Some basic statistical concepts

A.S. Lawless

University of Reading

Random variable A random variable is a variable which takes on values at random.

Probability distribution function A probability distribution function $P(x)$ describes the probability that x will take on a certain value. Thus the probability that x lies between x_1 and x_2 is given by

$$\int_{x_1}^{x_2} P(x)dx. \tag{1}$$

Expectation value Suppose that a random variable x can take on all values between $-\infty$ and ∞. Then the expectation value of x is given by

$$< x >= \int_{-\infty}^{\infty} xP(x)dx, \tag{2}$$

where $P(x)$ is the probability distribution function of x. The expectation value is a generalization of the mean. While the mean is calculated from a sum over a real data sample, the expectation value sums over a theoretical probability distribution. If a data sample is described by a theoretical distribution then as the size of the data sample tends to infinity, the mean tends to the expectation value. The definitions which follow can be applied to a finite data sample by replacing the expectation value with the arithmetic mean.

We note the properties

$$< x + y >=< x > + < y >, \tag{3}$$

but in general $< xy > \neq < x >< y >$.

Gaussian distribution The Gaussian distribution function (also known as the *normal* distribution) is a particularly important probability distribution function. It takes the form

$$P(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}. \tag{4}$$

It is a bell-shaped curve centred on $x = \mu$, with the width determined by σ. We find that μ is equal to the expectation value (or mean) of x and σ is the *standard deviation* of the distribution (see later definition).

The Gaussian distribution is important, since it describes well the distribution of errors, an important part of data assimilation. We often assume that errors have a Gaussian distribution.

Variance The variance of x, $V(x)$, is given by

$$V(x) = < (x - < x >)^2 > \tag{5}$$

$$= < x^2 > - < x >^2 . \tag{6}$$

The variance is a measure of the spread of x around the expectation value $< x >$.

Standard deviation The standard deviation is simply the square root of the variance and is usually denoted by the symbol σ, so that

$$\sigma = \sqrt{V(x)} \tag{7}$$

$$= \sqrt{< (x - < x >)^2 >}. \tag{8}$$

For a Gaussian distribution:

68.27% of the area lies with σ of the mean

95.45% of the area lies with 2σ of the mean

99.73% of the area lies with 3σ of the mean

Covariance Let x, y be two random variables. Then the covariance between x and y is defined as

$$\mathrm{cov}(x, y) = < (x - < x >)(y - < y >) > \tag{9}$$

$$= < xy > - < x >< y > . \tag{10}$$

The covariance measures the dependence between the two variables. If values of x above the expected value have a tendency to occur with values of y above the expected value, then both terms in (9) will have the same sign and the covariance will be positive. A similar situation occurs if both have lower than expected values together. If however values of x above the expected value occur with values of y below the expected value, then the terms will have the opposite sign and the covariance will be negative. If the variables x and y are independent then $x - < x >$ has an equal chance of being multiplied by a positive or negative $y- < y >$ and the covariance will be zero.

We note also that

$$\text{cov}(x, x) = V(x). \tag{11}$$

Covariance matrix Suppose we have n random variables $x_{(1)}, \ldots, x_{(n)}$. Then we can define a covariance between any two variables by

$$\text{cov}(x_{(i)}, x_{(j)}) =< (x_{(i)} - < x_{(i)} >)(x_{(j)} - < x_{(j)} >) > . \tag{12}$$

Then we can easily see that these covariances form an $n \times n$ matrix with entries

$$V_{ij} = \text{cov}(x_{(i)}, x_{(j)}). \tag{13}$$

This matrix is known as the *covariance matrix*. We note two important properties of this matrix:

1. The covariance matrix is symmetric, since $\text{cov}(x_{(i)}, x_{(j)}) = \text{cov}(x_{(j)}, x_{(i)})$.
2. Using (11) we see that the diagonal entries of the covariance matrix are just the variances.

Correlation coefficient The correlation coefficient ρ is a version of the covariance, normalized by the standard deviations to give a dimensionless quantity. It is defined for two variables x, y by

$$\rho(x, y) = \frac{\text{cov}(x, y)}{\sigma(x)\sigma(y)}. \tag{14}$$

The correlation coefficient varies beteen -1 and 1. If $\rho = 0$ then the variables are independent and are said to be uncorrelated. If $\rho = -1$ or $\rho = 1$ then the variables are completely correlated and one can be determined from the other.

Notes on statistics of errors

Let us suppose that $T_0(\mathbf{r}, t)$ is some variable which we are trying to measure (eg. temperature) and we have an estimate $T_e(\mathbf{r}, t)$ which has error $\epsilon(\mathbf{r}, t)$. Hence

$$T_e(\mathbf{r}, t) = T_0(\mathbf{r}, t) + \epsilon(\mathbf{r}, t). \tag{15}$$

Then we say that

- The measurement is *unbiased* if $< \epsilon(\mathbf{r}, t) >= 0$.

- The error is not spatially correlated if $< \epsilon(\mathbf{r}_i, t)\epsilon(\mathbf{r}_j, t) >= 0$ for $i \neq j$.

- The error is not temporally correlated if $< \epsilon(\mathbf{r}, t_1)\epsilon(\mathbf{r}, t_2) >= 0$ for $t_1 \neq t_2$.

References

Barlow, R.J. (1989), *Statistics - A guide to the use of statistical methods in the physical sciences*, John Wiley and Sons.

Daley, R (1991), *Atmospheric Data Analysis*, Cambridge University Press.