# Probability & Statistics Primer

**Gregory J. Hakim**
University of Washington

2 January 2009 v2.0

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 $A$, denoted by $P(A)$, can be viewed as the frequency of occurrence of the event (the "frequentist perspective"). Equivalently, $P(A)$ represents the likelihood of the event occurring; this is called the "Bayesian perspective," which will prove useful for interpreting conditional probability. There are three basic probability axioms. Probability axiom #1 states that $P(A)$ is a real number between 0 and 1. The complement of $A$, $A^c$, is defined as all events in $S$ that are not in $A$; $P(A) = 1 - P(A^c)$. A familiar example derives from rolling a one on an unbiased six-side die, which occurs with probability 1/6. We can check this by rolling the die many times, counting the number of ones, and dividing by the number of rolls. "Probability axiom #2" says that $P(S) = 1$. "Probability axiom #3" will be defined below, but basically it says that if we have $n$ outcomes that don't overlap, the probability of all of them occurring is just the sum of the probabilities for each outcome.

For discrete random variables, we shall use probability mass functions (PMFs). PMFs return a probability for the discrete random variable, $X$. For the unbiased six-side die $P(X = 6) = 1/6$, etc. 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,

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

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

$$P(a \le x \le a + da) \,=\, \int_{a}^{a+da} p(x)\,dx \approx P(a)\,da \qquad (2)$$

Note that $P(a \le x \le a+da)$ 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

$$P(x = a) = \int_a^a p(x)\,dx = 0. \tag{3}$$

### Union and Intersection

**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. $P(S) = 1$.

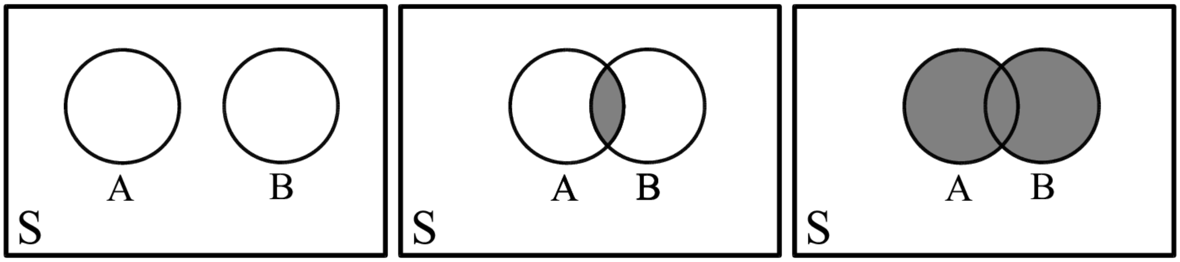


Figure 1: Venn diagrams for (left) **mutually exclusive** events (center) **intersection** or probability of $A$ and $B$ $[P(A \cap B)]$ and (right) **union** or probability of $A$ or $B$ $[P(A \cup B)]$.

**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:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{4}$$

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 $P(A \cap [B \cup C]) = P([A \cap B] \cup [A \cap C])$.

We may now state "Probability axiom #3:" given $n$ mutually exclusive events $A_1, A_2, \ldots, A_n$, then $P(A_1 \cup A_2 \cup \ldots A_n) \;=\; \sum_{i=1}^{n} P(A_i)$.

### Conditional probability

Conditional probability is denoted by $P(A \mid B)$ and read as "the probability of A given that B has occurred." Given that $B$ has occurred, the sample space shrinks from $S$ to $B$. We expect that $P(A \mid B) \propto P(A \cap B)$, since the sample space is now just $B$. Since $P(B \mid B)$ must be unity and $P(B \cap B) \;=\; P(B)$, the constant of proportionality must be $1/P(B)$:

$$P(A \mid B) \;=\; \frac{P(A \cap B)}{P(B)}. \tag{5}$$

Similarly

$$P(B \mid A) \;=\; \frac{P(A \cap B)}{P(A)} \tag{6}$$

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

$$P(A \mid B) \;=\; \frac{P(B \mid A)\; P(A)}{P(B)}. \tag{7}$$

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

Extending to three events (see Fig. 2 below), we expect that the probability of $A$ given that $B$ and $C$ have occurred, $P(A \mid B \;\cap\; C)$ or simply $P(A \mid B, C)$, is proportional to the probability of the intersection of $A$, $B$, and $C$, $P(A \cap B \cap C)$ or simply $P(A, B, C)$:

$$P(A \mid BC) \;=\; \frac{P(A, B, C)}{P(B, C)}, \tag{8}$$

where the constant of proportionality is deduced by similar logic to that for (5). A similar expression applies for $P(B \mid A, C)$:

$$P(B \mid A, C) \;=\; \frac{P(A, B, C)}{P(A, C)}. \tag{9}$$

Substituting $P(A, B, C)$ from (9) into (8) gives

$$P(A \mid B, C) \;=\; \frac{P(B \mid A, C)P(A, C)}{P(B, C)}. \tag{10}$$

Using the fact [i.e., (5)] that $P(A, C) = P(A \mid C)P(C)$ and $P(B, C) = P(B \mid C)P(C)$ yields

$$P(A \mid B, C) \;=\; \frac{P(B \mid A, C)P(A \mid C)}{P(B \mid C)}. \tag{11}$$

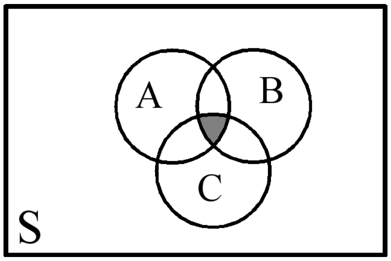


Figure 2: Probability of $A$ and $B$ and $C$ [ $P(A \cap B \cap C) = P(A, B, C)$].

More generally, the **chain rule** of conditional probability allows one to "factor" joint probabilities. Previously we discussed the "product rule" for two variables,

$$P(A, B) \;=\; P(A \mid B)P(B) \;=\; P(B \mid A)P(A), \tag{12}$$

and a similar result for three variables,

$$P(A, B, C) \;=\; P(A \mid B, C)P(B, C) \;=\; P(A \mid B, C)P(B \mid C)P(C). \tag{13}$$

Extending to $n$ variables,

$$P(A_1, A_2, \ldots, A_n) \;=\; P(A_1 \mid A_2, \ldots, A_n)P(A_2 \mid A_3, \ldots, A_n) \cdots P(A_{n-1} \mid A_n)P(A_n). \tag{14}$$

A **Markov Chain**, denoted $A \rightarrow B \rightarrow C$, is defined by a joint probability that factors as $P(A, B, C) \;=\; P(C \mid B)P(B \mid A)P(A)$, which implies that $C$ is independent of $A$. If $A, B, C$ occur sequentially in time (as implied by the arrows in the definition), this means that in order to determine the probability at the new time, $C$, we only need to know the probability of the current time, $B$.

**Marginal probability**

Marginal probability represents the unconditional probability of one or more independent random variables without considering information from other random variables; e.g. consider $P(A)$ instead of $P(A, B)$. For discrete random variables, we obtain $P(A)$ from $P(A, B)$ by summation over the unwanted variable in the joint probability mass function: $P(A) \;=\; \sum_B P(A, B) \;=\; \sum_B P(A \mid B)P(B)$. Similarly, for continuous random variables, marginal probability density functions are defined by integrating over the unwanted variables in the joint probability density function; e.g., $p(y) \;=\; \int_y p(x, y)dy \;=\; \int_y p(x \mid y)p(y)dy$.

**Likelihood**

With conditional probability, we are given specific information about an event that affects the probability of another event; e.g., knowing that $B = b$ allows us to "update" the probability of $A$ by $P(A \mid B = b)$ through (7). Conversely, if we view the conditional probability as a *function* of the second argument, $B$, we get a likelihood function,

$$L(b \mid A) \;=\; \alpha P(A \mid B = b), \tag{15}$$

where $\alpha$ is a constant parameter. For example, consider a sequence of observations of a two-sided coin with outcomes "H" and "T." If we flip the coin once and observe "H" we can ask, "what is the likelihood that the true probability, $P(H)$ is 0.5?" Fig. 3 (left panel) shows that $L = 0.5$, and in fact the most likely value is P(H) = 1; note that L(0) = 0 since "H" has been observed and therefore must have a nonzero probability. Suppose flipping the coin again also gives "H." Maximum likelihood is still located at $P(H) = 1$, and the likelihood of smaller $P(H)$ is diminished. Remember, that we can turn this around and ask, if we *know* that $P(H) = 0.5$ (i.e., the coin is fair), what is the probability of observing "HH?" This is conditional probability, which in this case is 0.25 (the same as $L(0.5 \mid ``HH'')$. Observing "HHH" gives a cubic function for $L$, and so one. If, on the other hand, one of the observations is also "T," then both $L(0)$ and $L(1)$ must be zero (Fig. 3; right panel).

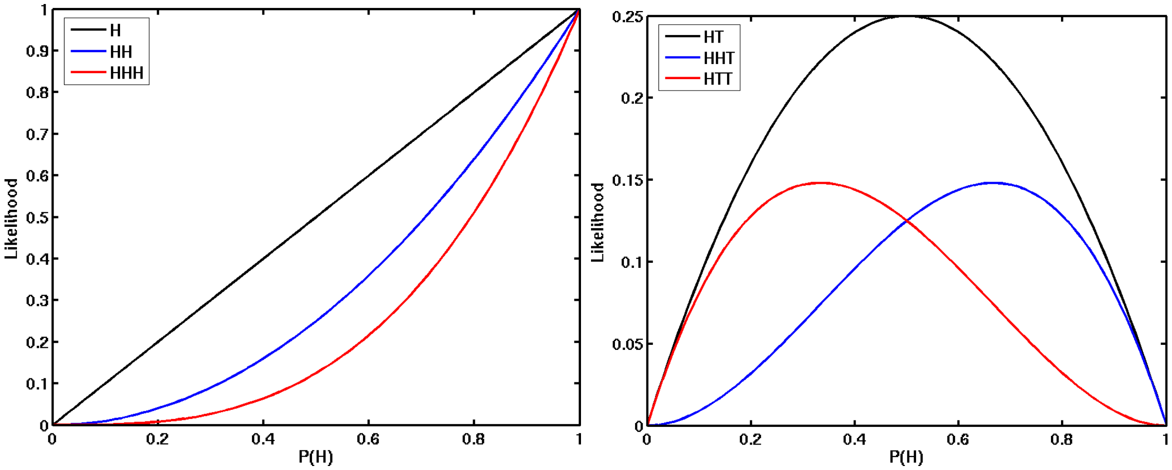


Figure 3: Likelihood as a function of the probability of "H" on a two-sided coin. On the left, the first three trials give only "H," whereas on the right "T" is observed as well.

Note that, unlike conditional probability, likelihood is **not a probability**, since it may take values larger than unity, and does not integrate to one.

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

$$P(A \mid B) \;=\; P(A). \tag{16}$$

Appealing to (5), (16) implies that

$$P(A \cap B) \;=\; P(A)P(B), \tag{17}$$

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) × (1/6). Note that independence can **not** be gleaned from examining a Venn Diagram.

A general "multiplication rule" for independent events following directly from (16):

$$P(A_1 \cap A_2 \cdots \cap A_n) = P(A_1)P(A_2)\cdots P(A_n) \tag{18}$$

Given $n$ mutually exclusive and exhaustive $[\sum_{i=1}^{n} P(A_i) = 1]$ events $A_1, \ldots, A_n$, the "Law of Total Probability." says that for some other event $B$

$$P(B) = P(B \mid A_1)P(A_1) + \cdots + P(B \mid A_n)P(A_n) \tag{19}$$

or, equivalently

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i). \tag{20}$$

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 $A_i$. Note that we may use this result in the denominators of (5) and (7).

### Properties of Probability Density Functions (PDFs)

**Cumulative distribution** function (CDF):

$$F(x) \;=\; \int_{-\infty}^{x} p(y)\, dy. \tag{21}$$

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

$$E[x] = \int_{-\infty}^{\infty} x\, p(x)\, dx. \tag{22}$$

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

$$\mu_k = \int_{-\infty}^{\infty} (x-\mu)^k \, p(x) \, dx, \tag{23}$$

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 $\sigma^2$; the **standard deviation** is defined as the square-root of the variance, $\sigma$.

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

$$cov(x,y) \; = \; E[(x-\mu)(y-\nu)] = E[xy] - \mu\nu, \tag{24}$$

where $\mu$ and $\nu$ 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**,

$$corr(x,y) \; = \; \frac{cov(x,y)}{\sigma_x \sigma_y}, \tag{25}$$

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 replace 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 (23) 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

$$p(x) \; = \; (2\pi)^{-1/2} (\sigma)^{-1} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \tag{26}$$

where $\mu$ is the expected value and $\sigma^2$ 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.

### Multivariate PDFs and Conditional Probability

Set union discussed on page 2 for discrete variables generalizes to joint probability density functions for continuous random variables, $p(\mathbf{x})$. For example, the $n$-dimensional generalization of the Gaussian distribution is

$$p(\mathbf{x}) \;=\; (2\pi)^{-n/2}(\mid \mathbf{B} \mid)^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{B}^{-1}(\mathbf{x}-\mu)}, \tag{27}$$

where $\mathbf{x}$ are $\mu$ 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 $\mu$, which represents the vector of expected values. The covariance matrix, real and symmetric, is defined by $\mathbf{B} \;=\; E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T]$, and $\mid \cdot \mid$ 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 $n = 1$, (27) recovers (26). Each property of the one-dimensional Gaussian has a multi-dimensional extension. One useful example is the linear transformation of the $n \times 1$ Gaussian random vector, $\mathbf{x} \sim N(\mu, \mathbf{B})$, by the $m \times n$ matrix operator, $\mathbf{A}$,

$$\mathbf{Y} \;=\; \mathbf{A}\mathbf{x}, \tag{28}$$

where $\mathbf{Y} \sim N(\mathbf{A}\mu, \mathbf{A}\mathbf{B}\mathbf{A}^{\mathrm{T}})$.

The generalization of Bayes Theorem to multivariate continuous random variables is given by

$$p(\mathbf{x}|\mathbf{y}) \;=\; \frac{f(\mathbf{x},\mathbf{y})}{f(\mathbf{y})} \;=\; \frac{p(\mathbf{y}|\mathbf{x})\,p(\mathbf{x})}{p(\mathbf{y})}. \tag{29}$$

Equivalently, using the definition for joint probability, (29) may be expressed as

$$p(\mathbf{x}|\mathbf{y}) \;=\; \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{y})}. \tag{30}$$

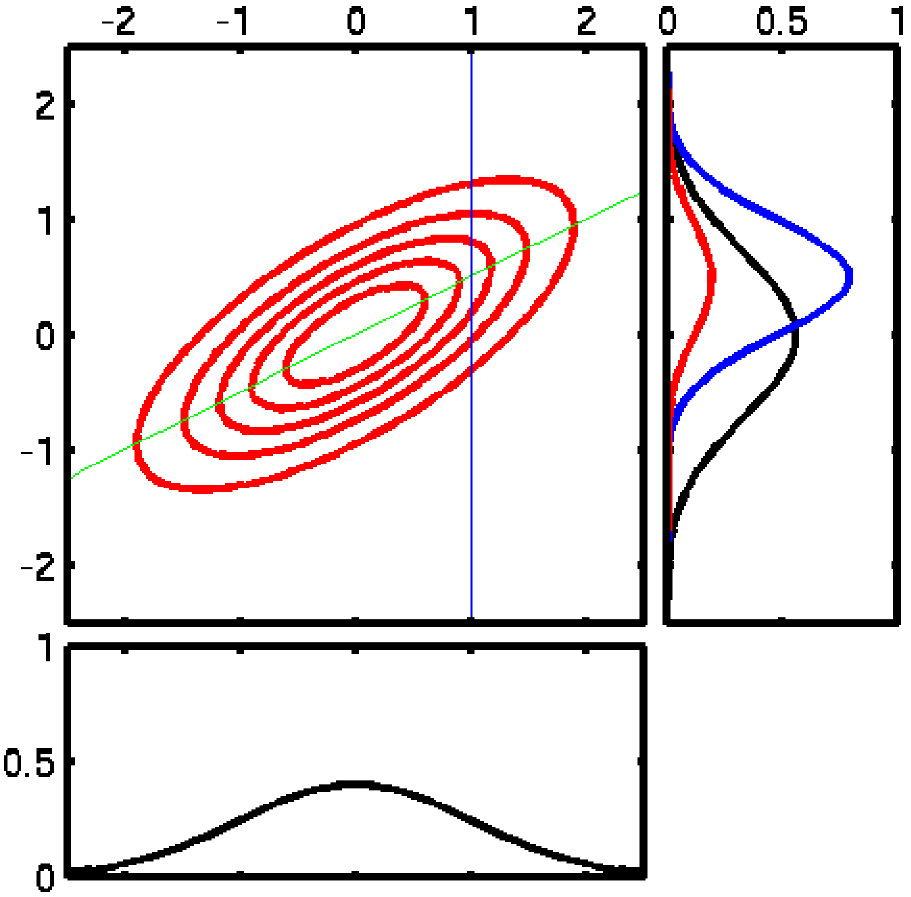
To help illustrate conditional probability, consider a two-dimensional example for Gaussian variables $x$ and $y$ having zero mean, correlation $\rho$, and standard deviation $\sigma_x$ and $\sigma_y$, respectively. The conditional probability of $y$ given $x$ can be shown from (30) and (27) to be

$$p(y \mid x) \;=\; \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}$$

where $\mu = \rho x \sigma_y / \sigma_x$ and $\sigma = \sigma_y (1 - \rho^2)^{1/2}$ are the mean and standard deviation, respectively, of the conditional probability density. Note that the from the definition of correlation, the mean can be re-written as $\mu = cov(x, y) x / {\sigma_x}^2$, which is the linear regression of $y$ on $x$. Note also that the standard deviation in the conditional probability density is smaller than in the original, unconditional, density by the factor $(1 - \rho^2)^{1/2}$, and therefore higher correlation results in lower variance in the conditional PDF.

These results are illustrated in the figure below. Red lines in the main panel show contours of constant probability density, which reflect greater variance in the $x$ variable compared to the $y$ variable. Black lines in the subplots below and to the right show the marginal probabilities; that is, the integral over the other dimension. Suppose that new information becomes available showing that $x = 1$, as indicated by the thin blue line in the main panel. Applying Bayes theorem gives the blue curve in the right panel. The green line shows the linear regression of $y$ on $x$, which intersects the blue line right at the peak of the conditional PDF shown in the right panel. Furthermore, note from the results derived above that the variance of the conditional PDF is independent of $x$, so that for other values of $x$, the conditional PDF shown in the right panel simply shifts along the $y$ axis, following the green line, without change of shape.

In contrast to the conditional PDF, the thin red line shows the value of the joint PDF at $x = 1$. First, notice that the area under the blue and red lines is different: the blue curve is a true probability density function, with unity integral, whereas the red line is not. The thin red line denotes the likelihood function, which is proportional to, but distinct from, conditional probability. Recall from the earlier discussion that, in the case of conditional probability, we are given information, in this case $x = 1$, that affects our knowledge of $y$ as expressed by $p(y \mid x = 1)$. For likelihood, the first argument is held fixed and we have a function of the second argument, $L(y = a \mid x) = p(x \mid y = a)$; i.e., if $x = 1$, what is the likelihood that $y = 0$? Notice for the example above that maximum likelihood occurs at the point of maximum conditional probability; they differ by a constant of proportionality.


-2
-1
0
1
2
0
0.5
1
2
1
0
-1
-2
1
0.5
0