1.2 : Statistical Independence

Let X and Y denote two sets of outcomes of an experiment. In keeping with the notation already introduced,

$$P(X) = \text{Prob(outcome } \in X), \quad P(Y) = \text{Prob(outcome } \in Y)$$

$$P(X\,Y) = \text{Prob(outcome } \in X \text{ and } Y)$$

We then say that X and Y are **independent** sets of outcomes if and only if

$$P(XY) \;=\; P(X)\,P(Y)$$

We can justify this result as follows. If X and Y are independent, then knowledge that the outcome belongs to X has no effect on the probability that the outcome belongs to Y. This means that

$$P(Y|X) \;=\; P(Y)$$

But since for *any* X and Y we have

$$P(XY) \;=\; P(X)P(Y|X)$$

it follows that, for X and Y independent

$$P(XY) \;=\; P(X)\,P(Y)$$

This equation defines statistical independence.

Extension to more than two sets of outcomes is straightforward. For example:-

$$P(XYZ) \;=\; P(X|YZ)\,P(YZ) \quad = P(X|YZ)\,P(Y|Z)\,P(Z)$$

If X, Y and Z are independent, then

$$P(XYZ) \;=\; P(X)\,P(Y)\,P(Z)$$

The outcomes which we have considered so far have been **qualitative** (head, tail, ace, spade etc). This is a useful means of introducing definitions of probability and independence, but we now need a description of probability which deals with **quantitative** outcomes.

1.3 : Probability Distributions

An observed event with several possible outcomes is called a *random* event. When the outcome is a numerical quantity (e.g. a physical measurement such as length, time, apparent magnitude, wavelength) it is called a **random variable** (RV).

1.3.1 : Discrete Probability Distributions

If a RV can take only a finite[1] number of values then it is a discrete RV. We can associate with each possible outcome, r, a probability, $p(r)$. The set of all $p(r)$ is called the **probability distribution** of the discrete random variable, r.

1.3.2 : Poisson Distribution

A **Poisson** RV is a discrete RV describing, e.g., the number of photons counted in a given time by a CCD. We denote the probability of counting r photons in time interval t by $p(r,t)$, although some textbooks use the notation $P_r(t)$. A Poisson RV is defined by the following three postulates.

a The probability of an event occuring in time interval, t, is independent of the past history of events prior to t

b For small interval, δt, there is an intrinsic **rate**, (i.e. number of events per unit time) $\mu\,(>0)$ such that the probability of a single event in δt, $p(1,\delta t) = \mu\delta t + o(\delta t)$.

c The probability of two or more events happening at the same time is zero, i.e. $p(r,\delta t) = o(\delta t)$, for all $r \geq 2$.

Here $o(\delta t)$ represents any function such that $o(\delta t)/\delta t \to 0 \quad \text{as } \delta t \to 0$.

These postulates imply that the probability distribution function of a Poisson RV takes the form

$$p(r,t) \quad = \quad \frac{(\mu t)^r}{r!}\,e^{-\mu t}$$

[1] or countably infinite, although this mathematical subtlety need not concern us in this course

We can prove this result by *induction*; although this proof is not examinable, a short summary is provided on a handout (see website). Note that

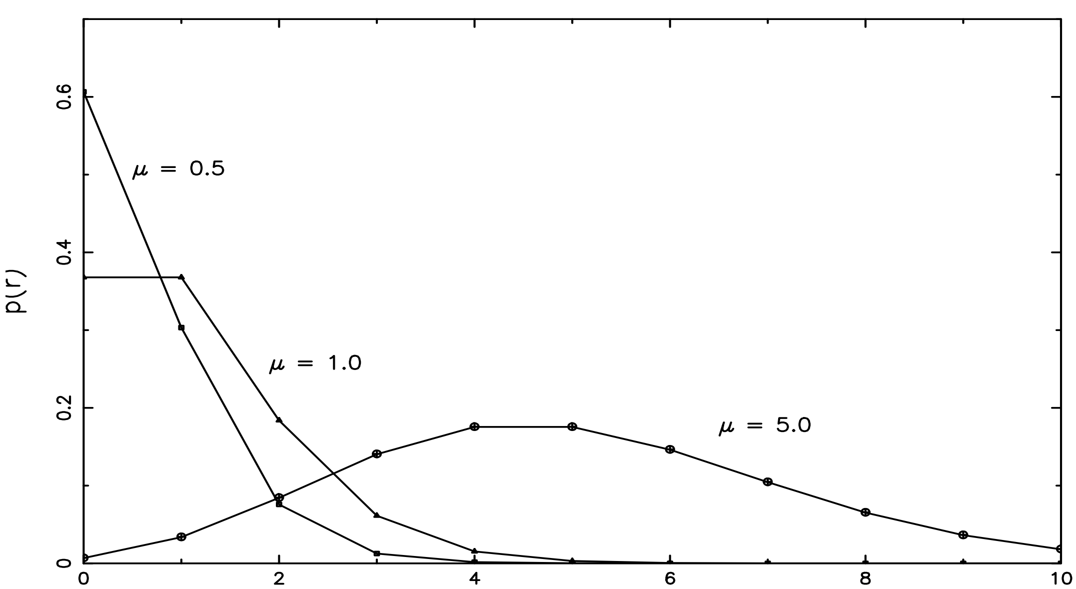
$$\sum_{r=0}^{\infty} p(r,t) \quad = \quad \sum_{r=0}^{\infty} \frac{(\mu t)^r}{r!} e^{-\mu t} \quad = e^{-\mu t} \sum_{r=0}^{\infty} \frac{(\mu t)^r}{r!} \quad = \quad 1$$

as required, since r must take *some* value between 0 and ∞. It is very often the case that the time interval, t, is simply taken to be be unity, in which case we can write

$$p(r) \quad = \quad \frac{\mu^r}{r!} e^{-\mu}$$

Figure 2 shows a plot of the Poisson distribution for several different values of μ. Note that the shape of the PDF changes significantly with increasing μ: for small values of μ the PDF is monotonic decreasing, whereas for larger values of μ it takes on more of a bell shape.

Figure 2: Poisson distribution, $p(r)$, for different values of μ



Note that we could also define the Poisson RV in space – e.g. the probability of finding r galaxies in a given volume or projected area of sky could be modelled as a spatial Poisson RV. In that case, the rate parameter, μ, would have dimensions of inverse volume, or inverse area, instead of inverse time. We will consider another common discrete probability distribution, the **binomial distribution**, later in the course.

1.3.3 : Continuous Distributions

Suppose a RV, X, can take *any* real value in a given interval – i.e. we have an uncountably infinite number of possible outcomes. We call X a **continuous RV**. Examples of continuous random variables include the apparent or absolute magnitude of stars or galaxies, distances, redshifts, orbital inclinations, etc. In fact, almost *all* quantitative physical measurements in astronomy can be regarded as continuous RVs.

Many textbooks denote a RV by a capital letter – often in bold face – and use the corresponding small letter to denote a particular *observed value*, or **realisation**, of the RV. Whenever convenient, we will adopt this notation.

What is $P(X = x)$? We have a potential paradox here. If we sum probabilities over x, we would have

$$\sum_x P(X = x) \quad = \infty \quad > 1$$

if $P(X = x) \neq 0$ for an infinite number of values of x. Of course, a probability cannot be greater than unity, far less equal to infinity!

This is simply telling us that the probability of X being *exactly* equal to any fixed value is zero. Instead we measure the probability of X lying in a small interval, $(x, x + dx)$. In the limit as $dx \to 0$, we have

$$P\left(X \in (x, x + dx)\right) \quad = \quad p(x)dx$$

Here $p(x)$ is known as the **probability density function** (PDF) but is **NOT** itself a probability. In particular, we can certainly have $p(x) > 1$, but always

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

Thus the probability that X lies in the interval (a, b) is given by

$$P(a < X < b) \quad = \quad \int_a^b p(x)dx$$

In general we can always define a RV, X, on the entire real line, $(-\infty, \infty)$. We simply define $p(x) = 0$ outside the range of physically meaningful values of x.

1.3.4 : Cumulative Distribution Function

Consider a RV, X. The function

$$P(t) = P(X < t) = \int_{-\infty}^{t} p(x)dx$$

is called the **cumuluative distribution function** (CDF) of X. Thus the CDF measures the probability that X takes a value less than t. Note that $P(-\infty) = 0\,, \quad P(\infty) = 1$.

1.3.5 : Examples of Continuous RVs

(1) Simplest example of a continuous RV is the **uniform distribution**, usually denoted by $U(a, b)$, defined on the interval (a, b), with $a \neq b$. The uniform distribution has PDF

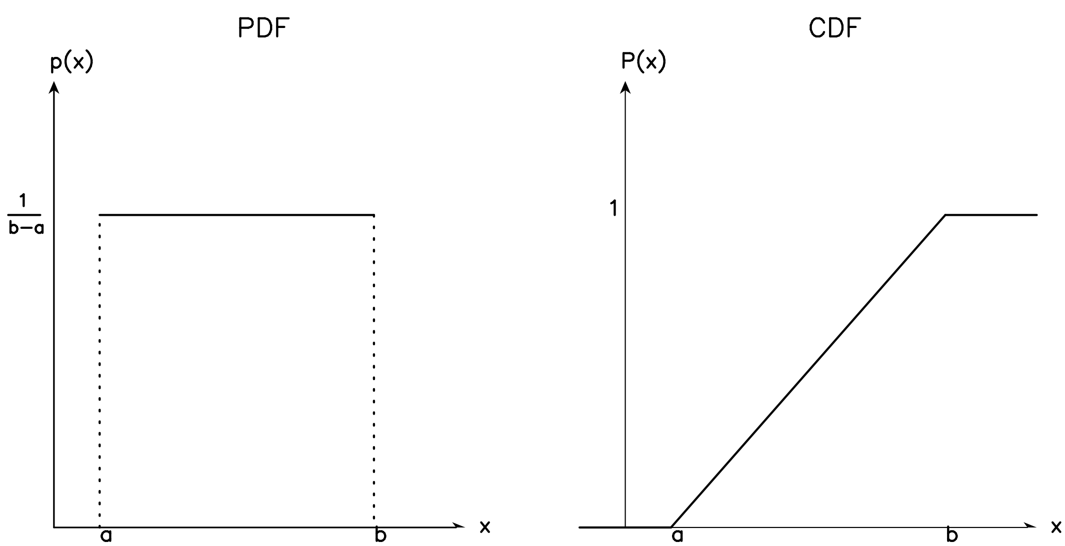
$$p(x) = \begin{cases} 1/(b-a) & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

and CDF

$$P(x) = \begin{cases} 0 & x \leq a \\ (x-a)/(b-a) & a < x < b \\ 1 & x \geq b \end{cases}$$

These functions are shown in Figure 3.

Figure 3: PDF and CDF of the uniform RV, $U(a, b)$



(2) The most important continuous RV is the normal, or Gaussian, distribution, usually denoted by $N(\mu, \sigma)$. It has PDF

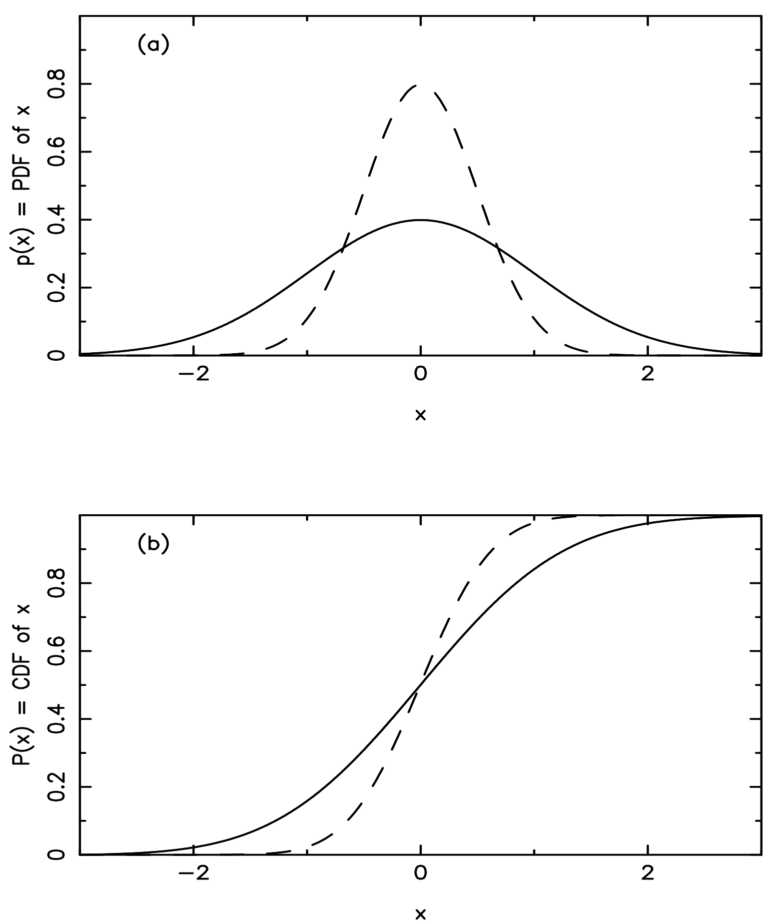
$$p(x) \quad = \quad \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$$

This is a bell-shaped curve, symmetrical about $x = \mu$. The parameter σ is a measure of the width of the PDF. There is no analytic form for the cdf of the normal distribution, although it is often denoted by $\Phi(t)$. Thus

$$\Phi(t) \quad = \quad \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{t} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]dx$$

The value of $\Phi(t)$ is tabulated in many statistics textbooks, or in numerical packages. The PDF and CDF of the normal distribution for several different values of σ is shown in Figure 4. Both the normal and uniform distributions are very important for theoretical reasons, as we will see later.

Figure 4: PDF (a) and CDF (b) of the normal distribution for $\sigma = 0.5$ (dashed) and $\sigma = 1.0$ (solid)



How, in practice, do we determine the PDF of a RV? The most intuitive approach is to make use of our 'frequentist' definition of probability. Suppose we make repeated measurements of our physical quantity, i.e. we repeat our experiment a very large number of times. We then record the measured values in a **histogram**, normalised so that the total area under the histogram is equal to unity. In the limit as the number of experimental 'trials' tends to infinity (and where the width of the histogram bins tends to zero), the heights of the histogram bins (the 'relative frequency' of the different outcomes) 'traces out' the PDF of the RV. This is illustrated for the simple case of a RV uniform on the interval $(0, 1)$ in Figure 5, below. (Here the sequence of 'experiments' have been generated on computer using a random number generator program. Note that as the number of trials increases the histogram more accurately approximates the 'flat' PDF.

Figure 5: Histogram approximations to a uniform RV, with PDF $U(0, 1)$