40.017 PROBABILITY & STATISTICS

Lecture Notes

Michael Hoon

February 22, 2024

Contents

1 Set Theory

1.1 Sample Spaces

The mathematical framework for probability is built around *sets*. The *sample space* S of an experiment is the set of all possible outcomes of the experiment. An *event* A is a subset of S, and we say that A occurred if the actual outcome is in A.

1.2 Naive Definition of Probability

Let A be an event for an experiment with a finite sample space S. A naive probability of A is

$$\mathbb{P}_{\text{naive}}(A) = \frac{|A|}{|S|} = \frac{\text{number of outcomes favorable to A}}{\text{total number of outcomes}} \tag{1}$$

In general, the result about complements always holds:

$$\mathbb{P}_{\text{naive}}(A^c) = \frac{|A^c|}{|S|} = \frac{|S| - |A|}{|S|} = 1 - \frac{|A|}{|S|} = 1 - \mathbb{P}_{\text{naive}}(A)$$

An important factor about the naive definition is that it is restrictive in requiring S to be finite.

1.3 General Definition of Probability

Definition 1.1. A probability space consists of a sample space S and a probability function P which takes an event $A \subseteq S$ as input and returns $P(A)$, where $P(A) \in \mathbb{R},\ P(A) \in [0, 1]$. The function must satisfy the following axioms:

1. $\mathbb{P}(\emptyset) = 1,\ \mathbb{P}(S) = 1$
2. $\mathbb{P}(A) \geq 0$
3. If $A_1, A_2, \ldots$ are **disjoint events**, then:

$$\mathbb{P}\left(\bigcup_{j=1}^{\infty}\right) = \sum_{j=1}^{\infty} \mathbb{P}(A_j)$$

Disjoint events are **mutually exclusive** (i.e. $A_i \cap A_j = \emptyset\ \forall\ i \neq j$).

1.3.1 Properties of Probability

Theorem 1.2. Probability has the following properties, for any events A and B:

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
2. If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

1.3.2 Inclusion-Exclusion Principle

For any events $A_1, \ldots A_n$,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1}\mathbb{P}(A_1 \cap \cdots \cap A_n) \tag{2}$$

For $n = 2$, we have a nicer result:

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$$

1.4 Conditional Probability

Definition 1.3. If A and B are events with $\mathbb{P}(B) > 0$, then the *conditional probability* of A given B, denoted by $\mathbb{P}(A \mid B)$ is defined as:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Here A is the event whose uncertainty we want to update, and B is the evidence we observe. $\mathbb{P}(A)$ is the *prior* probability of A and $\mathbb{P}(A|B)$ is the *posterior* probability of A. (For any event A, $\mathbb{P}(A|A) = \frac{\mathbb{P}(A \cap A)}{\mathbb{P}(A)}$).

2 Derangement

A derangement is a permutation of the elements of a set in which no element appears in its original position. We use D_n to denote the number of derangements of n distinct objects.

2.1 Counting Derangements

We consider the number of ways in which n hats $(h_1, \ldots, h_n)$ can be returned to n people $(P_1, \ldots, P_n)$ such that no hat makes it back to its owner.

We obtain the recursive formula:

$$D_n = (n-1)(D_{n-1} + D_{n-2}),\ \forall\, n \geq 2 \tag{3}$$

With the initial conditions $D_1 = 0$ and $D_2 = 1$, we can use the formula to recursively compute D_n for any n.

There are various other expressions for D_n, equivalent to formula 3:

$$D_n = n! \sum_{i=0}^{n} \frac{(-1)^i}{i!},\ \forall\, n \geq 0 \tag{4}$$

2.1.1 Limiting Growth

From Equation 4, and the taylor series expansion for e:

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!} \tag{5}$$

we substitute $x = -1$ and obtain the limiting value as $n \to \infty$:

$$\lim_{n\to\infty} \frac{D_n}{n!} = \lim_{n\to\infty} \sum_{i=0}^{n} \frac{(-1)^i}{i!} = e^{-1} \approx 0.367879\ldots$$

This is the limit of the probability that a randomly selected permutation of a large number of objects is a derangement. The probability converges to this limit extremely quickly as n increases, which is why D_n is the nearest integer to $\frac{n!}{e}$.

3 Discrete Random Variables

We formally define a random variable:

Definition 3.1. Given an experiment with sample space S, a *random variable* (r.v.) is a function from the sample space S to the real numbers $\mathbb{R}$. It is common to denote random variables by capital letters.

Thus, a random variable X assigns a numerical value $X(s)$ to each possible outcome s of the experiment. The randomness comes from the fact that we have a random experiment (with Probabilities described by the probability function P); the mapping itself is deterministic.

There are two main types of random variables used in practice: *discrete* and *continuous* r.v.s.

Definition 3.2. A random variable X is said to be *discrete* if there is a finite list of values $a_1, a_2, \dots, a_n$ or an infinite list of values $a_1, a_2, \dots$ such that $\mathbb{P}(X = a_j \text{ for some } j) = 1$. If X is a discrete r.v., then the finite or countably infinite set of values x such that $P(X = x) > 0$ is called the *support* of X.

3.1 Binomial

3.2 Hypergeometric

If we have an urn filled with w white and b black balls, then drawing n balls out of the urn *with replacement* yields a Binom$(n, \frac{w}{(w+b)})$. If we instead sample *without replacement*, then the number of white balls follow a **Hypergeometric** distribution.

Theorem 3.3. If $X \sim \text{hypgeo}(n, j, k)$, then the PMF of X is:

$$\mathbb{P}(X = x) = \frac{\binom{j}{x}\binom{k}{n-x}}{\binom{j+k}{n}}$$

$\forall x \in \mathbb{Z}$ satisfying $0 \leq x \leq n$ and $0 \leq n - x \leq j$, and $P(X = x) = 0$ otherwise.

If j and k are large compared to n, then selection without replacement can be approximated by selection with replacement. In that case, the hypergeometric RV $X \sim \text{hypgeo}(n, j, k)$ can be approximated by a binomial RV $Y \sim \text{binomial}(n, p)$, where $p := \frac{j}{j+k}$ is the probability of selecting a black marble.

We can also write X as the sum of (dependent) Bernoulli random variables:

$$X = X_1 + X_2 + \cdots + X_n$$

where each X_i equals 1 if the ith selected marble is black, and 0 otherwise.

3.2.1 Hypergeometric Symmetry

Theorem 3.4. The hypergeo(w, b, n) and hypergeo$(n, w + b - n, w)$ distributions are identical.

The proof follows from swapping the two sets of tags in the Hypergeometric story (white/black balls in urn) [3].

[3] The binomial and hypergeometric distributions are often confused. Note that in Binomial distributions, the Bernoulli trials are **independent**. The Bernoulli trials in Hypergeometric distribution are **dependent**, since the sampling is done *without replacement*.

3.3 Geometric

3.4 Negative Binomial

In a sequence of independent Bernoulli trials with success probability p, if X is the number of failures before the rth success, then X is said to have the Negative Binomial distribution with parameters r and p, denoted $X \sim \text{NBin}(r, p)$.

Both the Binomial and Negative Binomial distributions are based on independent Bernoulli trials; they differ in the *stopping rule* and in what they are counting. The Negative Binomial counts the **number of failures until a fixed number of successes**.

Theorem 3.5. If $X \sim \text{NBin}(r, p)$, then the PMF of X is

$$P(X = x) = \binom{x-1}{n-1}(1-p)^{x-n}p^n, \ \forall \ x \geq n \tag{6}$$

4 Law of Large numbers

Assume that we have i.i.d. $X_1, X_2, \ldots$ with finite mean μ and finite variance σ^2. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Definition 4.1. The (Weak) Law of Large Numbers (LLN) says that as n grows, the sample mean $\bar{X}_n$ converges to the true mean μ. Mathematically,

$$\forall \epsilon > 0, \ \mathbb{P}(|\bar{X}_n - \mu < \epsilon) = 1, \ \text{as } n \to \infty \tag{7}$$

For any positive margin ϵ, as n gets arbitrarily large, the probability that $\bar{X}_n$ is within ϵ of μ approaches 1.

Note that the LLN does not contradict the fact that a coin is memoryless (in the repeated coin toss experiment). The LLN states that the proportion of Heads converges to $\frac{1}{2}$, but this does not imply that after a long string of Heads, the coin is "due" for a Tails to "*balance things out*". Rather, the convergence takes place through *swamping*: past tosses are swamped by the infinitely many tosses that are yet to come.

4.1 Inequalities

The inequalities in this section provide bounds on the probability of an r.v. taking on an 'extreme' value in the right or left rail of a distribution.

4.1.1 Markov's Inequality

Definition 4.2. Let X be any random variable that takes only non-negative values, that is, $\mathbb{P}(X < 0) = 0$. Then for any constant $a > 0$, we have:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a} \tag{8}$$

For an intuitive interpretation, let X be the income of a randomly selected individual from a population. Taking $a = \mathbb{E}(X)$, Markov's Inequality says that $\mathbb{P}(\mathbb{X} \geq \nvDash\mathbb{E}(\mathbb{X})) \leq \frac{\nVdash}{\nvDash}$. i.e., it is impossible for more than half the population to make at least twice the average income.

4.1.2 Chebyshev's Inequality

Gives general bounds for the probability of being k standard deviations (SD) away from the mean.

Definition 4.3. Let Y be any random variable with mean $\mu < \infty$ and variance $\sigma^2 > 0$. Then for any constant $k > 0$, we have:

$$\mathbb{P}(|Y - \mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{9}$$

5 Central Limit Theorem

Let $X_1, X_2, \dots$ be i.i.d. with mean μ and variance σ^2.

Definition 5.1. The CLT states that for large n, the distribution of $\bar{X}_n$ after standardisation approaches a standard Normal distribution. By standardisation, we mean that we subtract μ, the mean of $\bar{X}_n$, and divide by $\frac{\sigma}{\sqrt{n}}$, the standard deviation of $\bar{X}_n$.

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x\right) = \Phi(x) \tag{10}$$

which is the cdf of the standard normal. Informally, when n is large (≥ 30), then $\bar{X}_n$ and $\sum_{i=1}^n X_i$ can each be approximated by a normal RV with the same mean and variance; the actual distribution of X_i becomes irrelevant:

$$\bar{X}_n \approx N(\mu, \frac{\sigma^2}{n}), \qquad \sum_{i=1}^n X_i \approx N(n\mu, n\sigma^2)$$

6 Moments

6.1 Interpreting Moments

Definition 6.1. Let X be an r.v. with mean μ and variance σ^2. For any positive integer n, the n^{th} moment of X is $\mathbb{E}(X^n)$, the n^{th} central moment is $\mathbb{E}((X - \mu)^n)$.

In particular, the mean is the first moment and the variance is the second central moment.

6.2 Moment Generating Functions

A moment generating function, as its name suggests, is a generating function that encodes the **moments** of a distribution. Starting with an infinite sequence $(a_0, a_1, a_2, \dots)$, we 'condense' or 'store' it as a single function g, the generating function of the sequence:

$$\sum_{n=0}^{\infty} a_n \frac{t^n}{n!} := g(t)$$

Definition 6.2. When we take $a_n = \mathbb{E}(X^n)$, the resulting generating function is known as the **moment generating function (MGF)** of X, and is denoted by $M_X(t)$.

The MGF of X can be computed as an expected value:

Note that $M_X(0) = 1$ for any valid MGF.

6.3 Formulas & Theorems

Some important formulas for the MGF of X:

$$\boxed{M_X(t) = \mathbb{E}(e^{tX})} \tag{11}$$

where if X is **discrete** with pmf f, then

$$M_X(t) = \sum_{all x_i} e^{tx_i} f(x_i) \tag{12}$$

and if X is **continuous** with pdf f, then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\,\mathrm{d}x \tag{13}$$

Theorem 6.3. Given the MGF of X, we can get the n^{th} moment of X by evaluating the n^{th} derivative of the MGF at 0:

$$\boxed{\mathbb{E}(X^n) = M_X^{(N)}(0)} \tag{14}$$

Theorem 6.4. If X and Y are independent, then the MGF of $X+Y$ is the product of the individual MGFs:

$$M_{X+Y}(t) = M_X(t) M_Y(t) \tag{15}$$

This is true because if X and Y are independent, then $\mathbb{E}(e^{t(X+Y)}) = \mathbb{E}(e^{tX})\mathbb{E}(e^{tY})$

Theorem 6.5. If two random variables have the same MGF, then they have the same distribution (same cdf, equivalently, same pdf or pmf) [a].

[a]For this to apply, the MGF needs to exist in an open interval around $t = 0$

6.4 Examples (Discrete)

6.4.1 Binomial MGF

We have $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$. The MGF can be found by:

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{n} \binom{n}{x} (\underbrace{e^t p}_{a})^x (\underbrace{1-p}_{b})^{n-x} \\ &= (e^t p + 1 - p)^n \end{aligned} \tag{16}$$

by using the fact that

$$\sum_x \binom{n}{x} a^x b^{n-x} = (a+b)^n$$

and from which we can obtain $\mathbb{E}(X) = M_X'(0) = n\,\overbrace{(e^t p + 1 - p)^{n-1} \cdot e^t p}^{p} |_{t=0} = np$

6.4.2 Poisson MGF

For a Poisson r.v., where $X \sim \text{Poisson}(\lambda)$ We have $f(x) = e^{-\lambda}\frac{\lambda^x}{x!}$. Then,

$$
\begin{aligned}
M_X(t) &= \sum_{x=0}^{\infty} e^{tx} e^{-\lambda} \frac{\lambda^x}{x!} \qquad (17) \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t \lambda)^x}{x!} \\
&= e^{-\lambda} e^{e^t \lambda} \\
&= e^{e^t \lambda - \lambda} \\
&= e^{\lambda(e^t - 1)}
\end{aligned}
$$

We can now find:

$$M'_X(t) = e^{\lambda(e^t-1)}(\lambda e^t)$$

and therefore

$$M'_X(0) = e^0(\lambda e^0) = \lambda$$

6.5 Examples (Continuous)

6.5.1 Standard Normal

If $Z \sim \mathcal{N}(0,1)$ is a standard normal r.v., then $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. For continuous distributions, we need to use the infinite integral:

$$
\begin{aligned}
M_Z(t) = \mathbb{E}(e^{tZ}) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{x^2}{2}} \, \mathrm{d}x \qquad (18) \\
&= \qquad (19)
\end{aligned}
$$

7 Gamma Distribution

The Gamma distribution is a **continuous distribution** on the positive real line, and is a generalisation of the Exponential distribution. While an Exponential r.v. represents the waiting time for the first success under conditions of **memorylessness**, the Gamma r.v. represents the total waiting time for *multiple successes*.

7.1 Gamma Function

Definition 7.1. The *gamma function* Γ is defined by

$$\Gamma(a) = \int_0^{\infty} x^a e^{-x} \frac{\mathrm{dx}}{x} \qquad (20)$$

for real numbers $a > 0$. It is possible to write the integrand as $x^{a-1}e^{-x}$, but it is left for convenience when we make the transformation $u = cx$, so that we have $\frac{du}{u} = \frac{dx}{x}$.

Some properties of the gamma function include:

1. $\Gamma(a+1) = a\Gamma(a) \ \forall a > 0$. This follows from integration by parts:

2. $\Gamma(n) = (n-1)!$ if n is a positive integer. Can be proved via induction, starting with $n = 1$ and using the recursive relation $\Gamma(a+1) = a\Gamma$.

7.2 Gamma Distribution

Definition 7.2. An r.v. Y is said to have the *Gamma distribution* with parameters α and λ, where $a > 0$ and $\lambda > 0$, if its PDF is:

$$f(x) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases} \tag{21}$$

We denote this by $X \sim \mathit{gamma}(\alpha, \lambda)$. We have:

$$\mathbb{E}(X) = \frac{\alpha}{\lambda}, \qquad \mathrm{Var}(X) = \frac{\alpha}{\lambda^2}$$

Taking $\alpha = 1$, the gamma$(1, \lambda)$ PDF is $f(x) = \lambda e^{-\lambda x}$, so gamma$(1, \lambda)$ and exp(λ) are the same. The extra parameter a allows Gamma PDFs to have a greater variety of shapes, refer to Figure **??** below.

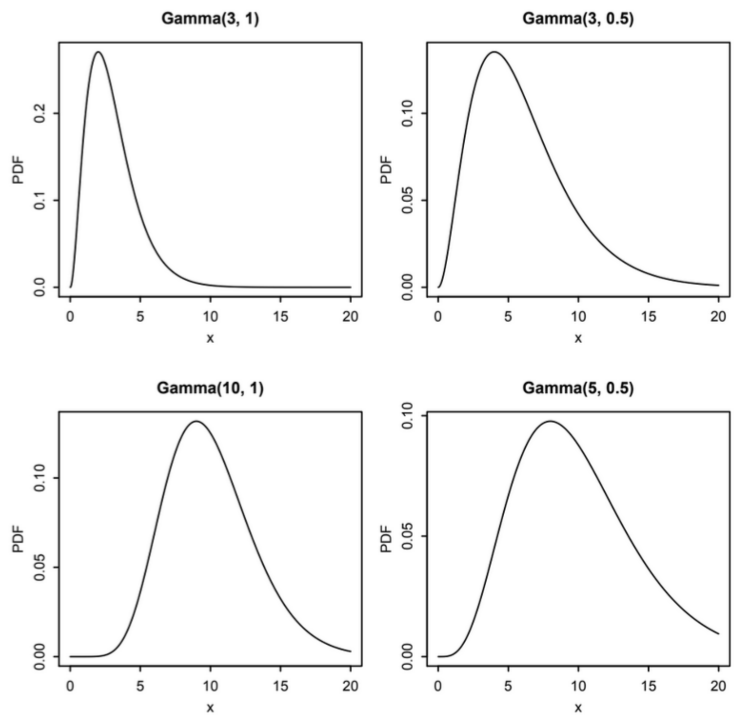

Figure 1: Gamma PDFs for various values of a and λ.

For small values of a, the PDF is skewed, but as a increases, the PDF starts to look more symmetrical and bell-shaped (following the LLN). The mean and variance are increasing in a and decreasing in λ.

7.3 MGF of Gamma Distribution

In week 3 lecture 1, we proved the MGF of the Gamma distribution:

$$\begin{aligned} M_X(t) &= \mathbb{E}(e^{tX}) \\ &= \dots \\ &= \frac{\lambda^\alpha (\lambda - t)^{-\alpha})}{\Gamma(a)} \underbrace{\int_0^\infty y^{\alpha-1} e^{-y}\, \mathrm{d}y}_{\Gamma(\alpha)} \\ &= \lambda^\alpha (\lambda - t)^{-\alpha} \end{aligned}$$

In the special case where a is an integer, we can represent a gamma(α, λ) r.v. as a sum (convolution) of i.i.d. $\exp(\lambda)$ r.v.s.

Theorem 7.3. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\exp(\lambda)$. Then

$$X_1 + X_2 + \cdots + X_n \sim \text{gamma}(, n\lambda)$$

Since $\alpha = n \in \mathbb{Z}^+$, then $\lambda^\alpha(\lambda - t)^{-\alpha} = \left(\frac{\lambda}{(\lambda - t)}\right)^n$.

Theorem 7.3 also allows us to connect the Gamma distribution to the story of the Poisson process. In Poisson processes of rate λ, the interarrival times are i.i.d. $\exp(\lambda)$ r.v.s but the total waiting time T_n for the nth arrival is the sum of the first n interarrival times, as shown in Figure 2 below. T_3 is the sum of the 3 interarrival times X_1, X_2, X_3. Therefore by the theorem, $T_n \sim \text{gamma}(n, \lambda)$. The interarrival times in a Poisson process are Exponential r.v.s, while the raw arrival times are Gamma r.v.s.

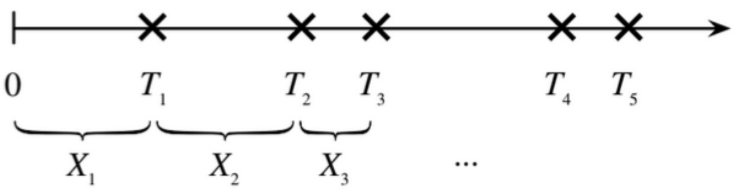

Figure 2: Poisson process, interarrival times X_j are i.i.d. $\exp(\lambda)$, while raw arrival times T_j are gamma(j, λ).

Note that unlike the X_j's, the T_j's are **not independent**, since they are constrained to be increasing, nor are they i.i.d. Now, we have an interpretation for the parameters of the gamma(α, λ) distribution. In the Poisson process story, α is the *number of successes* we are waiting for, and λ is the rate at which successes arrive. $Y \sim \text{gamma}(\alpha, \lambda)$ is the total waiting time for the ath arrival in a Poisson process of rate λ.

8 Conditionals

8.1 Bayes' Theorem Recap

Theorem 8.1. Recall Bayes' Theorem, which provides a link between $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \tag{22}$$

where $\mathbb{P}(B)$ is often computed from the **law of total probability**; for instance, when conditioned on A and A^c:

$$\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)$$

8.2 Law of Total Probability

Definition 8.2. Let $A_1, \dots, A_n$ be a partition of the sample space S (i.e. the A_i are disjoint events and their union is S), with $\mathbb{P}(A_i) > 0$, $\forall i$. Then:

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i) \tag{23}$$

The law of total probability tells us that to get the unconditional probability of B, we can divide the sample space into disjoint slices A_i, find the conditional probability of B within each of the slices, then take a weighted sum of the conditional probabilities, where the weights are probabilities $\mathbb{P}(A_i)$.

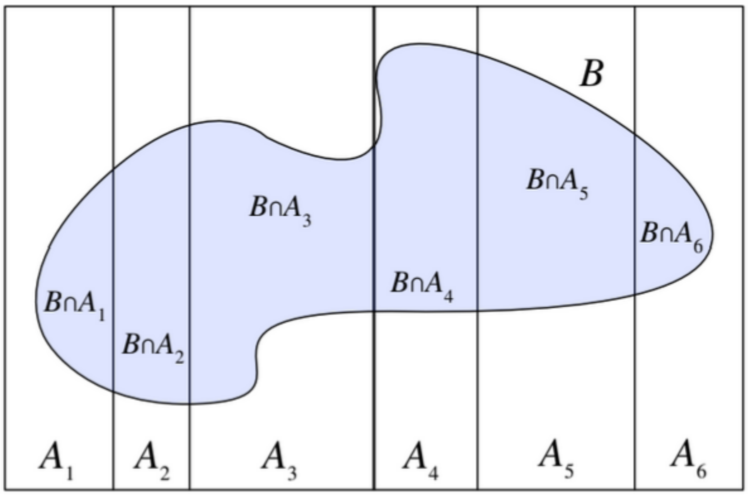

Figure 3: The A_i partition the sample space, $\mathbb{P}(B)$ is equal to $\sum_i \mathbb{P}(B \cap A_i)$

8.3 Independence of Events

The situation where events provide no information about each other is called independence.

Definition 8.3. Events A and B are *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

If $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, then this is equivalent to

$$\mathbb{P}(A|B) = \mathbb{P}(A), \qquad \mathbb{P}(B|A) = \mathbb{P}(B)$$

Note that independence [1] is a *symmetric relation*: if A is independent of B, then B is independent of A.

[1]Independence is completely different from *disjointness*. If A and B are disjoint, then $\mathbb{P}(A \cap B) = 0$, so disjoint events can be independent only if $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$. KNowing that A occurs tells us that B definitely did not occur, so A clearly conveys information about B, meaning the two events are not independent.

8.4 Conditional Independence

Definition 8.4. Events A and B are said to be *conditionally independent* given E if:

$$\mathbb{P}(A \cap B|E) = \mathbb{P}(A|E)\mathbb{P}(B|E) \tag{24}$$

In particular,

1. Two events can be conditionally independent given E, but not given E^c.
2. Two events can be conditionally independent given E, but not independent.
3. Two events can be independent, but not conditionally independent given E.

Equivalently, we have the result

$$\mathbb{P}(B|A \cap E) = \mathbb{P}(B|E) \tag{25}$$

In particular, $\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B)$ **does not** imply $\mathbb{P}(A, B|E) = \mathbb{P}(A|E)\mathbb{P}(B|E)$

8.5 Properties of the Conditional

Conditional probability satisfies all the properties of probability. In particular:

1. Conditional probabilities are between 0 and 1
2. $\mathbb{P}(S|E) = 1$, $\mathbb{P}(\emptyset|E) = 0$
3. If $A_1, A_2, \ldots$ are disjoint, then $\mathbb{P}(\bigcup_{j=1}^{\infty} A_j|E) = \sum_{j=1}^{\infty} \mathbb{P}(A_j|E)$
4. $\mathbb{P}(A^c|E) = 1 - \mathbb{P}(A|E)$
5. **Inclusion Exclusion**: $\mathbb{P}(A \cap B|E) = \mathbb{P}(A|E) + \mathbb{P}(B|E) - \mathbb{P}(A \cup B|E)$.

8.6 Discrete: Conditional P.M.F

8.6.1 Joint CDF

The most general description of the joint distribution of two r.v.s is the joint CDF.

Definition 8.5. The joint CDF of r.v.s X and Y is the function $F_{X,Y}$ given by

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y) \tag{26}$$

8.6.2 Joint PMF

Definition 8.6. The joint PMF of discrete r.v.s X and Y is the function

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) \tag{27}$$

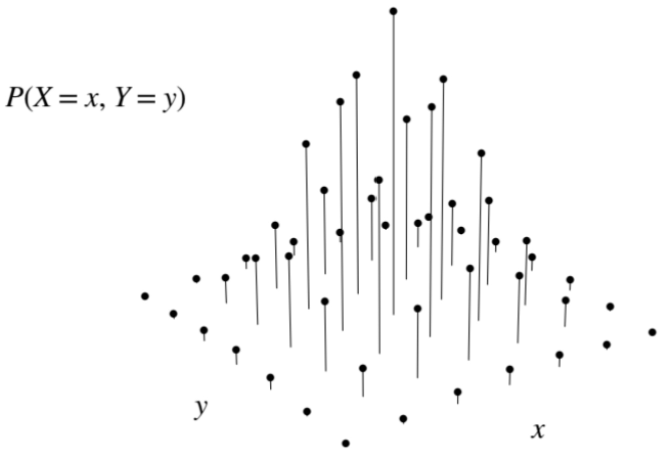

Figure 4: Joint PMF of discrete r.v.s X and Y

8.6.3 Marginal PMF

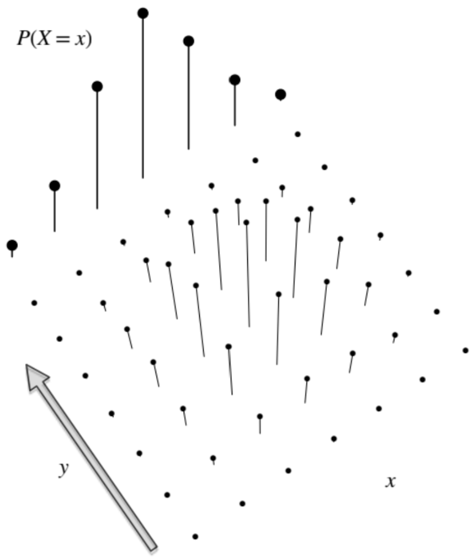

Figure 5: The marginal PMF $\mathbb{P}(X = x)$ is obtained by summing over the joint PMF in the y-direction.

8.6.4 Conditional PMF

Definition 8.7. For discrete r.v.s X and Y, if $f_Y(y) > 0$, then the **conditional pmf** of X given $Y = y$ is

$$f_{X|Y}(x|y) := \mathbb{P}((X = x)|(Y = y)) = \frac{f(x,y)}{f_Y(y)} \tag{28}$$

This is viewed as a function of y for fixed x. Think of $f_{X|Y}(x|y)$ as a function of x, when Y is fixed at y. It must be the case that

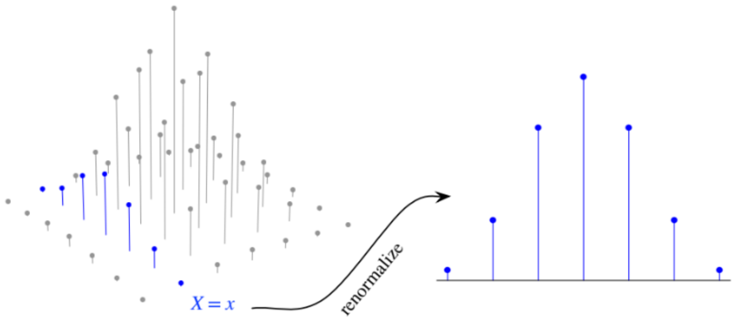

Figure 6: Conditional pmf of Y given $X = x$. The conditional pmf $\mathbb{P}(Y = y|X = x)$ is obtained by renormalising the column of the joint pmf that is compatible with the event $X = x$.

Figure illustrates the
We can also relate the conditional distribution to Bayes' theorem, which takes the form

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

and if X and Y are independent, then $\forall x, y$ with $f_Y(y) > 0$, we have

$$f_{X|Y}(x|y) = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x)$$

8.7 Continuous: Conditional P.M.F

8.7.1 Conditional PDF

Definition 8.8. For continuous r.v.s X and Y, if $f_Y(y) > 0$, then the **conditional pdf** of X given Y given $Y = y$ is

$$f_{X|Y}(x|y) := \frac{f(x, y)}{f_Y(y)} \tag{29}$$

9 Conditional Expectation

9.1 Discrete

The expectation $\mathbb{E}(Y)$ of a discrete r.v. Y is a weighted average of its possible values, where the weights are the PMF values $\mathbb{P}(Y = y)$. After learning that an event A occurred, we want to use weights that have been updated to reflect this new information.

Definition 9.1. For *discrete* r.v.s X and Y, the **conditional expectation** of X given $Y = y$ is

$$\mathbb{E}(X|Y = y) := \sum_{\text{all } x} x\mathbb{P}((X = x)|(Y = y)) = \sum_{\text{all } x} f_{X|Y}(x|y) \tag{30}$$

That is, $\mathbb{E}(X|Y = y)$ is the expectation of X given $Y = y$.

9.1.1 Law of Total Expectation (Discrete)

Definition 9.2. We have the law of total expectation:

$$\mathbb{E}(X) = \sum_{\text{all } y} \mathbb{E}(X|Y=y)\mathbb{P}(Y=y) \tag{31}$$

9.2 Continuous

Definition 9.3. For *continuous* r.v.s X and Y, the **conditional expectation** of X given $Y=y$:

$$\mathbb{E}(X|Y=y) := \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,\mathrm{d}x = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} x f(x,y)\,\mathrm{d}x \tag{32}$$

9.2.1 Law of Total Expectation (Continuous)

Definition 9.4. Similarly, we have the law of total expectation:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} \mathbb{E}(X|Y=y) f_Y(y)\,\mathrm{d}y \tag{33}$$

9.3 An Alternate Formulation

Notice that because we sum or integrate over x, $\mathbb{E}(X|Y=y)$ is a function of y only. Then, we let this be $g(y)$, so $\mathbb{E}(X|Y=y) = g(y)$. Then, the law of total expectation says:

$$\mathbb{E}(X) = \sum_{\text{all } y} g(y)\mathbb{P}(Y=y) = \mathbb{E}(g(Y)) \tag{34}$$

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} g(y) f_Y(y)\,\mathrm{d}y = \mathbb{E}(g(Y)) \tag{35}$$

Definition 9.5. Let $g(x) = \mathbb{E}(X|Y=y)$. Then the conditional expectation of X given Y, denoted $\mathbb{E}(X|Y)$ is defined to be the random variable $g(Y)$. Then the law of total expectation can be written:

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)) \tag{36}$$

In other words, if after doing the experiment X crystallises into x, then $\mathbb{E}(X|Y)$ crystallises into $g(y)$.

9.4 Conditional Expectation Given An Event

For any event A, we adapt the **law of total expectation** to compute $\mathbb{P}(A)$. We first define a random variable X, where $X=1$ if A occurs, and $X=0$ otherwise. Then, $\mathbb{E}(X) = \mathbb{P}(A)$, $\mathbb{E}(X|Y=y) = \mathbb{P}(A|Y=y)$.

Theorem 9.6. We apply the law to $\mathbb{E}(X)$ to obtain:

$$\mathbb{P}(A) = \int_{-\infty}^{\infty} \mathbb{P}(A|Y=y) f_Y(y)\,\mathrm{d}y \tag{37}$$

which is sort of like the continuous version of the law of total probability.

9.5 Properties of Conditional Expectation

Conditional expectation has some useful properties:

- If X and Y are *independent*, then $\mathbb{E}(\mathbb{X}|\mathbb{Y}) = \mathbb{E}(X)$
- For any function h, $\mathbb{E}(h(Y)X|Y) = h(Y)\mathbb{E}(X|Y)$.
- Linearity: $\mathbb{E}(X_1 + X_2|Y) = \mathbb{E}(X_1|Y) + \mathbb{E}(X_2|Y)$, and $\mathbb{E}(cX|Y) = c\mathbb{E}(X|Y)$, $\forall c \in \mathbb{R}$.
- Adam's Law: $\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X)$.
- Adam's Law with extra conditioning: For any r.v.s X, Y, Z, $\mathbb{E}(\mathbb{E}(X|Y,Z)|Z) = \mathbb{E}(X|Z)$. This is true because conditional probabilities are probabilities, so we are free to use Adam's Law here.

9.6 Conditional Variance

Definition 9.7. The **conditional variance** of X given $Y = y$ is denoted by $\text{Var}(X|Y = y)$, and is just the variance of X given that Y takes the value y. A fundamental result about variance (Eve's Law) is:

$$\text{Var}(X) = \mathbb{E}(\text{Var}(X|Y)) + \text{Var}(\mathbb{E}(X|Y)) \tag{38}$$

which is also known as the **law of total variance**.

Proof. Let $g(Y) = \mathbb{E}(X|Y)$. By Adam's law, $\mathbb{E}(g(Y)) = \mathbb{E}(Y)$. Then

$$\begin{aligned}\mathbb{E}(\text{Var}(X|Y)) &= \mathbb{E}(\mathbb{E}(X^2|Y) - g(Y)^2) = \mathbb{E}(X^2) - \color{red}{\mathbb{E}(g(Y)^2)} \\ \text{Var}(\mathbb{E}(X|Y)) &= \mathbb{E}(g(Y)^2) - (\mathbb{E}(X))^2 = \color{red}{\mathbb{E}(g(Y)^2)} - \mathbb{E}(X)^2\end{aligned}$$

Now adding these 2 equations (removing the red terms), we have Eve's Law. □

10 Covariance and Correlation

Covariance is a single-number summary of the Joint Distribution of two r.v.s, just like mean and variance for a single r.v. The covariance between r.v.s X and Y is:

$$\begin{aligned}\text{Cov}(X,Y) &= \mathbb{E}((X - \mu_X)(Y - \mu_Y)) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\end{aligned}$$

For general (not necessarily independent) r.v.s X and Y,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y) \tag{39}$$

If X and Y are *independent*, then $\text{Cov}(X,Y) = 0$ (however, reverse implication is not true).

Intuitively, if X and Y tend to move in the **same direction**, then $X - \mu_X$ and $Y - \mu_Y$ will tend to be either both positive or negative, so $(X - \mu_X)(Y - \mu_Y)$ will be positive on average, so covariance is positive. If they move in **opposite directions**, then they tend to have opposite signs, giving a negative covariance.

Theorem 10.1. If X and Y are independent, then they are uncorrelated.

Proof.

$$\begin{aligned}
\mathbb{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_X(x) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} y f_Y(y) \left(\int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x\right)\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x \int_{-\infty}^{\infty} y f_Y(y)\,\mathrm{d}y \\
&= \mathbb{E}(X)\mathbb{E}(Y)
\end{aligned}$$

□

10.1 Properties of Covariance

1. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
2. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
3. $\mathrm{Cov}(X, c) = 0$ for any constant c
4. $\mathrm{Cov}(aX, Y) = a\mathrm{Cov}(X, Y)$ for any constant a
5. $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$
6. $\mathrm{Cov}(X + Y, Z + W) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(X, W) + \mathrm{Cov}(Y, Z) + \mathrm{Cov}(Y, W)$
7. $\mathrm{Var}(X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j))$
8. $\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + 2ab\mathrm{Cov}(X, Y) + b^2\mathrm{Var}(Y)$

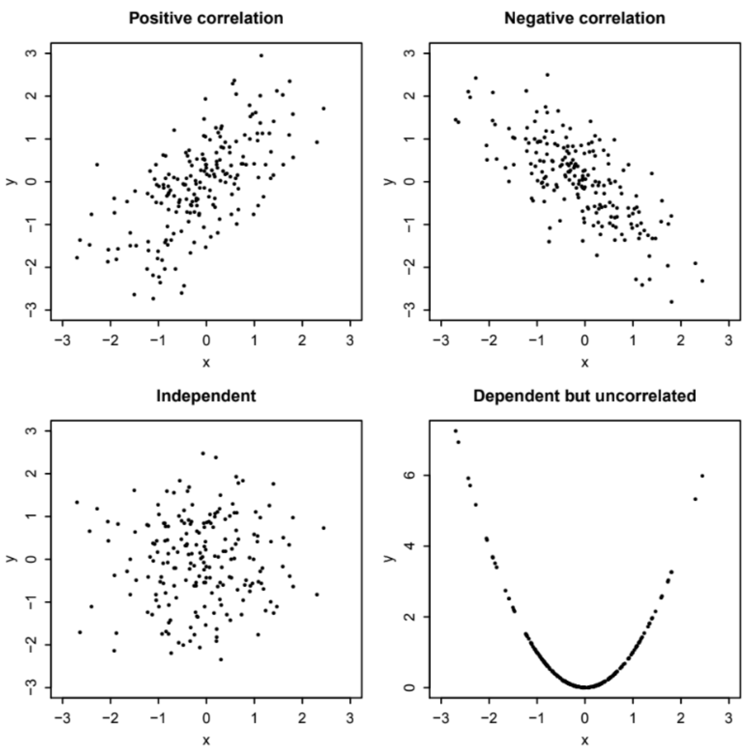

Figure 7: Joint Distribution of (X, Y) under various dependence structures.

Theorem 10.2. For **independent r.v.s**, the variance of the sum is the sum of the variance:

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) \tag{40}$$

Theorem 10.3. If X and Y are independent, then the properties of covariance gives

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) \tag{41}$$

10.2 Correlation

Definition 10.4. The correlation between r.v.s X and Y is

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \tag{42}$$

Theorem 10.5. Note that shifting and scaling X and Y has no effect on their correlation:

$$\text{Corr}(cX, y) = \frac{\text{Corr}(cX, Y)}{\sqrt{\text{Var}(cX)\text{Var}(Y)}} = \frac{c\text{Cov}(X, Y)}{\sqrt{c^2\text{Var}(X)\text{Var}(Y)}} = \text{Corr}(X, Y)$$

Theorem 10.6. (Correlation Bounds) For any r.v.s X and Y,

$$-1 \leq \text{Corr}(X, Y) \leq 1$$

11 Bivariate Normal

In order to fully specify a Bivariate Normal distribution for (X, Y), we need to know five parameters:

- The means $\mathbb{E}(X)$, $\mathbb{E}(Y)$
- The variances $\text{Var}(X)$, $\text{Var}(Y)$
- The correlation $\text{Corr}(X, Y)$

Definition 11.1. The r.v.s X and Y are said to have a **bivariate normal distribution** if their *joint pdf* for all real x and y is given by:

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}Q(x, y)\right] \tag{43}$$

where $Q(x, y) = \left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)$

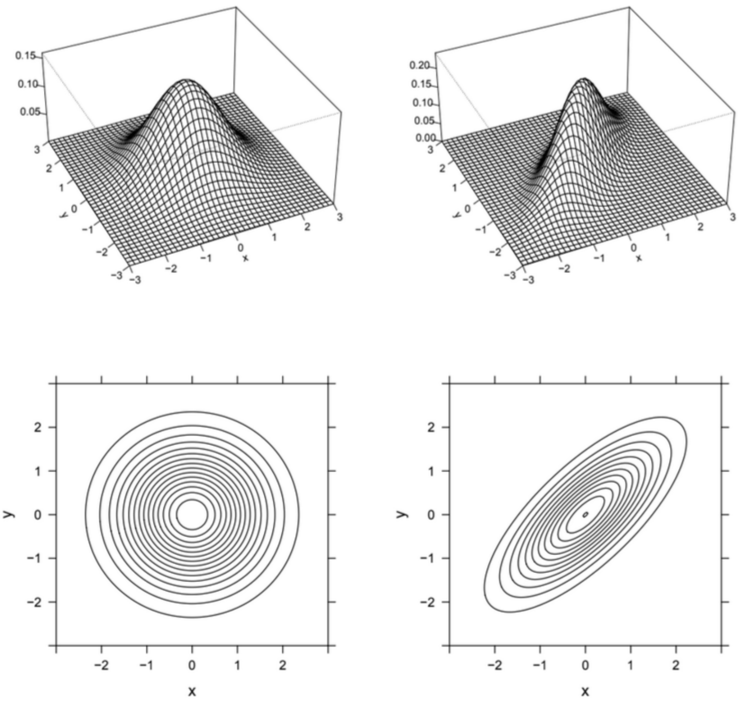

Figure 8: Joint PDFs of two Bivariate Normal Distributions. On the left, X and Y are marginally $\mathcal{N}(0,1)$ and have 0 correlation. On the right, they have correlation 0.75.

Theorem 11.2. If each of X and Y is a linear combination of independent normal r.v.s $U_1, U_2, \ldots, U_n$, then X and Y

Theorem 11.3. If X and Y have a bivariate normal distribution, then the **marginal** pdf's f_X and f_Y are **also normal**,

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2), \quad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2), \quad \rho_{X,Y} = \rho$$

Theorem 11.4. If X and Y have a bivariate normal distribution, then being **uncorrelated** ($\mathrm{Cov}(X,Y) = \rho_{X,Y} = 0$) is the same as being **independent** - From equation 42.

In other words, for bivariate normal X and Y, $\mathrm{Cov}(X,Y) = 0$ if and only if X and Y are independent.

Theorem 11.5. (Independence of sum and difference) Let $X, Y \sim^{i.i.d} \mathcal{N}(0,1)$. The joint distribution of $(X+Y, X-Y)$:

$$\mathrm{Cov}(X+Y, X-Y) = \mathrm{Var}(X) - \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,X) - \mathrm{Var}(Y) = 0$$

$X+Y$ is independent of $X-Y$. Furthermore, they are i.i.d $\mathcal{N}(0,2)$.

Theorem 11.6. If X an Y have a bivariate normal distribution, then the conditional pdf of X given $Y = y$ is also **normal** (vice versa).

The conditional pdf's can be visualised as cross-sections of the joint pdf.

Theorem 11.7. If X and Y have a bivariate normal distribution, then any *linear combination* of X and Y is also **normal**. That is, for constants a and b, if $W \sim aX + bY$, with $\mathbb{E}(W) = \mu$ and $\text{Var}(W) = \sigma^2$, then $W \sim \mathcal{N}(\mu, \sigma^2)$.

12 Poisson Processes

Definition 12.1. (1D Poisson Process) A sequence of arrivals in continuous time is a *Poisson Process* with rate λ if the following conditions hold:

1. The (average) number of arrivals in an interval of length t is distributed with $\text{Poi}(\lambda t)$ (The rate is scalable with time, and the expected number of occurrences in any interval of length is λt).
2. The numbers of arrivals in *disjoint* time intervals are **independent**.

A Poisson process describes the 'most random' way to distribute events in time.

Definition 12.2. Consider a Poisson process on $(0, \infty)$. Let $N(t)$ be the number of arrivals in $(0, t]$. It can be shown that $N(t) \sim \text{Poisson}(\lambda, t)$:

$$\mathbb{P}(N(t) = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!} \tag{44}$$

Note that $N(t)$ only depends on the **length of the interval** (and not when the interval starts or ends).

12.1 Interarrival Time

Definition 12.3. In a Poisson process, let T_1 be the time of the 1st arrival, T_2 be the time between the 1st and 2nd arrivals etc. The T_i's are called interarrival times.
In a Poisson process with rate λ, the interarrival times T_i's are i.i.d. exponential(λ) random variables.

So the waiting time between successive events is exponential(λ), while the *arrival time* of the nth event is gamma(n, λ).

12.2 Merging

Definition 12.4. Given 2 *independent* Poisson processes, where 1 process has rate λ_1 and process 2 has rate λ_2, their **merge** is another Poisson process, with rate $(\lambda_1 + \lambda_2)$.

References

Some references used in these notes:

Introduction to Probability, Joe Blitzstein & Jessica Hwang.