Machine Learning

Part 1: Mathematical Foundation of Machine Learning

Zengchang Qin (PhD)

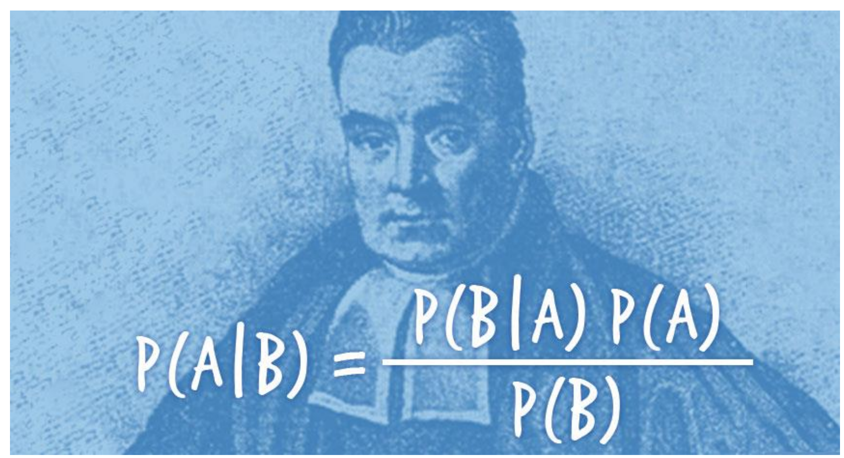
Probability & Statistics

Probability (Objective and Subjective)

The first approach is to define probability in terms of frequency of occurrence, as a percentage of successes in a moderately large number of similar situations.

Such an interpretation is often natural. For example, when we say that a perfectly manufactured coin lands on heads “with probability 50%,” we typically mean “roughly half of the time.”

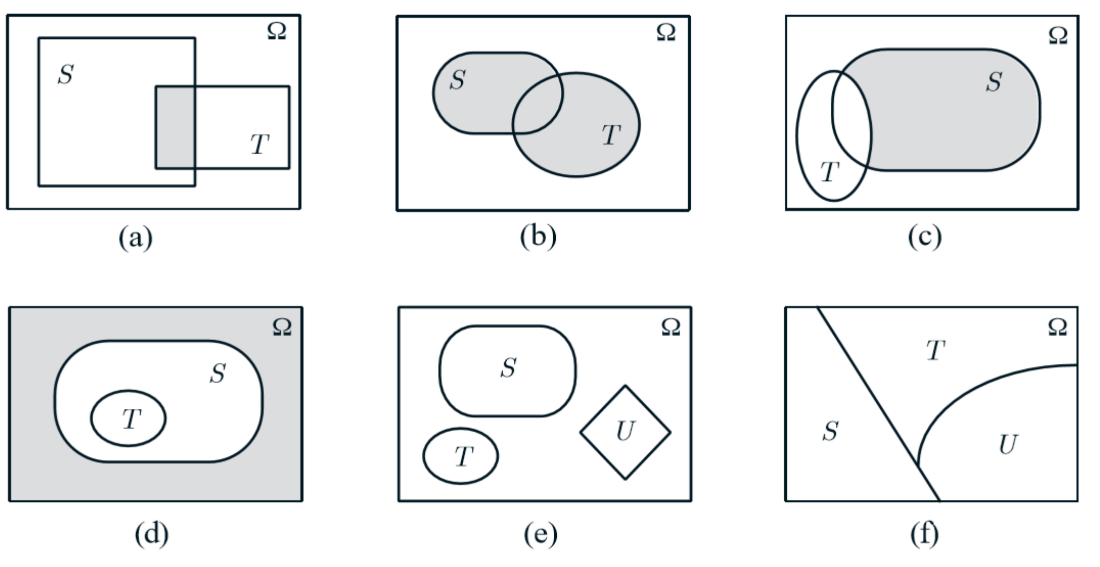
Consider, for example, a scholar who asserts that the Lliad and the Odyssey were composed by the same person, with probability 90%. Such an assertion conveys some information, but not in terms of frequencies, since the subject is a one-time event. Rather, it is an expression of the scholar’s subjective belief.



Set Operation

Examples of Venn diagrams.

(a) The shaded region is S ∩ T.
(b) The shaded region is S ∪ T.
(c) The shaded region is S ∩c(T) .
(d) Here, T ⊂ S. The shaded region is the complement of S.
(e) The sets S, T, and U are disjoint.
(f) The sets S, T, and U form a partition of the set **Ω**

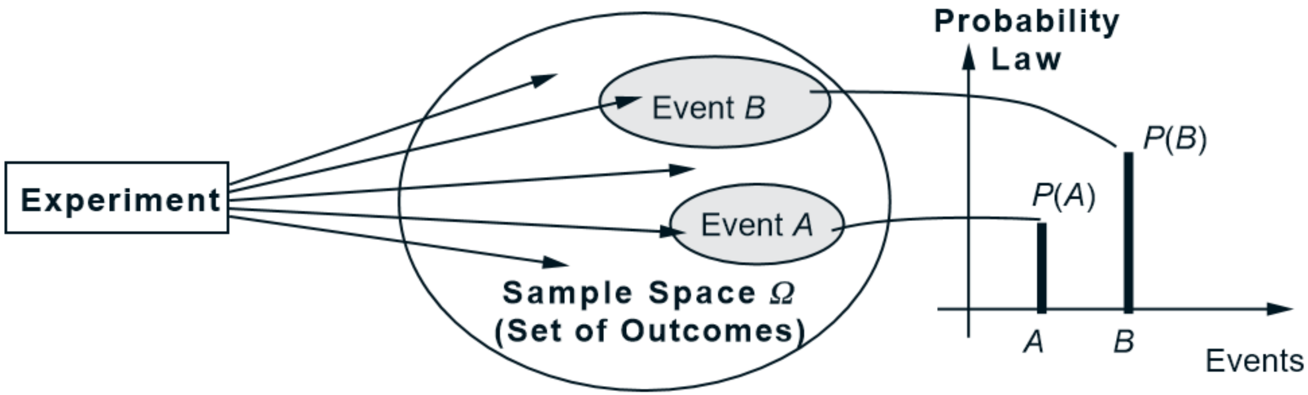


Probabilistic Models

Elements of a Probabilistic Model

The sample space Ω, which is the set of all possible outcomes of an experiment.

The **probability law,** which assigns to a set A of possible outcomes (also called an event) anonnegative number P(A) (called the probability of A) that encodes our knowledge or belief about the collective “likelihood” of the elements of A. The probability law must satisfy certain properties to be introduced shortly.



Probability Axioms

Probability Axioms

1. **(Nonnegativity)** $\mathbf{P}(A) \geq 0$, for every event A.
2. **(Additivity)** If A and B are two disjoint events, then the probability of their union satisfies

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

   Furthermore, if the sample space has an infinite number of elements and $A_1, A_2, \ldots$ is a sequence of disjoint events, then the probability of their union satisfies

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots$$

3. **(Normalization)** The probability of the entire sample space Ω is equal to 1, that is, $\mathbf{P}(\Omega) = 1$.

Conditional Probability

Properties of Conditional Probability

- The conditional probability of an event A, given an event B with $\mathbf{P}(B) > 0$, is defined by

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)},$$

  and specifies a new (conditional) probability law on the same sample space Ω. In particular, all known properties of probability laws remain valid for conditional probability laws.
- Conditional probabilities can also be viewed as a probability law on a new universe B, because all of the conditional probability is concentrated on B.
- In the case where the possible outcomes are finitely many and equally likely, we have

$$\mathbf{P}(A \mid B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}.$$

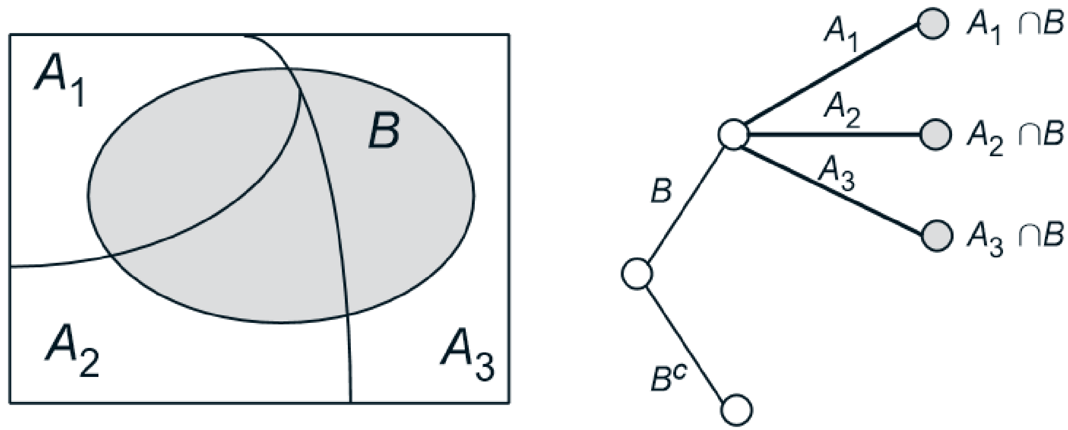
Let's consider a problem of conditional probability:

My neighbor John has two kids.

1. He told me that one of his two kids is a boy, what is the probability that the other one is a girl.

2. If I saw one's kids is playing outside, that is a boy, what is the probability that the other one is a girl.

Total Probability Theorem



Total Probability Theorem

Let $A_1, \ldots, A_n$ be disjoint events that form a partition of the sample space (each possible outcome is included in one and only one of the events $A_1, \ldots, A_n$) and assume that $\mathbf{P}(A_i) > 0$, for all $i = 1, \ldots, n$. Then, for any event B, we have

$$\begin{aligned}\mathbf{P}(B) &= \mathbf{P}(A_1 \cap B) + \cdots + \mathbf{P}(A_n \cap B) \\ &= \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n).\end{aligned}$$

Independence

Bayes' Rule

Let $A_1, A_2, \ldots, A_n$ be disjoint events that form a partition of space, and assume that $\mathbf{P}(A_i) > 0$, for all i. Then, for any ev that $\mathbf{P}(B) > 0$, we have

$$\begin{aligned}\mathbf{P}(A_i \mid B) &= \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n)}\end{aligned}$$

Rev. T. Bayes

Independence

- Two events A and B are said to independent if

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

  If in addition, $\mathbf{P}(B) > 0$, independence is equivalent to the condition

$$\mathbf{P}(A \mid B) = \mathbf{P}(A).$$

- If A and B are independent, so are A and B^c.
- Two events A and B are said to be conditionally independent, given another event C with $\mathbf{P}(C) > 0$, if

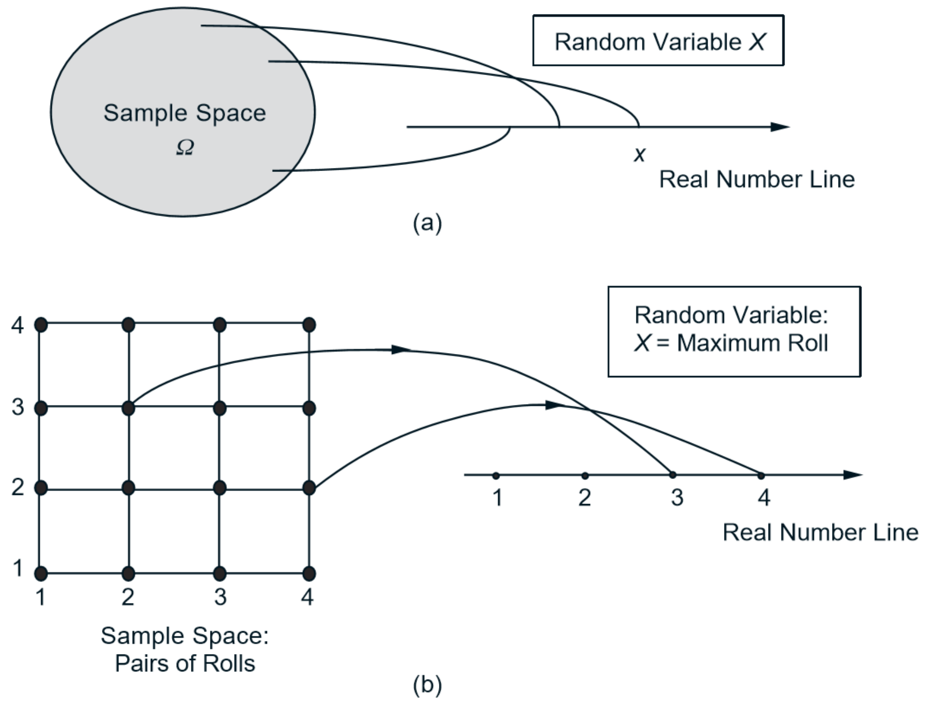
$$\mathbf{P}(A \cap B \mid C) = \mathbf{P}(A \mid C)\mathbf{P}(B \mid C).$$

  If in addition, $\mathbf{P}(B \cap C) > 0$, conditional independence is equivalent to the condition

$$\mathbf{P}(A \mid B \cap C) = \mathbf{P}(A \mid C).$$

- Independence does not imply conditional independence, and vice versa.

Random Variable



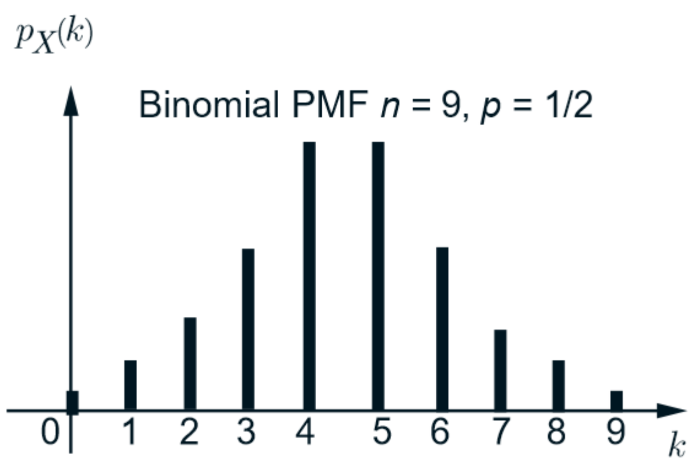
(a) Visualization of a random variable. It is a function that assigns a numerical value to each possible outcome of the experiment.

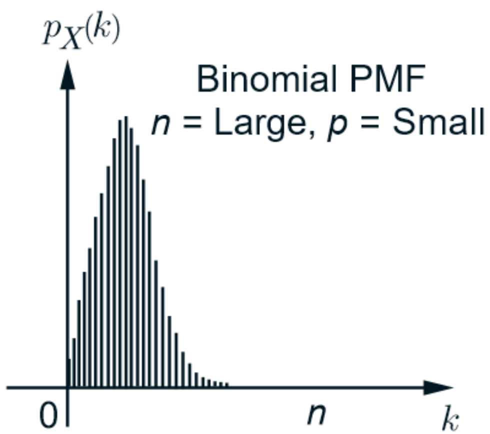
(b) An example of a random variable. The experiment consists of two rolls of a 4-sided die, and the random variable is the maximum of the two rolls. If the outcome of the experiment is (4,2), the experimental value of this random variable is 4.

Binomial Random Variable

At each toss, the coin comes up a head with probability p, and a tail with probability 1−p, independently of prior tosses. Let *X* be the number of heads in the *n*-toss sequence. We refer to X as a binomial random variable with parameters n and p.

$$p_X(k) = \mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n. \qquad \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$



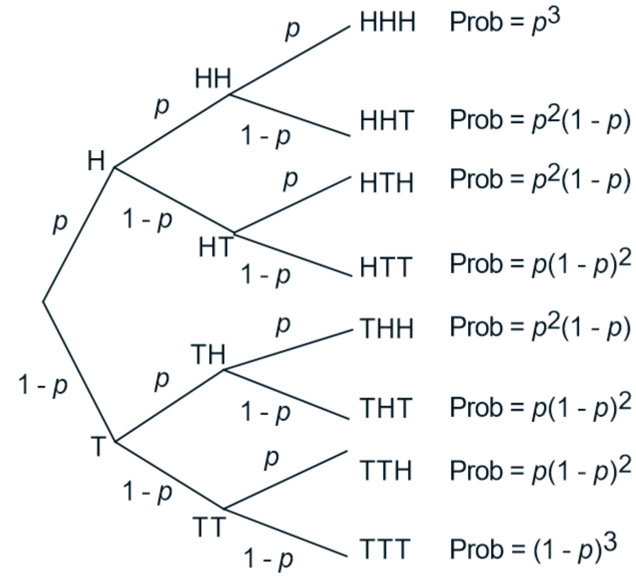


Binomial Probabilities

We showed above that the probability of any given sequence that contains k heads is $p^k(1-p)^{n-k}$, so we have

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where $\binom{n}{k}$ = number of distinct n-toss sequences that contain k heads.



The numbers $\binom{n}{k}$ (called "n choose k") are known as the **binomial coefficients**, while the probabilities $p(k)$ are known as the **binomial probabilities**. Using a counting argument, to be given in Section 1.6, one finds that

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}, \qquad k = 0, 1, \ldots, n,$$

Expectation of Random Variable and Function of Random Variables

Expectation

We define the **expected value** (also called the **expectation** or the **mean**) of a random variable X, with PMF $p_X(x)$, by[†]

$$\mathbf{E}[X] = \sum_x x p_X(x).$$

Expected Value Rule for Functions of Random Variables

Let X be a random variable with PMF $p_X(x)$, and let $g(X)$ be a real-valued function of X. Then, the expected value of the random variable $g(X)$ is given by

$$\mathbf{E}[g(X)] = \sum_x g(x) p_X(x).$$

Variance

Variance

The variance $\mathrm{var}(X)$ of a random variable X is defined by

$$\mathrm{var}(X) = \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)^2\big]$$

and can be calculated as

$$\mathrm{var}(X) = \sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x).$$

It is always nonnegative. Its square root is denoted by σ_X and is called the **standard deviation**.

Covariance

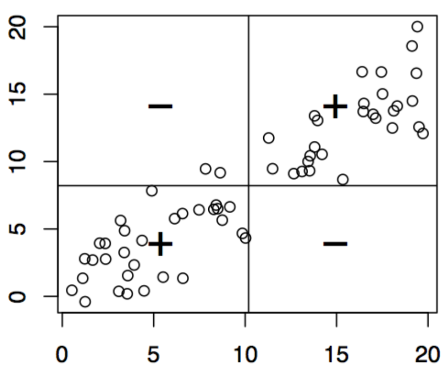
Covariance is a measure of the linear relationship between two random variables. We denote the covariance between X and Y as $\mathrm{Cov}(X, Y)$, and it is defined to be

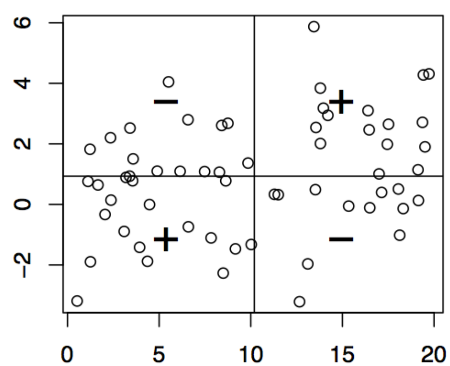
$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

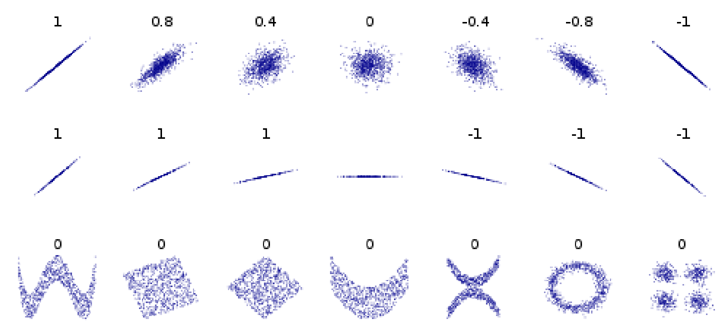
Note that the outer expectation must be taken over the joint distribution of X and Y.

Again, the linearity of expectation allows us to rewrite this as

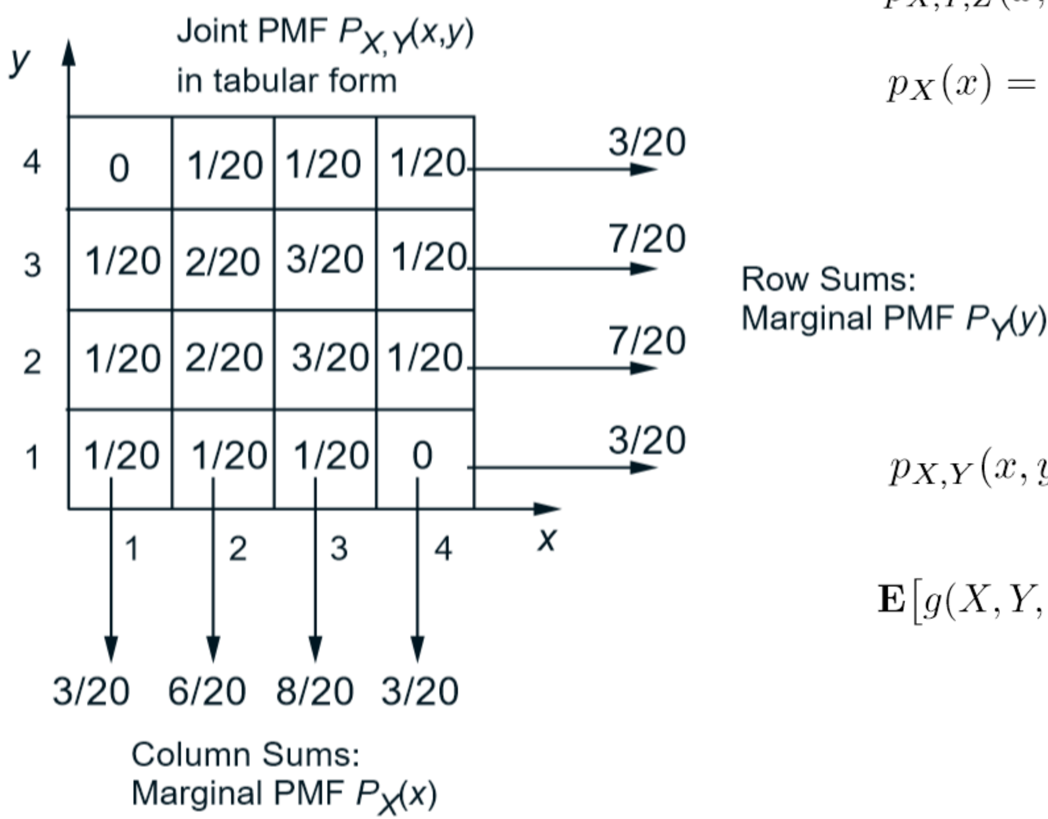
$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$







Joint Probability of More than One Random Variables



$$p_{X,Y,Z}(x,y,z) = \mathbf{P}(X=x, Y=y, Z=z),$$

$$p_X(x) = \sum_y \sum_z p_{X,Y,Z}(x,y,z).$$

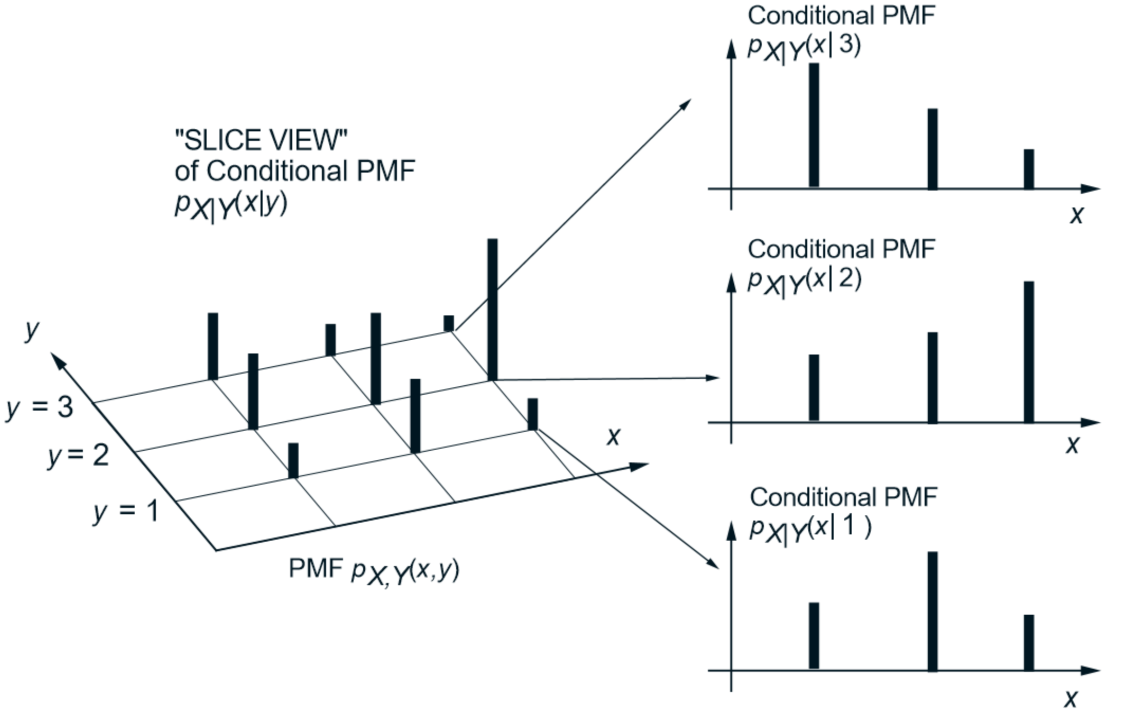
$$p_{X,Y}(x,y) = \sum_z p_{X,Y,Z}(x,y,z),$$

$$\mathbf{E}\big[g(X,Y,Z)\big] = \sum_{x,y,z} g(x,y,z) p_{X,Y,Z}(x,y,z),$$

Conditioning

The world is full of conditions, when we say independent, it usually implies "conditional independent".

Independent or dependent?



Visualization of the conditional PMF $p_{X|Y}(x \mid y)$. For each y, we view the joint PMF along the slice $Y = y$ and renormalize so that

$$\sum_x p_{X|Y}(x \mid y) = 1.$$

Continuous PDF

a **probability distribution** is a mathematical function that, stated in simple terms, can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment. From frequency to a continuous function.

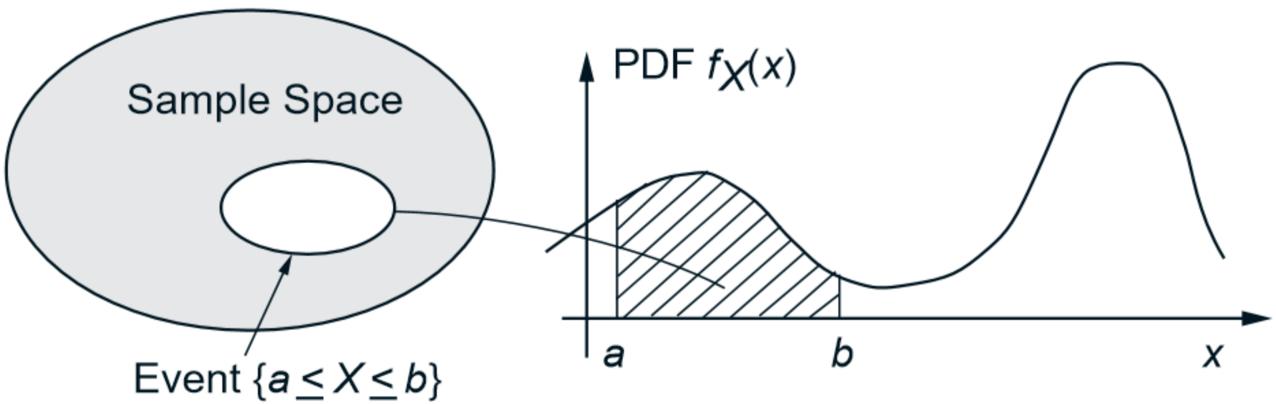
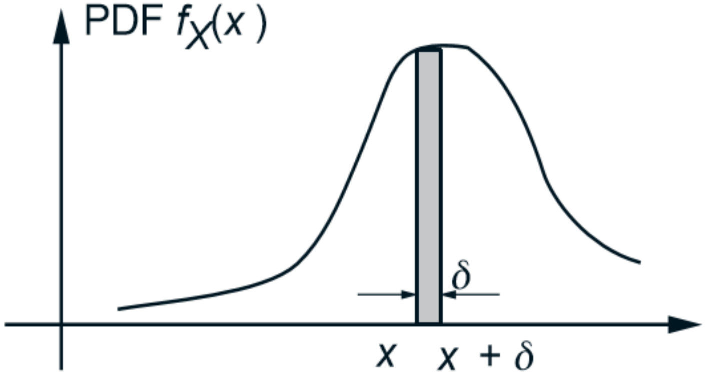
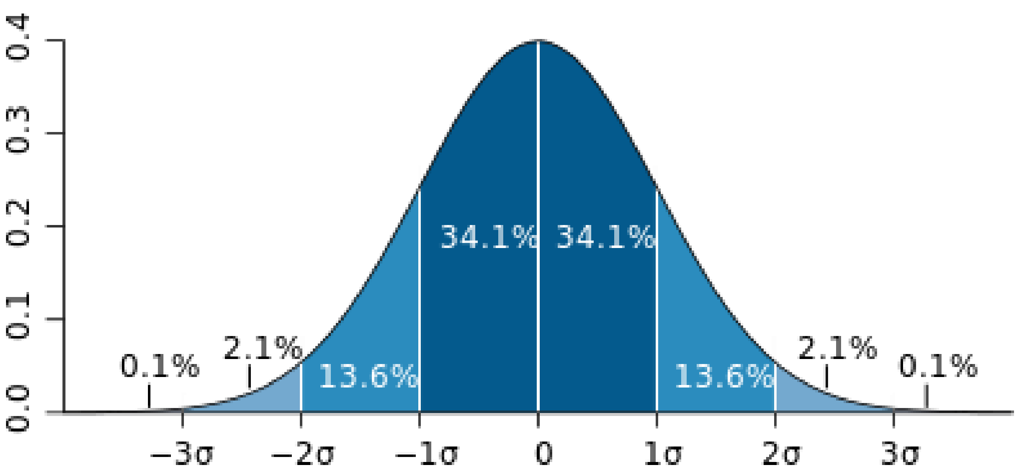


Illustration of a PDF. The probability that X takes value in an interval $[a, b]$ is $\int_a^b f_X(x)\,dx$, which is the shaded area in the figure.

Probability Distributions



Interpretation of the PDF $f_X(x)$ as "probability mass per unit length" around x. If δ is very small, the probability that X takes value in the interval $[x, x+\delta]$ is the shaded area in the figure, which is approximately equal to $f_X(x) \cdot \delta$.



Bernoulli Distribution

The probability distribution of any single experiment that asks a yes–no question; the question results in a Boolean valued outcome, a single bit of information whose value is success/yes/true/one with probability ***p***
and failure/no/false/zero with probability ***q***.

If X is a random variable with this distribution, we have:

$$\Pr(X=1) = p = 1 - \Pr(X=0) = 1 - q.$$

The probability mass function f of this distribution, over possible outcomes k, is

$$f(k;p) = \begin{cases} p & \text{if } k = 1, \\ 1-p & \text{if } k = 0. \end{cases}$$

This can also be expressed as

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

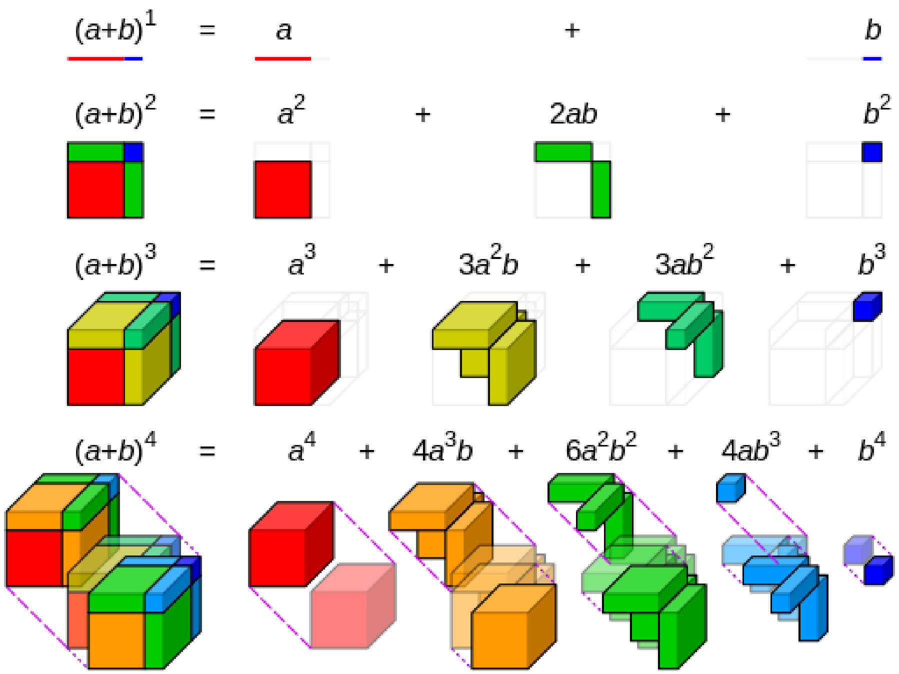
$$\mathrm{E}(X) = p \qquad \mathrm{E}[X^2] = \Pr(X=1)\cdot 1^2 + \Pr(X=0)\cdot 0^2 = p \cdot 1^2 + q \cdot 0^2 = p$$

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = p - p^2 = p(1-p) = pq$$

Jacob Bernoulli

Binomial Distribution

The binomial distribution with parameters ***n*** and ***p*** is the discrete probability distribution of the number of successes in a sequence of *n* independent Bernoulli trials of yes–no questions.



$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1

Pascal's triangle

Multinomial Distribution

$$f(x_1,\ldots,x_k;n,p_1,\ldots,p_k) = \Pr(X_1 = x_1 \text{ and } \ldots \text{ and } X_k = x_k)$$
$$= \begin{cases} \dfrac{n!}{x_1!\cdots x_k!} p_1^{x_1} \times \cdots \times p_k^{x_k}, & \text{when } \sum_{i=1}^k x_i = n \\ 0 & \text{otherwise,} \end{cases}$$

Suppose one does an experiment of extracting *n* balls of *k* different colours from a bag, replacing the extracted ball after each draw. Balls from the same colour are equivalent.

The probability mass function can be expressed using the gamma function as:

$$f(x_1,\ldots,x_k;p_1,\ldots,p_k) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^k p_i^{x_i}.$$

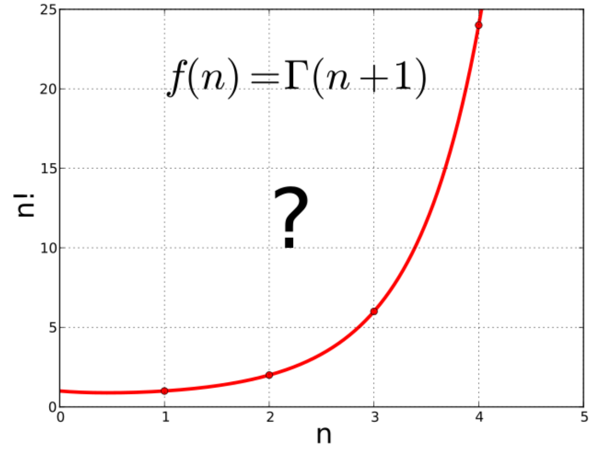
Where gamma function is an extension of factorial:

$$\Gamma(n) = (n-1)!$$

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$$

More on Gamma Function

It is easy graphically to interpolate the factorial function to non-integer values, but is there a formula that describes the resulting curve?



$$\begin{aligned}\Gamma(z+1) &= \int_0^\infty x^z e^{-x}\,dx \\ &= [-x^z e^{-x}]_0^\infty + \int_0^\infty z x^{z-1} e^{-x}\,dx \\ &= \lim_{x\to\infty}(-x^z e^{-x}) - (0e^{-0}) + z\int_0^\infty x^{z-1}e^{-x}\,dx\end{aligned}$$

$$\begin{aligned}\Gamma(1) &= \int_0^\infty x^{1-1} e^{-x}\,dx = [-e^{-x}]_0^\infty \\ &= \lim_{x\to\infty}(-e^{-x}) - (-e^{-0}) = 0-(-1) = 1\end{aligned}$$

Recognizing that as $x \to \infty, -x^z e^{-x} \to 0,$

$$\Gamma(z+1) = z\int_0^\infty x^{z-1} e^{-x}\,dx = z\Gamma(z)$$

Given that $\Gamma(1) = 1$ and $\Gamma(n+1) = n\Gamma(n)$

$$\Gamma(n) = 1\cdot 2\cdot 3\cdots(n-1) = (n-1)!$$

Gamma Distribution

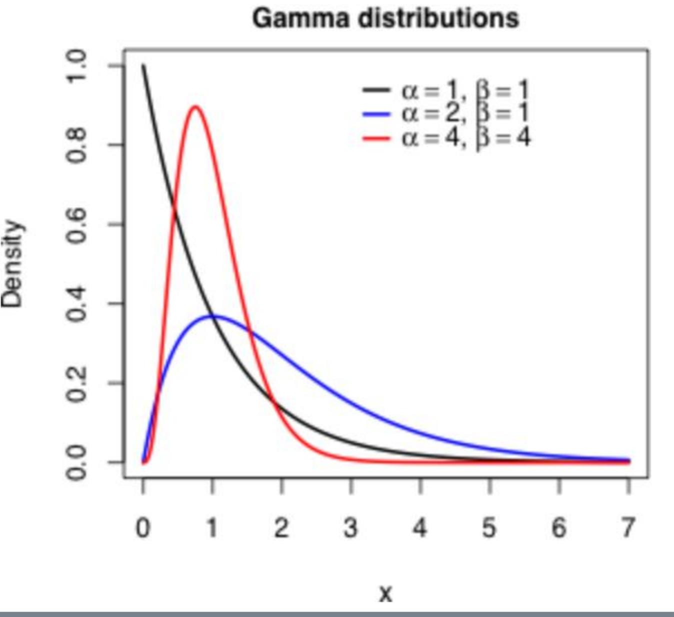
Like the lognormal the gamma distribution is unbounded on the right, defined for only positive *X*, and tends to yield skewed distributions.

$$X \sim \Gamma(\alpha, \beta) \equiv \mathrm{Gamma}(\alpha, \beta)$$

The corresponding **probability density function** in the shape-rate parametrization is

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad \text{for } x > 0 \text{ and } \alpha, \beta > 0,$$

where $\Gamma(\alpha)$ is a complete gamma function.



The gamma distribution is widely used as a conjugate prior in Bayesian statistics. It is the conjugate prior for the precision of a normal distribution. It is also the conjugate prior for the exponential distribution.

Beta Distribution

It is bounded on both sides. In this respect it resembles the binomial distribution. The standard beta distribution is constrained so that its domain is the interval (0, 1).

$$f(x;\alpha,\beta) = \text{constant}\cdot x^{\alpha-1}(1-x)^{\beta-1}$$

$$= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}$$

$$= \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

The beta function, **B** is a normalization constant to ensure that the total probability integrates to 1.

Poisson Distribution

A discrete random variable X is said to have a Poisson distribution with parameter $\lambda > 0$, if, for k = 0, 1, 2, ...,

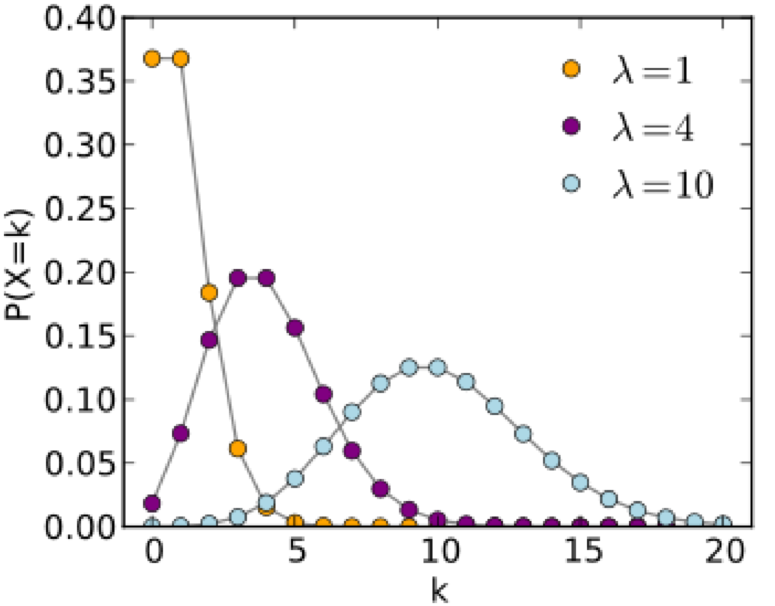
For example, on a particular river, overflow floods occur once every 100 years on average. Calculate the probability of k = 0, 1, 2, 3, 4, 5, or 6 overflow floods in a 100-year interval, assuming the Poisson model is appropriate. Because the average event rate is one overflow flood per 100 years, $\lambda = 1$, so that:

$$P(k \text{ overflow floods in 100 years}) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{1^k e^{-1}}{k!}$$

$$P(k = 0 \text{ overflow floods in 100 years}) = \frac{1^0 e^{-1}}{0!} = \frac{e^{-1}}{1} = 0.368$$

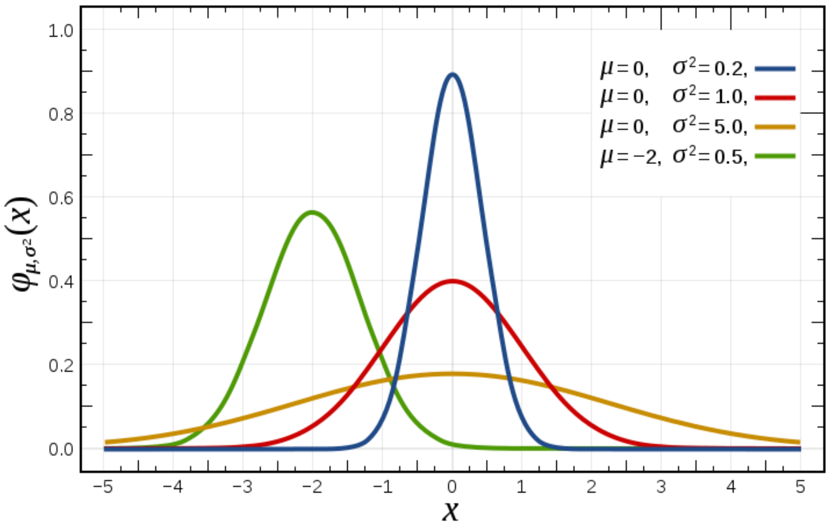
$$P(k = 1 \text{ overflow flood in 100 years}) = \frac{1^1 e^{-1}}{1!} = \frac{e^{-1}}{1} = 0.368$$

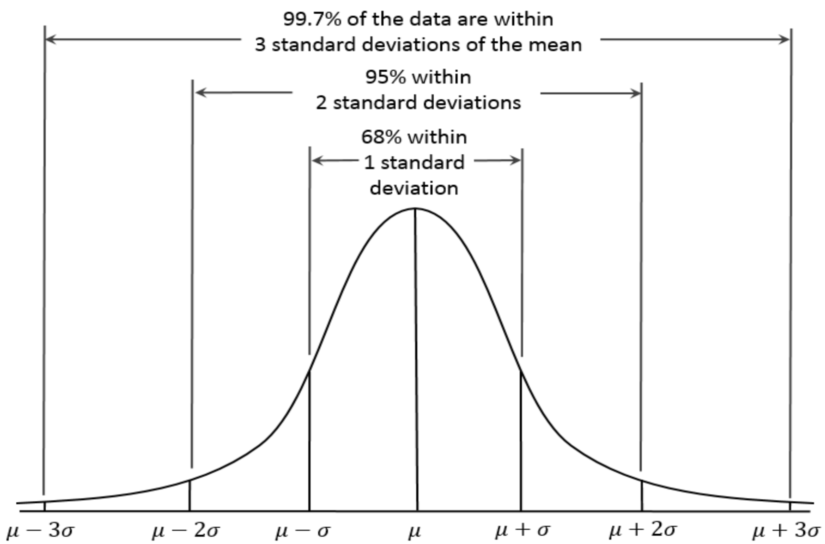
$$P(k = 2 \text{ overflow floods in 100 years}) = \frac{1^2 e^{-1}}{2!} = \frac{e^{-1}}{2} = 0.184$$



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

Gaussian (Normal) Distribution





The probability density of the normal distribution is:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

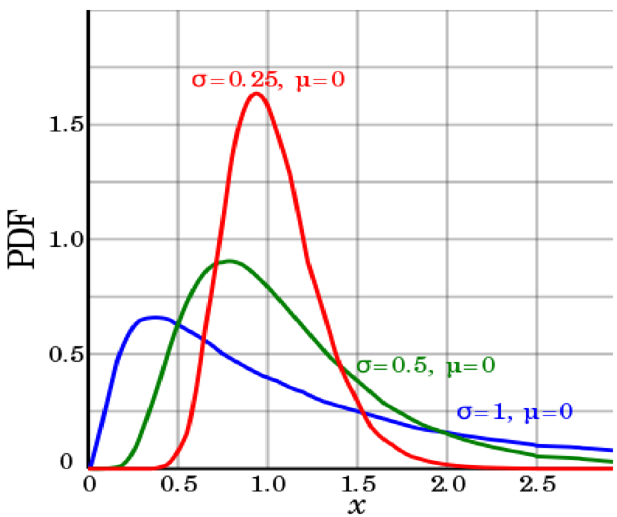
- μ is the mean or expectation of the distribution
- σ is the standard deviation
- σ^2 is the variance

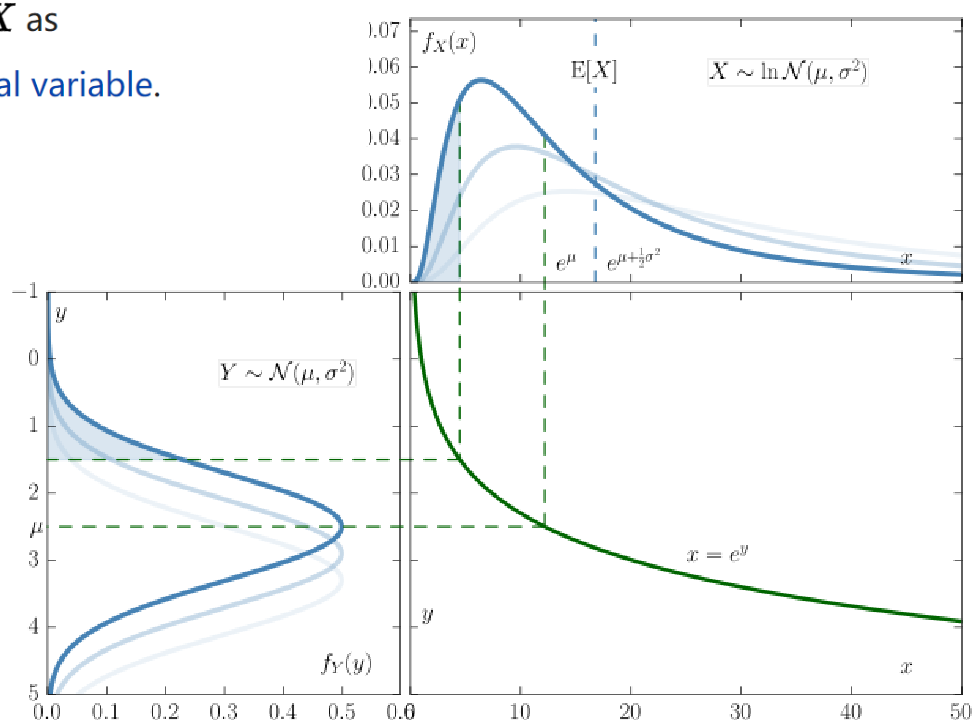
Log Normal Distribution

Given a log-normally distributed random variable X and two parameters μ and σ that are, respectively, the mean and standard deviation of the variable's natural logarithm, then the logarithm of X is normally distributed, and we can write X as

$$\boxed{X = e^{\mu+\sigma Z}}$$

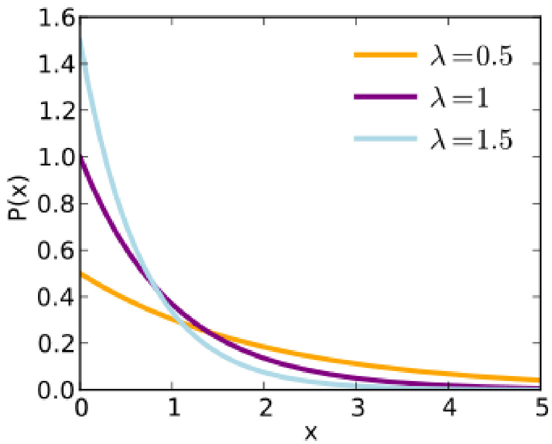
with Z a standard normal variable.





Exponential Distribution

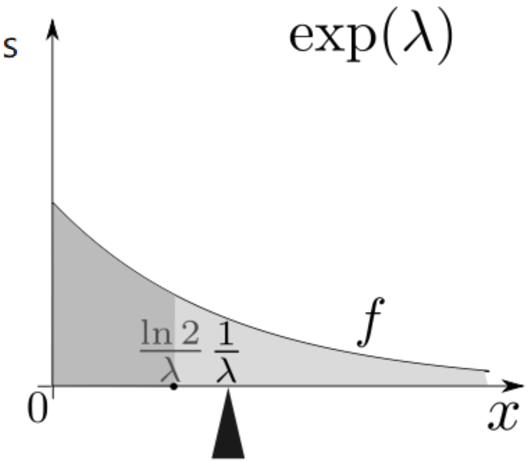
The exponential distribution (also known as negative exponential distribution) is the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. It is a particular case of the gamma distribution. It is the continuous analogue of the geometric distribution, and it has the key property of being memoryless.



The probability density function (pdf) of an exponential distribution is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0, \\ 0 & x < 0. \end{cases} \qquad \mathrm{E}[X] = \frac{1}{\lambda} \qquad \mathrm{Var}[X] = \frac{1}{\lambda^2}$$

The exponential distribution occurs naturally when describing the lengths of the inter-arrival times in a homogeneous Poisson process.



Bayesian Examples

Marie is getting married tomorrow at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman is forecasting rain for tomorrow. When it actually rains, the weatherman has forecast rain 90% of the time. When it doesn't rain, he has forecast rain 10% of the time. What is the probability it will rain on the day of Marie probability it will rain on the day of Marie s' wedding?

Event A: The weatherman has forecast rain.
Event B: It rains.

We know:

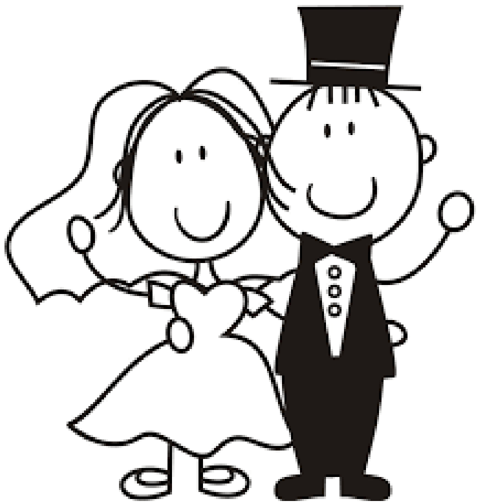
1. p(B) = 5 / 365 = 0.0137 [It rains 5 days out of the year.]
2. p (not B) = 360 / 365 = 0.9863
3. p(A | B) = 0.9 [When it rains, the weatherman has forecast rain 90% of the time.]
4. p (A | not B) = 0.1 [When it does not rain weatherman has forecast rain 10% of the time.]

Bayesian Example

We want to know p(B | A), the probability it will rain on the day of Marie's wedding, given a forecast for rain by the weatherman. The answer can be determine d from Bayes Rule:

1. $p (B | A) = p (A | B) \cdot p (B) / p (A)$
2. $p(A) = p(A | B) \cdot p(B) + p(A | \text{not } B) \cdot p(\text{not } B)$
   $= (0.9)(0.014) + (0.1)(0.986) = 0.111$
1. $p(B | A) = (0.9)(0.0137) / 0.111 = 0.111$

The special coin problem.



Simpson’s Paradox

Simpson's paradox, or the Yule–Simpson effect, is a phenomenon in probability and statistics, in which a trend appears in different groups of data but disappears or reverses when these groups are combined.

Department	Female Applicants	Female Admitted	%	Male Applicants	Male Admitted	%	All Applicants	All Admitted	Overall %
Business School	100	49	49%	20	15	75%	**120**	**64**	**53.3%**
Law School	20	1	5%	100	10	10%	**120**	**11**	**9.2%**
Both	120	50	42%	120	25	21%	**240**	**75**	**31.3%**

Suppose two people, Lisa and Bart, each edit articles for two weeks. In the first week, Lisa fails to improve the only article she edited, and Bart improves 1 of the 4 articles he edited. In the second week, Lisa improves 3 of 4 articles she edited, while Bart improves the only article he edited.

	Week 1	Week 2	Total
Lisa	0/1	3/4	**3/5**
Bart	**1/4**	**1/1**	2/5