Random Variables

Contents

3 Random Variables

We will now consider the case of experiments with non-deterministic outcomes where the result of the experiment is measurable, that is, quantifiable. In such a case we are dealing with a variable which is random in nature. In the case of a Random Variable:

- The experiment is producing and then measuring the variable
- The event is the variable taking on a particular value or range of values.

We will see that some of the examples we have already looked at can be recast as random variables.

3.1.1 Rolling Two Dice – Revisit

In the case of rolling two dice, calculate the probability that the difference is 2. This can now be treated as a random variable; by calculating the difference, we are treating the symbols on the dice as numbers. In this case:

- The experiment is the rolling of the dice, seeing which numbers come up and then finding the difference.
- The outcomes are whatever numbers show up.
- The Random variable is the difference of the two numbers, call it D.
- The event is $D = 2$.

From counting the possible ways in which this event comes about and dividing by the size of the sample space, exactly as before, the probability is

$$P[D = 2] = 8/36 = 2/9.$$

3.2 Discrete and Continuous Variables

We must first define two types of random variables, based on what values may arise as a result of the experiment.

3.2.1 Definition – Discrete Random Variable

A Random Variable is said to be *discrete* if it can only take on values from a finite or countably infinite set of distinct numbers, such as integers or a finite subset of numbers.

3.2.2 Definition – Continuous Random Variable

A Random Variable is said to be *continuous* if it can take on any real numbers within a certain range as a value.

Discrete random variables usually arise as the result of a counting process. An example might be 'the number of students who turn up for a lecture.' Continuous random variable are usually physical quantities, such as height, weight, distance, resistance, capacitance and times. When dealing with a continuous variable, we can no longer talk about the chances of the variable being equal to particular values; since there are an infinite number of possible values, the chances of one particular value coming up are zero. Instead we talk about the probability of the variable being in a particular range.

3.3 Some Discrete Distributions

Random variables usually have a distribution; this is a law governing the probability of particular values of the variable coming up. In the case of a discrete random variable, this could simply be a list of probabilities for each value. It can also be an equation giving the probability in terms of the value.

3.3.1 A Simple Example of a Distribution

Consider again the example of a random variable generated by rolling two dice and calculating the difference. This discreet random variable can only take on certain values, the integers 0 to 5. Recall we drew up the event space and wrote down the probability of all events:

$$P[D = 0] = 6/36 = 3/18.$$

$$P[D = 1] = 10/36 = 5/18.$$

$$P[D = 2] = 8/36 = 4/18.$$

$$P[D = 3] = 6/36 = 3/18.$$

$$P[D = 4] = 4/36 = 2/18.$$

$$P[D = 5] = 2/36 = 1/18.$$

3.3.2 Discrete Uniform Distribution

Here is fundamentally important but simple distribution. Let X be a random variable which takes on one of the k discrete values

$$1, 2, \ldots, k.$$

This list is in fact the sample space. This means that the simplest events involving this variable are of the form:

$$X = n, \text{ for each } n = 1, 2, \ldots, k.$$

Let us say that each value is equally likely to come up with a probability p:

$$P[X = n] = p, \text{ for each } n = 1, 2, \ldots, k.$$

This is the *discrete uniform distribution*. A simple application of the addition rule for probabilities gives p. Let U be the universal event, so that by definition

$$U = \{X = 1\} \text{ or } \{X = 2\} \ldots \text{ or } \{X = k\}.$$

Then since the probabilities are all p, then

$$\sum_{n=1}^{k} P[X = n] = kp,$$

giving $kp = 1$. An example we have already seen is that of a single dice, treating the faces as numbers, in which case $k = 6$ and so therefore $p = 1/6$.

3.3.3 The Probability Mass Function

Let X be a discrete random variable which takes on values from the countable set

$$U = \{x_i, i = 1, 2, \ldots\}.$$

We can assume that the index variable i has been assigned uniquely to each element of U, since it is countable. Let f be a function assigning a probability to each element of U. It is therefore a mapping from U to the interval [0, 1] :

$$f: U \rightarrow [0, 1], \text{ so that } P[X = x] = f(x)$$

This function is the probability mass function for variable X. This function will have several properties arising from its use assigning probabilities.

1. $\sum f(x_i) = 1$, where the summation takes place over every i.
2. If A is a subset of U we may regard the event A as being the occurrence that the variable X takes on a value x_i such that x_i is an element of A. Then

   $$P[A] = \sum f(x_i),$$

   where the summation takes place over every distinct i, such that x_i is in the set A.
3. If A and B are subsets of U and A is a distinct subset of B then, since we want $P[A] < P[B]$, then

   $$\sum f(x_i),\ x_i \text{ in } A, < \sum f(x_i)\ x_i \text{ in } B.$$

   This is the monotonicity property of the probability function expressed in the context of a random variable.

3.3.4 The Cumulative Distribution Function

Another, related, definition is that of the Cumulative Distribution function, defined as follows. Let X be a discrete random variable as described above. Let F be a function from U to the interval [0, 1] such that

$$F: U \to [0, 1], \text{ so that } P[X \le x] = F(x).$$

This function will have several properties:

1. Let a = min U and b = max U (these quantities are well defined since U is countable), then $F(x) = 0$ if $x < a$ and also $F(b) = 1$. In the limit where U is a countably infinite set of all the integers, $F(-\infty) = 1$ and $F(+\infty) = 1$.
2. If x and y are elements of U then if $x < y$

   $$F(x) < F(y).$$

   This is the monotonicity property again.
3. Since the elements of the set U are discrete, we can say that $F(x) = \sum f(y)$, if the summation takes place over all distinct values $y \le x$.

3.3.5 Another Discrete Random Variable

Consider the following random variable and its distribution. Let X be a variable which takes on any positive integer. The event space is therefore the set

$$\{1, 2, 3, \ldots \infty\}$$

The probability mass function is:

$$P[X = n] = f(n) = ½^n.$$

This means that the probability that the variable X takes on value n is simply ½ raised to that power so:

$$P[X = 1] = ½.$$

$$P[X = 2] = ¼, \text{ etc.}$$

The sum of all the probabilities is:

$$\sum_{n=1}^{\infty} P[X = n] = \sum_{n=1}^{\infty} \tfrac{1}{2}^n = \tfrac{1}{2}\frac{1}{1-\tfrac{1}{2}} = 1,$$

as the summation is a geometric series. The probability mass density function can be simply identified:

$$f(n) = ½^n.$$

The cumulative distribution function is then

$$F(n) = \sum_{i=1}^{n} \tfrac{1}{2}^n = \tfrac{1}{2}\frac{1-\tfrac{1}{2}^{n+1}}{1-\tfrac{1}{2}} = 1 - \tfrac{1}{2}^{n+1}.$$

3.3.6 The Bernoulli Trial

Consider the following case of a discrete distribution: let B be a random variable which takes on one of two discrete values

$$0, 1.$$

The value 1 can come up with a probability p:

$$P[B = 1] = p, \text{ so that } P[B = 0] = 1 - p.$$

This is the *Bernoulli trial*, so called because it is like a yes/no answer. The two functions are trivially identified:

$$f(0) = 1 - p, f(1) = p.$$

$$F(0) = 1 - p, F(0) = 1.$$

Although it is a very simple idea it will be very useful.

3.3.7 The Uniform Distribution

This distribution was introduced above; X is a random variable which takes on one of the k discrete values

$$1, 2, \ldots, k.$$

with equal probability p. So for $n = 1, 2, \ldots k$,

$$P[X = n] = p, \text{ where } pk = 1.$$

The probability mass function f of the variable is therefore

$$f(n) = p.$$

The CDF of this variable is found by working out the probability of the event $X \leq n$. This event is broken up as follows:

$$X \leq n \text{ is } X = 1 \text{ or } X = 2 \text{ etc. up to } X = n.$$

All these events are mutually exclusive and all have the same probability p. This means that the addition law of probability can be applied, which gives:

$$P[X \leq n] = np, \text{ so}$$

$$F(n) = np.$$

3.3.8 The Binomial Distribution

A trial is being repeated, with a possible result A. The following is known:

- Each time the trial is done, the probability of result A turning up is p.
- The trial is repeated n times.

Let X be the random variable of the number of times event A comes up; 'successes'. The probability of getting r results from n trials is:

$$P[X = r] = {}^{n}C_{r}p^{r}(1-p)^{n-r}.$$

This discrete random variable is the binomial distribution, so we can identify the probability mass function as

$$f(r) = 0 \text{ if } r < 0,$$
$$f(r) = {}^{n}C_{r}p^{r}(1-p)^{n-r},$$
$$f(r) = 0 \text{ if } r > n.$$

Using the binomial expansion ensures that

$$\sum_{r=0}^{n} f(r) = 1.$$

There is no useful closed form of the cumulative distribution function for the binomial distribution, it is simply

$$F(k) = P[X \le k] = \sum_{r=0}^{k} f(r) = \sum_{r=0}^{k} {}^{n}C_{r}p^{r}(1-p)^{n-r}.$$

The variable X may be regarded as the sum of n independent Bernoulli trials. Since the number r for variable X can only be produced if r of the Bernoulli trials come up as 1, this follows the same structure as the binomial concept. Therefore

$$X = \sum_{i=1}^{n} B_i .$$

This will be a very useful identification.

The example we studied before of the probability of 3 lefthanders in a group of 12 fits this distribution and the same logic shows where the equation for the distribution comes from:

- The trial being repeated is checking whether or not a person is left-handed.
- In the example we studied $n = 12$.
- The probability of this occurring for each individual 'tested' is $p = 0.11$. This number comes from the proportion of 0.11 of the wider population being left-handed and then from our definition of a probability.

Let L be the random variable of the number of left-handers in the group. The probability we are looking at is then:

$$P[L = 3].$$

Using the binomial distribution,

$$P[L = r] = {}^{n}C_{r}p^{r}(\,1 - p)^{n\,-\,r}.$$

$$P[L = 3] = {}^{12}C_{3}\;0.11^{3}\;0.89^{9} = 0.1026.$$

3.3.9 The Poisson Distribution

Consider a situation where a certain type of event is happening in time, and λ be the fixed (independent of the time) number of events happening per unit time. Let T be a certain interval of time. Let N be the random variable of the number of these events which occur in that time T. Then the random variable N has probability mass function

$$P[N = r] = e^{-\lambda T} \frac{(\lambda T)^r}{r!}.$$

This is usually written by defining a parameter μ as

$$\mu = \lambda T.$$

This number μ can be seen as the expected number of events in the time T. Then the equation defining this distribution is:

$$P[N = r] = e^{-\mu} \frac{\mu^r}{r!}.$$

Here is an example of a calculation for this distribution. Telephone calls are known to arrive in an exchange at a rate of 15 per minute. Calculate the probability that 3 calls come in a ten second period. If the calls are arriving at 15 per minute, then the rate (working with seconds as our unit if time) is $\lambda = ¼$. Then the parameter μ is given by

$$\mu = ¼\text{x}10 = 2.5.$$

Then the probability of getting 3 calls in this time period is:

$$P[N = 3] = e^{-\mu}\mu^n/n! = e^{-2.5}2.5^3/3! = 0.2135.$$

The Poisson distribution provides an approximation for calculating the binomial distribution if n is large, probability p is small and np is not large. If these conditions are met then the binomial probabilities can be written as Poisson probabilities, with $\mu = np$.

It is instructive to see where this distribution comes from. For the interval of time T, consider a partition into a large number n of segments, each of length Δt and so $n\Delta t = T$.

For each such segment, given the value of λ, assume n is large and Δt is small enough that that only one event occurs within the time interval Δt. The probability that a single event occurs in that segment will then be $\lambda\Delta t$. Then using the binomial distribution the probability we get r events in time T is approximately

$$P[N = r] = {}^nC_r(\lambda\Delta t)^r(\,1 - \lambda\Delta t)^{n-r}.$$

This approximation will converge on the true value as n increases without limit. We will need two results from mathematics, one from the definition of the exponential function:

$$\lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n = e^x.$$

The other result is the approximation of $n!$ as n gets very large:

$$n! \cong \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

This is known as Sterlings approximation. If this is used in the defining equation for the quantity ${}^nC_r$, we can produce the following expression:

$${}^nC_r = \frac{n!}{(n-r)!r!} \cong \sqrt{\frac{n}{n-r}}\left(\frac{n}{n-r}\right)^n \frac{e^{-r}(n-r)^r}{r!}.$$

Now substitute $\lambda\Delta t = \lambda T/n$ and the above result for ${}^nC_r$ into the equation for the probability to get

$$P[N=r] \cong \sqrt{\frac{n}{n-r}}\left(\frac{n}{n-r}\right)^n \frac{e^{-r}(n-r)^r}{r!}\left(\frac{\lambda T}{n}\right)^r\left(1-\frac{\lambda T}{n}\right)^{n-r}.$$

Carrying out some algebra on each term gives

$$P[N=r] \cong \sqrt{\frac{n}{n-r}}\,\frac{1}{(1-\frac{r}{n})^n}\,\frac{e^{-r}}{r!}\left(\frac{n-r}{n-\lambda T}\right)^r(\lambda T)^r\left(1-\frac{\lambda T}{n}\right)^n.$$

Now let n increase without limit so that:

$$\lim_{n\to\infty}\left(1-\frac{\lambda T}{n}\right)^n = e^{-\lambda T},\ \lim_{n\to\infty}\left(1-\frac{r}{n}\right)^n = e^{-r},\ \lim_{n\to\infty}\left(\frac{n-r}{n-\lambda T}\right)^r = 1.$$

Then we have

$$P[N=r] = e^{-\lambda T}\frac{(\lambda T)^r}{r!}, \text{ as required.}$$

The mathematical analysis used here can also show us that the binomial distribution can be approximated by the Poisson when n is large and p is very small.

3.3.10 Expected Values and Variance

Let X be a random variable with probability density function f and universal event set U. The expected value of a variable is denoted $E[X]$ and is defined as follows.

$$E[X] = \sum_{x\in U} x f(x).$$

It is each possible value that the variable may produce multiplied by the probability of that variable coming up; it may be regarded as the concept of a mean of a variable. We can make the following link between the two. Let S_n be a subset of the universal event U found by taking n values from U. Then the mean of the values in the subset S_n is an estimate of $E[X]$. Furthermore,

$$\lim_{n\to\infty} \bar{x} = \lim_{n\to\infty} \frac{\sum x_i}{n} = E[X].$$

This is not a trivial result. Because of this identity, the expected value $E[X]$ is often denoted as μ.

A second quantity, the variance, is defined as follows; it is also denoted as σ^2 for similar reasons as mentioned for μ:

$$\sigma^2 = E[(X - E[X])^2] = E[(X-\mu)^2] = \sum_{x\in U} (x-\mu)^2 f(x).$$

Some algebra on this identity leads to:

$$\sigma^2 = E[X^2] - E[X]^2 = E[X^2] - \mu^2 = \sum_{x\in U} x^2 f(x) - \mu^2.$$

Again we can say that

$$\lim_{n\to\infty} s^2 = \lim_{n\to\infty} \frac{\sum x_i^{\,2} - n\bar{x}^2}{n} = \sigma^2.$$

The number σ is known as the standard deviation. We therefore say that if the numbers

$$x_1, x_2, \ldots x_n$$

are a set of n values generated by the random variable X, then we regard them as a sample of the variable and our figures $\bar{x}$ and s are *estimates* of the true mean and standard deviation respectively. Let us look back at some of the basic distributions and calculate a mean and standard deviation for each one.

- The Bernoulli trial; it is a straightforward matter to show that in this case, the parameters are:

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

- The discrete uniform distribution taking on values 1 to k:

$$\mu = (k+1)/2,\ \sigma^2 = (k^2-1)/12.$$

- The Poisson distribution

$$\mu = \lambda T,\ \sigma^2 = \lambda T.$$

  This is why the parameter $\mu = \lambda T$ is used in the Poisson.

- The binomial distribution:

$$\mu = np,\ \sigma^2 = np(1-p).$$

  These two results may be proved by using some algebraic 'trickery', but they also follow directly from the observation that the binomial distribution with parameters n and p is equivalent to the sum of n Bernoulli trials with parameter p.

- For the cases of the roll of two dice, where we treated the sum S or the difference D as random variables, we can simply evaluate the expression

$$E[X] = \sum_{x \in U} x f(x),$$

  by hand. For example, $E[D] = 35/18$.

3.4 Continuous Random Variables

We will now look at a few continuous distributions and look at the equivalent definitions for them.

3.4.1 Cumulative Distribution Function

Let X be a continuous random variable. Then, for a given value x, the function F defined by

$$F(x) = P[X \leq x]$$

is called the cumulative Distribution function for the variable. In practice, the function F is defining the variable, so we usually write

$$P[X \leq x] = F(x).$$

As an immediate consequence, we can write the probability of the variable X giving a value between two numbers a and b:

$$P[a \leq X \leq b] = F(b) - F(a).$$

To see why this is the case, apply the laws of probability to the three events:

$$X < a, X \leq b, \text{ and } a \leq X \leq b.$$

A list of properties can be drawn of for the Cumulative Distribution function similar to those for discrete variables. In particular, for a variable defined on a range R a subset of the real numbers into the interval [0, 1], the function F is a bijection:

$$F(x) = F(y) \text{ if and only if } x = y.$$

3.4.2 Probability Density Function

For a continuous variable, there can be no equivalent of the probability mass function. We look instead at the Probability Density

Function, the PDF. This is defined as follows: with variable X and function F defined as above, so that

$$F(x) = P[X \leq x],$$

then the Probability density function is the derivative of F with respect to x:

$$f(x) = \frac{\mathrm{d}F}{\mathrm{d}x}(x).$$

It therefore also follows that F is the integral of f; more precisely, the probability $F(x)$ is found by integrating the function f from negative infinity (or the appropriate lowest value the variable X takes on) to x:

$$P[X \leq x] = F(x) = \int_{-\infty}^{x} f(u)\mathrm{d}u.$$

The probability of X taking on a value between two numbers is then:

$$P[a \leq X \leq b] = F(x) = \int_{a}^{b} f(u)\mathrm{d}u.$$

With this idea, the PDF can be thought of as measuring the *relative* likelihood a variable takes on certain values. The density function gives us our definitions of expected values.

3.4.3 Expected Values: Continuous Variables

Let X be a continuous random variable with probability density function f. Let R be the range of values which the variable can take on. The expected value of X, denoted by $E[X]$ is given by the equation:

$$E[X] = \int_R uf(u)\mathrm{d}u\,.$$

This means we are multiplying each possible value of the variable X and multiplying it by its relative probability. Therefore we define the mean for a continuous random variable as

$$\mu = E[X] = \int_R uf(u)\mathrm{d}u.$$

In the same way, we can define a variance and so a standard deviation:

$$\sigma^2 = Var[X] = \int_R (u-\mu)^2 f(u)\mathrm{d}u.$$

A bit of algebra shows that

$$\sigma^2 = Var[X] = \int_R u^2 f(u)\mathrm{d}u - \mu^2.$$

We retain the same understanding from the discrete case, that if the numbers

$$x_1, x_2, \ldots x_n$$

are a set of n values generated by the continuous random variable X, then we regard them as a sample of the variable and our figures $\bar{x}$ and s are *estimates* of the true mean and standard deviation respectively. We will now look at some examples of continuous variables.

3.4.4 The Continuous Uniform Distribution

Let X be a random variable which takes a value between two real number a and b, with every value equality likely. It therefore follows that for x between a and b, the probability density function is the constant function

$$f(x) = p, \text{ where } p(b - a) = 1.$$

The cumulative distribution function is:

$$F(x) = P[X \leq x] = p(x - a).$$

Simple algebra gives us the results

$$\mu = E[X] = \tfrac{1}{2}(b + a) \text{ and } \sigma^2 = (b^2 - a^2)/12.$$

The continuous uniform distribution can be simulated reasonably well with software and is therefore vitally important for generating random numbers for a general distribution. Let U be a random variable with uniform distribution between 0 and 1, so that it has cumulative distribution function $F_U(x) = x$. Let X be a random variable with cumulative distribution function F_X. Consider the random variable $Y = F_X^{-1}(U)$; consider the event $Y < x$ and its probability:

$$P[Y < x].$$

Since F is a bijection, we can say that

$$P[Y < x] = P[\ F_X(Y) < F_X(x)\ ].$$

Since $F_X(Y) = U$ we can say that

$$P[Y < x] = P[\ U < F_X(x)\ ] = F_X(x).$$

Since we have found that

$$P[Y < x] = F_X(x),$$

it therefore follows that Y is identical to X.

This means that if random numbers can be generated for the variable U they can be used to provide random numbers for any variable.

3.4.5 The Exponential Distribution

Let X be a random variable which can take on any positive number, and let λ be a positive parameter. The variable X has the Exponential Distribution if the cumulative distribution function for X is:

$$P[X < x] = 1 - e^{-\lambda x}.$$

Note that for x being any positive number, the probability goes from 0 to 1, as we would expect. So the probability we get a number less than 0 is:

$$P[X < 0] = 1 - e^{-\lambda 0} = 0.$$

We can also see that as x gets larger and larger, the probability gets closer to 1.

Consider a situation where a certain type of incident is happening in time and λ is the fixed number of incidents happening per unit time. Let T be the random variable of the time until the next incident happens. Then the probability that the incident will occur within a time t, in other words that $T < t$, is derived directly from the Poisson distribution by considering the event r =1, which gives the exponential distribution:

$$P[T < t] = 1 - e^{-\lambda t}.$$

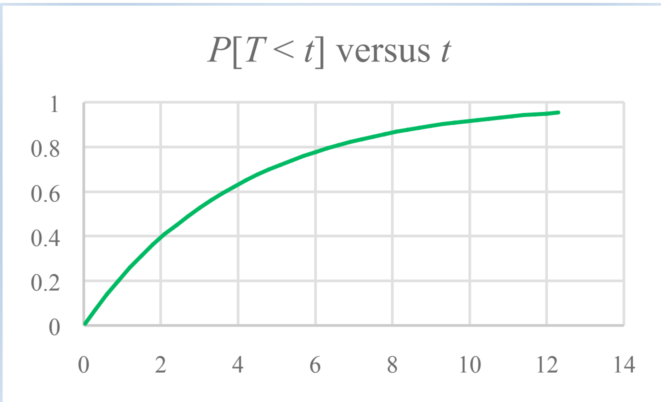
The exponential function is used to model certain types of failure rates and is the starting point for queuing theory because of this particular interpretation.

Consider the example of telephone calls known to arrive in an exchange at a rate of 15 per minute. We will calculate the probability that a call comes within 10 seconds.

If the calls are arriving at 15 per minute, then the rate (working with seconds as our unit of time) is λ = ¼. Then the probability is:

$$P[T < 10] = 1 - e^{-\frac{1}{4} .10} = 1 - e^{-2.5} = 0.918.$$

This result means that the probability of having an event within the ten seconds is very high, 0.918 as expected, if they average out at one call every 4 seconds. Here is a graph of $P[T < t]$ for λ = ¼:



For this distribution, we saw that the CDF is

$$F(x) = 1 - e^{-\lambda x}.$$

The Probability density function is then found by differentiating this function:

$$f(x) = \frac{\mathrm{d}F}{\mathrm{d}x}(x) = \lambda e^{-\lambda x}.$$

The mean is given by

$$\mu = E[X] = \int_R u f(u)\,\mathrm{d}u = \lambda \int_0^\infty u e^{-\lambda u}\,\mathrm{d}u.$$

Using a familiar result form integration by parts, this gives the result

$$\mu = 1/\lambda.$$

So the mean value for the exponential distribution is the inverse of the rate parameter λ, in other words the mean time between events. The exponential distribution is often written using this parameter instead:

$$P[X < x] = 1 - e^{-\lambda x} = 1 - e^{-\frac{x}{\mu}}.$$

For clarity we can write this as

$$P[X < x] = 1 - \exp(x/\mu).$$

The standard deviation is calculated with the same use of calculus to give

$$\sigma^2 = 1/\lambda^2.$$

Both the Poisson and exponential distributions are governing situations where incidents are happening at a fixed rate, independent of time; such a process is called a Poisson process. The exponential distribution modelling the time to the next incident has an interesting property we explore here. The probability an incident occurs within a time t is

$$P[T < t] = 1 - e^{-\lambda t}.$$

Let A be the event that the incident did not occurr before time s and let B be the event the incident occurs before the later time $s + t$. Now consider the probability the incident occurs before time $s + t$ but did not happen before time s: $P[B|A]$. Recall Bayes rule:

$$P[B \mid A] = \frac{P[B \cup A]}{P[A]}.$$

The probability of A is $e^{-\lambda t}$. The probability of B and A is then the probability an incident happens between time s and later time $s + t$:

$$P[s < T < s + t] = e^{-\lambda s}(1 - e^{-\lambda t}).$$

The application of Bayes rule then means that

$$P[B|A] = 1 - e^{-\lambda t}.$$

This is the same as the original exponential distribution. The distribution and the Poisson process it models are said to be *memoryless*.

3.5 The Normal Distribution

The normal or Gaussian distribution is one of the most important distributions, arising in many contexts in nature. It is characterised by two parameters and defined by the following probability distribution function.

A normally distributed continuous random variable X can take on any real number and has a probability density function given by the equation

$$f(x) = \frac{1}{b\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}\left(\tfrac{x-a}{b}\right)^2\right),$$

where a and b are positive real numbers. This means that the cumulative distribution function is given by

$$P[X < a] = \frac{1}{b\sqrt{2\pi}} \int_{-\infty}^{a} \exp\left[-\frac{1}{2}\left(\frac{u-a}{b}\right)^2\right] \mathrm{d}u.$$

The integral identity

$$\int_{-\infty}^{\infty} \exp\left(-\tfrac{1}{2}u^2\right) \mathrm{d}u = \sqrt{2\pi}\,,$$

ensures that the function f is a valid density function. Some work with these equations will show that

$$E[X] = a \text{ and } \mathrm{Var}[X] = b^2.$$

We therefore rewrite the definition as follows. For the normal distribution with mean μ and standard deviation σ, the probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}\left(\tfrac{x-\mu}{\sigma}\right)^2\right).$$

The cumulative distribution function is

$$P[X < a] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} \exp\left[-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right] \mathrm{d}u.$$

If a variable X is normally distributed, this means that if a large number of values are generated of the variable, then they will be more likely to be close to the mean μ, and unlikely, as determined by the standard deviation σ, to be far from it. In the particular case where the mean is 0 and the standard deviation is 1, the so-called standard normal variable usually denoted by Z, the equation for the density function reduces to

$$f(x) = \tfrac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}x^2\right),$$

And so the Cumulative Distribution function is given by

$$P[Z < x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\tfrac{1}{2}u^2\right) \mathrm{d}u.$$

This specific function is often called the 'error' function, denoted by $\Phi(x)$.

At this point we will note some basic properties of the normal distribution arising from the definition.

1. Let Z be a normally distributed random variable with $\mu = 0$ and $\sigma = 1$. This is called the *standard* normal distribution. Then $X = \sigma Z + \mu$, in other words
$$P[X < a] = P\left[Z < \frac{a-\mu}{\sigma}\right].$$
2. The distribution is symmetric: $P[Z < -a] = P[Z > a]$.
3. The probabilities $P[Z > a]$ fall away very rapidly. The normal distribution has the famous 'bell-curve' shape.

The integral defining $\Phi(x)$ does not deliver a closed form, so values must be produced by numerical integration. Before the advent of readily available computing power in spreadsheets and statistical packages, standard practice was to use probabilities calculated for the standard normal distribution in a table and use the equation linking probabilities to calculate those for any normal variable. The tables typically have, for example, the probabilities

$$P[\,Z > a],$$

for a positive real number a, in other words $1 - \Phi(a)$. Calculations for general normal distributions then use the properties of the distribution and the laws of probability, namely

$$P[\,Z > a] + P[Z < a] = 1.$$

Because of the main uses of this distribution in statistical testing, we will use the 'greater than' tables when required.

Notation

If a variable X is normally distributed, with mean μ and standard deviation σ, we denote this quickly by saying

$$X \sim \mathrm{N}(\mu\,,\sigma).$$

Thus we can say that variable Z follows the standard normal distribution by saying

$$Z \sim \mathrm{N}(0,\,1).$$

3.5.1 Example: Height in men

We will calculate the following probabilities using the information in the traditional tables with $P[\,Z > a]$. Let H be the random variable of height in men. This variable is normally distributed with mean and standard deviation are $\mu = 1.71$m, $\sigma = 0.11$m, so

$$H \sim \mathrm{N}(1.71\text{m}, 0.11\text{m}).$$

We will find the probability that the height of a man chosen at random is:

- Less than 1.74 metres.
- Less than 1.64 metres.

For the first event, we use the first property of the normal distribution and the laws of probability to see that

$$P[\,H < 1.74\text{m}] = P[\,Z < 0.27] = 1 - P[\,Z > 0.27].$$

This is the probability in the table so it is = 1 – 0.3936 = 0.6064. For the second part,

$$P[H < 1.64\text{m}] = P[Z < -0.64].$$

Then apply the symmetry property:

$$P[Z < -0.64] = P[Z > 0.64] = 0.2611.$$

3.5.2 Example: Middle Ranges

For the case of height in men, where $H \sim \mathrm{N}(1.71\text{m}, 0.11\text{m})$, we wish to find the middle 95% of heights. This means two values a and b such that $P[a < H < b] = 0.95$, but because it is the *middle* 95%, the probabilities of H being above or below the range are the same.

This gives us a way of tackling the calculation, since we have two probabilities:

$$P[H < a] = 0.025 \text{ and } P[H > b] = 0.025.$$

Looking at the upper limit, we need to find b where

$$P[H > b] = 0.05.$$

The value from the tables is 1.96, and this gives

$$b = 1.96\text{x}0.11\text{m} + 1.71\text{m} = 1.9256\text{m}.$$

The symmetry property means that the the calculation for a requires the figure –1.96;

$$a = -1.96\text{x}0.11\text{m} + 1.71\text{m} = 1.4944\text{m}.$$

3.6 The Central Limit Theorem

The importance of the normal distribution comes from the following fundamental theorem, called the Central Limit Theorem. The theorem is stated here, with two more refinements that make it increasing powerful and useful.

3.6.1 Central Limit Theorem, Part 1

Let X_1, X_2, … X_n be n random variables, all following a normal distribution with mean μ and standard deviation σ. Then the random variable of the sum of these variables is normally distributed, with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.

This theorem means that the variable Z given by

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}, \text{ where } \overline{X} = \frac{\sum X_i}{n},$$

follows the standard normal distribution. It is important that this version of the theorem states an equality; this is because of the requirement that the X_i are all identically normally distributed. The quantity $\overline{X}$ is of course the sample mean, but expressed as a random variable. The next, more powerful, version of the theorem broadens its usefulness; we first define another variable S as

$$S^2 = \frac{\sum X_i^2 - n\overline{X}^2}{n}.$$

The variable S is of course the same as the standard deviation of the variables X_i, with n instead of $n - 1$, though of course this makes no difference as we are discussing large n.

3.6.2 Central Limit Theorem, Part 2

For the n random variables X_1, X_2, … X_n, which are N(μ ,σ), the value of the variable S converges to the standard deviation σ as the number n becomes large. Then the mean of these n numbers is normally distributed, with mean μ and with standard deviation tending to $S/\sqrt{n}$.

This theorem means that the variable Z, given by

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}},$$

follows the standard normal distribution for large n. Now this theorem is useful if we are dealing with data values we knew to be normally distributed but with unknown standard deviation. The next part of the theorem addresses this last limitation.

3.6.3 Central Limit Theorem, Part 3

For the n random variables X_1, X_2, … X_n, with the same distribution and same mean of μ, the mean of these n variables is normally distributed, with mean of μ and standard deviation converging on $s/\sqrt{n}$.

This version of the theorem means that the variable Z given by

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}},$$

converges on the standard normal variable, for a sufficiently large value of n, irrespective of the distributions of the variables X_i.

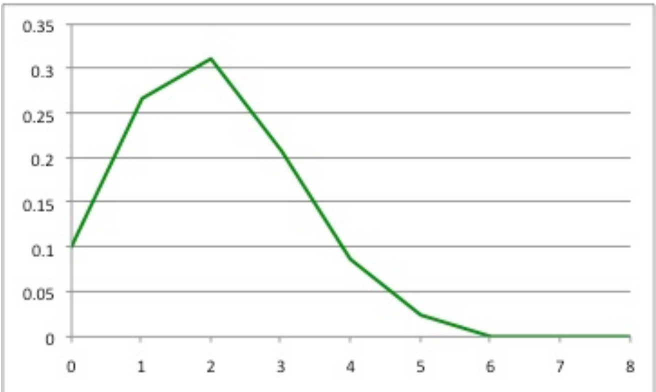
3.6.4 Origins

The origins of our understanding of both the normal distribution and the central limit theorem are naturally linked and are instructive. De Moivre arrived at a familiar looking bell curve by considering the number of heads resulting from a large number of throws of a coin.

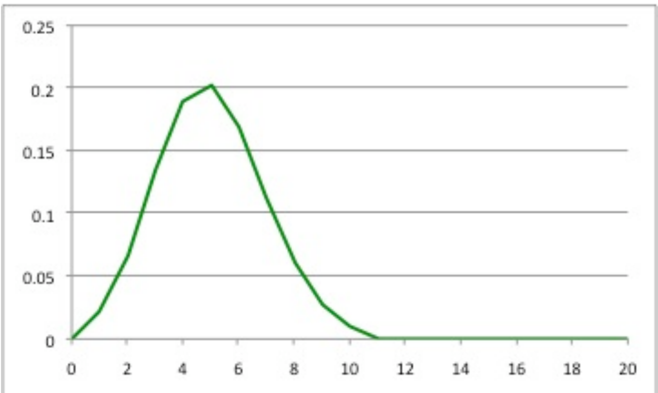
Laplace took this further and looked at the calculations for the binomial distribution when n is large and p is not close to 0 or 1. Here are some examples.

Let a multiple choice exam have n questions, each of which has 4 possible answers, only one of which is correct. A student chooses his answers at random. Let N be the random variable of the number of correct answers the student gets, this then follows the binomial distribution for a given n, with $p = ¼$. Here are graphs of the pdf against the number of correct answers:

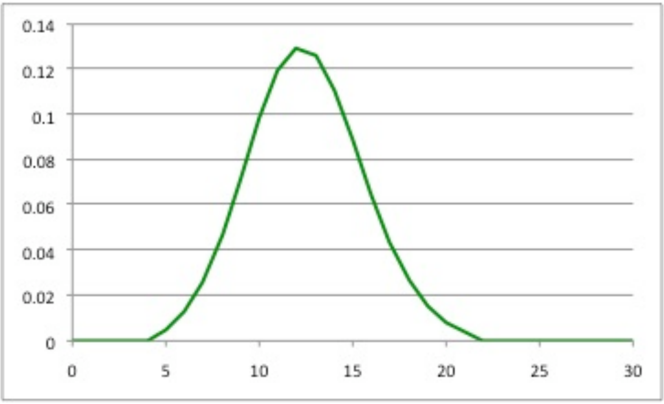
$n = 8$



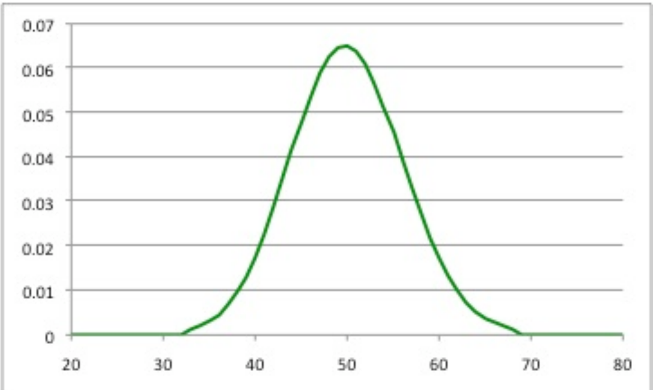
$n = 20$:



$n = 50$:



$n = 200$:



The familiar bell-shaped curve is seen to emerge as n becomes larger. Why should this happen? Recall that we can regard the binomial distribution as a sum of Bernoulli trials, so let Bi be a random variable which takes on one of two discrete values

$$0, 1, \text{ and}$$

$$P[B = 1] = p, \text{ so that } P[B = 0] = 1 - p.$$

The density function was:

$$f(0) = 1 - p, f(1) = p.$$

The parameters are:

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

The binomial distribution is then

$$N = \sum_{i=1}^{n} B_i\,,$$

We can now apply the central limit theorem, with one note; we do not need to use the sample standard deviation S since we know the actual standard deviation. We also write it in the form of a sum rather than a mean and state that, if μ and σ are the mean and standard deviation respectively of each B_i, then the theorem means that N is normally distributed with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$. We know what these values are, so then for the variable N:

$$\mu = np,\ \sigma^2 = np(1-p).$$

These are of course the same equations for mean and standard deviation as those for the binomial distribution, but now they apply to the normal distribution. For the case being considered,

$$\mu = n/4,\ \sigma^2 = 3n/16.$$

We therefore see the binomial distribution tend to the normal as n increases.

3.6.5 Confidence Intervals

The normal distribution will have a huge importance for the subject of Statistical Testing. We will finish our discussion here with one important use of the distribution arising directly from its origins in the central limit theorem.

Let X be a random variable, defined on the sample space of an experiment. It is required to estimate the mean of the variable X. the experiment is run n times, where n is a large value. Let X_i be the i-th representative of the variable; it is itself a random variable and so therefore we can apply the central limit theorem to establish that that the variable given by

$$\frac{\bar{X} - \mu}{S / \sqrt{n}},$$

converges on the standard normal variable Z. So for a sufficiently large value of n, irrespective of the distributions of the variables X_i, we can sat that

$$Z = \frac{\bar{X} - \mu}{S / \sqrt{n}} \sim \mathrm{N}(0, 1).$$

Now let α be a low probability and define a number z_α such that

$$P[Z > z_\alpha] = \alpha.$$

It then follows that

$$P\left[\left|\frac{\bar{X} - \mu}{S / \sqrt{n}}\right| > z_{\frac{1}{2}\alpha}\right] = \alpha,$$

which leads directly to the identity

$$P\left[\overline{X} - \frac{z_{\frac{1}{2}\alpha}}{S/\sqrt{n}} < \mu < \overline{X} + \frac{z_{\frac{1}{2}\alpha}}{S/\sqrt{n}}\right] = 1 - \alpha \cdot$$

Reading this event again, the probability the mean μ lies between these two values of the random variables shown is the figure α. Thus if the experiment is run a large number of times and values are produced for the variables delineating upper and lower limits for the mean μ, e then have n estimates for an interval within which the mean μ lies. Then for a proportion α of these intervals, this will be true, as n increases without limit.

This is a theoretical foundation for the practical concept of the confidence interval for a probability α. Typically α is chosen to be 0.05 or 0.01. As we develop the idea of the statistical test on a mean, we will encounter its limitations; the confidence interval for a mean is therefore often preferred as having a better foundation in the idea of a random variable and the all-powerful central limit theorem.