7. Statistics and Applied Probability

The concept of a random variable (RV), denoted by low case bold, eg **x**(ζ), is central. Loosely we can say that a RV is a number corresponding to one of a number of possible outcomes ζ. The set of possible outcomes, and defined subsets are usually indicated by curly brackets.

For example $\{x_1 \leq \mathbf{x} \leq x_2\}$ denotes the set of outcomes ζ such that $x_1 \leq \mathbf{x}(\zeta) \leq x_2$.

The *probability distribution function* or *cumulative probability distribution* of a RV **x** is defined as

$$F_x(x) = P\{\mathbf{x} \leq x\}$$

The *probability distribution* or *probability density function* (PDF) is formally defined as the derivative of the cumulative probability distribution, i.e.

$$f(x) = \frac{dF(x)}{dx}$$

It is sometimes also called the frequency function as it can be interpreted as the continuous limit of a frequency distribution, suitably normalised.

The probability of something happening is always unity and this is written formally as

$$\int_{-\infty}^{\infty} f(x)dx = 1 \qquad (7.1)$$

Also,

$$F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx \qquad (7.2)$$

and

$$P\{x_1 \leq \mathbf{x} \leq x_2\} = \int_{x_1}^{x_2} f(x)dx \qquad (7.3)$$

Although simple these are key equations for probabilistic analysis.

Perhaps the most important probability density function (distribution) is the so called *normal* or *gaussian* distribution.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-x - \eta)^2 / 2\sigma^2 \qquad (7.4)$$

The *expected value* or mean of an RV **x** is defined as

$$E\{x\} = \int_{-\infty}^{\infty} x f(x) dx \qquad (7.5)$$

It can be shown that the mean of a normal distribution as given in (7.4) is just η.

The variance of an RV **x** is defined by

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \eta)^2 f(x) dx \qquad (7.6)$$

Its square root, σ is known as the standard deviation. For the normal distribution it is just σ and hence the use of the notation in (7.4).

If M is a well defined event or group of events, the conditional density function $f(x|M)$ understood as 'probability of x given M' can be conceived. It may or may not be easy to calculate it depending on properties like statistical independence.

The variance is an example of a *central moment*. General moments can be defined by

$$E\{(x - a)^n\} = \int_{-\infty}^{\infty} (x - a)^n f(x) dx$$

This would be described as the nth moment of f(x) about a.

The concept of a probability distribution can be extended to more than one RV. We talk about *joint probability density* functions, f(x,y) related to $F_{xy}(x,y)$.

The covariance C_{xy} of two Rvs **x** and **y** is defined to be

$$C_{xy} = E\{(x - \eta_x)(y - \eta_y)\} \qquad (7.7)$$

and following from this we can define a *correlation coefficient* which is just a normalised covariance as

$$r = \frac{C_{xy}}{\sigma_x \sigma_y} \tag{7.8}$$

It takes values between 0 and 1.

We can now write out the joint normal probability distribution as

$$f(x, y) = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1 - r^2}} \exp\left\{ - \frac{1}{2(1 - r^2)} \left[\frac{(x - \eta_x)^2}{\sigma_x^2} - \frac{2r(x - \eta_x)(y - \eta_y)}{\sigma_x \sigma_y} + \frac{(y - \eta_y)^2}{\sigma_y^2} \right] \right\}$$

For short this is sometimes written as

$$N(\eta_x, \eta_y; \sigma_x, \sigma_y; r)$$

7.1 Autocovariance and autocorrelation

We can use the forgoing analysis tools to look at time series.

Time series are random variables which exist in time, eg x(t). Lags in time of a fixed length are often denoted by τ.

We are often concerned to look at how fast a time series changes in time. This motivates a function known as autocorrelation.

First we define *autocovariance* as

$$R(\tau) = E\{(x(t) - \eta_x)(x(t - \tau) - \eta_x)\} \tag{7.9}$$

For $\tau = 0$ this is equivalent the variance of x(t) treated as a simple RV.

The normalised version is known as the *autocorrellation*

$$r(\tau) = \frac{R(\tau)}{\sigma^2}$$

Clearly for $\tau = 0$ this always equals 1, for values of $\tau > 0$, $r(\tau) < 1$.

There is a simple link between autocovariance and spectral density in that the power spectrum turns out to be simply the Fourier transform of the autocovariance of a given process **x**(t).

$$S(\omega) = \int_{-\infty}^{\infty} R(\tau) e^{-i\omega\tau} d\tau \tag{7.10}$$

From the inverse transform we can regain R(τ)

$$R(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega) e^{i\omega\tau} d\omega \tag{7.11}$$

Care must be taken with the normalisation constants which appear in these formulae. An often used alternative in (7.10) and (7.11) is to use $1/\sqrt{(2\pi)}$ in both.

If **x**(t) is real then R(τ) is real and so is S(ω) and the latter two are also even functions. In this case

$$S(\omega) = \int_{-\infty}^{\infty} R(\tau) \cos(\omega\tau) d\tau$$

and

$$R(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega) \cos(\omega\tau) d\omega$$

Similarly the *cross spectral density* can be defined as the Fourier transform of the cross correlation function C_{xy}.

Finally a normalised function known as coherence can be defined

$$\lambda_{xy}^2(\omega) = \frac{|S_{xy}(\omega)|^2}{S_{xx}(\omega) S_{yy}(\omega)}$$