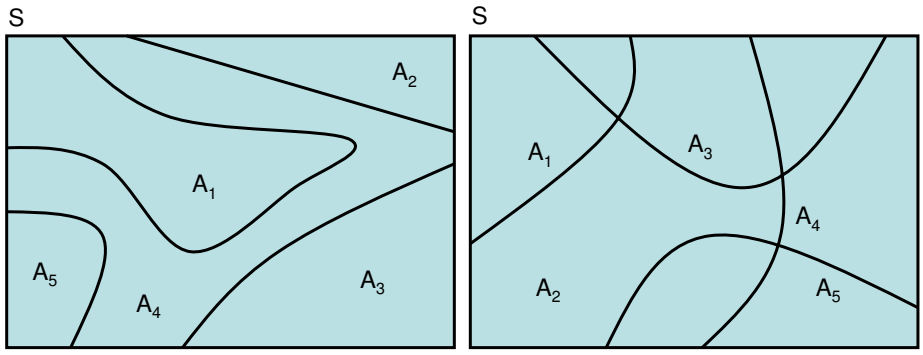
3 PROBABILITY

In this section, we discuss elements of probability, as a prerequisite for studying random processes.

3.1 Events

Define an event space S that has in it a number of events A_i. If the set of possible events A_i covers the space completely, then we will always get one of the events when we take a sample. On the other hand, if some of the space S is not covered with an A_i then it is possible that a sample is not classified as any of the events A_i. Events A_i may be overlapping in the event space, in which case they are *composite* events; a sample may invoke multiple events. But the A_i may not overlap, in which case they are *simple* events, and a sample brings only one event A_i, or none if the space S is not covered. In the drawing below, simple events cover the space on the left, and composite events cover the space on the right.



Intuitively, the probability of an event is the fraction of the number of positive outcomes to the total number of outcomes. Assign to each event a probability, so that we have

$$\begin{aligned} p_i &= p(A_i) \geq 0 \\ p(S) &= 1. \end{aligned}$$

That is, each defined event A_i has a probability of occurring that is greater than zero, and the probability of getting a sample from the entire event space is one. Hence, the probability has the interpretation of the area of the event A_i. It follows that the probability of A_i is exactly one minus the probability of A_i not occuring:

$$p(A_i) = 1 - p(\bar{A}_i).$$

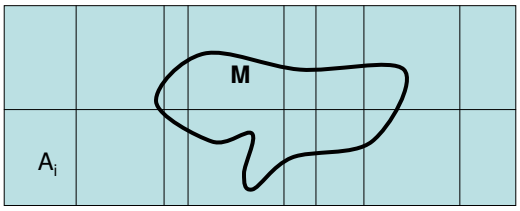
Furthermore, we say that if A_i and A_j are non-overlapping, then the probability of either A_i or A_j occuring is the same as the sum of the separate probabilities:

$$p(A_i \cup A_j) = p(A_i) + p(A_j).$$

Similarly if the A_i and A_j do overlap, then the probability of either or both occurring is the sum of the separate probabilities minus the sum of both occurring:

$$p(A_i \cup A_j) \quad = \quad p(A_i) + p(A_j) - p(A_i \cap A_j).$$

As a tangible example, consider a six-sided die. Here there are six events $A_1, A_2, A_3, A_4, A_5, A_6$, corresponding with the six possible values that occur in a sample, and $p(A_i) = 1/6$ for all i. The event that the sample is an even number is $M = A_2 \cup A_4 \cup A6$, and this is a composite event.



3.2 Conditional Probability

If a composite event M is known to have occurred, a question arises as to the probability that one of the constituent simple events A_i occurred. This is written as $P(A_j|M)$, read as "the probability of A_j, given M," and this is a conditional probability. The key concept here is that M replaces S as the event space, so that $p(M) = 1$. This will have the natural effect of inflating the probabilities of events that are part of event M, and in fact

$$p(A_j|M) = \frac{p(A_j \cap M)}{p(M)}.$$

Referring to our die example above, if M is the event of an even result, then we have

$$\begin{aligned} M &= A_2 \cup A_4 \cup A_6 \\ p(M \cap A_2) &= p(A_2) = 1/6 \\ p(M) &= 1/2 \longrightarrow \\ p(A_2|M) &= \frac{1/6}{1/2} = 1/3. \end{aligned}$$

Given that an event result was observed (composite event M), the probability that a two was rolled is 1/3. Now if all the A_j are independent (simple) events and M is a composite event, then we can write an opposing rule:

$$p(M) = p(M|A_1)p(A_1) + \cdots + p(M|A_n)p(A_n).$$

This relation collects conditional probabilities of M given each separate event A_i. Its logic is easily seen in a graph. Here is an example of how to use it in a practical problem. Box

A has 2000 items in it of which 5% are defective; box B has 500 items with 40% defective; boxes C and D each contain 1000 items with 10% defective. If a box is picked at random, and one item is taken from that box, what is the probability that it is defective? M is the composite event of a defective item, so we are after $p(M)$. We apply the formula above to find

$$p(M) = 0.05 \times 0.25 + 0.40 \times 0.25 + 0.10 \times 0.25 + 0.10 \times 0.25 = 0.1625.$$

3.3 Bayes' Rule

Consider a composite event M and a simple event A_i. We have from conditional probability above

$$p(A_i|M) = \frac{p(A_i \cap M)}{p(M)}$$
$$p(M|A_i) = \frac{p(A_i \cap M)}{p(A_i)},$$

and if we eliminate the denominator on the right-hand side, we find that

$$\begin{aligned} p(M|A_i) &= \frac{p(A_i|M)p(M)}{p(A_i)} \\ p(A_i|M) &= \frac{p(M|A_i)p(A_i)}{p(M)}. \end{aligned}$$

The second of these is most interesting - it gives the probability of a simple event, conditioned on the composite event, in terms of the composite event conditioned on the simple one! Recalling our above formula for $p(M)$, we thus derive Bayes' rule:

$$p(A_i|M) = \frac{p(M|A_i)p(A_i)}{p(M|A_1)p(A_1) + \cdots + p(M|A_n)p(A_n)}.$$

Here is an example of its use. Consider a medical test that is 99% accurate - it gives a negative result for people who do not have the disease 99% of the time, and it gives a positive result for people who do have the disease 99% of the time. Only one percent of the population has this disease. Joe just got a positive test result: What is the probability that he has the disease? The composite event M is that he has the disease, and the simple events are that he tested positive $(+)$ or he tested negative $(-)$. We apply

$$\begin{aligned} p(M|+) &= \frac{p(+|M)p(M)}{p(+)} \\ &= \frac{p(+|M)p(M)}{p(+|M)p(M) + p(+|\bar{M})p(\bar{M})} \\ &= \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.01 \times 0.99} \\ &= 1/2. \end{aligned}$$

This example is not well appreciated by many healthcare consumers!

Here is another example, without so many symmetries. Box A has nine red pillows in it and one white. Box B has six red pillows in it and nine white. Selecting a box at random and pulling out a pillow at random gives the result of a red pillow. What is the probability that it came from Box A? M is the composite event that it came from Box A; the simple event is that a red pillow was collected (R). We have

$$\begin{aligned} p(M|R) &= \frac{p(R|M)p(M)}{p(R)} \\ &= \frac{p(R|M)p(M)}{p(R|M)p(M) + p(R|\bar{M})p(\bar{M})} \\ &= \frac{0.9 \times 0.5}{0.9 \times 0.5 + 0.4 \times 0.5} \\ &= 0.692. \end{aligned}$$

3.4 Random Variables

Now we assign to each event A_i in the sample space a given value: each A_i corresponds with an x_i. For instance, a coin toss resulting in heads could be equated with a \$1 reward, and each tails could trigger a \$1 loss. Dollar figures could be assigned to each of the faces of a die. Hence we see that if each event A_i has a probability, then so will the numerical values x_i.

The average value of x_i can be approximated of course by sampling the space N times, summing all the x's, and dividing by N. As N becomes bigger, this computation will give an increasingly accurate result. In terms of probabilities the formula for the *expected value* is

$$\bar{x} = E(x) = \sum_{i=1}^{n} p(A_i)x_i.$$

The equivalence of this expected value with the numerical average is seen as follows: if the space is sampled N times, and the number of results $[A_i, x_i]$ is k_i, then $p(A_i) \simeq k_i/N$.

Superposition is an important property of the expectation operator:

$$E(x + y) = E(x) + E(y).$$

The mean of a function of x is defined using probabilities of the random variable x:

$$E\left[f(x(\xi))\right] = \sum_{i=1}^{n} f(x_i)p_i.$$

Another important property of a random variable is the variance - a measure of how much the x varies from its own mean:

$$\sigma^2 = E\left[(x - \bar{x})^2\right]$$

$$= \quad E(x^2) - \bar{x}^2.$$

The second line is apparent because $E(-2x\bar{x}) = -2\bar{x}^2$. Note we use the symbol σ^2 for variance; the standard deviation σ is just the square root, and has the same units as does the random variable x.

3.5 Continuous Random Variables and the Probability Density Function

Let us suppose now the random event has infinitely many outcomes: for example, the random variable x occurs *anywhere* in the range of $[0, 1]$. Clearly the probability of hitting any specific point is zero (although not impossible). We proceed this way:

$$p(x \text{ is in the range}[x_o, x_o + dx]) = \underline{p}(x_o)dx,$$

where $\underline{p}(x_o)$ is called the *probability density function*. Because all the probabilities that comprise it have to add up to one, we have

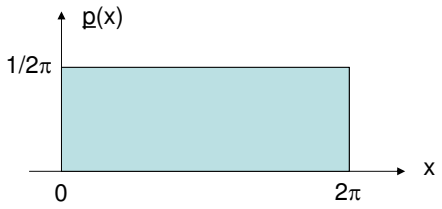
$$\int_{-\infty}^{\infty} \underline{p}(x)dx = 1.$$

With this definition, we can calculate the mean of the variable x and of a function of the variable $f(x)$:

$$\begin{aligned} E[x] &= \int_{-\infty}^{\infty} x\underline{p}(x)dx, \\ E[f(x)] &= \int_{-\infty}^{\infty} f(x)\underline{p}(x)dx. \end{aligned}$$

Here are a few examples. Consider a random variable that is equally likely to occur at any value between zero and 2π. Considering the area under $\underline{p}$ has to be one, we know then that $\underline{p}(x) = 1/2\pi$ for $x = [0, 2\pi]$ and it is zero everywhere else.

$$\begin{aligned} E(x) &= \pi \\ \sigma^2(x) &= \pi^2/3 \\ \sigma(x) &= \pi/\sqrt{3} \\ E(\cos x) &= \int_0^{2\pi} \frac{1}{2\pi} \cos x \; dx = 0 \\ E(\cos^2 x) &= \frac{1}{2}. \end{aligned}$$



The earlier concept of conditional probability carries over to random variables. For instance, considering this same example we can write

$$\begin{aligned} E\left[x|x>\pi\right] &= \int_0^{2\pi} x\underline{p}(x|x>\pi)\,dx \\ &= \int_\pi^{2\pi} x\frac{\underline{p}(x)}{p(x>\pi)}dx = \frac{3\pi}{2}. \end{aligned}$$

The denominator in the integral inflates the original pdf by a factor of two, and the limits of integration cause only values of x in the range of interest to be used.

3.6 The Gaussian PDF

The normal or Gaussian pdf is one of the most popular distributions for describing random variables, partly because many physical systems do exhibit Gaussian variability, and partly because the Gaussian pdf is amenable to some very powerful tools in design and analysis. It is

$$\underline{p}(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{(x-\bar{x})^2/2\sigma^2},$$

where σ and σ^2 are the standard deviation and variance, respectively, and $\bar{x}$ is the mean value. By design, this pdf always has area one. The cumulative probability function is

$$\begin{aligned} P(x) &= \frac{1}{2} + \mathrm{erf}\left(\frac{x-\bar{x}}{\sigma}\right), \text{ where} \\ \mathrm{erf}(\xi) &= \frac{1}{\sqrt{2\pi}}\int_0^\xi e^{-\xi^2/2}d\xi. \end{aligned}$$

Don't try to compute the *error function* erf(); look it up in a table or call a subroutine! The Guassian distribution has a shorthand: $N(\bar{x},\sigma^2)$. The arguments are the mean and variance.

3.7 The Cumulative Probability Function

The *cumulative probability function* is closely related to the pdf $\underline{p}(x)$:

$$\begin{aligned} P(x_o) &= p(x\le x_o) = \int_{-\infty}^{x_o}\underline{p}(x)dx, \text{ so that} \\ \underline{p}(x_o) &= \frac{dP(x_o)}{dx}. \end{aligned}$$

The probability density function is the derivative of the cumulative probability function. P is important because it lets us now transform the complete pdf of a random variable into the pdf of a function of the random variable. Let us say $y = f(x)$; the key idea is that for a monotonic function $f(x)$ (monotonic means the function is either strictly increasing or strictly decreasing with x),

$$p(x \leq x_o) = p(y \leq y_o = f(x_o));$$

these probabilities are the same, although we will see some subtleties to do with multiple values if the function is not monotonic. Here is a first example: let $y = ax + b$. In the case that $a > 0$, then

$$ax + b \leq y_o \text{ when } x \leq \frac{y_o - b}{a} \longrightarrow$$
$$p(y \leq y_o) = \int_{-\infty}^{\frac{y_o - b}{a}} \underline{p}(x)\, dx.$$

The case when $a < 0$ has simply

$$p(y \leq y_o) = 1 - \int_{-\infty}^{\frac{y_o - b}{a}} \underline{p}(x)\, dx.$$

All that we have done here is modify the upper limit of integration, to take the function into account. Now suppose that $y < y_o$ or $y > y_o$ over several disjoint regions of x. This will be the case if $f(x)$ is not monotonic. An example is $y = x^2$, which for a given value of y_o clearly has two corresponding x_o's. We have

$$\begin{aligned} p(y \geq y_o) &= p(x \leq -\sqrt{y_o}) + p(x \geq \sqrt{y_o}), \text{ or equivalently} \\ p(y \leq y_o) &= 1 - p(x \leq -\sqrt{y_o}) - p(x \geq \sqrt{y_o}) \end{aligned}$$

and there is of course no solution if $y_o < 0$. The use of pdf's for making these calculations, first in the case of monotonic $f(x)$, goes like this:

$$\begin{aligned} \underline{p}(y)|dy| &= \underline{p}(x)|dx|, \text{ so that} \\ \underline{p}(y) &= \underline{p}(x) / \left|\frac{dy}{dx}\right|. \end{aligned}$$

In the case of non-monotonic $f(x)$, a given value of y corresponds with $x_1, \cdots, x_n$. The correct extension of the above is

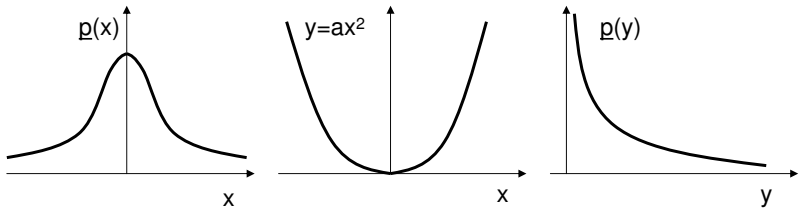
$$\underline{p}(y) = \underline{p}(x_1) / \left|\frac{dy(x_1)}{dx}\right| + \cdots + \underline{p}(x_n) / \left|\frac{dy(x_n)}{dx}\right|.$$

Here is a more detailed example. Consider the Gaussian or normal distribution $N(0, \sigma^2)$:

$$\underline{p}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2},$$

and let $y = ax^2$. For a given (positive) y, there are two solutions for x:

$$x_1 = -\sqrt{\frac{y}{a}}, \quad x_2 = \sqrt{\frac{y}{a}}.$$



Now $dy/dx = 2ax$ so that

$$\begin{aligned}
\left|\frac{dy(x_1)}{dx}\right| &= \left|\frac{dy(x_2)}{dx}\right| = 2a|x| = 2a\sqrt{\frac{y}{a}} = 2\sqrt{ay} \longrightarrow \\
\underline{p}(y) &= \underline{p}(x_1)/\left|\frac{dy(x_1)}{dx_1}\right| + \underline{p}(x_2)/\left|\frac{dy(x_2)}{dx_2}\right| \\
&= \frac{1}{\sigma\sqrt{2\pi}}\left\{\frac{1}{2\sqrt{ay}}e^{-y/2a\sigma^2} + \text{ same }\right\}, \text{ giving finally} \\
&= \frac{1}{\sigma\sqrt{2\pi ay}}e^{-y/2\sigma^2 a}.
\end{aligned}$$

3.8 Central Limit Theorem

A rather amazing property of random variables is captured in the central limit theorem; that a sum of random variables taken from distributions - even many different distributions - approaches a single Gaussian distribution as the number of samples gets large. To make this clear, let x_1 come from a distribution with mean $\bar{x}_1$ and variance σ_1^2, and so on up to x_n, where n is the number of samples. Let $y = \sum_{i=1}^{n} x_i$. As $n \to \infty$,

$$\begin{aligned}
\underline{p}(y) &= N(\bar{y}, \sigma_y^2), \text{ with} \\
\bar{y} &= \sum_{i=1}^{n} \bar{x}_i, \\
\sigma_y^2 &= \sum_{i=1}^{n} \sigma_i^2.
\end{aligned}$$

This is easy to verify numerically, and is at the heart of Monte Carlo simulation techniques. As a practical matter in using the theorem, it is important to remember that as the number of trials goes to infinity so will the variance, even if the mean does not (for example, if the underlying means are all zero). Taking more samples does not mean that the variance of the sum decreases, or even approaches any particular value.

2.017J Design of Electromechanical Robotic Systems
Fall 2009