2.3.1 Probability Densities, Moments.

Some statistical ideas are required, but the discussion is confined to stating some basic notions and to developing a notation.[20] We require the idea of a probability density for a random

Continued on next page...

variable x. This subject is a very deep one, but our approach will be heuristic.[21] Suppose that an arbitrarily large number of experiments can be conducted for the determination of the values of x, denoted X_i, $1 \leq i \leq M$, and a histogram of the experimental values found. The frequency function, or probability density, will be defined as the limit, supposing it exists, of the histogram of an arbitrarily large number of experiments, $M \to \infty$, divided into bins of arbitrarily small value ranges, and normalized by M, to produce the fraction of the total appearing in the ranges. Let the corresponding limiting frequency function be denoted $p_x(X)dX$, interpreted as the fraction (probability) of values of x lying in the range, $X \leq x \leq X + dX$. As a consequence of the definition, $p_x(X) \geq 0$ and,

$$\int_{\text{all } X} p_x(X)\, dX = \int_{-\infty}^{\infty} p_x(X)\, dX = 1. \tag{2.39}$$

The infinite integral is a convenient way of representing an integral over "all X", as p_x simply vanishes for impossible values of X. (It should be noted that this so-called frequentist approach has fallen out of favor, with Bayesian assumptions being regarded being ultimately more rigorous and fruitful. For present introductory purposes, however, empirical frequency functions appear to provide an adequate intuitive basis for proceeding.)

The "average," or "mean," or "expected value" is denoted $\langle x \rangle$ and defined as,

{32001}

$$\langle x \rangle \equiv \int_{\text{all } X} X p_x(X) dX \; = m_1. \tag{2.40}$$

The mean is the center of mass of the probability density. Knowledge of the true mean value of a random variable is commonly all that we are willing to assume known. If forced to "forecast" the numerical value of x under such circumstances, often the best we can do is to employ $\langle x \rangle$. If the deviation from the true mean is denoted x' so that $x = \langle x \rangle + x'$, such a forecast has the virtue that we are assured the average forecast error, $\langle x' \rangle$, would be zero if many such forecasts are made. The bracket operation is very important throughout this book; it has the property that if a is a non-random quantity, $\langle ax \rangle = a\langle x \rangle$ and $\langle ax + y \rangle = a \langle x \rangle + \langle y \rangle$.

Quantity $\langle x \rangle$ is the "first-moment" of the probability density. Higher order moments are defined as,

$$m_n = \langle x^n \rangle = \int_{-\infty}^{\infty} X^n p_x(X) dX,$$

where n are the non-negative integers. A useful theoretical result is that a knowledge of all the moments is usually enough to completely define the probability density itself. (There are troublesome situations with, e.g. non-existent moments, as with the so-called Cauchy distribution, $p_x(X) = (2/\pi)\left(1/\left(1 + X^2\right)\right)\ X \geq 0$, whose mean is infinite.) For many important probability densities, including the Gaussian, a knowledge of the first two moments $n = 1, 2$ is sufficient to

define all the others, and hence the full probability density. It is common to define the moments for $n > 1$ about the mean, so that one has,

$$\mu_n = \langle (x - \langle x \rangle)^n \rangle = \int_{-\infty}^{\infty} (X - \langle X \rangle)^n \, p_x(X) dX.$$

μ_2 is the variance and often written $\mu_2 = \sigma^2$, where σ is the "standard deviation."

2.3.2 Sample Estimates. Bias.

In observational sciences, one normally must estimate the values defining the probability density from the data itself. Thus the first moment, the mean, is often computed as the "sample average,"

$$\tilde{m}_1 = \langle x \rangle_M \equiv \frac{1}{M} \sum_{i=1}^{M} X_i. \tag{2.41}$$ {32002}

The notation $\tilde{m}_1$ is used to distinguish the sample estimate from the true value, m_1. On the other hand, if the experiment of computing $\tilde{m}_1$ from M samples could be repeated many times, the mean of the sample estimates would be the true mean. This conclusion is readily seen by considering the expected value of the difference from the true mean:

$$\begin{aligned} \langle \langle x \rangle_M - \langle x \rangle \rangle &= \left\langle \frac{1}{M} \sum_{i=1}^{M} X_i - \langle x \rangle \right\rangle \\ &= \frac{1}{M} \sum_{i=1}^{M} \langle X_i \rangle - \langle x \rangle = \frac{M}{M} \langle x \rangle - \langle x \rangle = 0. \end{aligned}$$

Such an estimate, is said to be "unbiased": its expected value is the quantity one seeks.

The interpretation is that for finite M, we do not expect that the sample mean will equal the true mean, but that if we could produce sample averages from distinct groups of observations, the sample averages would themselves have an average which will fluctuate about the true mean, with equal probability of being higher or lower. There are many sample estimates, however, some of which we encounter, where the expected value of the sample estimate is not equal to the true estimate. Such an estimator is said to be "biassed." A simple example of a biassed estimator is the "sample variance," defined as

$$s^2 \equiv \frac{1}{M} \sum_{i}^{M} (X_i - \langle x \rangle_M)^2 \tag{2.42}$$ {32003}

For reasons explained a bit later, (P. 40) one finds that

$$\langle s^2 \rangle = \frac{M-1}{M} \sigma^2 \neq \sigma^2$$

and thus the expected value is not the true variance. (This particular estimate is "asymptotically unbiased," as the bias vanishes as $M \to \infty$.)

We are assured that the sample mean is unbiased. But the probability that $\langle x \rangle_M = \langle x \rangle$, that is that we obtain exactly the true value, is very small. It helps to have a measure of the extent to which $\langle x \rangle_M$ is likely to be very far from $\langle x \rangle$. To do so, we need the idea of dispersion—the expected or average squared value of some quantity about some interesting value, like its mean. The most familiar measure of dispersion is the variance, already used above, the expected fluctuation of a random variable about its mean: {pagedispersio

$$\sigma^2 = \langle (x - \langle x \rangle)^2 \rangle .$$

More generally, define the dispersion of any random variable, z, as,

$$D^2(z) = \langle z^2 \rangle .$$

Thus, the variance of x is $D^2(x - \langle x \rangle)$.

We can thus ask for the variance of $\langle x \rangle_M$ about the correct value. A little algebra using the bracket notation produces,

{var1}
$$D^2 \left((\langle x \rangle_M - x)^2 \right) = \frac{\sigma^2}{M}. \tag{2.43}$$

This expression shows the well-known result that as M becomes large, any tendency of the sample mean to lie far from the true value will diminish. It does not prove that some particular value will not, by accident, be far away, merely that it becomes increasingly unlikely as M grows. (In statistics textbooks, the Chebyschev inequality is used to formalize this statement.)

An estimate which is unbiased and whose expected dispersion about the true value goes to zero with M is evidently desirable. In more interesting estimators, a bias is often present. Then for a fixed number of samples, M, there would be two distinct sources of deviation (error) from the true value: (1) the bias—how far, on average, it is expected to be from the true value, and (2) the tendency—from purely random events—for the value to differ from the true value (the random error). In numerous cases, one discovers that tolerating a small bias error can greatly reduce the random error—and thus the bias may well be worth accepting for that reason. In some cases therefore, a bias is deliberately introduced.

2.3.3 Functions and Sums of Random Variables

If the probability density of x is $p_x(x)$, then the mean of a function of x, $g(x)$ is just,

{32029}
$$\langle g(x) \rangle = \int_{-\infty}^{\infty} g(X) p_x(X) dX , \tag{2.44}$$

which follows from the definition of the probability density as the limit of the outcome of a number of trials.

The probability density for g regarded as a new random variable is obtained from

{32030}
$$p_g(G) = p_x(X(G)) \frac{dx}{dg} dG, \tag{2.45}$$

where dx/dg is the ratio of the differential intervals occupied by x and g and can be understood by reverting to the original definition of probability densities from histograms.

The Gaussian, or normal, probability density is one that is mathematically handy (but is potentially dangerous as a general model of the behavior of natural processes—many geophysical and fluid processes are demonstrably non-Gaussian). For a single random variable x, it is defined as,

$$p_x(X) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{(X - m_x)^2}{2\sigma_x^2}\right]$$

(sometimes abbreviated as $G(m_x, \sigma_x)$). It is readily confirmed that $\langle x \rangle = m_x$, $\langle (x - \langle x \rangle)^2 \rangle = \sigma_x^2$.

One important special case is the transformation of the Gaussian to another Gaussian of zero-mean and unit standard deviation,

$$z = \frac{x - m}{\sigma_x},$$

which can always be done, and thus,

$$p_z(Z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{Z^2}{2}\right]$$

A second important special case of a change of variable is $g = z^2$ where z is Gaussian of zero mean and unit variance. Then the probability density of g is,

{32031}
$$p_g(G) = \frac{1}{G^{1/2}\sqrt{2\pi}} \exp(-G/2)\,, \tag{2.46}$$

a special probability density usually denoted as χ_1^2 ("chi-square-sub-1"), the result for the square of a Gaussian.

2.3.4 Multivariable Probability Densities. Correlation

The idea of a frequency function generalizes easily to two or more random variables, x, y. We can, in concept, do an arbitrarily large number of experiments in which we count the occurrences of differing pair values, (X_i, Y_i), of x, y and make a histogram normalized by the total number of samples, taking the limit as the number of samples goes to infinity, and the bin sizes go to zero, to produce a joint probability density $p_{xy}(X, Y)$. $p_{xy}\,(X, Y)\,dXdY$ is then the fraction

of occurrences such that $X \leq x \leq X + dX, Y \leq y \leq Y + dY$. A simple example would be the probability density for the simultaneous measurement of the two components of horizontal velocity at a point in a fluid. Again, from the definition, $p_{xy}(X, Y) \geq 0$ and,

$$\int_{-\infty}^{\infty} p_{xy}(X, Y)\, dY = p_x(X), \tag{2.47}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{xy}(X, Y)\, dX dY = 1. \tag{2.48}$$

An important use of joint probability densities is in what is known as "conditional probability." Suppose that the joint probability density for x, y is known and furthermore, $y = Y$, that is, information is available concerning the actual value of y. What then is the probability density for x given that a particular value for y is known to have occurred? This new frequency function is usually written as $p_{x|y}(X|Y)$ and read as "the probability of x, given that y has occurred," or, "the probability of x conditioned on y." It follows immediately from the definition of the probability density that

{32004}
$$p_{x|y}(X|Y) = \frac{p_{xy}(X, Y)}{p_y(Y)} \tag{2.49}$$

(This equation is readily understood by going back to the original experimental concept, and understanding the restriction on x, given that y is known to lie within a strip paralleling the X axis).

Using the joint frequency function, define the average product as,

{32005}
$$\langle xy \rangle = \int\int_{\text{all } X,Y} XY p_{xy}(X, Y) dX\, dY\,. \tag{2.50}$$

Suppose that upon examining the joint frequency function, one finds that $p_{xy}(X, Y) = p_x(X)p_y(Y)$, that is it factors into two distinct functions. In that case, x, y are said to be "independent." Many important results follow including,

$$\langle xy \rangle = \langle x \rangle \langle y \rangle\,.$$

Non-zero mean values are often primarily a nuisance. One can always define modified variables, e.g. $x' = x - \langle x \rangle$ such that the new variables have zero mean. Alternatively, one computes statistics centered on the mean. Should the centered product $\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle$ be non-zero, x, y are said to "co-vary" or to be "correlated." If $\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle = 0$, then the two variables are "uncorrelated." If x, y are independent, then $\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle = 0$. Independence thus implies lack of correlation, but the reverse is not necessarily true. (These are theoretical relationships, and if $\langle x \rangle, \langle y \rangle$ are determined from observation, as described below, one must carefully distinguish estimated behavior from that expected theoretically.)

If the two variables are independent, then (2.49) is,

{32006}
$$p_{x|y}(X|Y) = p_x(X)\,, \tag{2.51}$$

that is, the probability of x given y does not depend upon Y, and thus

$$p_{xy}\,(X,Y) = p_x\,(X)\,p_y\,(Y)\,,$$

—and *there is then no predictive power for one variable given knowledge of the other.*

Suppose there are two random variables x, y between which there is anticipated to be some linear relationship,

$$x = ay + n, \tag{2.52}$$ {32008}

where n represents any contributions to x that remain unknown despite knowledge of y and a is a constant. Then,

$$\langle x\rangle = a\langle y\rangle + \langle n\rangle\,, \tag{2.53}$$ {32009a}

and (2.52) can be re-written as,

$$x - \langle x\rangle = a(y - \langle y\rangle) + (n - \langle n\rangle),$$

or

$$x' = ay' + n', \quad \text{where } x' = x - \langle x\rangle, \quad \text{etc.} \tag{2.54}$$ {32009b}

From this last equation,

$$a = \frac{\langle x'y'\rangle}{\langle y'^2\rangle} = \frac{\langle x'y'\rangle}{(\langle y'^2\rangle\langle x'^2\rangle)^{1/2}}\,\frac{\langle x'^2\rangle^{1/2}}{\langle y'^2\rangle^{1/2}} = \rho\,\frac{\langle x'^2\rangle^{1/2}}{\langle y'^2\rangle^{1/2}}\,, \tag{2.55}$$ {32010}

where it was supposed that $\langle y'n'\rangle = 0$, thus defining n'. The quantity

$$\rho \equiv \frac{\langle x'y'\rangle}{\langle y'^2\rangle^{1/2}\langle x'^2\rangle^{1/2}} \tag{2.56}$$ {32011}

is the "correlation coefficient" and is easily shown[22] to have the property $|\rho| \le 1$. If ρ should vanish, then so does a. If a vanishes, then knowledge of y' carries no information about the value of x'. If $\rho = \pm 1$, then it follows from the definitions that $n = 0$ and knowledge of a permits perfect prediction of x' from knowledge of y'. (Because probabilities are being used, rigorous usage would state "perfect prediction almost always," but this distinction will be ignored.)

A measure of how well the prediction of x' from y' will work can be obtained in terms of the variance of x'. We have,

$$\langle x'^2\rangle = a^2\langle y'^2\rangle + \langle n'^2\rangle = \rho^2\langle x'^2\rangle + \langle n'^2\rangle$$

or,

$$(1-\rho^2)\langle x'^2\rangle = \langle n'^2\rangle. \tag{2.57}$$ {32012}

That is, $(1-\rho^2)\langle x'^2\rangle$ is the fraction of the variance in x' that is *unpredictable* from knowledge of y' and is the "unpredictable power." Conversely, $\rho^2\langle x'^2\rangle$ is the "predictable" power in x' given knowledge of y'. The limits as $\rho \to 0, 1$ are readily apparent.

Thus we interpret the statement that two variables x', y' "are correlated" or "co-vary" to mean that knowledge of one permits at least a partial prediction of the other, the expected success of the prediction depending upon the size of ρ. If ρ is not zero, the variables cannot be independent, and the conditional probability $p_{x|y}(X|Y) \neq p_x(X)$. This result represents an implementation of the statement that if two variables are not independent, then knowledge of one permits some skill in the prediction of the other. If two variables do not co-vary, but are known not to be independent, a linear model like (2.52) would not be useful—a non-linear model would be required. Such non-linear methods are possible, and are touched on briefly later. The idea that correlation or covariance between various physical quantities carries useful predictive skill between them is an essential ingredient of many of the methods taken up in this book.

In most cases, quantities like $\rho, \left\langle x'^2 \right\rangle$, are determined from the available measurements, e.g. of the form,

$$ay_i + n_i = x_i\,, \tag{2.58}$$ {32013}

and are not known exactly. They are thus sample values, are not equal to the true values, and must be interpreted carefully in terms of their inevitable biases and variances. This large subject of regression analysis is left to the references.[23] {1}

2.3.5 Change of Variables

Suppose we have 2−random variables x, y with joint probability density $p_{xy}(X,Y)$. They are known as functions of two new variables $x = x(\xi_1,\xi_2), y = y(\xi_1,\xi_2)$ and and inverse mapping $\xi_1 = \xi_1(x,y), \xi_2 = \xi_2(x,y)$. What is the probability density for these new variables? The general rule for changes of variable in probability densities follows from area conservation in mapping from the x, y space to the ξ_1, ξ_2 space, that is,

$$p_{\xi_1\xi_2}(\Xi_1,\Xi_2) = p_{xy}(X(\Xi_1,\Xi_2),\, Y(\Xi_1,\Xi_2))\,\frac{\partial(X,Y)}{\partial(\Xi_1,\Xi_2)} \tag{2.59}$$ {32016}

where $\partial(X,Y)/\partial(\Xi_1,\Xi_2)$ is the Jacobian of the transformation between the two variable sets. As in any such transformation, one must be alert for zeros or infinities in the Jacobian, indicative of multiple valuedness in the transformation. Texts on multivariable calculus discuss such issues in great detail.

Example

Suppose x_1, x_2 *are independent Gaussian random variables of zero mean and variance* σ^2. *Then*

$$p_{\mathbf{x}}(\mathbf{X}) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{\left(X_1^2+X_2^2\right)}{2\sigma^2}\right).$$

Define new random variables

$$r = \left(x_1^2+x_2^2\right)^{1/2},\ \phi = \tan^{-1}(x_2/x_1), \tag{2.60}$$ {polar1}

whose mapping in the inverse direction is

$$x_1 = r\cos\phi,\ y_1 = r\sin\phi, \tag{2.61}$$ {polar2}

that is the mappings between polar and cartesian coordinates. The Jacobian of the transformation is $J_a = r$. *Thus*

$$p_{r,\phi}(R,\Phi) = \frac{R}{2\pi}\frac{2}{\sigma^2}\exp\left(-R^2/\sigma^2\right),\ 0\le r,\ -\pi\le\phi\le\pi \tag{2.62}$$ {polar3}

The probability density for r *alone is obtained by integrating*

$$p_r(R) = \int_{-\pi}^{\pi} p_{r,\phi}d\phi = \frac{R}{\sigma^2}\exp\left[-R^2/\left(2\sigma^2\right)\right], \tag{2.63}$$ {rayleigh1}

known as a Rayleigh distribution. By inspection then,

$$p_\phi(\Phi) = \frac{1}{2\pi},$$

which is the uniform distribution, independent of Φ. *(These results are very important in signal processing.)*

To generalize to $n-$dimensions, let there be $N-$variables, x_i, $1\le i\le N$, with known joint probablility density $p_{x_1\ldots x_N}$. Let there be $N-$new variables, ξ_i, that are known functions of the x_i, and an inverse mapping between them. Then the joint probability density for the new variables is just,

$$p_{\xi_1\ldots\xi_N}(\Xi_1,...,\Xi_N) = p_{x_1\ldots x_N}(\Xi_1(X_1,...,X_N)...,\Xi_N(X_1,....,X_N))\frac{\partial(X_1,...,X_N)}{\partial(\Xi_1,...,\Xi_N)} \tag{2.64}$$ {trans1}

Suppose that x, y are *independent* Gaussian variables $G(m_x, \sigma_x)$, $G(m_y, \sigma_y)$. Then their joint probability density is just the product of the two individual densities,

$$p_{x,y}(X,Y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{(X-m_x)^2}{2\sigma_x^2}-\frac{(Y-m_y)^2}{2\sigma_y^2}\right). \tag{2.65}$$ {32014}

Let two new random variables, ξ_1, ξ_2, be defined as a linear combination of x, y,

$$\begin{aligned}\xi_1 &= a_{11}(x - m_x) + a_{12}(y - m_y) + m_{\xi_1} \\ \xi_2 &= a_{21}(x - m_x) + a_{22}(y - m_y) + m_{\xi_2},\end{aligned} \tag{2.66}$$ {32015}

or in vector form,

$$\boldsymbol{\xi} = \mathbf{A}(\mathbf{x} - \mathbf{m}_x) + \mathbf{m}_\xi,$$

where $\mathbf{x} = [x, y]^T$, $\mathbf{m}_x = [m_x, m_y]^T$, $\mathbf{m}_y = [m_{\xi_1}, m_{\xi_2}]^T$, and the numerical values satisfy the corresponding functional relations,

$$\Xi_1 = a_{11}(X - m_x) + a_{12}(Y - m_y) + m_{\xi_1}\,,$$

etc. Suppose that the relationship (2.66) is invertible, that is, we can solve for,

$$\begin{aligned}x &= b_{11}(\xi_1 - m_{\xi_1}) + b_{12}(\xi_2 - m_{\xi_2}) + m_x \\ y &= b_{21}(\xi_1 - m_{\xi_1}) + b_{22}(\xi_2 - m_{\xi_2}) + m_y,\end{aligned}$$

or,

$$\mathbf{x} = \mathbf{B}(\boldsymbol{\xi} - \mathbf{m}_\xi) + \mathbf{m}_x. \tag{2.67}$$

Then the Jacobian of the transformation is,

$$\frac{\partial(X, Y)}{\partial(\Xi_1, \Xi_2)} = b_{11}b_{22} - b_{12}b_{21} = \det(\mathbf{B}) \tag{2.68}$$ {32018}

($\det(\mathbf{B})$ is the determinant). Eq. (2.66) produces,

$$\begin{aligned}\langle\xi_1\rangle &= m_{\xi_1} \\ \langle\xi_2\rangle &= m_{\xi_2} \\ \langle(\xi_1 - \langle\xi_1\rangle)^2\rangle &= a_{11}^2\sigma_x^2 + a_{12}^2\sigma_y^2, \quad \langle(\xi_2 - \langle\xi_2\rangle)^2\rangle = a_{21}^2\sigma_x^2 + a_{22}^2\sigma_y^2, \\ \langle(\xi_1 - \langle\xi_1\rangle)(\xi_2 - \langle\xi_2\rangle)\rangle &= a_{11}a_{21}\sigma_x^2 + a_{12}a_{22}\sigma_y^2 \neq 0\end{aligned} \tag{2.69}$$

In the special case,

$$\mathbf{A} = \begin{Bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{Bmatrix}, \qquad \mathbf{B} = \begin{Bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{Bmatrix}, \tag{2.70}$$ {32020}

the transformation (2.70) is a simple coordinate rotation through angle ϕ, and the Jacobian is 1. The new second-order moments are,

{32021a}
$$\langle(\xi_1 - \langle\xi_1\rangle)^2\rangle = \sigma^2_{\xi_1} = \cos^2\phi\,\sigma^2_x + \sin^2\phi\,\sigma^2_y\,, \tag{2.71}$$

{32021b}
$$\langle(\xi_2 - \langle\xi_2\rangle)^2\rangle = \sigma^2_{\xi_2} = \sin^2\phi\,\sigma^2_x + \cos^2\phi\,\sigma^2_y\,, \tag{2.72}$$

{32021c}
$$\langle(\xi_1 - \langle\xi_1\rangle)(\xi_2 - \langle\xi_2\rangle)\rangle \equiv \mu_{\xi_1\xi_2} = (\sigma^2_y - \sigma^2_x)\cos\phi\,\sin\phi\,. \tag{2.73}$$

The new probability density is

$$p_{\xi_1\xi_2}(\Xi_1,\Xi_2) = \frac{1}{2\pi\sigma_{\xi_1}\sigma_{\xi_2}\sqrt{1-\rho^2_\xi}} \tag{2.74}$$
$$\exp\left\{-\frac{1}{2\sqrt{1-\rho^2_\xi}}\left[\frac{(\Xi_1 - m_{\xi_1})^2}{\sigma^2_{\xi_1}} - \frac{2\rho_\xi(\Xi_1 - m_{\xi_1})(\Xi_2 - m_{\xi_2})}{\sigma_{\xi_1}\sigma_{\xi_2}} + \frac{(\Xi_2 - m_{\xi_2})^2}{\sigma^2_{\xi_2}}\right]\right\}$$

where $\rho_\xi = (\sigma^2_y - \sigma^2_x)\sin\phi\cos\phi/\left(\sigma^2_x + \sigma^2_y\right)^{1/2} = \mu_{\xi_1\xi_2}/\sigma_{\xi_1}\sigma_{\xi_2}$ is the correlation coefficient of the new variables. A probability density derived through a linear transformation from two independent variables which are Gaussian will be said to be "jointly Gaussian" and (2.74) is a canonical form. Because a coordinate rotation is invertible, it is important to note that if we had two random variables ξ_1, ξ_2 which were jointly Gaussian with $\rho \neq 1$, then we could find a pure rotation (2.70), which produces two other variables x, y which are uncorrelated, and therefore *independent*. Notice that (2.73) shows that two such uncorrelated variables x, y will necessarily have different variances, otherwise ξ_1, ξ_2 would have zero correlation, too, by Eq. (2.73) . {pagegauss}

As an important by-product, it is concluded that two jointly Gaussian random variables that are uncorrelated, are also independent. This property is one of the reasons Gaussians are so nice to work with; but it is not generally true of uncorrelated variables.

Vector Random Processes

Simultaneous discussion of two random processes, x, y can regarded as discussion of a vector random process $[x,\, y]^T$, and suggests a generalization to N dimensions. Let us label N random processes as x_i and define them as the elements of a vector $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$. Then the mean is a vector: $\langle\mathbf{x}\rangle = \mathbf{m}_x$, and the covariance is a matrix:

$$\mathbf{C}_{xx} = D^2(\mathbf{x} - \langle\mathbf{x}\rangle) = \langle(\mathbf{x} - \langle\mathbf{x}\rangle)(\mathbf{x} - \langle\mathbf{x}\rangle)^T\rangle\,, \tag{2.75}$$ {32023}

which is necessarily symmetric and positive semi-definite. The cross-covariance of two vector processes $\mathbf{x}$, $\mathbf{y}$ is,

$$\mathbf{C}_{xy} = \langle(\mathbf{x} - \langle\mathbf{x}\rangle)(\mathbf{y} - \langle\mathbf{y}\rangle)^T\rangle, \tag{2.76}$$ {32024}

and $\mathbf{C}_{xy} = \mathbf{C}_{yx}^T$.

It proves convenient to introduce two further moment matrices in addition to the covariance matrices. The "second moment" matrices will be defined as,

$$\mathbf{R}_{xx} \equiv D^2(\mathbf{x}) = \langle \mathbf{x}\mathbf{x}^T \rangle, \qquad \mathbf{R}_{xy} = \langle \mathbf{x}\mathbf{y}^T \rangle,$$

that is, not taken about the means. Note $\mathbf{R}_{xy} = \mathbf{R}_{yx}^T$, etc. Let $\tilde{\mathbf{x}}$ be an "estimate" of the true value, $\mathbf{x}$. Then the dispersion of $\tilde{\mathbf{x}}$ about the true value will be called the "uncertainty" (sometimes it is called the "error covariance") and is

$$\mathbf{P} \equiv D^2(\tilde{\mathbf{x}} - \mathbf{x}) = \langle (\tilde{\mathbf{x}} - \mathbf{x})(\tilde{\mathbf{x}} - \mathbf{x})^T \rangle \,.$$

$\mathbf{P}$ is similar to $\mathbf{C}$, but differs in being taken about the true value, rather than about the mean value; the distinction can be very important.

If there are N variables, ξ_i, $1 \le i \le N$, they will be said to have an "N-dimensional jointly normal probability density" if it is of the form,

{32025}
$$p_{\xi_1,\ldots,\xi_N}(\Xi_1, \ldots, \Xi_N) = \frac{\exp\left[-\frac{1}{2}(\mathbf{\Xi} - \mathbf{m})^T \mathbf{C}_{\xi\xi}^{-1} (\mathbf{\Xi} - \mathbf{m})\right]}{(2\pi)^{N/2}\sqrt{\det(\mathbf{C}_{\xi\xi})}} \,. \tag{2.77}$$

One finds $\langle \boldsymbol{\xi} \rangle = \mathbf{m}$, $\langle (\boldsymbol{\xi} - \mathbf{m})(\boldsymbol{\xi} - \mathbf{m})^T \rangle = \mathbf{C}_{\xi\xi}$. Eq. 2.74 is a special case for $N = 2$, and so the earlier forms are consistent with this general definition.

Positive definite symmetric matrices can be factored as,

{32026}
$$\mathbf{C}_{\xi\xi} = \mathbf{C}_{\xi\xi}^{T/2} \mathbf{C}_{\xi\xi}^{1/2} \,, \tag{2.78}$$

called the "Cholesky decomposition," where $\mathbf{C}_{\xi\xi}^{1/2}$ is an upper triangular matrix (all zeros below
{pagechol} the main diagonal) and non-singular.[24]. It follows that the transformation (a rotation and stretching),

{32027}
$$\mathbf{x} = \mathbf{C}_{\xi\xi}^{-T/2} (\boldsymbol{\xi} - \mathbf{m}) \,, \tag{2.79}$$

produces new variables $\mathbf{x}$ of zero mean, and diagonal covariance, that is, a probability density

{32028}
$$p_{x_1,\ldots,x_N}(X_1, \ldots X_N) = \frac{\exp -\frac{1}{2}(X_1^2 + \cdots + X_N^2)}{(2\pi)^{N/2}} = \frac{\exp\left(-\frac{1}{2}X_1^2\right)}{(2\pi)^{1/2}} \cdots \frac{\exp\left(-\frac{1}{2}X_N^2\right)}{(2\pi)^{1/2}} \,, \tag{2.80}$$

which factors into N-independent, normal variates of zero mean and unit variance ($\mathbf{C}_{xx} = \mathbf{R}_{xx} = \mathbf{I}$). Such a process is often called Gaussian "white noise," and has the property $\langle x_i x_j \rangle = 0$, $i \neq j$.[25]

2.3.6 Sums of Random Variables

It is often helpful to be able to compute the probability density of sums of independent random variables. The procedure for doing so is based upon (2.44). Let x be a random variable and consider the expected value of the function e^{ixt}:

{32032}
$$\langle e^{ixt} \rangle = \int_{-\infty}^{\infty} p_x\left(X\right) e^{iXt} dX \equiv \phi_x(t)\,, \tag{2.81}$$

which is the Fourier transform of $p_x(X)$; $\phi_x(t)$ is usually termed the "characteristic function" of x. Now consider the sum of two independent random variables x, y with probability densities p_x, p_y, respectively, and define a new random variable $z = x+y$. What is the probability density of z? One starts by first determining the characteristic function, $\phi_z(t)$ for z and then using the Fourier inversion theorem to obtain $p_x(Z)$. To obtain ϕ_z,

$$\phi_z(t) = \langle e^{izt} \rangle = \langle e^{i(x+y)t} \rangle = \langle e^{ixt} \rangle \langle e^{iyt} \rangle$$

where the last step depends upon the independence assumption. This last equation shows

$$\phi_z(t) = \phi_x(t)\phi_y(t)\,. \tag{2.82}$$ {32033}

That is, the characteristic function for a sum of two independent variables is the product of the characteristic functions. The "convolution theorem"[26] asserts that the Fourier transform (forward or inverse) of a product of two functions is the convolution of the Fourier transforms of the two functions. That is,

$$p_z\left(Z\right) = \int_{-\infty}^{\infty} p_x\left(r\right) p_y\left(Z-r\right) dr. \tag{2.83}$$ {conv}

We will not explore this relation in any detail, leaving the reader to pursue the subject in the references.[27] But it follows immediately that the multiplication of the characteristic functions of a sum of independent Gaussian variables produces a new variable, which is also Gaussian, with a mean equal to the sum of the means and a variance which is the sum of the variances ("sums of Gaussians are Gaussian"). It also follows immediately from Eq. 2.82) that if a variable ξ is defined as,

$$\xi = x_1^2 + x_2^2 + \cdots + x_\nu^2, \tag{2.84}$$ {32034}

where each x_i is Gaussian of zero mean and unit variance, that the probability density for ξ is,

$$p_\xi(\Xi) = \frac{\Xi^{\nu/2-1}}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} \exp(-\Xi/2)\,, \tag{2.85}$$ {32035}

known as χ^2_ν —"chi-square sub- ν." The chi-square probability density is central to the discussion of the expected sizes of vectors, such as $\tilde{\mathbf{n}}$, measured as $\tilde{\mathbf{n}}^T\tilde{\mathbf{n}} = \|\tilde{\mathbf{n}}\|^2 = \sum_i \tilde{n}_i^2$ if the elements of $\tilde{\mathbf{n}}$ can be assumed to be independent and Gaussian. Eq. (2.46) is the special case $\nu = 1$.

Degrees-of-Freedom

The number of independent variables described by a probability density is usually called the "number of degrees-of-freedom." Thus the densities in (2.77) and (2.80) have N-degrees of freedom and (2.85) has ν of them. If a sample average (2.41) is formed, it is said to have N-degrees of freedom if each of the x_j is independent. But what if the x_j have a covariance $\mathbf{C}_{xx}$ which is non-diagonal? This question of how to interpret averages of correlated variables will be explicitly discussed on P. 135.

Consider the special case of the sample variance Eq. (2.42)—which we claimed was biassed. The reason is that even if the sample values, x_i, are independent, the presence of the sample average in the sample variance means that there are only $N-1$ independent terms in the sum. That this is so is most readily seen by examining the two-term case. Two samples produce a sample mean, $\langle x\rangle_2 = (x_1 + x_2)/2$. The two-term sample variance is,

$$s^2 = \tfrac{1}{2}\left[(x_1 - \langle x\rangle_2)^2 + (x_2 - \langle x\rangle_2)^2\right],$$

but knowledge of x_1 and of the sample average, permits perfect prediction of $x_2 = 2\langle x\rangle_2 - x_1$. The second term in the sample variance as written is not independent of the first term, and thus there is just one independent piece of information in the two-term sample variance. To show it in general, assume without loss of generality that $\langle x\rangle = 0$, so that $\sigma^2 = \langle x^2\rangle$. The sample variance about the sample mean (which will not vanish) of independent samples is given by Eq. (2.42) and so,

{page39}

$$\begin{aligned}
\langle s^2\rangle &= \frac{1}{M}\sum_{i=1}^{M}\left\langle\left(x_i - \frac{1}{M}\sum_{j=1}^{M}x_j\right)\left(x_i - \frac{1}{M}\sum_{p=1}^{M}x_p\right)\right\rangle \\
&= \frac{1}{M}\sum_{i=1}^{M}\left\{\langle x_i^2\rangle - \frac{1}{M}\sum_{j=1}^{M}\langle x_i x_j\rangle - \frac{1}{M}\sum_{p=1}^{M}\langle x_i x_p\rangle + \frac{1}{M^2}\sum_{j=1}^{M}\sum_{p=1}^{M}\langle x_j x_p\rangle\right\} \\
&= \frac{1}{M}\sum_{i=1}^{M}\left\{\sigma^2 - \frac{\sigma^2}{M}\sum_j \delta_{ij} - \frac{\sigma^2}{M}\sum_p \delta_{ip} + \frac{\sigma^2}{M^2}\sum_j\sum_p \delta_{jp}\right\} \\
&= \frac{\sigma^2(M-1)}{M} \neq \sigma^2
\end{aligned}$$

Stationarity

Consider a vector random variable, with element x_i where the subscript i denotes a position in time or space. Then x_i, x_j are two different random variables—for example, the temperature at two different positions in a moving fluid, or the temperature at two different times at the same position. If the physics governing these two different random variables are independent of the parameter i (i.e., independent of time or space), then x_i is said to be "stationary"—meaning that all the underlying statistics are independent of i.[28] Specifically, $\langle x_i \rangle = \langle x_j \rangle \equiv \langle x \rangle$, $D^2(x_i) = D^2(x_j) = D^2(x)$, etc. Furthermore, x_i, x_j have a covariance

$$C_{xx}(i,j) = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \,, \tag{2.86}$$ {32036}

that is, independent of i, j, and might as well be written $C_{xx}(|i-j|)$, depending only upon the difference $|i-j|$. The distance, $|i-j|$, is often called the "lag." $C_{xx}(|i-j|)$ is called the "autocovariance" of $\mathbf{x}$ or just the covariance, because we now regard x_i, x_j as intrinsically the same process.[29] If C_{xx} does not vanish, then by the discussion above, knowledge of the numerical value of x_i implies some predictive skill for x_j and vice-versa—a result of great importance when we examine map-making and objective analysis. For stationary processes, all elements having the same $|i-j|$ are identical; it is seen that all diagonals of such a matrix $\{C_{xx}(i,j)\}$, are constant, for example, $\mathbf{C}_{\xi\xi}$ in Eq. (2.77). Matrices with constant diagonals are thus defined by the vector $C_{xx}(|i-j|)$, and are said to have a "Toeplitz form."