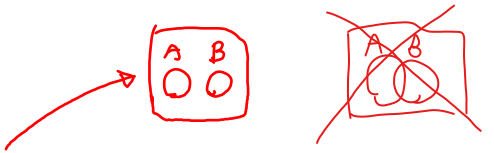


If the events are mutually exclusive (so they cannot happen at the same time), we get

$$\Pr(A \vee B) = \Pr(A) + \Pr(B) \tag{2.4}$$

For example, suppose X is chosen uniformly at random from the set $\mathcal{X} = \{1, 2, 3, 4\}$. Let A be the event that $X \in \{1, 2\}$ and B be the event that $X \in \{3\}$. Then we have $\Pr(A \vee B) = \frac{2}{4} + \frac{1}{4}$.

2.1.3.4 Conditional probability of one event given another

We define the **conditional probability** of event B happening given that A has occurred as follows:

$$\Pr(B|A) \triangleq \frac{\Pr(A, B)}{\Pr(A)} \tag{2.5}$$

This is not defined if $\Pr(A) = 0$, since we cannot condition on an impossible event.

2.1.3.5 Independence of events

We say that event A is **independent** of event B if

$$\Pr(A, B) = \Pr(A) \Pr(B) \tag{2.6}$$

2.1.3.6 Conditional independence of events

We say that events A and B are **conditionally independent** given event C if

$$\Pr(A, B|C) = \Pr(A|C) \Pr(B|C) \tag{2.7}$$

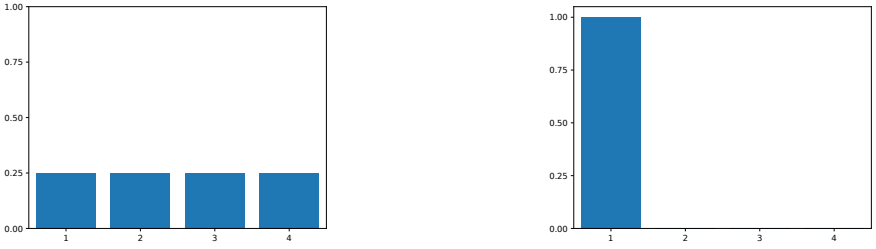
This is written as $A \perp B|C$. Events are often dependent on each other, but may be rendered independent if we condition on the relevant intermediate variables, as we discuss in more detail later in this chapter.

2.2 Random variables

Suppose X represents some unknown quantity of interest, such as which way a dice will land when we roll it, or the temperature outside your house at the current time. If the value of X is unknown and/or could change, we call it a **random variable** or **rv**. The set of possible values, denoted $\mathcal{X}$, is known as the **sample space** or **state space**. An **event** is a set of outcomes from a given sample space. For example, if X represents the face of a dice that is rolled, so $\mathcal{X} = \{1, 2, \ldots, 6\}$, the event of "seeing a 1" is denoted $X = 1$, the event of "seeing an odd number" is denoted $X \in \{1, 3, 5\}$, the event of "seeing a number between 1 and 3" is denoted $1 \leq X \leq 3$, etc.

2.2.1 Discrete random variables

If the sample space $\mathcal{X}$ is finite or countably infinite, then X is called a **discrete random variable**. In this case, we denote the probability of the event that X has value x by $\Pr(X = x)$. We define the



(a) (b)

Figure 2.1: Some discrete distributions on the state space $\mathcal{X} = \{1, 2, 3, 4\}$. (a) A uniform distribution with $p(x = k) = 1/4$. (b) A degenerate distribution (delta function) that puts all its mass on $x = 1$. Generated by discrete_prob_dist_plot.ipynb.

probability mass function or **pmf** as a function which computes the probability of events which correspond to setting the rv to each possible value:

$$p(x) \triangleq \Pr(X = x) \tag{2.8}$$

The pmf satisfies the properties $0 \leq p(x) \leq 1$ and $\sum_{x \in \mathcal{X}} p(x) = 1$.

If X has a finite number of values, say K, the pmf can be represented as a list of K numbers, which we can plot as a histogram. For example, Figure 2.1 shows two pmf's defined on $\mathcal{X} = \{1, 2, 3, 4\}$. On the left we have a uniform distribution, $p(x) = 1/4$, and on the right, we have a degenerate distribution, $p(x) = \mathbb{I}(x = 1)$, where $\mathbb{I}()$ is the binary indicator function. Thus the distribution in Figure 2.1(b) represents the fact that X is always equal to the value 1. (Thus we see that random variables can also be constant.)

2.2.2 Continuous random variables

If $X \in \mathbb{R}$ is a real-valued quantity, it is called a **continuous random variable**. In this case, we can no longer create a finite (or countable) set of distinct possible values it can take on. However, there are a countable number of *intervals* which we can partition the real line into. If we associate events with X being in each one of these intervals, we can use the methods discussed above for discrete random variables. Informally speaking, we can represent the probability of X taking on a specific real value by allowing the size of the intervals to shrink to zero, as we show below.

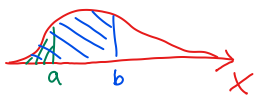
2.2.2.1 Cumulative distribution function (cdf)

Define the events $A = (X \leq a)$, $B = (X \leq b)$ and $C = (a < X \leq b)$, where $a < b$. We have that $B = A \vee C$, and since A and C are mutually exclusive, the sum rules gives

$$\Pr(B) = \Pr(A) + \Pr(C) \tag{2.9}$$

and hence the probability of being in interval C is given by

$$\Pr(C) = \Pr(B) - \Pr(A) \tag{2.10}$$



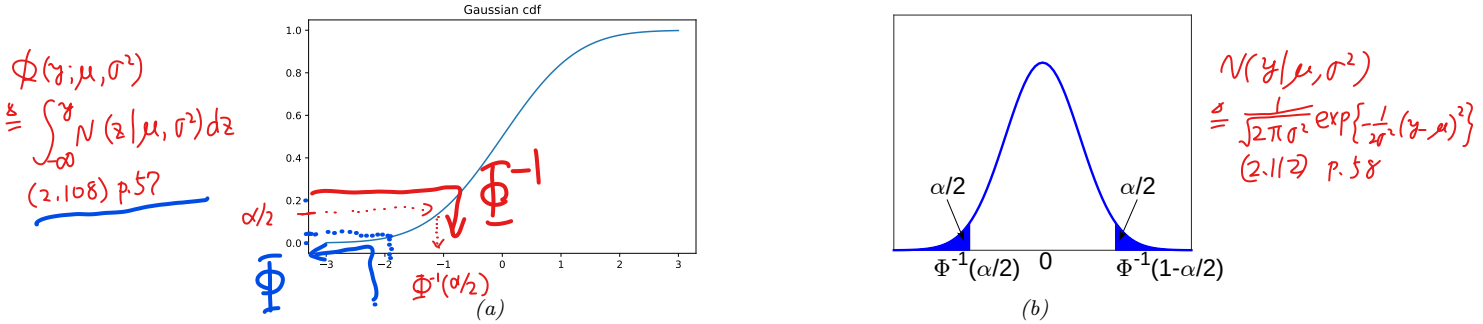


Figure 2.2: (a) Plot of the cdf for the standard normal, $\mathcal{N}(0,1)$. Generated by gauss_plot.ipynb. (b) Corresponding pdf. The shaded regions each contain $\alpha/2$ of the probability mass. Therefore the nonshaded region contains $1-\alpha$ of the probability mass. The leftmost cutoff point is $\Phi^{-1}(\alpha/2)$, where Φ is the cdf of the Gaussian. By symmetry, the rightmost cutoff point is $\Phi^{-1}(1-\alpha/2) = -\Phi^{-1}(\alpha/2)$. Generated by quantile_plot.ipynb.

In general, we define the **cumulative distribution function** or **cdf** of the rv X as follows:

$$P(x) \triangleq \Pr(X \leq x) \tag{2.11}$$

(Note that we use a capital P to represent the cdf.) Using this, we can compute the probability of being in any interval as follows:

$$\Pr(a < X \leq b) = P(b) - P(a) \tag{2.12}$$

Cdf's are monotonically non-decreasing functions. See Figure 2.2a for an example, where we illustrate the cdf of a standard normal distribution, $\mathcal{N}(x|0,1)$; see Section 2.6 for details.

2.2.2.2 Probability density function (pdf)

We define the **probability density function** or **pdf** as the derivative of the cdf:

$$p(x) \triangleq \frac{d}{dx} P(x) \tag{2.13}$$

(Note that this derivative does not always exist, in which case the pdf is not defined.) See Figure 2.2b for an example, where we illustrate the pdf of a univariate Gaussian (see Section 2.6 for details).

Given a pdf, we can compute the probability of a continuous variable being in a finite interval as follows:

$$\Pr(a < X \leq b) = \int_a^b p(x)dx = P(b) - P(a) \tag{2.14}$$

As the size of the interval gets smaller, we can write

$$\Pr(x < X \leq x + dx) \approx p(x)dx \tag{2.15}$$

Intuitively, this says the probability of X being in a small interval around x is the density at x times the width of the interval.

Author: Kevin P. Murphy. (C) MIT Press. CC-BY-NC-ND license

2.2.2.3 Quantiles 分位数

If the cdf P is strictly monotonically increasing, it has an inverse, called the **inverse cdf**, or **percent point function** (**ppf**), or **quantile function**.

If P is the cdf of X, then $P^{-1}(q)$ is the value x_q such that $\Pr(X \leq x_q) = q$; this is called the q'th **quantile** of P. The value $P^{-1}(0.5)$ is the **median** of the distribution, with half of the probability mass on the left, and half on the right. The values $P^{-1}(0.25)$ and $P^{-1}(0.75)$ are the lower and upper **quartiles**.

第1 四分位数 第3.

For example, let Φ be the cdf of the Gaussian distribution $\mathcal{N}(0,1)$, and Φ^{-1} be the inverse cdf. Then points to the left of $\Phi^{-1}(\alpha/2)$ contain $\alpha/2$ of the probability mass, as illustrated in Figure 2.2b. By symmetry, points to the right of $\Phi^{-1}(1-\alpha/2)$ also contain $\alpha/2$ of the mass. Hence the central interval $(\Phi^{-1}(\alpha/2), \Phi^{-1}(1-\alpha/2))$ contains $1-\alpha$ of the mass. If we set $\alpha = 0.05$, the central 95% interval is covered by the range

$$(\Phi^{-1}(0.025), \Phi^{-1}(0.975)) = (-1.96, 1.96) \tag{2.16}$$

If the distribution is $\mathcal{N}(\mu, \sigma^2)$, then the 95% interval becomes $(\mu - 1.96\sigma, \mu + 1.96\sigma)$. This is often approximated by writing $\mu \pm 2\sigma$.

$\mu \pm \sigma$ 68% $\mu \pm 3\sigma$ 99.7%

2.2.3 Sets of related random variables

同時分布

In this section, we discuss distributions over sets of related random variables.

Suppose, to start, that we have two random variables, X and Y. We can define the **joint distribution** of two random variables using $p(x,y) = p(X=x, Y=y)$ for all possible values of X and Y. If both variables have finite cardinality, we can represent the joint distribution as a 2d table, all of whose entries sum to one. For example, consider the following example with two binary variables:

濃度 基数 取る数の数. 例. Xのcardinalityは2

$p(X,Y)$	$Y=0$	$Y=1$
$X=0$	0.2	0.3
$X=1$	0.3	0.2

0.5 ← margin

If two variables are independent, we can represent the joint as the product of the two marginals. If both variables have finite cardinality, we can factorize the 2d joint table into a product of two 1d vectors, as shown in Figure 2.3.

Given a joint distribution, we define the **marginal distribution** of an rv as follows:

周辺 分布

$$p(X=x) = \sum_y p(X=x, Y=y) \tag{2.17}$$

和則

全確率の法則

where we are summing over all possible states of Y. This is sometimes called the **sum rule** or the **rule of total probability**. We define $p(Y=y)$ similarly. For example, from the above 2d table, we see $p(X=0) = 0.2 + 0.3 = 0.5$ and $p(Y=0) = 0.2 + 0.3 = 0.5$. (The term "marginal" comes from the accounting practice of writing the sums of rows and columns on the side, or margin, of a table.)

We define the **conditional distribution** of an rv using

条件付 分布

$$p(Y=y|X=x) = \frac{p(X=x, Y=y)}{p(X=x)} \tag{2.18}$$

We can rearrange this equation to get

$$p(x,y) = p(x)p(y|x) \tag{2.19}$$

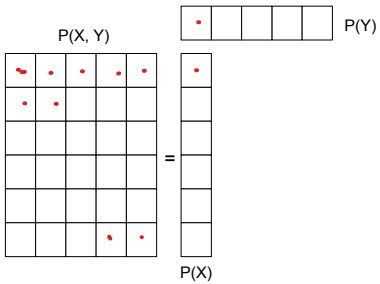


Figure 2.3: Computing $p(x, y) = p(x)p(y)$, where $X \perp Y$. Here X and Y are discrete random variables; X has 6 possible states (values) and Y has 5 possible states. A general joint distribution on two such variables would require $(6 \times 5) - 1 = 29$ parameters to define it (we subtract 1 because of the sum-to-one constraint). By assuming (unconditional) independence, we only need $(6 - 1) + (5 - 1) = 9$ parameters to define $p(x, y)$.

This is called the **product rule**.

By extending the product rule to D variables, we get the **chain rule of probability**:

$$p(\boldsymbol{x}_{1:D}) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3) \dots p(x_D|\boldsymbol{x}_{1:D-1}) \tag{2.20}$$

This provides a way to create a high dimensional joint distribution from a set of conditional distributions. We discuss this in more detail in Section 3.6.

2.2.4 Independence and conditional independence

We say X and Y are **unconditionally independent** or **marginally independent**, denoted $X \perp Y$, if we can represent the joint as the product of the two marginals (see Figure 2.3), i.e.,

$$X \perp Y \iff p(X, Y) = p(X)p(Y) \tag{2.21}$$

In general, we say a set of variables $X_1, \dots, X_n$ is (mutually) **independent** if the joint can be written as a product of marginals for all subsets $\{X_1, \dots, X_m\} \subseteq \{X_1, \dots, X_n\}$: i.e.,

$$p(X_1, \dots, X_m) = \prod_{i=1}^{m} p(X_i) \tag{2.22}$$

For example, we say X_1, X_2, X_3 are mutually independent if the following conditions hold: $p(X_1, X_2, X_3) = p(X_1)p(X_2)p(X_3)$, $p(X_1, X_2) = p(X_1)p(X_2)$, $p(X_2, X_3) = p(X_2)p(X_3)$, and $p(X_1, X_3) = p(X_1)p(X_3)$.[2]

Unfortunately, unconditional independence is rare, because most variables can influence most other variables. However, usually this influence is mediated via other variables rather than being direct. We therefore say X and Y are **conditionally independent** (CI) given Z iff the conditional joint can be written as a product of conditional marginals:

$$X \perp Y \mid Z \iff p(X, Y|Z) = p(X|Z)p(Y|Z) \tag{2.23}$$

2. For further discussion, see https://github.com/probml/pml-book/issues/353#issuecomment-1120327442.