Figure 5.4. PROBABILITY MODELLING: Linear Combinations of Random Variables

The matters presented in this Figure extend appreciably the types of probability calculations we can undertake.

Suppose the random variable T is given by: $T = \mathrm{a}U + \mathrm{b}V + \mathrm{c}W$ where: U, V, W are random variables, and: a, b, c are given constants.

We call T a *linear combination* of U, V and W; we confine *our* attention to three special cases of linear combinations:
– a sum (*e.g.*, when a = b = c = 1); – a difference (*e.g.*, when a = 1, b = –1, c = 0); – an average (*e.g.*, when a = b = c = ⅓).

To describe the probabilistic behaviour of T, we need to know *three* of its characteristics:
- its mean;
- its standard deviation;
- its distribution.

We want to relate these characteristics of T to the corresponding characteristics of U, V and W; in general, the mean is the *easiest* to deal with, the distribution the *hardest*. We state the following results without proof.

1. A Sum or Difference

- **Mean:** $E(\mathrm{a}U + \mathrm{b}V + \mathrm{c}W) = \mathrm{a}E(U) + \mathrm{b}E(V) + \mathrm{c}E(W).$ -----(5.4.1)

Examples: $E(U + V + W) = E(U) + E(V) + E(W)$
and similarly for sums of *more* than three random variables;
$E(U + V) = E(U) + E(V)$
$E(U - V) = E(U) - E(V).$

Thus, the mean of a sum or difference is the same sum or difference of the individual means; this is the analogue of the corresponding familiar behaviour of averages.

- **S.d. (and variance):** $s.d.(\mathrm{a}U + \mathrm{b}V + \mathrm{c}W) = \sqrt{\mathrm{a}^2 \cdot s.d.(U)^2 + \mathrm{b}^2 \cdot s.d.(V)^2 + \mathrm{c}^2 \cdot s.d.(W)^2}.$ -----(5.4.2)

Examples:
$s.d.(U + V + W) = \sqrt{s.d.(U)^2 + s.d.(V)^2 + s.d.(W)^2}$
$var(U + V + W) = var(U) + var(V) + var(W)$
and similarly for sums of *more* than three random variables;
$s.d.(U + V) = \sqrt{s.d.(U)^2 + s.d.(V)^2}$
$var(U + V) = var(U) + var(V)$
$s.d.(U - V) = \sqrt{s.d.(U)^2 + s.d.(V)^2}$
$var(U - V) = var(U) + var(V).$

provided U, V, W are *probabilistically independent* random variables

Thus, *provided* the individual random variables are *probabilistically independent*, the standard deviation of a *sum* is the square root of the sum of the individual standard deviations squared, and the standard deviation of a *difference* is this *same* sum. This latter result reminds us that the values arising from a measuring process for a *difference* generally show *more* variation than the values which yield the differences; this behaviour has important implications for the precision of the laboratory procedure of weighing liquids, for example, by *difference*.

NOTES: 1. As the expressions above indicate, standard deviations can be combined *only* by squaring, adding, and taking the (overall) square root; a memory aid is to say *standard deviations add like Pythagoras.*

2. If the individual random variables are *not* probabilistically independent, the s.d. (and variance) expressions above need additional term(s) involving a quantity called *covariance* (see the Appendix overleaf on page 5.12); dealing with standard deviations for *probabilistically dependent* random variables is beyond our present concern.

3. Random variables are (mutually) probabilistically independent if their *joint* probability density function can be written as the *product* of the probability density functions of the individual random variables.

- **Distribution**: if U, V and W have independent *Gaussian* distributions, T *also* has a Gaussian distribution. -----(5.4.3)

[We limit consideration to cases involving independent *Gaussian* random variables because, for many *other* distributions, there is no simple relationship between the distributions of U, V and W and that of T.]

2. An Average

For a data set consisting of the values ($\mathrm{y_j}$) for some response variate of n units, we are familiar with the average ($\overline{\mathrm{y}}$) as given in equation (5.4.4) at the right. If the n units have been selected *equiprobably* from the (study or respondent) population, so each $\mathrm{y_j}$ can be regarded as y_j, the value of random variable Y_j, the random variable representing the sample average $\overline{y}$ is:

$$\overline{\mathrm{y}} = \frac{\sum_{j=1}^{n} \mathrm{y_j}}{\mathrm{n}}$$ -----(5.4.4)

(*continued overleaf*)

$$\overline{Y} = \frac{Y_1+Y_2+Y_3+.....+Y_n}{n} \equiv \frac{\sum_{j=1}^{n} Y_j}{n} \equiv \frac{1}{n}(Y_1+Y_2+Y_3+.....+Y_n). \qquad \text{-----(5.4.5)}$$

Thus, the random variable $\overline{Y}$ is a *sum* of n random variables, multiplied by (a constant) 1/n.

We now place four *restrictions* on the random variables Y_1, Y_2, Y_3, ..., Y_n:

- ⁎ they all have the same *mean* (μ);
- ⁎ they all have the same *standard deviation* (σ);
- ⁎ they are mutually *probabilistically independent*;
- ⁎ they each have a *Gaussian* distribution.

[These restrictions may be met in practice, at least to a reasonable degree, in the case of:

- ○ repeated independent measurements of the same quantity;
- ○ variate values for units selected equiprobably.]

Then, applying to the expression (5.4.5) for $\overline{Y}$ the results (5.4.1), (5.4.2) and (5.4.3) given overleaf on page 5.11:

- **Mean:** $E(\overline{Y}) = \frac{1}{n}(\mu+\mu+.....+\mu) = \frac{1}{n}(n\mu) = \mu;$ -----(5.4.6)
- **S.d.:** $s.d.(\overline{Y}) = \frac{1}{n}\sqrt{\sigma^2+\sigma^2+.....+\sigma^2} = \frac{1}{n}\sqrt{n\sigma^2} = \sigma\sqrt{\frac{1}{n}};$ -----(5.4.7)
- **Distribution:** $\overline{Y}$ has a *Gaussian* distribution; -----(5.4.8)

$$\left.\begin{array}{l}(5.4.6)\\(5.4.7)\\(5.4.8)\end{array}\right\} \; i.e., \; \overline{Y} \sim G(\mu, \sigma\sqrt{\tfrac{1}{n}}); \qquad \text{-----(5.4.9)}$$

IN WORDS: (the random variable representing) the average of n independent $G(\mu, \sigma)$ random variables has a *Gaussian* distribution with the *same* mean as the individual variables but with *root one over* n of their standard deviation.

NOTES: 4. Equation (5.4.9) for $\overline{Y}$ is the reason why, when we have repeated independent measurements of a quantity, the 'best' estimate of the value of the quantity is the *average* of the measurements – the average has the same mean as the individual measurements but one root n*th* of their standard deviation (*i.e.*, the measuring process for the *average* is *root n* times *less* imprecise than the process for the *individual* measurements).

5. The discussion in Note 4 shows that, to decrease the imprecision of an average by a factor of 3, say, we need to take 9 (*not* 3) times as many observations from which to calculate the average; *i.e.*, imprecision decreases only as the *square root* of the number of (independent) repetitions of a measuring process.

6. The standard deviation of $\overline{Y}$, $\sigma\sqrt{\frac{1}{n}}$, is sometimes called *the standard error of the mean* and abbreviated *S.E.M.*; *if* such a term (with 'standard error' rather than 'standard deviation') is to be used, *we* would prefer to call it the standard error of the *average* (*S.E.A.*) but, unfortunately, use of *S.E.M.* is well-established.

3. Appendix: Covariance

The *covariance* of the random variables U and V, mentioned overleaf in Note 2 on page 5.11, is defined in equation (5.4.10) at the right; it is a measure of relationship between U and V, in the sense of quantifying their degree of *probabilistic dependence*. Covariance takes values in the interval $(-\infty, \infty)$.

The more general expression than equation (5.4.2) overleaf on page 5.11 for the standard deviation of a linear combination of *two* (probabilistically dependent) random variables is equation (5.4.11).

$$cov(U,V) = E(UV) - E(U)E(V) \qquad \text{-----(5.4.10)}$$

$$s.d.(aU+bV) = \sqrt{a^2 \cdot s.d.(U)^2 + b^2 \cdot s.d.(V)^2 + 2ab \cdot cov(U,V)} \qquad \text{-----(5.4.11)}$$

$$cor(U,V) = \frac{cov(U,V)}{s.d.(U) \cdot s.d.(V)} \qquad \text{-----(5.4.12)}$$

$$s.d.(U) = \sqrt{E[U-E(U)]^2} \qquad \text{-----(5.4.13)}$$

$$s.d.(V) = \sqrt{E[V-E(V)]^2} \qquad \text{-----(5.4.14)}$$

Two other comments about covariance are:

- Covariance is involved in the more familiar measure of (linear) relationship called (probabilistic) *correlation* [see equation (5.4.12) at the right above], which has the convenience over covariance that it takes values in the interval [–1, 1].
  - (Data) correlation is discussed in detail in Figure 4.5 of these Course Materials.
- When the random variables U and V are *probabilistically independent*, $E(UV)$ factors into $E(U)E(V)$ because the *joint* probability density function of U and V factors into the product of the (marginal) probability density functions of U and V – see Note 3 overleaf on page 5.11; hence, $cov(U,V) = 0$ when U and V are independent random variables. However, the converse is *not* true – zero covariance does *not* necessarily imply probabilistic independence.

1 Discuss briefly the implications, mentioned overleaf on page 5.11, for weighing *small* liquid samples by difference in an analytical laboratory; include a *quantitative* illustration in your discussion.

2 *Independent* measurements of the same quantity are mentioned in the *first* of the two circled (○) points near the top of the page. Discuss briefly the factors which determine whether repeated measurements *can* reasonably be considered independent.

3 Independence is not mentioned *explicitly* in the *second* circled (○) point above. Explain briefly whether this means that independence is not required in this context or whether it enters in another guise.