

2.4. Random Variables

Lecture based on <https://github.com/gwthomas/math4ml> (Garrett Thomas, 2018)

Prof. Dr. Christoph Lippert

Digital Health & Machine Learning

Random Variables

A **random variable** is some uncertain quantity with an associated probability distribution over the values it can assume.

Formally, a random variable on a probability space (Ω, \mathcal{F}, P) is a function $X : \Omega \rightarrow \mathbb{R}$.

We denote the range of X by $X(\Omega) = \{X(\omega) : \omega \in \Omega\}$.

Example

suppose X is the number of heads in two tosses of a fair coin.

The sample space is

$$\Omega = \{hh, tt, ht, th\}$$

and X is determined completely by the outcome ω , i.e. $X = X(\omega)$.

For example, the event $X = 1$ is the set of outcomes $\{ht, th\}$.

The values of a random variable and Ω are related as follows:

the event that the value of X lies in some set $S \subseteq \mathbb{R}$ is

$$X \in S = \{\omega \in \Omega : X(\omega) \in S\}$$

Note that special cases of this definition include X being equal to, less than, or greater than some specified value.

For example

$$P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\})$$

Notation:

- $p(X)$ denotes the entire **probability distribution** of X
- $p(x)$ for the evaluation of the function p at a particular value $x \in X(\Omega)$.

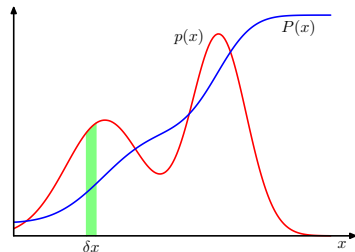
If p is parameterized by θ , we write $p(X; \theta)$ or $p(x; \theta)$

The **cumulative distribution function** (c.d.f.) gives the probability that a random variable is at most a certain value:

$$F(x) = P(X \leq x)$$

The c.d.f. can be used to give the probability that a variable lies within a certain range:

$$P(a < X \leq b) = F(b) - F(a)$$



(C.M. Bishop, Pattern Recognition and Machine Learning)

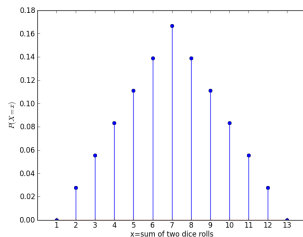
A **discrete random variable** is a random variable that has a countable range and assumes each value in this range with positive probability.

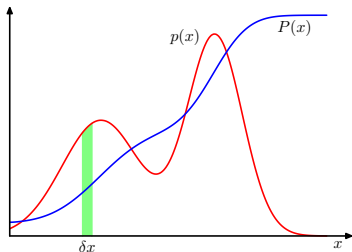
Discrete random variables are completely specified by their **probability mass function** (p.m.f.) $p : X(\Omega) \rightarrow [0, 1]$ which satisfies

$$\sum_{x \in X(\Omega)} p(x) = 1$$

For a discrete X , the probability of a particular value is given exactly by its p.m.f.:

$$P(X = x) = p(x)$$





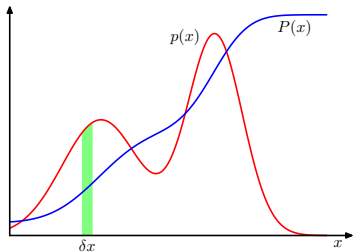
(C.M. Bishop, Pattern Recognition and Machine Learning)

A **continuous random variable** is a random variable that has an uncountable range and assumes each value in this range with probability zero.

Usually there exists a function $p : \mathbb{R} \rightarrow [0, \infty)$ that satisfies

$$F(x) \equiv \int_{-\infty}^x p(z) \, dz$$

p is called a **probability density function** (p.d.f.).

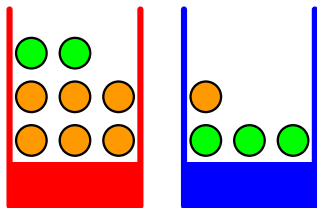


Hence, the p.d.f. must satisfy

$$\int_{-\infty}^{\infty} p(x) dx = 1$$

The values of this function are not themselves probabilities, since they could exceed 1.

Multiple Random Variables



(C.M. Bishop, Pattern Recognition and Machine Learning)

Example

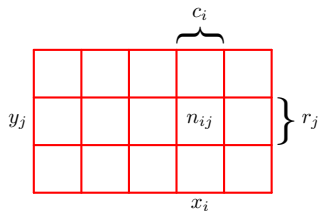
We repeat ten times:

- ① We pick one of the jars with equal probabilities.
- ② We sample (with replacement) a fruit from the jar.

We define the random variables

- X : Number of times we chose the red jar (R).
- Y : Number of times we picked an orange (O).

For some random variables X_1, \dots, X_n , the **joint distribution** is written $p(X_1, \dots, X_n)$ and gives probabilities over entire assignments to all the X_i simultaneously.



(C.M. Bishop, Pattern Recognition and Machine Learning) For $N \rightarrow \infty$:

- $p(x_i, y_j) = \frac{n_{ij}}{N}$

Example

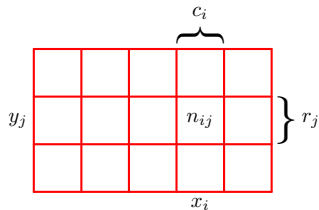
Random variables $X \in [1, \dots, M]$, $Y \in [1, \dots, L]$.

After N draws, we define:

- n_{ij} : number of instances, where $X = x_i$ and $Y = y_j$.
- c_i : number of instances, where $X = x_i$.
- r_j : number of instances, where $Y = y_j$.

Joint distributions

Sum Rule



(C.M. Bishop, Pattern Recognition and Machine Learning)

For $N \rightarrow \infty$:

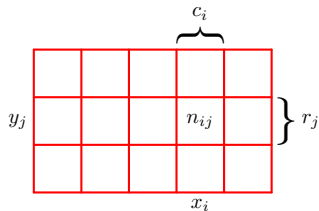
- $p(x_i, y_j) = \frac{n_{ij}}{N}$
- $p(x_i) = \frac{c_i}{N} = \sum_{j=1}^L p(x_i, y_j)$

If we have a joint distribution over some set of random variables, it is possible to obtain a distribution for a subset of them by “summing out” (or “integrating out” in the continuous case) the variables we don’t care about:

$$p(X) = \sum_y p(X, y)$$

Joint distributions

Product Rule



(C.M. Bishop, Pattern Recognition and Machine Learning) For $N \rightarrow \infty$:

- $p(x_i, y_j) = \frac{n_{ij}}{N}$
- $p(x_i) = \frac{c_i}{N} = \sum_{j=1}^L p(x_i, y_j)$
- $p(y_j|x_i) = \frac{n_{ij}}{c_i}$
- $\Rightarrow p(x_i, y_j) = \frac{n_{ij}}{c_i} \frac{c_i}{N}$

$$p(X, Y) = p(Y|X)p(X)$$

It follows **Bayes' theorem**:

$$p(Y|X) = \frac{p(X, Y)}{p(X)}$$

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

Or, using the sum rule,

$$p(y|X) = \frac{p(X|y)p(y)}{\sum_y p(X, y)}$$

We say that two variables X and Y are **independent** if their joint distribution factors into their respective distributions, i.e.

$$p(X, Y) = p(X)p(Y)$$

It is often convenient to assume that several random variables are **independent and identically distributed** (i.i.d.), so that their joint distribution can be factored entirely:

$$p(X_1, \dots, X_n) = \prod_{i=1}^n p(X_i)$$

where X_1, \dots, X_n all share the same p.m.f./p.d.f.