MA2116 - Probability

Liew Zhao Wei

Semester 1, 2023-2024

Abstract

This is a set of incomplete notes for an introductory course on probability.

1 Preliminaries

1.1 Taylor series expansions

We should know the following Taylor series expansions by heart:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

$$\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}$$

$$\sin x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}$$

$$\cos x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!}$$

1.2 Combinatorics

Here are some useful combinatorial identities.

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

$$i\binom{n}{i} = n\binom{n-1}{i-1}$$

1.3 Calculus

The following theorem in calculus will be important for studying continuous random variables.

> **Theorem 1.1** (First Fundamental Theorem of Calculus)
> Let f be continuous on $[a, b]$ and define F on $[a, b]$ by
>
> $$F(x) = \int_a^x f(t)\, dt.$$
>
> Then, F is differentiable on (a, b) and
>
> $$F'(x) = f(x)$$
>
> for all x in (a, b).

2   Axioms of Probability

TODO.

3   Discrete random variables

Let's start simple – with discrete random variables. A random variable X is *discrete* if it takes values in a countable set. $p_X(a) = \Pr(X = a)$ is the *probability mass function* of X. $F_X(a) = \Pr(X \le a) = \sum_{x \le a} p(x)$ is the *cumulative distribution function* of X.

3.1   Expectation

When given a random variable, we are often interested in its *average* value. This is captured by the notion of *expectation*.

Definition 3.1 (Expected value of a discrete r.v.)
Let X be a discrete random variable. Then the *expected value*, or *expectation*, of X, denoted by $\mathrm{E}[X]$, is defined as
$$\mathrm{E}[X] = \sum_{x \in \Omega_X} x \Pr(X = x).$$

We also have the following intuitive results for the expectated value of a discrete random variable.

Proposition 3.2 (Law of the unconscious statistician)
Let X be a discrete random variable and g be real-valued a function. Then
$$\mathrm{E}[g(X)] = \sum_{x \in \Omega_X} g(x) \Pr(X = x).$$

Corollary 3.3
Let X be a discrete random variable and $a, b \in \mathbb{R}$. Then
$$\mathrm{E}[aX + b] = a \, \mathrm{E}[X] + b.$$

Proposition 3.4 (Alternative formula for expectation of nonnegative discrete r.v.)
Let X be a *nonnegative* discrete random variable and g be real-valued a function. Then
$$\mathrm{E}[X] = \sum_{x=0} \Pr(X > x).$$

One important property of expectations of (general) random variables is that it is *linear*.

Theorem 3.5
Linearity of expectation Let $X_1, X_2, \ldots, X_n$ be random variables (not necessarily discrete). Then
$$\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{E}[X_i].$$

Another useful result to use is the *law of total expectation*.

Theorem 3.6 (Law of total expectation)
Let X be a discrete random variable and Y be a discrete random variable that depends on X. Then

$$E[Y] = E[E[Y \mid X]] = \sum_{x \in \Omega_X} E[Y \mid X = x] \Pr(X = x).$$

3.2 Variance

Expectation alone doesn't reveal much about the distribution of a random variable. For example, compare these two slot machines:

1. The first slot machine gives you $100 with probability 1.

2. The second slot machine gives you $10000 with probability 0.01 and $0 with probability 0.99.

The expected payout of both slot machines is $100, but the chances of winning are very different. This difference is captured by the notion of *variance*.

Definition 3.7 (Variance of a discrete r.v.)
Let X be a discrete random variable. Then the *variance* of X, denoted by $\mathrm{Var}(X)$, is defined as

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

Remark 3.8 (Alternative formula). *The variance can also be computed using the following formula:*

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

Remark 3.9. *Since variance is non-negative, we have*

$$E[X^2] \geq E[X]^2.$$

This can be used to prove the friendship paradox.

Proposition 3.10 (Variance of linear combination)
Let X be a discrete random variable and $a, b \in \mathbb{R}$. Then

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X).$$

3.3 Bernoulli and binomial random variables

Perhaps the most basic and well-known random variable is the *binomial* random variable and its little brother, the *Bernoulli* random variable. We motivate this with the age old example of a coin toss.

Example 3.11 (Coin toss)
Suppose we toss a coin n times that comes up heads with probability p and tails with probability $1 - p$. Let X be the number of heads that appear. Then X is a *binomial* random variable with parameters n and p. It takes values in $\{0, 1, \ldots, n\}$.

Definition 3.12 (Binomial random variable)
A random variable X is said to be *binomial* with parameters n and p, denoted by $X \sim \mathrm{Bin}(n, p)$, if X has the following probability mass function:

$$\Pr(X = i) = \binom{n}{i} p^i (1 - p)^{n-i}, \quad i = 0, 1, \ldots, n.$$

Remark 3.13 (Bernoulli random variable). *A Bernoulli random variable is a special case of a binomial random variable with $n = 1$.*

When $n = 1$, we can easily compute that $E[X] = p$ and $\text{Var}(X) = p(1 - p)$. By the additivity of expectation and variance, it is also easy to compute the expectation and variance for general n.

Theorem 3.14 (Properties of binomial random variables)
Let $X \sim \text{Bin}(n, p)$. Then

1. $E[X] = np$

2. $\text{Var}(X) = np(1 - p)$

3. $E[X^k] = np\,E[(Y + 1)^{k-1}]$ where $Y \sim \text{Bin}(n - 1, p)$

Proof. A proof of $E[X^k]$ is done with some algebraic manipulation of the definition of $E[X^k]$ and using the identity

$$i\binom{n}{i} = n\binom{n - 1}{i - 1}$$

$\square$

The following relationship is useful for computing the distribution function of a binomial random variable. It can also be used to find the maximum of the probability mass function.

Proposition 3.15 (Relationship between $\Pr(X = k + 1)$ and $\Pr(X = k)$)
Let $X \sim \text{Bin}(n, p)$. Then

$$\Pr(X = k + 1) = \frac{p}{1 - p}\frac{n - k}{k + 1}\Pr(X = k).$$

Corollary 3.16 (Condition for monotonic increase)
Let $X \sim \text{Bin}(n, p)$. Then, as k goes from 0 to n, $\Pr(X = k)$ first increases monotonically and then decreases monotonically. In fact, $\Pr(X = k) \geq \Pr(X = k + 1)$ if and only if $k \leq (n + 1)p$.

3.4 Poisson random variable

Besides the binomial random variable, another important discrete random variable is the *Poisson* random variable.

Definition 3.17 (Poisson random variable)
A random variable X is said to be *Poisson* with parameter $\lambda > 0$, denoted by $X \sim \text{Pois}(\lambda)$, if X has the following probability mass function:

$$\Pr(X = i) = e^{-\lambda}\frac{\lambda^i}{i!}, \quad i = 0, 1, 2, \ldots.$$

Observe that the ith term is $e^{-\lambda}$ times the ith term of the Taylor series expansion of e^λ.

Remark 3.18. *By the Taylor series expansion of e^x, we have*

$$\sum_{i=0}^{\infty}\Pr(X = i) = e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!} = e^{-\lambda}e^\lambda = 1.$$

The Poisson random variable is useful for approximating the binomial random variable when n is large and p is small enough so that $\lambda = np$ is of moderate size.

Example 3.19 (Examples that can be approximated by a Poisson r.v.)
We can think of each of these examples as a long series of independent trial, each with a small chance of success. Thus, we can approximate the number of successes by a Poisson random variable.

1. The number of misprints on a page of a book.

2. The number of people in a community who survive to age 100.

3. The number of wrong telephone numbers dialed in a day.

4. The number of packages of biscuits sold in a store in a day.

5. The customers centering a post office on a given day.

Remark 3.20. *This approximation works decently even with n as small as 10 and p as large as 0.1. For example, with $X \sim \text{Bin}(10, 0.1)$ and $Y \sim \text{Pois}(1)$, we have*

$$\Pr(X \leq 1) \approx 0.7361$$

$$\Pr(Y \leq 1) \approx 0.7358$$

Theorem 3.21 (Properties of Poisson r.v.)
Let $X \sim \text{Pois}(\lambda)$. Then

1. $\text{E}[X] = \lambda$

2. $\text{Var}(X) = \lambda$

These properties should be intuitive by comparing Poisson random variables with binomial random variables, but with large n.

We have shown that the Poisson distribution approximates the binomial distribution when n is large and p is small. In fact, we do not need the events be independent or to occur with equal probability. We just need the probability of the events to be small and the events to be at most weakly dependent. This is explained in what we call the *Poisson paradigm*.

Definition 3.22 (Weak dependence)
Events E and F are weakly dependent if $\Pr(E) \approx \Pr(E \mid F)$.

Theorem 3.23 (Poisson Paradigm)
Consider n events where the ith event occurs with probability p_i. If all the p_i are small and the trials are either independent or at most weakly dependent, then the number of events that occur approximately has a Poisson distribution with mean $\lambda = \sum_{i=1}^{n} p_i$.

Proposition 3.24 (Relationship between $\Pr(X = k + 1)$ and $\Pr(X = k)$)
Let $X \sim \text{Pois}(\lambda)$. Then
$$\Pr(X = k + 1) = \frac{\lambda}{k + 1} \Pr(X = k).$$

One other use of the Poisson distribution is in *poisson processes*.

Theorem 3.25 (Poisson process)

Suppose events occur at random points in time, and let $N(t)$ denote the number of events that occur in an interval of length t. $N(t)$ is a *Poisson process* having rate $\lambda > 0$ if

1. The process begins at time 0, that is, $N(0) = 0$.

2. $\Pr(N(h) = 1) = \lambda h + o(h)$

3. $\Pr(N(h) \geq 2) = o(h)$

4. The number of events occurring in disjoint time intervals are independent (otherwise called the independent increment assumption).

5. The distribution of the number of events occurring in a time interval of length t depends only on the length of the interval and not on its location (otherwise called the stationary increment assumption).

$N(t)$ has a Poisson distribution with mean λt.

Example 3.26 (Examples of poisson processes)

Poisson processes can approximate phonemona like:

1. The number of earthquakes occurring in a fixed time span.

2. The number of wars per year.

3. The number of electrons emitted from a heated cathode during a fixed time period.

4. The number of deaths, in a given period of time, of the policyholders of a life insurance company.

3.5   Geometric random variable

Imagine tossing a coin repeatedly until we get a head. Let X be the number of tosses required. Then X is a *geometric* random variable.

Definition 3.27 (Geometric random variable)

A random variable X is said to be *geometric* with parameter p if X has the following probability mass function:

$$\Pr(X = i) = (1 - p)^{i-1} p, \quad i = 1, 2, \dots.$$

Remark 3.28. *By the geometric series, we have*

$$\sum_{i=1}^{\infty} \Pr(X = i) = p \sum_{i=1}^{\infty} (1 - p)^{i-1} = 1$$

Proposition 3.29 (Properties of geometric r.v.)

Let X be a geometric random variable with parameter p. Then

1. $\mathrm{E}[X] = 1/p$

2. $\mathrm{Var}(X) = (1 - p)/p^2$

The expectation should be unsurprising. The variance should also be somewhat intuitive since

1. When p is small, X is approximately an exponential random variable with $\lambda = p$ and so $\mathrm{Var}(X) \approx 1/p^2$.

2. When p is large, $\mathrm{Var}(X) \approx 0$.

Remark 3.30. *The exponential distribution is the continuous analogue of the geometric distribution.*

3.6 Negative binomial random variable

Suppose we toss a coin repeatedly until we get r heads. Let X be the number of tosses required. Then X is a *negative binomial* random variable.

Definition 3.31 (Negative binomial random variable)
A random variable X is said to be *negative binomial* with parameters r and p if X has the following probability mass function:
$$\Pr(X = i) = \binom{i-1}{r-1} p^r (1-p)^{i-r}, \quad i = r, r+1, \dots.$$

Remark 3.32. *A geometric random variable is a special case of a negative binomial random variable with $r = 1$.*

The following properties should not be surprising since the negative binomial random variable is a sum of r independent geometric random variables.

Proposition 3.33 (Properties of negative binomial r.v.)
Let X be a negative binomial random variable with parameters r and p. Then

1. $\mathrm{E}[X] = r/p$

2. $\mathrm{Var}(X) = r(1-p)/p^2$

3.7 Hypergeometric random variable

Suppose we randomly choose n balls from an urn containing m white balls and $N - m$ black balls without replacement. Let X be the number of white balls chosen. Then X is a *hypergeometric* random variable.

Definition 3.34 (Hypergeometric random variable)
A random variable X is said to be *hypergeometric* with parameters n, N, and m if X has the following probability mass function:
$$\Pr(X = i) = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}, \quad i = 0, 1, \dots, n.$$

Proposition 3.35 (Properties of hypergeometric r.v.)
Let X be a hypergeometric random variable with parameters n, N, and m. Let $p = m/N$. Then

1. $\mathrm{E}[X] = np$

2. $\mathrm{Var}(X) = np(1-p)\left(1 - \frac{n-1}{N-1}\right)$

The following remark is intuitive since binomial is like hypergeometric, but with replacement.

Remark 3.36. *When $N \gg n$, X is approximately a binomial random variable with parameters n and p.*

4 Continuous random variables

Now, onto something more interesting – continuous random variables. A random variable X is *continuous* if there exists a non-negative function f such that for any measurable set $B<$

$$\Pr(X \in B) = \int_B f(x)\, dx$$

f is called the *probability density function* of X. Observe that it is meaningless to discuss the probability of X taking a particular value.

The cumulative distribution function F_X of X is defined as

$$F_X(a) = \Pr(X \leq a) = \int_{-\infty}^{a} f(x)\,dx = \Pr(X < a)$$

It is clear that the density is the derivative of the cumulative distribution function. An alternative intuitive interpretation is that $f(a)$ is how likely X will be near a, as given by

$$\Pr\left(a - \frac{\epsilon}{2} < X < a + \frac{\epsilon}{2}\right) = \int_{a-\epsilon/2}^{a+\epsilon/2} f(x)\,dx \approx \epsilon f(a)$$

when ϵ is small and $f(\cdot)$ is continuous at $x = a$.

4.1 Expectation

Recall that the expected value of a discrete random variable is defined as

$$\mathrm{E}[X] = \sum_{x \in \Omega_X} x \Pr(X = x).$$

We have an analogous definition of expectation for continuous random variables.

Definition 4.1 (Expected value of a continuous r.v.)
Let X be a continuous random variable. Then the *expected value*, or *expectation*, of X, denoted by $\mathrm{E}[X]$, is defined as
$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\,dx.$$

This is intuitive since we can imagine partitioning the domain of x into small subintervals dx, and for small dx, we have

$$f(x)\,dx \approx \Pr(x \leq X \leq x + dx)$$

Just as in the discrete case, we have the following results for continuous random variables.

Proposition 4.2 (Law of the unconscious statistician)
Let X be a continuous random variable and g be real-valued a function. Then

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx.$$

Corollary 4.3
Let X be a continuous random variable and $a, b \in \mathbb{R}$. Then,

$$\mathrm{E}[aX + b] = a\,\mathrm{E}[X] + b$$

Lemma 4.4 (Alternative formula for expectation of nonnegative continuous r.v.)
Let Y be a *nonnegative* continuous random variable. Then,

$$\mathrm{E}[Y] = \int_{0}^{\infty} \Pr(Y > y)\,dy.$$

By a similar proof, the *law of total expectation* holds for continuous random variables.

Theorem 4.5 (Law of total expectation)
Let X be a continuous random variable and Y be a continuous random variable that depends on X. Then

$$E[Y] = E[E[Y \mid X]] = \int_{-\infty}^{\infty} E[Y \mid X = x] f(x) \, dx.$$

4.2 Variance

Variance for continuous random variables is defined exactly as it is in the discrete case.

Definition 4.6 (Variance of a continuous r.v.)
Let X be a continuous random variable. Then the *variance* of X, denoted by $\text{Var}(X)$, is defined as

$$\text{Var}(X) = E[(X - E[X])^2].$$

We restate the same results for variance as in the discrete case. They are proved similarly.

Remark 4.7 (Alternative formula). *The variance can also be computed using the following formula:*
$$\text{Var}(X) = E[X^2] - E[X]^2.$$

Remark 4.8. *Since variance is non-negative, we have*
$$E[X^2] \geq E[X]^2.$$

Proposition 4.9 (Variance of linear combination)
Let X be a discrete random variable and $a, b \in \mathbb{R}$. Then

$$\text{Var}(aX + b) = a^2 \, \text{Var}(X).$$

4.3 Uniform random variable

The simplest continuous random variable is the *uniform* random variable.

Definition 4.10 (Uniform random variable)
A random variable X is said to be *uniform* on the interval $[a, b]$ if X has the following probability density function:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

The cumulative distribution function of X is

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq a \\ \dfrac{x-a}{b-a} & \text{if } a < x < b \\ 1 & \text{if } x \geq b \end{cases}$$

Proposition 4.11 (Properties of uniform r.v.)
Let X be a uniform random variable on the interval (a, b). Then

1. $E[X] = \dfrac{a+b}{2}$

2. $\text{Var}(X) = \dfrac{(b-a)^2}{12}$

Remark 4.12. *Given the uniform distribution X over $(0,1)$, we can compute a uniform distribution over (a,b) using the transformation $Y = a + (b-a)X$.*

4.4 Normal random variable

The most widely used continuous random variable is the *normal* random variable.

Definition 4.13 (Normal random variable)

A random variable X is said to be *normal* with parameters μ and variance σ^2 if X has the following probability density function:

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

The cumulative distribution function of X is

$$F_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \, dt$$

Remark 4.14. *Passing to two dimensions and using polar coordinates, we have*

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} \, dx = \sqrt{2\pi}$$

which helps us show that f is a probability density function.

Definition 4.15 (Standard normal random variable)

It is customary to call the normal random variable with parameters $\mu = 0$ and $\sigma^2 = 1$ the *standard normal* random variable. It is denoted as $Z \sim N(0,1)$. The cumulative distribution function of Z is denoted by Φ.

Remark 4.16. *For any x, we have*

$$\Phi(-x) = 1 - \Phi(x)$$

Remark 4.17. *Given the standard normal distribution Z, we can compute a normal distribution with mean μ and variance σ^2 using the transformation $Y = \mu + \sigma Z$.*

Proposition 4.18 (Properties of normal r.v.)

Let X be a normal random variable with parameters μ and σ^2. Then

1. $E[X] = \mu$

2. $\text{Var}(X) = \sigma^2$

Recall that we can approximate a binomial random variable with a Poisson random variable when n is large and p is small so that $\lambda = np$ is of moderate size. Similarly, the following result shows that we can approximate a binomial random variable with a normal random variable when $np(1-p)$ is large (typically at least 10).

Theorem 4.19 (De Moivre-Laplace limit theorem)

Let $S_n \sim \text{Bin}(n,p)$ and $X \sim N(np, np(1-p))$. Then, for any $a < b$,

$$\Pr(a \le S_n \le b) \to \Pr(a \le X \le b)$$

as $n \to \infty$.

This result is a consequence of the central limit theorem, which we will see later.

Remark 4.20. *Let $S_n \sim \text{Bin}(n,p)$ and $X \sim N(np, np(1-p))$. Then, $P(S_n = i)$ is approximated by $P(i - 0.5 \le X \le i + 0.5)$.*

4.5   Exponential random variable

Consider a Poisson process $N(t)$ with rate λ. If X is the waiting time until the first event occurs, then

$$P(X > t) = P(N(t) = 0) = e^{-\lambda t}$$

We call X an *exponential* random variable with parameter λ. It is the continuous analogue of the geometric random variable.

Definition 4.21 (Exponential random variable)
A random variable X is said to be *exponential* with parameter $\lambda > 0$ if X has the following probability density function:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The cumulative distribution function of X is

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Just like geometric distributions, the exponential distribution are memoryless. In fact, it is the only continuous distribution that is memoryless.

Definition 4.22 (Memorylessness)
A random variable X is said to be *memoryless* if for all $s, t \geq 0$,

$$\Pr(X > s + t \mid X > s) = \Pr(X > t)$$

Remark 4.23. *Given an exponential distribution X with parameter 1, we can compute an exponential distribution with parameter λ using the transformation $Y = \lambda X$.*

Proposition 4.24 (Properties of exponential r.v.)
Let X be an exponential random variable with parameter λ. Then

1. $\mathrm{E}[X^n] = \frac{n}{\lambda} \mathrm{E}[X^{n-1}]$

2. $\mathrm{E}[X] = 1/\lambda$

3. $\mathrm{Var}(X) = 1/\lambda^2$

4.6   Laplace random variable

The *Laplace* random variable a variation of the exponential random variable that is equally likely to be positive or negative, and whose absolute value is exponentially distributed with parameter $\lambda \geq 0$.

Definition 4.25 (Laplace random variable)
A random variable X is said to be *Laplace* or *double exponential* with parameter $\lambda > 0$ if X has the following probability density function:

$$f(x) = \frac{\lambda}{2} e^{-\lambda |x|}$$

The cumulative distribution function of X is

$$F_X(x) = \begin{cases} \dfrac{1}{2} e^{-\lambda x} & \text{if } x < 0 \\ 1 - \dfrac{1}{2} e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

Remark 4.26. *If X is a Laplace random variable with parameter λ, then $|X|$ is an exponential random variable with parameter λ.*

4.7 Hazard rate function

Definition 4.27 (Hazard rate function)
Let X be a continuous random variable denoting the lifetime of an item, with cumulative distribution function F and probability density function f. Then the *hazard rate function* of X is defined as

$$h(x) = \frac{f(x)}{1 - F(x)}$$

It represents the instantaneous probability intensity of failure at time x, given that the object has survived up to time x.

Since the exponential distribution is memoryless, we should verify that its hazard rate function is constant.
It turns out that the hazard rate function uniquely determines the distribution function F.

Remark 4.28. *Let $\lambda(s), s \geq 0$ be the hazard rate function of a non-negative random variable X. Then,*

$$F(x) = 1 - \exp\left(-\int_0^x \lambda(s)\, ds\right)$$

4.8 Gamma random variable

A *Gamma* random variable represents the time until the nth event in a Poisson process with rate λ. It is a generalization of the exponential random variable.

Definition 4.29 (Gamma random variable)
A random variable X is said to be *gamma* with parameters $\alpha > 0$ and $\lambda > 0$ if X has the following probability density function:

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

The cumulative distribution function of X is

$$F_X(x) = \begin{cases} \dfrac{\gamma(\alpha, \lambda x)}{\Gamma(\alpha)} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\gamma(\alpha, \lambda x)$ is the lower incomplete gamma function.

Note that α does not need to be an integer, but if when it is, we have $\Gamma(\alpha) = (\alpha - 1)!$.

Remark 4.30 (*n*-Erlang distribution). *Let $n = \alpha$ be an integer. Then, the Gamma distribution with parameters (n, λ) is known as the n-Erlang distribution.*

Given the similarities between the Gamma and exponential distributions, the following properties should not be surprising.

Proposition 4.31 (Properties of gamma r.v.)
Let X be a gamma random variable with parameters α and λ. Then

1. $\mathrm{E}[X] = \dfrac{\alpha}{\lambda}$

2. $\mathrm{Var}(X) = \dfrac{\alpha}{\lambda^2}$

4.9   Beta random variable

Definition 4.32 (Beta function)
The *beta function* is defined as
$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

A beta random variable represents the probability density of attaining $\alpha - 1$ successes and $\beta - 1$ failures in $\alpha + \beta - 2$ Bernoulli trials with probability x.

Definition 4.33 (Beta random variable)
A random variable X is said to be *beta* with parameters $\alpha > 0$ and $\beta > 0$ if X has the following probability density function:
$$f(x) = \begin{cases} \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$
The cumulative distribution function of X is
$$F_X(x) = \begin{cases} \dfrac{B(x; \alpha, \beta)}{B(\alpha, \beta)} & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

where $B(\alpha, \beta)$ is the beta function.

Proposition 4.34 (Properties of beta r.v.)
Let X be a beta random variable with parameters α and β. Then

1. $\text{E}[X] = \dfrac{\alpha}{\alpha + \beta}$

2. $\text{Var}(X) = \dfrac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$

4.10   Distributions of functions of random variables

When composing functions of continuous random variables, we want to be able to compute the distribution of the resulting random variable. There are some techniques we can use.

Theorem 4.35
Let X be a continuous random variable having probability density funciton f_X. Suppose that $g(x)$ is a strictly monotonic (increasing or decreasing), differentiable function of x. Then, the random variable Y defined by $Y = g(X)$ has probability density function
$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \dfrac{d}{dy} g^{-1}(y) \right| & \text{if } g^{-1}(y) \text{ exists} \\ 0 & \text{otherwise} \end{cases}$$

5   Jointly Distributed Random Variables

We are often interested in probability statements concerning multiple random variables. Let's define some terms for this.

Definition 5.1 (Joint cumulative probability distribution function)

Let X and Y be random variables. The *joint cumulative probability distribution function* of X and Y is defined as
$$F(a, b) = \Pr(X \leq a, Y \leq b)$$

Any joint probabiltiy statement can be answered in terms of F. For example,
$$\Pr(a_1 \leq X \leq a_2, b_1 \leq Y \leq b_2) = F(a_2, b_2) + F(a_1, b_1) - F(a_1, b_2) - F(a_2, b_1)$$

When X and Y are discrete random variables, it is convenient to define their *joint probability mass function*.

Definition 5.2 (Joint probability mass function)

Let X and Y be discrete random variables. The *joint probability mass function* of X and Y is defined as
$$p(x, y) = \Pr(X = x, Y = y)$$

The probability mass function of X is then given by
$$p_X(x) = \sum_j p(x, y_j)$$

Similarly, when X and Y are continuous, we can define their *joint probability density function*.

Definition 5.3 (Jointly continuous)

Let X and Y be continuous random variables. Then X and Y are said to be *jointly continuous* if there exists a function $f(x, y)$ such that for any set C of pairs (x, y),
$$\Pr((X, Y) \in C) = \iint_C f(x, y) \, dx \, dy$$

f is called the *joint probability density function* of X and Y.

Remark 5.4. *In particular, we have*
$$F(a, b) = \int_{-\infty}^b \int_{-\infty}^a f(x, y) \, dx \, dy$$
$$f(a, b) = \frac{\partial^2}{\partial a \partial b} F(a, b)$$

Remark 5.5. *Based on the derived equation below, we can also interpret $f(a, b)$ as the probability that (X, Y) lies near (a, b).*
$$\Pr(a \leq X \leq a + \Delta a, b \leq Y \leq b + \Delta b) \approx f(a, b) \Delta a \Delta b$$

Remark 5.6. *The probability density function of X is given by*
$$\Pr(X \in A) = \int_A \int_{-\infty}^\infty f(x, y) \, dy \, dx = \int_A f_X(x) \, dx$$

Generalizing these definitions to n random variables, we have the following definitions.

Definition 5.7 (Definitions for n joint r.v.s)

Let $X_1, X_2, \ldots, X_n$ be random variables. The joint cumulative probability distribution function is
$$F(a_1, a_2, \ldots, a_n) = \Pr(X_1 \leq a_1, X_2 \leq a_2, \ldots, X_n \leq a_n)$$

They are *jointly continuous* if there exists a function $f(x_1, x_2, \ldots, x_n)$ such that for any set C of pairs $(x_1, x_2, \ldots, x_n)$,
$$\Pr((X_1, X_2, \ldots, X_n) \in C) = \iint \cdots \int_C f(x_1, x_2, \ldots, x_n) \, dx_1 \, dx_2 \ldots dx_n$$

> **Example 5.8**
> The multinomial distribution is a joint distribution.

6 References

1. A First Course in Probability (10th edition) by Sheldon Ross

2. Stanford CS109 readings

3. Stanford CS109 course resources