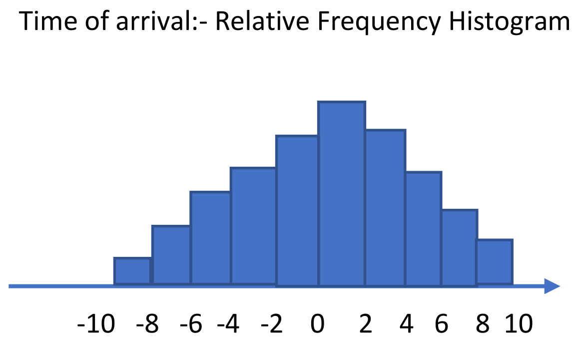
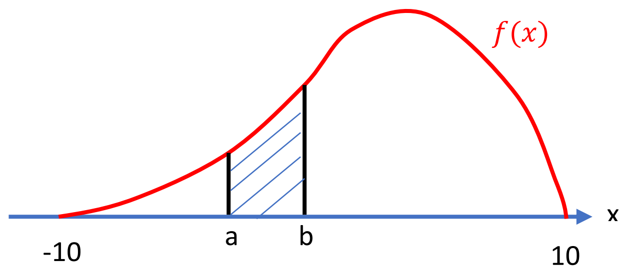
Continuous Random variables

Introduction

So far, we have discussed discrete random variables, i.e., the outcomes are points. In many applications the outcome could be a measurement, or time to an event etc, and so can have values anywhere over a continuous range. The measurement of the weight of an adult could take any value in the range [7,30] stones say, or the time of arrival of a bus could differ by that on the timetable by a time in the interval [-10,10] minutes. Such variables are referred to as continuous variables since they can take any value in a specified interval, but how do we assign the probabilities. We could divide the interval [-10,10] interval into a number of none overlapping intervals and determine the interval probabilities by analysing collected data. This might generate the relative frequency histogram shown. Because the plot is a histogram the area of each bar represents the relative frequency of the event.



The more data you collect the narrower the width of the bars and so the probability profile becomes smoother leading eventually to a smooth curve as shown. The probability that the arrival time lies in the interval [a,b] becomes the shaded area. If we know the mathematical form of $f(x)$ we can determine this area by integration, i.e.,



$$P(a \le X \le b) = \int_a^b f(x)dx.$$

Our basic rules of probability ensure that

$f(x) \ge 0$ for all x and $\int_{-\infty}^{+\infty} f(x)dx = 1$

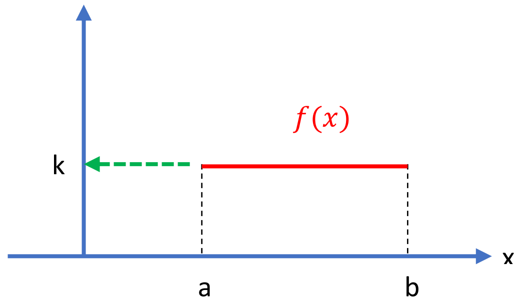
Note:- $f(x)$ will be 0 for most of the interval $(-\infty, +\infty)$. The name given to the smooth curve $f(x)$ is the **probability density function(pdf)**. We consider two popular continuous distributions – there are many more.

Continuous uniform distribution

Suppose the continuous random variable X can take any value in the range [a,b] with equal probability. Then we call it **a uniform continuous distribution** and denote it by U(a,b). Since there is no preferred value for X the probability density function must be flat, as shown in the figure alongside.

To find the value of k we know that the total probability is 1 and so we can write

$$\int_a^b k dx = k[x]_a^b = k(b-a) = 1$$



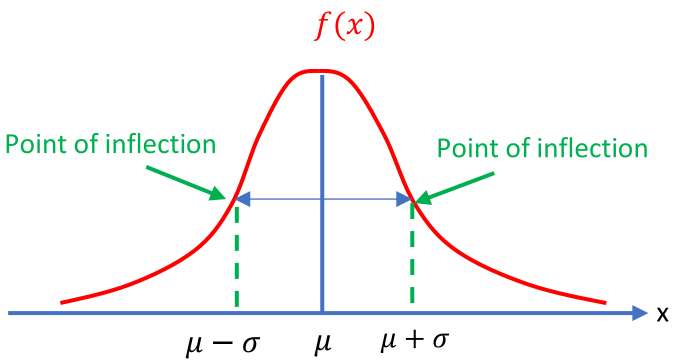
Therefore $k = \frac{1}{(b-a)}$. The mean and variance are as follows

$$\text{Mean} = \mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx = \frac{1}{(b-a)} \int_a^b x dx = \frac{(a+b)}{2}$$

$$\text{Variance(X)} = E(X^2) - E^2(X) = \frac{(b-a)^2}{12}$$

Normal Distribution

This is one of the most important probability density functions in statistics and probability theory. It is a perfectly symmetric pdf about its mean as shown alongside. The points of inflection are 1 standard deviation away from the mean. It is sometimes referred to as a "bell" shaped curve.



Mathematically the pdf is $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for $-\infty < x < \infty$, and notationally we would say X has a $N(\mu, \sigma^2)$ distribution.

A number of useful facts are listed below

(1) A **standard normal distribution** is a normal distribution with a mean of 0 and a standard deviation of 1, i.e., *N*(0,1).

(2) If the random variable X is $N(\mu, \sigma^2)$ the random variable $Z = \frac{(X-\mu)}{\sigma}$ is *N*(0,1), i.e. a standard normal variable.

(3) The cumulative distribution function for a standard normal variable Z is

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds,$$ where s is a dummy integration variable.

(4) If X is $N(\mu, \sigma^2)$ then $P(X \le x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$

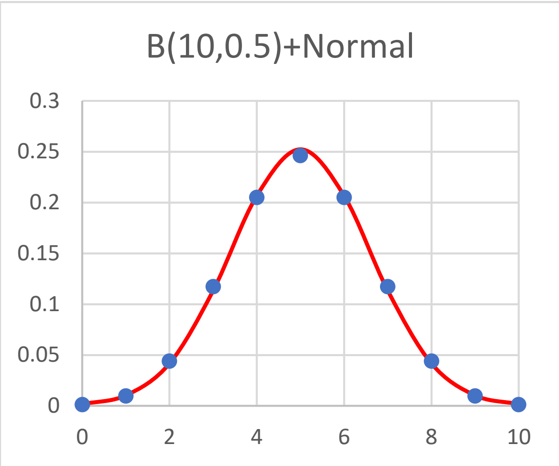
(5) The probability $P(\mu - \sigma < X < \mu + \sigma)$ =0.6827

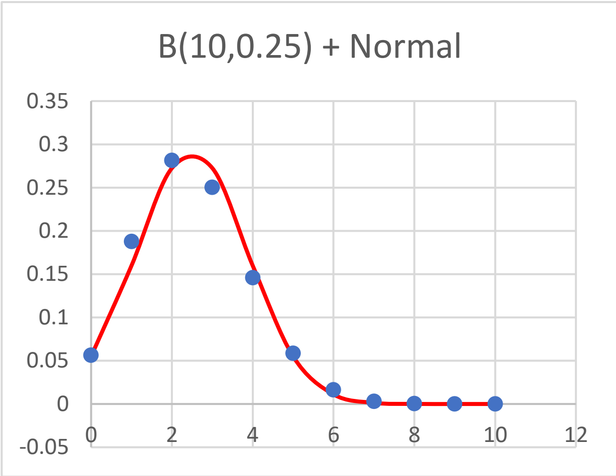
(6) The probability $P(\mu - 1.96\sigma < X < \mu + 1.96\sigma) = 0.95$

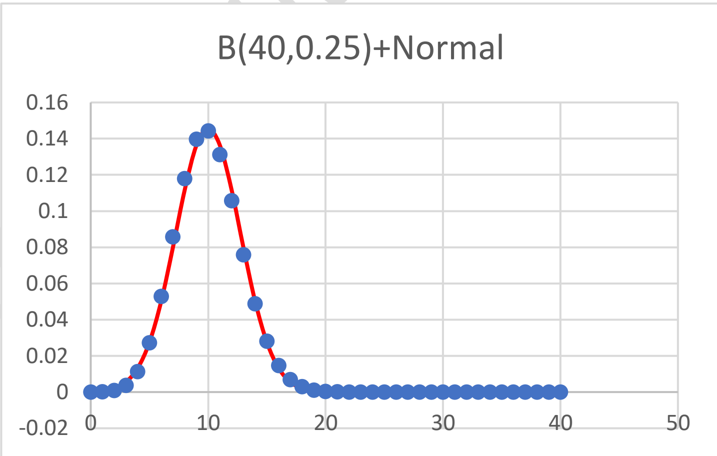
(5) and (6) give an interpretation of σ, as a measure of spread.

Normal Approximation to the Binomial

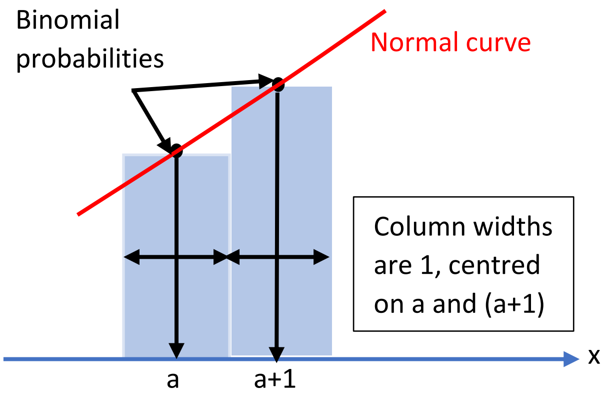
Here we investigate the possibility that the Normal pdf can be used to evaluate binomial probabilities. You would expect that the approximating normal distribution should have the same mean and standard deviation as the binomial that it is approximating. Below are a number of plots comparing the two distributions. The exact binomial probability is the blue dots and the approximating normal is the red curve.







You can see that the approximation is very good when n=10 and p = 0.5. However, reducing p to 0.25 makes the approximation worse for the same n value of 10. Increasing the n to 40 with p at 0.25 recovers a reasonable fit. This suggests that a criterion for an adequate fit would be related to both n and p. In fact, the working rule is that if the product $np > 10$ then the approximation should be reasonable. Obviously the larger the n the better. There is one final adjustment necessary to improve probability calculations, which is called the continuity correction. The probability for a binomial variable is at a point, since it is a discrete random variable, whereas the probability for a continuous random variable is the area under a curve and therefore an interval needs specifying. Consider the sketch alongside. For the probability at the point a, you place a column of width 1, centred on a and of height equal to the probability at that point. Numerically the area of each column is then the same as the value of the probability. The normal curve is then approximating these areas rather than the points. A consequence of this is that you need to adjust the specified limits when using the normal approximation. An



event such as $X_B \geq a$ for the binomial distribution becomes $X_N \geq a - 0.5$ when using the normal approximation and $X_B \leq a$ becomes $X_N \leq a + 0.5$. Just think of the columns and you will know whether to add 0.5 or subtract 0.5.

Note:- The continuity correction of 0.5, should actually be a half of the width between neighbouring probabilities, which could be different to 0.5, be careful.