Introduction to Probability

David Owen
{dowen@csee.wvu.edu}

1 Probability

For our purposes, probability is concerned with the following steps:

1. We conduct an **experiment**.
2. Experiments are broadly defined—they are simply anything with **outcomes**.
3. We are interested in a subset of the outcomes:

   What is the **probability** that something in that subset of outcomes occurs? For example, if we roll a six-sided die (that's the experiment) we might like to know the probability that we get an odd number ("odd numbers" is a subset of outcomes).

To speak precisely about experiments, outcomes and probability we need several definitions.

Definition: 1.1 *The set of all possible outcomes is called the* **sample space**, *or* S.

Definition: 1.2 *Any subset of* S *is called an* **event**.

Suppose we toss two coins. Our sample space $S = \{(h,h),(h,t),(t,h),(t,t)\}$ (h means "heads"; t means "tails"). We might be interested in the event (a subset of S) $\{(h,h),(t,t)\}$; that is, the event in which the coins match.

Definition: 1.3 *Given events* E *and* F, $E \cup F$ *is the union of the two events (with no duplicates).*

Definition: 1.4 *Given events* E *and* F, $E \cap F$ *(also written* EF*) is the intersection of the two events.*

For example, suppose $E_1 = \{(h,h),(h,t)\}$ and $E_2 = \{(h,t)\}$. $E_1 \cup E_2 = \{(h,h),(h,t)\}$ (in words: the first coin comes up "heads"), and $E_1 E_2 = \{(h,t)\}$.

Definition: 1.5 *Events* E *and* F *are* **mutually exclusive** *if* $EF = \emptyset$.

Definition: 1.6 E^c *= the* **complement** *of event* $E = S - E$ *(or* $S \setminus E$*).*

Definition: 1.7 *A real-valued function* $Pr : S \longrightarrow R$ *is called a* **probability function** *if:*

i. *The probability of* S, $\mathbf{Pr}(S) = 1$.

ii. $0 \leq \mathbf{Pr}(E) \leq 1$.

iii. *If events* E_1 *and* E_2 *are mutually exclusive,* $\mathbf{Pr}(E_1 \cup E_2) = \mathbf{Pr}(E_1) + \mathbf{Pr}(E_2)$.

Suppose we flip a single coin, so that $S = \{h, t\}$. Based on the above definitions, what is the probability we will get "heads?" The reader may be tempted to answer $\frac{1}{2}$, but this is not necessarily correct. What if we are flipping a strangely shaped or weighted coin, which for some reason tends to land more often on one side than the other? The definition of probability above states that probabilities may be *assigned* to events. This is an important point. It has been said that the assignment of probabilities is more an art than a science. We must be clear about the fact that we assign probabilities to events based on our experience or intuition about the experiment, and, based on the definitions above, derive results consistent with our initial assignments.

An alternate view of probability is sometimes used by statisticians. In the alternate view probability is the "relative frequency" of an event. This view of probability may in some cases conflict the set theoretic (or "axiomatic") definition of probability presented above:

$$\mathbf{Pr}(A) = \lim_{n\to\infty} \frac{n_A}{n},$$

where n is the number of times the experiment is conducted, and n_A is the number of times the experiment's result is A.

We may define the probability of the union of two events $\mathbf{Pr}(E \cup F)$ in terms of the individual probabilities E and F:

$$\mathbf{Pr}(E \cup F) = \mathbf{Pr}(E) + \mathbf{Pr}(F) - \mathbf{Pr}(EF).$$

The above may be expanded to define the probability of the union of many events in terms of their individual probabilities [2]:

Theorem: 1.1

$$\mathbf{Pr}(E_1 \cup E_2 \ldots \cup E_n) = \sum_{n=1}^{i} \mathbf{Pr}(E_i) - \sum_{i_1<i_2} \mathbf{Pr}(E_{i_1}E_{i_2}) + \ldots$$

$$+ (-1)^{r+1} \sum_{i_1<i_2<i_r} \mathbf{Pr}(E_{i_1}E_{i_2}\ldots E_{i_r}) + \ldots + (-1)^{n+1}\mathbf{Pr}(E_{i_1}E_{i_2}\ldots E_n).$$

The summation above, $\sum_{i_1<i_2<i_r} \mathbf{Pr}(E_{i_1}E_{i_2}\ldots E_{i_r})$, is taken over all $\binom{n}{r}$ possible subsets of the set $\{1, 2, \ldots n\}$. In words, Theorem 1.1 says that the probability of the union of n events is equal to the sum of probabilities of each of these events, minus the sum of probabilites of pairs of these events, plus the sum of probabilities of these events in groups of three, and so on alternating up to n [2].

2 Conditional Probability

2.1 Background

Definition: 2.1 *Given two events E and F, the* **conditional probability**, *which is the probability of E given F has occurred, is written* $\mathbf{Pr}(E|F)$.

For example, suppose we roll one fair die (by "fair" we mean the probability of each number between 1 and 6 is equal), so that our sample space $S = \{1, 2, 3, 4, 5, 6\}$. If the event $E = \{2, 4, 6\}$ (even numbers) and event $F = \{4\}$, $\mathbf{Pr}(E) = \frac{1}{2}$ and $\mathbf{Pr}(F) = \frac{1}{6}$, but what is $\mathbf{Pr}(E|F)$? In this case, it is easy to see that $\mathbf{Pr}(E|F) = \frac{1}{3}$.

Suppose we have a deck of cards numbered from 1 to 10; consider the event in which the card drawn is numbered ≥ 8. We are told that the card drawn is numbered ≥ 5. So, given that the card drawn is ≥ 5, what is the conditional probability that the card is numbered ≥ 8? We let event $E =$ *the card drawn is numbered* ≥ 8 and event $F =$ *the card drawn is numbered* ≥ 5. So what is the conditional probability of E given F? We use the following the definition of conditional probability:

$$\mathbf{Pr}(E|F) = \frac{\mathbf{Pr}(EF)}{\mathbf{Pr}(F)}.$$

$E = \{8, 9, 10\}$, $F = \{5, 6, 7, 8, 9, 10\}$, and $EF = \{8, 9, 10\}$; therefore $\mathbf{Pr}(EF) = \frac{3}{10}$, $\mathbf{Pr}(F) = \frac{6}{10}$, and:

$$\mathbf{Pr}(E|F) = \frac{\mathbf{Pr}(EF)}{\mathbf{Pr}(F)} = \frac{\frac{3}{10}}{\frac{6}{10}} = \frac{1}{2}.$$

If there are two children in a family, the probability of a child being a boy $= \frac{1}{2}$, and we know that one of the children is a boy, what is the conditional probability both are boys? Our initial sample space $S = \{(g, g), (g, b), (b, g), (b, b)\}$, the event in which one of the children is a boy $F = \{(g, b), (b, g), (b, b)\}$, and the event in which both are boys is $E = \{(b, b)\}$; therefore the conditional probability both are boys is given by:

$$\mathbf{Pr}(E|F) = \frac{\mathbf{Pr}(EF)}{\mathbf{Pr}(F)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

Definition: 2.2 *Two events A and B are said to be* **independent** *if* $\mathbf{Pr}(A|B) = \mathbf{Pr}(A)$.

As the definition above states, for two independent events A and B, $\mathbf{Pr}(A|B) = \mathbf{Pr}(A)$. Since $\mathbf{Pr}(A|B) = \frac{\mathbf{Pr}(AB)}{\mathbf{Pr}(A)}$, we can also say that for two independent events $\mathbf{Pr}(AB) = \mathbf{Pr}(A) \cdot \mathbf{Pr}(B)$ (substituting $\mathbf{Pr}(A)$ for $\mathbf{Pr}(A|B)$). Note that there is a difference between events being mutually exclusive and events being independent (neither property implies the other):

Mutually Exclusive	$\mathbf{Pr}(A \cup B) = \mathbf{Pr}(A) + \mathbf{Pr}(B)$
Independent	$\mathbf{Pr}(AB) = \mathbf{Pr}(A) \cdot \mathbf{Pr}(B)$

It is not always easy to tell whether two events are independent, and a small change in the statement of the problem can make the difference. Suppose we roll two fair dice. We define event $E_1 =$ *the sum of the dice is* 6 and event $F =$ *the first die rolled shows* 4. $\mathbf{Pr}(F) = \frac{1}{6}$.

$$E_1 = \textit{the sum of the dice is } 6 = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}.$$

So $\mathbf{Pr}(E_1) = \frac{5}{36}$. Our sample space $S = \{(1, 1), (1, 2), \ldots (2, 1), (2, 2), \ldots (6, 6)\}$—their are 36 different possibilities, so $\mathbf{Pr}(E_1 F) = \frac{1}{36}$. Events E_1 and F are not independent:

$$\mathbf{Pr}(E_1 F) \neq \mathbf{Pr}(E_1) \cdot \mathbf{Pr}(F)$$

$$\frac{1}{36} \neq \frac{5}{36} \cdot \frac{1}{6}.$$

But if we consider event E_2:

$$E_2 = \textit{the sum of the dice is } 7 = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}.$$

$\mathbf{Pr}(E_2) = \frac{1}{6}$, $\mathbf{Pr}(E_2 F) = \frac{1}{36}$, and events E_2 and F are independent (observe that in this case we can get a sum of 7 no matter what we roll with the first die; hence the events are independent):

$$\mathbf{Pr}(E_2 F) = \mathbf{Pr}(E_2) \cdot \mathbf{Pr}(F)$$

$$\frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6}.$$

There are two different notions of independence for an arbitrary set of events:

Definition: 2.3 *Events $E_1 \ldots E_n$ are said to be* **independent** *if, for every subset $E_{1'}, E_{2'}, \ldots E_{k'}$,* $\mathbf{Pr}(E_{1'} E_{2'} \ldots E_{k'}) = \mathbf{Pr}(E_{1'}) \mathbf{Pr}(E_{2'}) \ldots \mathbf{Pr}(E_{k'})$.

Definition: 2.4 *Events $E_1 \ldots E_n$ are said to be* **pairwise independent** *if* $\mathbf{Pr}(E_i E_j) = \mathbf{Pr}(E_i) \mathbf{Pr}(E_j)$ $\forall i, j$.

Why do we need two different definitions? The following example shows that events may be pairwise independent but not independent. Suppose we have an urn, in which are four balls numbered 1 to 4 ($S = \{1, 2, 3, 4\}$), and the probability of choosing any particular ball is $\frac{1}{4}$. We define three events and indicate their probabilities and the probabilities of their intersections:

Event	$E = \{1, 2\}$	$F = \{1, 3\}$	$G = \{1, 4\}$	$EF = \{1\}$
Probability	$\mathbf{Pr}(E) = \frac{1}{2}$	$\mathbf{Pr}(F) = \frac{1}{2}$	$\mathbf{Pr}(G) = \frac{1}{2}$	$\mathbf{Pr}(EF) = \frac{1}{4}$

Event	$EG = \{1\}$	$FG = \{1\}$	$EFG = \{1\}$
Probability	$\mathbf{Pr}(EG) = \frac{1}{4}$	$\mathbf{Pr}(FG) = \frac{1}{4}$	$\mathbf{Pr}(EFG) = \frac{1}{4}$

E, F, and G are pairwise independent:

$$\mathbf{Pr}(EF) = \mathbf{Pr}(E) \cdot \mathbf{Pr}(F)$$

$$\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}$$

$$\mathbf{Pr}(EG) = \mathbf{Pr}(E) \cdot \mathbf{Pr}(G)$$

$$\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}$$

$$\mathbf{Pr}(FG) = \mathbf{Pr}(F) \cdot \mathbf{Pr}(G)$$

$$\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2},$$

but not independent (by counterexample):

$$\mathbf{Pr}(EFG) \neq \mathbf{Pr}(E) \cdot \mathbf{Pr}(F) \cdot \mathbf{Pr}(G)$$

$$\frac{1}{4} \neq \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}$$

(of course independence trivially implies pairwise independence).

2.2 Bayes' Formula

For two events E and F defined under the same sample space S, anything in E must be in either F or F's complement, and cannot belong to both F and F's complement; therefore:

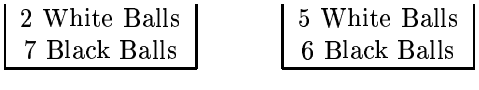
$$E = EF \cup EF^c\,.$$

Since events EF and EF^c are (obviously) mutually exclusive,

$$\begin{aligned}\mathbf{Pr}(E) &= \mathbf{Pr}(EF) + \mathbf{Pr}(EF^c) \\ &= \mathbf{Pr}(E|F)\mathbf{Pr}(F) + \mathbf{Pr}(E|F^c)\mathbf{Pr}(F^c) \\ &= \mathbf{Pr}(E|F)\mathbf{Pr}(F) + \mathbf{Pr}(EF^c)(E|F^c)(1 - \mathbf{Pr}(F))\,.\end{aligned}$$

This is Bayes' formula (attributed to Bayes but apparently due to Cardan); in words, the probability of E is a weighted average of: 1) the conditional probability of E, given F has occurred, and 2) the conditional probability of E, given F has not occurred, with 1) and 2) weighted according to the probability of F and F^c [2].

For example, we have two urns, A and B:



A B

We flip a fair coin. If the coin shows "heads," we choose a ball from urn A; if it shows "tails," we choose from B. Suppose we have done the experiment and chosen a white ball (W = *we chose a white ball*; H = *our coin showed heads*). What is the conditional probability our coin showed "heads"[1]?

$$\mathbf{Pr}(H|W) = \frac{\mathbf{Pr}(HW)}{\mathbf{Pr}(W)}$$

$$= \frac{\mathbf{Pr}(W|H)\mathbf{Pr}(H)}{\mathbf{Pr}(W|H)\mathbf{Pr}(H) + \mathbf{Pr}(W|H^c)\mathbf{Pr}(H^c)}$$

$$= \frac{\frac{2}{9} \cdot \frac{1}{2}}{\frac{2}{9} \cdot \frac{1}{2} + \frac{5}{11} \cdot \frac{1}{2}}$$

$$= \frac{22}{67}.$$

3 Random Variables

Often we are interested in a function of an experiment's outputs (where the sample space S = the set of outputs). Such a function is called a "random variable":

Definition: 3.1 *A real-valued function of the outcome of an experiment is called a* **random variable**.

Definition: 3.2 *The* **cumulative distribution function** *(***CDF***) of the random variable X is defined as $F(b) = \mathbf{Pr}(X \leq b)$, $-\infty < b < \infty$, such that:*

i. $F(b)$ *is a non-decreasing function of* b.

ii. $\lim_{(b \to \infty)} F(b) = 1$.

iii. $\lim_{(b \to -\infty)} F(b) = 0$.

3.1 Discrete Random Variables

If we restrict our discussion to integer-valued functions, we are dealing with **discrete random variables**. For example, we roll two fair dice. Let random variable X = *the sum of the two dice.* Since there is only one way to get a sum of 2, $\mathbf{Pr}(X = 2) = \frac{1}{36}$; $\mathbf{Pr}(X = 3) = \frac{2}{36}$, etc. And because we assign probabilities to the values taken by a random variable, the conditions stated in Definiton 1.7 must be satisfied. To show that they are for this example, we begin with the first condition, that $\mathbf{Pr}(S) = 1$:

$$\mathbf{Pr}(S) = \sum_{i=2}^{12} \mathbf{Pr}(X = i)$$

$$= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36} + \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36}$$

$$= \frac{1+2+3+4+5+6+5+4+3+2+1}{36} = \frac{36}{36} = 1\,.$$

The other two conditions are also met. All the probabilities are ≥ 0, and mutually exclusive events add in way described in Definition 1.7.

For a second example, suppose we toss two coins, and the random variable Y = *the number of "heads."* $\mathbf{Pr}(Y = 0) = \frac{1}{4}$, $\mathbf{Pr}(Y = 1) = \frac{1}{2}$, and $\mathbf{Pr}(Y = 2) = \frac{1}{4}$; therefore the sum of probabilities = 1 (again all probabilities are ≥ 0 and mutually exclusive events add in the proper way).

[1] The word *conditional* is important. Without it the probability our coin showed "heads" is simply $\frac{1}{2}$.

If we repeatedly toss a single coin, we can define a random variable Z as the number of tosses required to get "heads." Let $p =$ *the probability we get "heads" in a single toss.* The probability we require two tosses will be the product of the probability we fail the first time and succeed the second. The probably we require n tosses will be the probability we fail $n-1$ times multiplied by the probability we succeed the nth time:

$$\mathbf{Pr}(Z=1) = p$$

$$\mathbf{Pr}(Z=2) = (1-p)p$$

$$\mathbf{Pr}(Z=3) = (1-p)^2 p$$

$$\vdots$$

$$\mathbf{Pr}(Z=n) = (1-p)^{(n-1)} p\,.$$

Again, the sum of all possible Z values $= 1$:

$$\sum_{i=0}^{\infty} (1-p)^i p$$

$$= p \cdot \sum_{i=0}^{\infty} (1-p)^i$$

$$= \frac{p}{1-(1-p)} = 1\,.$$

The simplification above makes use of the following rule for infinite decreasing geometric series, valid for all $|x| < 1$; for more on this and other series simplification, see chapter three on summations in [1]:

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}\,.$$

We also note that $\mathbf{Pr}(a < X \leq b) = F(b) - F(a)$. What about $\mathbf{Pr}(X < b)$? We can use a limit to express the strict inequality ("<") in terms of the definition's less-than-or-equal ("≤"):

$$\mathbf{Pr}(X < b) = \lim_{h \to 0} \mathbf{Pr}(X \leq b - h)\,.$$

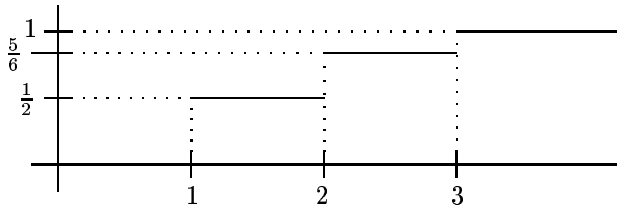
Definition: 3.3 *Given a discrete random variable X that may take on values $x_1, x_2, \ldots x_k$, we define the* **Probability Mass Function** *(***PMF***) such that:*

i. $p(x_i) > 0, i = \{1, 2, \ldots k\}$.

ii. $p(x) = 0$, *for all other x values.*

iii. $\sum_{i=1}^{k} p(x_i) = 1$ [3].

The CDF for a discrete random variable may be expressed in terms of the PMF $p(a)$ as $F(a) = \sum_{x_i < a} p(x_i)$. Suppose a random variable X can take on values 1, 2 and 3 with probability $\mathbf{Pr}(X=1) = \frac{1}{2}$, $\mathbf{Pr}(X=2) = \frac{1}{3}$ and $\mathbf{Pr}(X=3) = \frac{1}{6}$. The graph below illustrates the CDF for X:



Discrete random variables are classified according to their PMF.

3.2 PMF's and Examples for Discrete Random Variables

Note—much of the material for the remaining sections was taken directly from [3].

Definition: 3.4 *A* **Bernoulli trial** *is an experiment with precisely two possible outcomes, "success" and "failure." A random variable X that may take on values 0 and 1 (0 indicates "failure"; 1 indicates "success") is called a* **Bernoulli** *random variable.*

The PMF for a Bernoulli random variable X is:

$$\mathbf{Pr}(X = 0) = p(0) = 1 - p\,;$$

$$\mathbf{Pr}(X = 1) = p(1) = p\,.$$

Definition: 3.5 *A* **binomial random variable** *takes values representing the number of successes in n independent Bernoulli trials.*

If the probability of success in a single Bernoulli trial is p, and there are n trials, the PMF for a binomial random variable is:

$$\mathbf{Pr}(X = i) = p(i) = \binom{n}{i} p^i (1-p)^{n-i},\ i = 0, 1, \ldots n\,.$$

Note—$\binom{n}{i}$, which is the number possible sequences with i successes in n trials, is:

$$\binom{n}{i} = \frac{n!}{(n-i)!\,i!}\,.$$

If we sum the PMF of a binomial random variable from 0 to n, we find (as we would expect):

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = (p + (1-p))^n = 1\,.$$

As an example of the binomial random variable, we conduct an experiment in which 4 fair coins are flipped. Let $X =$ *the number of "heads"*. What is the probability $X = 2$ (we get two "heads")?

$$\mathbf{Pr}(X = i) = p(i) = \binom{n}{i} p^i (1-p)^{n-i}$$

$$p(2) = \binom{4}{2} \left(\frac{1}{2}\right)^2 \left(1 - \frac{1}{2}\right)^{(4-2)} = \left(\frac{4!}{(4-2)!\,2!}\right) \left(\frac{1}{4}\right) \left(\frac{1}{4}\right) = \left(\frac{24}{2 \cdot 2}\right) \left(\frac{1}{16}\right) = \frac{3}{8}\,.$$

Or suppose we have an airplane with an even number of engines, and the plane is only able to land if at least $\frac{1}{2}$ of the engines are intact. Would we be better off with a 4-engine plane or a 2-engine plane? Does it depend on the probability of engine failure? Or would we always be better with 4 engines?

Let $1 - p =$ *the probability of engine failure.* Then the probability of success for the 4-engine plane (the probability 2 engines fail plus the probability 1 engine fails plus the probability 0 engines fail) is:

$$\binom{4}{2} p^2 (1-p)^{4-2} + \binom{4}{3} p^3 (1-p)^{4-3} + \binom{4}{4} p^4 (1-p)^{4-4}$$

$$= 6p^2(1-p)^2 + 4p^3(1-p) + p^4\,.$$

And the probability of success for the 2-engine plane is the probability 1 engine fails plus the probability 0 engines fail:

$$\binom{2}{1} p^1 (1-p)^{2-1} + \binom{2}{2} p^2 (1-p)^{2-2}$$

$$= 2p(1-p) + p^2 \,.$$

The 4-engine plane is safer if:

$$6p^2(1-p)^2 + 4p^3(1-p) + p^4 \geq 2p(1-p) + p^2$$

$$6p(1-p)^2 + 4p^2(1-p) + p^3 \geq 2 - p$$

$$3p^3 - 8p^2 + 7p - 2 \geq 0$$

$$(p-1)^2(3p-2) \geq 0$$

$$p \geq \frac{2}{3} \,.$$

The 4-engine plane is safer only when the probability of an engine not failing is $\geq \frac{2}{3}$. When the probability is $< \frac{2}{3}$, the 2-engine plane is actually safer.

Definition: 3.6 *The geometric random variable takes values representing the number of Bernoulli trials required to get the first successful result.*

If we conduct n independent Bernoulli trials, the probability that we fail $n-1$ times $= (1-p)^{n-1}$; the probability that we succeed in the nth trial $= p$. Therefore the PMF for the geometric random variable is:

$$\mathbf{Pr}(X = n) = p(n) = (1-p)^{n-1} p \,.$$

As we would expect, the sum of probabilities for $n \geq 1 = 1$:

$$\sum_{n=1}^{\infty} (1-p)^{n-1} p = p \sum_{n=1}^{\infty} (1-p)^{n-1}$$

$$= p \sum_{k=0}^{\infty} (1-p)^k = p \left(\frac{1}{1-(1-p)} \right) = 1 \,.$$

3.3 The Normal (Continuous) Random Variable

Definition: 3.7 *if X is a* **continous random variable**, *there exists a non-negative function $f(x)$ defined for all $x \in (-\infty, \infty)$ having the property that for any set B of real numbers:*

i. $\mathbf{Pr}(X \in B) = \int_B f(x)dx$.

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

iii. $\mathbf{Pr}(a \leq X \leq b) = \int_a^b f(x)dx$.

iv. $\mathbf{Pr}(X = a) = 0$.

For continous random variables, the cumulative distribution function is:

$$F_X(a) = \int_{-\infty}^{a} f(x)dx.$$

Definition: 3.8 *X is a* **Normal Random Variable** *with parameters μ and σ^2 if:*

$$f(x) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right) e^{(-(x-\mu)^2/2\sigma^2)},\ -\infty < x < \infty\,.$$

Theorem: 3.1 *If X is normally distributed with parameters μ and σ^2, then $Y = \alpha X + \beta$ is normally distributed with parameters $\alpha\mu + \beta$ and $\alpha^2\sigma^2$.*

Proof: *Suppose $\alpha > 0$ and the cumulative distribution function of Y is given by*

$$F_Y(a) = \mathbf{Pr}(Y \leq a).$$

Substitute $\alpha X + \beta$ for Y:

$$\begin{aligned} F_Y(a) &= \mathbf{Pr}(\alpha X + \beta \leq a) \\ &= X \leq \frac{a-\beta}{\alpha} \\ &= F_X\left(\frac{a-\beta}{\alpha}\right), \end{aligned}$$

apply the definition of the cumulative distribution function for the normal distribution:

$$= \int_{-\infty}^{a-\beta/\alpha} \left(\frac{1}{\sigma\sqrt{2\pi}}\right) e^{(-(x-\mu)^2/2\sigma^2)} dx\,,$$

and substitute in $v = \alpha x + \beta$:

$$\int_{-\infty}^{a} \left(\frac{1}{\alpha\sigma\sqrt{2\pi}}\right) \exp\left\{\frac{-(v-(\alpha\mu+\beta))^2}{2\alpha^2\sigma^2}\right\} dv\,.$$

Now, since $F_Y(a) = \int_{-\infty}^{a} f_Y(v)dv$:

$$f_Y(v) = \left(\frac{1}{\alpha\sigma\sqrt{2\pi}}\right) \exp\left\{\frac{-(v-(\alpha\mu+\beta))^2}{2\alpha^2\sigma^2}\right\}.$$

The right-hand side of this equation is just the normal distribution with parameters $\alpha\mu + \beta$ and $\alpha^2\sigma^2$. For similar reasons, we would come to the same conclusion for $\alpha < 0$. □

Corollary: 3.1 *A continuous random variable Y with a* **standard normal distribution** *has parameters $\mu = 0$ and $\sigma^2 = 1$ and is given by:*

$$Y = \frac{x-\mu}{\sigma}\,.$$

3.4 Expectation (Expected Value)

Definition: 3.9 *If X is a random variable, we define the* **expected value** *$E(X)$ for discrete X as:*

$$E(X) = \sum_{x:p(x)>0} x \cdot p(x)$$

and for continuous X as:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx\,.$$

The expected value is in some qualitative sense an average. It has no mathematical relationship to the average we are familiar with, but describes what value we expect a random variable to take. For example, if we roll a fair die, what is the expected value for a random variable X representing the outcome?

$$\begin{aligned} E(X) &= 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} \\ &= \frac{21}{6} = \frac{7}{2}\,. \end{aligned}$$

3.4.1 Expectation of a Bernoulli Random Variable

We can calculate the expected value of a Bernoulli variable according to the definition above:

$$E(X) = \sum_{x \cdot p(x) > 0} x \cdot p(x) = 0(1-p) + 1 \cdot p = p\,.$$

3.4.2 Expectation of a Binomial Random Variable

The expected value of a binomial random variable is more complicated (recall that for a binomial random variable with n independent trials $\mathbf{Pr}(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$):

$$E(X) = \sum_{i=0}^{n} i \binom{n}{i} p^i (1-p)^{n-i}\,.$$

If $i = 0$, $i \binom{n}{i} p^i (1-p)^{n-i} = 0$, so we have:

$$E(X) = \sum_{i=1}^{n} i \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=1}^{n} \left(\frac{i \cdot n!}{(n-i)!\, i!} \right) p^i (1-p)^{n-i}\,,$$

and because $i! = i \cdot (i-1)!$, we divide by i:

$$= \sum_{i=1}^{n} \left(\frac{n!}{(n-i)!(i-1)!} \right) p^i (1-p)^{n-i}\,.$$

$$= np \sum_{i=1}^{n} \frac{(n-1)!}{(n-i)!(i-1)!} p^{i-1} (1-p)^{n-i}\,,$$

$$= np \sum_{i=1}^{n} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i}\,.$$

We let $k = i - 1$ and substitute:

$$= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} = np(p + (1-p))^{n-1} = np\,.$$

So, for a binomial random variable, the expected number of successes in n independent trials is n times the probability that a single trial is successful.

3.4.3 Expectation of a Geometric Random Variable

We now calculate the expected value of a geometric random variable (recall that for a geometric random variable $\mathbf{Pr}(X = i) = p(1-p)^{i-1}$). Since at least 1 trial must take place, we begin the summation with $n = 1$:

$$E(X) = \sum_{n=1}^{\infty} np(1-p)^{n-1}\,.$$

We let $q = 1 - p$ and substitute:

$$= p \sum_{n=1}^{\infty} n q^{n-1} .$$

The derivative of q^n, $\frac{d}{dq}(q^n) = nq^{n-1}$:

$$= p \sum_{n=1}^{\infty} \frac{d}{dq}(q^n) = p \frac{d}{dq} \left(\sum_{n=1}^{\infty} q^n \right) = p \frac{d}{dq} \left(\frac{q}{1-q} \right) = \frac{p}{(1-q)^2} = \frac{1}{p} .$$

So the number of trials we expect to require is equal to the reciprocal of the probability a single trial is successful.

3.4.4 Expectation of a Normal Random Variable

As stated above, the expected value of a continuous random variable is $\int_{-\infty}^{\infty} x \cdot f(x) dx$. A normal random variable with parameters μ and σ^2 is defined as:

$$f(x) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right) e^{(-(x-\mu)^2/2\sigma^2)} ,$$

so the expected value of a normal random variable with these parameters is:

$$E(X) = \int_{-\infty}^{\infty} x \left(\left(\frac{1}{\sigma\sqrt{2\pi}} \right) e^{(-(x-\mu)^2/2\sigma^2)} \right) dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x \left(e^{(-(x-\mu)^2/2\sigma^2)} \right) dx .$$

Since $x = (x - \mu) + \mu$, we can write:

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu) \left(e^{(-(x-\mu)^2/2\sigma^2)} \right) dx + \mu \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(e^{(-(x-\mu)^2/2\sigma^2)} \right) dx .$$

We note that the last term $\mu \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(e^{(-(x-\mu)^2/2\sigma^2)} \right) dx = \mu \int_{-\infty}^{\infty} f(x) dx$; we let $y = x - \mu$ and substitute:

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} y \left(e^{(-y^2/2\sigma^2)} \right) dy + \mu \int_{-\infty}^{\infty} f(x) dx .$$

Since $y \left(e^{(-y^2/2\sigma^2)} \right)$ is an odd function, $\int_{-\infty}^{\infty} y \left(e^{(-y^2/2\sigma^2)} \right) dy = 0$; therefore:

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

and because X is a random variable, the sum of its probabilities taken from $-\infty$ to ∞ must be $= 1$, so:

$$E(X) = \mu .$$

This is why μ is commonly called the "mean" value of the normal random variable.

3.5 More on Expectation

Expectation may also be defined for a function of a random variable (for example X^2, $5X + 6$, etc.). Suppose X is a random variable that can take on values $0, 1$, or 2, with probabilities $\mathbf{Pr}(X = 0) = 0.2$, $\mathbf{Pr}(X = 1) = 0.5$, and $\mathbf{Pr}(X = 2) = 0.3$. The expected value of X is:

$$E(X) = 0(0.2) + 1(0.5) + 2(0.3) = 0.5 + 0.3 = 0.8\,.$$

The expected value of Y, where $Y = X^2$, is:

$$E(Y) = 0^2(0.2) + 1^2(0.5) + 2^2(0.3) = 0.5 + 1.2 = 1.7\,.$$

Definition: 3.10 *In general we define expectation for a real-valued function of a discrete random variable $g(X)$:*

$$E(g(X)) = \sum_{x:p(x)>0} g(x)p(x)\,,$$

and of a continuous random variable:

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx\,.$$

Theorem: 3.2 *For a random variable X and constants a and b, $E(aX + b) = a \cdot E(X) + b$.*

Proof: *For discrete random variables:*

$$E(aX + b) = \sum_{x:p(x)>0} (ax + b)p(x)$$

$$= a\left(\sum_{x:p(x)>0} xp(x)\right) + b\left(\sum_{x:p(x)>0} p(x)\right) = a \cdot E(X) + b\,,$$

and for continuous random variables:

$$E(aX + b) = \int_{-\infty}^{\infty} (ax + b)f(x)dx$$

$$= a\left(\int_{-\infty}^{\infty} x \cdot f(x)dx\right) + b\left(\int_{-\infty}^{\infty} f(x)dx\right)$$

$$= a \cdot E(X) + b\,.$$

□

Definition: 3.11 *We define the n***th moment of** *X for discrete random variables as:*

$$E(X^n) = \sum_{x:p(x)>0} x^n p(x)\,,$$

and for continuous random variables as

$$E(X^n) = \int_{-\infty}^{\infty} x^n \cdot f(x)dx\,.$$

Definition: 3.12 *We define the* **variance** $Var(X) = E[(X - E(X))^2]$.

The variance of X measures the expected square of the deviation of X from X's expected value—in very loose terminology, the variance represents how much X "varies" from its expected value. As an example, we calculate the variance of a normal random variable X with parameters μ and σ^2. Since $E(X) = \mu$,

$$Var(X) = E[(x-\mu)^2]$$
$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 \left(e^{-(x-\mu)^2/2\sigma^2} dx \right) .$$

We let $y = (x-\mu)/\sigma$ and substitute:

$$Var(x) = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 \left(e^{-y^2/2} dy \right) ,$$

and by the integration identity $\int_{-\infty}^{\infty} y^2 \left(e^{-y^2/2} dy \right) = \sqrt{2\pi}$, we have:

$$Var(X) = \sigma^2 .$$

Theorem: 3.3 *Given a random variable X,*

$$Var(X) = E(X^2) - (E(X))^2 .$$

Proof: *(here we prove Theorem 3.3 for the continuous case; a similar proof holds for the discrete case) Suppose X is a continuous random variable with density function f.*

$$Var(X) = E((X - E(X))^2)$$
$$= E(X^2 - 2E(X) \cdot X + E(X)^2)$$
$$= \int_{-\infty}^{\infty} (x^2 - 2E(X) \cdot x + E(X)^2) f(x) dx$$
$$= \left(\int_{-\infty}^{\infty} x^2 f(x) dx \right) - 2E(X) \left(\int_{-\infty}^{\infty} x \cdot f(x) dx \right) + E(X)^2 \left(\int_{-\infty}^{\infty} f(x) dx \right)$$
$$= E(X^2) - 2E(X)(E(X)) + E(X)^2$$
$$= E(X^2) - E(X)^2 .$$

□

Definition: 3.13 *For probability statements concerning two or more random variables, which we call* **jointly distributed random variables**, *we define the* **joint cumulative distribution function** *F as follows. For random variables $X_1, X_2, \ldots X_n$:*

$$F(a_1, a_2, \ldots a_n) = \mathbf{Pr}(X_1 \leq a_1, X_2 \leq a_2, \ldots X_n \leq a_n), \ -\infty < a_i < \infty .$$

For two random variables X and Y, $F(a,b) = \mathbf{Pr}(X \leq a, Y \leq b)$. We note here but do not prove a very important theorem, called the **Linearity of Expectation Rule**:

Theorem: 3.4 *Given two random variables X and Y defined for the same sample space S:*

$$E[aX + bY] = a \cdot E[X] + b \cdot E[Y] ,$$

where a and b are arbitrary constansts;
Linearity of Expectation also holds for an arbitrary number of random variables $X_1, X_2, \ldots X_n$ defined for the same sample space S:

$$E[c_1X_1 + c_2X_2 + \ldots c_nX_n] = c_1 \cdot E[X_1] + c_2 \cdot E[X_2] + \ldots c_n \cdot E[X_n] ,$$

where $c_1, c_2, \ldots c_n$ are arbitrary constants.

Note that Theorem 3.4 applies events $E[X]$ and $E[Y]$ regardless of their relationship to each other—they may or may not be independent (or anything else). This is a very important point, as it makes the theorem very useful.

We can use Theorem 3.4 to calculate the expected sum when three fair dice are rolled. Suppose $X = \textit{the sum of the three dice.}$ Since we have already calculated the expected value for one fair die, which is $\frac{7}{2}$,

$$E(X) = E(X_{\text{first die}}) + E(X_{\text{second die}}) + E(X_{\text{third die}}) = 3\left(\frac{7}{2}\right) = \frac{21}{2}.$$

References

[1] T. Cormen, C. Leiserson and R. Rivest. *Introduction to Algorithms.* McGraw-Hill, 1999.

[2] S. Ross. *A First Course in Probability.* Macmillan, 1976.

[3] S. Ross *Introduction to Probability Models.* Harcourt, 2000.