $\frac{33}{36}$ | $\frac{35}{36}$ | $\frac{36}{36}$ |

Remember, though, that the domain of the distribution function  $F$  is the set of *all* real numbers. Hence, we must find the value  $F(x)$  for *all* numbers  $x$ , not just those in the distribution table. For example, to find  $F(2.6)$  we note that the event  $X \leq 2.6$  is the subset  $\{(1, 1)\}$ , since the sum of the numbers on the dice is less than or equal to 2.6 if and only if the sum is exactly 2. Therefore,

$$F(2.6) = P(X \leq 2.6) = \frac{1}{36}.$$

In fact,  $F(x) = \frac{1}{36}$  for all  $x$  in the interval  $2 \leq x < 3$ , since for any such  $x$  the event  $X \leq x$  is the same subset, namely  $\{(1, 1)\}$ . Note that this interval contains  $x = 2$ , but does not contain  $x = 3$ , since  $F(3) = \frac{3}{36}$ . Similarly, we find  $F(x) = \frac{3}{36}$  for all  $x$  in the interval  $3 \leq x < 4$ , but a jump occurs at  $x = 4$ , since  $F(4) = \frac{6}{36}$ .

These facts are shown on the following graph of the distribution function.



The graph consists entirely of horizontal line segments (i.e. it is a step function). We use a heavy dot to indicate which of the two horizontal segments should be read at each jump (step) in the graph. Note that the magnitude of the jump at  $x = 2$  is  $f(2) = \frac{1}{36}$ , the jump at  $x = 3$  is  $f(3) = \frac{2}{36}$ , the jump at  $x = 4$  is  $f(4) = \frac{6}{36}$ , etc.

Finally, since the sum of all numbers on the dice is never less than 2 and always at most 12, we have  $F(x) = 0$  if  $x < 2$  and  $F(x) = 1$  if  $x \geq 12$ .

If one knows the height of the graph of  $F$  at all points where jumps occur, then the entire graph of  $F$  is easily drawn. It is for this reason that we shall always list in the distribution table only those  $x$ -values at which jumps of  $F$  occur.

If we are given the graph of the distribution function  $F$  of a random variable  $X$ , then reading its height at any number  $x$ , we find  $F(x)$ , the probability that the value of  $X$  is less than or equal to  $x$ .

Also, we can determine the places where jumps in the graph occur, as well as the magnitude of each jump, and so we can construct the probability function of  $X$ . Thus, we can obtain the probability function from the distribution function, or vice versa!

## Probability Distributions

We have made our observations up to this point on the basis of some special examples, especially the two-dice example. I now turn to some general statements that apply to all probability and distribution functions of random variables defined on *finite* sample spaces.

Let  $X$  be a *finite random variable* on a sample space  $\Omega$ , that is,  $X$  assigns only a finite number of values to  $\Omega$ . Say,

$$R_X = \{x_1, x_2, \dots, x_n\}$$

(We assume that  $x_1 < x_2 < \dots < x_n$ .) Then,  $X$  induces a function  $f$  which assigns probabilities to the points in  $R_X$  as follows:

$$f(x_k) = P(X = x_k) = P(\{\omega \in \Omega : X(\omega) = x_k\})$$

The set of ordered pairs,  $[x_i, f(x_i)]$  is usually given in the form of a table as follows:

|        |          |          |          |         |          |
|--------|----------|----------|----------|---------|----------|
| $x$    | $x_1$    | $x_2$    | $x_3$    | $\dots$ | $x_n$    |
| $f(x)$ | $f(x_1)$ | $f(x_2)$ | $f(x_3)$ | $\dots$ | $f(x_n)$ |

The function  $f$  is called the *probability distribution* or, simply, *distribution*, of the random variable  $X$ ; it satisfies the following two conditions:

(i)  $f(x) \geq 0$  ( $x = 0, \pm 1, \pm 2, \dots$ )

(ii)  $\sum_{x=-\infty}^{\infty} f(x) = 1$ .

The second condition expresses the requirement that it is certain that  $X$  will take one of the available values of  $x$ . Observe also that

$$Prob(a \leq X \leq b) = \sum_{x=a}^b f(x).$$

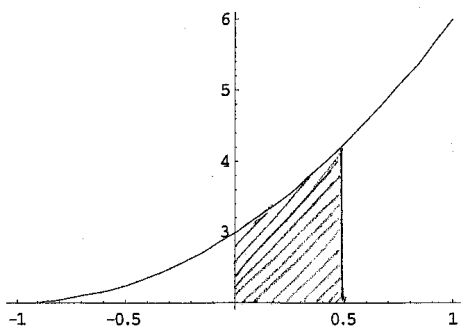
This latter observation leads us to the consideration of random variables which may take any real value.

Such random variables are called *continuous*. For the continuous case, the probability associated with any particular point is zero, and we can only assign positive probabilities to intervals in the range of  $x$ .

In particular, suppose that  $X$  is a random variable on a sample space  $\Omega$  whose range space  $R_X$  is a continuum of numbers such as an interval. We assume that there is a continuous function  $f: \mathbf{R} \rightarrow \mathbf{R}$  such that  $Prob(a \leq X \leq b)$  is equal to the area under the graph of  $f$  between  $x = a$  and  $x = b$ .

**Example 7** Suppose  $f(x) = x^2 + 2x + 3$ .

Then  $P(0 \leq X \leq 0.5)$  is the area under the graph of  $f$  between  $x = 0$  and  $x = 0.5$ .



In the language of calculus,

$$Prob(a \leq X \leq b) = \int_a^b f(x) dx$$

In this case, the function  $f$  is called the *probability density function (pdf)* of the continuous random variable  $X$ ; it satisfies the conditions

- (i)  $f(x) \geq 0$  (*all*  $x$ )
- (ii)  $\int_{-\infty}^{\infty} f(x) dx = 1$ .

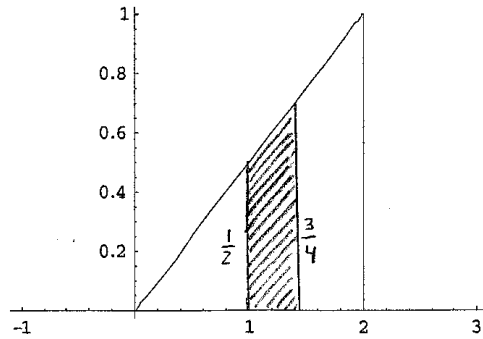
That is,  $f$  is nonnegative and the total area under its graph is 1.

The second condition expresses the requirement that it is certain that  $X$  will take some real value. If the range of  $X$  is not infinite, it is understood that  $f(x) = 0$  anywhere outside the appropriate range.

**Example 8** Let  $X$  be a random variable with the following pdf:

$$f(x) = \begin{cases} \frac{1}{2}x & \text{if } 0 \leq x \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

The graph of  $f$  looks like this:



Then, the probability  $P(1 \leq X \leq 1.5)$  is equal to the area of shaded region in diagram:

$$\begin{aligned} P(1 \leq X \leq 1.5) &= \int_1^{1.5} f(x) dx \\ &= \int_1^{1.5} \frac{1}{2}x dx \\ &= \frac{x^2}{4} \Big|_1^{1.5} = \frac{5}{16} \end{aligned}$$

Let  $X$  be a random variable (discrete or continuous). The *cumulative distribution function*  $F$  of  $X$  is the function  $F : \mathbf{R} \rightarrow \mathbf{R}$  defined by

$$F(a) = P(X \leq a).$$

Suppose  $X$  is a discrete random variable with distribution  $f$ . Then  $F$  is the “step function” defined by

$$F(x) = \sum_{x_i \leq x} f(x_i).$$



On the other hand, suppose  $X$  is a continuous random variable with distribution  $f$ . Then

$$F(x) = \int_{-\infty}^x f(t) dt,$$

In either case,  $F(x)$  must satisfy the following properties:

(i)  $F$  is *monotonically increasing*, that is,

$$F(a) \leq F(b)$$

whenever  $a \leq b$ .

(ii) The limit of  $F$  to the left is 0 and to the right is 1:

$$\lim_{x \rightarrow -\infty} F(x) = 0 \text{ and } \lim_{x \rightarrow \infty} F(x) = 1.$$

Finally, from the definition of the cdf,

$$\text{Prob}(a < X \leq b) = F(b) - F(a).$$

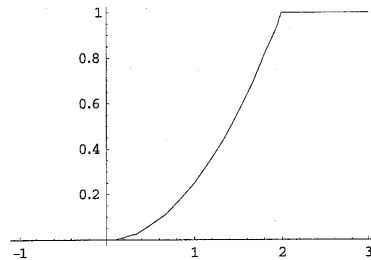
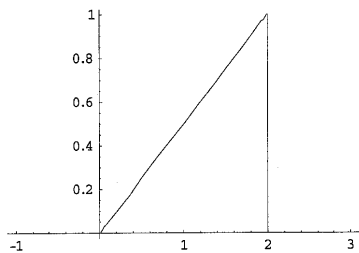
Any valid pdf will imply a valid cdf, so there is no need to verify this conditions separately.

**Example 9** Let  $X$  be a continuous random variable with the following pdf

$$f(x) = \begin{cases} \frac{1}{2}x & \text{if } 0 \leq x \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

The cdf of  $X$  follows

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{4}x^2 & \text{if } 0 \leq x \leq 2 \\ 1 & \text{if } x > 2 \end{cases}$$



Here we use the fact that, for  $0 \leq x \leq 2$ ,

$$F(x) = \int_0^x \frac{1}{2}x dx = \frac{1}{4}x^2$$

## Expectation

Let  $X$  be a finite random variable, and suppose the following is its distribution:

|        |          |          |          |         |          |
|--------|----------|----------|----------|---------|----------|
| $x$    | $x_1$    | $x_2$    | $x_3$    | $\dots$ | $x_n$    |
| $f(x)$ | $f(x_1)$ | $f(x_2)$ | $f(x_3)$ | $\dots$ | $f(x_n)$ |

Then the *mean*, or *expectation* (or *expected value*) of  $X$ , denoted by  $E(X)$ , or simply  $E$ , is defined by

$$E = E(X) = x_1f(x_1) + x_2f(x_2) + \dots + x_nf(x_n) = \sum x_i f(x_i)$$

Roughly speaking, if the  $x_i$  are numerical outcomes of an experiment, then  $E$  is the expected value of the experiment. We may also view  $E$  as the *weighted average* of the outcomes where each outcome is weighted by its probability.

So, suppose that  $X$  is a random variable with  $n$  distinct values  $x_1, x_2, \dots, x_n$  and suppose  $x_i$  occurs with the same probability  $p_i$ . Then  $p_i = \frac{1}{n}$ . Accordingly

$$E = E(X) = x_1\left(\frac{1}{n}\right) + x_2\left(\frac{1}{n}\right) + \dots + x_n\left(\frac{1}{n}\right) = \frac{x_1 + x_2 + \dots + x_n}{n}$$

This is precisely the *average* or *mean* value of the numbers  $x_1, x_2, \dots, x_n$ . For this reason  $E(X)$  is called the *mean* of the random variable  $X$ . Furthermore, since the Greek letter  $\mu$  is used for the mean value of a population, we also use  $\mu$  for the expectation of  $X$ . That is,

$$\boxed{\mu = \mu_X = E(X)}$$

Finally, the *expectation*  $E(X)$  for a continuous random variable  $X$  is defined by the following integral when it exists:

$$\int_{-\infty}^{\infty} xf(x) dx$$

## Variance

The mean of a random variable  $X$  measures, in a certain sense, the “average” value of  $X$ . The next two concepts, variance and standard deviation, measure the “spread” or “dispersion” of  $X$ .

Let  $X$  be a random variable with mean  $\mu = E(X)$  and the following probability distribution:

|        |          |          |          |         |          |
|--------|----------|----------|----------|---------|----------|
| $x$    | $x_1$    | $x_2$    | $x_3$    | $\dots$ | $x_n$    |
| $f(x)$ | $f(x_1)$ | $f(x_2)$ | $f(x_3)$ | $\dots$ | $f(x_n)$ |

The *variance* of  $X$ , denoted by  $var(X)$ , is defined by

$$\begin{aligned} var(X) &= (x_1 - \mu)^2 f(x_1) + (x_2 - \mu)^2 f(x_2) + \dots + (x_n - \mu)^2 f(x_n) \\ &= \sum (x_i - \mu)^2 f(x_i) \\ &= E((X - \mu)^2) \end{aligned}$$

The *standard deviation* of  $X$ , denoted by  $\sigma_X$  or simply  $\sigma$  is the nonnegative square root of  $var(X)$ , that is

$$\sigma_X = \sqrt{var(X)}$$

Accordingly,  $var(X) = \sigma_X^2$ . Both  $var(X)$  and  $\sigma^2$  are used to denote the variance of a random variable  $X$ .

The next theorem gives us an alternate and sometimes more useful formula for calculating the variance of a random variable  $X$ .

**Theorem 1**  $var(X) = x_1^2 f(x_1) + x_2^2 f(x_2) + \dots + x_n^2 f(x_n) - \mu^2 = \sum x_i^2 f(x_i) - \mu^2 = E(X^2) - \mu^2$

**Proof.** Using  $\sum x_i f(x_i) = \mu$  and  $\sum f(x_i) = 1$ , we obtain

$$\begin{aligned} \sum (x_i - \mu)^2 f(x_i) &= \sum (x_i^2 - 2\mu x_i + \mu^2) f(x_i) \\ &= \sum x_i^2 f(x_i) - 2\mu \sum x_i f(x_i) + \mu^2 \sum f(x_i) \\ &= \sum x_i^2 f(x_i) - 2\mu^2 + \mu^2 \\ &= \sum x_i^2 f(x_i) - \mu^2 \end{aligned}$$

This proves the theorem.

### Standardized Random Variable

Let  $X$  be a random variable with mean  $\mu$  and standard deviation  $\sigma > 0$ . The *standardized random variable*  $Z$  is defined by

$$Z = \frac{X - \mu}{\sigma}$$

The standardized random variable  $Z$  has mean  $\mu_Z = 0$  and standard deviation  $\sigma_Z = 1$ .

**Example 10** Suppose a random variable  $X$  has the following distribution:

|        |     |     |     |     |
|--------|-----|-----|-----|-----|
| $x$    | 2   | 4   | 6   | 8   |
| $f(x)$ | 0.1 | 0.2 | 0.3 | 0.4 |

The mean of  $X$  is

$$\mu = E(X) = \sum x_i f(x_i) = 2(0.1) + 4(0.2) + 6(0.3) + 8(0.4) = 6$$

and

$$E(X^2) = \sum x_i^2 f(x_i) = 2^2(0.1) + 4^2(0.2) + 6^2(0.3) + 8^2(0.4) = 40$$

Now using the last theorem, we obtain

$$\sigma^2 = \text{var}(X) = E(X^2) - \mu^2 = 40 - 6^2 = 4 \text{ and } \sigma = 2$$

Using  $z = \frac{x-\mu}{\sigma} = \frac{x-6}{2}$  and  $f(z) = f(x)$ , we obtain the following distribution for  $Z$ :

|        |     |     |     |     |
|--------|-----|-----|-----|-----|
| $z$    | -2  | -1  | 0   | 1   |
| $f(z)$ | 0.1 | 0.2 | 0.3 | 0.4 |

Then

$$\mu_Z = E(Z) = \sum z_i f(z_i) = -2(0.1) - 1(0.2) + 0(0.3) + 1(0.4) = 0$$

$$E(Z^2) = \sum z_i^2 f(z_i) = (-2)^2(0.1) + (-1)^2(0.2) + 0^2(0.3) + 1^2(0.4) = 1$$

And again, using the last theorem, we obtain

$$\sigma_Z^2 = \text{var}(Z) = E(Z^2) - \mu^2 = 1 - 0^2 = 1 \text{ and } \sigma_Z = 1$$

The *variance*  $\text{var}(X)$  for a continuous random variable  $X$  is defined by the following integral when it exists:

$$\text{var}(X) = E((X - \mu)^2) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Just as in the discrete case, it can be shown that  $\text{var}(X)$  exists if and only if  $\mu = E(X)$  and  $E(X^2)$  both exist and then

$$\text{var}(X) = E(X^2) - \mu^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

When  $\text{var}(X)$  does exist, the *standard deviation*  $\sigma_X$  is defined as in the discrete case by

$$\sigma_X = \sqrt{\text{var}(X)}$$

**Example 11** Let  $X$  be a continuous random variable with the following pdf

$$f(x) = \begin{cases} \frac{1}{2}x & \text{if } 0 \leq x \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

Using calculus we can compute the expectation, variance, and standard deviation of  $X$ :

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x) dx \\ &= \int_0^2 \frac{1}{2}x^2 dx \\ &= \left. \frac{x^3}{6} \right|_0^2 = \frac{4}{3} \end{aligned}$$

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2f(x) dx \\ &= \int_0^2 \frac{1}{2}x^3 dx \\ &= \left. \frac{x^4}{8} \right|_0^2 = 2 \end{aligned}$$

$$\text{var}(X) = E(X^2) - \mu^2 = 2 - \frac{16}{9} = \frac{2}{9} \text{ and } \sigma_X = \sqrt{\frac{2}{9}}$$

## Joint Distribution of Random Variables

Earlier in this course we have seen that certain experiments can be analyzed in terms of compounds of simple experiments. Often, though, is not the resulting set of ordered pairs (or triples, etc.) which is of prime interest to the experimenter.

For example, sampling two ball bearings produced a set of ordered pairs as elementary events, but the experimenter may be interested only in the *number* of good ball bearings. Public opinion polls produce a sequence of responses, some favorable some unfavorable, but interest usually centers in the *proportion* of favorable responses rather than the ordered sequence of responses.

Many such situations can be viewed as a case of adding random variables. In mathematical terms, we are given a sample space  $\Omega$  and  $n$  random variables defined on  $\Omega$ , where  $n$  is an integer greater than or equal to 2. Lets look at the bivariate case ( $n = 2$ ):

**Example 12** A fair coin is tossed three independent times. We choose the familiar set

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

as sample space and assign probability  $\frac{1}{8}$  to each simple event. We define the following random variables:

$$X = \begin{cases} 0 & \text{if the first toss is a tail} \\ 1 & \text{if the first toss is a head,} \end{cases}$$

$Y =$  the total number of heads,

$Z =$  the absolute value of the difference between the number of heads and tails

We can list the values of these three random variables for each element of the sample space  $\Omega$ :

| Element of $\Omega$ | Value of $X$ | Value of $Y$ | Value of $Z$ |
|---------------------|--------------|--------------|--------------|
| <i>HHH</i>          | 1            | 3            | 3            |
| <i>HHT</i>          | 1            | 2            | 1            |
| <i>HTH</i>          | 1            | 2            | 1            |
| <i>THH</i>          | 0            | 2            | 1            |
| <i>HTT</i>          | 1            | 1            | 1            |
| <i>THT</i>          | 0            | 1            | 1            |
| <i>TTH</i>          | 0            | 1            | 1            |
| <i>TTT</i>          | 0            | 0            | 3            |

Consider now the first pair  $X, Y$ . We want to determine not only the possible pairs of values of  $X$  and  $Y$ , but also the probability with which each such pair occurs.

To say, for example, that  $X$  has the value 0 and  $Y$  the value 1 is to say that the event  $\{THT, TTH\}$  occurs. The probability of this event is therefore  $\frac{2}{8}$  or  $\frac{1}{4}$ . We write

$$P(X = 0, Y = 1) = \frac{1}{4},$$

adopting the convention in which a comma is used in place of  $\cap$  to denote the *intersection* of the two events  $X = 0$  and  $Y = 1$ . We similarly find

$$P(X = 0, Y = 0) = P(\{TTT\}) = \frac{1}{8},$$

$$P(X = 1, Y = 0) = P(\emptyset) = 0, \text{ etc.}$$

In this way, we obtain the probabilities of all possible pairs of values of  $X$  and  $Y$ . These probabilities can be arranged in the following table, the so-called *joint probability table* of  $X$  and  $Y$ .

|                  |               |               |               |               |               |
|------------------|---------------|---------------|---------------|---------------|---------------|
| $x \backslash y$ | 0             | 1             | 2             | 3             | $P(X = x)$    |
| 0                | $\frac{1}{8}$ | $\frac{1}{4}$ | $\frac{1}{8}$ | 0             | $\frac{1}{2}$ |
| 1                | 0             | $\frac{1}{8}$ | $\frac{1}{4}$ | $\frac{1}{8}$ | $\frac{1}{2}$ |
| $P(Y = y)$       | $\frac{1}{8}$ | $\frac{3}{8}$ | $\frac{3}{8}$ | $\frac{1}{8}$ | 1             |

Notice that the event  $Y = 0$  is the union of the mutually exclusive events  $(X = 0, Y = 0)$  and  $(X = 1, Y = 0)$ . Hence

$$P(Y = 0) = P(X = 0, Y = 0) + P(X = 1, Y = 0) = \frac{1}{8} + 0 = \frac{1}{8}.$$

In the table, this probability is obtained as the sum of the entries in the column headed  $y = 0$ . By adding the entries in the other columns, we similarly find

$$P(Y = 1) = \frac{3}{8}, P(Y = 2) = \frac{3}{8}, P(Y = 3) = \frac{1}{8}.$$

In this way, we obtain the probability function of the random variable  $Y$  from the joint probability table of  $X$  and  $Y$ . This function is commonly called the *marginal* probability function of  $Y$ . By adding across the rows in the joint table, one similarly obtains the (marginal) probability function of  $X$ .

Notice that knowing the value of  $X$  changes the probability that a given value of  $Y$  occurs. For example,  $P(Y = 2) = \frac{3}{8}$ . But if we are told that the value of  $X$  is 1, then the conditional probability of the event  $Y = 2$  becomes  $\frac{1}{2}$ . This follows from the definition of conditional probability:

$$P(Y = 2 | X = 1) = \frac{P(X = 1, Y = 2)}{P(X = 1)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}.$$

In other words, the events  $X = 1$  and  $Y = 2$  are *not* independent: knowing that the first toss results in a head *increases* the probability of obtaining exactly two heads in three tosses.

What we have done for the pair  $X, Y$  can also be done for  $X$  and  $Z$ . In this case, the joint probability table looks like this:

|                 |               |               |               |
|-----------------|---------------|---------------|---------------|
| $x \setminus z$ | 1             | 3             | $P(X = x)$    |
| 0               | $\frac{3}{8}$ | $\frac{1}{8}$ | $\frac{1}{2}$ |
| 1               | $\frac{3}{8}$ | $\frac{1}{8}$ | $\frac{1}{2}$ |
| $P(Z = z)$      | $\frac{3}{4}$ | $\frac{1}{4}$ | 1             |

Notice that the events  $X = 0$  and  $Z = 1$  are independent:

$P(X = 0, Z = 1) = \frac{3}{8}$ , and this is equal to the product of  $P(X = 0) = \frac{1}{2}$  and  $P(Z = 1) = \frac{3}{4}$ .

In general, two random variables  $X$  and  $Z$  are independent if each entry in the joint distribution table is the product of the marginal entries.

**Definition 4** Let  $X$  and  $Y$  be random variables on the same sample space  $\Omega$  with respective range spaces

$$R_X = \{x_1, x_2, \dots, x_n\} \text{ and } R_Y = \{y_1, y_2, \dots, y_m\}$$

The joint distribution or joint probability function of  $X$  and  $Y$  is the function  $h$  on the product space  $R_X \times R_Y$  defined by

$$h(x_i, y_j) \equiv P(X = x_i, Y = y_j) \equiv P(\{\omega \in \Omega : X(\omega) = x_i, Y(\omega) = y_j\})$$

The function  $h$  is usually given in the form of a table, and has the following properties:

- (i)  $h(x_i, y_j) \geq 0$ ,
- (ii)  $\sum_i \sum_j h(x_i, y_j) = 1$ .

Thus,  $h$  defines a probability space on the product space  $R_X \times R_Y$ .

### Mean of Sums of Random Variables

It follows from the preceding definition that if two random variables  $X$  and  $Y$  are defined on a sample space  $\Omega$ , then there are many other random variables also defined on  $\Omega$ .

Consider the random variables  $X$  and  $Y$  in our previous example. The possible values of  $X$  and  $Y$ , together with their joint probabilities were given in the first table.

Let  $z(x, y) = x + y$  so that  $U = z(X, Y) = X + Y$ . From the joint probability table, we can determine the possible values of  $U$  as well as the probability with which each value occurs. For example,

$$P(U = 2) = P(X = 0, Y = 2) + P(X = 1, Y = 1) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}.$$



In this way, we obtain the entries in the following probability table for the random variable  $U = X + Y$ :

|            |               |               |               |               |               |
|------------|---------------|---------------|---------------|---------------|---------------|
| $u$        | 0             | 1             | 2             | 3             | 4             |
| $P(U = u)$ | $\frac{1}{8}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{8}$ |

From this table we can calculate the mean of  $U$ :

$$E(U) = E(X + Y) = 0\left(\frac{1}{8}\right) + 1\left(\frac{1}{4}\right) + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{4}\right) + 4\left(\frac{1}{8}\right) = 2.$$

From the marginal probability functions of  $X$  and  $Y$ , we find that

$$E(X) = 0\left(\frac{1}{2}\right) + 1\left(\frac{1}{2}\right) = \frac{1}{2}, \quad E(Y) = 0\left(\frac{1}{8}\right) + 1\left(\frac{3}{8}\right) + 2\left(\frac{3}{8}\right) + 3\left(\frac{1}{8}\right) = \frac{3}{2}.$$

Observe that  $E(X + Y) = E(X) + E(Y)$ .

**Theorem 2** *Let  $X$  and  $Y$  be any random variables defined on a sample space  $\Omega$ . Then*

$$E(X + Y) = E(X) + E(Y)$$

*In words, the mean of the sum of two random variables is equal to the sum of their means.*

We can extend this result noting that for any constants  $a$  and  $b$ ,

$$E(aX + bY) = aE(X) + bE(Y).$$

Lets define now  $z(x, y)$  as the product rather than the sum of  $x$  and  $y$ .

Then,  $V = z(X, Y) = XY$  is a random variable and its following probability table is:

|            |               |               |               |               |
|------------|---------------|---------------|---------------|---------------|
| $v$        | 0             | 1             | 2             | 3             |
| $P(V = v)$ | $\frac{1}{2}$ | $\frac{1}{8}$ | $\frac{1}{4}$ | $\frac{1}{8}$ |

Now we compute the mean of  $V$ ,

$$E(V) = E(XY) = 0\left(\frac{1}{2}\right) + 1\left(\frac{1}{8}\right) + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{8}\right) = 1.$$

Observe that  $E(XY) \neq E(X)E(Y)$ .

**Theorem 3** *Let  $X$  and  $Y$  be independent random variables defined on a sample space  $\Omega$ . Then*

$$E(XY) = E(X)E(Y)$$

*In words, the mean of the product of two independent random variables is equal to the product of their means.*

But, what happens if  $X$  and  $Y$  are not independent?

**Example 13** Suppose  $X$  has probability table:

|            |               |               |               |
|------------|---------------|---------------|---------------|
| $x$        | $-1$          | $0$           | $1$           |
| $P(X = x)$ | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ |

Let  $Y = X^2$ . Then  $X$  and  $Y$  are dependent.

This dependence can be seen in the joint probability table:

|                  |               |               |               |
|------------------|---------------|---------------|---------------|
| $x \backslash y$ | $0$           | $1$           | $P(X = x)$    |
| $-1$             | $0$           | $\frac{1}{4}$ | $\frac{1}{4}$ |
| $0$              | $\frac{1}{2}$ | $0$           | $\frac{1}{2}$ |
| $1$              | $0$           | $\frac{1}{4}$ | $\frac{1}{4}$ |
| $P(Y = y)$       | $\frac{1}{2}$ | $\frac{1}{2}$ | $1$           |

Note that  $E(X) = 0$ ,  $E(Y) = \frac{1}{2}$ , and  $E(XY) = E(X^3) = 0$ , so that the previous theorem holds.

However, the theorem did not hold for the dependent random variables in the previous example.

We conclude that the preceding theorem holds for *all* pairs of *independent* random variables and *some* but *not all* pairs of *dependent* random variables.

### Variance of Sums of Random Variables

We turn now to some results leading to a formula for the variance of a sum of random variables. First, the following identity can be established:

$$\begin{aligned} E[(X - \mu_X)(Y - \mu_Y)] &= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y. \end{aligned}$$

Except for the sign the last three terms are equal. Hence,

$$E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y$$

Notice that if  $X$  and  $Y$  are independent, then

$$E[(X - \mu_X)(Y - \mu_Y)] = 0$$

**Theorem 4** *Let  $X$  and  $Y$  be independent random variables defined on a sample space  $\Omega$ . Then:*

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y).$$

*In words, the variance of the sum of two independent random variables is equal to the sum of their variances.*

**Proof.** By definition of variance we have

$$\begin{aligned} \text{var}(X + Y) &= E([(X + Y) - E(X + Y)]^2) \\ &= E([(X - \mu_X) + (Y - \mu_Y)]^2), \end{aligned}$$

where we have rearranged terms in the bracket using Theorem 2. Now we perform the indicated squaring operation to obtain

$$\begin{aligned} \text{var}(X + Y) &= E[(X - \mu_X)^2 + 2(X - \mu_X)(Y - \mu_Y) + (Y - \mu_Y)^2] \\ &= E[(X - \mu_X)^2] + 2E[(X - \mu_X)(Y - \mu_Y)] + E[(Y - \mu_Y)^2], \end{aligned}$$

Note that if  $X$  and  $Y$  are independent, the middle term on the right hand side vanishes. The other two terms are, by definition, precisely  $\text{var}(X)$  and  $\text{var}(Y)$ . Q.E.D.

Now if  $X$  and  $Y$  are independent, then so are  $aX$  and  $bY$  for any constants  $a$  and  $b$ . Thus, we can extend the previous result to  $aX$  and  $bY$ :

$$\text{var}(aX + bY) = \text{var}(aX) + \text{var}(bY).$$

### Covariance and Correlation

Let  $X$  and  $Y$  be random variables with the joint distribution  $h(x, y)$ , and suppose now that we want to measure how the possible values of  $X$  are related to the possible values of  $Y$ .

In our last theorem we showed that

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2E[(X - \mu_X)(Y - \mu_Y)],$$

and since  $X$  and  $Y$  were assumed to be independent, we concluded that the last term vanishes. However, this last expression should be studied more carefully. In particular, we are now going to pay attention to the last term in the preceding expression.

The *covariance* of  $X$  and  $Y$ , denoted by  $cov(X, Y)$ , is defined by

$$cov(X, Y) = \sum_{i,j} (x_i - \mu_X)(y_j - \mu_Y) h(x_i, y_j) = E[(X - \mu_X)(Y - \mu_Y)]$$

or equivalently,

$$cov(X, Y) = \sum_{i,j} x_i y_j h(x_i, y_j) - \mu_X \mu_Y = E(XY) - \mu_X \mu_Y$$

Using this notation, we can express the variance of the sum of the two random variables  $X$  and  $Y$  as

$$var(X + Y) = var(X) + var(Y) + 2cov(X, Y)$$

Notice that we treat  $X$  and  $Y$  symmetrically, i.e.

$$cov(X, Y) = cov(Y, X),$$

and since  $X - \mu_X$  and  $Y - \mu_Y$  each have mean zero,

$$cov(X - \mu_X, Y - \mu_Y) = cov(X, Y),$$

The covariance is thus a measure of the extent to which the values of  $X$  and  $Y$  tend to increase or decrease together. If  $X$  has values greater than its mean  $\mu_X$  whenever  $Y$  has values greater than its mean  $\mu_Y$  and  $X$  has values less than  $\mu_X$  whenever  $Y$  has values less than  $\mu_Y$ , then  $(X - \mu_X)(Y - \mu_Y)$  has positive values and  $cov(X, Y) > 0$ .

On the other hand, if values of  $X$  are above  $\mu_X$  whenever values of  $Y$  are below  $\mu_Y$  and vice versa, then  $cov(X, Y) < 0$ .

By a suitable choice of two random variables, we can make their covariance any number we like. For example, if  $a$  and  $b$  are constants, then

$$\begin{aligned} cov(aX, bY) &= E(aXbY) - E(aX)E(bY) \\ &= abE(XY) - (a\mu_X)(b\mu_Y), \end{aligned}$$

from which follows that

$$cov(aX, bY) = ab cov(X, Y).$$

It should be clear from this last equation that if  $cov(X, Y) \neq 0$ , then by varying  $a$  and  $b$  we can make  $cov(aX, bY)$  positive or negative, as small or as large as we please.

It is more convenient to have a measure of the relation that cannot vary so widely. The standardized random variable  $X^*$  is defined by

$$X^* = \frac{X - \mu_X}{\sigma_X}$$

Similarly, the standardized random variable  $Y^*$  is defined by

$$Y^* = \frac{Y - \mu_Y}{\sigma_Y}$$

Thus,

$$\begin{aligned} \text{cov}(X^*, Y^*) &= \text{cov}\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right) \\ &= \frac{1}{\sigma_X \sigma_Y} \text{cov}(X - \mu_X, Y - \mu_Y) \\ &= \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}, \end{aligned}$$

this last equality follows from the definition of covariance (see above).

The *correlation* of  $X$  and  $Y$ , denoted by  $\rho(X, Y)$ , is defined by

$$\rho(X, Y) = \text{cov}(X^*, Y^*) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

If  $\sigma_X = 0$  or if  $\sigma_Y = 0$ , we define  $\rho(X, Y) = 0$ . The random variables  $X$  and  $Y$  are said to be *uncorrelated* if and only if  $\rho(X, Y) = 0$ ; otherwise they are said to be *correlated*.

If  $\sigma_X > 0$  and  $\sigma_Y > 0$ , then  $\rho(X, Y) = 0$  if and only if  $\text{cov}(X, Y) = 0$ .

Note that if  $X$  and  $Y$  are independent random variables, then they are uncorrelated. But the opposite is not true (i.e. we can find two random variable that are uncorrelated but not independent).

Finally, it is worth noting the following properties of  $\rho$ :

- (i)  $\rho(X, Y) = \rho(Y, X)$ ,
- (ii)  $-1 \leq \rho \leq 1$ ,
- (iii)  $\rho(X, X) = 1$ ,  $\rho(X, -X) = -1$

**Example 14** A pair of dice is tossed. The sample space  $\Omega$  consists of the 36 ordered pairs  $(a, b)$  where  $a$  and  $b$  can be any integers between 1 and 6. Let  $X$  assign to each point  $(a, b)$  the maximum of its numbers, that is,  $X(a, b) = \max(a, b)$ . Now let  $Y$  assign to each point  $(a, b)$  the sum of its numbers, that is,  $Y(a, b) = a + b$ .

So, for example  $X(1, 1) = 1$ ,  $X(3, 4) = 4$ ,  $X(5, 2) = 5$ ,  $X(6, 6) = 6$ ; and in the case of the random variable  $Y$ ,  $Y(1, 1) = 2$ ,  $Y(3, 4) = 7$ ,  $Y(6, 3) = 9$ ,  $Y(6, 6) = 12$ .

Then  $X$  is a random variable where any number between 1 and 6 could occur, and no other number can occur. Thus, the range space  $R_X$  of  $X$  is as follows:

$$R_X = \{1, 2, 3, 4, 5, 6\}$$

And,  $Y$  is a random variable where any number between 2 and 12 could occur, and no other number can occur. Thus, the range space  $R_Y$  of  $Y$  is as follows:

$$R_Y = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

The joint distribution appears in the following table:

| $x \ y$    | 2              | 3              | 4              | 5              | 6              | 7              | 8              | 9              | 10             | 11             | 12             | $P(X = x)$      |
|------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|-----------------|
| 1          | $\frac{1}{36}$ | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | $\frac{1}{36}$  |
| 2          | 0              | $\frac{2}{36}$ | $\frac{1}{36}$ | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | $\frac{3}{36}$  |
| 3          | 0              | 0              | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ | 0              | 0              | 0              | 0              | 0              | 0              | $\frac{5}{36}$  |
| 4          | 0              | 0              | 0              | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ | 0              | 0              | 0              | 0              | $\frac{7}{36}$  |
| 5          | 0              | 0              | 0              | 0              | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ | 0              | 0              | $\frac{9}{36}$  |
| 6          | 0              | 0              | 0              | 0              | 0              | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ | $\frac{11}{36}$ |
| $P(Y = y)$ | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |                 |

The entry  $h(3, 5) = \frac{2}{36}$  comes from the fact that  $(3, 2)$  and  $(2, 3)$  are the only points in  $\Omega$  whose maximum number is 3 and whose sum is 5, that is,

$$h(3, 5) \equiv P(X = 3, Y = 5) = P\{(3, 2), (2, 3)\} = \frac{2}{36}$$

The other entries are obtained in a similar manner.

Notice first that the right side column gives the distribution  $f$  of  $X$ , and the bottom row gives the distribution  $g$  of  $Y$ .

Now we are going to compute the covariance and correlation of  $X$  and  $Y$ . First we compute the expectation of  $X$  and  $Y$  as follows:

$$E(X) = 1\left(\frac{1}{36}\right) + 2\left(\frac{3}{36}\right) + 3\left(\frac{5}{36}\right) + 4\left(\frac{7}{36}\right) + 5\left(\frac{9}{36}\right) + 6\left(\frac{11}{36}\right) = \frac{161}{36} \approx 4.47$$

$$E(Y) = 2\left(\frac{1}{36}\right) + 3\left(\frac{2}{36}\right) + 4\left(\frac{3}{36}\right) + \dots + 12\left(\frac{1}{36}\right) = \frac{252}{36} = 7$$

Next, we compute  $\sigma_X$  and  $\sigma_Y$  as follows:

$$E(X^2) = \sum x_i^2 f(x_i)$$

$$= 1^2\left(\frac{1}{36}\right) + 2^2\left(\frac{3}{36}\right) + 3^2\left(\frac{5}{36}\right) + 4^2\left(\frac{7}{36}\right) + 5^2\left(\frac{9}{36}\right) + 6^2\left(\frac{11}{36}\right) = \frac{791}{36} \approx 21.97$$

Hence

$$\text{var}(X) = E(X^2) - \mu_X^2 = 21.97 - 19.98 = 1.99 \text{ and } \sigma_X = \sqrt{1.99} \approx 1.4$$

Similarly,

$$E(Y^2) = \sum y_i^2 g(y_i)$$

$$= 2^2\left(\frac{1}{36}\right) + 3^2\left(\frac{2}{36}\right) + 4^2\left(\frac{3}{36}\right) + \dots + 12^2\left(\frac{1}{36}\right) = \frac{1974}{36} \approx 54.8$$

Hence

$$\text{var}(Y) = E(Y^2) - \mu_Y^2 = 54.8 - 49 = 5.8 \text{ and } \sigma_Y = \sqrt{5.8} \approx 2.4$$

Now we compute  $E(XY)$  as follows:

$$E(XY) = \sum x_i y_j h(x_i, y_j)$$

$$= 1(2)\left(\frac{1}{36}\right) + 2(3)\left(\frac{2}{36}\right) + 2(4)\left(\frac{1}{36}\right) + \dots + 6(12)\left(\frac{1}{36}\right) = \frac{1232}{36} \approx 34.2$$

So, the covariance of  $X$  and  $Y$  is computed as:

$$\text{cov}(X, Y) = E(XY) - \mu_X \mu_Y = 34.2 - (4.47)(7) \approx 2.9$$

and

$$\begin{aligned} \rho(X, Y) &= \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \\ &= \frac{2.9}{(1.4)(2.4)} \approx 0.86. \end{aligned}$$

## Chebyshev's Inequality and Law of Large Numbers

As we just learned, the standard deviation  $\sigma$  of a random variable  $X$  measures the spread of the values about the mean  $\mu$  of  $X$ . Accordingly, for smaller values of  $\sigma$ , we should expect that  $X$  will be closer to its mean  $\mu$ .

This intuitive expectation is made more precise by the following inequality, named after the Russian mathematician P.L. Chebyshev (1921-1994):

**Theorem 5 (Chebyshev's Inequality):** *Let  $X$  be a random variable with mean  $\mu$  and standard deviation  $\sigma$ . Then, for any positive number  $k$ , the probability that the value of  $X$  lies in the interval  $[\mu - k\sigma, \mu + k\sigma]$  is at least  $1 - \frac{1}{k^2}$ . That is,*

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

**Proof.** Note first that

$$P(|X - \mu| > k\sigma) = 1 - P(|X - \mu| \leq k\sigma) = 1 - P(\mu - k\sigma \leq X \leq \mu + k\sigma)$$

By definition

$$\sigma^2 = \text{var}(X) = \sum (x_i - \mu)^2 f(x_i)$$

Delete all terms from the summation for which  $x_i$  is in the interval  $[\mu - k\sigma, \mu + k\sigma]$ , that is, delete all terms for which  $|x_i - \mu| \leq k\sigma$ . Denote the summation of the remaining terms by  $\sum^*(x_i - \mu)^2 f(x_i)$ . Then

$$\begin{aligned} \sigma^2 &\geq \sum^*(x_i - \mu)^2 f(x_i) \geq \sum^* k^2 \sigma^2 f(x_i) = k^2 \sigma^2 \sum^* f(x_i) \\ &= k^2 \sigma^2 P(|X - \mu| > k\sigma) \\ &= k^2 \sigma^2 [1 - P(\mu - k\sigma \leq X \leq \mu + k\sigma)] \end{aligned}$$

If  $\sigma > 0$ , then dividing by  $k^2 \sigma^2$  gives

$$\frac{1}{k^2} \geq 1 - P(\mu - k\sigma \leq X \leq \mu + k\sigma)$$

or

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

which proves Chebyshev's inequality for  $\sigma > 0$ . If  $\sigma = 0$ , then  $x_i = \mu$  for all  $f(x_i) > 0$  and

$$P(\mu - k \times 0 \leq X \leq \mu + k \times 0) = P(X = \mu) = 1 > 1 - \frac{1}{k^2}$$

which completes the proof.



**Example 15** Suppose  $X$  is a random variable with mean  $\mu = 100$  and standard deviation  $\sigma = 5$ . Let  $k = 2$ .

Setting  $k = 2$  we get

$$\begin{aligned}\mu - k\sigma &= 100 - 2(5) = 90, \\ \mu + k\sigma &= 100 + 2(5) = 110, \\ 1 - \frac{1}{k^2} &= 1 - \frac{1}{4} = \frac{3}{4}.\end{aligned}$$

Thus, from Chebyshev's inequality we can conclude that the probability that  $X$  lies between 90 and 110 is at least  $\frac{3}{4}$ .

**Example 16** Suppose  $X$  is a random variable with mean  $\mu = 100$  and standard deviation  $\sigma = 5$ . What is the probability that  $X$  lies between 80 and 120?

Here  $k\sigma = 20$ , and since  $\sigma = 5$ , we get  $5k = 20$ , or  $k = 4$ . Thus, by Chebyshev's inequality:

$$P(80 \leq X \leq 120) \geq 1 - \frac{1}{k^2} = 1 - \frac{1}{4^2} = \frac{15}{16} \approx 0.94.$$

**Example 17** Suppose  $X$  is a random variable with mean  $\mu = 100$  and standard deviation  $\sigma = 5$ . Find an interval  $[a, b]$  about the mean  $\mu = 100$  for which the probability that  $X$  lies in the interval is at least 99 percent.

Here we set  $1 - \frac{1}{k^2} = 0.99$  and solve for  $k$ . This yields

$$1 - 0.99 = \frac{1}{k^2} \text{ or } 0.01 = \frac{1}{k^2} \text{ or } k^2 = \frac{1}{0.01} = 100 \text{ or } k = 10.$$

Thus, the desired interval is

$$[a, b] = [\mu - k\sigma, \mu + k\sigma] = [100 - 10(5), 100 + 10(5)] = [50, 150].$$

### Law of Large Numbers

One winter night during one of the many German air raids on Moscow in World War II, a distinguished Soviet professor of statistics showed up in his local air-raid shelter. He had never appeared before. "There are seven million people in Moscow," he used to say. "Why should I expect them to hit me?" His friends were astonished to see him and asked what had happened to change his mind. "Look," he explained, "there are seven million people in Moscow and one elephant. Last night they got the elephant."

This story illuminates the dual character that runs throughout everything that has to do with probability: past frequencies can collide with degrees of belief when risky choices must be made. In this case, the statistics professor was keenly aware of the mathematical probability of being hit by a bomb. However, after one elephant was killed by a Nazi bomb, he decided that time had to come to go to the air-raid shelter.

Real-life situations, like the one described by this anecdote, often require us to measure probability in precisely this fashion – from sample to universe. In only rare cases does life replicate games of chance, for which we can determine the probability of an outcome *before* an event even occurs. In most instances, we have to estimate probabilities from what happened *after* the fact – *a posteriori*. The very notion of *a posteriori* implies experimentation and changing degrees of belief.

So how do we develop probabilities from limited amounts of real-life information? The answer to this question was one of Jacob Bernoulli's contribution to probability theory. His theorem for calculating probabilities *a posteriori* is known as the Law of Large Numbers.

Let  $X$  be the random variable and  $n$  the number of independent trials corresponding to some experiment. We may view the numerical value of each particular trial to be a random variable with the same mean as  $X$ .

Specifically, we let  $X_k$  denote the outcome of the  $k^{\text{th}}$  trial where  $k = 1, 2, \dots, n$ . The average value of all  $n$  outcomes is also a random variable, denoted by  $\bar{X}_n$  and called the *sample mean*. That is,

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$$

The law of large numbers says that as  $n$  increases, the probability that the value of sample mean  $\bar{X}_n$  is close to  $\mu$  approaches 1.

**Example 18** Suppose a fair die is tossed 8 times with the following outcomes:  $x_1 = 2, x_2 = 5, x_3 = 4, x_4 = 1, x_5 = 4, x_6 = 6, x_7 = 3, x_8 = 2$ .

We calculate the sample mean  $\bar{X}_8$  as follows:

$$\bar{X}_8 = \frac{2 + 5 + 4 + 1 + 4 + 6 + 3 + 2}{8} = \frac{27}{8} = 3.375$$

For a fair die, the mean  $\mu = 3.5$ . The law of large numbers tells us that as  $n$  gets larger, the probability that the sample mean  $\bar{X}_n$  will get close to 3.5 becomes larger, and, in fact, approaches one.

Now, contrary to popular view this law does not provide a method for validating observed facts, which are only an incomplete representation of the whole truth.

Nor does it say that an increasing number of observations will increase the probability that what you see is what you are going to get.

Suppose we toss a coin over and over. The law of large numbers does not tell us that the average of our throws will approach 50% as we increase the number of throws. Rather, the law states that increasing the number of throws will increase the probability that the ratio of heads thrown to total throws will vary from 50% by less than some stated amount, no matter how small.

The word “vary” is what matters. The search is not for the true mean of 50% but for the probability that the difference between the observed average and the true average will be less than, say 2%.

Namely, all the law tells us is that the average of a large number of throws will be more likely than the average of a small number of throws to differ from the true average by less than some stated amount. Let’s examine this statement formally:

**Theorem 6 (*Law of Large Numbers*):** *For any positive number  $\alpha$ , no matter how small,*

$$P(\mu - \alpha \leq \bar{X}_n \leq \mu + \alpha) \rightarrow 1 \text{ as } n \rightarrow \infty$$

In words, the probability that the sample mean has a value in the interval  $[\mu - \alpha, \mu + \alpha]$  approaches 1 as  $n$  approaches infinity.

**Proof.** Note first the following. Let  $n$  be any positive integer and let  $X_1, X_2, \dots, X_n$  be  $n$  independent, identically distributed random variables, each with mean  $\mu_X$  and variance  $\sigma_X^2$ .

If

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n},$$

then

$$\mu_{\bar{X}} = \mu_X \text{ and } \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}.$$

Now we apply Chebyshev’s inequality to the random variable  $\bar{X}$  and find that

$$P(|\bar{X} - \mu_{\bar{X}}| > \alpha) < \frac{\sigma_{\bar{X}}^2}{\alpha^2}$$

Then

$$P(|\bar{X} - \mu_X| > \alpha) < \frac{\sigma_X^2}{n\alpha^2}$$

or

$$P(|\bar{X} - \mu_X| \leq \alpha) > 1 - \frac{\sigma_X^2}{n\alpha^2}$$

Note that as  $n$  increases, the quantity  $\frac{\sigma_X^2}{n\alpha^2}$  decreases and approaches zero. Hence  $1 - \frac{\sigma_X^2}{n\alpha^2}$  approaches 1 as  $n$  gets larger and larger, and so  $P(|\bar{X} - \mu_X| \leq \alpha)$ , can be made as close as 1 as we like by choosing  $n$  sufficiently large. This completes the proof.

## Probability Distributions: Binomial and Normal Distributions

Earlier today we defined a random variable  $X$  on a probability space  $\Omega$  and its probability distribution  $f$ . You possibly noticed that one can discuss  $X$  and  $f(x)$  without referring to the original probability space  $\Omega$ . In fact, there are many applications of probability theory which give rise to the same probability distribution (i.e. infinitely many different random variables can have the same probability function).

Also given that certain kind of experiments and associated random variables occur time and again in the theory of probability, they are made the object of special study. The properties of these probability distributions are explored, values of frequently needed probabilities are tabulated, and so on.

Now we will discuss two such important distributions in probability – the binomial distribution and the normal distribution. In addition, we will also briefly discuss other distributions, including the uniform and Poisson distributions.

Finally, I will also try to indicate how each distribution might be an appropriate probability model for some applications. Keep in mind, though, that while some experimental situations naturally give rise to specific probability distributions, in the majority of cases in the social sciences the distributions used are merely models of the observed phenomena.

### Binomial Distribution

A number of times in this class we looked at experiments made up of a number, say  $n$ , of individual trials. Each trial was in itself an arbitrary experiment, and therefore, we defined it mathematically by some sample space and assignment of probabilities to its sample points.

Although each trial may have many possible outcomes, we may be interested only in whether a certain result occurs or not. For example, a card is selected from a standard deck and it is an ace or not an ace; two dice are rolled and the sum of the numbers showing is seven or is different from seven.

The convention in these cases is to call one of the two possible results a *success* ( $S$ ) and the other a *failure* ( $F$ ). It is also convenient to make the sample space defining the trial represent this fact by containing just two elements:  $\{S, F\}$ .

Consider an experiment  $\varepsilon$  with only two outcomes,  $\{S, F\}$ . Let  $p$  denote the probability of success in such experiment and let  $q = 1 - p$  denote the probability of failure. Then, given an acceptable assignment of probabilities,  $p + q = 1$ .

Suppose the experiment  $\varepsilon$  is repeated and suppose the trials are independent, that is, suppose the outcome of any trial does not depend on any previous outcomes, such as tossing a coin. Such independent repeated trials of an experiment with two outcomes are called *Bernoulli trials*, named after our good friend, the Swiss mathematician Jacob Bernoulli (1654-1705) <sup>1</sup>.

**Example 19** *Let the experiment  $\varepsilon$  be made up of three Bernoulli trials with probability  $p$  for success on each trial.*

The sample space for the experiment  $\varepsilon$  is the Cartesian product set  $\{S, F\} \times \{S, F\} \times \{S, F\}$  containing  $2^3 = 8$  tree-tuples as elements. Notice that since the trials are independent, the probabilities of the simple events corresponding to these three-tuples can be obtained using the product rule.

Denote by  $S_3$  the random variable indicating the number of successes in the experiment  $\varepsilon$ . The possible values for  $S_3$  are  $k = 0, 1, 2, 3$ . Then  $P(S_3 = k)$  is the probability function of the random variable  $S_3$ .

The following table summarizes this information:

| Outcome of $\varepsilon$ | Corresponding Probability | $S_3 = k$ | $P(S_3 = k)$ |
|--------------------------|---------------------------|-----------|--------------|
| FFF                      | $qqq = q^3$               | 0         | $q^3$        |
| FFS                      | $qqp = pq^2$              | 1         | $3pq^2$      |
| FSF                      | $qpq = pq^2$              |           |              |
| SFF                      | $pqq = pq^2$              |           |              |
| FSS                      | $qpp = p^2q$              | 2         | $3p^2q$      |
| SFS                      | $pqp = p^2q$              |           |              |
| SSF                      | $ppq = p^2q$              |           |              |
| SSS                      | $ppp = p^3$               | 3         | $p^3$        |

Notice that the probabilities in the last column are the terms in the binomial expansion <sup>2</sup> of  $(q+p)^3$ . Since  $p+q = 1$ , it follows that the sum of these probabilities, as expected, is indeed 1.

<sup>1</sup>Daniel Bernoulli, as in “expected utility” was Jacob’s nephew.

<sup>2</sup>If  $a$  and  $b$  are two numbers, powers of their sum such as  $(a+b)^2$ ,  $(a+b)^3$ , ... are computed by multiplying

A *binomial experiment* consists of a fixed number, say  $n$ , of Bernoulli trials. (The use of the term “binomial” will soon be apparent.) Such a binomial experiment will be denoted by

$$B(n, p)$$

That is,  $B(n, p)$  denotes a binomial experiment with  $n$  trials and probability  $p$  of success. The sample space of the  $n$  repeated trials consist of all  $n$ -tuples (that is,  $n$ -element sequences) whose components are either  $S$  or  $F$ .

In general, we are interested in the probability of a certain number of successes in a binomial experiment and not necessarily in the order in which they occur.

Let  $A$  be the event of exactly  $k$  successes. Then  $A$  consists of all  $n$ -tuples of which  $k$  components are  $S$  and  $n - k$  components are  $F$ .

The number of such  $n$ -tuples in the event  $A$  is equal to the number of ways that  $k$  letters  $S$  can be distributed among the  $n$  components of an  $n$ -tuple. Therefore  $A$  consists of  $C(n, k) = \binom{n}{k}$  sample points, where  $\binom{n}{k}$  is the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

(recall that the symbol  $\binom{n}{k}$  reads as “ $n$  choose  $k$ ,” and that sometimes is presented as  $C(n, k)$ ).

Each point in  $A$  has the same probability, namely  $p^k q^{n-k}$ ; hence

$$P(A) = \binom{n}{k} p^k q^{n-k}.$$

Notice also that the probability of no success is

$$P(0) = \binom{n}{0} p^0 q^n,$$

Thus, the probability of one or more successes is  $1 - q^n$ .

We have proved the following result:

---

and then combining similar terms:

$$(a + b)^2 = a^2 + 2ab + b^2$$

$$(a + b)^3 = (a + b)(a + b)^2 = (a + b)(a^2 + 2ab + b^2) = a^3 + 3a^2b + 3ab^2 + b^3$$

and so on. The formula telling us how to multiply out an arbitrary power

$$(a + b)^n = a^n + na^{n-1}b + \frac{n(n-1)}{2}a^{n-2}b^2 + \dots + b^n$$

is well known in algebra as the “binomial formula”.

**Theorem 7** *The probability of exactly  $k$  success in a binomial experiment  $B(n, p)$  is given by:*

$$P(k) = P(k \text{ successes}) = \binom{n}{k} p^k q^{n-k}.$$

*The probability of one or more successes is  $1 - q^n$ .*

Observe that the probability of getting at least  $k$  successes, that is,  $k$  or more successes is given by

$$P(k) + P(k + 1) + P(k + 2) + \dots P(n)$$

This follows from the fact that the events of getting  $k$  and  $k'$  successes are disjoint for  $k \neq k'$ .

For given values of  $n$  and  $p$ , the probability function defined by the formula in Theorem 3 is called the *binomial probability function* or the *binomial distribution* with parameters  $n$  and  $p$ .

**Example 20** *The probability that a marksman hits a target at any time is  $p = \frac{1}{3}$ , hence he misses with probability  $q = 1 - p = \frac{2}{3}$ . Suppose he fires at a target 7 times. What is the probability that he hits the target exactly 3 times?*

This is a binomial experiment with  $n = 7$  and  $p = \frac{1}{3}$ . We are interested in  $k = 3$ . By theorem 3, the probability that he hits the target exactly 3 times is

$$\begin{aligned} P(3) &= \binom{7}{3} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^4 \\ &= \frac{7!}{3!(4)!} \left(\frac{1}{27}\right) \left(\frac{16}{81}\right) \\ &= \frac{5040}{(6)(24)} \left(\frac{16}{2187}\right) \\ &= \frac{560}{2187} \approx 0.26 \end{aligned}$$

**Example 21** *Suppose we are looking at the same experiment as the one in the previous example. What is the probability that the marksman hits the target at least 1 time?*

The probability that he never hits the target, that is, all failures is:

$$\begin{aligned} P(0) &= q^7 = \left(\frac{2}{3}\right)^7 \\ &= \frac{128}{2187} \approx 0.06 \end{aligned}$$

Thus, the probability that he hits the target at least once is

$$1 - q^7 = \frac{2059}{2187} \approx 0.94.$$

**Example 22** A fair coin is tossed 6 times; call heads a success. What is the probability that exactly 2 heads occur?

This is a binomial experiment with  $n = 6$  and  $p = q = \frac{1}{2}$ . We are interested in  $k = 2$ . By theorem 3, the probability that exactly 2 heads occur is

$$\begin{aligned} P(2) &= \binom{6}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^4 \\ &= \frac{15}{64} \approx 0.23 \end{aligned}$$

**Example 23** Suppose we are looking at the same experiment as the one in the previous example. What is the probability that at least 4 heads occur?

Now we want to calculate the probability of getting at least 4 heads, that is,  $k = 4$ ,  $k = 5$ , or  $k = 6$ . Hence,

$$\begin{aligned} P(4) + P(5) + P(6) &= \binom{6}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^2 + \binom{6}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right) + \binom{6}{6} \left(\frac{1}{2}\right)^6 \\ &= \frac{15}{64} + \frac{6}{64} + \frac{1}{64} \\ &= \frac{22}{64} \approx 0.34 \end{aligned}$$

As the examples show, the formula in Theorem 3 defines not just one binomial distribution, but a whole family of binomial distributions, one for every possible pair of values for  $n$  and  $p$ .

**Definition 5** Consider a binomial experiment  $B(n, p)$ . That is,  $B(n, p)$  consists of  $n$  independent repeated trials with two outcomes, success or failure, and  $p$  is the probability of success and  $q = 1 - p$  is the probability of failure. Let  $X$  denote the number of successes in such experiment. Then  $X$  is a random variable with the following distribution:

$$\begin{array}{c|cccccc} k & 0 & 1 & 2 & \dots & n \\ \hline P(k) & q^n & \binom{n}{1}q^{n-1}p & \binom{n}{2}q^{n-2}p^2 & \dots & p^n \end{array}$$

**Example 24** Suppose a fair coin is tossed 6 times and heads is call a success. This is a binomial experiment with  $n = 6$  and  $p = q = \frac{1}{2}$ . What is the binomial distribution  $B(6, \frac{1}{2})$ ?

We already know that  $P(2) = \frac{15}{64}$ ,  $P(4) = \frac{15}{64}$ ,  $P(5) = \frac{6}{64}$ , and  $P(6) = \frac{1}{64}$ . Using the formula in Theorem 3 we can also calculate  $P(0) = \frac{1}{64}$ ,  $P(1) = \frac{6}{64}$ , and  $P(3) = \frac{20}{64}$ . Thus, the binomial distribution  $B(6, \frac{1}{2})$  follows:

$$\begin{array}{c|cccccc} k & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline P(k) & \frac{1}{64} & \frac{6}{64} & \frac{15}{64} & \frac{20}{64} & \frac{15}{64} & \frac{6}{64} & \frac{1}{64} \end{array}$$



Calculating the probabilities of particular events of binomial experiments can be tedious. In particular, sometimes we want to compute the probability not of exactly  $k$  successes, but *at least*  $k$  or *at most*  $k$  successes (as in example 23). Since such cumulative probabilities are obtained by computing all the included individual probabilities and adding, this task soon becomes laborious.

As I mentioned a few times in this class, mathematicians are lazy people. So some of them have developed extensive tables to lighten the task of such computations.

| $n$ | $r$ | $p = .01$ | $p = .05$ | $p = .10$ | $p = .20$ | $p = .30$ | $p = .40$ | $p = .50$ |
|-----|-----|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 1   | 1   | .010      | .050      | .100      | .200      | .300      | .400      | .500      |
| 2   | 1   | .020      | .098      | .190      | .360      | .510      | .640      | .750      |
|     | 2   |           | .002      | .010      | .040      | .090      | .160      | .250      |
| 3   | 1   | .030      | .143      | .271      | .488      | .657      | .784      | .875      |
|     | 2   |           | .007      | .028      | .104      | .216      | .352      | .500      |
|     | 3   |           |           | .001      | .008      | .027      | .064      | .125      |
| 4   | 1   | .039      | .185      | .344      | .590      | .760      | .870      | .938      |
|     | 2   | .001      | .014      | .052      | .181      | .348      | .525      | .688      |
|     | 3   |           |           | .004      | .027      | .084      | .179      | .312      |
|     | 4   |           |           |           | .002      | .008      | .026      | .062      |
| 5   | 1   | .049      | .226      | .410      | .672      | .832      | .922      | .969      |
|     | 2   | .001      | .023      | .081      | .263      | .472      | .663      | .812      |
|     | 3   |           | .001      | .009      | .058      | .163      | .317      | .500      |
|     | 4   |           |           |           | .007      | .031      | .087      | .188      |
|     | 5   |           |           |           |           | .002      | .010      | .031      |
| 6   | 1   | .059      | .265      | .469      | .738      | .882      | .953      | .984      |
|     | 2   | .001      | .033      | .114      | .345      | .580      | .767      | .891      |
|     | 3   |           | .002      | .016      | .099      | .256      | .456      | .656      |
|     | 4   |           |           | .001      | .017      | .070      | .179      | .344      |
|     | 5   |           |           |           | .002      | .011      | .041      | .109      |
|     | 6   |           |           |           |           | .001      | .004      | .016      |
| 7   | 1   | .068      | .302      | .522      | .790      | .918      | .972      | .992      |
|     | 2   | .002      | .044      | .150      | .423      | .671      | .841      | .938      |
|     | 3   |           | .004      | .026      | .148      | .353      | .580      | .773      |
|     | 4   |           |           | .003      | .033      | .126      | .290      | .500      |
|     | 5   |           |           |           | .005      | .029      | .096      | .227      |
|     | 6   |           |           |           |           | .004      | .019      | .062      |
|     | 7   |           |           |           |           |           | .002      | .008      |
| 8   | 1   | .077      | .337      | .570      | .832      | .942      | .983      | .996      |
|     | 2   | .003      | .057      | .187      | .497      | .745      | .894      | .965      |
|     | 3   |           | .006      | .038      | .203      | .448      | .685      | .855      |
|     | 4   |           |           | .005      | .056      | .194      | .406      | .637      |
|     | 5   |           |           |           | .010      | .058      | .174      | .363      |
|     | 6   |           |           |           | .001      | .011      | .050      | .145      |
|     | 7   |           |           |           |           | .001      | .009      | .035      |
|     | 8   |           |           |           |           |           | .001      | .004      |

In example 24 we found that the probability of getting 6 heads if a fair coin is tossed 6 times is  $P(k = 6) = \frac{1}{64}$ . Look at the last column and the entry for  $n = 6$  and  $k = 6$  (actually  $r$  in this table) with probability  $p = .5$ . You will read 0.016, which agrees with our calculated probability.

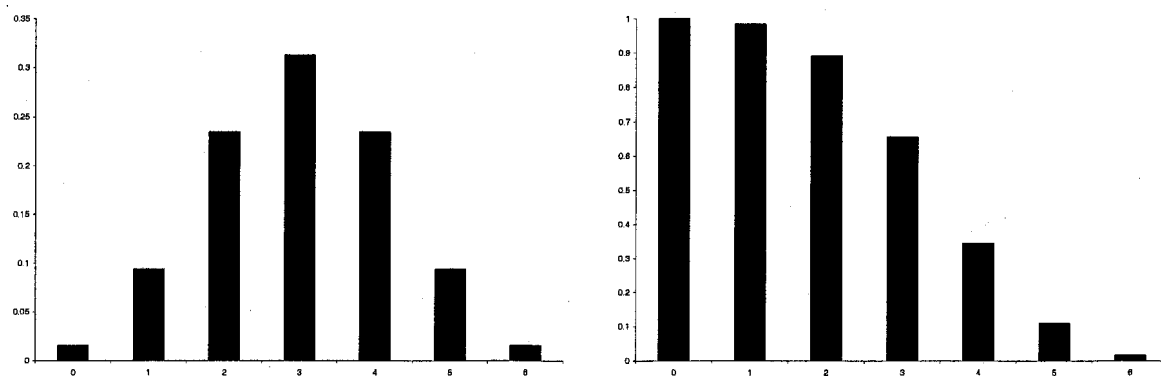
In the example 23 we found that the probability of getting exactly 2 heads if a fair coin is tossed 6 times is  $P(k = 2) = \frac{15}{64}$ . To find  $P(k = 2)$  in the table first note that

$$P(2) = P(k \geq 2) - P(k \geq 3),$$

These cumulative probabilities can be read directly from the table, so we find

$$P(2) = .891 - .656 = 0.235$$

which is similar to the answer computed in example 22.



$$P(X = k) \text{ and } P(X \geq k).$$

Lets look now at some properties of the binomial distribution:

Let  $X$  be the binomial random variable  $B(n, p)$ . We can use the definitions of mean and variance we learned last week to compute  $E(X)$  and  $var(X)$ . That is, we have to evaluate the sums

$$E(X) = \sum_{k=0}^n k \binom{n}{k} p^k q^{n-k}$$

and

$$E(X^2) = \sum_{k=0}^n k^2 \binom{n}{k} p^k q^{n-k}$$

from which we compute the variance of  $X$  by use of the formula

$$var(X) = E(X^2) - [E(X)]^2$$

**Theorem 6** *A binomially distributed random variable with parameters  $n$  and  $p$  has mean  $np$ , variance  $npq$ , and standard deviation  $\sqrt{npq}$ .*

**Proof.** On the sample space of  $n$  Bernoulli trials, let  $X_i$  (for  $i = 1, 2, \dots, n$ ) be the random variable which has the value of 1 or 0 according as the  $i^{\text{th}}$  trial is a success or a failure. Then each  $X_i$  has the following distribution:

$$\begin{array}{c|cc} x & 0 & 1 \\ \hline P(x) & q & p \end{array}$$

and the total number of successes is  $X = X_1 + X_2 + \dots + X_n$ .

For each  $i$  we have

$$E(X_i) = 0(q) + 1(p) = p$$

Using the linearity property of  $E$ , we have

$$\begin{aligned} E(X) &= E(X_1 + X_2 + \dots + X_n) \\ &= E(X_1) + E(X_2) + \dots + E(X_n) \\ &= p + p + \dots + p = np \end{aligned}$$

For each  $i$  we have

$$E(X_i^2) = 0^2(q) + 1^2(p) = p$$

and

$$\text{var}(X_i) = E(X_i^2) - [E(X_i)]^2 = p - p^2 = p(1 - p) = pq$$

The  $n$  random variables  $X_i$  are independent. Therefore

$$\begin{aligned} \text{var}(X) &= \text{var}(X_1 + X_2 + \dots + X_n) \\ &= \text{var}(X_1) + \text{var}(X_2) + \dots + \text{var}(X_n) \\ &= pq + pq + \dots + pq = npq \end{aligned}$$

Finally, we know that  $\sigma$  is the nonnegative square root of  $\text{var}(X)$ , that is

$$\sigma = \sqrt{npq}$$

This completes the proof.

**Example 25** *The probability that a marksman hits a target is  $p = \frac{1}{4}$ . She fires 100 times. What is the expected number  $\mu$  of times she will hit the target and the standard deviation  $\sigma$ ?*

Here  $p = \frac{1}{4}$ . Hence,

$$\mu = np = 100 \times \frac{1}{4} = 25$$

and

$$\sigma = \sqrt{npq} = \sqrt{100 \times \frac{1}{4} \times \frac{3}{4}} = 2.5$$

**Example 26** You take a 30-question true-false test after a night of partying, so you decide to just answer the questions by guessing. The expected number of correct answers will be around \_\_\_\_\_ give or take \_\_\_\_\_.

Here  $p = \frac{1}{2}$ . Hence,

$$\mu = np = 30 \times \frac{1}{2} = 15$$

and

$$\sigma = \sqrt{npq} = \sqrt{30 \times \frac{1}{2} \times \frac{1}{2}} \approx 2.7$$

So, the expected number of correct answers will be around 15 give or take 3.

### Normal Distribution

Let  $X$  be a random variable on a sample space  $\Omega$  where, by definition,  $\{a \leq X \leq b\}$  is an event in  $\Omega$ . Recall that  $X$  is said to be continuous if there is a function  $f(x)$  defined on the real line  $\mathbf{R} = (-\infty, \infty)$  such that

- (i)  $f(x) \geq 0$  ( $f$  is non-negative).
- (ii)  $\int_{-\infty}^{\infty} f(x) dx = 1$  (The area under the curve of  $f$  is one).
- (iii)  $P(a \leq X \leq b) = \int_a^b f(x) dx$  (The probability that  $X$  lies in the interval  $[a, b]$  is equal to the area under  $f$  between  $x = a$  and  $x = b$ ).

The amount of mass in an arbitrary interval  $a < X \leq b$ , which corresponds to the probability that the variable  $X$  will assume a value belonging to this interval, will be

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x) dx$$

If, in particular, we take here  $a = -\infty$ , we obtain

$$F(b) = \int_{-\infty}^b f(x) dx$$

and for  $b = \infty$

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

which means that the total mass of the distribution is unity. On the other hand, we obtain by differentiation of  $F(b) = \int_{-\infty}^b f(x) dx$ ,

$$F'(x) = f(x)$$

Therefore the pdf is the derivative of the cdf.

From  $P(a < X \leq b) = F(b) - F(a)$  we can also find that if we keep  $b$  fixed and allow  $a$  to tend to  $b$ ,

$$F(b) - F(b - 0) = P(X = b),$$

$$F(b + 0) - F(b) = 0.$$

So if  $F(x)$  is continuous in a certain point  $x = a$ , then  $P(X = a) = 0$ .

The most important example of a continuous random variable  $X$  is the *normal* random variable, whose pdf has the familiar bell-shaped curve. This distribution was discovered by De Moivre in 1733 as the limiting form of the binomial distribution.

The normal distribution is sometimes called the “Gaussian distribution” after Gauss who discussed it in 1809, it was actually already known in 1774 by Laplace.

Formally, a random variable  $X$  is said to be *normally distributed* if its pdf  $f$  has the following form:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}[(x-\mu)^2/\sigma^2]}.$$

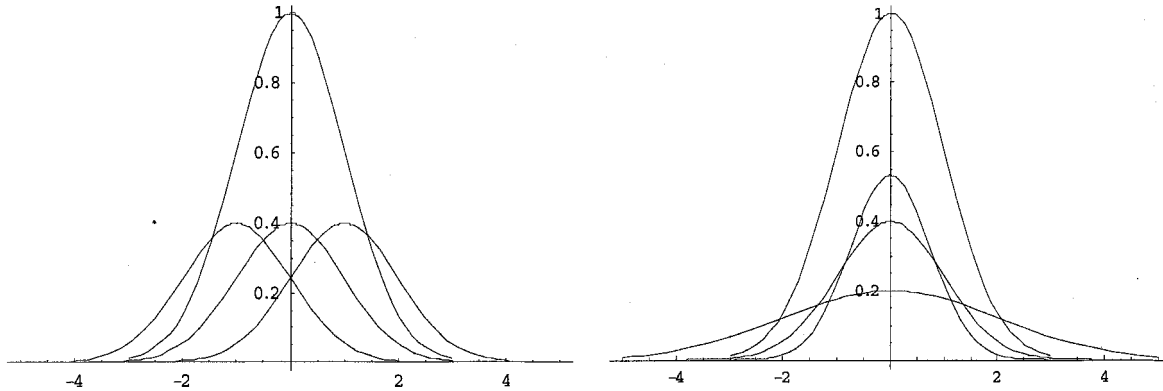
where  $\mu$  is any real number and  $\sigma$  is any positive number.

The above distribution, which depends on the *parameters*  $\mu$  and  $\sigma$ , is usually denoted by

$$N(\mu, \sigma^2)$$

Thus, we say that  $X \sim N(\mu, \sigma^2)$ , where the standard notation  $X \sim f(x)$  means that “ $X$  has probability distribution  $f(x)$ .”

The two diagrams below show the changes in the bell-shaped curves as  $\mu$  and  $\sigma$  vary. The one on the left shows the distribution of for  $\mu = -1, \mu = 0, \mu = 1$  and a constant value of  $\sigma$ , ( $\sigma = 1$ ). In the other one,  $\mu = 0$  and  $\sigma = .75, \sigma = 1, \sigma = 2$ .



Observe that each curve reaches its highest point at  $x = \mu$  and that the curve is symmetric about  $x = \mu$ . The *inflection* points, where the direction of the bend of the curve changes, occur when  $x = \mu + \sigma$  and  $x = \mu - \sigma$ .

Properties of the normal distribution follow:

**Normal Distribution  $N(\mu, \sigma^2)$ .**

- Mean or expected value,  $\mu$ .
- Variance,  $\sigma^2$ .
- Standard Deviation,  $\sigma$ .

That is, the mean, variance, and standard deviation of the normal distribution are  $\mu$ ,  $\sigma^2$  and  $\sigma$  respectively. That is why the symbols  $\mu$  and  $\sigma$  are used as parameters in the definition of the above pdf.

Among the most useful properties of the normal distribution is its preservation under linear transformation. Suppose that  $X \sim N(\mu, \sigma)$ . Recall that the standardized random variable corresponding to  $X$  is defined by

$$Z = \frac{X - \mu}{\sigma}$$

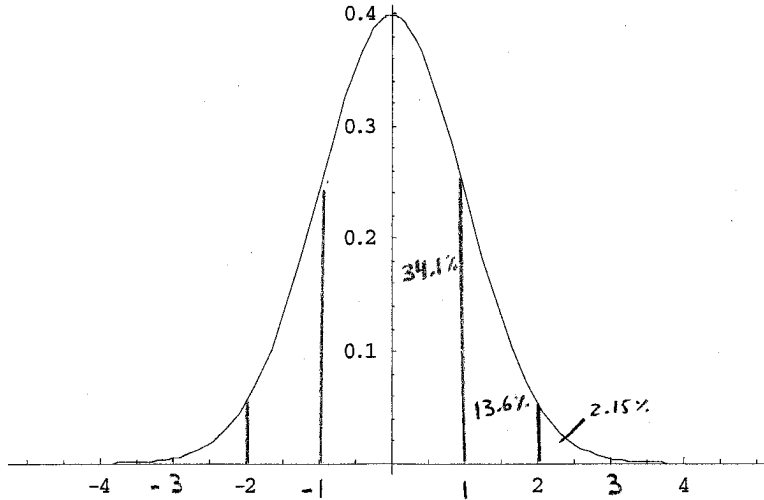
We note that  $Z$  is also a normal distribution and that  $\mu = 0$  and  $\sigma = 1$ , that is  $Z \sim N(0, 1)$ .

The pdf for  $Z$  obtained by setting  $z = \frac{(x-\mu)}{\sigma}$  in the above formula for  $N(\mu, \sigma)$ , follows:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

The specific notation  $\phi(z)$  is often used for this distribution and  $\Phi(z)$  for its cdf.

The graph of this function is:



The figure also tells us the percentage of area under the standardized normal curve and hence also under any normal distribution as follows:

68.2% for  $-1 \leq z \leq 1$  and for  $\mu - \sigma \leq x \leq \mu + \sigma$

95.4% for  $-2 \leq z \leq 2$  and for  $\mu - 2\sigma \leq x \leq \mu + 2\sigma$

99.7% for  $-3 \leq z \leq 3$  and for  $\mu - 3\sigma \leq x \leq \mu + 3\sigma$

This gives rise to the so-called:

68 – 95 – 99.7 rule

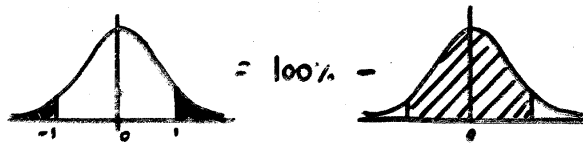
This rule says that, in a normally distributed population, 68 percent (approximately) of the population falls within one standard deviation of the mean, 95 percent falls within two standard deviations of the mean, and 99.7 percent falls within three standard deviations of the mean.

Tables of the standard normal cdf appear in most statistics textbooks. Because the form of the distribution does not change under a linear transformation, it is not necessary to tabulate the distribution for other values of  $\mu$  and  $\sigma$ .

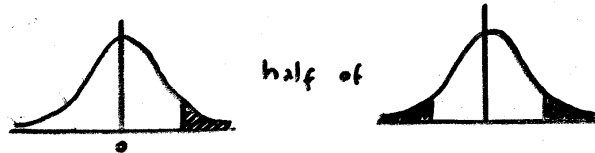
In addition, because the distribution is symmetric,  $\Phi(-z) = 1 - \Phi(z)$ , it is not necessary to tabulate both the negative and positive halves of the distribution.

**Example 27** *Lets say we want to find the area to the right of 1 under the normal curve.*

We go to the table, and we find that the area between -1 and 1 is roughly 68%. That means that the area outside this interval is 32%.



By symmetry, the area to the right of 1 is half of this, or 16%.



**Example 28** *Lets say we want to find now the area to the left of 2 under the normal curve.*

The area to the left of 2 is the sum of the area to the left of 2, and the area between 0 and 2.



The area to the left of 0 is half the total area: 50%. The area between 0 and 2 is about 48%. The sum is 98%.



## Uniform Distribution

A uniform distribution is constant over a bounded interval  $a \leq x \leq b$  and zero outside the interval. A random variable that has a uniform density function is said to be *uniformly distributed*.

A continuous random variable is uniformly distributed if the probability that its value will be in a particular subinterval of the bounded interval is equal to the probability that it will be in any other subinterval that has the same length. In other words, a uniformly distributed random variable is one for which all the values in some interval are “equally likely”.

Let  $k$  be the constant value of a uniform density function  $f(x)$  on the interval  $a \leq x \leq b$ . The value of  $k$  is then determined by the requirement that the total area under the graph of  $f$  be equal to 1. In particular, since  $f(x) = 0$  outside the interval  $a \leq x \leq b$ ,

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(x) \, dx = \int_a^b f(x) \, dx \\ &= \int_a^b k \, dx = kx \Big|_a^b = k(b - a) \end{aligned}$$

and so

$$k = \frac{1}{b - a}.$$

This last observation leads to the following formula for a uniform distribution. The pdf for the support  $X = [a, b]$  is

$$f(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{if } x > b \end{cases}$$

and the cdf is

$$f(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases}$$

**Example 29** *A certain traffic light remains red for 40 seconds at a time. You arrive (at random) at the light and find it red. What is the probability that you will have to wait at least 15 seconds for the light to turn green?*

Let  $x$  denote the time (in seconds) that you must wait. Since all waiting times between 0 and 40 are “equally likely,”  $x$  is uniformly distributed over the interval  $0 \leq x \leq 40$ . The corresponding uniform pdf is

$$f(x) = \begin{cases} \frac{1}{40} & \text{if } 0 \leq x \leq 40 \\ 0 & \text{otherwise} \end{cases}$$

and the desired probability is

$$P(15 \leq x \leq 40) = \int_{15}^{40} \frac{1}{40} dx = \frac{x}{40} \Big|_{15}^{40} = \frac{40 - 15}{40} = \frac{5}{8}.$$