# 18 Continuous probability and normal distribution

## 18.1 Discrete and continuous distributions

Binomial and Poisson[1] distributions are examples of discrete probability distributions, where the sample space is a discrete set of points, and individual points have positive probability. There are also continuous distributions, where the probability of any particular point is zero.

## 18.2 Uniform distribution on on $[a, b]$, pdfs, random variables, and expectation

The uniform distribution on the interval $[a, b]$

$$[a, b] = \{x : a \leqslant x \leqslant b\}$$

There are infinitely many points, so each point has probability zero. One must use *integration* to compute probabilities. Events are sets of points.

**Example.** Let $X$ be a random variable uniformly distributed on $[0, 1]$. Let $E$ be the event: $x \leqslant 0.1 \vee x \geqslant 0.6$. In other words,

$$E = [0, 0.1] \cup [0.6, 1]$$

For this specific distribution on this specific interval, if an event $F$ is a single interval then $P(F)$ (Prob$(F)$) is its length. The intervals are disjoint. Hence $P(E) = 0.1 + 0.4 = 0.5$.

**More generally,** a continuous probability distribution is defined by a 'probability density function' $p(x)$ (defined on $\mathbb{R}$ or a subset of $\mathbb{R}$), and the probability of an interval $[A, B]$ is

$$\int_A^B p(x)dx.$$

**Notation: PDFs** A 'pdf' is short for 'probability density function.'

The uniform distribution on $[a, b]$ has constant probability

$$p(x) = \frac{1}{b-a}$$

and this can be evaluated on disjoint unions of intervals.

One can define random variables and expectation in connection with continuous distributions. Expectation is defined by integration. So the simple formula (see earlier sections)

$$E(Y) = \sum_i p_i f(x_i)$$

becomes

$$E(Y) = \int p(x) f(x) dx.$$

The *mean* is $\mu = \int x p(x) dx$ and *variance* is $\sigma^2 = \int (x - \mu)^2 p(x) dx$.

[1]The Poisson distribution is not in this module.

**Mean** of uniform distribution on $[a, b]$ is

$$\int_a^b \frac{x}{b-a} dx = \frac{a+b}{2},$$

and its variance is

$$\int_a^b \left( \frac{x - \frac{a+b}{2}}{b-a} \right)^2 dx = \int_{\frac{a-b}{2}}^{\frac{b-a}{2}} \frac{x^2 \, dx}{b-a} = \frac{(b-a)^2}{12}$$

## 18.3 Distribution of $X_1 + X_2$

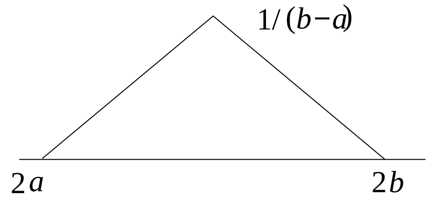
**(18.1) Definition** *A list $X_1, \ldots, X_n$ of random variables on the same sample space is* independent and identically distributed *if they all have the same distribution and,* **roughly speaking,** *if all of them are sampled, the outcome of sampling any one of them has no effect on the other outcomes.*

*Abbreviation:*

**iid**

Suppose $X_1$ and $X_2$ are independent and both uniformly distributed on $[a, b]$. What is the distribution of $X_1 + X_2$?

**Answer** linear ascending, $2a \leqslant x \leqslant a + b$, and then descending $a + b \leqslant x \leqslant 2b$.




In other words: the sum of two independent random variables uninformly distributed in $[a, b]$ has sample space $[2a, 2b]$ and is *not uniformly distributed.*

The distribution for

$$\frac{X_1 + X_2}{2},$$

the *sample average* with two uniformly distributed iid random variables, is got by scaling the above — the range is scaled by 1/2 and the height scaled by 2. So it is certainly *not* uniformly distributed. See Figure 1.

The distribution for

$$\frac{X_1 + X_2 + X_3 + X_4}{4}$$

the average of *4* has the distribution illustrated. Actually, it is made up of 4 cubic curves.

Notice that even for this small number of random variables, the pdf of the average already resembles a bell curve (normal distribution).

## 18.4 Normal distribution

**The most important distribution** is the *Gaussian* or *normal* distribution, written

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

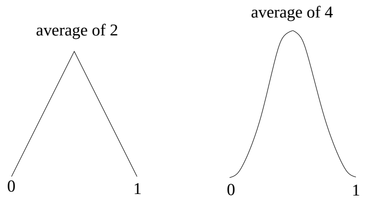



Figure 1: Sample average of 2 and 4 iid uniformly distributed random variables

In the formulae, we use $\exp(\ldots)$ rather than $e^{\cdots}$ because exponents are hard to read. The Probability Density function, abbreviated

**PDF** or **pdf**

is

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x-\mu)^2}{2\sigma^2}$$

**(18.2) Proposition** *$N(\mu, \sigma^2)$ has total probability 1 (as is required). $N(\mu, \sigma^2)$ has mean $\mu$ and variance $\sigma^2$.*

*(It is hard to prove that the total probability is 1, the others are probably easier.)* ∎

The total probability is 1, based on the curious fact that

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}.$$

Again, **events** are *intervals of real numbers* or unions of intervals. The probability of a single interval $[a, b]$ is $\int_a^b p(x)dx$.

There is no way to simplify this integral in terms of log, exp, or trig functions. Its values are given in statistical tables.

**The normal distribution** is the most important. For example, in taking measurements in a laboratory, the error distribution will be normal or close to normal.

Unlike the uniform distribution, the average of normally distributed random variables is again normally distributed:

**(18.3) Lemma** *If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ random variables, then $\overline{X}$ has the distribution*

$$N(\mu, \frac{\sigma^2}{n}). \quad ∎$$

There is a very striking way in which the normal distribution is the limit to which all distributions tend:

**(18.4) Theorem (Central Limit Theorem.)** *If $X_1, \ldots, X_n$ are iid (independent identically distributed) random variables with mean $\mu$ and variance $\sigma^2$, and $\overline{X}$ is the sample average, then for large values of $n$, the probability distribution of*

$$\sqrt{n}(\overline{X} - \mu)$$

*converges to $N(0, \sigma^2)$.* ∎

**(18.5) Corollary** *Given sample* $X_1, \ldots, X_n$ *of iid random variables (any distribution, with mean* $\mu$*), if* $n$ *is large 'in some sense' then* $\overline{X}$ *is almost certainly almost exactly* $\mu$*, 'in some sense.'* $\blacksquare$