EART266 Geological Signal Processing
Lecture #2
Important PDFs

Uniform: This distribution is often used when only the range of a random variable is known. Few random processes actually mimic it.

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \le x < b \\ 0 & \text{else} \end{cases}$$

This has moments:

$$\mu_x = \frac{b+a}{2}$$
$$\sigma_x^2 = \frac{(b-a)^2}{12}$$

It is symmetric about the mean. The primary application of uniform distributions is as an a priori distribution for a range-limited quantity.

Dirac's Delta: This distribution is best represented by its cumulative density function:

$$F(x) = \begin{cases} 1 & x > 0 \\ 0 & x < 0 \end{cases}$$

The value at $x = 0$ is up for debate (most people say 1/2). The pdf is typically denoted by a delta:

$$F(x) = \int_{-\infty}^{x} \delta(s)ds$$

You can picture the pdf as a spike of infinitesimal width but unit area (meaning it is very tall) centered at $x = 0$. The position of the spike is easily moved to accommodate other values. For instance:

$$f(x) = \delta(x-a)$$

positions the spike at $x = a$. Dirac's delta function represents a non-random variable: it can only take on one value!

Moments:

$$\mu_x = a$$
$$\sigma_x^2 = 0$$

The primary application of Dirac's Delta has nothing to do with probability, but rather as a "comb" in integration. For example:

$$\int_{-\infty}^{\infty} g(x)\delta(x-a)dx = g(a)$$

Gaussian (Normal): This distribution is described by two parameters and has the form:

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

The moments depend directly on the parameters:

$$\mu_x = \mu$$
$$\sigma_x^2 = \sigma^2$$

All higher moments are functions of σ:

$$\sigma_{2p+1}^{2p+1} = 0 \text{ (it's symmetric)}$$
$$\sigma_{2p}^{2p} = \frac{(2p)!}{2^p p!} \sigma_x^{2p}$$

For example,

$$\sigma_4^4 = \frac{4!}{2^2 2!} \sigma_x^4 = 3\sigma_x^4$$

Gaussians occur frequently in nature and many non-Gaussian distributions can be made approximately Gaussian or are intrinsically approximately Gaussian. One reason Gaussians are so common is the **Central Limit Theorem** which states that the sum of n random variables tends to a Gaussian pdf as $n \to \infty$ regardless of their pdfs. The sum of Gaussian random variables is always Gaussian. Why this makes the Gaussian common is simple to understand: any process that combines many random variables (grain transport efficiency for example: current direction and velocity, particle size distribution, density distribution, bedform topography) will produce random variables that are approximately Gaussian.

Let $z = x + y$ where x and y are independent Gaussian random variables with means μ_x and μ_y and variances σ_x^2 and σ_y^2 respectively. Then z is a Gaussian random variable with mean $\mu_x + \mu_y$ and variance $\sigma_x^2 + \sigma_y^2$. In words, the means and variances of independent Gaussian random variables are additive (this is true of all sums of statistically independent random variables). More importantly, the sum of two Gaussian random variables is itself Gaussian. The proof of this is fairly difficult (there is a neat tool that greatly simplifies it known as the moment generating function):

$$f(z) = \int_{-\infty}^{\infty} f_x(x) f_y(z-x)\,dx = \frac{1}{\sigma_x \sigma_y 2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right) \exp\left(-\frac{((z-x)-\mu_y)^2}{2\sigma_y^2}\right) dx$$

This is an example of convolution. We impose that the sum of *x* and *y* equal *z*. That "conditions" the two variables, hence I have substituted *z* - *x* for *y* and am integrating over all possible values of *x*. The mathematics can be made much simpler if we let both *x* and *y* be zero mean, unit-variant variables, then:

$$f(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2} - \frac{(z-x)^2}{2}\right) dx$$

$$f(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-x^2 + zx - \frac{z^2}{2}\right) dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-x^2 + zx - \frac{z^2}{4} - \frac{z^2}{4}\right) dx$$

$$f(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\left(x - \frac{z}{2}\right)^2 - \frac{z^2}{4}\right) dx = \frac{1}{2\pi} \exp\left(-\frac{z^2}{4}\right) \int_{-\infty}^{\infty} \exp\left(-\left(x - \frac{z}{2}\right)^2\right) dx$$

$$f(z) = \frac{1}{4\pi} \exp\left(-\frac{z^2}{4}\right) \int_{-\infty}^{\infty} \exp\left(-\left(\frac{s}{2}\right)^2\right) ds = \frac{1}{4\pi} \exp\left(-\frac{z^2}{4}\right) \sqrt{2\pi}$$

$$f(z) = \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{z^2}{2^2}\right)$$

where I've utilized the fact that the integral of the Gaussian over the real number line is unity. The answer comes out to have zero mean and a variance of two, as predicted.

Another reason that Gaussians are so important is that of all pdfs with variance σ_x^2, the Gaussian has the greatest spread, *i.e.*, the least information. Proof of this is outside the scope of this class, but since we often know only the mean and variance of a random variable, assuming a Gaussian is justified–it implies the least additional information.

Using the definitions of marginal pdf, variance and covariance, we can examine joint Gaussian distributions. The simplest is the joint pdf of two independent Gaussian variables, *x* and *y*. Because *x* and *y* are independent, their joint pdf is simply the product of marginal pdfs:

$$f(x,y) = f_x(x) f_y(y)$$

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(\frac{-(x-\mu_x)^2}{2\sigma_x^2}\right) \exp\left(\frac{-(y-\mu_y)^2}{2\sigma_y^2}\right)$$

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{1}{2}\left(\frac{(x-\mu_x)^2}{2\sigma_x^2} + \frac{(y-\mu_y)^2}{2\sigma_y^2}\right)\right)$$

For covariant *x* and *y* the more general form is:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}} \exp\left(-\frac{1}{2\left(1-\rho_{xy}^2\right)}\left(\left(\frac{(x-\mu_x)}{\sigma_x}\right)^2 - 2\rho_{xy}\left(\frac{(x-\mu_x)}{\sigma_x}\right)\left(\frac{(y-\mu_y)}{\sigma_y}\right) + \left(\frac{(y-\mu_y)}{\sigma_y}\right)^2\right)\right)$$

You should confirm to yourself that this reduces to the previous joint pdf for $\rho = 0$.

To extend this to three or more random variables (with finite covariance) is less difficult that it might seem. However, we do need a few new quantities. The first is the covariance matrix:

$$C_{ij} = \int_{-\infty}^{\infty} d\mathbf{x}\left(x_i - \mu_i\right)\left(x_j - \mu_j\right) f(\mathbf{x})$$

Here $\mathbf{x}$ is a vector whose components, x_i for $i = 1, 2, \ldots, n$ are n random variables with individual mean values μ_i. The integral notation here is a little deceptive:

$$\int_{-\infty}^{\infty} d\mathbf{x} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} dx_1 dx_2 \ldots dx_n$$

in words, we are integrating over the whole of sample space for all the random variables.

The diagonal elements of **C**, namely the C_{ii}, are the variances of the x_i. The off-diagonal terms are covariances. **C** has three important properties:

1. $C_{ij} = \sigma_{ij}$ the covariance of x_i and x_j.

2. **C** is a symmetric matrix (*i.e.*, $C_{ij} = C_{ji}$).

3. **C** is positive semi-definite, such that:

$$\mathbf{x}^T \mathbf{C} \mathbf{x} \geq 0 \text{ for all } \mathbf{x}$$

Because it is symmetric, **C** and $\mathbf{C}^{-1}$ share eigenvectors. The eigenvalues of $\mathbf{C}^{-1}$ are simply the reciprocals of **C**'s, such that:

$$\mathbf{x}^T \mathbf{C}^{-1} \mathbf{x} \geq 0 \text{ for all } \mathbf{x}$$

is also true.

Using **C** for the n random variables $x_1, x_2, \ldots, x_n$, we define the n-variant Gaussian joint pdf as:

$$f(\mathbf{x}) = \left[\frac{1}{(2\pi)^n \det \mathbf{C}} \right]^{1/2} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu)^T \mathbf{C}^{-1} (\mathbf{x} - \mu) \right)$$

(Note that the μ's in the above are vectors–Word won't let me make them bold-face–containing as elements the means of the n random variables.) Since **C** is positive semi-definite, the argument of the exponential is always ≤ 0. In the exponential, the inverse of **C** serves to scale squared distance from the mean, as variance does in the univariate Gaussian.

This compact form is incredibly useful and is the basis of many statistical estimation schemes in common use.