

Definition 1. Data that is counted is **discrete** while data that is measured is **continuous** (Gates et al., 2018b).

Definition 2. A **combination** is an arrangement of items without regard to the order. A **permutation** is an arrangement of items in a specific order. (Gates et al., 2018a).

Definition 3. An **experiment** is a procedure that yields one of a given set of possible outcomes (Rosen, 2002).

Definition 4. The **sample space** of the experiment is the set of possible outcomes (Rosen, 2002).

Definition 5. An **event** is a subset of the sample space (Rosen, 2002).

Definition 6. Probability is the measure of the likelihood that an event will occur. Probability is quantified as a number between 0 and 1, where 0 indicates impossibility and 1 certainty.

Definition 7. Finite probability. If S is a finite nonempty sample space of equally likely outcomes, and E is an event, that is, a subset of S , then the probability of E is $p(E) = \frac{|E|}{|S|}$ (Rosen, 2002).

Definition 8. A **random variable**¹ is a function from the sample space of an experiment to the set of real numbers.

$$X : S \rightarrow \mathbb{R}$$

A random variable $X(t)$ maps each outcome t to a real number (Rosen, 2002).

Note. It is worth clarifying that “random variable” is a misnomer. A random variable is a function. It is not a variable, and it is not random! It is a representation of an underlying random system. The name *random variable* (the translation of *variabile casuale*) was introduced by the Italian mathematician F. P. Cantelli in 1916 (Rosen, 2002).

1 Discrete Probability Distributions

Definition 9. The **distribution** of a random variable X on a sample space S is the set of pairs $(r, p(X = r))$ for all $r \in X(S)$, where $p(X = r)$ is the probability that X takes the value r . The set of pairs in this distribution is determined by the probabilities $p(X = r)$ for $r \in X(S)$ (Rosen, 2002).

1.1 Bernoulli Distribution

Definition 10. Discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$.

¹ Random variables translate outcomes in the sample space of an experiment to real numbers so that we can reason about them with mathematics.

Random variable

X = The outcome of the **Bernoulli experiment** is “success”.

Probability mass function

$$f(k; p) = \Pr(k; p) = \Pr(X = k) = p^k(1 - p)^{1-k} \quad \text{for } k \in \{0, 1\}$$

Parameter

p = Probability of the outcome of a **Bernoulli experiment** being a success.

1.2 Binomial Distribution

Definition 11. Discrete probability distribution of the number of successes in a sequence of n independent Bernoulli experiments. For a single trial, that is, $n = 1$, the binomial distribution is a Bernoulli distribution.

Note. The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N . If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a **hypergeometric** distribution, not a binomial one. However, for N much larger than n , the binomial distribution remains a good approximation, and is widely used.

Random variable

X = The number of successes in a sequence of n independent **Bernoulli experiments** is k .

Probability mass function of getting k successes in n trials.

$$f(k; n, p) = \Pr(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad \text{for } k \in \{0, 1, \dots, n\}$$

Parameters

p = Probability of the outcome of a **Bernoulli experiment** being a success.

n = Number of trials (experiemnts).

Definition 12. The binomial coefficient, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, is the coefficient of the x^k term in the polynomial expansion of the binomial power $(1 + x)^n = \binom{n}{0}x^0 + \binom{n}{1}x^1 + \dots + \binom{n}{n}x^n$, that is, the **binomial theorem**.

Explanation k successes occur with probability p^k and $n - k$ failures occur with probability $(1 - p)^{n-k}$. The k successes can occur anywhere among the n trials, and there are $\binom{n}{k}$ different ways of distributing k successes in a sequence of n trials.

1.3 Poisson Distribution

Definition 13. Discrete probability distribution of a given number of events occurring in a fixed space interval, usually time, and these events occur with a known constant rate and independently of the time since the last event. The Poisson distribution is derived from the **binomial distribution** by splitting up the interval into n subintervals, each of which is so small that at most one event could occur in it with non-zero probability (Wackerly et al., 2008).

Random variable

X = The number of events in an interval is k .

Probability mass function of getting k events in an interval.

$$f(k; \lambda) = \Pr(k; \lambda) = \Pr(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Parameter

λ = event rate, the average number of events in an interval.

1.4 Geometric Distribution

Geometric random variable has two alternate formulations:

X_1 = The number of **Bernoulli experiments** needed to get the first success.

X_2 = The number of failed **Bernoulli experiments** before the first success.

Probability mass function for each corresponding random variable.

$$\begin{aligned} f_1(k; p) &= \Pr_1(k; p) = \Pr(X_1 = k) = (1 - p)^{k-1}p && \text{for } k = 1, 2, \dots, 0 \leq p \leq 1 \\ f_2(k; p) &= \Pr_2(k; p) = \Pr(X_2 = k) = (1 - p)^k p && \text{for } k = 0, 1, 2, \dots, 0 \leq p < 1 \end{aligned}$$

Parameter

p = Probability of the outcome of a Bernoulli experiment being success.

Example Probability that the first coin flip turns up heads:

$$p(X_1 = 1) = (1 - p)^{1-1}p = (1 - p)^0 p = p(X_2 = 0)$$

2 Continuous Probability Distributions

2.1 Exponential Distribution

Definition 14. The exponential distribution (also known as negative exponential distribution) is the probability distribution that describes the time between events in a **Poisson point process**, i.e., a process in which events occur continuously and independently at a constant average rate. It is a particular case of the gamma distribution. It is the continuous analogue of the **geometric distribution**, and it has the key property of being memoryless. In addition to being used for the analysis of Poisson point processes it is found in various other contexts.

Exponential random variable

X = The amount of time that passes before the next event.

Probability density function

$$f(x; \lambda) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$

Parameter

λ = event rate, the average number of events in an interval.

Derivation from the geometric distribution.

1. Split up each interval into n sub-intervals such that the probability of an event occurring in a certain sub-interval is $p = \frac{\lambda}{n}$
2. For a **geometric random variable** Y = The number of trials needed to get the first success, $p(X \leq b) \approx P(Y \leq b \cdot n)$

$$\begin{aligned} P(Y \leq b \cdot n) &= \sum_{k=1}^{b \cdot n} P(Y = k) \\ &= \sum_{k=1}^{b \cdot n} (1-p)^{k-1} p = \sum_{k=1}^{b \cdot n} \left(1 - \frac{\lambda}{n}\right)^{k-1} \cdot \frac{\lambda}{n} \\ &= \frac{\lambda}{n} \cdot \sum_{k=1}^{b \cdot n} \left(1 - \frac{\lambda}{n}\right)^{k-1} \\ &= \frac{\lambda}{n} \cdot \sum_{k=0}^{b \cdot n - 1} \left(1 - \frac{\lambda}{n}\right)^k \end{aligned}$$

$$\sum_{k=0}^m a^k = \frac{1 - a^{m+1}}{1 - a} \quad \text{finite geometric sum}$$

$$m = b \cdot n - 1 \text{ and } a = 1 - \frac{\lambda}{n}$$

$$\begin{aligned}
P(Y \leq b \cdot n) &= \frac{\lambda}{n} \cdot \frac{1 - \left(1 - \frac{\lambda}{n}\right)^{b \cdot n}}{1 - \left(1 - \frac{\lambda}{n}\right)} \\
&= 1 - \left(1 - \frac{\lambda}{n}\right)^{b \cdot n} \\
&= 1 - \left(\left(1 - \frac{\lambda}{n}\right)^n\right)^b
\end{aligned}$$

Yue (2018) present an alternate derivation from a **Markov process** representation of the exponential distribution.

3 Stochastic (Random) Process

Definition 15. A **stochastic process** is a random phenomenon that arises through a process which is developing in time in a manner controlled by probabilistic laws (Parzen, 1999). From a point of view of the mathematical theory of probability a stochastic process is best defined as a collection of random variables, indexed by points in some space.

$$\{X(t), t \in T\}$$

A more appropriate name in mathematics is **random field**. The set used to index the random variables is called the **index set**. Historically, the index set was some subset of the real line, such as the natural numbers, giving the index set the interpretation of time. Each random variable in the collection takes values from the same mathematical space known as the **state space**. An **increment** is the amount that a stochastic process changes between two index values, often interpreted as two points in time.

Note. Even though the index set can be any set in any space, generally more results and theorems are possible for stochastic processes when the index set is ordered. Most commonly, random variables in a stochastic process are indexed by the positive numbers along the real number line, interpreted as time. Among these, the **Brownian motion process** and the **Poisson process** are considered the most important and central in the theory of stochastic processes. Other stochastic processes include **random walk** and **Markov Chain** and **martingales**. A martingale models a fair game. A simple random walk is a Markov chain and also a martingale.

4 Point Process

Definition 16. A point process (or **point field**) is a collection of mathematical points randomly located on some underlying mathematical space. Point processes on the real line form an important special case that is particularly amenable to study because the points are ordered in a natural way, and the whole point process can be described completely by the (random) intervals between the points.

Note. Different from stochastic processes, point processes are defined as a collection points randomly located on some space and not as a collection of random variables indexed by some index space. Various interpretations of a point process exist:

- a random object that arises from or is associated with a stochastic process
- a stochastic process indexed by sets of the underlying space on which it is defined

4.1 Renewal (Point) Process

Definition 17. Historically the first point processes that were studied had the real half line $\mathbb{R}_+ = [0, \infty)$ as their state space, which in this context is usually interpreted as time. These studies were motivated by the wish to model telecommunication systems, in which the points represented events in time, such as calls to a telephone exchange. Point processes on \mathbb{R}_+ are typically described by giving the sequence of their (random) inter-event times (T_1, T_2, \dots) , from which the actual sequence (X_1, X_2, \dots) of event times can be obtained as

$$X_k = \sum_{j=1}^k T_j \quad \text{for } k \geq 1.$$

If inter-event times are independent and identically distributed, the point process is called a **renewal process**.

Note. The **intensity** $\lambda(t|H_t)$ of a point process on the real half-line with respect to a filtration H_t is defined as

$$\lambda(t|H_t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \Pr(\text{One event occurs in time interval } [t, t + \Delta t] | H_t)$$

H_t can denote the history of event-point times preceding time t but can also correspond to other *filtrations* (for example in the case of a Cox process). The *compensator* of a point process, also known as the dual-predictable projection, is the integrated conditional intensity function defined by

$$\Lambda(s, u) = \int_s^u \lambda(t|H_t) dt$$

4.2 Poisson (Point)² Process

Definition 18. The number of points in a region of finite size within a Poisson process is a random variable with a **Poisson distribution**. The Poisson point process is often defined on the real line, where it can be considered as a **stochastic process**. The Poisson process is a point process with convenient mathematical properties. The Poisson point process has the property that each point is stochastically independent to all the other points in the process. The Poisson process has been used to build other point processes where the points are not independent of each other.

² The word point is often omitted, but there are other Poisson processes of objects, which, instead of points, consist of more complicated mathematical objects such as lines and polygons, and such processes can be based on the Poisson point process.

Note. **Poisson Counting Process**

Definition 19. A homogeneous Poisson point process on the positive half-line that represents the total number of occurrences or events that have happened up to and including time t .

$$\{N(t), t \geq 0\}$$

The probability of random variable $N(t)$ being equal to n :

$$P\{N(t) = n\} = e^{-\lambda t} \frac{(\lambda t)^n}{n!}$$

4.3 PCIM

Definition 20. Model for learning nonlinear temporal dependencies in event streams. Events carry both information about their timing and their type. Used to predict users' interests from their web queries and the type of system from supercomputer system error log. The dependencies between events can be due to both their timing and their types.

The **Piecewise-Constant Conditional Intensity Model** (PCIM) is a class of marked point processes that can model the types and timing of events. This model captures the dependencies of each type of event on events in the past through a set of piecewise-constant conditional intensity functions. Decision trees represent these dependencies. The decision trees are the piecewise-constant conditional intensity functions (Gunawardana et al., 2011)

A conjugate prior for this model allows for closed-form computation of the marginal likelihood and parameter posteriors. Model selection then becomes a problem of choosing a decision tree. Decision tree induction can be done efficiently because of the closed form for the marginal likelihood.

Poisson Networks (Rajaram et al., 2005) (also piece-wise constant conditional intensity models) are closely related to PCIMs.

Conditional Intensity Model (Didelez, 2008; Daley and Vere-Jones, 2003):

$$p(x|\theta) = \prod_{l \in \mathcal{L}} \prod_{i=1}^n \lambda_l(t_i|h_i, \theta)^{\mathbf{1}_l(l_i)} e^{-\Lambda_l(t_i|h_i; \theta)}$$

where $\Lambda_l(t|x; \theta) = \int_{-\infty}^t \lambda_l(\tau|x; \theta) d\tau$ for each event sequence x and the indicator function $\mathbf{1}_l(l')$ is one if $l' = l$ and zero otherwise. The rate parameter changes as a function of time and history and is computed by the conditional intensity function $\lambda_l(t_i|h_i, \theta)$.

Note. The *Conditional Intensity Model* is analogous to the likelihood of the sequence of samples from the exponential distribution: Let X be exponential distributed with rate parameter λ and x_1, \dots, x_n are n independent samples from X . The likelihood function is given by:

$$L(\lambda; x) = p(x; \lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i}$$

When there is only one label, the conditional intensity model becomes

$$p(x|\theta) = \prod_{i=1}^n \lambda(t_i|h_i, \theta) e^{-\Lambda(t_i|h_i; \theta)}$$

and when the rate parameter is fixed $\lambda(t_i|h_i, \theta) = \lambda$ it becomes $L(\lambda; x)$.

References

D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2003. ISBN 0-387-95541-0.

- Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(1):245–264, 2008. URL <http://www.jstor.org/stable/20203821>.
- Les Gates, Dianne Gentry, and David Sevilla. *Combinations and Permutations*, 2018a. URL <https://www.mathsisfun.com/combinatorics/combinations-permutations.html>.
- Les Gates, Dianne Gentry, and David Sevilla. *Discrete and Continuous Data*, 2018b. URL <https://www.mathsisfun.com/data/data-discrete-continuous.html>.
- Asela Gunawardana, Christopher Meek, and Puyang Xu. A model for temporal dependencies in event streams. In *Neural Information Processing Systems*. Neural Information Processing Systems Foundation, 2011. URL <https://www.microsoft.com/en-us/research/publication/a-model-for-temporal-dependencies-in-event-streams/>.
- Emanuel Parzen. *Stochastic Processes*. SIAM, 1999.
- Shyamsundar Rajaram, Thore Graepel, and Ralf Herbrich. Poisson-networks: A model for structured point processes. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics, January 2005. URL <https://www.microsoft.com/en-us/research/publication/poisson-networks-a-model-for-structured-point-processes/>.
- Kenneth H. Rosen. *Discrete Mathematics and Its Applications*. McGraw-Hill Higher Education, 7th edition, 2002.
- Dennis D. Wackerly, William Mendenhall, and Richard L. Scheaffer. *Mathematical Statistics with Applications*. Thomson Learning, Inc., 7th edition, 2008.
- Dick K.P. Yue. *Derivation of Exponential Distribution*, 2018. URL <https://ocw.mit.edu/courses/mechanical-engineering/2-854-introduction-to-manufacturing-systems-fall-2016/lecture-notes/derivation-of-exponential-distribution/>.