Random variables

§1. Random variable (or process)

A random variable is a variable whose value is subject to variations due to chance (i.e. randomness, in a mathematical sense). As opposed to other mathematical variables, a random variable conceptually does not have a single, fixed value (even if unknown); rather, it can take on a set of possible different values, each with an associated probability. Conceptually, this "uncertainty" can be the result of *incomplete* knowledge (ex. the number of jelly beans in a jar that you are not allowed to open), *imprecise* measurements (ex. you don't hold the thermometer under your tongue in exactly the same way each time), or *intrinsic* variability (ex. randomness inherent to rolling a die). A random variable's possible values might represent the possible outcomes of a yet-to-be-performed experiment, or the possible outcomes of a past experiment whose already-existing value is uncertain. Examples are reaction times from a (future or past) behavioral task, or neuronal firing rates (to be measure or already recorded) in response to a stimulus, etc.

The mathematical function describing the possible values of a random variable and their associated probabilities is known as a probability distribution. A probability distribution can either be univariate or multivariate. A univariate distribution gives the probabilities of a single random variable taking on various alternative values; a multivariate distribution (a joint probability distribution) gives the probabilities of a random vector—a set of two or more random variables—taking on various combinations of values.

Random variables can be discrete, continuous, or a mixture of both types.

Discrete random variable

A discrete random variable takes any of a specified finite or countably infinite ("countable") list of values, and has a **probability mass function** (**pmf**) as its probability distribution. The probability mass function is a function that gives the probability that a discrete random variable is exactly equal to some value (e.g., Pr(X=u) = p_u).

For a discrete random variable,

$$\sum_u \Pr(X = u) = 1$$

as *u* runs through the set of all possible values of *X*.

The probability mass function is often the primary means of defining a discrete probability distribution, and such functions exist for either scalar or multivariate random variables whose *domain* is discrete. Among the most well-known discrete probability distributions that are used for statistical modeling are the Poisson distribution, the Bernoulli distribution, the binomial distribution, the geometric distribution, and the negative binomial distribution

Continuous random variable

Intuitively, a continuous random variable is one that can take a continuous range of values (as opposed to a discrete distribution, where the set of possible values for the random variable is at most countable). Formally, a continuous random variable takes any numerical value in an interval or collection of intervals, and has a **probability density function** (**pdf**) describing the probability distribution. The probability density function, or density, is a function that describes the relative likelihood for a continuous random variable to take on a given value. The values of a continuous pdf are not probabilities as such: *a pdf must be integrated over an interval to yield a probability.* Thus, the probability of the random variable falling within a particular range of values (say a to b) is

$$\Pr[a \leq X \leq b] = \int_a^b f(x)\,dx$$

This corresponds to the area under the density function but above the horizontal axis and between a and b. The probability density function is nonnegative everywhere, and its integral over the entire space is equal to one.

While for a discrete distribution an event with probability zero corresponds to an impossible event (e.g. rolling 3½ on a standard die is impossible, and has probability zero), this is not so in the case of a continuous random variable. For example, if one measures the width of an oak leaf, the result of 3½ cm is possible, however it has probability zero because there are uncountably many other potential values even between 3 cm and 4 cm. Each of these individual outcomes has probability zero, yet the probability that the outcome will fall into the interval (3 cm, 4 cm) is nonzero. This apparent paradox is resolved by the fact that the probability that *X* attains some value within an infinite set, such as an interval, cannot be found by naively adding the probabilities for individual values. Formally, each value has an infinitesimally small probability, which statistically is equivalent to zero. In particular, the probability for *X* to take any single value *a* (that is $a \leq X \leq a$) is zero, because an integral with coinciding upper and lower limits is always equal to zero.

Cumulative distribution function

The cumulative distribution function (**CDF**), or just distribution function, describes the probability that a real-valued random variable *X* with a given probability distribution will be found at a value less than or equal to *x*.

$$F(x) = \Pr[X \leq x] \qquad \text{for all } x \in \mathbb{R}.$$

If *X* is a purely discrete random variable, then it attains values x_1, x_2, ... with probability $p_i = P(x_i)$, and the CDF of *X* will be:

$$F(x) = \mathrm{P}(X \leq x) = \sum_{x_i \leq x} \mathrm{P}(X = x_i) = \sum_{x_i \leq x} p(x_i).$$

Note that this function will look like a series of steps, with steps occurring at the points x_i.

The CDF of a continuous random variable *X* can be expressed as the integral of its probability density function f_X, i.e, the area under the probability density function from minus infinity to *x*.

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt.$$

Example 1: Suppose that *X* takes only the discrete values 0 and 1, with equal probability. Then the CDF of X is given by

$$F(x) = \begin{cases} 0 & : \ x < 0 \\ 1/2 & : \ 0 \leq x < 1 \\ 1 & : \ 1 \leq x. \end{cases}$$

Example 2: Suppose *X* is uniformly distributed on the unit interval [0, 1]. Then the CDF of X is given by

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

Asides

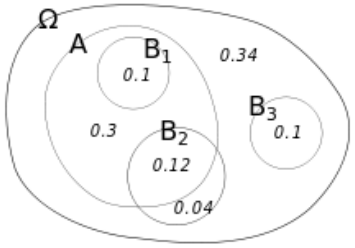
1. The basic concept of "random variable" in statistics is real-valued, and therefore expected values, variances and other measures can be computed. However, one can consider arbitrary types of such as boolean values (binary; 0=FALSE, 1=TRUE), categorical variables, complex numbers, vectors, matrices, sequences, trees (graph theory), sets, shapes, manifolds, functions, and processes. The term *random element* is used to encompass all such related concepts.
2. Symmetry is a property of some distributions in which the portion of the distribution to the left of a specific value is a mirror image of the portion to its right. Skewness: a measure of the extent to which a pmf or pdf "leans" to one side of its mean.
3. A probability distribution on the real line is **completely characterized** by its cumulative distribution function.

Conditional probability

P(X∩Y) = P(X/Y) P(Y) = P(X) P(Y/X)

(Read: P of A intersection B = P of X given Y times P of Y)

Illustration of conditional probabilities with an Euler diagram. The unconditional probability $P(A) = 0.52$. However, the conditional probability $P(A \mid B_1) = 1$, $P(A \mid B_2) \approx 0.75$, and $P(A \mid B_3) = 0$.

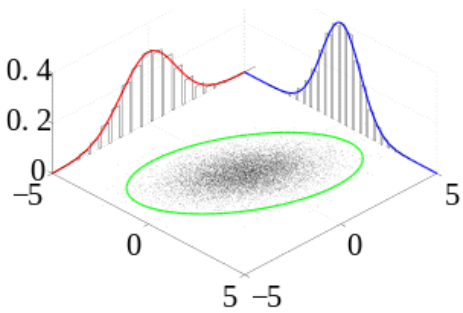


Joint distributions

Given at least two random variables *X*, *Y*, ..., that are defined on a probability space, **the joint probability distribution** for *X*, *Y*, ... is a probability distribution that gives the probability that each of *X*, *Y*, ... falls in any particular range or discrete set of values specified for that variable. In the case of only two random variables, this is called a bivariate distribution, but the concept generalizes to any number of random variables, giving a multivariate distribution.

The joint probability distribution can be expressed either in terms of a joint cumulative distribution function or in terms of a joint probability density function (in the case of continuous variables) or joint probability mass function (in the case of discrete variables). These in turn can be used to find two other types of distributions: the marginal distribution giving the probabilities for any one of the variables with no reference to any specific ranges of values for the other variables, and the conditional probability distribution giving the probabilities for any subset of the variables conditional on particular values of the remaining variables.

The joint probability distribution for a pair of random variables can be expressed



in terms of their cumulative distribution function

$$F(x, y) = P(X \leq x, Y \leq y).$$

For two discrete random variables, the joint probability mass function is equal to:

$$\mathrm{P}(X = x \text{ and } Y = y) = \mathrm{P}(Y = y \mid X = x) \cdot \mathrm{P}(X = x) = \mathrm{P}(X = x \mid Y = y) \cdot \mathrm{P}(Y = y).$$

For continuous random variables, the **joint probability density function** $f_{X,Y}(x, y)$ is equal to:

$$f_{X,Y}(x, y) = f_{Y|X}(y|x) f_X(x) = f_{X|Y}(x \mid y) f_Y(y)$$

…where $f_{Y|X}(y|x)$ and $f_{X|Y}(x|y)$ give the conditional distributions of Y given $X = x$ and of X given $Y = y$ respectively, and $f_X(x)$ and $f_Y(y)$ give the marginal distributions for X and Y respectively.

Jointly normal:

Two random variables X and Y are said to be jointly normal if they can be expressed in the form

X = aU + bV,

Y = cU + dV,

where U and V are independent normal random variables.

Many sample observations (black) are shown from a joint probability distribution. The marginal densities are shown as well (in red and blue).

Mean

The mean of a random variable provides the *long-run average* of the variable, or the expected average outcome over many observations, i.e., the weighted average of the possible values. For a discrete random variable *X*, the mean is a weighted average of the possible values that the random variable can take. Unlike the sample mean of a group of observations, which gives each observation equal weight, the mean of a random variable weights each outcome x according to its probability p(x). Common symbols for the mean (also known as the expected value of X) are μ, or E(X).

The mean of a discrete random variable X which takes on values x_i and has a pmf $p(x_i)=p_i$ is

$$\mu = \sum_{i=1}^{n} p_i \cdot x_i$$

The mean of a continuous random variable X with a pdf f(x) is

$$\mu = \int x\, f(x)\, dx$$

where the integrals are definite integrals taken for *x* ranging over the range of *X*.

Properties of Means

If a random variable *X* is adjusted by multiplying by the value *b* and adding the value *a*, then the mean is affected as follows:

$$\mu_{a+bX} = a + b\mu_X$$

The mean of the sum of two random variables *X* and *Y* is the sum of their means:

$$\mu_{X+Y} = \mu_X + \mu_Y$$

Variance

The variance of a random variable X measures the spread (with respect to the mean), or variability, of the distribution. It is defined as the mean square deviation, i.e.,

$$\begin{aligned}
\operatorname{Var}(X) &= \mathrm{E}\left[(X - \mu)^2\right]. \\
\operatorname{Var}(X) &= \mathrm{E}\left[X^2 - 2X\,\mathrm{E}[X] + (\mathrm{E}[X])^2\right] \\
&= \mathrm{E}\left[X^2\right] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + (\mathrm{E}[X])^2 \\
&= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2
\end{aligned}$$

A mnemonic for the above expression is "mean of square minus square of mean".

This definition encompasses random variables that are discrete, continuous, neither, or mixed.

Discrete random variable

The *variance* of a discrete random variable *X* with probability mass function $x_1 \mapsto p_1$, ..., $xn \mapsto pn$ is defined by

$$\operatorname{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2 = \sum_{i=1}^{n} (p_i \cdot x_i^2) - \mu^2$$

where ***μ*** *is the expected value, i.e.*

$$\mu = \sum_{i=1}^{n} p_i \cdot x_i$$

The standard deviation **σ** is the square root of the variance.

Continuous random variable

If the random variable *X* is continuous with probability density function *f*(*x*), then the variance is given by

$$\operatorname{Var}(X) = \sigma^2 = \int (x-\mu)^2 f(x)\,dx = \int x^2 f(x)\,dx - \mu^2$$

Properties of Variances

Variance is invariant with respect to changes in a location parameter. That is, if a constant is added to all values of the variable, the variance is unchanged.

$$\operatorname{Var}(X + a) = \operatorname{Var}(X).$$

If all values are scaled by a constant, the variance is scaled by the square of that constant.

$$\operatorname{Var}(aX) = a^2 \operatorname{Var}(X).$$

The variance of a sum of two random variables is given by:

$$\operatorname{Var}(aX + bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2ab \operatorname{Cov}(X, Y),$$
$$\operatorname{Var}(X - Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) - 2 \operatorname{Cov}(X, Y),$$

where **Cov** stands for covariance (we will get to its definition in just a little bit).

$$\begin{aligned}\operatorname{Var}(XY) &= E[X^2Y^2] - [E(XY)]^2 \\ &= \operatorname{Cov}(X^2, Y^2) + E(X^2)E(Y^2) - [E(XY)]^2 \\ &= \operatorname{Cov}(X^2, Y^2) + (\operatorname{Var}(X) + [E(X)]^2)(\operatorname{Var}(Y) + [E(Y)]^2) - [\operatorname{Cov}(X, Y) + E(X)E(Y)]^2\end{aligned}$$

Sample variance

In many practical situations, the true variance of a population is not known *a priori* and must be computed somehow. When dealing with extremely large populations, it is not possible to count every object in the population, so the computation must be performed on a sample of the population. Sample variance can also be applied to the estimation of the variance of a continuous distribution from a sample of that distribution.

We take a sample with replacement of *n* values $y_1, \ldots, y_n$ from the population, where $n < N$, and estimate the variance on the basis of this sample. Directly taking the variance of the sample gives:

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Here, $\overline{y}$ denotes the sample mean:

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Since the y_i are selected randomly, both $\overline{y}$ and σ_y^2 are random variables. Their expected values can be evaluated by summing over the ensemble of all possible samples $\{y_i\}$ from the population. For σ_y^2 this gives:

$$\begin{aligned}E[\sigma_y^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{j=1}^{n} y_j\right)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[y_i^2 - \frac{2}{n}y_i\sum_{j=1}^{n} y_j + \frac{1}{n^2}\sum_{j=1}^{n} y_j \sum_{k=1}^{n} y_k\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}E[y_i^2] - \frac{2}{n}\sum_{j\neq i} E[y_i y_j] + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k\neq j} E[y_j y_k] + \frac{1}{n^2}\sum_{j=1}^{n} E[y_j^2]\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}(\sigma^2 + \mu^2) - \frac{2}{n}(n-1)\mu^2 + \frac{1}{n^2}n(n-1)\mu^2 + \frac{1}{n}(\sigma^2 + \mu^2)\right] \\ &= \frac{n-1}{n}\sigma^2.\end{aligned}$$

Because y_j and y_k are independent, and $E(y_j) = E(y_k) = \mu$.

Hence σ_y^2 gives an estimate of the population variance that is biased by a factor of (*n*-1)/*n*. For this reason, σ_y^2 is referred to as the *biased sample variance*. Correcting for this bias yields the *unbiased sample variance*:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

Either estimator may be simply referred to as the *sample variance* when the version can be determined by context. The same proof is also applicable for samples taken from a continuous probability distribution.

Moments

If f(x) is a probability density function, then the nth moment of the probability distribution about a value c is

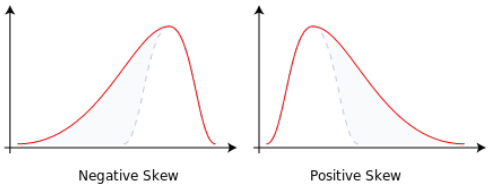
$$\mu_n' = \int_{-\infty}^{\infty}(x-c)^n f(x)\,dx.$$

The *n*th moment about zero is the expected value of X^n ($E(X^n)$)and is called a *raw moment* or *crude moment*. The moments about its mean μ are called *central* moments (denoted by μ_n); these describe the shape of the function, independently of translation. ***The mean of a distribution is therefore the first raw moment, and the variance the second central moment.***

The *n*th standardized moment of a probability distribution is μ_n/σ^n where μ_n is the *n*th central moment and σ is the standard deviation.

Skewness

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or even undefined.



(Optional) The skewness of a random variable *X*, denoted γ_1 is defined as

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}$$

where μ_3 is the third central moment μ, σ is the standard deviation.
The skewness is also sometimes denoted Skew[*X*].

Sample skewness

For a sample of *n* values the *sample skewness* is

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{3/2}},$$

where $\overline{x}$ is the sample mean, m_3 is the sample third central moment, and m_2 is the sample variance.
Aside: Given samples from a population, the equation for the sample skewness g_1 above is a biased estimator of the population skewness. $g_1 = (n-2)/\sqrt{(n(n-1))} * G_1$, where G_1 is the population skewness.

Kurtosis

Kurtosis (from Greek, meaning curved, arching) is any measure of the "heavy-tailedness" of the probability distribution of a real-valued random variable. In a similar way to the concept of skewness, *kurtosis* is a descriptor of the shape of a probability distribution and, just as for skewness, there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population. There are different interpretations of kurtosis, and of how particular measures should be interpreted: tail weight, lack of shoulders, etc.

(Optional) One common measure of kurtosis, originating with Karl Pearson, is based on a scaled version of the fourth moment of the data or population, and it has been argued that this really measures heavy tails. For this measure, higher kurtosis means more of the variance is the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations.

The fourth standardized moment is defined as

$$\beta_2 = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2} = \frac{\mu_4}{\sigma^4}$$

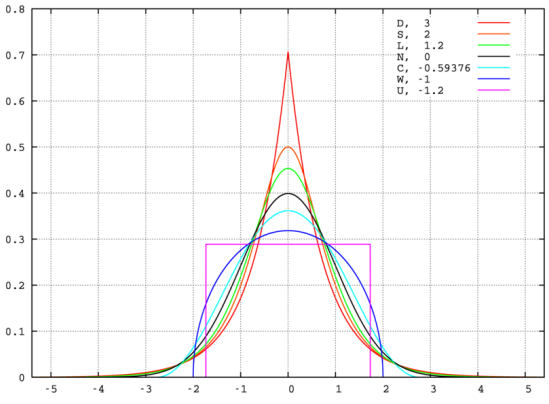
where μ_4 is the fourth moment about the mean and σ is the standard deviation.

It is common practice to use an adjusted version of Pearson's kurtosis, the **excess kurtosis**, to provide a comparison of the shape of a given distribution to that of the normal distribution. Distributions with negative or positive excess kurtosis are called platykurtic distributions or leptokurtic distributions respectively.

Kurtosis (γ_2) is more commonly defined as the fourth moment around the mean divided by the square of the variance of the probability distribution minus 3,

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

which is also known as **excess kurtosis**. The "minus 3" at the end of this formula is often explained as a correction to make the kurtosis of the normal distribution equal to zero.



Several well-known, unimodal and symmetric distributions from different parametric families are compared here. Each has a mean and skewness of zero. The parameters have been chosen to result in a variance equal to 1 in each case. The images on the right show curves for the following seven densities, on a linear scale and logarithmic scale:

- L: logistic distribution, green curve, excess kurtosis = 1.2
- N: normal distribution, black curve (inverted parabola in the log-scale plot), excess kurtosis = 0
- W: Wigner semicircle distribution, blue curve, excess kurtosis = −1
- U: uniform distribution, magenta curve (shown for clarity as a rectangle), excess kurtosis = −1.2.

§2. Uncorrelatedness, independence, and orthogonality

Covariance

The covariance between two jointly distributed real-valued random variables *x* & *y* with finite second moments is defined:

Cov(X,Y) = $\sigma(x,y) = \mathrm{E}\left[(x - \mathrm{E}[x])(y - \mathrm{E}[y])\right],$

By using the linearity property of expectations, this can be simplified to

$$\begin{aligned}\sigma(x,y) &= \mathrm{E}\left[(x - \mathrm{E}\left[x\right])\left(y - \mathrm{E}\left[y\right]\right)\right] \\ &= \mathrm{E}\left[xy - x\,\mathrm{E}\left[y\right] - \mathrm{E}\left[x\right]y + \mathrm{E}\left[x\right]\mathrm{E}\left[y\right]\right] \\ &= \mathrm{E}\left[xy\right] - \mathrm{E}\left[x\right]\mathrm{E}\left[y\right] - \mathrm{E}\left[x\right]\mathrm{E}\left[y\right] + \mathrm{E}\left[x\right]\mathrm{E}\left[y\right] \\ &= \mathrm{E}\left[xy\right] - \mathrm{E}\left[x\right]\mathrm{E}\left[y\right].\end{aligned}$$

Consequently, the variance of a random variable *X* can also be thought of as the covariance of a random variable with itself:

$$\mathrm{Var}(X) = \mathrm{Cov}(X,X).$$

For random vectors $\mathbf{X}$ and $\mathbf{Y}$ (of dimension m and n respectively) the m×n cross covariance matrix (also known as dispersion matrix or variance–covariance matrix, or simply called covariance matrix) is equal to

$$\begin{aligned}\sigma(\mathbf{x},\mathbf{y}) &= \mathrm{E}\left[(\mathbf{x} - \mathrm{E}[\mathbf{x}])(\mathbf{y} - \mathrm{E}[\mathbf{y}])^{\mathrm{T}}\right] \\ &= \mathrm{E}\left[\mathbf{x}\mathbf{y}^{\mathrm{T}}\right] - \mathrm{E}[\mathbf{x}]\,\mathrm{E}[\mathbf{y}]^{\mathrm{T}},\end{aligned}$$

where m^T is the transpose of the vector (or matrix) m.

Dependence (and correlation)

In statistics, dependence is any statistical relationship between two random variables or two sets of data. Formally, *dependence* refers to any situation in which random variables do not satisfy a mathematical condition of probabilistic independence (which we will get to in a little bit).

Correlation refers to a specialized class of statistical relationships or dependence, namely, linear relationships. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice; however, the existence of a predictive relationship is not sufficient to demonstrate that it is a causal one (i.e., correlation does not imply causation). Note that whereas in ***loose usage***, correlation can refer to any departure of two or more random variables from independence, in technical terms it refers to linear relationships. There are several correlation coefficients measuring the degree of correlation. The most common of these is the Pearson correlation coefficient, also called the Pearson product-moment correlation coefficient, or simply "the correlation coefficient".

Pearson's correlation

Pearson's correlation coefficient (Pearson's r) is a measure of the linear relationship (linear dependence) between two variables (which *may or may not* exist when one is a nonlinear function of the other). Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. It takes values between -1 and 1 (inclusive).

For a population

The population correlation coefficient $\rho_{X,Y}$ between two random variables *X* and *Y* with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as:

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

where *corr* a widely used alternative notation for the correlation coefficient.

The Pearson correlation is defined only if both of the standard deviations are finite and both of them are nonzero.

For a sample

Pearson's correlation coefficient when applied to a sample is commonly represented by the letter *r* and may be referred to as the *sample correlation coefficient* or the *sample Pearson correlation coefficient*. We can obtain a formula for *r* by substituting estimates of the covariances and variances based on a sample into the formula above. That formula for *r* is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Properties:

A key mathematical property of the Pearson correlation coefficient is that it is invariant (up to a sign) to separate changes in location and scale in the two variables. That is, we may transform *X* to *a* + *bX* and transform *Y* to *c* + *dY*, where *a*, *b*, *c*, and *d* are constants, without changing the correlation coefficient (this fact holds for both the population and sample Pearson correlation coefficients). Note that more general linear transformations do change the correlation: see a later section for an application of this.

Uncorrelatedness

Two real-valued random variables, *X*,*Y*, are said to be uncorrelated ***if and only if*** their covariance is zero, i.e.,

Cov (X,Y) = E(*XY*) - E(*X*)E(*Y*) = 0. In other words, E(XY) = E(X) E(Y).

Uncorrelated random variables, therefore, have a Pearson correlation coefficient of zero, except in the trivial case when either variable has zero variance (is a constant). In this case the correlation is undefined.

A set of two or more random variables is called uncorrelated if each pair of them is uncorrelated.

Conceptually: If two variables are uncorrelated, there is **no linear relationship** between them.

Independence

Two events are independent (alternatively called statistically independent or stochastically independent) if the occurrence of one does not affect the probability of the other. Similarly, two random variables are independent if the realization of one does not affect the probability distribution of the other.

Formally, two events *A* and *B* (or random variables X and Y) are independent ***if and only if*** their joint probability equals the product of their probabilities:

$P(X \cap Y) = P(X)P(Y)$ ($P(X \cap Y)$ can be denoted in many ways: P(XY), P(X,Y), $P(X \cap Y)$, $P(X \wedge Y)$, P(X & Y).)

The concept of independence extends to dealing with collections of more than two events or random variables. For instance, P(ABC) = P(A)P(B)P(C).

Conceptually: Iff two variables are independent, there is **no relationship of any kind** between them.

Property of independent RVs: If two RVs are independent, then E(XY) = E(X)E(Y). (Note, this is just ***if***, not ***if and only if***.)

Uncorrelatedness & Independence

- If *X* and *Y* are independent, then they are uncorrelated (lack of dependence => lack of linear dependence)
  Proof: If X & Y are independent, then E(XY) = E(X)E(Y) =>E(XY)-E(X)E(Y)=0 => Cov(XY)=0 => X & Y are uncorrelated.
- ***However***, not all uncorrelated variables are independent. (Lack of linear dependence does not mean lack of all forms of dependence).
  Example: If *X* is a continuous random variable uniformly distributed on [−1, 1] and $Y = X^2$, (cup shaped) then *X* and *Y* are uncorrelated but they are not independent (because *X* determines *Y* and a particular value of *Y* can be produced by only one or two values of *X*).

When uncorrelatedness ***does*** imply independence

It is sometimes mistakenly thought that one context in which uncorrelatedness implies independence is when the random variables involved are normally distributed. Here is the complete picture:

- If two variables are ***jointly*** normally distributed, then uncorrelatedness => independence (* see note below)
- If two variables are ***merely marginally*** normally distributed but not jointly normally distributed, then uncorrelatedness ≠> (does not imply) independence

*Note: Suppose two random variables *X* and *Y* are *jointly* normally distributed (i.e., the random vector (*X*, *Y*) has a multivariate (in this case, bivariate) normal distribution). Based on "jointly normal" from page 3, this means that the joint probability distribution of *X* and *Y* is such that each linear combination of *X* and *Y* is normally distributed. In other words, for any two constant (i.e., non-random) scalars *a* and *b*, the random variable $aX + bY$ is normally distributed. *In that case* if *X* and *Y* are uncorrelated, i.e., their covariance cov(*X*, *Y*) is zero, *then* they are independent.

Some properties of independent random variables

If two variables X and Y are independent, the variance of their sum is given by

$Var(aX \pm bY) = a^2Var(X) + b^2Var(Y)$

Variances are added for both the sum *and* difference of two independent random variables because the variation in each variable contributes to the variation in each case *(If the variables are not independent, then variability in one variable is related to variability in the other).*

If two variables X and Y are independent, the variance of their product is given by

$$\begin{aligned}\operatorname{Var}(XY) &= [E(X)]^2 \operatorname{Var}(Y) + [E(Y)]^2 \operatorname{Var}(X) + \operatorname{Var}(X) \operatorname{Var}(Y) \\ &= E(X^2)E(Y^2) - [E(X)]^2[E(Y)]^2.\end{aligned}$$

Orthogonality

Two random variables are orthogonal if and only if their joint expectation is zero, i.e., E(XY) = 0.

Uncorrelatedness & Orthogonality

In general, uncorrelatedness (*E(XY)=E(X)E(Y)*) is not the same as orthogonality (*E(XY)=0)*). The exception is the special case where either *X* or *Y* has an expected value of 0. In this case, the covariance is the expectation of the product (Cov(XY) = E(XY) – E(X)E(Y) = E(XY)) , and *X* and *Y* are uncorrelated if and only if E(*XY*) = 0.

Review

Think about the conceptual (non-formula-based) meanings of uncorrelated, independent, and orthogonal.

§3. (Optional) Appendix:

Normal distribution

The normal distribution with parameters μ and σ is a continuous distribution whose probability density function is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

It has mean μ and variance equal to:

$$\operatorname{Var}(X) = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \sigma^2.$$

Exponential distribution

The exponential distribution with parameter λ is a continuous distribution whose support is the semi-infinite interval [0,∞). Its probability density function is given by:

$$f(x) = \lambda e^{-\lambda x},$$

and it has expected value $\mu = \lambda^{-1}$. The variance is equal to:

$$\operatorname{Var}(X) = \int_0^\infty (x - \lambda^{-1})^2 \, \lambda e^{-\lambda x} dx = \lambda^{-2}.$$

So for an exponentially distributed random variable $\sigma^2 = \mu^2$.

Poisson distribution

The Poisson distribution with parameter λ is a discrete distribution for *k* = 0, 1, 2, ... Its probability mass function is given by:

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

and it has expected value μ = λ. The variance is equal to:

$$\operatorname{Var}(X) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} (k - \lambda)^2 = \lambda,$$

So for a Poisson-distributed random variable $\sigma^2 = \mu$.

Binomial distribution

The binomial distribution with parameters *n* and *p* is a discrete distribution for *k* = 0, 1, 2, ..., *n*. Its probability mass function is given by:

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

and it has expected value μ = *np*. The variance is equal to:

$$\operatorname{Var}(X) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} (k - np)^2 = np(1-p),$$

Coin toss example: The binomial distribution with p=0.5 describes the probability of getting k heads in n tosses.