Stochastic Processes 1

STAT 333

Fall 2021 (1219)[1]

Cameron Roopnarine[2] Steve Drekic[3] Mirabelle Huynh[4]

17th October 2021

[1] Online Course

[2] $\LaTeX$er

[3] Main Instructor

[4] Assisting Instructor

Contents

Chapter 1

Review of Elementary Probability

Week 1
8th to 15th September

Fundamental Definition of a Probability Function

Probability Model: A probability model consists of 3 essential components: a *sample space*, a collection of *events*, and a *probability function (measure)*.

- **Sample Space**: For a random experiment in which all possible outcomes are known, the set of all possible outcomes is called the sample space (denoted by Ω).
- **Event**: Every subset A of a sample space Ω is an event.
- **Probability Function**: For each event A of Ω, $\mathbb{P}(A)$ is defined as the *probability of an event A*, satisfying 3 conditions:
  (i) $0 \leq \mathbb{P}(A) \leq 1$,
  (ii) $\mathbb{P}(\Omega) = 1$, or equivalently, $\mathbb{P}(\emptyset) = 0$, where $\emptyset$ is the *null event*,
  (iii) For $n \in \mathbb{Z}^+$ (in fact, $n = \infty$ as well), $\mathbb{P}(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}(A_i)$ if the sequence of events $\{A_i\}_{i=1}^n$ is *mutually exclusive* (i.e., $A_i \cap A_j = \emptyset \; \forall i \neq j$).

As a result of conditions (ii) and (iii), and noting that A^c is the complement of A, it follows that

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c) \implies \mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

Conditional Probability

Conditional Probability: The *conditional probability of event A given event B occurs* is defined as

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

provided that $\mathbb{P}(B) > 0$.

Remarks:

(1) When $B = \Omega$, $\mathbb{P}(A \mid \Omega) = \mathbb{P}(A \cap \Omega)/\mathbb{P}(\Omega) = \mathbb{P}(A)/1 = \mathbb{P}(A)$, as one would expect.

(2) Rewriting the above formula, $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B)$, which is often referred to as the basic "multiplication rule." For a sequence of events $\{A_i\}_{i=1}^n$, the generalized multiplication rule is given by

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1) \cdots \mathbb{P}(A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Example 1.1. Suppose that we roll a fair six-sided die once (i.e., $\Omega = \{1, 2, 3, 4, 5, 6\}$). Let A denote the event of rolling a number less than 4 (i.e., $A = \{1, 2, 3\}$), and let B denote the event of rolling an odd number (i.e., $B = \{1, 3, 5\}$). Given that the roll is odd, what is the probability that number rolled is less than 4?

Solution: Since the die is fair, it immediately follows that $\mathbb{P}(A) = 3/6 = 1/2$ and $\mathbb{P}(B) = 3/6 = 1/2$. Moreover,

$$\begin{aligned} \mathbb{P}(A \cap B) &= \mathbb{P}\big(\{1, 2, 3\} \cap \{1, 3, 5\}\big) && (1.1) \\ &= \mathbb{P}\big(\{1, 3\}\big) && (1.2) \\ &= \frac{2}{6} && (1.3) \\ &= \frac{1}{3}. && (1.4) \end{aligned}$$

Therefore,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/3}{1/2} = \frac{2}{3}.$$

Independence of Events

Independence of Events: Two events A and B are *independent* if and only if (iff)

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

In general, if an experiment consists of a sequence of independent trials, and $A_1, A_2, \ldots, A_n$ are events such that A_i depends only on the i^{th} trial, then $A_1, A_2, \ldots, A_n$ are independent events and

$$\mathbb{P}(\cap_{i=1}^n A_i) = \prod_{i=1}^n \mathbb{P}(A_i).$$

Law of Total Probability

Law of Total Probability: For $n \in \mathbb{Z}^+$ (and even $n = \infty$), suppose that $\Omega = \cup_{i=1}^n B_i$, where the sequence of events $\{B_i\}_{i=1}^n$ is mutually exclusive. Then,

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap \Omega)\\ &= \mathbb{P}\big(A \cap \{\cup_{i=1}^n B_i\}\big)\\ &= \mathbb{P}\big(\cup_{i=1}^n \{A \cap B_i\}\big)\\ &= \sum_{i=1}^n \mathbb{P}(A \cap B_i)\\ &= \sum_{i=1}^n \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i),\end{aligned}$$

where the second last equality follows from the fact that the sequence of events $\{A \cap B_i\}_{i=1}^n$ is also mutually exclusive.

Bayes' Formula

Bayes' Formula: Under the same assumptions as in the previous slide,

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(A \cap B_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \mid B_j)\,\mathbb{P}(B_j)}{\sum_{i=1}^n \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i)}.$$

Definition of a Random Variable

Definition: A *random variable* (rv) X is a real-valued function which maps a sample space Ω onto a state space $\mathcal{S} \subseteq \mathbb{R}$ (i.e., $X\colon \Omega \to \mathcal{S}$).

Discrete type: $\mathcal{S}$ consists of a finite or countable number of possible values. Important functions include:

$$\begin{aligned}p(a) &= \mathbb{P}(X = a) && \text{(pmf)},\\ F(a) &= \mathbb{P}(X \leq a) = \sum_{x \leq a} p(x) && \text{(cdf)},\\ \bar{F}(a) &= \mathbb{P}(X > a) = 1 - F(a) && \text{(tpf)},\end{aligned}$$

where pmf stands for *probability mass function*, cdf stands for *cumulative distribution function*, and tpf stands for *tail probability function*.

Remark: If X takes on values in the set $\mathcal{S} = \{a_1, a_2, a_3, \ldots\}$ where $a_1 < a_2 < a_3 < \cdots$ such that $p(a_i) > 0$ $\forall i$, then we can recover the pmf from knowledge of the cdf via

$$\begin{aligned}p(a_1) &= F(a_1),\\ p(a_i) &= F(a_i) - F(a_{i-1}),\ i = 2, 3, 4, \ldots\end{aligned}$$

Discrete Distributions

Special Discrete Distributions:

1. **Bernoulli**: If we consider a *Bernoulli trial*, which is a random trial with probability p of being a "success" (denoted by 1) and a probability $1-p$ of being a "failure" (denoted by 0), then X is *Bernoulli* (i.e., $X \sim$ BERN(p)) with pmf

$$p(x) = p^x(1-p)^{1-x}, \ x = 0, 1.$$

2. **Binomial**: If X denotes the number of successes in $n \in \mathbb{Z}^+$ independent Bernoulli trials, each with probability p of being a success, then X is Binomial (i.e., $X \sim$ BIN(n, p)) with pmf

$$p(x) = \binom{n}{x} p^x(1-p)^{n-x}, \ x = 0, 1, \ldots, n,$$

where

$$\binom{n}{x} = \frac{n!}{(n-x)!x!} = \frac{(n)_x}{x!} = \frac{n(n-1)\cdots(n-x+1)}{x!}$$

is the number of distinct groups of x objects chosen from a set of n objects.

Remarks:

(1) A BIN$(1, p)$ distribution simplifies to become the BERN(p) distribution.

(2) The binomial pmf is even defined for $n = 0$, in which case $p(x) = 1$ for $x = 0$. Such a distribution is said to be degenerate at 0.

(3) Note that $\binom{n}{x} = 0$ if $n, x \in \mathbb{N}$ with $n < x$.

3. **Negative Binomial**: If X denotes the number of Bernoulli trials (each with success probability p) required to observe $k \in \mathbb{Z}^+$ successes, then X is *Negative Binomial* (i.e., $X \sim$ NB$_t(k, p)$) with pmf

$$p(x) = \binom{x-1}{k-1} p^k(1-p)^{x-k}, \ x = k, k+1, k+2, \ldots.$$

Remarks:

(1) In the above pmf, $\binom{x-1}{k-1}$ appears rather than $\binom{x}{k}$ since the final trial must always be a success.

(2) Sometimes, a negative binomial distribution is alternatively defined as the number of failures observed to achieve k successes. If Y denotes such a rv and $X \sim$ NB$_t(k, p)$, then we clearly have the relationship $X = Y + k$, which immediately leads to the following pmf for Y:

$$p_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(X = y + k) = \binom{y+k-1}{k-1} p^k(1-p)^y, \ y = 0, 1, 2, \ldots.$$

To refer to this negative binomial distribution, we will write $Y \sim$ NB$_f(k, p)$.

4. **Geometric**: If $X \sim$ NB$_t(1, p)$, then X is *Geometric* (i.e., $X \sim$ GEO$_t(p)$) with pmf

$$p(x) = p(1-p)^{x-1}, \ x = 1, 2, 3 \ldots.$$

In other words, the geometric distribution models the number of Bernoulli trials required to observe the first success.

Remark: Similarly, if $X \sim \text{NB}_f(1,p)$ then we obtain an alternative geometric distribution (denoted by $X \sim \text{GEO}_f(p)$) which models the number of failures observed prior to the first success.

5. **Discrete Uniform**: If X is equally likely to take on values in the (finite) set $\{a, a+1, \ldots, b\}$ where $a, b \in \mathbb{Z}$ with $a \leq b$, then X is *Discrete Uniform* (i.e., $X \sim \text{DU}(a,b)$) with pmf

$$p(x) = \frac{1}{b-a+1},\ x = a, a+1, \ldots, b.$$

6. **Hypergeometric**: If X denotes the number of success objects in n draws without replacement from a finite population of size N containing exactly r success objects, then X is *Hypergeometric* (i.e., $X \sim \text{HG}(N,r,n)$) with pmf

$$p(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}},\ x = \max\{0, n-N+r\}, \ldots, \min\{n, r\}.$$

7. **Poisson**: A rv X is *Poisson* (i.e., $X \sim \text{POI}(\lambda)$) with parameter $\lambda > 0$ if its pmf is one of the form

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!},\ x = 0, 1, 2, \ldots.$$

Remark: The pmf is even defined for $\lambda = 0$ (if we use the standard convention that $0^0 = 1$), in which case $p(x) = 1$ for $x = 0$ (i.e., X is degenerate at 0).

Example 1.2. Show that when n is large and p is small, the $\text{BIN}(n,p)$ distribution may be approximated by a $\text{POI}(\lambda)$ distribution where $\lambda = np$.

Solution: Recall $e^z = \lim\limits_{n\to\infty}(1+z/n)^n$, $z \in \mathbb{R}$. Letting $X \sim \text{BIN}(n,p)$, we have

$$\begin{aligned}
\mathbb{P}(X = x) &= \binom{n}{x}p^x(1-p)^{n-x} \\
&= \frac{n(n-1)\cdots(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \frac{n}{n}\frac{n-1}{n}\cdots\frac{n-x+1}{n}\frac{\lambda^x}{x!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^x} \\
&\simeq (1)(1)\cdots(1)\frac{\lambda^x}{x!}\frac{e^{-\lambda}}{1} && \text{when } n \text{ is large} \\
&= \frac{e^{-\lambda}\lambda^x}{x!}
\end{aligned}$$

Continuous Random Variables

Continuous type: A rv X takes on a continuum of possible values (which is uncountable) with cdf

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y,$$

where $f(x)$ denotes the *probability density function* (pdf) of X, which is a non-negative real-valued function that satisfies

$$\mathbb{P}(X \in B) = \int_{x \in B} f(x)\,\mathrm{d}x,$$

where B is the set of real numbers (e.g., an interval).

Remarks:

(1) If $F(x)$ (or the tpf $\bar{F}(x) = 1 - F(x)$) is known, we can recover the pdf using the relation

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x) = F'(x) = -\bar{F}'(x),$$

which holds by the *Fundamental Theorem of Calculus*.

(2) When working with pdfs in general, it is usually not necessary to be precise about specifying whether a range of numbers includes the endpoints. This is quite different from the situation we encounter with discrete rvs. Throughout this course, however, we will adopt the convention of **not including** the endpoints when specifying the range of values for pdfs.

Continuous Distributions

Special Continuous Distributions:

1. **Uniform**: A rv X is *Uniform* on the real interval (a, b) (i.e., $X \sim \mathrm{U}(a, b)$) if it has pdf

$$f(x) = \frac{1}{b-a},\ a < x < b,$$

where $a, b \in \mathbb{R}$ with $a < b$.

Remark: The choice of name is because X takes on values in (a, b) with all subintervals of a fixed length being equally likely.

2. **Beta**: A rv X is *Beta* with parameters $m \in \mathbb{Z}^+$ and $n \in \mathbb{Z}^+$ (i.e., $X \sim \mathrm{Beta}(m, n)$) if it has pdf

$$f(x) = \frac{(m+n-1)!}{(m-1)!(n-1)!} x^{m-1}(1-x)^{n-1},\ 0 < x < 1.$$

Remark: A $\mathrm{Beta}(1, 1)$ distribution simplifies to become the $\mathrm{U}(0, 1)$ distribution.

3. **Erlang**: A rv X is *Erlang* with parameters $n \in \mathbb{Z}^+$ and $\lambda > 0$ (i.e., $X \sim \text{Erlang}(n, \lambda)$) if it has pdf

$$f(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}, \; x > 0.$$

Remark: The Erlang(n, λ) distribution is actually a special case of the more general Gamma distribution in which n is extended to be any positive real number.

4. **Exponential**: A rv X is *Exponential* with parameter $\lambda > 0$ (i.e., $X \sim \text{EXP}(\lambda)$) if it has pdf

$$f(x) = \lambda e^{-\lambda x}, \; x > 0.$$

Remark: An Erlang$(1, \lambda)$ distribution actually simplifies to become the EXP(λ) distribution.

Expectation

Expectation: If $g(\,\cdot\,)$ is an arbitrary real-valued function, then

$$\mathbb{E}\big[g(X)\big] = \begin{cases} \sum_x g(x)p(x) & \text{, if } X \text{ is a discrete rv,} \\ \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x & \text{, if } X \text{ is a continuous rv.} \end{cases}$$

Special choices of $g(\,\cdot\,)$:

1. $g(X) = X^n$, $n \in \mathbb{N} \implies \mathbb{E}\big[g(X)\big] = \mathbb{E}[X^n]$ is the n^{th} moment of X. In general, moments serve to describe the shape of a distribution. If $n = 0$, then $\mathbb{E}[X^0] = 1$. If $n = 1$, then $\mathbb{E}[X] = \mu_X$ is the *mean* of X.
2. $g(X) = \big(X - \mathbb{E}[X]\big)^2 \implies \mathbb{E}\big[g(X)\big] = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big]$ is the *variance* of X. Note that

$$\text{Var}(X) = \sigma_X^2 = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big] = \mathbb{E}[X^2] - \mathbb{E}[X]^2,$$

or equivalently

$$\sigma_X^2 = \mathbb{E}\big[X(X-1)\big] + \mathbb{E}[X] - \mathbb{E}[X]^2.$$

Related to this quantity, the *standard deviation* of X is $\sqrt{\text{Var}(X)} = \sigma_X$.

3. $g(X) = aX + b$, $a, b \in \mathbb{R}$ (i.e., $g(X)$ is a linear function of X). Note that

$$\begin{aligned} \mu_{aX+b} &= \mathbb{E}[aX+b] = a\mu_X + b, \\ \sigma_{aX+b}^2 &= \text{Var}(aX+b) = a^2\sigma_X^2, \\ \sigma_{aX+b} &= \sqrt{\text{Var}(aX+b)} = |a|\sigma_X. \end{aligned}$$

Moment Generating Function

4. $g(X) = e^{tX}$, $t \in \mathbb{R} \implies \mathbb{E}\big[g(X)\big] = \mathbb{E}[e^{tX}]$ is the *moment generating function* (mgf) of X. This quantity is a function of t and is denoted by

$$\phi_X(t) = \mathbb{E}[e^{tX}].$$

First, $\phi_X(0) = \mathbb{E}[e^{0X}] = \mathbb{E}[1] = 1$. Moreover, making use of the linearity property of the expected value operator, note that

$$\begin{aligned}
\phi_X(t) &= \mathbb{E}[e^{tX}] \\
&= \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right] \\
&= \mathbb{E}\left[\frac{t^0X^0}{0!} + \frac{t^1X^1}{1!} + \frac{t^2X^2}{2!} + \cdots + \frac{t^nX^n}{n!} + \cdots\right] \\
&= \mathbb{E}[X^0]\frac{t^0}{0!} + \mathbb{E}[X]\frac{t^1}{1!} + \mathbb{E}[X^2]\frac{t^2}{2!} + \cdots + \mathbb{E}[X^n]\frac{t^n}{n!} + \cdots ,
\end{aligned}$$

implying that the n^{th} moment of X is simply the coefficient of $t^n/n!$ in the above series expansion.

We have: $\phi_X(t) = \mathbb{E}[t^X] = \mathbb{E}[X^0]\frac{t^0}{0!} + \mathbb{E}[X]\frac{t^1}{1!} + \mathbb{E}[X^2]\frac{t^2}{2!} + \cdots + \mathbb{E}[X^n]\frac{t^n}{n!} + \cdots$.

Remarks:

(1) Given the mgf of X, we can extract its n^{th} moment via

$$\mathbb{E}[X^n] = \phi_X^{(n)}(0) = \frac{\mathrm{d}^n}{\mathrm{d}t^n}\phi_X(t)\bigg|_{t=0} = \lim_{t\to 0}\frac{\mathrm{d}^n}{\mathrm{d}t^n}\phi_X(t),\ n \in \mathbb{N}.$$

Note that the 0^{th} derivative of a function is simply the function itself.

(2) A mgf uniquely characterizes the probability distribution of a rv (i.e., there exists a one-to-one correspondence between the mgf and the pmf/pdf of a rv). In other words, if two rvs X and Y have the same mgf, then they must have the same probability distribution (which we denote by $X \sim Y$). Thus, by finding the mgf of a rv, one has indeed determined its probability distribution.

Example 1.3. Suppose that $X \sim \text{BIN}(n,p)$. Find the mgf of X and use it to find $\mathbb{E}[X]$ and $\text{Var}(X)$.

Solution: Recall the binomial series formula

$$(a+b)^m = \sum_{x=0}^{m}\binom{m}{x}a^x b^{m-x},\ a, b \in \mathbb{R},\ m \in \mathbb{N}.$$

Using this formula, we obtain

$$\begin{aligned}\phi_X(t) &= \mathbb{E}[e^{tX}] \\ &= \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x(1-p)^{n-x} \\ &= \sum_{x=0}^{n}\binom{n}{x}(pe^t)^x(1-p)^{n-x} \\ &= (pe^t+1-p)^n,\ t\in\mathbb{R}.\end{aligned}$$

Then,

$$\phi'_X(t) = n(pe^t+1-p)^{n-1}pe^t \quad \text{and} \quad Q''_X(t) = n(pe^t+1-p)^{n-1}pe^t + npe^t(n-1)(pe^t+1-p)^{n-2}pe^t.$$

Thus,

$$\mathbb{E}[X] = \phi'_X(0) = n(pe^0+1-p)^{n-1}pe^0 = np,$$
$$\text{Var}(X) = \mathbb{E}[X^2]-\mathbb{E}[X]^2 = \phi''_X(0) - n^2p^2 = np + np(n-1)p - n^2p^2 = np.$$

Joint Distributions

Joint Distributions: The following results are presented for the bivariate case mostly, but these ideas extend naturally to an arbitrary number of rvs.

Definition: The *joint cdf* of X and Y is

$$\begin{aligned}F(a,b) &= \mathbb{P}(X\le a, Y\le b) \\ &= \mathbb{P}\big(\{X\le a\}\cap\{Y\le b\}\big),\ a,b\in\mathbb{R}.\end{aligned}$$

Remark: If the joint cdf is known, then we can recover their marginal counterparts as follows:

$$F_X(a) = \mathbb{P}(X\le a) = F(a,\infty) = \lim_{b\to\infty}F(a,b),$$
$$F_Y(a) = \mathbb{P}(Y\le b) = F(\infty,b) = \lim_{a\to\infty}F(a,b).$$

Jointly Discrete Case:
Joint pmf:

$$p(x,y) = \mathbb{P}(X=x, Y=y)$$

Marginals:

$$p_X(x) = \mathbb{P}(X=x) = \sum_y p(x,y)$$
$$p_Y(y) = \mathbb{P}(Y=y) = \sum_x p(x,y)$$

Multinomial Distribution: Consider an experiment which is repeated $n\in\mathbb{Z}^+$ times, with one of $k\ge 2$ distinct outcomes possible each time. Let $p_1,p_2,\ldots,p_k$ denote the probabilities of the k types of outcomes (with $\sum_{i=1}^k p_i = 1$). If X_i, $i=1,2,\ldots,k$, counts the number of type-i outcomes to occur, then

$(X_1, X_2, \ldots, X_k)$ is *Multinomial* (i.e., $(X_1, X_2, \ldots, X_k) \sim \text{MN}(n, p_1, p_2, \ldots, p_k)$) with joint pmf

$$p(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}, \ x_i = 0, 1, \ldots, n \ \forall i \text{ and } \sum_{i=1}^{k} x_i = n$$

Remark: A $\text{MN}(n, p_1, 1 - p_1)$ distribution simplifies to become the $\text{BIN}(n, p_1)$ distribution.

Jointly Continuous Case:
Joint pdf: The joint pdf $f(x, y)$ is a non-negative real-valued function which enables one to calculate probabilities of the form

$$\mathbb{P}(X \in A, Y \in B) = \int_B \int_A f(x, y) \, \mathrm{d}x \, \mathrm{d}y = \int_A \int_B f(x, y) \, \mathrm{d}x \, \mathrm{d}y$$

where A and B are sets of real numbers (e.g., intervals). As a result,

$$F(a, b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x, y) \, \mathrm{d}x \, \mathrm{d}y = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y) \, \mathrm{d}y \, \mathrm{d}x$$

Marginals:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) \, \mathrm{d}y$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) \, \mathrm{d}x$$

Jointly Continuous Case:
Important Relationship:

$$f(x, y) = \frac{\partial^2}{\partial x \, \partial y} F(x, y)$$

Transformations: Let (X, Y) be jointly continuous with joint pdf $f(x, y0)$ and region of support $\mathcal{S}(X, Y)$. Suppose that the rvs V and W are given by $V = b_1(X, Y)$ and $W = b_2(X, Y)$, where the functions $v = b_1(x, y)$ and $w = b_2(x, y)$ defined a one-to-one transformation that maps the set $\mathcal{S}(X, Y)$ onto the set $\mathcal{S}(V, W)$. If x and y are expressed in terms of v and w (i.e., $x = h_1(v, w)$ and $y = h_2(v, w)$), then the joint pdf of V and W is given by

$$g(v, w) = \begin{cases} f\big(h_1(v, w), h_2(v, w)\big)|J|, & \text{if } (v, w) \in \mathcal{S}(V, W), \\ 0, & \text{elsewhere,} \end{cases}$$

where J is the *Jacobian* of the transformation given by

$$J = \frac{\partial x}{\partial v} \frac{\partial y}{\partial w} - \frac{\partial x}{\partial w} \frac{\partial y}{\partial v}.$$

Expectation

Expectation: If $g(\,\cdot\,,\,\cdot\,)$ denotes an arbitrary real-valued function, then

$$\mathbb{E}\big[g(X,Y)\big] = \begin{cases} \sum_x \sum_y g(x,y)p(x,y) & \text{, if } X \text{ and } Y \text{ are jointly discrete,} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,\mathrm{d}y\,\mathrm{d}x & \text{, if } X \text{ and } Y \text{ are jointly continuous.} \end{cases}$$

Remark: The order of summation/integration is irrelevant and can be interchanged.

Special choices of $g(\,\cdot\,)$:

1. $g(X,Y) = \big(X-\mathbb{E}[X]\big)\big(Y-\mathbb{E}[Y]\big) \implies \mathbb{E}\big[g(X,Y)\big] = \mathbb{E}\big[\big(X-\mathbb{E}[X]\big)\big(Y-\mathbb{E}[Y]\big)\big]$ is the *covariance* of X and Y. Note that

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

   and $\operatorname{Cov}(X,X) = \operatorname{Var}(X)$.

2. $g(X,Y) = aX + bY$, $a,b \in \mathbb{R}$ (i.e., $g(X,Y)$ is a linear combination of X and Y). Note that:

$$\begin{aligned} \mathbb{E}[aX+bY] &= a\,\mathbb{E}[X] + b\,\mathbb{E}[Y], \\ \operatorname{Var}(aX+bY) &= a^2\operatorname{Var}(X) + b^2\operatorname{Var}(Y) + 2ab\operatorname{Cov}(X,Y). \end{aligned}$$

3. $g(X,Y) = e^{sX+tY}$, $s,t \in \mathbb{R} \implies \mathbb{E}\big[g(X,Y)\big] = \mathbb{E}[e^{sX+tY}]$ is the joint mgf of X and Y. A joint mgf (denoted by $\phi(s,t)$) also uniquely characterizes a joint probability distribution and can be used to calculate joint moments of X and Y via the formula

$$\mathbb{E}[X^mY^n] = \phi^{(m,n)}(0,0) = \left(\frac{\partial^{m+n}}{\partial s^m\,\partial t^n}\phi(s,t)\right)_{s=0,t=0} = \lim_{s\to 0,t\to 0}\frac{\partial^{m+n}}{\partial s^m\,\partial t^n}\phi(s,t),\;\; m,n\in\mathbb{N}$$

Independence of Random Variables

Formal Definition: If X and Y are *independent* rvs if

$$\begin{aligned} F(a,b) &= \mathbb{P}(X\le a, Y\le b) \\ &= \mathbb{P}(X\le a)\,\mathbb{P}(Y\le b) \\ &= F_X(a)F_Y(b)\;\forall a,b\in\mathbb{R}. \end{aligned}$$

Equivalently, independence exists iff $p(x,y) = p_X(x)p_Y(y)$ (in the jointly discrete case) or $f(x,y) = f_X(x)f_Y(y)$ (in the jointly continuous case) $\forall x,y\in\mathbb{R}$.

Important Property: For arbitrary real-valued functions $g(\,\cdot\,)$ and $h(\,\cdot\,)$, if X and Y are independent, then

$$\mathbb{E}\big[g(X)h(Y)\big] = \mathbb{E}\big[g(X)\big]\,\mathbb{E}\big[h(Y)\big].$$

Remark: As a consequence of this property, $\operatorname{Cov}(X,Y)=0$ if X and Y are independent, implying that $\operatorname{Var}(aX+bY) = a^2\operatorname{Var}(X) + b^2\operatorname{Var}(Y)$. However, if $\operatorname{Cov}(X,Y)=0$, we cannot conclude that X and Y are independent (we can only say that X and Y are *uncorrelated*).

Example 1.4. Suppose that X and Y have joint pmf (and corresponding marginals) of the form

		y		
	$p(x,y)$	0	1	$p_X(x)$
x	0	0.2	0	0.2
	1	0	0.6	0.6
	2	0.2	0	0.2
	$p_Y(y)$	0.4	0.6	1

Show that $\text{Cov}(X,Y) = 0$ holds, but X and Y are not independent.

Solution: Recall that $\text{Cov}(X,X) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$. Note that

$$\begin{aligned}\mathbb{E}[XY] &= \sum_x \sum_y xyp(x,y)\\ &= (0)(0)(0.2) + (0)(1)(0) + (1)(0)(0) + (1)(1)(0.6) + (2)(0)(0.2) + (2)(1)(0)\\ &= 0.6,\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X] &= \sum_x xp_X(x) = (0)(0.2) + (1)(0.6) + (2)(0.2) = 1,\\ \mathbb{E}[Y] &= \sum_y yp_Y(y) = (0)(0.4) + (1)(0.6) = 0.6.\end{aligned}$$

Thus,

$$\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0.6 - (1)(0.6) = 0.$$

However, from the given table, it is clear that $p(2,0) = 0.2 \neq 0.08 = (0.2)(0.4) = p_X(2)p_Y(0)$. Thus, we conclude that while $\text{Cov}(X,Y) = 0$, X and Y are not independent.

Theorem 1.1. If $X_1, X_2, \ldots, X_n$ are independent rvs where $\phi_{X_i}(t)$ is the mgf of X_i, $i = 1, 2, \ldots, n$, then $T = \sum_{i=1}^n X_i$ has mgf $\phi_T(t) = \prod_{i=1}^n \phi_{X_i}(t)$.

Proof: Note that the mgf of T given by

$$\begin{aligned}\phi_T(t) &= \mathbb{E}[e^{tT}]\\ &= \mathbb{E}[e^{t(X_1+X_2+\cdots+X_n)}]\\ &= \mathbb{E}[e^{tX_1}e^{tX_2}\cdots e^{tX_n}]\\ &= \mathbb{E}[e^{tX_1}]\,\mathbb{E}[e^{tX_2}]\cdots\mathbb{E}[e^{tX_n}] && \text{by independence of } \{X_i\}_{i=1}^n\\ &= \phi_{X_1}(t)\phi_{X_2}(t)\cdots\phi_{X_n}(t)\\ &= \prod_{i=1}^n \phi_{X_i}(t).\end{aligned}$$

<u>Remarks</u>:

(1) Simply put, Theorem 1.1 states that the mgf of a sum of independent rvs is just the product of their individual mgfs.

(2) As a special case of the above result, note that $\phi_T(t) = \phi_{X_1}(t)^n$ if $X_1, X_2, \ldots, X_n$ is an independent and identically distributed (iid) sequence of rvs.

Example 1.5. Let $X_1, X_2, \dots, X_m$ be an independent sequence of rvs where $X_i \sim \text{BIN}(n_i, p)$, $i = 1, 2, \dots, m$. Find the distribution of $T = \sum_{i=1}^m X_i$.

Solution: Looking at the mgf of T, note that

$$\begin{aligned} \phi_T(t) &= \prod_{i=1}^m \phi_{X_i}(t) && \text{by Theorem 1.1} \\ &= \prod_{i=1}^m (pe^t + 1 - p)^{n_i} && \text{using the result of Example of 1.3} \\ &= (pe^t + 1 - p)^{\sum_{i=1}^m n_i}, \ t \in \mathbb{R}. \end{aligned}$$

By the mgf uniqueness property we recognize that $T = \sum_{i=1}^m X_i \sim \text{BIN}\big(\sum_{i=1}^m n_i, p\big)$.
Remark: As a special case of the above example, if $X_1, X_2, \dots, X_m$ are iid BERN(p) rvs, then $T = \sum_{i=1}^m X_i \sim \text{BIN}(m, p)$.

Convergence of Random Variables

Modes of Convergence: If X_n, $n \in \mathbb{Z}^+$, and X are rvs, then

1. $X_n \to X$ *in distribution* iff

$$\lim_{n\to\infty} \mathbb{P}(X_n \le x) = \mathbb{P}(X \le x), \ \forall x \in \mathbb{R} \text{ at which } \mathbb{P}(X \le x) \text{ is continuous,}$$

2. $X_n \to X$ *in probability*, iff $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(|X_n - X| > \varepsilon\big) = 0,$$

3. $X_n \to X$ *almost surely (a.s.)* iff

$$\mathbb{P}\big(\lim_{n\to\infty} X_n = X\big) = 1.$$

Remarks:

(1) In probability theory, an event is said to happen a.s. if it happens with probability 1.

(2) The following implications hold true in general:

$$X_n \to X \text{ a.s.} \implies X_n \to X \text{ in probability} \implies X_n \to X \text{ in distribution.}$$

Strong Law of Large Numbers

Strong Law of Large Numbers (SLLN): If $X_1, X_2, \dots, X_n$ is an iid sequence of rvs with common mean μ and $\mathbb{E}\big[|X_1|\big] < \infty$, then

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu \text{ a.s. as } n \to \infty.$$

Remark: The SLLN is one of the most important results in probability and statistics, indicating that the sample mean will, with probability 1, converge to the true mean of the underlying distribution as the sample size approaches infinity. In other words, if the same experiment or study is repeated independently many times, the average of the results of the trials must be close to the mean. The result gets closer to the mean as the number of trials is increased.

Chapter 2

Conditional Distributions and Conditional Expectation

Week 2
15th to 22nd September

2.1 Definitions and Construction

Jointly Discrete Case

Formulation: If X_1 and X_2 are both discrete rvs with joint pmf $p(x_1, x_2)$ and marginal pmfs $p_1(x_1)$ and $p_2(x_2)$, respectively, then the conditional distribution of X_1 given $X_2 = x_2$, denoted by $X_1 \mid (X_2 = x_2)$, is defined via its *conditional pmf*

$$p_{1|2}(x_1 \mid x_2) = \mathbb{P}(X_1 = x_1 \mid X_2 = x_2) = \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2)}{\mathbb{P}(X_2 = x_2)} = \frac{p(x_1, x_2)}{p_2(x_2)},$$

provided that $p_2(x_2) > 0$. Similarly, the conditional distribution of $X_2 \mid (X_1 = x_1)$ is defined via its conditional pmf

$$p_{2|1}(x_2 \mid x_1) = \mathbb{P}(X_2 = x_2 \mid X_1 = x_1) = \frac{p(x_1, x_2)}{p_1(x_1)}, \text{ provided that } p_1(x_1) > 0.$$

Remarks:

(1) If X_1 and X_2 are independent, then $p(x_1, x_2) = p_1(x_1)p_2(x_2)\ \forall x_1, x_2 \in \mathbb{R}$, and so $p_{1|2}(x_1 \mid x_2) = p_1(x_1)$ and $p_{2|1}(x_2 \mid x_1) = p_2(x_2)$.

(2) These ideas extend beyond the simple bivariate case naturally. For example, suppose that X_1, X_2, and X_3 are discrete rvs. We can define the conditional distribution of (X_1, X_2) given $X_3 = x_3$ via its conditional pmf as follows:

$$p_{12|3}(x_1, x_2 \mid x_3) = \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3)}{\mathbb{P}(X_3 = x_3)} = \frac{p(x_1, x_2, x_3)}{p_3(x_3)},$$

provided that $p_3(x_3) > 0$. Alternatively, we can define the conditional distribution of X_2 given $(X_1 = x_1, X_3 = x_3)$ via its conditional pmf given by

$$p_{2|13}(x_2 \mid x_1, x_3) = \frac{p(x_1, x_2, x_3)}{p_{13}(x_1, x_3)}, \text{ provided that } p_{13}(x_1, x_3) > 0,$$

where $p_{13}(x_1, x_3)$ is the joint pmf of X_1 and X_3.

Conditional Expectation: The *conditional mean* of $X_1 \mid (X_2 = x_2)$ is

$$\mathbb{E}[X_1 \mid X_2 = x_2] = \sum_{x_1} x_1 p_{1|2}(x_1 \mid x_2).$$

More generally, if $w(\,\cdot\,,\,\cdot\,)$, $h(\,\cdot\,)$, and $g(\,\cdot\,)$ are arbitrary real-valued functions, then

$$\mathbb{E}\big[w(X_1, X_2) \bigm| X_2 = x_2\big] = \mathbb{E}\big[w(X_1, x_2) \bigm| X_2 = x_2\big] = \sum_{x_1} w(x_1, x_2) p_{1|2}(x_1 \mid x_2)$$

and

$$\mathbb{E}\big[g(X_1)h(X_2) \bigm| X_2 = x_2\big] = \mathbb{E}\big[g(X_1)h(x_2) \bigm| X_2 = x_2\big] = h(x_2)\,\mathbb{E}\big[g(X_1) \bigm| X_2 = x_2\big].$$

As an immediate consequence, if $a, b \in \mathbb{R}$, then we obtain

$$\mathbb{E}\big[ag(X_1) + bh(X_1) \bigm| X_2 = x_2\big] = a\,\mathbb{E}\big[g(X_1) \bigm| X_2 = x_2\big] + b\,\mathbb{E}\big[h(X_1) \bigm| X_2 = x_2\big].$$

Furthermore, if we recall that $\mathbb{E}[X_1 + X_2] = \sum_{x_1} \sum_{x_2} (x_1 + x_2) p(x_1, x_2)$, then it correspondingly follows that

$$\begin{aligned}
\mathbb{E}[X_1 + X_2 \mid X_3 = x_3] &= \sum_{x_1}\sum_{x_2}(x_1 + x_2)p_{12|3}(x_1, x_2 \mid x_3) \\
&= \sum_{x_1}\sum_{x_2}(x_1 + x_2)\frac{p(x_1, x_2, x_3)}{p_3(x_3)} \\
&= \sum_{x_1}\sum_{x_2} x_1 \cdot \frac{p(x_1, x_2, x_3)}{p_3(x_3)} + \sum_{x_1}\sum_{x_2} x_2 \cdot \frac{p(x_1, x_2, x_3)}{p_3(x_3)} \\
&= \sum_{x_1} \frac{x_1}{p_3(x_3)} \sum_{x_2} p(x_1, x_2, x_3) + \sum_{x_2} \frac{x_2}{p_3(x_3)} \sum_{x_1} p(x_1, x_2, x_3) \\
&= \sum_{x_1} \frac{x_1}{p_3(x_3)} p_{13}(x_1, x_3) + \sum_{x_2} \frac{x_2}{p_3(x_3)} p_{23}(x_2, x_3) \\
&= \sum_{x_1} x_1 p_{1|3}(x_1 \mid x_3) + \sum_{x_2} x_2 p_{2|3}(x_2 \mid x_3) \\
&= \mathbb{E}[X_1 \mid X_3 = x_3] + \mathbb{E}[X_2 \mid X_3 = x_3].
\end{aligned}$$

We have: $\mathbb{E}[X_1 + X_2 \mid X_3 = x_3] = \mathbb{E}[X_1 \mid X_3 = x_3] + \mathbb{E}[X_2 \mid X_3 = x_3]$. In other words, the conditional expected value is also a **linear** operator. In fact, more generally, if $a_i \in \mathbb{R}$, $i = 1, 2, \dots, n$, then the same essential approach can be used to show that

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i \,\middle|\, Y = y\right] = \sum_{i=1}^{n} a_i \, \mathbb{E}[X_i \mid Y = y].$$

Conditional Variance: If we take $g(X_1) = \big(X_1 - \mathbb{E}[X_1 \mid X_2 = x_2]\big)^2$, then

$$\mathbb{E}\big[g(X_1) \mid X_2 = x_2\big] = \mathbb{E}\Big[\big(X_1 - \mathbb{E}[X_1 \mid X_2 = x_2]\big)^2 \,\Big|\, X_2 = x_2\Big] = \mathrm{Var}(X_1 \mid X_2 = x_2)$$

is the *conditional variance* of $X_1 \mid (X_2 = x_2)$.

As with the calculation of variance, the following result provides an alternative (and often times preferred) way to calculate $\mathrm{Var}(X_1 \mid X_2 = x_2)$.

Theorem 2.1. $\mathrm{Var}(X_1 \mid X_2 = x_2) = \mathbb{E}[X_1^2 \mid X_2 = x_2] - \mathbb{E}[X_1 \mid X_2 = x_2]^2$.

Proof:

$$\begin{aligned}
\mathrm{Var}(X_1 \mid X_2 = x_2) &= \mathbb{E}\Big[\big(X_1 - \mathbb{E}[X_1 \mid X_2 = x_2]\big)^2 \,\Big|\, X_2 = x_2\Big] \\
&= \mathbb{E}\big[X_1^2 - 2X_1 \, \mathbb{E}[X_1 \mid X_2 = x_2] + \mathbb{E}[X_1 \mid X_2 = x_2]^2 \,\big|\, X_2 = x_2\big] \\
&= \mathbb{E}[X_1^2] - 2\,\mathbb{E}[X_1 \mid X_2 = x_2]^2 + \mathbb{E}[X_1 \mid X_2 = x_2]^2 \\
&= \mathbb{E}[X_1^2 \mid X_2 = x_2] - \mathbb{E}[X_1 \mid X_2 = x_2]^2
\end{aligned}$$

Example 2.1. Suppose that X_1 and X_2 are discrete rvs having joint pmf of the form

$$p(x_1, x_2) = \begin{cases} 1/5 & \text{, if } x_1 = 1 \text{ and } x_2 = 0, \\ 2/15 & \text{, if } x_1 = 0 \text{ and } x_2 = 1, \\ 1/15 & \text{, if } x_1 = 1 \text{ and } x_2 = 2, \\ 1/5 & \text{, if } x_1 = 2 \text{ and } x_2 = 0, \\ 2/5 & \text{, if } x_1 = 1 \text{ and } x_2 = 1, \\ 0 & \text{, otherwise.} \end{cases}$$

Find the conditional distribution of $X_1 \mid (X_2 = 1)$. Also, calculate $\mathbb{E}[X_1 \mid X_2 = 1]$ and $\text{Var}(X_1 \mid X_2 = 1)$.

Solution: Note that for problems of this nature, it often helps to create a table summarizing the information:

			x_2		
	$p(x_1, x_2)$	0	1	2	$p_1(x_1)$
x_1	0	0	$2/15$	0	$2/15$
	1	$1/5$	$2/5$	$1/15$	$2/3$
	2	$1/5$	0	0	$1/5$
	$p_2(x_2)$	$2/5$	$8/15$	$1/15$	1

Then,

- $p_{1|2}(0 \mid 1) = \mathbb{P}(X_1 = 0 \mid X_2 = 1) = (2/15)/(8/15) = 1/4$, and
- $p_{1|2}(1 \mid 1) = \mathbb{P}(X_1 = 1 \mid X_2 = 1) = (2/5)/(8/15) = 3/4$.

Thus, the conditional pmf of $X_1 \mid (X_2 = 1)$ can be represented as follows:

x_1	0	1
$p_{1\|2}(x_1 \mid 1)$	$1/4$	$3/4$

Note that $X_1 \mid (X_2 = 1) \sim$ BERN$(3/4)$. Thus, $\mathbb{E}[X_1 \mid X_2 = 1] = 3/4$ and $\text{Var}(X_1 \mid X_2 = 1) = 3/4(1 - 3/4) = 3/16$.

Example 2.2. For $i = 1, 2$, suppose that $X_i \sim$ BIN(n_i, p) where X_1 and X_2 are independent. Find the conditional distribution of X_1 given $X_1 + X_2 = m$.

Solution: We want to find the conditional pmf of $X_1 \mid (Y = m)$, where $Y = X_1 + X_2$. Let this conditional pmf be denoted by $p_{X_1|Y}(x_1 \mid m) = \mathbb{P}(X_1 = x_1 \mid Y = m)$. Recall from Example 1.5 that

$X_1 + X_2 \sim$ BIN$(n_1 + n_2, p)$.

$$\begin{aligned}
p_{X_1|Y}(x_1 \mid m) &= \frac{\mathbb{P}(X_1 = x_1, Y = m)}{\mathbb{P}(Y = m)} \\
&= \frac{\mathbb{P}(X_1 = x_1, X_1 + X_2 = m)}{\mathbb{P}(X_1 + X_2 = m)} \\
&= \frac{\mathbb{P}(X_1 = x_1, X_2 = m - x_1)}{\binom{n_1+n_2}{m} p^m (1-p)^{n_1+n_2-m}} \\
&= \frac{p_1(x_1) p_2(m - x_1)}{\binom{n_1+n_2}{m} p^m (1-p)^{n_1+n_2-m}} \\
&= \frac{\binom{n_1}{x_1} p^{x_1} (1-p)^{n_1-x_1} \binom{n_2}{m-x_1} p^{m-x_1} (1-p)^{n_2-(m-x_1)}}{\binom{n_1+n_2}{m} p^m (1-p)^{n_1+n_2-m}}
\end{aligned}$$

provided that $0 \le x_1 \le n_1$, and $0 \le m - x_1 \le n_2$ (i.e., $m - n_2 \le x \le m$). Simplifying,

$$p_{X_1|Y}(x_1 \mid m) = \frac{\binom{n_1}{x_1}\binom{n_2}{m-x_1}}{\binom{n_1+n_2}{m}},$$

for $x_1 = \max\{0, m - n_2\}, \ldots, \min\{n_1, m\}$.

Remark: Looking at the conditional pmf we just obtained, we recognize that $X_1 \mid (X_1 + X_2 = m) \sim$ HG$(n_1 + n_2, n_1, m)$. The result that $X_1 \mid (X_1 + X_2 = m)$ has a hypergeometric distribution should not be all that surprising. Consider the sequence of $n_1 + n_2$ Bernoulli trials represented visually as follows:

$$\underbrace{\boxed{1}\quad\boxed{2}\quad\cdots\quad\boxed{n_1}\quad\boxed{1}\quad\boxed{2}\quad\cdots\quad\boxed{n_2}}_{n_1 + n_2 \text{ trials}}$$

Of these $n_1 + n_2$ trials in which m of them were known to be successes, we want x_1 successes to have occurred among the first n_1 trials (thereby implying that $m - x_1$ successes are obtained during the final n_2 trials). Since any of these trials were equally likely to be a success (i.e., the same success probability p is assumed), the desired result ends up being the obtained hypergeometric probability.

Example 2.3. Let $X_1, X_2, \ldots, X_m$ be independent rvs where $X_i \sim$ POI(λ_i), $i = 1, 2, \ldots, m$. Define $Y = \sum_{i=1}^m X_i$. Find the conditional distribution of $X_j \mid (Y = n)$.

Solution: We are interested in the conditional pmf of $X_j \mid (Y = n)$, to be denoted by

$$\begin{aligned}
p_{X_j|Y}(x_j \mid n) &= \mathbb{P}(X_j = x_j \mid Y = n) \\
&= \frac{\mathbb{P}(X_j = x_j, Y = n)}{\mathbb{P}(Y = n)} \\
&= \frac{\mathbb{P}\Big(X_j = x_j, \sum_{i=1}^m X_i = n\Big)}{\mathbb{P}(Y = n)}
\end{aligned}$$

First, we investigate the numerator:

$$\begin{aligned}\mathbb{P}\bigg(X_j = x_j, \sum_{i=1}^{m} X_i = n\bigg) &= \mathbb{P}\bigg(X_j = x_j, X_j + \sum_{i=1, i\neq j}^{m} X_i = n\bigg) \\ &= \mathbb{P}\bigg(X_j = x_j, \sum_{i=1, i\neq j}^{m} X_i = n - x_j\bigg) \\ &= \mathbb{P}(X_j = x_j)\, \mathbb{P}\bigg(\sum_{i=1, i\neq j}^{m} X_i = n - x_j\bigg)\end{aligned}$$

where the last equality follows due to the independence of $\{X_i\}_{i=1}^{m}$. We are given that $X_j \sim$ POI(λ_j). Due to the result of Exercise 1.1, it follows that

$$\sum_{i=1, i\neq j}^{m} X_i \sim \text{POI}\bigg(\sum_{i=1, i\neq j}^{m} \lambda_i\bigg).$$

By the same result, we also have that

$$Y = \sum_{i=1}^{m} X_i \sim \text{POI}\bigg(\sum_{i=1}^{m} \lambda_i\bigg).$$

Therefore,

$$p_{X_j|Y}(x_j \mid n) = \frac{\frac{e^{-\lambda_j}\lambda_j^{x_j}}{x_j!}\, \frac{e^{-\sum_{i=1,i\neq j}^{m}\lambda_i}(\sum_{i=1,i\neq j}^{m}\lambda_i)^{n-x_j}}{(n-x_j)}}{\frac{e^{-\sum_{i=1}^{m}\lambda_i}(\sum_{i=1}^{m}\lambda_i)^{n}}{n!}}$$

provided that $x_j \geq 0$ and $n - x_j \geq 0$ which implies $0 \leq x_j \leq n$. Thus,

$$\begin{aligned}p_{X_j|Y}(x_j \mid n) &= \binom{n}{x_j}\frac{\lambda_j^{x_j}(\lambda_Y - \lambda_j)^{n-x_j}}{\lambda_Y^{n}} \\ &= \binom{n}{x_j}\bigg(\frac{\lambda_j}{\lambda_Y}\bigg)^{x_j}\bigg(1 - \frac{\lambda_j}{\lambda_Y}\bigg)^{n-x_j},\ x_j = 0, 1, \ldots, n\end{aligned}$$

where $\lambda_Y = \sum_{i=1}^{m} \lambda_i$ and note that $\lambda_Y^{x_j}\lambda_Y^{n-x_j} = \lambda_Y$. We see that

$$X_j \mid (Y = n) \sim \text{BIN}\bigg(n, \frac{\lambda_j}{\sum_{i=1}^{m}\lambda_i}\bigg).$$

Example 2.4. Suppose that $X \sim$ POI(λ) and $Y \mid (X = x) \sim$ BIN(x, p). Find the conditional distribution of $X \mid (Y = y)$.

Solution: We want to calculate the conditional pmf of $X \mid (Y = y)$, to be denoted by

$$p_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}.$$

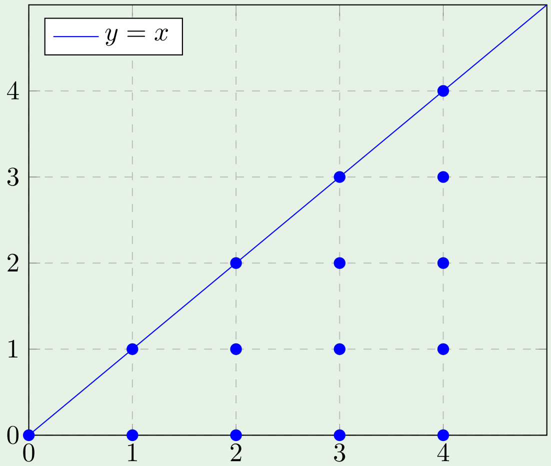
First, note that

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)},$$

which implies that

$$\mathbb{P}(X=x,Y=y)=\mathbb{P}(Y=y\mid X=x)\,\mathbb{P}(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}\binom{x}{y}p^y(1-p)^{x-y},$$

for $x=0,1,2,\ldots$ and $y=0,1,\ldots,x$. Note that the range of y depends on the values of x. A graphical display of the region is given below:



We may rewrite this region with the range of x depending on the values of y. Specifically, note that $x=0,1,2,\ldots$ and $y=0,1,\ldots,x$ is equivalent to $y=0,1,2,\ldots$ and $x=y,y+1,y+2,\ldots$. We use this alternative region to find the marginal pmf of Y.

$$\begin{aligned}
\mathbb{P}(Y=y)&=\sum_x \mathbb{P}(X=x,Y=y)\\
&=\sum_{x=y}^{\infty}e^{-\lambda}\frac{\lambda^x}{x!}\binom{x}{y}p^y(1-p)^{x-y}\\
&=\sum_{x=y}^{\infty}e^{-\lambda}\frac{\lambda^x}{x!}\frac{x!}{y!(x-y)!}p^y(1-p)^{x-y}\\
&=\frac{e^{-\lambda}}{y!}p^y\sum_{x=y}^{\infty}\frac{\lambda^x(1-p)^{x-y}}{(x-y)!}\lambda^{-y}\lambda^y\\
&=\frac{e^{-\lambda}(\lambda p)^y}{y!}\sum_{x=y}^{\infty}\frac{\big(\lambda(1-p)\big)^{x-y}}{(x-y)!} && \text{let } z=x-y\\
&=\frac{e^{-\lambda}(\lambda p)^y}{y!}e^{\lambda(1-p)}\\
&=\frac{e^{-\lambda p}(\lambda p)^y}{y!} && y=0,1,2,\ldots
\end{aligned}$$

In fact, $Y\sim$ POI(λp). Therefore,

$$\begin{aligned}
p_{X|Y}(x\mid y)&=\frac{\frac{e^{-\lambda}\lambda^x}{x!}\frac{x!}{y!(x-y)!}p^y(1-p)^{x-y}}{\frac{e^{-\lambda p}(\lambda p)^y}{y!}}\\
&=\frac{e^{-\lambda(1-p)}\big(\lambda(1-p)\big)^{x-y}}{(x-y)!},
\end{aligned}$$

for $x=y,y+1,\ldots$.

Remark: The above conditional pmf is recognized as that of a **shifted** Poisson distribution (y units to the right). Specifically, we have that

$$X \mid (Y = y) \sim W + y$$

where $W \sim \text{POI}\big(\lambda(1-p)\big)$.

Formulation: In the jointly discrete case, it was natural to define:

$$p_{X|Y}(x \mid y) = \mathbb{P}(X = x \,|\, Y = y) = \mathbb{P}(X = x, Y = y)/\,\mathbb{P}(Y = y).$$

Strictly speaking, this no longer makes sense in a continuous context since $f(x, y) \neq \mathbb{P}(X = x, Y = y)$ and $f_Y(y) \neq \mathbb{P}(Y = y)$. However, for small positive values of $\mathrm{d}y$ (as the figure below shows), $\mathbb{P}(y \leq Y \leq y + \mathrm{d}y) \approx f_Y(y)\,\mathrm{d}y$.

Formally,

$$f_Y(y) = \lim_{\mathrm{d}y \to 0} \frac{\mathbb{P}(y \leq Y \leq y + \mathrm{d}y)}{\mathrm{d}y}.$$

Similarly,

$$f(x, y) = \lim_{\mathrm{d}x \to 0, \mathrm{d}y \to 0} \frac{\mathbb{P}(x \leq X \leq x + \mathrm{d}x, y \leq Y \leq y + \mathrm{d}y)}{\mathrm{d}x\,\mathrm{d}y},$$

which implies that $\mathbb{P}(x \leq X \leq x + \mathrm{d}x, y \leq Y \leq y + \mathrm{d}y) \approx f(x, y)\,\mathrm{d}x\,\mathrm{d}y$. For small positive values of $\mathrm{d}x$ and $\mathrm{d}y$, consider now

$$\begin{aligned}
\mathbb{P}(x \leq X \leq x + \mathrm{d}x \,|\, y \leq Y \leq y + \mathrm{d}y) &= \frac{\mathbb{P}(x \leq X \leq x + \mathrm{d}x \,|\, y \leq Y \leq y + \mathrm{d}y)}{\mathbb{P}(y \leq Y \leq y + \mathrm{d}y)} \\
&\approx \frac{f(x, y)\,\mathrm{d}x\,\mathrm{d}y}{f_Y(y)\,\mathrm{d}y} \\
&= \frac{f(x, y)}{f_Y(y)}\,\mathrm{d}x.
\end{aligned}$$

As a result, we formally define the *conditional pdf* of X given $Y = y$ (again to be denoted by $X \mid (Y = y)$) as

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)} = \lim_{\mathrm{d}x \to 0, \mathrm{d}y \to 0} \frac{\mathbb{P}(x \leq X \leq x + \mathrm{d}x, y \leq Y \leq y + \mathrm{d}y)}{\mathrm{d}x}.$$

Remark: In the jointly continuous case, the conditional probability of an event of the form $\{a \leq X \leq b\}$ given $Y = y$ would be calculated as

$$\mathbb{P}(a \leq X \leq b \,|\, Y = y) = \int_a^b f_{X|Y}(x \mid y)\,\mathrm{d}x = \frac{\int_a^b f(x, y)\,\mathrm{d}x}{f_Y(y)},$$

which we can also express as

$$\mathbb{P}(a \leq X \leq b \mid Y = y) = \frac{\int_a^b f(x, y)\,\mathrm{d}x}{\int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x}.$$

In other words, we could view this as a way of assigning probability to an event $\{a \leq X \leq b\}$ over a "slice," $Y = y$, of the (joint) region of support for the pair of rvs X and Y.

Example 2.5. Suppose that the joint pdf of X and Y is given by

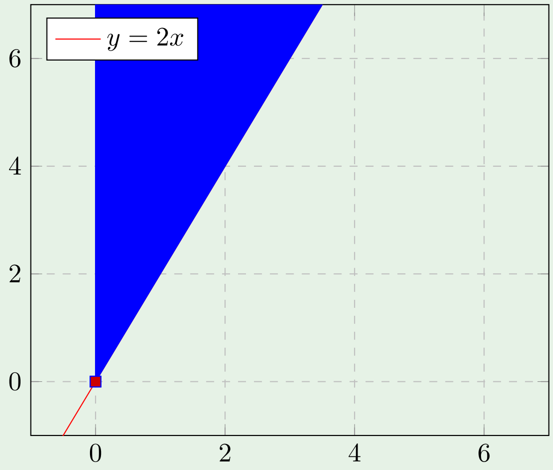
$$f(x, y) = \begin{cases} 5e^{-3x-y}, & \text{if } 0 \leq 2x \leq y < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

Determine the conditional distribution of $Y \mid (X = x)$ where $0 \leq x < \infty$.

Solution: We wish to find the conditional pdf of $Y \mid (X = x)$ given by

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)}$$

The region of support for this joint distribution looks like:



For $0 < x < \infty$:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \\ &= \int_{2x}^{\infty} 5e^{-3x-y}\,\mathrm{d}y \\ &= \left[5e^{-3x}(-e^{-y})\right]_{y=2x}^{y=\infty} \\ &= 5e^{-3x}e^{-2x} \\ &= 5e^{-5x} \end{aligned}$$

Note that $X \sim \text{EXP}(5)$. Finally, we get:

$$f_{Y|X}(y \mid x) = \frac{5e^{-3x-y}}{5e^{-5x}} = e^{-y+2x}, \; y > 2x.$$

Remark: The conditional pdf of $Y \mid (X = x)$ is recognized as that of a *shifted exponential distribution* ($2x$ units to the right). Specifically, we have that $Y \mid (X = x) \sim W + 2x$, where $W \sim \text{EXP}(1)$.

Conditional Expectation: If X and Y are jointly continuous rvs and $g(\,\cdot\,)$ is an arbitrary real-valued function, then the *conditional expectation* of $g(X)$ given $Y = y$ is

$$\mathbb{E}[g(X) \mid Y = y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x \mid y)\,\mathrm{d}x,$$

and so the conditional mean of $X \mid (Y = y)$ is given by

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x.$$

Example 2.6. Suppose that the joint pdf of X and Y is given by

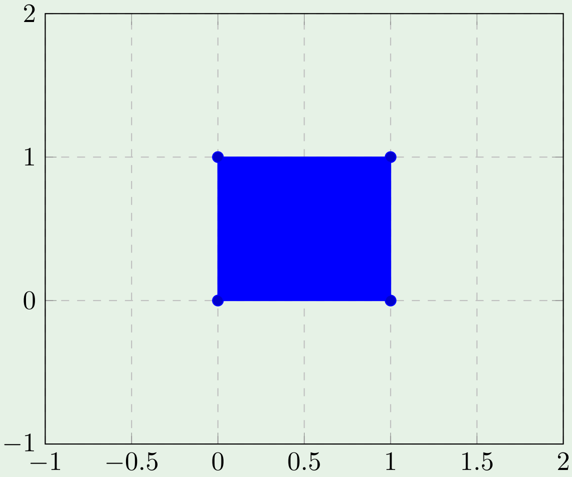
$$f(x,y) = \begin{cases} \frac{12}{5}x(2-x-y), & \text{if } 0 < x < 1 \text{ and } 0 < y < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the conditional distribution of X given $Y = y$ where $0 < y < 1$, and use it to calculate its conditional mean.

Solution: Using our earlier theory, we wish to find the conditional pdf of $X \mid (Y = y)$ given by

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}.$$

The region of support for this joint distribution of X and Y look like:



For $0 < y < 1$,

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x \\ &= \int_0^1 \frac{12}{5}x(2-x-y)\,\mathrm{d}x \\ &= \frac{12}{5}\int_0^1 (2x - x^2 - xy)\,\mathrm{d}x \\ &= \frac{12}{5}\left[x^2 - \frac{x^3}{3} - \frac{x^2 y}{3}\right]_{x=0}^{x=1} \\ &= \frac{12}{5}\left(1 - \frac{1}{3} - \frac{y}{2}\right) \\ &= \frac{2(4-3y)}{5} \end{aligned}$$

You can verify this by integrating $f_Y(y)$ over the support of Y (to get 1). Thus,

$$f_{X|Y}(x \mid y) = \frac{12/5x(2-x-y)}{2/5(4-3y)} = \frac{6x(2-x-y)}{4-3y}, \; 0 < x < 1$$

The conditional mean of X given $Y = y$ is:

$$\begin{aligned}\mathbb{E}[X \mid Y = y] &= \int_0^1 x \frac{6x(2-x-y)}{4-3y}\,\mathrm{d}x \\ &= \frac{6}{4-3y}\int_0^1 (2x^2 - x^3 - x^2 y)\,\mathrm{d}x \\ &= \frac{6}{4-3y}\left[\frac{2x^3}{3} - \frac{x^4}{4} - \frac{x^3 y}{3}\right]_{x=0}^{x=1} \\ &= \frac{6}{4-3y}\left(\frac{2}{3} - \frac{1}{4} - \frac{y}{3}\right) \\ &= \frac{5-4y}{2(4-3y)}\end{aligned}$$

Conditional Variance: Likewise, as in the jointly discrete case, we can also consider the notion of conditional variance, which retains the same definition as before:

$$\operatorname{Var}(X \mid Y = y) = \mathbb{E}\Big[\big(X - \mathbb{E}[X \mid Y = y]\big)^2 \,\Big|\, Y = y\Big] = \mathbb{E}[X^2 \mid Y = y] - \mathbb{E}[X \mid Y = y]^2.$$

A fact that is becoming more and more evident is that conditional expectation inherits many of the properties from regular expectation. Moreover, the same properties concerning conditional expectation that held in the jointly discrete case continue to hold true in the jointly continuous case (as we are effectively replacing summation with integration).

Example 2.6. (*continued*) Calculate $\operatorname{Var}(X \mid Y = y)$ where $0 < y < 1$ and the joint pdf of X and Y is given by

$$f(x,y) = \begin{cases} \dfrac{12}{5}x(2-x-5), & \text{if } 0 < x < 1 \text{ and } 0 < y < 1, \\ 0, & \text{elsewhere.}\end{cases}$$

Solution: Our earlier results tell us that

$$\begin{aligned}\mathbb{E}[X^2 \mid Y = y] &= \int_0^1 x^2 \frac{6x(2-x-y)}{4-3y}\,\mathrm{d}x \\ &= \frac{6}{4-3y}\int_0^1 (2x^3 - x^4 - x^3 y)\,\mathrm{d}x \\ &= \frac{6}{4-3y}\left[\frac{x^4}{2} - \frac{x^5}{5} - \frac{x^4 y}{4}\right]_{x=0}^{x=1} \\ &= \frac{6}{4-3y}\left(\frac{1}{2} - \frac{1}{5} - \frac{y}{4}\right) \\ &= \frac{3(6-5y)}{10(4-3y)}.\end{aligned}$$

Therefore, this leads to

$$\begin{aligned}\operatorname{Var}(X \mid Y = y) &= \mathbb{E}[X^2 \mid Y = y] - \mathbb{E}[X \mid Y = y]^2 \\ &= \frac{3(6-5y)}{10(4-3y)} - \frac{(5-4y)^2}{4(4-3y)^2} \\ &= \frac{19 + 2y(5y-14)}{20(4-3y)^2}.\end{aligned}$$

Mixed Case

We can also consider conditional distributions where the rvs are neither jointly continuous nor jointly discrete. To consider such a situation, suppose X is a continuous rv having pdf $f_X(x)$ and Y is a discrete rv having pmf $p_Y(y)$.
If we focus on the conditional distribution of X given $Y = y$, then let us look at the following quantity:

$$\begin{aligned}\frac{\mathbb{P}(x \le X \le x + \mathrm{d}x \mid Y = y)}{\mathrm{d}x} &= \frac{\mathbb{P}(x \le X \le x + \mathrm{d}x, Y = y)}{\mathrm{d}x\,\mathbb{P}(Y = y)} \\ &= \frac{\mathbb{P}(x \le X \le x + \mathrm{d}x)\,\mathbb{P}(Y = y \mid x \le X \le x + \mathrm{d}x)}{\mathrm{d}x\,\mathbb{P}(Y = y)} \\ &= \frac{\mathbb{P}(Y = y \mid x \le X \le x + \mathrm{d}x)}{\mathbb{P}(Y = y)}\frac{\mathbb{P}(x \le X \le x + \mathrm{d}x)}{\mathrm{d}x},\end{aligned}$$

where $\mathrm{d}x$ is again, a small positive value.
By letting $\mathrm{d}x \to 0$, we can formally define the conditional pdf of $X \mid (Y = y)$ as follows:

$$\begin{aligned}f(x \mid y) &= \lim_{\mathrm{d}x \to 0} \frac{\mathbb{P}(x \le X \le x + \mathrm{d}x \mid Y = y)}{\mathrm{d}x} \\ &= \lim_{\mathrm{d}x \to 0} \frac{\mathbb{P}(Y = y \mid x \le X \le x + \mathrm{d}x)}{\mathbb{P}(Y = y)}\frac{\mathbb{P}(x \le X \le x + \mathrm{d}x)}{\mathrm{d}x} \\ &= \frac{\mathbb{P}(Y = y \mid X = x)}{\mathbb{P}(Y = y)} f_X(x) \\ &= \frac{p(y \mid x) f_X(x)}{p_Y(y)},\end{aligned}$$

where $p(y \mid x) = \mathbb{P}(Y = y \mid X = x)$ is defined as the conditional pmf of $Y \mid (X = x)$. Note that since $f(x \mid y)$ is a pdf, it follows that

$$\int_{-\infty}^{\infty} f(x \mid y)\,\mathrm{d}x = 1 \implies p_Y(y) = \int_{-\infty}^{\infty} p(y \mid x) f_X(x)\,\mathrm{d}x.$$

Similarly, we can also write

$$p(y \mid x) = \frac{f(x \mid y) p_Y(y)}{f_X(x)}.$$

Since $p(y \mid x)$ is a pmf, we have that

$$\sum_y p(y \mid x) = 1 \implies f_X(x) = \sum_y f(x \mid y) p_Y(y).$$

Example 2.7. Suppose that $X \sim \text{U}(0, 1)$ and $Y \mid (X = x) \sim \text{BERN}(x)$. Find the conditional distribution of $X \mid (Y = y)$.

Solution: We wish to find the conditional pdf of $X \mid (Y = y)$ given by

$$f(x \mid y) = \frac{p(y \mid x) f_X(x)}{p_Y(y)}$$

Based on the given information, we have

$$\begin{gathered} f_X(x) = 1,\ 0 < x < 1, \\ p(y \mid x) = x^y(1 - x)^{1-y},\ y = 0, 1. \end{gathered}$$

For $y = 0, 1$, note that

$$\begin{aligned} p_Y(y) &= \int_{-\infty}^{\infty} p(y \mid x) f_X(x) \, \mathrm{d}x \\ &= \int_0^1 x^y (1-x)^{1-y} (1) \, \mathrm{d}x \end{aligned}$$

- For $y = 0 \implies p_Y(0) = \int_0^1 (1-x) \, \mathrm{d}x = \big[x - x^2/2\big]_{x=0}^{x=1} = 1/2$.
- For $y = 1 \implies p_Y(1) = \int_0^1 x \, \mathrm{d}x = \big[x^2/2\big]_{x=0}^{x=1} = 1/2$.

In other words, we have that

$$p_Y(y) = \frac{1}{2}, \ y = 0, 1 \implies Y \sim \text{BERN}\left(\frac{1}{2}\right)$$

Thus, for $y = 0, 1$, we ultimately obtain

$$f(x \mid y) = \frac{x^y (1-x)^{1-y} (1)}{1/2} = 2x^y (1-x)^{1-y}, \ 0 < x < 1.$$

2.2 Computing Expectation by Conditioning

An Important Observation

As before, let $g(\,\cdot\,)$ be an arbitrary real-valued function. In general, we recognize that $\mathbb{E}\big[g(X) \bigm| Y = y\big] = v(y)$, where $v(y)$ is some function of y. With this in mind, let us make the following definition:

$$\mathbb{E}\big[g(X) \bigm| Y\big] = \mathbb{E}\big[g(X) \bigm| Y = y\big]\Big|_{y=Y} = v(Y).$$

Functions of rvs are, once again, rvs themselves. Therefore, it makes sense to consider the expected value of $v(Y)$. In this regard, we would obtain:

$$\begin{aligned} \mathbb{E}\big[\mathbb{E}\big[g(X) \bigm| Y\big]\big] &= \mathbb{E}\big[v(Y)\big] \\ &= \begin{cases} \sum_y v(y) p_Y(y) & \text{, if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} v(y) f_Y(y) \, \mathrm{d}y & \text{, if } Y \text{ is continuous,} \end{cases} \\ &= \begin{cases} \sum_y \mathbb{E}\big[g(X) \bigm| Y = y\big] p_Y(y) & \text{, if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} \mathbb{E}\big[g(X) \bigm| Y = y\big] f_Y(y) \, \mathrm{d}y & \text{, if } Y \text{ is continuous.} \end{cases} \end{aligned}$$

Law of Total Expectation

The following important result is regarded as the *law of total expectation*.

Theorem 2.2. For rvs X and Y, $\mathbb{E}\big[g(X)\big] = \mathbb{E}\big[\mathbb{E}\big[g(X) \bigm| Y\big]\big]$.

Proof: Without loss of generality, assume that X and Y are jointly continuous rvs. From above, we have

$$\begin{aligned}\mathbb{E}\big[\mathbb{E}\big[g(X)\bigm| Y\big]\big] &= \int_{-\infty}^{\infty} \mathbb{E}\big[g(X)\bigm| Y=y\big] f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f_{X|Y}(x \mid y)\,\mathrm{d}x f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)\frac{f(x,y)}{f_Y(y)} f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} g(x)\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x \\ &= \mathbb{E}\big[g(X)\big]\end{aligned}$$

<u>Remark</u>: Using a similar method of proof, the result of Theorem 2.2 can naturally be extended as follows:

$$\mathbb{E}\big[g(X,Y)\big] = \mathbb{E}\big[\mathbb{E}\big[g(X,Y)\bigm| Y\big]\big].$$

The usefulness of the law of total expectation is well-demonstrated in the following example.

Example 2.8. Suppose that $X \sim \text{GEO}_t(p)$ with pmf $p_X(x) = (1-p)^{x-1}p$, $x = 1, 2, 3, \ldots$. Calculate $\mathbb{E}[X]$ and $\text{Var}(X)$ using the law of total expectation.

Solution: With $X \sim \text{GEO}_t(p)$, recall that X actually models the number of (independent) trials necessary to obtain the first success. Define:

$$Y = \begin{cases} 0 & \text{, if the 1}^{\text{st}}\text{ trial is a failure,} \\ 1 & \text{, if the 1}^{\text{st}}\text{ trial is a success.}\end{cases}$$

We observe that $Y \sim \text{BERN}(p)$, so that $p_Y(0) = 1-p$ and $p_Y(1) = p$.
<u>Note</u>:

- $X \mid (Y=1)$ is degenerate at 1 (i.e., X given $Y=1$ is equal to 1 with probability 1).
- $X \mid (Y=0)$ is equivalent in distribution $1+X$ (i.e., $X \mid (Y=0) \sim 1+X$).

By the law of total expectation, we obtain:

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\big[\mathbb{E}[X \mid Y]\big] \\ &= \sum_{y=0}^{1} \mathbb{E}[X \mid Y=y] p_Y(y) \\ &= (1-p)\,\mathbb{E}[X \mid Y=y] p_Y(y) \\ &= (1-p)\,\mathbb{E}[X \mid Y=0] + p\,\mathbb{E}[X \mid Y=1] \\ &= (1-p)\,\mathbb{E}[1+X] + p \\ &= (1-p) + (1-p)\,\mathbb{E}[X] + p \\ &= 1 + (1-p)\,\mathbb{E}[X],\end{aligned}$$

which implies that $(1-(1-p))\,\mathbb{E}[X] = 1$, or simply $\mathbb{E}[X] = 1/p$. Similarly, we use the law of total expectation to get

$$\begin{aligned}
\mathbb{E}[X^2] &= \mathbb{E}\big[\mathbb{E}[X^2 \mid Y]\big] \\
&= \sum_{y=0}^{1} \mathbb{E}[X^2 \mid Y = y] p_Y(y) \\
&= (1-p)\,\mathbb{E}[X^2 \mid Y = 0] + p\,\mathbb{E}[X^2 \mid Y = 1] \\
&= (1-p)\,\mathbb{E}\big[(1+X)^2\big] + p \\
&= (1-p)\big(\mathbb{E}[X^2] + 2\,\mathbb{E}[X] + 1\big) + p \\
&= 1 + (1-p)\,\mathbb{E}[X^2] + \frac{2(1-p)}{p},
\end{aligned}$$

which implies that

$$\big(1-(1-p)\big)\,\mathbb{E}[X] = \frac{p + 2(1-p)}{p}$$

or simply

$$\mathbb{E}[X^2] = \frac{p+2-2p}{p^2} = \frac{2-p}{p^2}$$

Finally,

$$\mathrm{Var}(X) = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}$$

Remarks:

(1) Note that the obtained mean and variance agree with known results. Moreover, the above procedure relied only on basic manipulations and did not involve any complicated sums or the differentiation of a mgf.

(2) As part of the above solution, we claimed that $X \mid (Y = 0) \sim Z$ where $Z = 1 + X$, and this implied that $\mathbb{E}[X^2 \mid Y = 0] = \mathbb{E}\big[(1+X)^2\big]$. To see why this holds true formally, consider first

$$p_{X|Y}(x \mid 0) = \mathbb{P}(X = x \mid Y = 0) = \frac{\mathbb{P}(X = x, Y = 0)}{\mathbb{P}(Y = 0)} = \frac{\mathbb{P}(X = x, Y = 0)}{1-p}.$$

Note that

$$\begin{aligned}
\mathbb{P}(X = x, Y = 0) &= \mathbb{P}(1^{\text{st}} \text{ trial is a failure and } x \text{ total trials needed to get } 1^{\text{st}} \text{ success}) \\
&= \mathbb{P}(1^{\text{st}} \text{ trial is a failure, next } x-2 \text{ trials are failures, and } x^{\text{th}} \text{ trial is a success}) \\
&= (1-p)(1-p)^{x-2}p \text{ due to independence of trials.}
\end{aligned}$$

Thus,

$$p_{X|Y}(x \mid 0) = \frac{(1-p)(1-p)^{x-2}p}{1-p} = (1-p)^{x-2}p,\ x = 2, 3, 4, \ldots.$$

On the other hand, note that

$$\begin{aligned}
p_Z(z) &= \mathbb{P}(Z = z) \\
&= \mathbb{P}(1 + X = z) \\
&= \mathbb{P}(X = z - 1) \\
&= (1-p)^{(z-1)-1}p \\
&= (1-p)^{z-2}p,\ z = 2, 3, 4, \ldots.
\end{aligned}$$

Since these two pmfs are identical, it follows that $X \mid (Y = 0) \sim Z$. As a further consequence, for an arbitrary real-valued function $g(\,\cdot\,)$, we must have that

$$\mathbb{E}\big[g(X) \bigm| Y = 0\big] = \mathbb{E}\big[g(Z)\big] = \mathbb{E}\big[g(1+X)\big].$$

Computing Variances by Conditioning

In recognizing that $\mathbb{E}\big[g(X) \,\big|\, Y = y\big]$ is a function of y, it similarly follows that $\mathrm{Var}(X \mid Y = y)$ is also a function of y. Therefore, we can make the following definition:

$$\mathrm{Var}(X \mid Y) = \mathrm{Var}(X \mid Y = y)\big|_{y=Y}.$$

Since $\mathrm{Var}(X \mid Y)$ is a function of Y, it is a rv as well, meaning that we could take its expected value. The following result, usually referred to as the *conditional variance formula*, provides a convenient way to calculate variance through the use of conditioning.

Theorem 2.3. For rvs X and Y, $\mathrm{Var}(X) = \mathbb{E}\big[\mathrm{Var}(X \mid Y)\big] + \mathrm{Var}\big(\mathbb{E}[X \mid Y]\big)$.

Proof: First, consider the term $\mathbb{E}\big[\mathrm{Var}(X \mid Y)\big]$. Since

$$\mathrm{Var}(X \mid Y = y) = \mathbb{E}[X^2 \mid Y = y] - \mathbb{E}[X \mid Y = y]^2,$$

it follows that

$$\mathrm{Var}(X \mid Y) = \mathbb{E}[X^2 \mid Y] - \mathbb{E}[X \mid Y]^2,$$

which yields (by Theorem 2.2)

$$\begin{aligned}\mathbb{E}\big[\mathrm{Var}(X \mid Y)\big] &= \mathbb{E}\big[\mathbb{E}[X^2 \mid Y] - \mathbb{E}[X \mid Y]^2\big] \\ &= \mathbb{E}\big[\mathbb{E}[X^2 \mid Y]\big] - \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big] \\ &= \mathbb{E}[X^2] - \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big].\end{aligned}$$

Next, recall

$$\mathrm{Var}(v(Y)) = \mathbb{E}\big[v(Y)^2\big] - \mathbb{E}\big[v(Y)\big]^2.$$

Applying Theorem 2.2 once more,

$$\begin{aligned}\mathrm{Var}\big(\mathbb{E}[X \mid Y]\big) &= \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big] - \mathbb{E}\big[\mathbb{E}[X \mid Y]\big]^2 \\ &= \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big] - \mathbb{E}[X]^2.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}\big[\mathrm{Var}(X \mid Y)\big] + \mathrm{Var}\big(\mathbb{E}[X \mid Y]\big) &= \mathbb{E}[X^2] - \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big] + \mathbb{E}\big[\mathbb{E}[X \mid Y]^2\big] - \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \mathrm{Var}(X).\end{aligned}$$

Example 2.9. Suppose that $\{X_i\}_{i=1}^{\infty}$ is an iid sequence of rvs with common mean μ and common variance σ^2. Let N be a discrete, non-negative integer-valued rv that is independent of each X_i. Find the mean and variance of $T = \sum_{i=1}^{N} X_i$ (referred to as a *random sum*).

Solution: By the law of total expectation,

$$\mathbb{E}[T] = \mathbb{E}\big[\mathbb{E}[T \mid N]\big].$$

Note that

$$\begin{aligned}
\mathbb{E}[T \mid N = n] &= \mathbb{E}\Big[\sum_{i=1}^{N} X_i \,\Big|\, N = n\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{n} X_i \,\Big|\, N = n\Big] \\
&= \sum_{i=1}^{n} \mathbb{E}[X_i \mid N = n] \\
&= \sum_{i=1}^{n} \mathbb{E}[X_i] && \text{since } N \text{ is independent of } \{X_i\}_{i=1}^{\infty} \\
&= n\mu.
\end{aligned}$$

Thus,

$$\mathbb{E}[T \mid N] = \mathbb{E}[T \mid N = n]\big|_{n=N} = N\mu,$$

and so $\mathbb{E}[T] = \mathbb{E}[N\mu] = \mu\,\mathbb{E}[N]$. To calculate $\operatorname{Var}(T)$, we employ Theorem 2.3 to obtain

$$\begin{aligned}
\operatorname{Var}(T) &= \mathbb{E}\big[\operatorname{Var}(T \mid N)\big] + \operatorname{Var}\big(\mathbb{E}[T \mid N]\big) \\
&= \mathbb{E}\big[\operatorname{Var}(T \mid N)\big] + \operatorname{Var}(N\mu) \\
&= \mathbb{E}\big[\operatorname{Var}(T \mid N)\big] + \mu^2 \operatorname{Var}(N).
\end{aligned}$$

Now,

$$\begin{aligned}
\operatorname{Var}(T \mid N = n) &= \operatorname{Var}\Big(\sum_{i=1}^{N} X_i \,\Big|\, N = n\Big) \\
&= \operatorname{Var}\Big(\sum_{i=1}^{n} X_i \,\Big|\, N = n\Big) \\
&= \operatorname{Var}\Big(\sum_{i=1}^{n} X_i\Big) && \text{since } N \text{ is independent of } \{X_i\}_{i=1}^{\infty} \\
&= \sum_{i=1}^{n} \operatorname{Var}(X_i) \\
&= n\sigma^2.
\end{aligned}$$

Thus, $\operatorname{Var}(T \mid N) = \operatorname{Var}(T \mid N = n)\big|_{N=n} = N\sigma^2$. Finally,

$$\begin{aligned}
\operatorname{Var}(T) &= \mathbb{E}[N\sigma^2] + \mu^2 \operatorname{Var}(N) \\
&= \sigma^2\,\mathbb{E}[N] + \mu^2 \operatorname{Var}(N).
\end{aligned}$$

WEEK 3
22nd to 29th September

2.3 Computing Probabilities by Conditioning

For any two rvs, recall that

$$\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big] = \begin{cases} \sum_y \mathbb{E}[X \mid Y = y] p_Y(y) & \text{, if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} \mathbb{E}[X \mid Y = y] f_Y(y) \, \mathrm{d}y & \text{, if } Y \text{ is continuous.} \end{cases} \tag{2.1}$$

Now suppose that A represents some event of interest, and we wish to determine $\mathbb{P}(A)$. Define an indicator rv X such that

$$X = \begin{cases} 0 & \text{, if event } A^c \text{ occurs,} \\ 1 & \text{, if event } A \text{ occurs.} \end{cases}$$

Clearly, $\mathbb{P}(X = 1) = \mathbb{P}(A)$ and $\mathbb{P}(X = 0) = 1 - \mathbb{P}(A)$, so that $X \sim \text{BERN}\big(\mathbb{P}(A)\big)$. Thus,

$$\begin{aligned} \mathbb{E}[X \mid Y = y] &= \sum_x x \, \mathbb{P}(X = x \mid Y = y) \\ &= 0 \, \mathbb{P}(X = 0 \mid Y = y) + 1 \, \mathbb{P}(X = 1 \mid Y = y) \\ &= \mathbb{P}(X = 1 \mid Y = y) \\ &= \mathbb{P}(A \mid Y = y). \end{aligned}$$

Therefore, (2.1) becomes

$$\mathbb{P}(A) = \begin{cases} \sum_y \mathbb{P}(A \mid Y = y) p_Y(y) & \text{, if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} \mathbb{P}(A \mid Y = y) f_Y(y) \, \mathrm{d}y & \text{, if } Y \text{ is continuous,} \end{cases} \tag{2.2}$$

which are analogues of the law of total probability. In other words, the expectation formula (2.1) can also be used to calculate probabilities of interest as indicated by (2.2).

Example 2.10. Suppose that X and Y are independent continuous rvs. Find an expression for $\mathbb{P}(X < Y)$.

Solution: With the event defined as $A = \{X < Y\}$, we apply (2.2) to get

$$\begin{aligned} \mathbb{P}(X < Y) &= \mathbb{P}(A) \\ &= \int_{-\infty}^{\infty} \mathbb{P}(A \mid Y = y) f_Y(y) \, \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X < Y \mid Y = y) f_Y(y) \, \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X < y \mid Y = y) f_Y(y) \, \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X < y) f_Y(y) \, \mathrm{d}y && \text{since } X \text{ and } Y \text{ are independent rvs} \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X \leq y) f_Y(y) \, \mathrm{d}y && \text{since } X \text{ is a continuous rv} \\ &= \int_{-\infty}^{\infty} F_X(y) f_Y(y) \, \mathrm{d}y \end{aligned} \tag{2.3}$$

Remark: If, in addition, X and Y are identically distributed, then the pdf $f_Y(y)$ is equal to $f_X(y)$ and the result of Example 2.10 simplifies to become

$$\begin{aligned}\mathbb{P}(X<Y) &= \int_{-\infty}^{\infty} F_X(y) f_X(y)\,\mathrm{d}y \\ &= \int_0^1 u\,\mathrm{d}u \qquad\qquad \text{where } u = F_X(y) \implies \frac{\mathrm{d}u}{\mathrm{d}y} = f_X(y) \implies \mathrm{d}u = f_X(y)\,\mathrm{d}y \\ &= \left[\frac{u^2}{2}\right]_{u=0}^{u=1} \\ &= \frac{1}{2},\end{aligned}$$

as one would expect.

Example 2.11. Suppose that $X \sim$ EXP(λ_1) and $Y \sim$ EXP(λ_2) are independent exponential rvs. Show that

$$\mathbb{P}(X<Y) = \frac{\lambda_1}{\lambda_1+\lambda_2}.$$

Solution: Since X and Y are both exponential rvs, it immediately follows that

$$\begin{aligned}f_Y(y) &= \lambda_2 e^{-\lambda_2 y},\ y>0, \\ F_X(y) &= \int_0^y \lambda_1 e^{-\lambda_1 x}\,\mathrm{d}x \\ &= \lambda_1\left[-\frac{1}{\lambda_1}e^{-\lambda_1 x}\right]_{x=0}^{x=y} \\ &= 1-e^{-\lambda_1 y},\ y \geq 0.\end{aligned}$$

Therefore, (2.3) becomes

$$\begin{aligned}\mathbb{P}(X<Y) &= \int_0^\infty (1-e^{-\lambda_1 y})\lambda_2 e^{-\lambda_2 y}\,\mathrm{d}y \\ &= \int_0^\infty \lambda_2 e^{-\lambda_2 y}\,\mathrm{d}y - \lambda_2 \int_0^\infty e^{-(\lambda_1+\lambda_2)y}\,\mathrm{d}y \\ &= 1 - \frac{\lambda_2}{(\lambda_1+\lambda_2)}\int_0^\infty (\lambda_1+\lambda_2)e^{-(\lambda_1+\lambda_2)y}\,\mathrm{d}y \\ &= 1-\frac{\lambda_2}{\lambda_1+\lambda_2} \\ &= \frac{\lambda_1}{\lambda_1+\lambda_2}.\end{aligned}$$

Remark: As a matter of interest, this particular result will be featured quite prominently in Chapter 4.

Example 2.12. Suppose W, X, and Y are independent continuous rvs on $(0,\infty)$. If $Z = X \mid (X<Y)$, then show that $(W,X) \mid (W<X<Y)$ and $(W,Z) \mid (W<Z)$ are identically distributed.

Solution: Let us first consider the joint conditional cdf of $(W, X) \mid (W < X < Y)$:

$$\begin{aligned} G(w,x) &= \mathbb{P}(W \le w, X \le x \mid W < X < Y) \\ &= \frac{\mathbb{P}(W \le w, X \le x, W < X < Y)}{\mathbb{P}(W < X < Y)} \\ &= \frac{\mathbb{P}(W \le w, X \le x, W < X, X < Y)}{\mathbb{P}(W < X, X < Y)}, \ w, x \ge 0. \end{aligned}$$

Conditioning on the rv X and noting that W, X, and Y are independent rvs, it follows that

$$\begin{aligned} \mathbb{P}(W < X, X < Y) &= \int_0^\infty \mathbb{P}(W < X, X < Y \mid X = s) f_X(s)\,\mathrm{d}s \\ &= \int_0^\infty \mathbb{P}(W < s, Y > s \mid X = s) f_X(s)\,\mathrm{d}s \\ &= \int_0^\infty \mathbb{P}(W < s, Y > s) f_X(s)\,\mathrm{d}s \\ &= \int_0^\infty \mathbb{P}(W < s)\,\mathbb{P}(Y > s) f_X(s)\,\mathrm{d}s \qquad (2.4) \end{aligned}$$

and

$$\begin{aligned} \mathbb{P}(W \le w, X \le x, W < X, X < Y) &= \int_0^\infty \mathbb{P}(W \le w, X \le x, W < X, X < Y \mid X = s) f_X(s)\,\mathrm{d}s \\ &= \int_0^\infty \mathbb{P}(W \le w, s \le x, W < s, Y > s \mid X = s) f_X(s)\,\mathrm{d}s \\ &= \int_0^\infty \mathbb{P}(W \le w, s \le x, W < s, Y > s) f_X(s)\,\mathrm{d}s \\ &= \int_0^x \mathbb{P}(W \le w, W < s, Y > s) f_X(s)\,\mathrm{d}s \\ &= \int_0^x \mathbb{P}\big(W \le \min\{w, s\}, Y > s\big) f_X(s)\,\mathrm{d}s \\ &= \int_0^x \mathbb{P}\big(W \le \min\{w, s\}\big)\,\mathbb{P}(Y > s) f_X(s)\,\mathrm{d}s \qquad (2.5) \end{aligned}$$

Next, consider the conditional rv $Z = X \mid (X < Y)$.

$$\begin{aligned} \mathbb{P}(Z \le z) &= \mathbb{P}(X \le z \mid X < Y) \\ &= \frac{\mathbb{P}(X \le z, X < Y)}{\mathbb{P}(X < Y)} \\ &= \frac{\int_0^\infty \mathbb{P}(X \le z, X < Y \mid X = s) f_X(s)\,\mathrm{d}s}{\mathbb{P}(X < Y)} \\ &= \frac{\int_0^\infty \mathbb{P}(s \le z, s < Y \mid X = s) f_X(s)\,\mathrm{d}s}{\mathbb{P}(X < Y)} \\ &= \frac{\int_0^\infty \mathbb{P}(s \le z, s < Y) f_X(s)\,\mathrm{d}s}{\mathbb{P}(X < Y)} \\ &= \frac{\int_0^z \mathbb{P}(Y > s) f_X(s)\,\mathrm{d}s}{\mathbb{P}(X < Y)} \end{aligned}$$

and so the pdf of Z is given by

$$\begin{aligned} h_Z(z) &= \frac{\mathrm{d}}{\mathrm{d}z}\,\mathbb{P}(Z \le z) \\ &= \frac{\frac{\mathrm{d}}{\mathrm{d}z}\int_0^z \mathbb{P}(Y > s) f_X(s)\,\mathrm{d}s}{\mathbb{P}(X < Y)} \\ &= \frac{\mathbb{P}(Y > z) f_X(z)}{\mathbb{P}(X < Y)},\ z > 0. \end{aligned}$$

Now, the joint conditional cdf of $(W, Z) \mid (W < Z)$ is given by

$$\mathbb{P}(W \le w, Z \le z \,|\, W < Z) = \frac{\mathbb{P}(W \le w, Z \le z, W < Z)}{\mathbb{P}(W < Z)},\ w, z \ge 0$$

Due to the independence of W with X and Y,

$$\begin{aligned} \mathbb{P}(W < Z) &= \int_0^\infty \mathbb{P}(W < Z \,|\, Z = s) h_Z(s)\,\mathrm{d}s \\ &= \int_0^\infty \mathbb{P}(W < z \,|\, Z = s) h_Z(s)\,\mathrm{d}s \\ &= \int_0^\infty \mathbb{P}(W < s) h_Z(s)\,\mathrm{d}s \\ &= \int_0^\infty \mathbb{P}(W < s)\frac{\mathbb{P}(Y > s) f_X(s)}{\mathbb{P}(X < Y)}\,\mathrm{d}s \\ &= \frac{\mathbb{P}(W < X, X < Y)}{\mathbb{P}(X < Y)} \qquad \text{from (2.4)} \end{aligned}$$

Next,

$$\begin{aligned} \mathbb{P}(W \le w, Z \le z, W < Z) &= \int_0^\infty \mathbb{P}(W \le w, Z \le z, W < Z \,|\, Z = s) h_Z(s)\,\mathrm{d}s \\ &= \int_0^\infty \mathbb{P}(W \le w, s \le z, W < s) h_Z(s)\,\mathrm{d}s \\ &= \int_0^z \mathbb{P}(W \le w, W < s) h_Z(s)\,\mathrm{d}s \\ &= \int_0^z \mathbb{P}(W \le \min\{w, s\})\frac{\mathbb{P}(Y > s) f_X(s)}{\mathbb{P}(X < Y)}\,\mathrm{d}s \\ &= \mathbb{P}(W \le w, X \le z, W < X, X < Y) \qquad \text{from (2.5)} \end{aligned}$$

Therefore, we ultimately obtain:

$$\mathbb{P}(W \le w, Z \le z, W < Z) = \frac{\mathbb{P}(W \le w, X \le z, W < X, X < Y)}{\mathbb{P}(W < X, X < Y)} = G(w, z),\ w, z \ge 0.$$

This implies that

$$(W, X) \mid (W < X < Y) \sim (W, Z) \mid (W < Z).$$

Remark: It can likewise be shown that if $V = X \mid (W < X)$, then $(X, Y) \mid (W < X < Y)$ and $(V, Y) \mid (V < Y)$ are identically distributed (left as an upcoming exercise).

2.4 Some Further Extensions

If you consider our treatment of the conditional expectation $\mathbb{E}[X \mid Y = y]$, then one detail you should notice is that this kind of expectation behaves *exactly* the same as the regular (i.e., unconditional) expectation *except* that all pmfs/pdfs used now are conditional on the event $Y = y$. In this sense, conditional expectations essentially satisfy all the properties of regular expectation. Thus, for an arbitrary real-valued function $g(\,\cdot\,)$, a corresponding analogue of

$$\mathbb{E}\big[g(X)\big] = \begin{cases} \sum_w \mathbb{E}\big[g(X) \bigm| W = w\big] p_W(w) & \text{, if } W \text{ is discrete,} \\ \int_{-\infty}^{\infty} \mathbb{E}\big[g(X) \bigm| W = w\big] f_W(w)\,\mathrm{d}w & \text{, if } W \text{ is continuous,} \end{cases}$$

would be

$$\mathbb{E}\big[g(X) \bigm| Y = y\big] = \begin{cases} \sum_w \mathbb{E}\big[g(X) \bigm| W = w, Y = y\big] p_{W|Y}(w \mid y) & \text{, if } W \text{ is discrete,} \\ \int_{-\infty}^{\infty} \mathbb{E}\big[g(X) \bigm| W = w, Y = y\big] f_{W|Y}(w \mid y)\,\mathrm{d}w & \text{, if } W \text{ is continuous.} \end{cases}$$

We remark that the above relation makes sense, since if we assume (without loss of generality) that X and Y are discrete rvs, then we obtain (in the case when W is discrete too):

$$\begin{aligned} \sum_w \mathbb{E}\big[g(X) \bigm| W = w, Y = y\big] p_{W|Y}(w \mid y) &= \sum_w \sum_x g(x) p_{X|WY}(x \mid w, y) p_{W|Y}(w \mid y) \\ &= \sum_w \sum_x g(x) \frac{p_{XWY}(x, w, y)}{p_{WY}(w, y)} \frac{p_{WY}(w, y)}{p_Y(y)} \\ &= \sum_x \frac{g(x)}{p_Y(y)} \sum_w p_{XWY}(x, w, y) \\ &= \sum_x g(x) \frac{p_{XY}(x, y)}{p_Y(y)} \\ &= \sum_x g(x) p_{X|Y}(x, y) \\ &= \mathbb{E}\big[g(X) \bigm| Y = y\big]. \end{aligned}$$

Similarly, if one introduces an event of interest A and defines

$$g(X) = \begin{cases} 0 & \text{, if event } A^c \text{ occurs,} \\ 1 & \text{, if event } A \text{ occurs,} \end{cases}$$

then we obtain

$$\mathbb{E}\big[A \bigm| Y = y\big] = \begin{cases} \sum_w \mathbb{E}\big[A \bigm| W = w, Y = y\big] p_{W|Y}(w \mid y) & \text{, if } W \text{ is discrete,} \\ \int_{-\infty}^{\infty} \mathbb{E}\big[A \bigm| W = w, Y = y\big] f_{W|Y}(w \mid y)\,\mathrm{d}w & \text{, if } W \text{ is continuous.} \end{cases}$$

Furthermore, if we now define

$$\mathbb{E}\big[g(X) \bigm| W, Y\big] = \mathbb{E}\big[g(X) \bigm| W = w, Y = y\big]\big|_{w=W, y=Y},$$

then the law of total expectation extends to become

$$\mathbb{E}\big[g(X)\big] = \mathbb{E}\big[\mathbb{E}[g(X) \mid Y]\big] = \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[g(X) \mid W, Y] \bigm| Y\big]\Big].$$

Example 2.13. Consider an experiment in which independent trials, each having success probability $p \in (0, 1)$, are performed until k consecutive successes are achieved where $k \in \mathbb{Z}^+$. Determine the expected number of trials needed to achieve k consecutive successes.

Solution: Let N_k represent the number of trials needed to get k consecutive successes. We wish to determine $\mathbb{E}[N_k]$. For $k = 1$, note that $N_1 \sim \text{GEO}_t(p)$, therefore $\mathbb{E}[N_k] = \frac{1}{p}$. For arbitrary $k \geq 2$, let us consider conditioning on the outcome of the first trial, represented by W, such that

$$W = \begin{cases} 0 & \text{, if first trial is a failure,} \\ 1 & \text{, if first trial is a success.} \end{cases}$$

Thus,

$$\begin{aligned} \mathbb{E}[N_k] &= \mathbb{E}\big[\mathbb{E}[N_k \mid W]\big] \\ &= \mathbb{P}(W = 0)\,\mathbb{E}[N_k \mid W = 0] + \mathbb{P}(W = 1)\,\mathbb{E}[N_k \mid W = 1] \\ &= (1 - p)\,\mathbb{E}[N_k \mid W = 0] + p\,\mathbb{E}[N_k \mid W = 1] \end{aligned}$$

Now, it is clear $N_k \mid (W = 0) \sim 1 + N_k$, but unfortunately we do not have a nice corresponding result for $N_k \mid (W = 1)$. It does not hold true that $N_k \mid (W = 0) \sim 1 + N_{k-1}$. What else can we try?
Idea: Let's try $\mathbb{E}[N_k] = \mathbb{E}\big[\mathbb{E}[N_k \mid N_{k-1}]\big]$, i.e., to get k in a row, we must first get $k - 1$ in a row. Define

$$Y \mid (N_{k-1} = n) = \begin{cases} 0 & \text{, if } (n+1)^{\text{th}} \text{ trial is a failure,} \\ 1 & \text{, if } (n+1)^{\text{th}} \text{ trial is a success.} \end{cases}$$

By independence of the trials,

$$\begin{aligned} \mathbb{P}(Y = 0 \mid N_{k-1} = n) &= 1 - p, \\ \mathbb{P}(Y = 1 \mid N_{k-1} = n) &= p. \end{aligned}$$

As a result, we get:

$$\begin{aligned} \mathbb{E}[N_k \mid N_{k-1} = n] &= \sum_{y=0}^{1} \mathbb{E}[N_k \mid N_{k-1} = n, Y = y]\,\mathbb{P}(Y = y \mid N_{k-1} = n) \\ &= (1 - p)\,\mathbb{E}[N_k \mid N_{k-1} = n, Y = 0] + p\,\mathbb{E}[N_k \mid N_{k-1} = n, Y = 1]. \end{aligned}$$

Note that $N_k \mid (N_{k-1} = n, Y = 0) \sim n + 1 + N_k$ (i.e., given that we know it took n trials to get $k - 1$ consecutive successes, and then on the next trial we got a failure, what happens?). Also, $N_k \mid (N_{k-1} = n, Y = 1)$ is equal to $n + 1$ with probability 1. Therefore,

$$\begin{aligned} \mathbb{E}[N_k \mid N_{k-1} = n] &= (1 - p)(n + 1 + \mathbb{E}[N_k]) + p(n + 1) \\ &= n + 1 + (1 - p)\,\mathbb{E}[N_k]. \end{aligned}$$

Therefore,

$$\mathbb{E}[N_k \mid N_{k-1}] = \mathbb{E}[N_k \mid N_{k-1} = n]\big|_{n = N_{k-1}} = N_{k-1} + 1 + (1 - p)\,\mathbb{E}[N_k].$$

Now, our whole idea was to apply $\mathbb{E}[N_k] = \mathbb{E}\big[\mathbb{E}[N_k \mid N_{k-1}]\big]$, and now we have the inner piece, so

$$\begin{aligned} \mathbb{E}[N_k] &= \mathbb{E}\big[N_{k-1} + 1 + (1 - p)\,\mathbb{E}[N_k]\big] \\ &= \mathbb{E}[N_{k-1}] + 1 + (1 - p)\,\mathbb{E}[N_k] \end{aligned}$$

Therefore,

$$\big(1 - (1 - p)\big)\,\mathbb{E}[N_k] = 1 + \mathbb{E}[N_{k-1}] \implies \mathbb{E}[N_k] = \frac{1}{p} + \frac{\mathbb{E}[N_{k-1}]}{p},\ k \geq 2,$$

which is a recursive equation for $\mathbb{E}[N_k]$. Take $k = 2$:

$$\mathbb{E}[N_2] = \frac{1}{p} + \frac{\mathbb{E}[N_1]}{p} = \frac{1}{p} + \frac{(1/p)}{p} = \frac{1}{p} + \frac{1}{p^2}.$$

Take $k = 3$:

$$\mathbb{E}[N_3] = \frac{1}{p} + \frac{\mathbb{E}[N_2]}{p} = \frac{1}{p} + \frac{1}{p^2} + \frac{1}{p^3}.$$

Take $k = 4$:

$$\mathbb{E}[N_4] = \frac{1}{p} + \frac{\mathbb{E}[N_3]}{p} = \frac{1}{p} + \frac{1}{p^2} + \frac{1}{p^3} + \frac{1}{p^4}.$$

Continuing inductively, we actually have

$$\mathbb{E}[N_k] = \frac{1}{p} + \frac{1}{p^2} + \cdots + \frac{1}{p^k},$$

which is a finite geometric series, therefore,

$$\mathbb{E}[N_k] = \frac{(1/p) - (1/p^{k+1})}{1 - (1/p)} = \frac{p^{-k} - 1}{1 - p},\ k \geq 2.$$

Actually, this holds true for $k \in \mathbb{Z}^+$ (try it).

Chapter 3

Discrete-time Markov Chains

WEEK 4
0929 to 6th October

3.1 Definitions and Basic Concepts

Stochastic Process

Definition: $\{X(t), t \in \mathcal{T}\}$ is called a *stochastic process* if $X(t)$ is a rv (or possibly a random vector) for any given $t \in \mathcal{T}$. $\mathcal{T}$ is referred to as the *index set* and is often interpreted in the context of time. As such, $X(t)$ is often called the *state of the process at time t*. We note that:

$$\text{Index set } \mathcal{T} \begin{cases} \text{can be a } \underline{\text{continuum}} \text{ of values such as } \mathcal{T} = \{t : t \geq 0\}, \\ \text{can be a set of } \underline{\text{discrete}} \text{ points such as } \mathcal{T} = \{t_0, t_1, t_2, \ldots\}. \end{cases}$$

Since there is a one-to-one correspondence between the sets $\mathcal{T} = \{t_0, t_1, t_2, \ldots\}$ and $\mathbb{N} = \{0, 1, 2, \ldots\}$, we will use $\mathcal{T} = \mathbb{N}$ as the general index set for a discrete-time stochastic process (unless otherwise stated). In other words, $\{X(n), n \in \mathbb{N}\}$ or $\{X_n, n \in \mathbb{N}\}$ will represent a general discrete-time stochastic process.

Discrete-time Stochastic Process

Some examples of a discrete-time stochastic process $\{X_n, n \in \mathbb{N}\}$ might include:

(1) X_n represents the outcome of the n^{th} toss of a die,

(2) X_n represents the price of a stock at the end of day n trading,

(3) X_n represents the maximum temperature in Waterloo during the n^{th} month,

(4) X_n represents the number of goals scored in game n by the varsity hockey team,

(5) X_n represents the number of STAT 333 students in class for the n^{th} lecture.

Discrete-time Markov Chain

Definition: A stochastic process $\{X_n, n \in \mathbb{N}\}$ is said to be a *discrete-time Markov chain* (DTMC) if the following two conditions hold true:

(1) For $n \in \mathbb{N}$, X_n is a <u>discrete</u> rv (i.e., the state space $\mathcal{S}$ of X_n is of discrete type).

(2) For $n \in \mathbb{N}$ and all states $x_0, x_1, \dots, x_{n+1} \in \mathcal{S}$, the *Markov property* must hold:

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_1 = x_1, X_0 = x_0) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n).$$

In mathematical terms, this property states that the conditional distribution of any *future* state X_{n+1} given the past states $X_0, X_1, \dots, X_{n-1}$ and the *present* state X_n is independent of the past states.

In a more informal way, the Markov property tells us, for a random process, that if we know the value taken by the process at a given time, we will not get any additional information about the future behaviour of the process by gathering more knowledge about the past.

Remarks:

(1) Unless otherwise stated, the state space $\mathcal{S}$ of a DTMC $\{X_n, n \in \mathbb{N}\}$ will be assumed to be $\mathbb{N}$.

(2) In general, the sequence of rvs $\{X_n\}_{n=0}^{\infty}$ are neither independent nor identically distributed.

(3) The Markov property does not require "full" information on the past to ensure independence. For example, consider the following conditional probability:

$$\begin{aligned}
&\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0)\\
&= \frac{\mathbb{P}(X_{n+1} = x_{n+1}, X_n = x_n, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0)}{\mathbb{P}(X_n = x_n, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0)},
\end{aligned}$$

which is "missing" the information for X_{n-1}.

However, note that:

$$\begin{aligned}
&\mathbb{P}(X_{n+1} = x_{n+1}, X_n = x_n, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0)\\
&= \sum_{x_{n-1}=0}^{\infty} \mathbb{P}(X_{n+1} = x_{n+1}, X_n = x_n, X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0)\\
&= \sum_{x_{n-1}=0}^{\infty} \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0)\\
&\quad \times \mathbb{P}(X_n = x_n, X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0)\\
&= \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n)\\
&\quad \times \sum_{x_{n-1}=0}^{\infty} \mathbb{P}(X_n = x_n, X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0) \text{ Markov property}\\
&= \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n)\, \mathbb{P}(X_n = x_n, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0).
\end{aligned}$$

We have:

$$\begin{aligned}
&\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0)\\
&= \frac{\mathbb{P}(X_{n+1} = x_{n+1}, X_n = x_n, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0)}{\mathbb{P}(X_n = x_n, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0)}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}(X_{n+1} = x_{n+1}, X_n = x_n, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0)\\
&= \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n)\, \mathbb{P}(X_n = x_n, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0).
\end{aligned}$$

Substituting this latter expression into the numerator of the top equation yields

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-2} = x_{n-2}, \dots, X_1 = x_1, X_0 = x_0) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n).$$

It is straightforward to extend the above result to any number of previous time points from $0, 1, \dots, n-1$ with "missing" information. This is the essence of the Markov property.

One-step Transition Probability Matrix

Definition: For any pair of states i and j, the *transition probability* from state i at time n to state j at time $n+1$ is given by

$$P_{n,i,j} = \mathbb{P}(X_{n+1} = j \mid X_n = i), \; n \in \mathbb{N}.$$

Let P_n be the associated matrix containing all these transition probabilities, referred to as the *one-step transition probability matrix* (TPM) from time n to time $n+1$. It looks like

$$P_n = [P_{n,i,j}] = \begin{array}{c|cccccc} & 0 & 1 & 2 & \cdots & j & \cdots \\ \hline 0 & P_{n,0,0} & P_{n,0,1} & P_{n,0,2} & \cdots & P_{n,0,j} & \cdots \\ 1 & P_{n,1,0} & P_{n,1,1} & P_{n,1,2} & \cdots & P_{n,1,j} & \cdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \\ i & P_{n,i,0} & P_{n,i,1} & P_{n,i,2} & \cdots & P_{n,i,j} & \cdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \end{array},$$

where, for convenience, the states of the DTMC are labelled along the margins of the matrix.
For each pair of states i and j, if $P_{n,i,j} = P_{i,j}$ $\forall n \in \mathbb{N}$, then we say that the DTMC is *stationary* or *homogenous*. In this situation, the one-step TPM becomes:

$$P = [P_{i,j}] = \begin{array}{c|cccccc} & 0 & 1 & 2 & \cdots & j & \cdots \\ \hline 0 & P_{0,0} & P_{0,1} & P_{0,2} & \cdots & P_{0,j} & \cdots \\ 1 & P_{1,0} & P_{1,1} & P_{1,2} & \cdots & P_{1,j} & \cdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \\ i & P_{i,0} & P_{i,1} & P_{i,2} & \cdots & P_{i,j} & \cdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \end{array}.$$

Remark: In STAT 333, we only consider stationary DTMCs. Moreover, from the construction of the TPM P, it is clear that $P_{i,j} \geq 0$ $\forall i, j \in \mathbb{N}$ and $\sum_{j=0}^{\infty} P_{i,j} = 1$ $\forall i \in \mathbb{N}$ (i.e., each row sum of P must be 1). Such a matrix whose elements are non-negative and whose row sums are equal to 1 is said to be *stochastic*.

Example 3.1. On a given day, the weather is either clear, overcast, or raining. If the weather is clear today, then it will be clear, overcast, or raining tomorrow with respective probabilities 0.6, 0.3, and 0.1. If the weather is overcast today, then it will be clear, overcast, or raining tomorrow with respective probabilities 0.2, 0.5, and 0.3. If the weather is raining today, then it will be clear, overcast, or raining tomorrow with respective probabilities 0.4, 0.2, and 0.4. Construct the underlying DTMC and determine its TPM.

Solution: Note that the weather tomorrow only depends on the weather today, implying that the Markov property holds true. Thus, letting X_n denote the state of the weather on the n^{th} day, $\{X_n, n \in \mathbb{N}\}$ is a three-state DTMC.
If we let state 0 correspond to clear weather, state 1 correspond to overcast, and state 2 correspond to raining, then the state space of the DTMC is $\mathcal{S} = \{0, 1, 2\}$ and its TPM is given by

$$P = \begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 0.6 & 0.3 & 0.1 \\ 1 & 0.2 & 0.5 & 0.3 \\ 2 & 0.4 & 0.2 & 0.4 \end{array}.$$

n-step Transition Probability Matrix

Definition: For any pair of states i and j, the *n-step transition probability* is given by

$$P_{i,j}^{(n)} = \mathbb{P}(X_{m+n} = j \mid X_m = i), \ m, n \in \mathbb{N}.$$

Due to the stationary assumption, this quantity is actually independent of m (which is why we do not include m in its notation). Thus, $P_{i,j}^{(n)} = \mathbb{P}(X_n = j \mid X_0 = i)$, $n \in \mathbb{N}$. Furthermore, it is evident that

$$P_{i,j}^{(0)} = \mathbb{P}(X_0 = j \mid X_0 = i) = \begin{cases} 0, & \text{if } i \neq j, \\ 1, & \text{if } i = j. \end{cases}$$

Similarly, let $P^{(n)} = \left[P_{i,j}^{(n)}\right]$ represent the associated n-step TPM. Clearly, when $n = 1$, $P^{(1)} = P$. When $n = 0$, $P^{(0)} = I$, where I represents the identity matrix. Just as with the one-step TPM, it follows that the row sums of $P^{(n)}$ must equal 1 as well.

Chapman-Kolmogorov Equations

For $n \in \mathbb{Z}^+$, let us consider

$$\begin{aligned} P_{i,j}^{(n)} &= \mathbb{P}(X_n = j \mid X_0 = i) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(X_n = j \mid X_{n-1} = k, X_0 = i)\, \mathbb{P}(X_{n-1} = k \mid X_0 = i) \\ &= \sum_{k=0}^{\infty} P_{i,k}^{(n-1)}\, \mathbb{P}(X_n = j \mid X_{n-1} = k, X_0 = i) \\ &= \sum_{k=0}^{\infty} P_{i,k}^{(n-1)}\, \mathbb{P}(X_n = j \mid X_{n-1} = k) \text{ due to the Markov property} \\ &= \sum_{k=0}^{\infty} P_{i,k}^{(n-1)} P_{k,j}. \end{aligned} \tag{3.1}$$

We have: $P_{i,j}^{(n)} = \sum_{k=0}^{\infty} P_{i,k}^{(n-1)} P_{k,j} \leftarrow$ (3.1).

Recall: If $A = [a_{i,j}]$, $B = [b_{i,j}]$, and $C = AB$ where $C = [c_{i,j}]$, then $c_{ij} = \sum_k a_{i,k} b_{k,j}$.
As a result, note that (3.1) implies that $P^{(n)} = P^{(n-1)}P$, $n \in \mathbb{Z}^+$. More generally, $P_{i,j}^{(n)}$ can be expressed as

$$P_{i,j}^{(n)} = \sum_{k=0}^{\infty} P_{i,k}^{(n)} P_{k,j}^{(n-m)} \ \forall i, j \in \mathbb{N} \text{ and } 0 \leq m \leq n,$$

which are referred to as the *Chapman-Kolmogorov* equations for a DTMC. In matrix form, this translates to

$$P^{(n)} = P^{(m)} P^{(n-m)}, \ 0 \leq m \leq n.$$

Coming back to $P^{(n)} = P^{(n-1)}P$, $n \in \mathbb{Z}^+$, let us look at a few values of n:

$$\begin{aligned} &\text{Take } n = 2 \implies P^{(2)} = P^{(1)}P = PP = P^2, \\ &\text{Take } n = 3 \implies P^{(3)} = P^{(2)}P = P^2 P = P^3, \\ &\text{Take } n = 4 \implies P^{(4)} = P^{(3)}P = P^3 P = P^4. \end{aligned}$$

Clearly, we see that

$$P^{(n)} = P^n,$$

and so the n-step TPM is simply the one-step TPM multiplied by itself n times.

Marginal pmf of X_n

For $n \in \mathbb{N}$, let us now introduce a particular row vector, which we will denote as either

$$\underline{\alpha}_n = (\alpha_{n,0}, \alpha_{n,1}, \ldots, \alpha_{n,k}, \ldots),$$

or

$$\underline{\alpha}_n = \begin{bmatrix} \alpha_{n,0} & \alpha_{n,1} & \cdots & \alpha_{n,k} & \cdots \end{bmatrix},$$

where $\alpha_{n,k} = \mathbb{P}(X_n = k)\ \forall k \in \mathbb{N}$. In other words, $\alpha_{n,k}$ represents the marginal pmf of X_n, $n \in \mathbb{N}$. As a consequence, it follows that $\sum_{k=0}^{\infty} \alpha_{n,k} = 1\ \forall n \in \mathbb{N}$.
If we focus on the case when $n = 0$, then $\underline{\alpha}_0$ is referred to as the *initial probability row vector* of the DTMC, or simply the *initial conditions* of the DTMC.
For $n \in \mathbb{Z}^+$, note that

$$\begin{aligned}
\alpha_{n,k} &= \mathbb{P}(X_n = k) \\
&= \sum_{i=0}^{\infty} \mathbb{P}(X_n = k \mid X_m = i)\, \mathbb{P}(X_m = i) \text{ where } 0 \le m \le n \\
&= \sum_{i=0}^{\infty} \alpha_{m,i}\, \mathbb{P}(X_{n-m} = k \mid X_0 = i) \text{ due to the stationary assumption} \\
&= \sum_{i=0}^{\infty} \alpha_{m,i} P_{i,k}^{(n-m)}.
\end{aligned}$$

In matrix form, the above relation implies that

$$\underline{\alpha}_n = \underline{\alpha}_m P^{(n-m)} = \underline{\alpha}_m P^{n-m},\ 0 \le m \le n,$$

which subsequently leads to

$$\underline{\alpha}_n = \underline{\alpha}_0 P^{(n)} = \underline{\alpha}_0 P^n,\ n \in \mathbb{N}.$$

Probabilities of Interest

Having knowledge of the initial conditions and the one-step transition probabilities, one can readily calculate various probabilities of possible interest such as

$$\begin{aligned}
&\mathbb{P}(X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0) \\
&= \mathbb{P}(X_0 = x_0)\, \mathbb{P}(X_1 = x_1 \mid X_0 = x_0)\, \mathbb{P}(X_2 = x_2 \mid X_1 = x_1, X_0 = x_0) \times \cdots \\
&\quad \times \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_0 = x_0) \\
&= \mathbb{P}(X_0 = x_0)\, \mathbb{P}(X_1 = x_1 \mid X_0 = x_0)\, \mathbb{P}(X_2 = x_2 \mid X_1 = x_1) \cdots \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}) \\
&= \alpha_{0,x_0} P_{x_0,x_1} P_{x_1,x_2} \cdots P_{x_{n-1},x_n},
\end{aligned}$$

where the second last equality follows due to the Markov property.

Similarly,

$$\begin{aligned}
&\mathbb{P}(X_{n+m} = x_{n+m}, X_{n+m-1} = x_{n+m-1}, \ldots, X_{n+1} = x_{n+1} \mid X_n = x_n) \\
&= \frac{\mathbb{P}(X_{n+m} = x_{n+m}, X_{n+m-1} = x_{n+m-1}, \ldots, X_{n+1} = x_{n+1}, X_n = x_n)}{\mathbb{P}(X_n = x_n)} \\
&= \tfrac{\mathbb{P}(X_n=x_n)\,\mathbb{P}(X_{n+1}=x_{n+1} \mid X_n=x_n)\cdots\mathbb{P}(X_{n+m}=x_{n+m} \mid X_{n+m-1}=x_{n+m-1}, X_{n+m-2}=x_{n+m-2}, \ldots, X_n=x_n)}{\mathbb{P}(X_n=x_n)} \\
&= P_{x_n,x_{n+1}} P_{x_{n+1},x_{n+2}} \cdots P_{x_{n+m-1},x_{n+m}}.
\end{aligned}$$

The key observation here is that the DTMC is *completely characterized* by its one-step TPM P and the initial conditions $\underline{\alpha}_0$.

Example 3.2. A particle moves along the states $\{0, 1, 2\}$ according to a DTMC whose TPM is given by

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0.5 & 0 & 0.5 \end{bmatrix} \end{array}.$$

Let X_n denote the position of the particle after the n^{th} move. Suppose that the particle is equally likely to start in any of the three positions.

(a) Calculate $\mathbb{P}(X_3 = 1 \mid X_0 = 0)$.

Solution: We wish to determine $P^{(3)}_{0,1}$. To get this, we proceed to calculate $P^{(3)} = P^3$. First,

$$P^{(2)} = P^2 = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0.5 & 0 & 0.5 \end{bmatrix} \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0.5 & 0 & 0.5 \end{bmatrix} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0.54 & 0.26 & 0.2 \\ 0.2 & 0.36 & 0.44 \\ 0.6 & 0.1 & 0.3 \end{bmatrix} \end{array}.$$

Then,

$$P^{(3)} = P^{(2)}P = \begin{bmatrix} 0.54 & 0.26 & 0.2 \\ 0.2 & 0.36 & 0.44 \\ 0.6 & 0.1 & 0.3 \end{bmatrix} \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0.5 & 0 & 0.5 \end{bmatrix} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0.478 & 0.264 & 0.258 \\ 0.36 & 0.256 & 0.384 \\ 0.57 & 0.18 & 0.25 \end{bmatrix} \end{array}.$$

Thus, $\mathbb{P}(X_3 = 1 \mid X_0 = 0) = P^{(3)}_{0,1} = 0.264$.

(b) Calculate $\mathbb{P}(X_4 = 2)$.

Solution: We wish to calculate $\alpha_{4,2} = \mathbb{P}(X_4 = 2)$. Note that

$$\begin{aligned}
\underline{\alpha}_4 &= \begin{bmatrix} \alpha_{4,0} & \alpha_{4,1} & \alpha_{4,2} \end{bmatrix} \\
&= \underline{\alpha}_0 P^{(4)} \\
&= \underline{\alpha}_0 P^{(3)} P \\
&= \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 0.478 & 0.264 & 0.258 \\ 0.36 & 0.256 & 0.384 \\ 0.57 & 0.18 & 0.25 \end{bmatrix} \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0.5 & 0 & 0.5 \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 0.4636 & 0.254 & 0.2824 \\ 0.444 & 0.2256 & 0.3304 \\ 0.524 & 0.222 & 0.254 \end{bmatrix} \\
&= \begin{bmatrix} 0.4772 & 0.233867 & 0.288933 \end{bmatrix}.
\end{aligned}$$

Thus, $\mathbb{P}(X_4 = 2) = 0.288933 \approx 0.289$.

(c) Calculate $\mathbb{P}(X_6 = 0, X_4 = 2)$.

Solution: We have

$$\begin{aligned}\mathbb{P}(X_6 = 0, X_4 = 2) &= \mathbb{P}(X_4 = 2)\,\mathbb{P}(X_6 = 0 \mid X_4 = 2)\\ &= \alpha_{4,2} P_{2,0}^{(2)}\\ &= (0.288933)(0.6)\\ &= 0.17336\\ &\approx 0.173.\end{aligned}$$

(d) Calculate $\mathbb{P}(X_9 = 0, X_7 = 2 \mid X_5 = 1, X_2 = 0)$.

Solution: We have

$$\begin{aligned}&\mathbb{P}(X_9 = 0, X_7 = 2 \mid X_5 = 1, X_2 = 0)\\ &= \mathbb{P}(X_7 = 2 \mid X_5 = 1, X_2 = 0)\,\mathbb{P}(X_9 = 0 \mid X_7 = 2, X_5 = 1, X_2 = 0)\\ &= \mathbb{P}(X_7 = 2 \mid X_5 = 1)\,\mathbb{P}(X_9 = 0 \mid X_7 = 2) \qquad \text{Markov property}\\ &= P_{1,2}^{(2)} P_{2,0}^{(2)}\\ &= (0.44)(0.6)\\ &= 0.264.\end{aligned}$$

Accessibility and Communication

With these basic results in place, we next consider the classification of states in a DTMC.

Definition: State j is *accessible* from state i (denoted by $i \to j$) if $\exists n \in \mathbb{N}$ such that $P_{i,j}^{(n)} > 0$. If states i and j are accessible from each other, then the states i and j *communicate* (denoted by $i \leftrightarrow j$). In other words, $i \leftrightarrow j$ iff $\exists m, n \in \mathbb{N}$ such that $P_{i,j}^{(n)} > 0$ and $P_{j,i}^{(m)} > 0$.

In terms of accessibility, note that the size of the components of P do not matter. All that matters is which are positive and which are 0. In particular, if state j is not accessible from state i, then $P_{i,j}^{(n)} = 0 \; \forall n \in \mathbb{N}$ and

$$\begin{aligned}&\mathbb{P}(\text{DTMC ever visits state } j \mid X_0 = i)\\ &= \mathbb{P}\big(\cup_{n=0}^{\infty}\{X_n = j\} \bigm| X_0 = i\big)\\ &\leq \sum_{n=0}^{\infty} \mathbb{P}(X_n = j \mid X_0 = i) \text{ due to Boole's inequality (see Exercise 1.1.1)}\\ &= \sum_{n=0}^{\infty} P_{i,j}^{(n)}\\ &= 0,\end{aligned}$$

implying that $\mathbb{P}(\text{DTMC ever visits state } j \mid X_0 = i) = 0$.

Equivalence Relation

The concept of communication forms what is known as an equivalence relation, satisfying the following criteria:

(i) **Reflexivity**: $i \leftrightarrow i$.

Clearly true since $P_{i,i}^{(0)} = 1 > 0$.

(ii) **Symmetry**: $i \leftrightarrow j \implies j \leftrightarrow i$.

This is obviously true by definition.

(iii) **Transitivity**: $i \leftrightarrow j$ and $j \leftrightarrow k \implies i \leftrightarrow k$.

To see this holds formally, we know that $\exists n \in \mathbb{N}$ such that $P_{i,j}^{(n)} > 0$. Also, $\exists m \in \mathbb{N}$ such that $P_{j,k}^{(m)} > 0$. Using the Chapman-Kolmogorov equations, we have that

$$P_{i,k}^{(n+m)} = \sum_{\ell=0}^{\infty} P_{i,\ell}^{(n)} P_{\ell,k}^{(m)} \geq P_{i,j}^{(n)} P_{j,k}^{(m)} > 0.$$

Therefore, $P_{i,k}^{(n+m)} > 0$, implying that $i \to k$. Using precisely the same logic, it is straightforward to show that $k \to i$. Thus, by definition, $i \leftrightarrow k$

Communication Classes

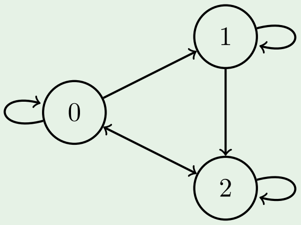
The fact that communication forms an equivalence relation allows us to *partition* all the states of a DTMC into various communication classes, so that within each class, all states communicate. However, if states i and j belong to *different classes*, then $i \leftrightarrow j$ is not true (i.e., at most one of $i \to j$ or $j \to i$ can be true).

Definition: A DTMC that has only one communication class is said to be *irreducible*. On the other hand, a DTMC is called *reducible* if there are two or more communication classes.

Example 3.2. (*continued*) What are the communication classes of the DTMC?

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 0 & \;\;1 & \;\;2 \end{array} \\ \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0.5 & 0 & 0.5 \end{bmatrix} \end{array}.$$

Solution: To answer questions of this nature, it is often useful to draw a state transition diagram.



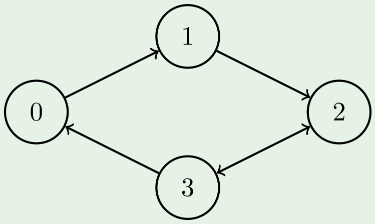
It is clear from this diagram that there is only one class for this DTMC, namely $\{0, 1, 2\}$. Therefore, this DTMC is irreducible.

Example 3.3. Consider a DTMC with TPM

$$P = \begin{array}{cc} & \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \end{array} & \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0.5 & 0 \end{bmatrix} \end{array}.$$

What are the communication classes of this DTMC?

Solution:



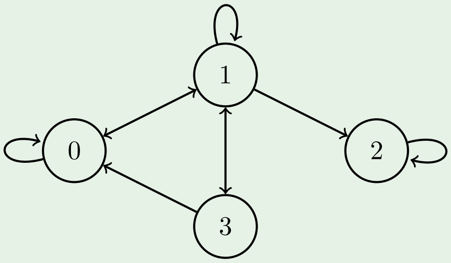
The above diagram indicates that there is only one communication class for this DTMC, namely $\{0, 1, 2, 3\}$. Therefore, this DTMC is irreducible.

Example 3.4. Consider a DTMC with TPM

$$P = \begin{array}{cc} & \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \end{array} & \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \\ 0 & 0 & 1 & 0 \\ \frac{3}{4} & \frac{1}{4} & 0 & 0 \end{bmatrix} \end{array}.$$

What are the communication classes of this DTMC?

Solution: State Transition Diagram



The above diagram indicates that the communication classes are $\{0, 1, 3\}$ and $\{2\}$. Thus, this DTMC is reducible.

Periodicity

Definition: The *period* of state i is given by $d(i) = \gcd\{n \in \mathbb{Z}^+ : P_{i,i}^{(n)} > 0\}$, where $\gcd\{\,\cdot\,\}$ denotes the greatest common divisor of a set of positive integers.

Remark: If $d(i) = 1$, then state i is said to be aperiodic. In fact, a DTMC is said to be *aperiodic* if $d(i) = 1 \ \forall i \in \mathbb{N}$. Furthermore, if $P_{i,i}^{(n)} = 0 \ \forall n \in \mathbb{Z}^+$, then we set $d(i) = \infty$.

Example 3.5. Consider a DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} \frac{1}{3} & 0 & 0 & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 1 & 0 \\ \frac{3}{4} & 0 & 0 & \frac{1}{4} \end{bmatrix} \end{array}.$$

Determine the communication classes of this DTMC and the period of each state.

Solution:

Example 3.3. (*continued*) Recall that $\{0, 1, 2, 3\}$ is the only communication class for the DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0.5 & 0 \end{bmatrix} \end{array}.$$

Determine the period of each state.

Solution:

Example 3.6. Consider the DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{array}.$$

Find the communication classes of this DTMC and determine the period of each state.

Solution:

The above examples illustrate some kinds of periodic behaviour that can be exhibited by DTMCs. However, we do observe that among the states within a given communication class, it seems as though the periodic behaviour is consistent. This is not a coincidence, as the next theorem indicates.

Theorem 3.1. If $i \leftrightarrow j$, then $d(i) = d(j)$.

Proof:

Example 3.7. Consider a DTMC with TPM

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \end{array}.$$

Find the communication classes of this DTMC and determine the period of each state.

Solution:

Remark: As the previous example demonstrates, it is still possible to observe aperiodic behaviour even though the main diagonal components of P are all zero. More generally, if $d(i) = k$, then this does not necessarily imply that $P_{i,i}^{(k)} > 0$ Instead, it implies that if the DTMC is in state i at time 0, then it is impossible to observe the DTMC in state i at time $n \in \mathbb{Z}^+$ if n is not a multiple of k (i.e., $P_{i,i}^{(n)} = 0$ for such n).

Appendix A

Useful Documents

A.1 List of Acronyms

iff	if and only if
gcd	greatest common divisor
rv	random variable
pmf	probability mass function
pdf	probability density function
cdf	cumulative distribution function
tpf	tail probability function
mgf	moment generating function
iid	independent and identically distributed
SLLN	Strong Law of Large Numbers
a.s.	almost surely
DTMC	discrete-time Markov chain
TPM	transition probability matrix
BLT	Basic Limit Theorem (for DTMCs)

A.2 Special Symbols

$\to$	approaches
$\implies$	implies
Ω	sample space for a probability model
$\mathbb{P}(\,\cdot\,)$	probability function
$\varnothing$	null event (or empty set)
$\cup$	union operator
$\cap$	intersection operator
A^c	complement of A
$\subseteq$	is a subset of
$\mathbb{R}$	set of all real numbers
$\mathbb{Z}$	set of all integers $\{0, \pm 1, \pm 2, \ldots\}$
$\mathbb{Z}^+$	set of positive integers $\{1, 2, \ldots\}$
$\mathbb{N}$	set of non-negative integers $\{0, 1, 2, \ldots\}$
$\mathcal{S}$	state space of a rv (or a DTMC)
$\approx$	approximately equal to
$\sim$	has the probability distribution of
$\mathbb{E}[\,\cdot\,]$	expected value operator
$\underline{a}$	row vector notation
$\underline{a}^\top$	column vector notation
$[A_{i,j}]$	matrix A with the elements of the form $A_{i,j}$
I	identity matrix (of appropriate dimension)
$\mathbf{0}$	zero matrix (of appropriate dimension)
$\underline{e}^\top$	vector of ones (of appropriate dimension)
$\mathcal{T}$	index set of a stochastic process
$n!$	n factorial
$\binom{n}{x}$	n choose x
$(n)_x$	n taken to x terms
$\delta_{i,j}$	1 if $i = j$, 0 if $i \neq j$ (Kronecker delta)
$\lvert x\rvert$	absolute value of x
$\lfloor x \rfloor$	greatest integer less than or equal to x
$\exp\{x\}$	exponential function e^x

A.3 Results for Some Fundamental Probability Distributions

Discrete Distribution	Probability Mass Function of X	Mean $\mathbb{E}[X]$	Variance $\text{Var}(X)$
DU(a,b)	$p(x) = \frac{1}{b-a+1}$, $x = a, a+1, \ldots, b$	$\frac{a+b}{2}$	$\frac{(b-a)(b-a+2)}{12}$
BIN(n,p)	$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, \ldots, n$	np	$np(1-p)$
BERN(p)	$p(x) = p^x (1-p)^{1-x}$, $x = 0, 1$	p	$p(1-p)$
HG(N,r,n)	$p(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$, $x = \max\{0, n-N+r\}, \ldots, \min\{n, r\}$	$\frac{nr}{N}$	$\frac{nr(N-r)(N-n)}{N^2(N-1)}$
POI(λ)	$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$	λ	λ
NB$_t(k,p)$	$p(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$, $x = k, k+1, k+2, \ldots$	$\frac{k}{p}$	$\frac{k(1-p)}{p^2}$
GEO$_t(p)$	$p(x) = (1-p)^{x-1} p$, $x = 1, 2, 3, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
NB$_f(k,p)$	$p(x) = \binom{x+k-1}{k-1} p^k (1-p)^x$, $x = 0, 1, 2, \ldots$	$\frac{k(1-p)}{p}$	$\frac{k(1-p)}{p^2}$
GEO$_f(p)$	$p(x) = (1-p)^x p$, $x = 0, 1, 2, \ldots$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$

Continuous Distribution	Probability Mass Function of X	Mean $\mathbb{E}[X]$	Variance $\text{Var}(X)$
U(a,b)	$f(x) = \frac{1}{b-a}$, $a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Beta(m,n)	$f(x) = \frac{(m+n-1)!}{(m-1)!(n-1)!} x^{m-1}(1-x)^{n-1}$, $0 < x < 1$	$\frac{m}{m+n}$	$\frac{mn}{(m+n)^2(m+n+1)}$
Erlang(n,λ)	$f(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}$, $x > 0$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$
EXP(λ)	$f(x) = \lambda e^{-\lambda x}$, $x > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

"DU"	stands for	*Discrete Uniform*	"BIN"	stands for	*Binomial*
"BERN"	stands for	*Bernoulli*	"HG"	stands for	*Hypergeometric*
"POI"	stands for	*Poisson*	"NB$_t$"	stands for	*Negative Binomial (for trials)*
"GEO$_t$"	stands for	*Geometric (for trials)*	"NB$_f$"	stands for	*Negative Binomial (for failures)*
"GEO$_f$"	stands for	*Geometric (for failures)*	"U"	stands for	*(Continuous) Uniform*
"EXP"	stands for	*Exponential*			