

Probability & Statistics Primer
Gregory J. Hakim
University of Washington
3 January 2005 v1.1

This primer provides an overview of basic concepts and definitions in probability and statistics. We shall denote a **sample space** by S , and define a **random variable** by the result of a rule (function) that associates a real number with each outcome in S . For simplicity, consider two discrete random variables that are associated with events in the sample space, A and B . The probability of A is denoted by $P(A)$, which represents the frequency of occurrence of the event (a real number between 0 and 1; “probability axiom #1”). For example, we expect that the probability of rolling a one on an unbiased six-side die is $1/6$; we can check this by rolling the die many times, counting the number of ones, and dividing by the number of rolls. Equivalently, $P(A)$ represents the likelihood of the event occurring; this is called the “Bayesian perspective” (cf. the “frequentist perspective” defined above), which will prove useful for interpreting conditional probabilities. “Probability axiom #2” says that $P(S) = 1$. The complement of A , A^c , is defined as all events in S that are not in A ; $P(A) = 1 - P(A^c)$.

For continuous variables, we shall use probability density functions (PDFs), denoted by $p(x)$. A PDF is related to a probability through an integral relationship, which is unity over the whole sample space. For example, in one dimension,

$$P(-\infty < x < \infty) = \int_{-\infty}^{\infty} p(x) dx = 1. \quad (1)$$

The probability that x is between a and $a + da$ is

$$P(a \leq x \leq a + da) = \int_a^{a+da} p(x) dx \approx P(a) da \quad (2)$$

Note that $P(a \leq x \leq a+da)$ may be a very small number even at the “peak” of a distribution if da is small; however, it will be larger than all others for similarly chosen da and different x . Furthermore, it doesn’t really make sense to say that, “the probability of $x = a$ is $P(a)$ ” because as (2) shows, this limiting case gives

$$P(x = a) = \int_a^a p(x) dx = 0. \quad (3)$$

More definitions:

Sample space (S): the set of all possible outcomes. Example: the sample space for a six-sided die consists of the integer values 1,2,3,4,5, and 6. $P(S) = 1$.

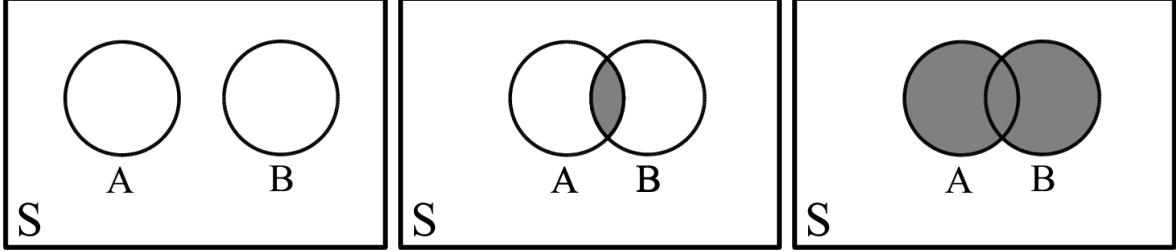


Figure 1: Venn diagrams for (left) **mutually exclusive** events (center) **intersection** or probability of A and B [$P(A \cap B)$] and (right) **union** or probability of A or B [$P(A \cup B)$].

Mutually exclusive events: A and B share no common outcomes (Fig. 1, left panel).

Intersection $P(A \cap B)$, “ A and B ”: outcome is both A and B (Fig. 1, center panel). If A and B are *mutually exclusive*, then $P(A \cap B) = 0$. The intersection is commutative, $P(A \cap B) = P(B \cap A)$, and associative, $P(A \cap B \cap C) = P([A \cap B] \cap C) = P(A \cap [B \cap C])$.

Union $P(A \cup B)$, “ A or B ”: outcome is either A or B . From the Venn diagram in Fig. 1 (right panel) we can see that the probability of A or B is the sum of the individual probabilities, minus the *intersection*, which gets added twice:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \quad (4)$$

If A and B are *mutually exclusive*, then $P(A \cup B) = P(A) + P(B)$; i.e. sum the probabilities. The union is commutative, $P(A \cup B) = P(B \cup A)$, and associative, $P(A \cup B \cup C) = P([A \cup B] \cup C) = P(A \cup [B \cup C])$.

NOTE: Union *and* intersection are not associative: $P([A \cup B] \cap C) \neq P(A \cup [B \cap C])$, but they do satisfy the distributive property: $P(A \cup [B \cap C]) = P([A \cup B] \cap [A \cup C])$ and $P(A \cap [B \cup C]) = P([A \cap B] \cup [A \cap C])$.

Conditional probability $P(A | B)$, “probability of A given that B has occurred”: given that B has occurred, the sample space shrinks from S to B . We expect that $P(A | B) \propto P(A \cap B)$, since the sample space is now just B . Since $P(B | B)$ must be unity and $P(B \cap B) = P(B)$, the constant of proportionality must be $1/P(B)$:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}. \quad (5)$$

Similarly

$$P(B | A) = \frac{P(A \cap B)}{P(A)} \quad (6)$$

Using (6) to replace $P(A \cap B)$ in (5) gives **Bayes' Theorem**,

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}. \quad (7)$$

Note that if two events are *mutually exclusive*, then $P(A \cap B) = P(A | B) = 0$.

Independence. Events A and B are independent if the outcome of one does not depend on the outcome of the other; i.e.

$$P(A | B) = P(A). \quad (8)$$

Appealing to (5), (8) implies that

$$P(A \cap B) = P(A)P(B), \quad (9)$$

i.e., the outcome “A and B” is determined by a product of probabilities. This is a great simplification that is often made as a leading-order approximation. Of course it is perfect in some cases, as for example in rolling two dice, where the outcome on one die has no affect on the other; the probability of rolling a six on each is $(1/6) \times (1/6)$.

A general “multiplication rule” for independent events following directly from (8):

$$P(A_1 \cap A_2 \cdots \cap A_n) = P(A_1)P(A_2) \cdots P(A_n) \quad (10)$$

The final probability axiom (#3) defines the “Law of Total Probability.” Given n mutually exclusive and exhaustive $[\sum_{i=1}^n P(A_i) = 1]$ events $A_1, \dots, A_n$, then for some other event B

$$P(B) = P(B | A_1)P(A_1) + \cdots + P(B | A_n)P(A_n) \quad (11)$$

or, equivalently

$$P(B) = \sum_{i=1}^n P(B | A_i)P(A_i). \quad (12)$$

Since the A_i are mutually exclusive, we just need to add up the intersections of each with B , and since they are exhaustive there is no probability unaccounted for by the A_i .

Properties of Probability Density Functions (PDFs)

Cumulative distribution function (CDF):

$$F(x) = \int_{-\infty}^x p(y) dy. \quad (13)$$

The **expected value** of a PDF is equivalent to the “center of mass” in classical physics. It is determined by the expectation operator, $E[\cdot]$, defined by

$$E[x] = \int_{-\infty}^{\infty} x p(x) dx. \quad (14)$$

The expectation operator is linear, e.g. $E[ax + by] = aE[x] + bE[y]$.

The k^{th} **moment** about the expected value is defined by

$$\mu_k = \int_{-\infty}^{\infty} (x - \mu)^k p(x) dx, \quad (15)$$

where $\mu = E[x]$. Note that the first moment is zero, and that the second moment is called the **variance**, and is usually denoted by σ^2 ; the **standard deviation** is defined as the square-root of the variance, σ .

The **covariance** between two random variables is defined by

$$cov(x, y) = E[(x - \mu)(y - \nu)] = E[xy] - \mu\nu, \quad (16)$$

where μ and ν are $E[x]$ and $E[y]$, respectively. The covariance determines the linear relationship between two quantities, and the strength of this relationship is measured by the **correlation coefficient**,

$$corr(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y}, \quad (17)$$

which ranges between -1 and +1. If x and y are independent, then $cov(x, y) = corr(x, y) = 0$.

Note that these definitions for PDFs extend naturally to multidimensional settings, and also to discrete random variables; in the discrete case integrals are replaced by sums.

Gaussian (Normal) Probability Density Function

Although there are many analytically defined PDFs, the Gaussian is particularly important in many areas of science. The assumption of Gaussian statistics is often invoked because

it greatly simplifies the analysis, since all moments (15) of the PDF may be determined from the mean and covariance. A theoretical underpinning for this widespread use of the Gaussian PDF is given by the Central Limit Theorem, which states that the sum of many independent, identically distributed random variables is normally distributed (i.e., the mean value approaches $n\mu$ and the variance approaches $n\sigma^2$ for sufficiently large n); note that “identically distributed” means “have the same distribution.”

In one dimension, the Gaussian (Normal) distribution is defined by

$$p(x) = (2\pi)^{-1/2}(\sigma)^{-1}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad (18)$$

where μ is the expected value and σ^2 is the variance. A shorthand notation to denote a variable having a Gaussian distribution is $x \sim N(\mu, \sigma^2)$.

Properties of the Gaussian include:

- If x is normally distributed, then so is a linear function of x . If $x \sim N(\mu, \sigma^2)$ then $y = ax + b \sim N(a\mu + b, a^2\sigma^2)$.
- The normal sum (difference) distribution: adding (subtracting) *different* Gaussian distributions yields a Gaussian.

The n -dimensional generalization of the Gaussian distribution is

$$p(\mathbf{x}) = (2\pi)^{-n/2}(|\mathbf{B}|)^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{B}^{-1}(\mathbf{x}-\mu)}, \quad (19)$$

where $\mathbf{x}$ and μ are vectors and $\mathbf{B}$ is the covariance matrix. The vector $\mathbf{x}$ has one entry for each of the n degrees of freedom, as does μ , which represents the vectors of expected values. The covariance matrix is defined by $\mathbf{B} = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T]$, and $|\cdot|$ is the determinant. Diagonal elements of $\mathbf{B}$ represent the variance in each degree of freedom, and the off-diagonal elements represent $cov(x_i, x_j)$; in the case where all degrees of freedom are independent, $\mathbf{B}$ is diagonal. The superscripted “T” denotes the transpose. In the limiting case of $n = 1$, (19) recovers (18). Each property of the one-dimensional Gaussian has a multi-dimensional extension.