Random Variables (Ghahramani 4.1)

In many random experiments we are interested in some function of the outcome rather than the actual outcome itself.

For instance, in tossing two dice (as in Monopoly) we may be interested in the sum of the two dice (e.g. 7) and not in the actual outcome (e.g. (1,6), (2,5), (3,4), (4,3), (5,2) or (6,1)).

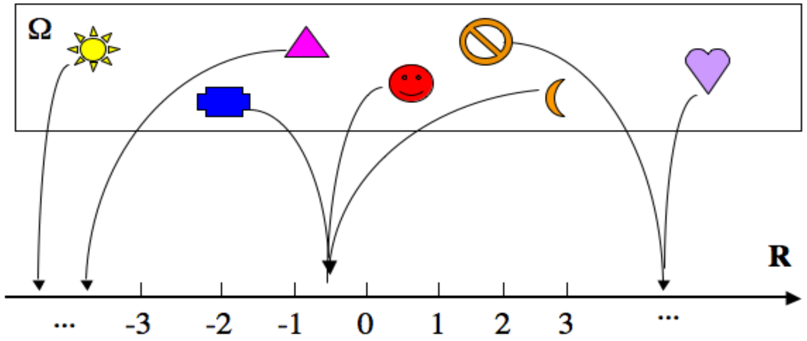
In these cases we wish to assign a real number x to each outcome ω in the sample space Ω. That is

$$x = X(\omega)$$

is the value of a *function* X from Ω to the real numbers $\mathbb{R}$.

Definition

Consider a random experiment with sample space Ω. A *function* X which assigns to every outcome $\omega \in \Omega$ a real number $X(\omega)$ is called a *random variable*.



NB. In more advanced courses (e.g., mast30020), there are some restrictions on the function X, but we won't worry about them here.

- The terminology “random variable” is unfortunate because X is neither random nor a variable. However it is universally accepted.

- It is standard to denote random variables by capital letters X, Y etc. and the values they take by lower case letters x, y etc.

- We shall denote the *set of possible values* (or *state space*) of X by $S_X \subseteq \mathbb{R}$ (this differs from the notation in Ghahramani)

Example

Suppose we toss two coins. The sample space is

$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}.$$

Let $X(\omega),\ \omega \in \Omega$ be the number of heads in ω. Then

$$\begin{aligned} X((H,H)) &= 2 \\ X((H,T)) &= X((T,H)) = 1 \\ X((T,T)) &= 0 \end{aligned}$$

Here $S_X = \{0,1,2\}$.

- X is not necessarily a 1-1 function. Different ω's may lead to the same value of $X(\omega)$, e.g. $X((H,T)) = X((T,H))$.
- The sets

$$\begin{aligned} A_2 &= \{\omega : X(\omega) = 2\} = \{(H,H)\} \\ A_1 &= \{\omega : X(\omega) = 1\} = \{(T,H),(H,T)\} \\ A_0 &= \{\omega : X(\omega) = 0\} = \{(T,T)\} \end{aligned}$$

are subsets of Ω and hence are *events* of the random experiment.

So we can see that a probability function defined on events leads to a distribution of probabilities across the possible values of the random variable. We formalise this as follows.

Definition: Consider a random experiment with sample space Ω. Let X be a random variable defined on Ω. Then, for $x \in S_X$ the probability that X is equal to x, denoted $\mathbb{P}(X = x)$, is the probability of the event $A_x := \{\omega : X(\omega) = x\}$ (in math, $:=$ means "the symbol at the LHS is defined as the RHS"). Thus

$$\mathbb{P}(X = x) = \mathbb{P}(A_x).$$

Consequently we can think of statements involving random variables as a form of shorthand eg

- $X = x$ for $\{\omega : X(\omega) = x\}$.
- $X \leq x$ for $\{\omega : X(\omega) \leq x\}$.
- $x < X \leq y$ for $\{\omega : x < X(\omega) \leq y\}$.

This shorthand reflects a shift in our interest from the random experiment as a whole $(\Omega, \mathbb{P})$ towards the distribution of the random variable of interest $(X, \mathbb{P}(X = x))$.

Example

Toss two dice. The sample space is

$$\Omega = \{(1,1), \ldots, (6,6)\}.$$

Let X denote the random variable whose value is the sum of the two faces. Assuming each outcome in Ω is equally likely,

$$\mathbb{P}(X = 2) = \mathbb{P}(\{\omega : X(\omega) = 2\}) = \mathbb{P}(\{(1,1)\}) = 1/36$$
$$\mathbb{P}(X = 3) = \mathbb{P}(\{\omega : X(\omega) = 3\}) = \mathbb{P}(\{(1,2),(2,1)\}) = 2/36$$
$$\mathbb{P}(X = 4) = \mathbb{P}(\{\omega : X(\omega) = 4\}) = \mathbb{P}(\{(1,3),(2,2),(3,1)\}) = 3/36$$

etc. Of course other random variables may be of interest eg the minimum number or the maximum.

Definition

A set is said to be *countable* if it is either finite or can be put into a 1-1 correspondence with the set of natural numbers $\{1, 2, 3, \ldots\}$. That is, a set is countable if it is possible to list its elements in the form $x_1, x_2, \ldots$. Otherwise a set is *uncountable*.

$$\begin{aligned} \mathbb{N} &= \{1, 2, \ldots\} \\ \mathbb{Z} &= \{0, 1, -1, 2, -2, \ldots\} \\ \mathbb{Z} \times \mathbb{Z} &= \{(0,0), (0,1), (1,0), (0,-1), (-1,0), \ldots\} \end{aligned}$$

are all countable sets.

It is known, via a very elegant proof, that $[0, 1]$ and $\mathbb{R}$ are uncountable.

Discrete Random Variables (Ghahramani 4.3, 4.2)

A *discrete random variable* is one for which the set of possible values S_X is countable. That is, X can take only a countable number of values.

Definition

Let X be a discrete random variable. The *probability mass function (pmf)* $p_X(x)$ of X is the function from S_X to $[0,1]$ defined by

$$p_X(x) = \mathbb{P}(X = x).$$

You can think of the $p_X(x)$ as discrete *masses* of probability assigned to each possible $x \in S_X$.

In the above, x is a dummy variable: we could use t or ζ or ξ or anything else. However, it is common to use x as a reminder that X is the random variable, and if it is clear that the pmf of X is intended, the subscript X may then be omitted.

We talk about the *probability mass function (pmf)* determining the *probability distribution* (or just *distribution* for short) of the discrete random variable X.

Note in Ghahramani, the *pmf* is defined on the domain $\mathbb{R}$, but as $p_X(x) = 0$ for all $x \notin S_X$ we prefer to restrict the domain to S_X.

Example

Let X be the sum of the numbers shown on the toss of two fair dice. Then $S_X = \{2, \ldots, 12\}$ and $p(x)$ is given by

$$\begin{array}{lcl lcl}
p(2) & = & \mathbb{P}(X = 2) = 1/36, & p(8) & = & 5/36, \\
p(3) & = & \mathbb{P}(X = 3) = 2/36, & p(9) & = & 4/36, \\
p(4) & = & 3/36, & p(10) & = & 3/36, \\
p(5) & = & 4/36, & p(11) & = & 2/36, \\
p(6) & = & 5/36, & p(12) & = & 1/36. \\
p(7) & = & 6/36, & & &
\end{array}$$

Theorem : The *probability mass function* $p_X(x)$ of a discrete random variable X satisfies the following

1. $p_X(x) \geq 0, \quad \forall\, x.$

2. $$\sum_{x \in S_X} p_X(x) = 1.$$

Indeed any function satisfying (1) and (2) can be thought of as the pmf for some random variable.

Proof

Part (1) is obvious as:

$$\begin{aligned} p(x) &= \mathbb{P}(X = x) \\ &= \mathbb{P}(\{\omega : X(\omega) = x\}) \end{aligned}$$

and $0 \leq \mathbb{P}(A) \leq 1$ for all events A.

For (2) first note that for $x_1 \neq x_2$, the events

$$\begin{aligned} & \{\omega : X(\omega) = x_1\} \\ \text{and} \qquad & \{\omega : X(\omega) = x_2\} \end{aligned}$$

are disjoint. So

$$\begin{aligned} \mathbb{P}(X = x_1 \text{ or } x_2) &= \mathbb{P}(\{\omega : X(\omega) = x_1 \text{ or } x_2\}) \\ &= \mathbb{P}(\{\omega : X(\omega) = x_1\} \cup \{\omega : X(\omega) = x_2\}) \\ &= \mathbb{P}(\{\omega : X(\omega) = x_1\}) + \mathbb{P}(\{\omega : X(\omega) = x_2\}) \\ &= \mathbb{P}(X = x_1) + \mathbb{P}(X = x_2). \end{aligned}$$

As S_X is the set of possible values of $X(\omega)$ for $\omega \in \Omega$, it follows that

$$\mathbb{P}(\{\omega : X(\omega) \in S_X\}) = \mathbb{P}(\Omega) = 1$$

but also

$$\begin{aligned}
\mathbb{P}(\{\omega : X(\omega) \in S_X\}) &= \mathbb{P}\left(\bigcup_{x \in S_X} \{\omega : X(\omega) = x\}\right) \\
&= \sum_{x \in S_X} \mathbb{P}(\{\omega : X(\omega) = x\}) \\
&= \sum_{x \in S_X} \mathbb{P}(X = x)
\end{aligned}$$

Hence

$$\sum_{x \in S_X} \mathbb{P}(X = x) = 1. \qquad \blacksquare$$

From the proof we can see that, for any set $B \subseteq \mathbb{R}$, given the *pmf*, we can compute the probability that $X \in B$ via

$$\mathbb{P}(X \in B) = \sum_{x \in B \cap S_X} p_X(x).$$

In particular

$$\mathbb{P}(X \leq x) = \sum_{y \leq x} p_X(y).$$

Example

Suppose that the discrete random variable X has pmf given by:

x	1	2	3	4	5
$p_X(x)$	α	2α	3α	4α	5α

Calculate α and $\mathbb{P}(2 \leq X \leq 4)$.

Sol

$$\sum p_X(x) = 1 \;\Rightarrow\; \alpha+2\alpha+3\alpha+4\alpha+5\alpha = 15\alpha = 1 \;\Rightarrow\; \alpha = 1/15$$

$$\mathbb{P}(2 \leq X \leq 4) = 2/15 + 3/15 + 4/15 = 9/15.$$

Distribution function (Ghahramani 4.2)

Definition: Let X be a random variable. The *distribution function* $F_X(x)$ of X is the function from $\mathbb{R}$ to $[0,1]$ defined by

$$F_X(x) = \mathbb{P}(X \leq x), \; x \in \mathbb{R}.$$

Particularly in the statistical literature, the distribution function is sometimes referred to as the *cumulative distribution function (Cdf)*.

Example (cont)

Derive the cdf of the random variable given in slide 101.

Properties of the distribution function (Ghahramani 4.2)

1. $0 \leq F_X(x) \leq 1$, since it is a probability.
2. $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$, if $a < b$ since:

   $$\{X \leq a\} \cup \{a < X \leq b\} = \{X \leq b\}$$

   and the events on the LHS are mutually disjoint.
3. $F_X(x)$ is non-decreasing. This follows from Property 2, since if $b > a$, then $F_X(b) - F_X(a) = \mathbb{P}(a < X \leq b) \geq 0$.

4. $F_X(-\infty) = 0, F_X(\infty) = 1$. In fact, let $A_n = \{X \leq -n\}$, then $A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots$ and $B := \cap_n A_n = \emptyset$. Since F_X is non-decreasing, using Property 10 on Slide 30,

$$F_X(-\infty) = \lim_{n\to\infty} F_X(-n) = \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(B) = 0.$$

Likewise, let $A_n = \{X \leq n\}$, then $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$ and $B := \cup_n A_n = \Omega$, so

$$F_X(\infty) = \lim_{n\to\infty} F_X(n) = \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(B) = 1.$$

5. $F_X(\cdot)$ is continuous on the right, that is $\lim_{h\downarrow 0} F_X(x+h) = F_X(x)$. If $h > 0$, then

$$[F_X(x+h) - F_X(x)] = \mathbb{P}(x < X \leq x+h).$$

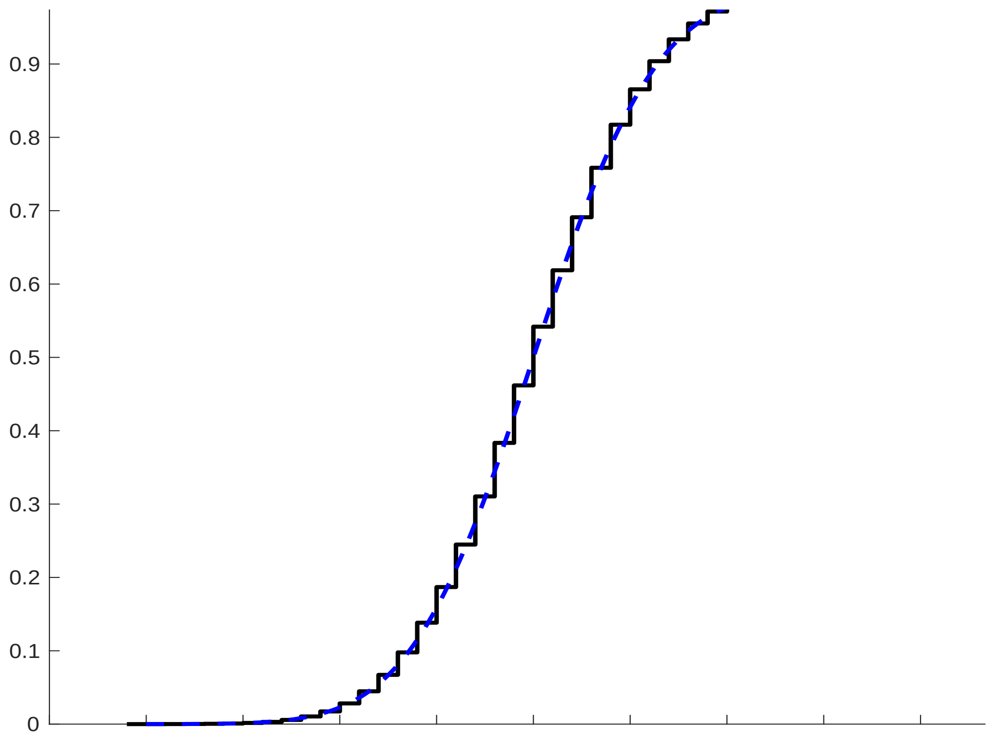
Since F_X is non-decreasing, using Property 10 on Slide 30, as $n \to \infty$, $h = \frac{1}{n} \downarrow 0$, $A_n := \{x < X \leq x + \frac{1}{n}\}$ satisfy $A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$ and $\cap_{n=1}^{\infty} A_n = \emptyset$ so the probability on the right hand side approaches zero.

6. $\mathbb{P}(X = x)$ is the jump in F_X at x. That is $\mathbb{P}(X = x) = F_X(x) - \lim_{h \downarrow 0} F_X(x - h)$. Again if $h > 0$, then

$$F_X(x) - F_X(x - h) = \mathbb{P}(x - h < X \leq x).$$

As $h = \frac{1}{n} \downarrow 0$, $A_n := \{x - \frac{1}{n} < X \leq x\}$ satisfy $A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots$ and $\cap_{n=1}^{\infty} A_n = \{X = x\}$.

Large number of values with small prob



When there are many values with small prob, the CDF can be approximated by a continuous curve, here shown in blue

Continuous random variables (Ghahramani 6.1, 4.2)

- If the cdf F_X of a random variable X is continuous, then we call X a *continuous random variable*. In this case,
  - The state space S_X is *uncountable*.
  - Assigning probability masses directly to any possible value is useless because $\mathbb{P}(X = x) = 0$ for all x!
  - We deal with this by assigning probabilities to intervals.

Probability density function

Let X be a continuous random variable. We say that X *has a density*, if there is a function $f_X(x)$ defined on $\mathbb{R}$, such that

$$\int_{-\infty}^{x} f_X(y)dy = F_X(x) \quad \text{for all } x \in \mathbb{R}.$$

The function $f_X(x)$ is called the *probability density function (pdf)* of X.

Remarks

- If a density function exists, then it is unique:

  $$f_X(x) = \frac{dF_X(x)}{dx}.$$

  Note. The above relation is only true for ”almost every” $x \in \mathbb{R}$ unless $f_X(x)$ is a continuous function (very deep theoretical point).

- It is not true that every continuous random variable has a density (even deeper theoretical point).

Properties of the pdf

1. $f_X(x) \geq 0$ since $F_X(x)$ is non-decreasing.
2. $\int_a^b f_X(t)dt = F_X(b) - F_X(a) = \mathbb{P}(a < X \leq b)$, i.e., probability is represented by the area under the graph of $f_X(x)$. For a random variable that has a density function

   $$\begin{aligned} \mathbb{P}(a < X \leq b) &= \mathbb{P}(a \leq X < b) \\ = \mathbb{P}(a \leq X \leq b) &= \mathbb{P}(a < X < b) \end{aligned}$$

   since the end points have zero probability.
3. $\int_{-\infty}^{\infty} f_X(t)dt = 1$ since $F_X(\infty) = 1$ and $F_X(-\infty) = 0$.

Note. Any function that satisfies Properties 1 and 3 is the pdf of some random variable.

Discrete vs Continuous

Discrete	Continuous
pmf $\quad p_X(x)$	*pdf* $\quad f_X(x)$
prob. masses $p_X(x)$ at x	no positive masses $p_X(x) = 0 \quad \forall\, x$
$\sum_{x \in S_X} p_X(x) = 1$	$\int_{-\infty}^{\infty} f_X(t)dt = 1$
$\mathbb{P}(X \in I) = \sum_{x \in I} p_X(x)$	$\mathbb{P}(X \in I) = \int_a^b f_X(t)dt$
$0 \leq p_X(x) \leq 1$	$f_X(x) \geq 0$

where $I = [a, b]$

There is no need to have $f_X(x) \leq 1$ since areas, not the value of the density function, represent probabilities. Thus, for example

$$f_X(x) = \begin{cases} 10^6 & 0 \leq x \leq 10^{-6} \\ 0 & \text{otherwise} \end{cases}$$

is a pdf since $f_X(x) \geq 0$ and $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

Pdf interpretation

Whilst the value of the *pdf* $f_X(x)$ is not the probability at x it can be interpreted as a probability density “around” x.

Define “$\{X \approx x\}$” to mean “$\{x - \frac{1}{2}\delta < X \leq x + \frac{1}{2}\delta\}$”. Then

$$\begin{aligned}
\mathbb{P}(X \approx x) &= \mathbb{P}\left(x - \frac{1}{2}\delta < X \leq x + \frac{1}{2}\delta\right) \\
&= \int_{x-\frac{1}{2}\delta}^{x+\frac{1}{2}\delta} f_X(u)du \\
&\approx f_X(x)\delta.
\end{aligned}$$

So we have $\mathbb{P}(X \approx x) \approx f_X(x)\delta$.

The Story So Far

It is important to make sure that you are fully aware of the subtle distinctions between probability functions, random variables, probability mass and density functions and cumulative distribution functions. Now that we have seen them all, we will take a moment to review them and point out the differences. Consider a random experiment with sample space Ω.

Then

1. The probability function $\mathbb{P}$ maps the set of events (i.e. subsets of Ω) to $[0, 1]$.
2. A random variable X maps Ω to $\mathbb{R}$.
3. For a discrete random variable, the probability mass function maps the set of possible values S_X to $[0, 1]$.
4. The distribution function maps $\mathbb{R}$ to $[0, 1]$.
5. For a continuous random variable with density, the probability density function maps $\mathbb{R}$ to $[0, \infty)$.

We often talk about random variables and their probability mass and distribution functions without explicit reference to the underlying sample space. For example, in talking about an experiment in which a coin is tossed n times we may define the random variable X to be the number of heads that turns up, and then go on to talk about the probability mass and distribution functions of X (which are?).

This is an example of shorthand expression which mathematicians often use. However they only use it when they fully understand the situation. In a case such as that described above, it is understood that the underlying sample space is the set of sequences of H and T of length n without this fact having to be mentioned explicitly.

Expectation (Ghahramani 4.4)

The distribution function contains all the information about the likelihood of the values of a random variable. However, this information can be difficult to digest. Because of this, we often summarise the information by reducing it in some way.

The most common such measure is the expected value.

The concept of expectation first arose in gambling problems: Is a particular game a good investment? Consider the game where the winnings $\$W$ has pmf

w	-1	1	10
$\mathbb{P}(W = w)$	0.75	0.20	0.05

Is it worthwhile? If you played the game 1000 times, you would expect to lose $\$1$ about 750 times, to win $\$1$ about 200 times and to win $\$10$ about 50 times. Thus you will win about

$$\$\left(\frac{-1 \times 750 + 1 \times 200 + 10 \times 50}{1000}\right)$$

per game.

Your “expected winnings” are -5 cents per game. We say that the “expected value of W” is -0.05.

This gives an indication of the worth of the game: in the long run, you can expect to lose an average of about 5 cents per game.

Expectations of Discrete RV’s (Ghahramani 4.4)

Let X be a discrete random variable with possible values in the set S_X, and probability mass function $p_X(x)$.

The *expected value* or *expectation* or *mean* of X, denoted by $\mathbb{E}[X]$, is defined by

$$\mathbb{E}[X] = \sum_{x \in S_X} x p_X(x)$$

provided the sum on the right hand side converges absolutely. We often denote the expected value by μ or μ_X to be clear which random variable is involved.

Why absolutely convergence?

We toss a fair coin repeatedly until we get a head and let Y be the number of tosses needed to get the first head. The "reward" is $X = (-2)^Y$, find the pmf of X and its mean if exists.

Example

Find $\mathbb{E}[X]$ if X is the value of the upturned face after a toss of a fair die.

Solution $S_X = \{1, \cdots, 6\}$ and $p_X(1) = p_X(2) = \cdots = p_X(6) = 1/6$. Hence

$$\mathbb{E}[X] = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + \ldots + 6\left(\frac{1}{6}\right) = \frac{7}{2}.$$

Remarks

- $\mathbb{E}[X]$ is not necessarily a possible value of X. It isn't in the example. We can never get $7/2$ to show on the face of a die.

- If we toss a die n times and let X_i denote the result of the ith toss, then we would expect that

  $$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mathbb{E}[X]$$

  (we will elaborate on this in Slide 512). Thus, after a large number of tosses we expect the average of all the observed values of X to be close to $\mathbb{E}[X]$.

More generally, suppose any random experiment is repeated a large number of times, and the random variable X observed each time, then the average of the observed values should approximately equal $\mathbb{E}[X]$.

Another way to think of the expected value of a random variable is as the location of the “centre of mass” of its probability distribution.

Example

A manufacturer produces items of which 10% are defective and 90% are non-defective. If a defective item is produced the manufacturer loses $\$1$, while a non-defective item yields a profit of $\$5$. If X is the profit on a single item, find $\mathbb{E}[X]$.

For any given item the manufacturer will either lose $\$1$ or make $\$5$. The interpretation of $\mathbb{E}[X]$ is that if the manufacturer makes a lot of items he or she can expect to make an average $\$4.40$ per item.

Expectations of Continuous RV's (Ghahramani 6.3)

The definition of the expected value of a continuous random variable is analogous to that for a discrete random variable.

Let X be a continuous random variable with probability density function $f_X(x)$.

The *expected value* or *mean* of X, denoted by $\mathbb{E}[X]$ is defined by

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

provided the integral on the right hand side converges absolutely.

Example

If X has pdf $f_X(x) = cx^2(1-x) \quad (0 < x < 1)$, find c and $\mathbb{E}[X]$.

The connection with the definition of the expected value of a discrete random variable can be seen by approximating the integral with a Riemann sum. Assume we partition $\mathbb{R}$ into intervals $[x_i, x_i + \delta)$ of length δ. Then

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &\approx \sum_i x_i f_X(x_i)\delta \\ &\approx \sum_i x_i \mathbb{P}(x_i \leq X < x_i + \delta). \end{aligned}$$

Expectation of functions (Ghahramani 4.4, 6.3)

In many situations we are interested in calculating the expected value of a function $\psi(X)$ of a random variable X. We need an accounting trick to do this.

Example

A couple plans to have three children. Assume that a child is equally likely to be a boy or a girl, find the expected number of girls in the three children.

Method 1 Let X be the number of girls in the three children, then $p_X(0) = 1/8$, $p_X(1) = 3/8$, $p_X(2) = 3/8$, $p_X(3) = 1/8$ with

$$\mathbb{E}(X) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = 1.5.$$

Method 2 We write

$\Omega = \{bbb, bbg, bgb, gbb, ggb, gbg, bgg, ggg\}$, then

$$\begin{aligned}\mathbb{E}(X) &= X(bbb)\mathbb{P}(bbb) + X(bbg)\mathbb{P}(bbg) + X(bgb)\mathbb{P}(bgb) + X(gbb)\mathbb{P}(gbb) \\ &\quad + X(ggb)\mathbb{P}(ggb) + X(gbg)\mathbb{P}(gbg) + X(bgg)\mathbb{P}(bgg) \\ &\quad + X(ggg)\mathbb{P}(ggg) \\ &= \cdots = 1.5\end{aligned}$$

An accounting trick

If Ω is countable, then

$$\mathbb{E}(X) = \sum_{\text{all } \omega \in \Omega} X(\omega)\mathbb{P}(\omega).$$

NB If Ω is uncountable, the formula is still correct provided we replace $\sum$ with $\int$ if necessary.

Theorem

If X is a discrete random variable with the set of possible values S_X and probability mass function $p_X(x)$, then for any real-valued function ψ, we have

$$\mathbb{E}[\psi(X)] = \sum_{x \in S_X} \psi(x) p_X(x)$$

provided the sum converges absolutely.

Proof

$\psi(X)$ is a discrete rv, let $A_x = \{\omega : X(\omega) = x\}$ for $x \in S_X$, then by the accounting trick, ,

$$\begin{aligned}
\mathbb{E}[\psi(X)] &= \sum_{\text{all } \omega} \psi(X(\omega))\mathbb{P}(\omega) \\
&= \sum_{x \in S_X} \sum_{\omega \in A_x} \psi(X(\omega))\mathbb{P}(\omega) \\
&= \sum_{x \in S_X} \psi(x)\mathbb{P}(X = x) \\
&= \sum_{x \in S_X} \psi(x)p_X(x). \qquad \blacksquare
\end{aligned}$$

Example

Let's return to the toss of a fair die with X the number on the upturned face, find $\mathbb{E}[X^2]$ and $\mathbb{E}[X]^2$.

Theorem

If X is a continuous random variable with probability density function $f_X(x)$, then for any real-valued function ψ, we have

$$\mathbb{E}[\psi(X)] = \int_{\infty}^{\infty} \psi(x) f_X(x) dx$$

provided the integral converges absolutely.

Example

If X has pdf $f_X(x) = 2x \quad (0 < x < 1)$, find $\mathbb{E}(X)$ and $\mathbb{E}\left[\frac{1}{X}\right]$.

Sol:

Note that

$$\mathbb{E}\left[\frac{1}{X}\right] \neq \frac{1}{\mathbb{E}[X]} = \frac{3}{2}.$$

Generally, $\mathbb{E}[\psi(X)] \neq \psi(\mathbb{E}[X])$, with one important exception, when ψ is a linear function.

Theorem If X is a random variable and a and b are constants, then

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

Proof We do the discrete case. The continuous is similar.

$$\begin{aligned}
\mathbb{E}[aX + b] &= \sum_{x \in S_X} (ax + b) p_X(x) \\
&= \sum_{x \in S_X} ax p_X(x) + \sum_{x \in S_X} b p_X(x) \\
&= a \textstyle\sum_{x \in S_X} x p_X(x) + b \sum_{x \in S_X} p_X(x) \\
&= a\mathbb{E}[X] + b. \qquad \blacksquare
\end{aligned}$$

Variance (Ghahramani 4.5, 6.3)

The mean of the distribution of a rv is the centre of the distribution and an equally important measure is the "spread" of the distribution.

Definition The *variance* $V(X)$ or $\mathrm{Var}(X)$ of a random variable X is defined by

$$V(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

$V(X)$ measures the *consistency* of outcome – a small value of $V(X)$ implies that X is more often near $\mathbb{E}[X]$, whereas a large value of $V(X)$ means that X varies around $\mathbb{E}[X]$ quite a lot.

Example

Consider the batting performance of two cricketers, one of whom hits a century (exactly 100) with probability $1/2$ or gets a duck (0) with probability $1/2$. The other scores 50 every time.

Let X_1 be the random variable giving number of runs scored by the first batsman and X_2 the number of runs scored by the second batsman. Then

$$\begin{aligned} \mathbb{E}[X_1] &= \tfrac{1}{2} \times 0 + \tfrac{1}{2} \times 100 = 50 \\ \mathbb{E}[X_2] &= 1 \times 50 = 50. \end{aligned}$$

However

$$\begin{aligned} V(X_1) &= \tfrac{1}{2}(0-50)^2 + \tfrac{1}{2}(100-50)^2 \\ &= \frac{1}{2}\,(2500) + \frac{1}{2}\,(2500) \\ &= 2500 \\ V(X_2) &= 1 \cdot (50-50)^2 = 0. \end{aligned}$$

This reflects the fact that the second player is more consistent.

From the definition of the variance we can see that the more widespread the likely values of X, the larger the likely values of $(X - \mu)^2$ and hence the larger the value of $V(X)$.

This is why the variance is a measure of spread. We often denote the variance by σ^2, or σ_X^2 to be clear which random variable is involved.

The square root of $V(X)$ is called the *standard deviation* and is denoted by σ_X, $sd(X)$ or just σ if the random variable involved is clear. As the units of the standard deviation and the random variable are the same, spread is often measured in standard deviation units.

There are alternative measures of spread.

For example the *mean deviation*, $d = \mathbb{E}(|X - \mu|)$. However for various mathematical reasons the variance (and standard deviation) are by far the most frequently used.

Notes on variance

1. $V(X) \geq 0$ since $(X - \mu)^2 \geq 0$.
2. $V(X) = 0 \;\Leftrightarrow\; \mathbb{P}(X = \mu) = 1$.
3. $V(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. This form is almost always the best to use when evaluating $V(X)$.
4. If $Y = aX + b$, then $V(Y) = a^2 V(X)$ and $sd(Y) = |a|\; sd(X)$.
5. If X has mean μ and variance σ^2, then $X_s = \frac{X-\mu}{\sigma}$ has mean 0 and variance 1. X_s is called a standardised random variable.
6. The mean and variance do not determine the distribution - they just give some idea of the centre and spread.

Theorem

$V(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$

Proof: Let $\mu = \mathbb{E}[X]$, then

$$\begin{aligned} V(X) &= \mathbb{E}[(X-\mu)^2] \\ &= \mathbb{E}[X^2 - 2X\mu + \mu^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X\mu] + \mathbb{E}[\mu^2] \\ &= \mathbb{E}[X^2] - 2\mu\mu + \mu^2 \\ &= \mathbb{E}[X^2] - 2\mu^2 + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2. \quad \blacksquare \end{aligned}$$

Example

Calculate $V(X)$ where X represents the roll of a fair die.

Solution We saw before that

$$\begin{aligned}
\mathbb{E}[X] &= \frac{7}{2} \\
\mathbb{E}[X^2] &= \frac{91}{6} \\
\text{and so } V(X) &= \frac{91}{6} - \frac{49}{4} = \frac{35}{12}.
\end{aligned}$$

Higher moments of a random variable (Ghahramani 4.5, 11.1)

The k^{th} moment (about the origin) of a random variable X is given by $\mu_k = \mathbb{E}(X^k)$.

The k^{th} central moment (about the mean) of a random variable X is given by $\nu_k = \mathbb{E}\big((X - \mu)^k\big)$.

So the mean $\mathbb{E}(X)$ is the first moment of X and the variance $V(X)$ is the second central moment of X.

Computing moments via tail probabilities

Suppose that $\mathbb{P}(X \geq 0) = 1$. Then for each $n \geqslant 1$, we have

$$\mathbb{E}[X^n] = n \int_0^\infty x^{n-1}[1 - F_X(x)]dx.$$

In particular,

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)dx.$$

Proof. The proof of the formula requires more advanced tools. We only give a proof when X is bounded by a constant $M > 0$ (namely $0 \leqslant X \leqslant M$) and at the same time X has a pdf $f_X(x)$. Note that in this case $f_X(x) = 0$ when $x > M$.

$$\int_0^\infty nx^{n-1}(1 - F_X(x))dx$$

$$= \int_0^M nx^{n-1}(1 - F_X(x))dx \quad (F_X(x) = 1 \text{ when } x > M)$$

$$\int_0^M (1 - F_X(x))d(x^n)$$

$$= x^n(1 - F_X(x))|_0^M - \int_0^M x^n(1 - F_X(x))'dx \quad \text{(integration by parts)}$$

$$= \int_0^M x^n f_X(x)dx$$

$$= \int_0^\infty x^n f_X(x)dx \quad (f_X(x) = 0 \text{ when } x > M).$$

Example

Let X be a random variable with pdf

$$f(x) = \begin{cases} 2(x+1)^{-3}, & \text{if } x \geq 0, \\ 0, & \text{else.} \end{cases}$$

Compute $\mathbb{E}[X]$.

Sol. We have

$$F(x) = \begin{cases} 1-(x+1)^{-2}, & \text{if } x \geq 0, \\ 0, & \text{else,} \end{cases}$$

so

$$\int_0^\infty (1-F(x))dx = \int_0^\infty (x+1)^{-2}dx = -(x+1)^{-1}\Big|_0^\infty = 1 < \infty.$$

Using the formula in Slide 150, we obtain $\mathbb{E}[X] = 1$.

St. Petersburg Paradox (Ctd)

- Toss a fair coin, bet on Tail.
- Bet \$1 in the first toss. End game if I win.
- If I lose, bet \$2 in the second toss. End game if I win.
- If I lose again, bet \$4 in the third toss. End game if I win.
- Keep playing until first Tail appears.
- Should I play this game?

Define

$$\begin{aligned} N &: \text{number of games played up to first win} \\ L &: \text{total loss} \\ W &: \text{total winning} \end{aligned}$$

Note that L and W are the functions of N:

$$\begin{aligned} L &= 1 + 2^1 + 2^2 + \cdots + 2^{N-2} = 2^{N-1} - 1, \\ W &= 2^{N-1}. \end{aligned}$$

At first glance, this is a stupid game that everyone should play because

$$W - L = 1.$$

Let's compute the expected loss. First note that

$$\mathbb{P}(N = n) = \left(\frac{1}{2}\right)^{n-1} \times \frac{1}{2} = \frac{1}{2^n}.$$

Therefore,

$$\begin{aligned}\mathbb{E}[L] &= \sum_{n=1}^{\infty}(2^{n-1} - 1) \times \mathbb{P}(N = n) = \sum_{n=1}^{\infty}(2^{n-1} - 1) \times \frac{1}{2^n} \\ &= \sum_{n=1}^{\infty}\left(\frac{1}{2} - \frac{1}{2^n}\right) = +\infty.\end{aligned}$$

You are expected to lose $+\infty$ before winning $+\infty + \$1$!