Answer (Ex. 5.2) —
We want $1-\alpha = 0.95$, and from the standard Normal Table we know that the corresponding $z_{\alpha/2} = 1.96$. Then we can get the right sample size n from the CLT implied Equation (61) in the lecture notes, which is,

$$n = \left(\sqrt{(V(X_1)}z_{\alpha/2})/\epsilon\right)^2 \ ,$$

as follows:

$$\begin{aligned} n &= \left(\sqrt{(V(X_1)}z_{\alpha/2})/\epsilon\right)^2 = \left((\sqrt{1/4}\times 1.96)/(1/10)\right)^2 \\ &= (((1/2)\times 1.96)/(1/10))^2 = (0.98\times 10)^2 = 9.8^2 = 96.04 \end{aligned}$$

Finally, by rounding 96.04 up to the next largest integer we need $n = 97$ measurements to meet the specifications of your boss (at least up to the approximation provided by the CLT).

Answer (Ex. 5.3) —
By CLT, $\frac{\sqrt{n}(\overline{X}_n - \mathbf{E}(X_1))}{\sqrt{\mathbf{V}(X_1)}} \rightsquigarrow Z \sim \text{Normal}(0,1)$. So we need to apply the "standardization" to both sides of the inequality that is defining the event of interest:

$$\{\overline{X}_n < 5.5\} \ ,$$

in order to find its probability $\mathbf{P}(\overline{X}_n < 5.5)$.

$$\begin{aligned} \mathbf{P}(\overline{X}_n < 5.5) &= P\left(\frac{\sqrt{n}(\overline{X}_n - \mathbf{E}(X_1))}{\sqrt{\mathbf{V}(X_1)}} < \frac{\sqrt{n}(5.5 - \mathbf{E}(X_1))}{\sqrt{\mathbf{V}(X_1)}}\right) \\ &\approxeq P\left(Z < \frac{\sqrt{n}(5.5-\lambda)}{\sqrt{\lambda}}\right) \qquad [\text{ since we know/assume that } \mathbf{E}(X_1) = \mathbf{V}(X_1) = \lambda] \\ &= P\left(Z < \frac{\sqrt{125}(5.5-5)}{\sqrt{5}}\right) \qquad [\text{Since, } \lambda = 5 \text{ and } n = 125 \text{ in this Example}] \\ &= \mathbf{P}(Z \leq 2.5) = \Phi(2.5) = 0.9938 \ . \end{aligned}$$

(source: Wasserman, *All of Statistics*, Springer, p. 78, 2003)

Answer (Ex. 5.4) — HINT: Use the LLN after finding the population mean of X_i.

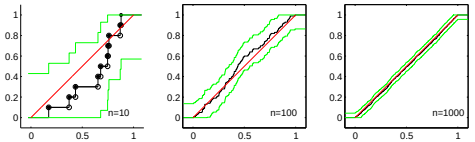
Answer (Ex. 5.5) — HINT: Use the CLT after finding the population mean and variance of X_i.

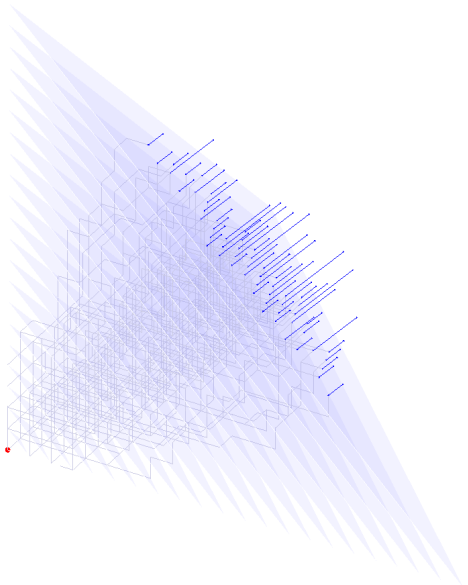
Probability Theory 1

Raazesh Sainudiin* and Dominic Lee†,
*Laboratory for Mathematical Statistical Experiments, Uppsala Centre, and
*Department of Mathematics, Uppsala University, Uppsala, Sweden
†School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

Version Date: October 22, 2019

This work was partially supported by NSF grant DMS-03-06497 and NSF/NIGMS grant DMS-02-01037.





Answer (Ex. 3.53) —

1.
$$\varphi_X(t) = E\left(e^{\imath tX}\right) = \sum_{x=0}^{\infty} e^{\imath tx}\frac{\lambda^x}{x!}e^{-\lambda} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{e^{\imath tx}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\left(\lambda e^{\imath t}\right)^x}{x!} = e^{-\lambda}e^{\lambda e^{\imath t}} = e^{\lambda e^{\imath t}-\lambda} \ .$$

The second-last equality above is using $e^{\alpha} = \sum_{x=0}^{\infty}\frac{\alpha^x}{x!}$ with $\alpha = \lambda e^{\imath t}$.

2. To find $V(X)$ using $\varphi_X(t)$ we need $E(X)$ and $E(X^2)$.
$$\begin{aligned} E(X) &= \frac{1}{\imath}\left[\frac{d}{dt}\varphi_X(t)\right]_{t=0} = \frac{1}{\imath}\left[\frac{d}{dt}e^{\lambda e^{\imath t}-\lambda}\right]_{t=0} = \frac{1}{\imath}\left[e^{\lambda e^{\imath t}-\lambda}\frac{d}{dt}\left(\lambda e^{\imath t}-\lambda\right)\right]_{t=0} \\ &= \frac{1}{\imath}\left[e^{\lambda e^{\imath t}-\lambda}\lambda\imath e^{\imath t}\right]_{t=0} = \frac{1}{\imath}\left(e^{\lambda-\lambda}\lambda\imath\right) = \frac{1}{\imath}\left(\lambda\imath\right) = \lambda \ . \end{aligned}$$

$$\begin{aligned} E(X^2) &= \frac{1}{\imath^2}\left[\frac{d^2}{dt^2}\varphi_X(t)\right]_{t=0} = \frac{1}{\imath^2}\left[\frac{d}{dt}\left(e^{\lambda e^{\imath t}-\lambda}\lambda\imath e^{\imath t}\right)\right]_{t=0} = \frac{1}{\imath^2}\left[\frac{d}{dt}\left(\lambda\imath e^{\lambda e^{\imath t}-\lambda+\imath t}\right)\right]_{t=0} \\ &= \frac{1}{\imath^2}\left[\lambda\imath e^{\lambda e^{\imath t}-\lambda+\imath t}\frac{d}{dt}\left(\lambda e^{\imath t}-\lambda+\imath t\right)\right]_{t=0} = \frac{1}{\imath^2}\left[\lambda\imath e^{\lambda e^{\imath t}-\lambda+\imath t}\left(\lambda\imath e^{\imath t}-0+\imath\right)\right]_{t=0} \\ &= \frac{1}{\imath^2}\left(\lambda\imath e^{\lambda e^{0}-\lambda+0}\left(\lambda\imath e^{0}+\imath\right)\right) = \frac{1}{\imath^2}\left(\lambda\imath\left(\lambda\imath+\imath\right)\right) = \frac{1}{\imath^2}\left(\lambda\imath^2\left(\lambda+1\right)\right) = \lambda^2+\lambda \ . \end{aligned}$$

Finally,
$$V(X) = E(X^2) - (E(X))^2 = \lambda^2+\lambda-\lambda^2 = \lambda \ .$$

Answer (Ex. 3.54) —
$$\varphi_W(t) = \varphi_{X+Y}(t) = \varphi_X(t)\times\varphi_Y(t) = e^{\lambda e^{\imath t}-\lambda}\times e^{\mu e^{\imath t}-\mu} = e^{\lambda e^{\imath t}-\lambda+\mu e^{\imath t}-\mu} = e^{(\lambda+\mu)e^{\imath t}-(\lambda+\mu)} \ .$$
So, W is a Poisson$(\lambda+\mu)$ RV. Thus the sum of two independent Poisson RVs is also a Poisson RV with parameter given by the sum of the parameters of the two RVs being added. The same idea generalizes to the sum of more than two Poisson RVs.

Answer (Ex. 3.55) —
We can use the facts by noting $Y = -Z = 0+(-1)\times Z$, with $a=0$ and $b=-1$ in $Y = a+bX$ and get
$$\varphi_Y(t) = e^{\imath\times 0\times t}\varphi_Z(-1\times t) = \varphi_Z(-t) = e^{-((-1\times t))^2/2} = e^{-t^2/2} = \varphi_Z(t)$$
Thus, $\varphi_{-Z}(t) = \varphi_Z(t)$ and therefore the distributions of Z and $-Z$ are the same. This should make sense because by switching signs of a symmetric (about 0) RV you have not changed its distribution! Note: we are not saying $Z = -Z$ but just that their distibutions are the same, i.e., $F_Z(z) = F_{-Z}(z)$ for every $z\in\mathbb{R}$.

Answer (Ex. 3.56) — Apply the formulas for the sample mean, $\overline{X_7}$, and sample variance.

Answer (Ex. 4.1) — This is nothing but the inversion sampler for the standard Cauchy RV X.

2. To find $V(X)$ using $V(X) = E(X^2) - (E(X))^2$ we need the first two moments of X. First, differentiate $\varphi_X(t)$ w.r.t. t

$$\frac{d}{dt}\varphi_X(t) = \frac{d}{dt}\left(\frac{1}{3}\left(1 + e^{\imath t} + e^{\imath 2t}\right)\right) = \frac{1}{3}\left(0 + \imath e^{\imath t} + 2\imath e^{\imath 2t}\right) = \frac{\imath}{3}\left(e^{\imath t} + 2e^{\imath 2t}\right)$$

We get the k-th moment $E(X^k)$ by multiplying the k-th derivative of $\varphi_X(t)$ evaluated at $t = 0$ by $\frac{1}{\imath^k}$ as follows:

$$E(X) = \frac{1}{\imath}\left[\frac{d}{dt}\varphi_X(t)\right]_{t=0} = \frac{1}{\imath}\left[\frac{\imath}{3}\left(e^{\imath t} + 2e^{\imath 2t}\right)\right]_{t=0} = \frac{1}{\imath}\frac{\imath}{3}\left(e^0 + 2e^0\right) = \frac{1}{3}(1+2) = 1 \ .$$

$$\begin{aligned} E(X^2) &= \frac{1}{\imath^2}\left[\frac{d^2}{dt^2}\varphi_X(t)\right]_{t=0} = \frac{1}{\imath^2}\left[\frac{d}{dt}\frac{\imath}{3}\left(e^{\imath t} + 2e^{\imath 2t}\right)\right]_{t=0} = \frac{1}{\imath^2}\left[\frac{\imath}{3}\left(\imath e^{\imath t} + 4\imath e^{\imath 2t}\right)\right]_{t=0} \\ &= \frac{1}{3}\left(e^0 + 4e^0\right) = \frac{5}{3} \ . \end{aligned}$$

Finally,

$$V(X) = E(X^2) - (E(X))^2 = \frac{5}{3} - 1^2 = \frac{5-3}{3} = \frac{2}{3} \ .$$

Answer (Ex. 3.51) —

1.

$$\varphi_X(t) = E\left(e^{\imath tX}\right) = \sum_{x=0}^{\infty} e^{\imath tx}\theta(1-\theta)^x = \sum_{x=0}^{\infty} \theta\left(e^{\imath t}(1-\theta)\right)^x = \frac{\theta}{1 - e^{\imath t}(1-\theta)} \ .$$

The last equality is due to $\sum_{x=0}^{\infty} ar^x = \frac{a}{1-r}$ with $a = \theta$ and $r = e^{\imath t}(1-\theta)$.

2.

$$\begin{aligned} E(X) &= \frac{1}{\imath}\left[\frac{d}{dt}\left(\frac{\theta}{1 - e^{\imath t}(1-\theta)}\right)\right]_{t=0} = \frac{1}{\imath}\theta\left[\frac{d}{dt}\left(\left(1 - e^{\imath t}(1-\theta)\right)^{-1}\right)\right]_{t=0} \\ &= \frac{1}{\imath}\theta\left[-\left(1 - e^{\imath t}(1-\theta)\right)^{-2}\frac{d}{dt}\left(1 - e^{\imath t}(1-\theta)\right)\right]_{t=0} \\ &= \frac{1}{\imath}\theta\left[-\left(1 - e^{\imath t}(1-\theta)\right)^{-2}\left(0 - \imath e^{\imath t}(1-\theta)\right)\right]_{t=0} \\ &= \frac{1}{\imath}\theta\left(-\left(1 - e^0(1-\theta)\right)^{-2}\left(-\imath e^0(1-\theta)\right)\right) \\ &= \frac{1}{\imath}\theta\left(-\left(1 - 1 + \theta\right)^{-2}\left(-\imath(1-\theta)\right)\right) \\ &= \frac{1-\theta}{\theta} \ . \end{aligned}$$

Answer (Ex. 3.52) —

$$\begin{aligned} \varphi_X(t) = E\left(e^{\imath tX}\right) &= \int_{-\infty}^{\infty} e^{\imath tx} f_X(x;a,b)dx = \int_a^b e^{\imath tx}\frac{1}{b-a}dx = \frac{1}{b-a}\int_a^b e^{\imath tx}dx = \frac{1}{b-a}\left[\frac{e^{\imath tx}}{\imath t}\right]_{x=a}^{b} \\ &= \frac{1}{(b-a)\imath t}\left(e^{\imath tb} - e^{\imath ta}\right) \ . \end{aligned}$$

Course Syllabus and Overview

Course Code: 1MS034		Semester: Autumn 2019
Report Code: 10504		week: 36 – 44
33%, DAG, NML	$\overset{\text{work}}{\Longrightarrow}$	$12 = \lfloor \frac{33}{100} 37.5 \rfloor$ hours / week
Uppsala University		2019-09-02 – 2019-11-03

Probability Theory I, 5.0 c
Course Syllabus: Friday, 04th of October 2019
Course Coordinator: Raazesh Sainudiin

Official Course Syllabus

See `https://www.uu.se/en/admissions/master/selma/kursplan/?kpid=38971&type=1`.

LEARNING OUTCOMES
On completion of the course, the student should be able to

- account for the axiomatic basis of the probability theory;
- carry out probability calculations by means of combinatorial principles and be able to use methods for independent events;
- account for the concepts of stochastic variable and expectation and be able to calculate probabilities, expectations and variance for given distributions;
- account for the most common probability distributions and how to do simulations with them;
- handle conditioned probabilities, distributions and expectations as well as moment generating and characteristic functions;
- apply the law of large numbers and the central limit theorem;
- account for probabilistic models within different application fields.

CONTENT
Combinatorics. The probability concept. Calculation of probabilities. Stochastic variable. Probability distributions. Independent and conditioned distributions. Expectation and variance. Conditioned expectations. Moment generating function. The central limit theorem. The law of large numbers. Practical examples of design of probability models.

ASSESSMENT
Written examination at the end of the course combined with written assignments during the course according to instructions delivered at course start.

Time Table & Course Overview – In Progress

(KEY,VALUE): (EX, Exercise), (RD, Read), (RW, Review), (UD, Understand), (PM, Program)

Table 1: Time Table for Virtual Student of Probability Theory I

Lec.	Lab.	Week	Topics	Comprehension ↦ Action × Content
01		36	Preliminaries: Set Theory, Numbers, Functions ,...	RW Sec. 1.1,1.3,1.4
				EX 1.2; RW Table. 1.1
			Elementary Combinatorics & Number Theory	RD 1.6; UD 6,7; EX 1.5; RD 1.9
	*		[optional] Introduction to MATLAB	PM 5,9,10,11,21
02			Probability	RD 2.1,2.2; EXs 2.3
03			Probability and Conditional Probability	RD 2.2,2.4; UD 31
04		37	Conditional Probability & Bayes Theorem	RD 2.4.1; UD 32,33,34; RD 2.4.2
			Conditional Probability & Independence	RD 2.4.2; UD 35,36,37; EXs 2.5
05			Random variables	RD 3.1; EX 3.1; UD 1,41i, EX 3.2
06			Discrete Random variables and IID Bernoulli Trials	RD 3.1, 3.2.1, 3.2.3
			Common Discrete Random variables	RD 3.2.3; UD 45; UD 4 ;EX 3.3;
07		38	Common Discrete Random variables	UD 46, 5, 47, 48, 6, 49, 50, 51
			Common Discrete Random variables	EX 3.4 & EXs 3.3;
08			Continuous RVs	RD 3.4; UD 52, 53, 54, RD 3.4.1
			Common Continuous RVs	RD 3.4.2; UD 55, 56; EX 3.17, 3.18
			Common Continuous RVs	EX 57, 58; UD 59; EXs 3.5
09			Transformations of RVs – Discrete	RD 3.6; UD 60, 61; RW 3.6.1
10		39	Transformations of RVs – Discrete	RD 3.6.2, UD 62, 63, 64
			Transformations of RVs – Continuous (1-to-1 & monotone)	RD 3.6.3; UD 65, 66, 67, 68
			Transformations of RVs – Continuous (Direct method)	RD 3.6.3; UD 69, 70; EXs 3.7
11			Expectations	RD 3.8, 3.8.1; UD 71, 72, 74, 75, 76;
				EX 3.29; UD 77, 78, 79, 80; EXs 3.9
12		40	Multivariate Random Variables	RD 3.10; UD 83, 84, 85, 86,
				UD 87, 88, 89, 90, 91, 92
13				RD 3.10.2, 3.10.3; UD 95, 96
			Common $\mathbb{R}^m$-valued RVs	RD 3.10.4; UD 97, 98;
				EX 3.36, 3.37, EXs 3.11
14			Characteristic Functions	RD 3.12; UD 101, 102, 103, 104, 105;
				EXs 3.13; [non-examinable] UD 169, 170
15		41	Statistics & Random Number Generation	RD 3.14, 4.2, 4.3
16			Statistics	EXs 3.15
17			Simulation – Inversion & Rejection Samplers	EXs 4.4
18		42	Convergence of RVs & Limit Laws	RD 5.1, UD 157, 158, 159, 160
19			Basic Inequalities & Law of Large Numbers	UD 161, 163, 164; EX 5.1
			Law of Large Numbers & Central Limit Theorem via CFs	UD 165, 166, 167, 168; EXs 5.4
20			Model exam with live-scribed solutions	

Table 2: Time Table for Inference Theory I

Lec.	Lab.	Week	Topics	Section/Labworks/Simulations

The values for the distribution function $\Phi(z)$ of the Normal$(0,1)$ RV Z are in the table on page 67. The thickness of the coating layer represented by Y_3 has the least probability (0.7888) of meeting specifications. Consequently, a priority should be to reduce variability in this part of the process.

Answer (Ex. 3.48) —
Let $X_1, X_2, \ldots, X_{10}$ denote the fill volumes of 10 cans. The average fill volume is the sample mean

$$\overline{X}_{10} = \frac{1}{10}\sum_{i=1}^{10} X_i$$

By property of Expectations and Variances for linear combinations

$$E(\overline{X}_{10}) = E\left(\frac{1}{10}\sum_{i=1}^{10} X_i\right) = \frac{1}{10}\sum_{i=1}^{n} E(X_i) = \frac{1}{10}\sum_{i=1}^{n} E(X_1) = \frac{1}{10}\times 10 \times E(X_1) = E(X_1) = 12.1$$

Or by directly using the "formula" $E(\overline{X}_{10}) = E(X_1) = 12.1$ for these 10 identically distributed RVs. Similarly,

$$V\left(\overline{X}_{10}\right) = V\left(\frac{1}{10}\sum_{i=1}^{10} X_i\right) = 10 \times \frac{1}{10^2} V(X_1) = \frac{1}{10} \times 0.01 = 0.001$$

Or by directly using the "formula" $V(\overline{X}_{10}) = V(X_1)/10$ for these 10 independently and identically distributed RVs.
By the special property of Normal RVs – a linear combination of independent normal RVs is also normal – we know that $\overline{X}_{10}$ is a Normal$(12.1, 0.001)$ RV. Consequetly, the probability of interest is

$$\begin{aligned}\mathbf{P}(\overline{X}_{10} < 12.01) &= P\left(\frac{\overline{X}_{10} - E(\overline{X}_{10})}{\sqrt{0.001}} < \frac{12.01 - E(\overline{X}_{10})}{\sqrt{0.001}}\right) = P\left(Z < \frac{12.01 - 12.1}{0.0316}\right) \\ &\approxeq \mathbf{P}(Z < -2.85) = 1 - \mathbf{P}(Z < 2.85) = 1 - \Phi(2.85) = 1 - 0.9978 = 0.0022\end{aligned}$$

Answer (Ex. 3.49) —
Using the Multinomial$(n = 10, \theta_1 = 0.6, \theta_2 = 0.3, \theta_3 = 0.08, \theta_4 = 0.02)$ $\vec{\text{RV}}$ as our model

$$\begin{aligned}&P\left((X_1, X_2, X_3, X_4) = (6, 2, 2, 0); n = 10, \theta_1 = 0.6, \theta_2 = 0.3, \theta_3 = 0.08, \theta_4 = 0.02\right) \\ &= \frac{10!}{6! \times 2! \times 2! \times 0!} \times 0.6^6 \times 0.3^2 \times 0.08^2 \times 0.02^0 \\ &= \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(6 \times 5 \times 4 \times 3 \times 2 \times 1) \times (2 \times 1) \times (2 \times 1) \times 1} \times 0.6^6 \times 0.3^2 \times 0.08^2 \times 1 \approxeq 0.03386\end{aligned}$$

Answer (Ex. 3.50) —

1. The CF of the discrete RV X is

$$\begin{aligned}\varphi_X(t) = E(e^{\imath tX}) &= \sum_{x\in\{0,1,2\}} e^{\imath tx} f_X(x) = e^{\imath t\times 0} \times \frac{1}{3} + e^{\imath t\times 1} \times \frac{1}{3} + e^{\imath t\times 2} \times \frac{1}{3} \\ &= \frac{1}{3}\left(1 + e^{\imath t} + e^{\imath 2t}\right) \ .\end{aligned}$$

1. If $x \notin [0,\infty)$ or $y \notin [2,3]$ then $f_X(x)f_Y(y) = 0 = f_{X,Y}(x,y)$
2. If $x \in [0,\infty)$ and $y \in [2,3]$ then $f_X(x)f_Y(y) = e^{-x} \times 1 = e^{-x} = f_{X,Y}(x,y)$.

Thus, X and Y are independent.

Answer (Ex. 3.46) —

$$\begin{aligned}
&\mathbf{P}(X_1 > 1000, X_2 > 1000, X_3 > 1000, X_4 > 1000) \\
&\quad = \int_{1000}^{\infty}\int_{1000}^{\infty}\int_{1000}^{\infty}\int_{1000}^{\infty} 9\times 10^{-12} e^{-0.001x_1 - 0.002x_2 - 0.0015x_3 - 0.003x_4} dx_1 dx_2 dx_3 dx_4 \\
&\quad = 9\times 10^{-12}\int_{1000}^{\infty}\int_{1000}^{\infty}\int_{1000}^{\infty}\int_{1000}^{\infty} e^{-0.001x_1} e^{-0.002x_2} e^{-0.0015x_3} e^{-0.003x_4} dx_1 dx_2 dx_3 dx_4 \\
&\quad = 9\times 10^{-12}\int_{1000}^{\infty} e^{-0.001x_1}\int_{1000}^{\infty} e^{-0.002x_2}\int_{1000}^{\infty} e^{-0.0015x_3}\int_{1000}^{\infty} e^{-0.003x_4} dx_4 dx_3 dx_2 dx_1
\end{aligned}$$

Since

$$\int_{1000}^{\infty} e^{-ax_i} dx_i = \left[\frac{e^{-ax_i}}{-a}\right]_{1000}^{\infty} = 0 + \frac{e^{-1000\times a}}{a} \ ,$$

the above quadruply iterated integral becomes

$$\begin{aligned}
&9\times 10^{-12} \times \frac{e^{-1000\times 0.001}}{0.001} \times \frac{e^{-1000\times 0.002}}{0.002} \times \frac{e^{-1000\times 0.0015}}{0.0015} \times \frac{e^{-1000\times 0.003}}{0.003} \\
&= 9\times 10^{-12} \times \frac{1000}{1} \times \frac{1000}{2} \times \frac{1000}{1.5} \times \frac{1000}{3} \times e^{-1} \times e^{-2} \times e^{-1.5} \times e^{-3} \\
&= 9\times 10^{-12} \times \frac{1}{9} \times 10^{12} \times e^{-7.5} = e^{-7.5} \approxeq 0.00055 \ .
\end{aligned}$$

Answer (Ex. 3.47) —
Due to independence of Y_1, Y_2 and Y_3

$$\begin{aligned}
&\mathbf{P}(9500 < Y_1 < 10500, 950 < Y_2 < 1050, 75 < Y_3 < 85) \\
&\qquad = \mathbf{P}(9500 < Y_1 < 10500)\mathbf{P}(950 < Y_2 < 1050)\mathbf{P}(75 < Y_3 < 85)
\end{aligned}$$

After standardizing each Normal RV (subtracting its mean and dividing by its standard deviation) we get

$$\begin{aligned}
&\mathbf{P}(9500 < Y_1 < 10500)\mathbf{P}(950 < Y_2 < 1050)\mathbf{P}(75 < Y_3 < 85) \\
&= P\left(\frac{9500-10000}{250} < Z < \frac{10500-10000}{250}\right) P\left(\frac{950-1000}{20} < Z < \frac{1050-1000}{20}\right) \\
&\qquad P\left(\frac{75-80}{4} < Z < \frac{85-80}{4}\right) \\
&= \mathbf{P}(-2.0 < Z < 2.0)\mathbf{P}(-2.5 < Z < 2.5)\mathbf{P}(-1.25 < Z < 1.25) \\
&= (\Phi(2.0) - (1-\Phi(2.0))) \times (\Phi(2.5) - (1-\Phi(2.5))) \times (\Phi(1.25) - (1-\Phi(1.25))) \\
&= (2\Phi(2.0) - 1) \times (2\Phi(2.5) - 1) \times (2\Phi(1.25) - 1) \\
&= ((2\times 0.9772) - 1) \times ((2\times 0.9938) - 1) \times ((2\times 0.8944) - 1) \qquad \text{using Table for } \Phi(z) \\
&= 0.9544 \times 0.9876 \times 0.7888 = 0.7435
\end{aligned}$$

Contents

$$E((X,Y)) = \sum_{(x,y)\in\mathcal{S}_{X,Y}} (x,y)\times f_{X,Y}(x,y)$$
$$= (0,0)\times 0.2 + (1,1)\times 0.1 + (1,2)\times 0.1 + (2,1)\times 0.1 + (2,2)\times 0.1 + (3,3)\times 0.4$$
$$= (0,0) + (0.1,0.1) + (0.1,0.2) + (0.2,0.1) + (0.2,0.2) + (1.2,1.2)$$
$$= (1.8,1.8)$$

Since addition is component-wise $E((X,Y)) = (E(X),E(Y))$ and therefore $E(X) = E(Y) = 1.8$. Alternatively, you can first find the marginal PMFs f_X and f_Y for X and Y and then take the expectations $E(X) = \sum_x x \times f_X(x)$ and $E(Y) = \sum_y y \times f_Y(y)$.
Finally,

$$\mathbf{Cov}(X,Y) = E(XY) - E(X)E(Y) = 4.5 - 1.8^2 = 1.26 \ .$$

Answer (Ex. 3.44) —
Since X and Y are independent, $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for all $(x,y)\in\mathbb{R}^2$, and we get:

$$F_{X,Y}(x,y) = \begin{cases} 0 & \text{if } x<1 \text{ or } y<0 \\ \frac{1}{2}\left(1-\frac{1}{2}e^{-y}-\frac{1}{2}e^{-2y}\right) & \text{if } 1\le x<2 \text{ and } y\ge 0 \\ 1-\frac{1}{2}e^{-y}-\frac{1}{2}e^{-2y} & \text{if } x\ge 2 \text{ and } y\ge 0 \end{cases}$$

You can arrive at the answer by partitioning x-axis into $(-\infty,1)$, $[1,2)$ and $[2,\infty)$ where $F_X(x)$ takes distinct values. Similarly, partition the y-axis into $(-\infty,0)$ and $[0,\infty)$ where $F_Y(y)$ takes distinct values. Now (x,y) can take values in one of these $3\times 2 = 6$ partitions of the $x\times y$ plane as follows (make a picture!):

$$(-\infty,1)\times(-\infty,0),\ [1,2)\times(-\infty,0),\ [2,\infty)\times(-\infty,0),\ (-\infty,1)\times[0,\infty),\ [1,2)\times[0,\infty),\ [2,\infty)\times[0,\infty) \ .$$

Now work out what $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ is for (x,y) in each of the above six partitions of the plane and you will get the the expression for $F_{X,Y}(x,y)$ given above.

Answer (Ex. 3.45) —
First obtain marginal PDF of Y. If $y\in[2,3]$ then

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_0^{\infty} e^{-x}dx = \left[-e^{-x}\right]_0^{\infty} = 0-(-1) = 1 \ .$$

Therefore,

$$f_Y(y) = \begin{cases} 1 & \text{if } y\in[2,3] \\ 0 & \text{otherwise} \ . \end{cases}$$

Now, obtain the marginal PDF of X. If $x\in[0,\infty)$ then

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_2^3 e^{-x}dy = e^{-x}\int_2^3 1dy = e^{-x}\left[y\right]_2^3 = e^{-x}(3-2) = e^{-x} \ .$$

Therefore,

$$f_X(x) = \begin{cases} e^{-x} & \text{if } x\in[0,\infty) \\ 0 & \text{otherwise} \ . \end{cases}$$

Finally, verifying that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for any $(x,y)\in\mathbb{R}^2$ is done case by case. Draw a picture on the plane to work out the cases from the distinct expressions taken by $f_{X,Y}(x,y)$. There are only two cases to consider (when $f_{X,Y}(x,y)$ takes zero values and when $f_{X,Y}(x,y)$ takes non-zero values):

Answer (Ex. 3.41) —
First derive the marginal PMF of X and Y and then check if the JPMF is the product of the marginal PMFs.

$$f_X(0) = \sum_{y \in \mathcal{S}_{X,Y}} f_{X,Y}(0,y) = f_{X,Y}(0,0) + f_{X,Y}(0,1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

and

$$f_X(1) = \sum_{y \in \mathcal{S}_{X,Y}} f_{X,Y}(1,y) = f_{X,Y}(1,0) + f_{X,Y}(1,1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Thus,

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{if } x = 0 \\ \frac{1}{2} & \text{if } x = 1 \\ 0 & \text{otherwise } . \end{cases}$$

Similarly,

$$f_Y(y) = \begin{cases} \sum_{x \in \mathcal{S}_{X,Y}} f_{X,Y}(x,0) = f_{X,Y}(0,0) + f_{X,Y}(1,0) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} & \text{if } y = 0 \\ \sum_{x \in \mathcal{S}_{X,Y}} f_{X,Y}(x,1) = f_{X,Y}(0,1) + f_{X,Y}(1,1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} & \text{if } y = 1 \\ 0 & \text{otherwise } . \end{cases}$$

Finally, the product of $f_X(x)$ and $f_Y(y)$ is

$$f_X(x) \times f_Y(y) = \begin{cases} \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} & \text{if } (x,y) = (0,0) \\ \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} & \text{if } (x,y) = (0,1) \\ \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} & \text{if } (x,y) = (1,0) \\ \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} & \text{if } (x,y) = (1,1) \\ 0 & \text{otherwise } . \end{cases}$$

which in turn is equal to the JPMF $f_{X,Y}(x,y)$ in the question. Therefore we have shown that the component RVs X and Y in the R$\vec{\text{V}}$ (X,Y) are indeed indepedent.

Answer (Ex. 3.42) —
Let X_1, X_2, X_3 be independent RVs that denote the thickness of the first, second and third layer, respectively. Let X denote the thickness of the final product. Then

$$X = X_1 + X_2 + X_3$$

By the property that $V\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 V(X_i)$, Variance of X is

$$V(X) = 1^2 V(X_1) + 1^2 V(X_2) + 1^2 V(X_3) = 25 + 40 + 30 = 95nm^2 \ .$$

This shows how the variance in each layer is propagated to the variance of the final product.

Answer (Ex. 3.43) —
Find $E(XY)$, $E(X)$ and $E(Y)$ to get $\mathbf{Cov}(X,Y) = E(XY) - E(X)E(Y)$ as follows:

$$\begin{aligned} E(XY) &= \sum_{(x,y) \in \mathcal{S}_{X,Y}} x \times y \times f_{X,Y}(x,y) \\ &= 0 \times 0 \times 0.2 + 1 \times 1 \times 0.1 + 1 \times 2 \times 0.1 + 2 \times 1 \times 0.1 + 2 \times 2 \times 0.1 + 3 \times 3 \times 0.4 = 4.5 \end{aligned}$$

List of Tables

8.

$$
\begin{aligned}
E(Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{X,Y}(x,y)dxdy \\
&= \int_0^1\int_0^1 y\frac{6}{5}\left(x^2+y\right)dxdy = \frac{6}{5}\int_0^1\int_0^1\left(x^2y+y^2\right)dxdy = \frac{6}{5}\int_0^1\left[\frac{x^3y}{3}+y^2x\right]_{x=0}^{1}dy \\
&= \frac{6}{5}\int_0^1\left(\frac{y}{3}+y^2-0-0\right)dy = \frac{6}{5}\left[\frac{y^2}{6}+\frac{y^3}{3}\right]_{y=0}^{1} = \frac{6}{5}\left(\frac{1}{6}+\frac{1}{3}+-0-0\right) = \frac{6}{5}\times\frac{3}{6} = \frac{3}{5}
\end{aligned}
$$

9.

$$
\begin{aligned}
E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)dxdy \\
&= \frac{6}{5}\int_0^1\int_0^1 xy\left(x^2+y\right)dxdy = \frac{6}{5}\int_0^1\int_0^1 x^3y+xy^2dxdy = \frac{6}{5}\int_0^1\left[\frac{x^4}{4}y+\frac{x^2y^2}{2}\right]_{x=0}^{1}dy \\
&= \frac{6}{5}\int_0^1\left(\frac{y}{4}+\frac{y^2}{2}-0-0\right)dy = \frac{6}{5}\left[\frac{y^2}{8}+\frac{y^3}{6}\right]_{y=0}^{1} = \frac{6}{5}\left(\frac{1}{8}+\frac{1}{6}-0-0\right) \\
&= \frac{6}{5}\left(\frac{3}{24}+\frac{4}{24}\right) = \frac{6}{5}\times\frac{7}{24} = \frac{7}{20}
\end{aligned}
$$

10.

$$
\mathbf{Cov}(X,Y) = E(XY)-E(X)E(Y) = \frac{7}{20}-\left(\frac{3}{5}\times\frac{3}{5}\right) = \frac{7}{20}-\frac{9}{25} = \frac{35}{100}-\frac{36}{100} = -\frac{1}{100}
$$

Answer (Ex. 3.40) — Note that $Y=(X-\mu)^2$ is not on-to-one so it is better to use the direct method by differentiating the distribution function of Y, $F_Y(y)$, to obtain $f_Y(y)$.
If $y \geqslant 0$,

$$
\begin{aligned}
F_Y(y) &= \mathbf{P}(Y \leqslant y) \\
&= \mathbf{P}((X-\mu)^2 \leqslant y) \\
&= \mathbf{P}(-\sqrt{y} \leqslant X-\mu \leqslant \sqrt{y}) \\
&= \mathbf{P}(\mu-\sqrt{y} \leqslant X \leqslant \mu+\sqrt{y}) \\
&= F_X(\mu+\sqrt{y}) - F_X(\mu+\sqrt{y})
\end{aligned}
$$

Differentiating this expression gives

$$
\begin{aligned}
f_Y(y) &= \frac{d}{dy}\left(F_X(\mu+\sqrt{y}) - F_X(\mu-\sqrt{y})\right) \\
&= \frac{1}{2}y^{-\frac{1}{2}}f_X(\mu+\sqrt{y}) - \left(-\frac{1}{2}y^{-\frac{1}{2}}\right)f_X(\mu-\sqrt{y}) \\
&= \frac{1}{2\sqrt{y}}\left(f_X(\mu+\sqrt{y})+f_X(\mu-\sqrt{y})\right)
\end{aligned}
$$

Note: If $y<0$ then $f_Y(y)=0$ since $F_Y(y)=0$ in this case.

Finally, the marginal PDF of the RV X in the first component of the R$\vec{\text{V}}$ (X,Y) is

$$f_X(x) = \begin{cases} \frac{6}{5}\left(x^2+\frac{1}{2}\right) & \text{if } 0<x<1 \\ 0 & \text{otherwise .} \end{cases}$$

3. Similarly, the marginal PDF $f_Y(y)$ for any $y \in (0,1)$ by integrating over x is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_0^1 \frac{6}{5}(x^2+y)dx = \frac{6}{5}\left[x^3/3+yx\right]_{x=0}^1 = \frac{6}{5}(y+1/3).$$

Finally, the marginal PDF of the RV Y in the second component of the R$\vec{\text{V}}$ (X,Y) is

$$f_Y(y) = \begin{cases} \frac{6}{5}\left(y+\frac{1}{3}\right) & \text{if } 0<y<1 \\ 0 & \text{otherwise .} \end{cases}$$

4.The product of marginal PDFs of X and Y does not equal the joint PDF of (X,Y) for values of $(x,y) \in (0,1)^2$

$$f_X(x)f_Y(y) = \frac{6}{5}\frac{6}{5}\left(y+\frac{1}{3}\right)\left(x^2+\frac{1}{2}\right) = \frac{6}{25}\left(6x^2y+2x^2+3y+1\right) \neq \frac{6}{5}(x^2+y) = f_{X,Y}(x,y)$$

Therefore X and Y are not independent random variables (they are dependent!).

5. The joint distribution function $F_{X,Y}(x,y)$ for any $(x,y) \in (0,1)^2$ is

$$\begin{aligned} F_{X,Y}(x,y) &= \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)dudv = \int_0^y\int_0^x \frac{6}{5}(u^2+v)dudv = \frac{6}{5}\int_0^y\left[\frac{u^3}{3}+vu\right)\Big]_{u=0}^{x} dv \\ &= \frac{6}{5}\int_0^y\left(\frac{x^3}{3}+vx-0\right)dv = \frac{6}{5}\left[\frac{x^3v}{3}+\frac{v^2x}{2}\right]_{v=0}^{y} = \frac{6}{5}\left(\frac{x^3y}{3}+\frac{y^2x}{2}-0\right) \\ &= \frac{6}{5}\left(\frac{x^3y}{3}+\frac{y^2x}{2}-0\right) \end{aligned}$$

6.

$$\begin{aligned} \mathbf{P}(X>0.5, Y<0.6) &= \int_{-\infty}^{0.6}\int_{0.5}^{\infty} f_{X,Y}(x,y)dxdy = \int_0^{0.6}\int_{0.5}^1 \frac{6}{5}(x^2+y)dxdy \\ &= \frac{6}{5}\int_0^{0.6}\left[\frac{x^3}{3}+yx\right]_{x=0.5}^{1} dy = \frac{6}{5}\int_0^{0.6}\left(\frac{7}{24}+\frac{y}{2}\right)dy \\ &= \frac{6}{5}\left[\frac{7}{24}y+\frac{y^2}{2}\right]_{y=0}^{0.6} = \frac{6}{5}\left(\frac{7}{24}\times\frac{6}{10}+\frac{36}{400}\right) = 0.318 \end{aligned}$$

7.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)dxdy \\ &= \frac{6}{5}\int_0^1\int_0^1 x\left(x^2+y\right)dxdy = \frac{6}{5}\int_0^1\int_0^1\left(x^3+xy\right)dxdy = \frac{6}{5}\int_0^1\left[\frac{x^4}{4}+\frac{1}{2}x^2y\right]_{x=0}^{1} dy \\ &= \frac{6}{5}\int_0^1\left(\frac{1}{4}+\frac{y}{2}-0-0\right)dy = \frac{6}{5}\left[\frac{y}{4}+\frac{y^2}{4}\right]_{y=0}^{1} = \frac{6}{5}\left(\frac{1}{4}+\frac{1}{4}-0-0\right) = \frac{6}{5}\times\frac{1}{2} = \frac{3}{5} \end{aligned}$$

List of Figures

Answer (Ex. 3.38) — The probability that *one* light bulb doesn't need to be replaced in 1200 hours is:

$$\begin{aligned}
\mathbf{P}(X > 1.2) &= 1 - \mathbf{P}(X < 1.2) \\
&= 1 - \int_1^{1.2} 6(0.25 - (x-1.5)^2)\,dx \\
&= 1 - \int_1^{1.2} 6(0.25 - x^2 + 3x - 2.25)\,dx \\
&= 1 - \int_1^{1.2} (-6x^2 + 18x - 12)\,dx \\
&= 1 - \left[-2x^3 + 9x^2 - 12x\right]_1^{1.2} \\
&= 1 - 0.1040 \\
&= 0.8960
\end{aligned}$$

Assuming that the three light bulbs function independently of each other, the probability that none of them need to be replaced in the first 1200 hours is

$$\mathbf{P}(\{X_1 > 1.2\} \cap \{X_2 > 1.2\} \cap \{X_3 > 1.2\}) = 0.8960^3 = 0.7193$$

where X_i is the length of time that bulb i lasts.

Answer (Ex. 3.39) —

1. To find a we simply set $1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy$ and solve for a as follows:

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = \int_0^1\int_0^1 a(x^2+y)dxdy \\
&= a\int_0^1\int_0^1 (x^2+y)dxdy = a\int_0^1 \left[\frac{1}{3}x^3 + yx\right]_{x=0^1} dy \\
&= a\int_0^1 \left(\frac{1}{3} + y - 0\right) dy = a\left[\frac{y}{3} + \frac{1}{2}y^2\right]_{y=0}^1 \\
&= a\left(0 - \left(\frac{1}{3} + \frac{1}{2}1^2\right)\right) = a\left(\frac{1}{3} + \frac{1}{2}\right) \\
&= a\left(\frac{5}{6}\right)
\end{aligned}$$

   Therefore $a = 6/5$ and the joint PDF is

$$f_{X,Y}(x,y) = \begin{cases} \frac{6}{5}(x^2+y) & \text{if } 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{otherwise .} \end{cases}$$

2. First compute the marginal PDF $f_X(x)$ for any $x \in (0,1)$ by integrating over y

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_0^1 \frac{6}{5}(x^2+y)dy = \left[\frac{6}{5}(yx^2 + y^2/2)\right]_{y=0}^1 \\
&= \frac{6}{5}\left((1\times x^2 + 1^2/2) - 0\right) = \frac{6}{5}\left(x^2 + \frac{1}{2}\right)
\end{aligned}$$

Answer (Ex. 3.35) — 1.The probability mass function of X is

$$f(x) = \begin{cases} \frac{1}{6} & x = 1 \\ \frac{1}{6} & x = 2 \\ \frac{1}{6} & x = 3 \\ \frac{1}{6} & x = 4 \\ \frac{1}{6} & x = 5 \\ \frac{1}{6} & x = 6 \end{cases}$$

and so

$$\mathbf{E}(X) \;=\; \sum_{i=1}^{6} x_i f(x_i) \;=\; \frac{1}{6}+\frac{2}{6}+\frac{3}{6}+\frac{4}{6}+\frac{5}{6}+\frac{6}{6} \;=\; 3.5\,.$$

Now,

$$\mathbf{E}(X^2) \;=\; \sum_{i=1}^{6} x_i^2 f(x_i) \;=\; \frac{1}{6}+\frac{4}{6}+\frac{9}{6}+\frac{16}{6}+\frac{25}{6}+\frac{36}{6} \;=\; \frac{91}{6} \;=\; 15.1667$$

and so the variance is

$$V(X) \;=\; \mathbf{E}(X^2) - (\mathbf{E}(X))^2 \;=\; 15.1667 - 3.5^2 \;=\; 2.9167\,.$$

2.The density function of the uniform distribution, X, on $[0, 8]$ is

$$f(x) = \begin{cases} \frac{1}{8} & 0 < x < 8 \\ 0 & \text{otherwise} \end{cases}$$

Therefore,

$$\mathbf{E}(X) \;=\; \frac{8+0}{2} \;=\; 4\ ,$$

and

$$\mathbf{V}(X) \;=\; \frac{(8-0)^2}{12} \;=\; \frac{16}{3}$$

3.Since $f(x)$ is the density function of an Exponential(λ) random variable with parameter $\lambda = 2$, we can use earlier results to get

$$\mathbf{E}(X) \;=\; \frac{1}{2} \quad \text{and} \quad \mathbf{V}(X) \;=\; \frac{1}{4}\ .$$

(You should also be able to do this by integration!)

Answer (Exercise 3.36) —

- The number of paths that lead to a (x_1, x_2) with $x_1 + x_2 = n$ is equal to $\binom{n}{x_1}$. We have already seen this as random walks in Manhatta.
- $\binom{n}{x_1}\theta^{x_1}(1-\theta)^{x_2}$

Answer (Exercise 3.37) — The probability of going east or north in the first quadrant (idealized Manhattan with streets and avenues) is the same as a ball falling left or right in the Galton's Quincunx. The buckets that collect the balls after dropping through n levels of nails are labelled by the number of right turns as $0, 1, \ldots, n$ when modelled by a Binomial(n, θ), and this is analogous to the number of steps taken north in the random walk case.

Chapter 1

Preliminaries

1.1 Elementary Set Theory

A **set** is a collection of distinct objects. We write a set by enclosing its elements with curly braces. For example, we denote a set of the two objects $\circ$ and $\bullet$ by:

$$\boxed{\{\circ, \bullet\}}\,.$$

Sometimes, we give names to sets. For instance, we might call the first example set A and write:

$$\boxed{A = \{\circ, \bullet\}}\,.$$

We do not care about the order of elements within a set, i.e. $A = \{\circ, \bullet\} = \{\bullet, \circ\}$. We do not allow a set to contain multiple copies of any of its elements unless the copies are distinguishable, say by labels. So, $B = \{\circ, \bullet, \bullet\}$ is not a set unless the two copies of $\bullet$ in B are labelled or marked to make them distinct, e.g. $B = \{\circ, \tilde{\bullet}, \bullet'\}$. Names for sets that arise in a mathematical discourse are given upper-case letters $(A, B, C, D, \ldots)$. Special symbols are reserved for commonly encountered sets.

Here is the set $\mathfrak{G}$ of twenty two Greek lower-case alphabets that we may encounter later:

$$\mathfrak{G} = \{\ \alpha,\ \beta,\ \gamma,\ \delta,\ \epsilon,\ \zeta,\ \eta,\ \theta,\ \kappa,\ \lambda,\ \mu,\ \nu,\ \xi,\ \pi,\ \rho,\ \sigma,\ \tau,\ \upsilon,\ \phi,\ \chi,\ \psi,\ \omega\ \}\quad.$$

They are respectively named alpha, beta, gamma, delta, epsilon, zeta, eta, theta, kappa, lambda, mu, nu, xi, pi, rho, sigma, tau, upsilon, phi, chi, psi and omega. LHS and RHS are abbreviations for objects on the Left and Right Hand Sides, respectively, of some binary relation. By the notation:

$$\boxed{LHS := RHS}\,,$$

we mean that LHS **is equal, by definition, to** RHS.

The set which does not contain any element (the collection of nothing) is called the **empty set**:

$$\boxed{\emptyset := \{\,\}}\,.$$

We say an element b **belongs to** a set B, or simply that b belongs to B or that b is an element of B, if b is one of the elements that make up the set B, and write:

$$\boxed{b \in B}\,.$$

When b **does not belong to** B, we write:

$$\boxed{b \notin B}\,.$$

2.The profit per conditioner is \$55, and so the expected daily profit given by

$$E(55\,X) \;=\; 55E(X) \;=\; 55 \times 11.7 \;=\; 643.50\,,$$

is \$643.50.

Answer (Ex. 3.31) — The expected value $\mathbf{E}(X)$ is

$$\begin{aligned}
\mathbf{E}(X) &= \int_0^1 6x(1-x)x\,dx \\
&= \int_0^1 (6x^2 - 6x^3)\,dx \\
&= 2x^3 - \frac{6}{4}x^4\Big]_0^1 \\
&= 2(1^3 - 0) - \frac{6}{4}(1^4 - 0) \\
&= 0.5
\end{aligned}$$

$$\begin{aligned}
\mathbf{E}(X^2) &= \int_0^1 6x(1-x)x^2\,dx \\
&= \int_0^1 6x^3 - 6x^4\,dx \\
&= \frac{6}{4}x^4 - \frac{6}{5}x^5\Big]_0^1 \\
&= \frac{6}{4}(1^4 - 0) - \frac{6}{5}(1^5 - 0) \\
&= 0.3
\end{aligned}$$

the variance is $\mathbf{V}(X) \;=\; \mathbf{E}(X^2) - (\mathbf{E}(X))^2 \;=\; 0.3 - 0.5^2 \;=\; 0.05$

Answer (Ex. 3.32) — This was already done in Sec. 3.8.2 on Properties of Expectation. Make sure you understand each step.

Answer (Ex. 3.33) — Using the definition of variance, expectations and by completing the square, we get:

$$\begin{aligned}
&V(aX+b) = E((aX+b)^2) - (E(aX+b))^2 = E\left((aX)^2 + 2aXb + b^2\right) - (aE(X)+b)^2 \\
&= a^2E\left(X^2\right) + 2abE(X) + b^2 - a^2(E(X))^2 - 2abE(X) - b^2 = a^2\left(E\left(X^2\right) - (E(X))^2\right) = a^2V(X)\,.
\end{aligned}$$

Answer (Ex. 3.27) — Since $y = g(x) = \sqrt{x}$ is a monotone increasing function for $x \geqslant 0$, we can apply the change of variable formula.
Now $x = g^{-1}(y) = y^2$ is a monotone increasing function for $y \geqslant 0$ so on this interval

$$\left|\frac{d}{dy}\left(g^{-1}(y)\right)\right| \;=\; \left|\frac{d}{dy}\left(y^2\right)\right| \;=\; 2y\,.$$

Therefore

$$f_Y(y) \;=\; f_X\left(y^2\right) \times |2y| \;=\; \lambda e^{-\lambda y^2} \times 2y \;=\; 2\lambda y e^{-\lambda y^2}\,.$$

So the probability density function of Y is given by

$$f_Y(y) \;=\; \begin{cases} 2\lambda y e^{-\lambda y^2} & y \geqslant 0 \\ 0 & y < 0 \end{cases}.$$

Answer (Ex. 3.28) — First note that $y = g(x) = \log_e(x)$ is a monotone increasing function over $a \leqslant x \leqslant b$, so we can apply the change of variable formula.
$x = g^{-1}(y) = e^y$ is a monotone increasing function over $\log_e(a) \leqslant \log_e(x) \leqslant \log_e(b)$, that is, over $\log_e(a) \leqslant y \leqslant \log_e(b)$.
For $\log_e(a) \leqslant y \leqslant \log_e(b)$,

$$\left|\frac{d}{dy}\left(g^{-1}(y)\right)\right| \;=\; \left|\frac{d}{dy}\left(e^y\right)\right| \;=\; e^y\,.$$

Therefore

$$f_Y(y) \;=\; f_X\left(g^{-1}(y)\right) \times \left|\frac{d}{dy}\left(g^{-1}(y)\right)\right| \;=\; \frac{1}{b-a} \times e^y\,.$$

So the probability density function of Y is given by

$$f_Y(y) \;=\; \begin{cases} \dfrac{e^y}{b-a} & \log_e(a) \leqslant x \leqslant \log_e(b) \\ 0 & \text{otherwise} \end{cases}.$$

Answer (Exercise 3.29) — Derive the answers from the definition of $\mathbf{E}(X)$ and $\mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2$ when $X \sim \text{Uniform}(\theta_1, \theta_2)$ with PDF given in Model 9.

$$\mathbf{E}(X) = \frac{\theta_1 + \theta_2}{2} \quad \mathbf{V}(\boldsymbol{X}) = \frac{(\theta_2 - \theta_1)^2}{12}$$

Answer (Ex. 3.30) — 1.The expected number of conditioners that the store sells daily is

$$\begin{aligned} \mathbf{E}(X) &= \sum_{i=1}^{n} x_i p_{x_i} \\ &= (10 \times 0.1 + 11 \times 0.3 + 12 \times 0.4 + 13 \times 0.2) \\ &= 1 + 3.3 + 4.8 + 2.6 \\ &= 11.7\ . \end{aligned}$$

For our example set $A = \{\circ, \bullet\}$, $\star \notin A$ but $\bullet \in A$.

We say that a set C is a **subset** of another set D and write:

$$\boxed{C \subset D}$$

if every element of C is also an element of D. By this definition, any set is a subset of itself.

We say that two sets C and D are **equal** (as sets) and write $C = D$ 'if and only if' ($\iff$) every element of C is also an element of D, and every element of D is also an element of C. This definition of set equality is notationally summarised as follows:

$$\boxed{C = D \quad \iff \quad C \subset D, D \subset C}\,.$$

When two sets C and D are not equal by the above definition, we say that C is **not equal** to D and write:

$$\boxed{C \neq D}\,.$$

The **union** of two sets C and D, written as $C \cup D$, is the set of elements that belong to C or D. We can formally express our definition of set union as:

$$\boxed{C \cup D := \{x : x \in C \quad \text{or} \quad x \in D\}}\,.$$

When a colon (:) appears inside a set, it stands for 'such that'. Thus, the above expression is read as 'C union D is equal by definition to the set of all elements x, such that x belongs to C or x belongs to D.'

Similarly, the **intersection** of two sets C and D, written as $C \cap D$, is the set of elements that belong to both C and D. Formally:

$$\boxed{C \cap D := \{x : x \in C \quad \text{and} \quad x \in D\}}\,.$$

Venn diagrams are visual aids for set operations as in the diagrams below.

Figure 1.1: Union and intersection of sets shown by Venn diagrams

The set-difference or **difference** of two sets C and D, written as $C \setminus D$, is the set of elements in C that do not belong to D. Formally:

$$\boxed{C \setminus D := \{x : x \in C \quad \text{and} \quad x \notin D\}}\,.$$

When a universal set, e.g. U is well-defined, the **complement** of a given set B denoted by B^c is the set of all elements of U that don't belong to B, i.e.:

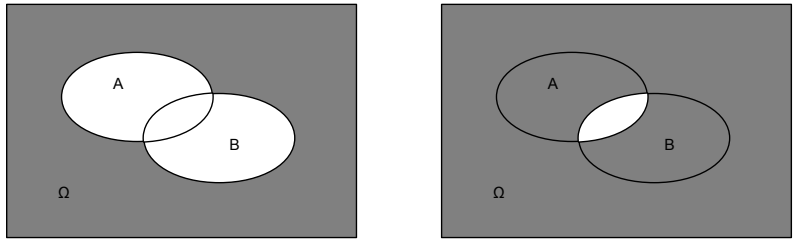
$$\boxed{B^c := U \setminus B}\,.$$

We say two sets C and D are **disjoint** if they have no elements in common, i.e. $C \cap D = \emptyset$.

By drawing Venn diagrams, let us check **De Morgan's Laws**:

$$\boxed{(A \cup B)^c = A^c \cap B^c \text{ and } (A \cap B)^c = A^c \cup B^c}$$

Classwork 1 (Fruits and colours) Consider a set of fruits $F = \{\text{orange}, \text{banana}, \text{apple}\}$ and a set of colours $C = \{\text{red}, \text{green}, \text{blue}, \text{orange}\}$. Then,



(a) $(A \cup B)^c = A^c \cap B^c$ (b) $(A \cap B)^c = A^c \cup B^c$

Figure 1.2: These Venn diagram illustrate De Morgan's Laws.

1. $F \cap C =$
2. $F \cup C =$
3. $F \setminus C =$
4. $C \setminus F =$

Classwork 2 (Subsets of a universal set) Suppose we are given a universal set U, and three of its subsets, A, B and C. Also suppose that $A \subset B \subset C$. Find the circumstances, if any, under which each of the following statements is true (T) and justify your answer:

(1) $C \subset B$	T when $B = C$	(2) $A \subset C$	T by assumption
(3) $C \subset \emptyset$	T when $A = B = C = \emptyset$	(4) $\emptyset \subset A$	T always
(5) $C \subset U$	T by assumption	(6) $U \subset A$	T when $A = B = C = U$

1.2 Exercises

Ex. 1.1 — Let Ω be the universal set of students, lecturers and tutors involved in a course. Now consider the following subsets:

- The set of 50 students, $S = \{S_1, S_2, S_2, \ldots S_{50}\}$.
- The set of 3 lecturers, $L = \{L_1, L_2, L_3\}$.
- The set of 4 tutors, $T = \{T_1, T_2, T_3, L_3\}$.

Note that one of the lecturers also tutors in the course. Find the following sets:

(a)$T \cap L$	(f)$S \cap L$
(b)$T \cap S$	(g)$S^c \cap L$
(c)$T \cup L$	(h)T^c
(d)$T \cup L \cup S$	(i)$T^c \cap L$
(e)S^c	(j)$T^c \cap T$

Ex. 1.2 — Using Venn diagram, sketch and check the rule:
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Note that the second equality above is emphasizing that the inverse image $g[-1](y)$ is indeed the inverse function $g^{-1}(y)$ for this $g : \{1, 2, 3, \ldots\} \to \{2^{-1}, 2^{-2}, 2^{-3}, \ldots\}$. Therefore,

$$f_Y(y) = \begin{cases} f_X(-\log_2(y)) = \theta(1-\theta)^{-\log_2(y)-1} & \text{if } y \in \{2^{-1}, 2^{-2}, 2^{-3}, \ldots\} \\ 0 & \text{otherwise}\,. \end{cases}$$

Answer (Ex. 3.25) — Since X is a Poisson(λ) random variable (by suppressing the '; λ' in the argument to $f_X(\cdot)$ for notational ease), we get

$$f_X(x) \;=\; \mathbf{P}(X = x) \;=\; \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{for } x = 0, 1, 2, \ldots \\ 0 & \text{otherwise}\,. \end{cases}$$

If $Y = (X+1)^{-2} = 1/(X+1)^2$ then

$$\{\ldots, 2, 1, 0\} \ni x \xrightarrow{(x+1)^{-2}} y \;\in\; \left\{1, \frac{1}{4}, \frac{1}{9}, \ldots\right\}$$

and since $y = g(x) = (x+1)^{-2}$ as it maps or associates each $y \in \{1, \frac{1}{4}, \frac{1}{9}, \ldots\}$ to exactly one $x \in \{0, 1, 2, \ldots\}$ given by $g^{-1}(y) = y^{-1/2} - 1 = \frac{1}{\sqrt{y}} - 1 = x$, its inverse function, this is because g is *injective* or *one-to-one* as explained here if you want to recall quickly `https://en.wikipedia.org/wiki/Injective_function` again, so we get:

$$f_Y(y) \;=\; \mathbf{P}(Y = y) \;=\; \sum_{\{x : g(x) = y\}} f_X(x) \;=\; f_X\left(\frac{1}{\sqrt{y}} - 1\right) \;=\; \frac{\lambda^{(\frac{1}{\sqrt{y}} - 1)}\, e^{-\lambda}}{(\frac{1}{\sqrt{y}} - 1)!}$$

for $y = 1, \frac{1}{4}, \frac{1}{9}, \ldots$, and 0 otherwise.

CAUTION: This is a discrete RV and so don't just blindly apply the change of variable formula that only applies to continuous RV with a monotone and one-to-one function g with inverse g^{-1}; it's just that in this discrete RV setting the inverse image also happens to satisfy these properties. But Poisson is discrete and 'change of variable formula" is hence inapplicable.

Answer (Ex. 3.26) — Since $y = g(x) = e^x$ is a monotone increasing function for $x \geqslant 0$, we can apply the change of variable formula.
Now $x = g^{-1}(y) = \log_e(y)$ is a monotone increasing function for y in $[1, \infty)$

$$\left|\frac{d}{dy}\left(g^{-1}(y)\right)\right| \;=\; \left|\frac{d}{dy}\left(\log_e(y)\right)\right| \;=\; \frac{1}{y}\,.$$

Therefore

$$f_Y(y) \;=\; f_X\left(\log_e(y)\right) \times \left|\frac{1}{y}\right| \;=\; \log_e(y)\, e^{-\log_e(y)} \times \frac{1}{y} \;=\; \log_e(y)\, \frac{1}{y^2}$$

since $e^{-\log_e(y)} = e^{\log_e(y^{-1})} = y^{-1}$.
So the probability density function of Y is given by

$$f_Y(y) \;=\; \begin{cases} \dfrac{\log_e(y)}{y^2} & \text{if } x \geqslant 1 \\ 0 & \text{otherwise} \end{cases}\,.$$

Answer (Ex. 3.21) — 1.

$$\begin{aligned}\int_0^2 k\,e^{-x}\,dx &= 1\\ \left[-k\,e^{-x}\right]_0^2 &= 1\\ k\,(-e^{-2}+1) &= 1\\ k &= \frac{1}{1-e^{-2}} \quad (\approx 1.1565)\end{aligned}$$

2.

$$\begin{aligned}\mathbf{P}(X \geq 1) &= 1 - \mathbf{P}(X < 1)\\ &= 1 - \int_0^1 ke^{-x}\,dx\\ &= 1 + k\left(e^{-x}\right]_0^1\\ &= 1 + \frac{e^{-1}-1}{1-e^{-2}}\\ &\approx 0.2689\end{aligned}$$

Answer (Ex. 3.22) — Using Equation (3.37), we can tabulate as follows:

y	0	1	4	9
$f_Y(y)$	$f_X(3) = \frac{1}{6}$	$f_X(2) + f_X(4) = \frac{2}{6}$	$f_X(1) + f_X(5) = \frac{2}{6}$	$f_X(6) = \frac{1}{6}$

Answer (Ex. 3.23) — The probability mass function $f_Y(y;n)$ for $Y = |X|$, the absolute value of X, comes from applying the formula:

$$f_Y(y;n) = \sum_{x\in\{x:g(x)=y\}} f_X(x;n)\ ,$$

as follows:

$$f_Y(y) = \begin{cases}\sum_{x\in\{x:|x|=0\}} f_X(x;n) = f_X(0;n) = \frac{1}{2n+1} & \text{if } y = 0\\ \sum_{x\in\{x:|x|=y\}} f_X(x;n) = (f_X(y;n) + f_X(-y;n)) = \frac{2}{2n+1} & \text{if } y \in \{1,2,\ldots,n\}\\ 0 & \text{otherwise}\end{cases}$$

Answer (Ex. 3.24) — We are given that $Y = 2^{-X}$. Define the function

$$g : \{1,2,3,\ldots\} \to \{2^{-1}, 2^{-2}, 2^{-3}, \ldots\}$$

by $y = g(x) = 2^{-x}$. Then g is one-to-one and onto and so by Equation (3.37),

$$f_Y(y) = \sum_{x\in g^{[-1]}(y)} f_X(x) = \sum_{x\in g^{-1}(y)} f_X(x) = \sum_{x\in\{-\log_2(y)\}} f_X(x) = f_X(-\log_2(y))\,.$$

Ex. 1.3 — Using Venn diagram, sketch and check the rule:
$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Ex. 1.4 — Using a Venn diagram, illustrate the idea that $A \subseteq B$ if and only if $A \cup B = B$.

SET SUMMARY

$\{a_1, a_2, \ldots, a_n\}$	–	a set containing the elements, $a_1, a_2, \ldots, a_n$.
$a \in A$	–	a is an element of the set A.
$A \subseteq B$	–	the set A is a subset of B.
$A \cup B$	–	"union", meaning the set of all elements which are in A or B, or both.
$A \cap B$	–	"intersection", meaning the set of all elements in both A and B.
$\{\}$ or $\emptyset$	–	empty set.
Ω	–	universal set.
A^c	–	the complement of A, meaning the set of all elements in Ω, the universal set, which are not in A.

1.3 Natural Numbers, Integers and Rational Numbers

We denote the number of elements in a set named B by:

$$\boxed{\#B := \text{Number of elements in the set } B}\ .$$

In fact, the Hindu-Arab numerals we have inherited are based on this intuition of the size of a collection. The elements of the set of **natural numbers**:

$$\boxed{\mathbb{N} := \{1,2,3,4,\ldots\}}\ ,\text{may be defined using } \# \text{ as follows:}$$

$$\begin{aligned}1 &:= \#\{\star\} = \#\{\bullet\} = \#\{\alpha\} = \#\{\{\bullet\}\} = \#\{\{\bullet,\bullet'\}\} = \ldots,\\ 2 &:= \#\{\star',\star\} = \#\{\bullet,\circ\} = \#\{\alpha,\omega\} = \#\{\{\circ\},\{\alpha,\star,\bullet\}\} = \ldots,\\ &\vdots\end{aligned}$$

For our example sets, $A = \{\circ,\bullet\}$ and the set of Greek alphabets $\mathfrak{G}$, $\#A = 2$ and $\#\mathfrak{G} = 22$. The number zero may be defined as the size of an empty set:

$$0 := \#\emptyset = \#\{\}$$

The set of **non-negative integers** is:

$$\boxed{\mathbb{Z}_+ := \mathbb{N} \cup \{0\} = \{0,1,2,3,\ldots\}}\ .$$

A **product set** is the **Cartesian product** ($\times$) of two or more possibly distinct sets:

$$\boxed{A \times B := \{(a,b) : a \in A \text{ and } b \in B\}}$$

For example, if $A = \{\circ,\bullet\}$ and $B = \{\star\}$, then $A \times B = \{(\circ,\star),(\bullet,\star)\}$. Elements of $A \times B$ are called **ordered pairs**.

The binary arithmetic operation of **addition** $(+)$ between a pair of non-negative integers $c, d \in \mathbb{Z}_+$ can be defined via sizes of disjoint sets. Suppose, $c = \#C$, $d = \#D$ and $C \cap D = \emptyset$, then:

$$c + d = \#C + \#D := \#(C \cup D) \ .$$

For example, if $A = \{\circ, \bullet\}$ and $B = \{\star\}$, then $A \cap B = \emptyset$ and $\#A + \#B = \#(A \cup B) \iff 2+1 = 3$.

The binary arithmetic operation of **multiplication** $(\cdot)$ between a pair of non-negative integers $c, d \in \mathbb{Z}_+$ can be defined via sizes of product sets. Suppose, $c = \#C$, $d = \#D$, then:

$$c \cdot d = \#C \cdot \#D := \#(C \times D) \ .$$

For example, if $A = \{\circ, \bullet\}$ and $B = \{\star\}$, then $\#A \cdot \#B = \#(A \times B) \iff 2 \cdot 1 = 2$.

More generally, a product set of $A_1, A_2, \ldots, A_m$ is:

$$\boxed{A_1 \times A_2 \times \cdots \times A_m := \{(a_1, a_2, \ldots, a_m) : a_1 \in A_1, a_2 \in A_2, \ldots, a_m \in A_m\}}$$

Elements of an m-product set are called **ordered m-tuples**. When we take the product of the same set we abbreviate as follows:

$$\boxed{A^m := \underbrace{A \times A \times \cdots \times A}_{m \text{ times}} := \{(a_1, a_2, \ldots, a_m) : a_1 \in A, a_2 \in A, \ldots, a_m \in A\}}$$

Classwork 3 (Cartesian product of sets) 1. Let $A = \{\circ, \bullet\}$. What are the elements of A^2? 2. Suppose $\#A = 2$ and $\#B = 3$. What is $\#(A \times B)$? 3. Suppose $\#A_1 = s_1, \#A_2 = s_2, \ldots, \#A_m = s_m$. What is $\#(A_1 \times A_2 \times \cdots \times A_m)$?

Now, let's recall the definition of a function. A **function** is a "mapping" that associates each element in some set $\mathbb{X}$ (the domain) to exactly one element in some set $\mathbb{Y}$ (the range). Two different elements in $\mathbb{X}$ can be mapped to or associated with the same element in $\mathbb{Y}$, and not every element in $\mathbb{Y}$ needs to be mapped. Suppose $x \in \mathbb{X}$. Then we say $f(x) = y \in \mathbb{Y}$ is the **image** of x. To emphasise that f is a **function** from $\mathbb{X} \ni x$ to $\mathbb{Y} \ni y$, we write:

$$\boxed{f(x) = y : \mathbb{X} \to \mathbb{Y}} \ .$$

And for some $y \in \mathbb{Y}$, we call the set:

$$\boxed{f^{[-1]}(y) := \{x \in \mathbb{X} : f(x) = y\} \subset \mathbb{X}} \ ,$$

the **pre-image** or **inverse image** of y, and

$$\boxed{f^{[-1]} := f^{[-1]}(y \in \mathbb{Y}) = X \subset \mathbb{X}} \ ,$$

That is,

$$\begin{aligned} k\,x\Big]_{-4}^{4} &= 1 \\ k(4-(-4)) &= 1 \\ 8k &= 1 \\ k &= \frac{1}{8} \end{aligned}$$

2.First note that if $x < -4$, then

$$F(x) \ = \ \int_{-\infty}^{x} 0\,dv \ = \ 0 \, .$$

If $-4 \le x \le 4$, then

$$\begin{aligned} F(x) &= \int_{-\infty}^{-4} 0\,dv \ + \ \int_{-4}^{x} \frac{1}{8}\,dv \\ &= 0 \ + \ \left[\frac{1}{8}v\right]_{-4}^{x} \\ &= \frac{1}{8}(x+4) \end{aligned}$$

If $x \ge 4$, then

$$\begin{aligned} F(x) &= \int_{-\infty}^{-4} 0\,dv \ + \ \int_{-4}^{4} \frac{1}{8}\,dv \ + \ \int_{4}^{x} 0\,dv \\ &= 0 \ + \ \left[\frac{1}{8}v\right]_{-4}^{4} \ + \ 0 \\ &= 1 \end{aligned}$$

Hence

$$F(x) \ = \ \begin{cases} 0 & x < -4 \\ \frac{1}{8}(x+4) & -4 \le x \le 4 \\ 1 & x \ge 4 \end{cases}$$

3.The graphs of $f(x)$ and $F(x)$ for random variable X are as follows:

Answer (Ex. 3.20) — 1.Since the distribution function is $F(t;\lambda) \ = \ 1 - \exp(-\lambda t)$,

$$\mathbf{P}(t > \tau) \ = \ 1 - \mathbf{P}(t < \tau) \ = \ 1 - F(\tau; \lambda = 0.01) \ = \ 1 - (1 - e^{-0.01\tau}) \ = \ e^{-0.01\tau} \, .$$

2.Set

$$\mathbf{P}(t > \tau) \ = \ e^{-0.01\tau} \ = \ \frac{1}{2}$$

and solve for τ to get then $\tau \ = \ -100 \times \log(0.5) \ = \ 69.3$ (3 sig. fig.) .

4.Occurrence of lacunae may not always be independent. For example, a machine malfunction may cause them to be clumped.

Answer (Exercise 3.17) — Let $X \sim \text{Exponential}(\lambda = 0.1)$ denote the time taken to serve any given customer in an IID manner. Let y denote the unknown time that the current customer being served has already been served before your arrival. By memorylessness of $\text{Exponential}(\lambda = 0.1)$ RV X, we know $\mathbf{P}(X > 2 + y | X > y) = \mathbf{P}(X > 2) = e^{-\lambda 2} = e^{-2/10}$.

Answer (Exercise 3.18) —

(a) Since $f(x)$ is a density function which integrates to one,

$$\begin{aligned} \int_2^6 f(x)\,dx &= \int_2^6 k\,dx \\ 1 &= kx\big]_2^6 \\ 1 &= 6k - 2k \\ 1 &= 4k \\ k &= \frac{1}{4} \end{aligned}$$

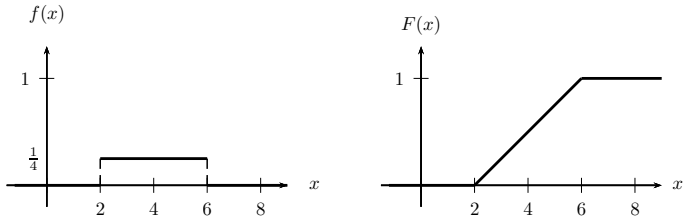
as expected!

(b) Now

$$F(x) = \begin{cases} 0 & x < 2 \\ \frac{1}{4}(x-2) & 2 \le x < 6 \\ 1 & x \ge 6 \ . \end{cases}$$

so the graphs are:

Graphs of $f(x)$ and $F(x)$.



Answer (Ex. 3.19) — 1.Since $f(x)$ is a (continuous) probability density function which integrates to one,

$$\int_{-4}^{4} k dx = 1\,.$$

Figure 1.3: A function f ("father of") from $\mathbb{X}$ (a set of children) to $\mathbb{Y}$ (their fathers) and its inverse ("children of").

as the **inverse** of f.

We motivated the non-negative integers $\mathbb{Z}_+$ via the size of a set. With the notion of two directions ($+$ and $-$) and the magnitude of the current position from the origin zero (0) of a dynamic entity, we can motivate the set of **integers**:

$$\boxed{\mathbb{Z} := \{\ldots, -3, -2, -1, 0, +1, +2, +3, \ldots\}}\,.$$

The integers with a **minus** or **negative sign** ($-$) before them are called negative integers and those with a **plus** or **positive sign** ($+$) before them are called positive integers. Conventionally, $+$ signs are dropped. Some examples of functions you may have encountered are **arithmetic operations** such as **addition** ($+$), **subtraction** ($-$), **multiplication** ($\cdot$) and **division** ($/$) of ordered pairs of integers. The reader is assumed to be familiar with such arithmetic operations with pairs of integers. Every integer is either positive, negative, or zero. In terms of this we define the notion of **order**. We say an integer a is **less than** an integer b and write $a < b$ if $b - a$ is positive. We say an integer a is **less than or equal to** an integer b and write $a \le b$ if $b - a$ is positive or zero. Finally, we say that a is greater than b and write $a > b$ if $b < a$. Similarly, a is greater than equal to b, i.e. $a \ge b$, if $b \le a$. The set of integers are **well-ordered**, i.e., for every integer a there is a next largest integer $a + 1$.

Classwork 4 (Addition over integers) Consider the set of integers $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$. Try to set up the arithmetic operation of addition as a function. The domain for addition is the Cartesian product of $\mathbb{Z}$:

$$\mathbb{Z}^2 := \mathbb{Z} \times \mathbb{Z} := \{(a, b) : a \in \mathbb{Z}, b \in \mathbb{Z}\}$$

What is its range ?

$$+ : \mathbb{Z} \times \mathbb{Z} \to$$

If the magnitude of the entity's position is measured in units (e.g. meters) that can be rationally divided into q pieces with $q \in \mathbb{N}$, then we have the set of rational numbers:

$$\mathbb{Q} := \{p/q : p \in \mathbb{Z}, q \in \mathbb{Z} \setminus \{0\}\}$$

The expressions p/q and p'/q' denote the same rational number if and only if $p \cdot q' = p' \cdot q$. Every rational number has a unique irreducible expression p/q, where q is positive and as small as possible. For example, 1/2, 2/4, 3/6, and 1001/2002 are different expressions for the same rational number whose irreducible unique expression is 1/2.

Figure 1.4: A pictorial depiction of addition and its inverse. The domain is plotted in orthogonal **Cartesian coordinates**.

Addition and multiplication are defined for rational numbers by:

$$\frac{p}{q} + \frac{p'}{q'} = \frac{p \cdot q' + p' \cdot q}{q \cdot q'} \qquad \text{and} \qquad \frac{p}{q} \cdot \frac{p'}{q'} = \frac{p \cdot p'}{q \cdot q'} \; .$$

The rational numbers form a **field** under the operations of addition and multiplication defined above in terms of addition and multiplication over integers. This means that the following properties are satisfied:

1. Addition and multiplication are each **commutative**

$$a + b = b + a, \qquad a \cdot b = b \cdot a \; ,$$

and associative

$$a + (b + c) = (a + b) + c, \qquad a \cdot (b \cdot c) = (a \cdot b) \cdot c \; .$$

2. Multiplication **distributes** over addition

$$a \cdot (b + c) = (a \cdot b) + (a \cdot c) \; .$$

3. 0 is the **additive identity** and 1 is the multiplicative identity

$$0 + a = a \qquad \text{and} \qquad 1 \cdot a = a \; .$$

4. Every rational number a has a negative, $a + (-a) = 0$ and every non-zero rational number a has a reciprocal, $a \cdot 1/a = 1$.

The field axioms imply the usual laws of arithmetic and allow subtraction and division to be defined in terms of addition and multiplication as follows:

$$\frac{p}{q} - \frac{p'}{q'} := \frac{p}{q} + \frac{-p'}{q'} \qquad \text{and} \qquad \frac{p}{q} \Big/ \frac{p'}{q'} := \frac{p}{q} \cdot \frac{q'}{p'}, \quad \text{provided } p' \neq 0 \; .$$

We will see later that the theory of finite fields is necessary for the study of pseudo-random number generators (PRNGs) and PRNGs are the heart-beat of randomness and statistics with computers.

Answer (Ex. 3.13) — This is a Binomial experiment with parameters $\theta = 0.1$ and $n = 10$, and so

$$\mathbf{P}(X \geq 1) = 1 - \mathbf{P}(X < 1) = 1 - \mathbf{P}(X = 0) \; ,$$

where

$$\mathbf{P}(X = 0) = \binom{10}{0} 0.1^0 0.9^{10} \approxeq 0.3487 \; .$$

Therefore, the probability that the target will be hit at least once is

$$1 - 0.3487 \approxeq 0.6513 \; .$$

Answer (Ex. 3.14) — Since 2 defects exist on every 100 meters, we would expect 6 defects on a 300 meter tape. If X is the number of defects on a 300 meter tape, then X is Poisson with $\lambda = 6$ and so the probability of zero defects is

$$\mathbf{P}(X = 0; 6) \; = \; \frac{6^0}{0!} e^{-6} \; = \; 0.0025 \; .$$

Answer (Ex. 3.15) — Since X is Poisson(λ) random variable with $\lambda = 0.5$, $\mathbf{P}(X \geq 2)$ is the probability of observing two or more particles during any given second.

$$\mathbf{P}(X \geq 2) \; = \; 1 - \mathbf{P}(X < 2) \; = \; 1 - \mathbf{P}(X = 1) - \mathbf{P}(X = 0) \; ,$$

where $\mathbf{P}(X = 1)$ and $\mathbf{P}(X = 0)$ can be carried out by the Poisson probability mass function

$$\mathbf{P}(X = x) \; = \; f(x) \; = \; \frac{\lambda^x}{x!} e^{-\lambda} \; .$$

Now

$$\mathbf{P}(X = 0) \; = \; \frac{0.5^0}{0!} \times e^{-0.5} \; = \; 0.6065$$

and

$$\mathbf{P}(X = 1) \; = \; \frac{0.5^1}{1!} \times e^{-0.5} \; = \; 0.3033$$

and so

$$\mathbf{P}(X \geq 2) \; = \; 1 - 0.9098 \; = \; 0.0902 \; .$$

Answer (Ex. 3.16) — 1.The Probability mass function for Poisson(λ) random variable X is

$$\mathbf{P}(X = x) \; = \; f(x; \lambda) \; = \; \frac{e^{-\lambda} \lambda^x}{x!}$$

where λ is the mean number of lacunae per specimen and X is the random variable "number of lacunae on a specimen".

2.If $x = 0$ then $x! = 0! = 1$ and $\lambda^x = \lambda^0 = 1$, and the formula becomes $\mathbf{P}(X = 0) \; = \; e^{-\lambda}$.

3.Since $\mathbf{P}(X \geq 1) = 0.1$,

$$\mathbf{P}(X = 0) \; = \; 1 - \mathbf{P}(X \geq 1) \; = \; 0.9 \, .$$

Using (b) and solving for λ gives:

$$e^{-\lambda} \; = \; 0.9 \quad \text{that is,} \quad \lambda \; = \; -\ln(0.9) \; = \; 0.1 \text{ (approximately .)}$$

Hence

$$\mathbf{P}(X = 2) \; = \; \frac{e^{-0.1}(0.1)^2}{2!} \; = \; 0.45\% \text{ (approximately .)}$$

$$\mathbf{P}(X \geq 1) = 1 - \mathbf{P}(X = 0) = 1 - f(0) = \frac{15}{16}$$
$$\mathbf{P}(X \leq 3) = f(0) + f(1) + f(2) + f(3) = \frac{15}{16}$$

Answer (Ex. 3.12) — 1.If the random variable X denotes the number of type AB blood donors in the sample of 15, then X has a binomial distribution with $n = 15$ and $\theta = 0.05$. Therefore

$$\mathbf{P}(X = 1) \;=\; \binom{15}{1}(0.05)^1(0.95)^{14} \;=\; 0.366 \quad \text{(3 sig. fig.)}\,.$$

2.If the random variable X denotes the number of type B blood donors in the sample of 15, then X has a binomial distribution with $n = 15$ and $\theta = 0.10$. Therefore

$$\begin{aligned}\mathbf{P}(X \geq 3) \;&=\; 1 \;-\; \mathbf{P}(X = 0) \;-\; \mathbf{P}(X = 1) \;-\; \mathbf{P}(X = 2)\\ &=\; 1 \;-\; \binom{15}{0}(0.1)^0(0.9)^{15} \;-\; \binom{15}{1}(0.1)^1(0.9)^{14} \;-\; \binom{15}{2}(0.1)^2(0.9)^{13}\\ &=\; 1 \;-\; 0.2059 \;-\; 0.3432 \;-\; 0.2669\\ &=\; 0.184 \quad \text{(to 3 sig. fig.)}\end{aligned}$$

3.If the random variable X denotes the number of type O or type A blood donors in the sample of 15, then X has a binomial distribution with $n = 15$ and $\theta = 0.85$. Therefore

$$\begin{aligned}\mathbf{P}(X > 10) \;&=\; \mathbf{P}(X = 11) \;+\; \mathbf{P}(X = 12) \;+\; \mathbf{P}(X = 13) \;+\; \mathbf{P}(X = 14) \;+\; \mathbf{P}(X = 15)\\ &=\; \binom{15}{11}(0.85)^{11}(0.15)^4 \;+\; \binom{15}{12}(0.85)^{12}(0.15)^3\\ &+\; \binom{15}{13}(0.85)^{13}(0.15)^2 \;+\; \binom{15}{14}(0.85)^{14}(0.15)^1 \;+\; \binom{15}{15}(0.85)^{15}(0.15)^0\\ &=\; 0.1156 \;+\; 0.2184 \;+\; 0.2856 \;+\; 0.2312 \;+\; 0.0874\\ &=\; 0.938 \quad \text{(to 3 sig. fig.)}\end{aligned}$$

4.If the random variable X denotes the number of blood donors that are *not* of type A blood donors in the sample of 15, then X has a binomial distribution with $n = 15$ and $\theta = 0.6$. Therefore

$$\begin{aligned}\mathbf{P}(X < 5) \;&=\; \mathbf{P}(X = 0) \;+\; \mathbf{P}(X = 1) \;+\; \mathbf{P}(X = 2) \;+\; \mathbf{P}(X = 3) \;+\; \mathbf{P}(X = 4)\\ &=\; \binom{15}{0}(0.6)^0(0.4)^{15} \;+\; \binom{15}{1}(0.6)^1(0.4)^{14} \;+\; \binom{15}{2}(0.6)^2(0.4)^{13}\\ &+\; \binom{15}{3}(0.6)^3(0.4)^{12} \;+\; \binom{15}{4}(0.6)^4(0.4)^{11}\\ &=\; 0.0000 \;+\; 0.0000 \;+\; 0.0003 \;+\; 0.0016 \;+\; 0.0074\\ &=\; 0.009 \quad \text{(to 3 DP.)}\end{aligned}$$

1.4 Real Numbers

Unlike rational numbers which are expressible in their reduced forms by p/q, it is fairly tricky to define or express real numbers. It is possible to define real numbers formally and constructively via equivalence classes of Cauchy sequence of rational numbers. For this all we need are notions of (1) infinity, (2) sequence of rational numbers and (3) distance between any two rational numbers in an infinite sequence of them. These are topics usually covered in an introductory course in real analysis and are necessary for a firm foundation in computational statistics. Instead of a formal constructive definition of real numbers, we give a more concrete one via decimal expansions. See Donald E. Knuth's treatment [*Art of Computer Programming, Vol. I, Fundamental Algorithms*, 3rd Ed., 1997, pp. 21-25] for a fuller story. A **real number** is a numerical quantity x that has a decimal expansion:

$$x = n + 0.d_1d_2d_3\ldots \;,\; \text{where, each } d_i \in \{0, 1, \ldots, 9\}, n \in \mathbb{Z}\;,$$

and the sequence $0.d_1d_2d_3\ldots$ does not terminate with infinitely many consecutive 9s. By the above decimal representation, the following arbitrarily accurate enclosure of the real number x by rational numbers is implied:

$$n + \frac{d_1}{10} + \frac{d_2}{100} + \cdots + \frac{d_k}{10^k} =: \underline{x}_k \leq x < \overline{x}_k := n + \frac{d_1}{10} + \frac{d_2}{100} + \cdots + \frac{d_k}{10^k} + \frac{1}{10^k}$$

for every $k \in \mathbb{N}$. Thus, rational arithmetic $(+, -, \cdot, /)$ can be extended with arbitrary precision to any ordered pair of real numbers x and y by operations on their rational enclosures $\underline{x}, \overline{x}$ and $\underline{y}, \overline{y}$.

Some examples of real numbers that are not rational (**irrational numbers**) are:

$$\begin{aligned}\sqrt{2} &= 1.41421356237309\ldots\text{.the side length of a square with area of 2 units}\\ \pi &= 3.14159265358979\ldots\text{.the ratio of the circumference to diameter of a circle}\\ e &= 2.71828182845904\ldots\text{.Euler's constant}\end{aligned}$$

We can think of π as being enclosed by the following pairs of rational numbers:

$$\begin{aligned}3 + \frac{1}{10} &=: \underline{\pi}_1 \leq \pi < \overline{\pi}_1 := 3 + \frac{1}{10} + \frac{1}{10^1}\\ 3 + \frac{1}{10} + \frac{4}{100} &=: \underline{\pi}_2 \leq \pi < \overline{\pi}_2 := 3 + \frac{1}{10} + \frac{4}{100} + \frac{1}{100}\\ 3 + \frac{1}{10} + \frac{4}{100} + \frac{1}{10^3} &=: \underline{\pi}_3 \leq \pi < \overline{\pi}_3 := 3 + \frac{1}{10} + \frac{4}{100} + \frac{1}{10^3} + \frac{1}{10^3}\\ &\vdots\\ 3.14159265358979 &=: \underline{\pi}_{14} \leq \pi < \overline{\pi}_{14} := 3.14159265358979 + \frac{1}{10^{14}}\\ &\vdots\end{aligned}$$

Think of the real number system as the continuum of points that make up a line, as shown in Figure 1.5.

Let y and z be two real numbers such that $y \leq z$. Then, the **closed interval** $[y, z]$ is the set of real numbers x such that $y \leq x \leq z$:

$$[y, z] := \{x : y \leq x \leq z\}\;.$$

Figure 1.5: A depiction of the real line segment $[-10, 10]$.

The **half-open interval** $(y, z]$ or $[y, z)$ and the **open interval** (y, z) are defined analogously:

$$(y, z] := \{x : y < x \leq z\} ,$$
$$[y, z) := \{x : y \leq x < z\} ,$$
$$(y, z) := \{x : y < x < z\} .$$

We also allow y to be **minus infinity** (denoted $-\infty$) or z to be **infinity** (denoted ∞) at an open endpoint of an interval, meaning that there is no lower or upper bound. With this allowance we get the set of **real numbers** $\mathbb{R} := (-\infty, \infty)$, the **non-negative real numbers** $\mathbb{R}_+ := [0, \infty)$ and the **positive real numbers** $\mathbb{R}_{>0}(0, \infty)$ as follows:

$$\mathbb{R} := (-\infty, \infty) = \{x : -\infty < x < \infty\} ,$$
$$\mathbb{R}_+ := [0, \infty) = \{x : 0 \leq x < \infty\} ,$$
$$\mathbb{R}_{>0} := [0, \infty) = \{x : 0 < x < \infty\} .$$

For a positive real number $b \in \mathbb{R}_{>0}$ and an integer $n \in \mathbb{Z}$, the n-th **power** or **exponent** of b is:

$$b^0 = 1, \qquad b^n = b^{n-1} \cdot b \quad \text{if } n > 0, \qquad b^n = b^{n+1}/b \quad \text{if } n < 0 .$$

The following **laws of exponents** hold by mathematical induction when $m, n \in \mathbb{Z}$:

$$b^{m+n} = b^m \cdot b^n, \qquad (b^m)^n = b^{m \cdot n} .$$

If $y \in \mathbb{R}$ and $m \in \mathbb{N}$, the unique positive real number $z \in \mathbb{R}_{>0}$ such that $z^m = y$ is called the **m-th root of** y and denoted by $\sqrt[m]{y}$, i.e.,

$$z^m = y \implies z = \sqrt[m]{y} .$$

For a rational number $r = p/q \in \mathbb{Q}$, we define the r-th power of $b \in \mathbb{R}$ as follows:

$$b^r = b^{p/q} := \sqrt[q]{b^p}$$

The laws of exponents hold for this definition and different expressions for the same rational number $r = ap/aq$ yield the same power, i.e., $b^{p/q} = b^{ap/aq}$. Recall that a real number $x = n + 0.d_1d_2d_3\ldots \in \mathbb{R}$ can be arbitrarily precisely enclosed by the rational numbers $\underline{x}_k := n + \frac{d_1}{10} + \frac{d_2}{100} + \cdots + \frac{d_k}{10^k}$ and $\overline{x}_k := n + \frac{d_1}{10} + \frac{d_2}{100} + \cdots + \frac{d_k}{10^k} + \frac{1}{10^k}$ by increasing k. Suppose first that $b > 1$. Then, using rational powers, we can enclose b^x,

$$b^{n+\frac{d_1}{10}+\frac{d_2}{100}+\cdots+\frac{d_k}{10^k}} =: b^{\underline{x}_k} \leq b^x < b^{\overline{x}_k} := b^{n+\frac{d_1}{10}+\frac{d_2}{100}+\cdots+\frac{d_k}{10^k}+\frac{1}{10^k}} ,$$

within an interval of width $b^{n+\frac{d_1}{10}+\frac{d_2}{100}+\cdots+\frac{d_k}{10^k}}\left(b^{\frac{1}{10^k}} - 1\right) < b^{n+1}(b-1)/10^k$. By taking a large enough k we can evaluate b^x to any accuracy. Finally, when $b < 1$ we define $b^x := (1/b)^x$ and when $b = 0$, $b^x := 1$.

Now $\sum_{x=0}^{\infty} \frac{1}{2^x}$ is a geometric series with common ratio $r = \frac{1}{2}$ and first term $a = 1$, and so has sum

$$S = \frac{a}{1-r} = \frac{1}{1-\frac{1}{2}} = 2$$

Therefore,

$$2k = 1 , \quad \text{that is,} \quad k = \frac{1}{2} .$$

2.From (a), the probability mass function of f is

$$f(x) = \frac{\frac{1}{2}}{2^x} = \frac{1}{2^{x+1}} . \quad (x = 0, 1, 2, \ldots)$$

Now

$$\mathbf{P}(X \geq 4) = 1 - \mathbf{P}(X < 4) = 1 - \mathbf{P}(X \leq 3)$$

where

$$\begin{aligned} \mathbf{P}(X \leq 3) &= \sum_{x=0}^{3} \frac{1}{2^{x+1}} \\ &= \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} \\ &= \frac{8}{16} + \frac{4}{16} + \frac{2}{16} + \frac{1}{16} \\ &= \frac{15}{16} . \end{aligned}$$

That is, $\mathbf{P}(X \geq 4) = \frac{1}{16}$.

Answer (Ex. 3.11) — Note that $\theta = \frac{1}{2}$ here.

1.X has probability mass function

$$f(x) = \begin{cases} \binom{4}{0}\frac{1}{2}^0\frac{1}{2}^4 = \frac{1}{16} & x = 0 \\ \binom{4}{1}\frac{1}{2}^1\frac{1}{2}^3 = \frac{4}{16} & x = 1 \\ \binom{4}{2}\frac{1}{2}^2\frac{1}{2}^2 = \frac{6}{16} & x = 2 \\ \binom{4}{3}\frac{1}{2}^3\frac{1}{2}^1 = \frac{4}{16} & x = 3 \\ \binom{4}{4}\frac{1}{2}^4\frac{1}{2}^0 = \frac{1}{16} & x = 4 \end{cases}$$

2.The required probabilities are:

$\mathbf{P}(X = 0) = f(0) = \frac{1}{16}$

$\mathbf{P}(X = 1) = f(1) = \frac{4}{16}$

Answer (Ex. 3.8) — (a)

x	3	4	5	6	7	8	9	10	11	12	13
$F(x) = \mathbf{P}(X \le x)$	0.07	0.08	0.17	0.18	0.34	0.59	0.79	0.82	0.84	0.95	1.00

(b) (i)$\mathbf{P}(X \le 5) = F(5) = 0.17$

(ii)$\mathbf{P}(X < 12) = \mathbf{P}(X \le 11) = F(11) = 0.84$

(iii)$\mathbf{P}(X > 9) = 1 - \mathbf{P}(X \le 9) = 1 - F(9) = 1 - 0.79 = 0.21$

(iv)$\mathbf{P}(X \ge 9) = 1 - \mathbf{P}(X < 9) = 1 - \mathbf{P}(X \le 8) = 1 - 0.59 = 0.41$

(v)$\mathbf{P}(4 < X \le 9) = F(9) - F(4) = 0.79 - 0.08 = 0.71$

(vi)$\mathbf{P}(4 < X < 11) = \mathbf{P}(4 < X \le 10) = F(10) - F(4) = 0.82 - 0.08 = 0.74$

Answer (Ex. 3.9) — Since we are sampling without replacement,

$$\begin{aligned}
\mathbf{P}(X=0) &= \frac{4}{10}\cdot\frac{3}{9} = \frac{2}{15} \quad \text{(one way of drawing two right screws)},\\
\mathbf{P}(X=1) &= \frac{6}{10}\cdot\frac{4}{9} + \frac{4}{10}\cdot\frac{6}{9} = \frac{8}{15} \quad \text{(two ways of drawing one left and one right screw)},\\
\mathbf{P}(X=2) &= \frac{6}{10}\cdot\frac{5}{9} = \frac{1}{3} \quad \text{(one way of drawing two left screws)}.
\end{aligned}$$

So the probability mass function of X is:

$$f(x) \;=\; \mathbf{P}(X=x) \;=\; \begin{cases} \frac{2}{15} & \text{if } x=0 \\ \frac{8}{15} & \text{if } x=1 \\ \frac{1}{3} & \text{if } x=2 \end{cases}$$

The required probabilities are:

1.
$$\mathbf{P}(X \le 1) \;=\; \mathbf{P}(X=0) + \mathbf{P}(X=1) \;=\; \frac{2}{15} + \frac{8}{15} \;=\; \frac{2}{3}$$

2.
$$\mathbf{P}(X \ge 1) \;=\; \mathbf{P}(X=1) + \mathbf{P}(X=2) \;=\; \frac{8}{15} + \frac{1}{3} \;=\; \frac{13}{15}$$

3.
$$\mathbf{P}(X > 1) \;=\; \mathbf{P}(X=2) \;=\; \frac{1}{3}$$

Answer (Ex. 3.10) — 1.Since f is a probability mass function,

$$\sum_{x=0}^{\infty} \frac{k}{2^x} \;=\; 1\,, \qquad \text{that is,} \qquad k\sum_{x=0}^{\infty} \frac{1}{2^x} \;=\; 1 \quad .$$

Suppose $y \in \mathbb{R}_{>0}$ and $b \in \mathbb{R} \setminus \{1\}$ then the real number x such that $y = b^x$ is called the **logarithm of y to the base** b and we write this as:

$$y = b^x \iff x = \log_b y$$

The definition implies:

$$x = \log_b(b^x) = b^{\log_b x}\,,$$

and the laws of exponents imply:

$$\begin{aligned}
&\log_b(xy) = \log_b x + \log_b y, \quad \text{if} \quad x>0,\ y>0 \ \text{ and}\\
&\log_b(c^y) = y\log_b c, \quad \text{if} \quad c>0\ .
\end{aligned}$$

The **common logarithm** is $\log_{10}(y)$, the **binary logarithm** is $\log_2(y)$ and the **natural logarithm** is $\log_e(y)$, where e is the Euler's constant. Since we will mostly work with $\log_e(y)$ we use $\log(y)$ to mean $\log_e(y)$. You are assumed to be familiar with trigonometric functions ($\sin(x)$, $\cos(x)$, $\tan(x)$, ...). We sometimes denote the special power function e^y by $\exp(y)$.

Familiar extremal elements of a set of real numbers, say A, are the following:

$$\boxed{\max A := \text{greatest element in } A}$$

For example, $\max\{1, 4, -9, 345\} = 345$, $\max[-93.8889, 1002.786] = 1002.786$.

$$\boxed{\min A := \text{least element in } A}$$

For example, $\min\{1, 4, -9, 345\} = -9$, $\min[-93.8889, 1002.786] = -93.8889$. We need a slightly more sophisticated notion for the extremal elements of a set A that may not belong to A. We say that a real number x is a **lower bound** for a non-empty set of real numbers A, provided $x \le a$ for every $a \in A$. We say that the set A is **bounded below** if it has at least one lower bound. A lower bound is the **greatest lower bound** if it is at least as large as any other lower bound. The greatest lower bound of a set of real numbers A is called the **infimum** of A and is denoted by:

$$\boxed{\inf A := \text{greatest lower bound of } A}$$

For example, $\inf(0, 1) = 0$ and $\inf\{10.333 \cup [-99, 1001.33)\} = -99$. We similarly define the **least upper bound** of a non-empty set of real numbers A to be the **supremum** of A and denote it as:

$$\boxed{\sup A := \text{least upper bound of } A}$$

For example, $\sup(0, 1) = 1$ and $\sup\{10.333 \cup [-99, 1001.33)\} = 1001.33$. By convention, we define $\inf\emptyset := \infty$, $\sup\emptyset := -\infty$. Finally, if a set A is not bounded below then $\inf A := -\infty$ and if a set A is not bounded above then $\sup A := \infty$.

Symbol	Meaning
$A = \{\star, \circ, \bullet\}$	A is a set containing the elements $\star$, $\circ$ and $\bullet$
$\circ \in A$	$\circ$ belongs to A or $\circ$ is an element of A
$A \ni \circ$	$\circ$ belongs to A or $\circ$ is an element of A
$\odot \notin A$	$\odot$ does not belong to A
$\#A$	Size of the set A, for e.g. $\#\{\star, \circ, \bullet, \odot\} = 4$
$\mathbb{N}$	The set of natural numbers $\{1, 2, 3, \ldots\}$
$\mathbb{Z}$	The set of integers $\{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$
$\mathbb{Z}_+$	The set of non-negative integers $\{0, 1, 2, 3, \ldots\}$
$\emptyset$	Empty set or the collection of nothing or $\{\}$
$A \subset B$	A is a subset of B or A is contained by B, e.g. $A = \{\circ\}, B = \{\bullet\}$
$A \supset B$	A is a superset of B or A contains B e.g. $A = \{\circ, \star, \bullet\}, B = \{\circ, \bullet\}$
$A = B$	A equals B, i.e. $A \subset B$ and $B \subset A$
$Q \implies R$	Statement Q implies statement R or If Q then R
$Q \iff R$	$Q \implies R$ and $R \implies Q$
$\{x : x \text{ satisfies property } R\}$	The set of all x such that x satisfies property R
$A \cup B$	A union B, i.e. $\{x : x \in A \text{ or } x \in B\}$
$A \cap B$	A intersection B, i.e. $\{x : x \in A \text{ and } x \in B\}$
$A \setminus B$	A minus B, i.e. $\{x : x \in A \text{ and } x \notin B\}$
$A := B$	A is equal to B by definition
$A =: B$	B is equal to A by definition
A^c	A complement, i.e. $\{x : x \in U, \text{ the universal set, but } x \notin A\}$
$A_1 \times A_2 \times \cdots \times A_m$	The m-product set $\{(a_1, a_2, \ldots, a_m) : a_1 \in A_1, a_2 \in A_2, \ldots, a_m \in A_m\}$
A^m	The m-product set $\{(a_1, a_2, \ldots, a_m) : a_1 \in A, a_2 \in A, \ldots, a_m \in A\}$
$f := f(x) = y : \mathbb{X} \to \mathbb{Y}$	A function f from domain $\mathbb{X}$ to range $\mathbb{Y}$
$f^{[-1]}(y)$	Inverse image of y
$f^{[-1]} := f^{[-1]}(y \in \mathbb{Y}) = X \subset \mathbb{X}$	Inverse of f
$a < b$ or $a \leq b$	a is less than b or a is less than or equal to b
$a > b$ or $a \geq b$	a is greater than b or a is greater than or equal to b
$\mathbb{Q}$	Rational numbers
(x, y)	the open interval (x, y), i.e. $\{r : x < r < y\}$
$[x, y]$	the closed interval (x, y), i.e. $\{r : x \leq r \leq y\}$
$(x, y]$	the half-open interval $(x, y]$, i.e. $\{r : x < r \leq y\}$
$[x, y)$	the half-open interval $[x, y)$, i.e. $\{r : x \leq r < y\}$
$\mathbb{R} := (-\infty, \infty)$	Real numbers, i.e. $\{r : -\infty < r < \infty\}$
$\mathbb{R}_+ := [0, \infty)$	Real numbers, i.e. $\{r : 0 \leq r < \infty\}$
$\mathbb{R}_{>0} := (0, \infty)$	Real numbers, i.e. $\{r : 0 < r < \infty\}$

Table 1.1: Symbol Table: Sets and Numbers

x	observed frequency	observed relative frequency	Prob of x hits
0	229	$229/576 = 0.398$	$f(0; 0.933) = 0.394$
1	211	$211/576 = 0.366$	$f(1; 0.933) = 0.367$
2	93	$93/576 = 0.161$	$f(2; 0.933) = 0.171$
3	35	$35/576 = 0.0608$	$f(3; 0.933) = 0.0532$
4	7	$7/576 = 0.0122$	$f(4; 0.933) = 0.0124$
≥ 5	1	$1/576 = 0.00174$	$1 - \sum_{x=0}^{4} f(x; 0.933) = 0.00275$

Answer (Ex. 3.5) — $\mathbf{P}(X = 3)$ does not satisfy the condition that $0 \leq \mathbf{P}(A) \leq 1$ for any event A. If Ω is the sample space, then $\mathbf{P}(\Omega) = 1$ and so the correct probability is

$$\mathbf{P}(X = 3) \;=\; 1 - 0.07 - 0.10 - 0.32 - 0.40 \;=\; 0.11 \;.$$

Answer (Ex. 3.6) — 1.Tabulate the values for the probability mass function as follows:

x	1	2	3	4	5
$\mathbf{P}(X = x)$	0.1	0.2	0.2	0.2	0.3

so the distribution function is:

$$F(x) \;=\; \mathbf{P}(X \leq x) \;=\; \begin{cases} 0 & \text{if } 0 \leq x < 1 \\ 0.1 & \text{if } 1 \leq x < 2 \\ 0.3 & \text{if } 2 \leq x < 3 \\ 0.5 & \text{if } 3 \leq x < 4 \\ 0.7 & \text{if } 4 \leq x < 5 \\ 1 & \text{if } x \geq 5 \end{cases}$$

2.The probability that the machine needs to be replaced during the first 3 years is:

$$\mathbf{P}(X \leq 3) \;=\; \mathbf{P}(X = 1) + \mathbf{P}(X = 2) + \mathbf{P}(X = 3) \;=\; 0.1 + 0.2 + 0.2 \;=\; 0.5\,.$$

(This answer is easily seen from the distribution function of X.)

3.The probability that the machine needs no replacement during the first three years is

$$\mathbf{P}(X > 3) \;=\; 1 - \mathbf{P}(X \leq 3) \;=\; 0.5\,.$$

Answer (Ex. 3.7) — Assuming that the probability model is being built from the observed relative frequencies, the probability mass function is:

$$f(x) \;=\; \begin{cases} \frac{176}{200} & x = 1 \\ \frac{22}{200} & x = 2 \\ \frac{2}{200} & x = 3 \end{cases}$$

Answer (Exercise 3.1) — The first mistake is the solid vertical lines (blue) from 0 in the domain or x-axis to $\mathbf{P}$('not A') in the range or y-axis and from 1 in the domain to 1 in the range. This is ill-defined for any function if we are to interpret that the elements in the domain, namely 0 and 1, are to be associated with the uncountably many image values in the range of the function, namely $[0, \mathbf{P}(\text{'not A'})]$ and $[\mathbf{P}(\text{'not A'}), 1]$, respectively. So we should first replace them by dotted lines which merely help us track where the function jumped to at 0 and 1.

The second mistake is failing to emphasise that the value taken by the function at 0 and 1 is not 0 and $\mathbf{P}$('not A'), respectively. So it is best to introduce an empty circle like $\circ$ at $(0, 0)$ and $(1, \mathbf{P}(\text{'not A'}))$ to indicate the points of discontinuity. The same mistakes should be fixed in the next Figure 3.2.

Answer (Exercise 3.2) — The probability that X takes on a specific value x is:

$$P(X = x) = P(\{\omega : X(\omega) = x\}) = \begin{cases} P(\emptyset) = 0, & \text{if } x \notin \{0, 1\} \\ P(\{\mathsf{T}\}) = \frac{1}{2}, & \text{if } x = 0 \\ P(\{\mathsf{H}\}) = \frac{1}{2}, & \text{if } x = 1 \end{cases}$$

or more simply,

$$P(X = x) = \begin{cases} \frac{1}{2} & \text{if } x = 0 \\ \frac{1}{2} & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

The distribution function for X is:

$$F(x) \;=\; P(X \leq x) \;=\; P(\{\omega : X(\omega) \leq x\}) \;=\; \begin{cases} P(\emptyset) = 0, & \text{if } -\infty < x < 0 \\ P(\{\mathsf{T}\}) = \frac{1}{2}, & \text{if } 0 \leq x < 1 \\ P(\{\mathsf{H}, \mathsf{T}\}) = P(\Omega) = 1, & \text{if } 1 \leq x < \infty \end{cases}$$

or more simply,

$$F(x) \;=\; P(X \leq x) \;=\; \begin{cases} 0, & \text{if } -\infty < x < 0 \\ \frac{1}{2}, & \text{if } 0 \leq x < 1 \\ 1, & \text{if } 1 \leq x < \infty \end{cases}$$

Answer (Exercise 3.3) — This was done in week 1. Get notes from your mates or wait until Raaz scribes for virtual convenience.

Answer (Exercise 3.4) — We are given that 537 flying bombs hit an area A of south London made up of $24 \times 24 = 576$ small equal-sized areas, say $A_1, A_2, \ldots, A_{576}$. Assuming the hits were purely random over A the probability that a particular bomb will hit a given small area, say A_i, is $\frac{1}{576}$. Let X denote the number of hits that a small area A_i receives in this German raid. Since 537 bombs fell over A, we can model X as Binomial$(n = 537, \theta = \frac{1}{576})$ that is counting the number of 'successes' (for German bombers) with probability θ in a sequence of $n = 537$ independent Bernoulli(θ) trials. Finally, we can approximate this Binomial$(n = 537, \theta = \frac{1}{576})$ random variable by Poisson(λ) random variable with $\lambda = n\theta = \frac{537}{576} \cong 0.933$. Using the probability mass function formula for Poisson$(\lambda = 0.933)$ random variable X we can obtain the probabilities and compare them with the relative frequencies from the data as follows:

1.5 Introduction to MATLAB

We use MATLAB to perform computations and visualisations. MATLAB is a numerical computing environment and programming language that is optimised for vector and matrix processing. STAT 218/313 students will have access to Maths & Stats Department's computers that are licensed to run MATLAB . You can remotely connect to these machines from home by following instructions at `http://www.math.canterbury.ac.nz/php/resources/compdocs/remote`.

Labwork 5 (Basics of MATLAB) Let us familiarize ourselves with MATLAB in this session. First, you need to launch MATLAB from your terminal. Since this is system dependent, ask your tutor for help. The command window within the MATLAB window is where you need to type commands. Here is a minimal set of commands you need to familiarize yourself with in this session.

1. Type the following command to add 2 numbers in the command window right after the command prompt `>>` .

```
>> 13+24
```

Upon hitting `Enter` or `Return` on your keyboard, you should see:

```
ans =

    37
```

The summand 37 of 13 and 24 is stored in the default variable called `ans` which is short for answer.

2. We can write **comments** in MATLAB following the `%` character. All the characters in a given line that follow the percent character `%` are ignored by MATLAB . It is very helpful to comment what is being done in the code. You won't get full credit without sufficient comments in your coding assignments. For example we could have added the comment to the previous addition. To save space in these notes, we suppress the blank lines and excessive line breaks present in MATLAB 's command window.

```
>> 13+24 % adding 13 to 24 using the binary arithmetic operator +
ans =    37
```

3. You can **create or reopen a diary file** in MATLAB to record your work. Everything you typed or input and the corresponding output in the command window will be recorded in the diary file. You can create or reopen a diary file by typing `diary filename.txt` in the command window. When you have finished recording, simply type `diary off` in the command window **to turn off the diary file**. The diary file with **.txt** extension is simply a text-file. It can be edited in different editors after the diary is turned off in MATLAB . You need to type `diary LabWeek1.txt` to start recording your work for electronic submission if needed.

```
>> diary blah.txt % start a diary file named blah.txt
>> 3+56
ans =    59
>> diary off % turn off the current diary file blah.txt
```

```
>> type blah.txt % this allows you to see the contents of blah.txt
3+56
ans =    59
diary off
>> diary blah.txt % reopen the existing diary file blah.txt
>> 45-54
ans =    -9
>> diary off % turn off the current diary file blah.txt again
>> type blah.txt % see its contents
3+56
ans =    59
diary off
45-54
ans =    -9
diary off
```

4. Let's learn to store values in variables of our choice. Type the following at the command prompt :

```
>> VariableCalledX = 12
```

Upon hitting enter you should see that the number 12 has been assigned to the variable named `VariableCalledX` :

```
VariableCalledX =    12
```

5. MATLAB stores default value for some variables, such as `pi` (π), `i` and `j` (complex numbers).

```
>> pi
ans =    3.1416
>> i
ans =        0 + 1.0000i
>> j
ans =        0 + 1.0000i
```

All predefined symbols (variables, constants, function names, operators, etc) in MATLAB are written in lower-case. Therefore, it is a good practice to name the variable you define using upper and mixed case letters in order to prevent an unintended overwrite of some predefined MATLAB symbol.

6. We could have stored the sum of 13 and 24 in the variable `X`, by entering:

```
>> X = 13 + 24
X =    37
```

7. Similarly, you can store the outcome of multiplication (via operation `*`), subtraction (via operation `-`), division (via `/`) and exponentiation (via `^`)of any two numbers of your choice in a variable name of your choice. Evaluate the following expressions in MATLAB :

$p = 45.89 * 1.00009$ $\qquad$ $d = 89.0/23.3454$

$m = 5376.0 - 6.00$ $\qquad$ $p = 2^{0.5}$

8. You may compose the elementary operations to obtain rational expressions by using parenthesis to specify the order of the operations. To obtain $\sqrt{2}$, you can type the following into MATLAB 's command window.

Answer (Ex. 2.11) — (a)Let $F1$ be the event a gale of force 1 occurs, let $F2$ be the event a gale of force 2 occurs and $F3$ be the event a gale of force 3 occurs. Now we know that

$$P(F1) = \frac{2}{3}, \quad P(F2) = \frac{1}{4}, \quad P(F3) = \frac{1}{12}.$$

If D is the event that a gale causes damage, then we also know the following conditional probabilities:

$$P(D|F1) = \frac{1}{4}, \quad P(D|F2) = \frac{2}{3}, \quad P(D|F3) = \frac{5}{6}.$$

The probability that a reported gale causes damage is

$$P(D) = P(D \cap F1) + P(D \cap F2) + P(D \cap F3)$$

where

$$P(D \cap F1) = P(D|F1)P(F1) = \frac{1}{4} \times \frac{2}{3} = \frac{1}{6},$$

$$P(D \cap F2) = P(D|F2)P(F2) = \frac{2}{3} \times \frac{1}{4} = \frac{1}{6},$$

and

$$P(D \cap F3) = P(D|F3)P(F3) = \frac{5}{6} \times \frac{1}{12} = \frac{5}{72}.$$

Hence

$$P(D) = \frac{1}{6} + \frac{1}{6} + \frac{5}{72} = \frac{29}{72}$$

(b)Knowing that the gale did cause damage we can calculate the probabilities that it was of the various forces using the probabilites in (a) as follows (Note: $P(D \cap F1) = P(F1 \cap D)$ etc.):

$$P(F1|D) = \frac{P(F1 \cap D)}{P(D)} = \frac{1/6}{29/72} = \frac{12}{29}$$

$$P(F2|D) = \frac{P(F2 \cap D)}{P(D)} = \frac{1/6}{29/72} = \frac{12}{29}$$

$$P(F3|D) = \frac{P(F3 \cap D)}{P(D)} = \frac{5/72}{29/72} = \frac{5}{29}$$

(c)First note that the probability that a reported gale does NOT cause damage is:

$$P(D^c) = 1 - P(D) = 1 - \frac{29}{72} = \frac{43}{72}.$$

Now we need to find probabilities like $P(F1 \cap D^c)$. The best way to do this is to use the partitioning idea of the "Total Probability Theorem", and write:

$$P(F1) = P(F1 \cap D^c) + P(F1 \cap D),$$

Rearranging this gives

$$P(F1 \cap D^c) = P(F1) - P(F1 \cap D)$$

and so

$$P(F1|D^c) = \frac{P(F1 \cap D^c)}{P(D^c)} = \frac{P(F1) - P(D \cap F1)}{P(D^c)} = \frac{2/3 - 1/6}{43/72} = \frac{36}{43}.$$

Similarly,

$$P(F2|D^c) = \frac{P(F2 \cap D^c)}{P(D^c)} = \frac{P(F2) - P(D \cap F2)}{P(D^c)} = \frac{1/4 - 1/6}{43/72} = \frac{6}{43},$$

and

$$P(F3|D^c) = \frac{P(F3 \cap D^c)}{P(D^c)} = \frac{P(F3) - P(D \cap F3)}{P(D^c)} = \frac{1/12 - 5/72}{43/72} = \frac{1}{43}.$$

Answer (Ex. 2.10) — Let the event that a micro-chip is defective be D, and the event that the test is correct be C. So the probability that the micro-chip is defective is $P(D) = 0.05$, and the probability that it is effective is $P(D^c) = 0.95$.
The probability that the test correctly detects a defective micro-chip is the conditional probability $P(C|D) = 0.8$, and the probability that if a good micro-chip is tested but the test declares it is defective is the conditional probability $P(C^c|D^c) = 0.1$. Therefore, we also have the probabilities $P(C^c|D) = 0.2$, and $P(C|D^c) = 0.9$.
Moreover, the probability that a micro-chip is defective, and has been declared as defective is

$$P(C \cap D) \;=\; P(C|D)P(D) \;=\; 0.8 \times 0.05 \;=\; 0.04\,.$$

The probability that a micro-chip is effective, and has been declared as effective is

$$P(C \cap D^c) \;=\; P(C|D^c)P(D^c) \;=\; 0.9 \times 0.95 \;=\; 0.855\,.$$

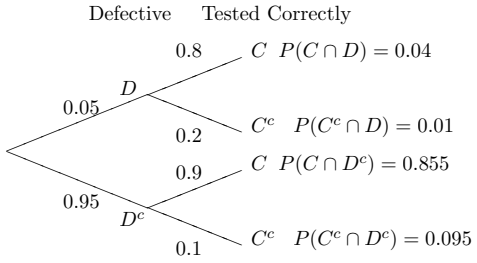
The probability that a micro-chip is defective, and has been declared as effective is

$$P(C^c \cap D) \;=\; P(C^c|D)P(D) \;=\; 0.2 \times 0.05 = 0.01\,.$$

The probability that a micro-chip is effective, and has been declared as defective is

$$P(C^c \cap D^c) \;=\; P(C^c|D^c)P(D^c) \;=\; 0.1 \times 0.95 \;=\; 0.095\,.$$

The tree diagram for these events and probabilities is:



(a) If a micro-chip is tested to be good, it could be defective but tested incorrectly, or it could be effective and tested correctly. Therefore, the probability that the micro-chip is tested good, but it is actually defective is

$$\frac{P(C^c \cap D)}{P(C^c \cap D) + P(C \cap D^c)} \;=\; \frac{0.01}{0.01 + 0.855} \;\approxeq\; 0.012$$

(b) Similarly, the probability that a micro-chip is tested to be defective, but it was good is

$$\frac{P(C^c \cap D^c)}{P(C \cap D) \;+\; P(C^c \cap D^c)} = \frac{0.095}{0.095 + 0.04} \approxeq 0.704$$

(c) The probability that both the micro-chips are effective, and have been tested and determined to be good, is

$$\left(\frac{P(C \cap D^c)}{P(C^c \cap D) + P(C \cap D^c)}\right)^2$$

and so the probability that at least one is defective is:

$$1 \;-\; \left(\frac{P(C \cap D^c)}{P(C^c \cap D) + P(C \cap D^c)}\right)^2 \;=\; 1 - \left(\frac{0.855}{0.01 + 0.855}\right)^2 \;\approxeq\; 0.023$$


```
>> 2^(1/2)
ans =    1.4142
```

The omission of parenthesis about 1/2 means something else and you get the following output:

```
>> 2^1/2
ans =     1
```

MATLAB first takes the 1st power of 2 and then divides it by 2 using its default precedence rules for binary operators in the absence of parenthesis. The order of operations or default precedence rule for arithmetic operations is 1. **b**rackets or parentheses; 2. **e**xponents (powers and roots); 3. **d**ivision and **m**ultiplication; 4. **a**ddition and **s**ubtraction. The mnemonic **bedmas** can be handy. When in doubt, use parenthesis to force the intended order of operations.

9. When you try to divide by 0, MATLAB returns `Inf` for infinity.

```
>> 10/0
ans =   Inf
```

10. We can clear the value we have assigned to a particular variable and reuse it. We demonstrate it by the following commands and their output:

```
>> X
X =    37
>> clear X
>> X
??? Undefined function or variable 'X'.
```

Entering `X` after `clear`ing it gives the above self-explanatory error message preceded by `???`.

11. We can suppress the output on the screen by ending the command with a semi-colon. Take a look at the simple command that sets X to sin(3.145678) with and without the '`;`' at the end:

```
>> X = sin(3.145678)
X =   -0.0041
>> X = sin(3.145678);
```

12. If you do not understand a MATLAB function or command then type `help` or `doc` followed by the function or command. For example:

```
>> help sin
 SIN    Sine of argument in radians.
    SIN(X) is the sine of the elements of X.
    See also asin, sind.
    Overloaded methods:
       darray/sin
    Reference page in Help browser
       doc sin
>> doc sin
```

It is a good idea to use the help files before you ask your tutor.

13. Set the variable `x` to equal 17.13 and evaluate $\cos(x)$, $\log(x)$, $\exp(x)$, $\arccos(x)$, $\text{abs}(x)$, $\text{sign}(x)$ using the MATLAB commands `cos`, `log`, `exp`, `acos`, `abs`, `sign`, respectively. Read the help files to understand what each function does.

14. When we work with real numbers (floating-point numbers) or really large numbers, we might want the output to be displayed in concise notation. This can be controlled in MATLAB using the `format` command with the `short` or `long` options with/without `e` for scientific notation. `format compact` is used for getting compacted output and `format` returns the default format. For example:

```
>> format compact
>> Y=15;
>> Y = Y + acos(-1)
Y =   18.1416
>> format short
>> Y
Y =   18.1416
>> format short e
>> Y
Y =  1.8142e+001
>> format long
>> Y
Y =  18.141592653589793
>> format long e
>> Y
Y =    1.814159265358979e+001
>> format
>> Y
Y =   18.1416
```

15. Finally, to quit from MATLAB just type `quit` or `exit` at the prompt.

```
>> quit
```

16. An **M-file** is a special text file with a `.m` extension that contains a set of code or instructions in MATLAB . In this course we will be using two types of M-files: **script** and **function** files. A script file is simply a list of commands that we want executed and saves us from retyping code modules we are pleased with. A function file allows us to write specific tasks as functions with input and output. These functions can be called from other script files, function files or command window. We will see such examples shortly.

By now, you are expected to be familiar with arithmetic operations, simple function evaluations, format control, starting and stopping a diary file and launching and quitting MATLAB .

1.6 Elementary Combinatorics

Combinatorics is the branch of mathematics that specialises in counting. We will give a more intuitive treatment with examples and then formally define the most primitive ideas called permutations and combinations. We also use several commonly encountered notations.

The most basic counting rule we use enables us to determine the number of distinct elements in a set that is constructed from taking two or more steps, where each step uses elements of another set. This is a lot easier than it sounds. Let's understand this through the analogy of performing several tasks.

To do this with formulae, partitioning according to the midterm test result and using the multiplication rule, we get:

$$\begin{aligned}\mathbf{P}(A) &= \mathbf{P}(A\cap B)+\mathbf{P}(A\cap B^c)\\ &= \mathbf{P}(A|B)\mathbf{P}(B)+\mathbf{P}(A|B^c)\mathbf{P}(B^c)\\ &= (0.8)(0.7)+(0.4)(0.3) \;=\; 0.68\end{aligned}$$

Answer (Ex. 2.9) — Let A be the event that bottles are produced by machine 1; and A^c is the event that bottles are produced by machine 2. R denotes the event that the bottles are rejected; and R^c denotes the event that the bottles are accepted. We know the following probabilities:

$$\mathbf{P}(A) \;=\; 0.75 \quad\text{and}\quad \mathbf{P}(A^c) \;=\; 0.25$$

$$\mathbf{P}(R|A) \;=\; \frac{1}{20} \quad\text{and}\quad \mathbf{P}(R^c|A) \;=\; \frac{19}{20}$$

$$\mathbf{P}(R|A^c) \;=\; \frac{1}{30} \quad\text{and}\quad \mathbf{P}(R^c|A^c) \;=\; \frac{29}{30}$$

We want $\mathbf{P}(A|R^c)$ which is give by

$$\mathbf{P}(A|R^c) \;=\; \frac{\mathbf{P}(R^c\cap A)}{\mathbf{P}(R^c)} \;=\; \frac{\mathbf{P}(R^c\cap A)}{\mathbf{P}(R^c\cap A)+\mathbf{P}(R^c\cap A^c)}$$

where,

$$\mathbf{P}(R^c\cap A) \;=\; \mathbf{P}(R^c|A)\mathbf{P}(A) \;=\; \frac{19}{20}\times 0.75$$

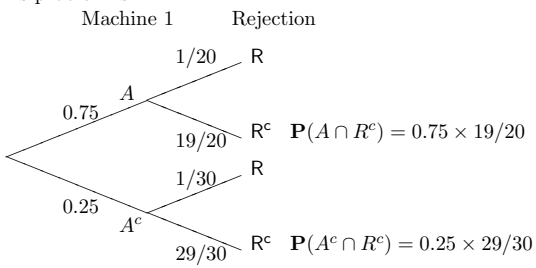
and,

$$\mathbf{P}(R^c\cap A^c) \;=\; \mathbf{P}(R^c|A^c)\mathbf{P}(A^c) \;=\; \frac{29}{30}\times 0.25$$

Therefore,

$$\mathbf{P}(A|R^c) \;=\; \frac{\frac{19}{20}\times 0.75}{\frac{19}{20}\times 0.75+\frac{29}{30}\times 0.25} \;\approxeq\; 0.747$$

The tree diagram for this problem is:



So the required probability is

$$\mathbf{P}(A|R^c) \;=\; \frac{\mathbf{P}(R^c\cap A)}{\mathbf{P}(R^c)} \;=\; \frac{\frac{19}{20}\times 0.75}{\frac{19}{20}\times 0.75+\frac{29}{30}\times 0.25} \;\approxeq\; 0.747$$

Answer (Ex. 2.7) — 1.The sample space is

$$\begin{array}{ll}\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6), & (2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ (3,1),(3,2),(3,3),(3,4),(3,5),(3,6), & (4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ (5,1),(5,2),(5,3),(5,4),(5,5),(5,6), & (6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}\end{array}$$

Note: Order matters here. For example, the outcome "16" refers to a "1" on the first die and a "6" on the second, whereas the outcome "61" refers to a "6" on the first die and a "1" on the second.

2.First tabulate all possible sums as follows:

+	1	2	3	4	5	6
1	2	3	4	**5**	**6**	7
2	3	4	**5**	**6**	7	8
3	4	**5**	**6**	7	8	9
4	**5**	**6**	7	8	9	10
5	**6**	7	8	9	10	11
6	7	8	9	10	11	12

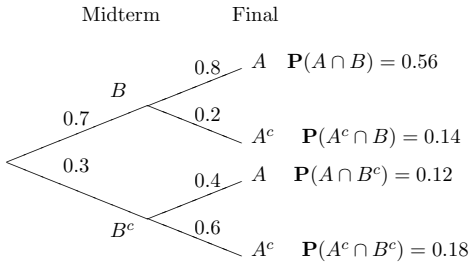
Let A be the event *the sum is 5* and B be the event *the sum is 6*, then A and B are mutually exclusive events with probabilities

$$\mathbf{P}(A) = \frac{4}{36} \quad \text{and} \quad \mathbf{P}(B) = \frac{5}{36}\,.$$

Therefore,

$$\mathbf{P}(4 < \text{sum } < 7) \;=\; \mathbf{P}(A \cup B) \;=\; \mathbf{P}(A) + \mathbf{P}(B) \;=\; \frac{4}{36} + \frac{5}{36} \;=\; \frac{1}{4}$$

Answer (Ex. 2.8) — First draw a tree with the first split based on the outcome of the midterm test and the second on the outcome of the final exam. Note that the probabilities involved in this second branch are *conditional* probabilities that depend on the outcome of the midterm test. Let A be the event that the student passes the final exam and let B be the event that the student passes the midterm test.



Then the probability of passing the final exam is:

$$\mathbf{P}(A) \;=\; 0.56 \,+\, 0.12 \;=\; 0.68\,.$$

The multiplication principle: If a task can be performed in n_1 ways, a second task in n_2 ways, a third task in n_3 ways, etc., then the total number of distinct ways of performing all tasks together is

$$n_1 \times n_2 \times n_3 \times \ldots$$

Example 6 Suppose that a Personal Identification Number (PIN) is a six-symbol code word in which the first four entries are letters (lowercase) and the last two entries are digits. How many PINS are there? There are six selections to be made:

First letter: 26 possibilities	Fourth letter: 26 possibilities
Second letter: 26 possibilities	First digit: 10 possibilities
Third letter: 26 possibilities	Second digit: 10 possibilities

So in total, the total number of possible PINS is:

$$26 \times 26 \times 26 \times 26 \times 10 \times 10 \;=\; 26^4 \times 10^2 \;=\; 45,697,600\,.$$

Example 7 Suppose we now put restrictions on the letters and digits we use. For example, we might say that the first digit cannot be zero, and letters cannot be repeated. This time the the total number of possible PINS is:

$$26 \times 25 \times 24 \times 23 \times 9 \times 10 \;=\; 32,292,000\,.$$

When does order matter? In English we use the word "combination" loosely. If I say

"I have 17 probability texts on my bottom shelf"

then I don't care (usually) about what order they are in, but in the statement

"The combination of my PIN is math99"

I do care about order. A different order gives a different PIN.

So in mathematics, we use more precise language:

- A selection of objects in which the order is important is called a **permutation**.
- A selection of objects in which the order is *not* important is called a **combination**.

Permutations: There are basically two types of permutations:

1. Repetition is allowed, as in choosing the letters (unrestricted choice) in the PIN of Example 6. More generally, when you have n objects to choose from, you have n choices each time, so when choosing r of them, the number of permutations are n^r.

2. No repetition is allowed, as in the restricted PIN Example 7. Here you have to reduce the number of choices. If we had a 26 letter PIN then the total permutations would be

$$26 \times 25 \times 24 \times 23 \times \ldots \times 3 \times 2 \times 1 = 26!$$

but since we want four letters only here, we have

$$\frac{26!}{22!} = 26 \times 25 \times 24 \times 23$$

choices.

The number of distinct **permutations** of n objects taking r at a time is given by

$$^nP_r = \frac{n!}{(n-r)!}$$

Combinations: There are also two types of combinations:

1. Repetition is allowed such as the coins in your pocket, say, $(10c, 50c, 50c, \$1, \$2, \$2)$.
2. No repetition is allowed as in the lottery numbers (2, 9, 11, 26, 29, 31). The numbers are drawn one at a time, and if you have the lucky numbers (no matter what order) you win!

The number of distinct **combinations** of n objects taking r at a time is given by

$$^nC_r = \binom{n}{r} = \frac{n!}{(n-r)!\,r!}$$

Example 8 Let us imagine being in the lower Manhattan in New York city with its perpendicular grid of streets and avenues. If you start at a given intersection and are asked to only proceed in a north-easterly direction then how may ways are there to reach another intersection by walking exactly two blocks or exactly three blocks?

Solution:
Let us answer this question of combinations by drawing Fig. 1.6. Let us denote the number of easterly turns you take by r and the total number of blocks you are allowed to walk either easterly or northerly by n. From Fig. 1.6(a) it is clear that the number of ways to reach each of the three intersections labeled by r is given by $\binom{n}{r}$, with $n = 2$ and $r \in \{0, 1, 2\}$. Similarly, from Fig. 1.6(b) it is clear that the number of ways to reach each of the four intersections labeled by r is given by $\binom{n}{r}$, with $n = 3$ and $r \in \{0, 1, 2, 3\}$.

Exercise 1.5 (Choosing Volunteers) Suppose we need three students to be the class representatives in this course. Assume that everyone wants to be selected to keep it simple. In how many ways can we choose these three people from the class of 50 students?

Now, we give more formal definitions and notations that will help us make precise arguments faster when we study sampling schemes in Inference Theory.

2.Since there are eleven letters in WAIMAKARIRI the probabilities are:

$$\mathbf{P}(\{\mathsf{W}\}) = \tfrac{1}{11}, \mathbf{P}(\{\mathsf{A}\}) = \tfrac{3}{11}, \mathbf{P}(\{\mathsf{I}\}) = \tfrac{3}{11}, \mathbf{P}(\{\mathsf{M}\}) = \tfrac{1}{11}, \mathbf{P}(\{\mathsf{K}\}) = \tfrac{1}{11}, \mathbf{P}(\{\mathsf{R}\}) = \tfrac{2}{11}\,.$$

3.By the complementation rule, the probability of not choosing the letter R is:

$$1 - \mathbf{P}(\text{choosing the letter R}) = 1 - \frac{2}{11} = \frac{9}{11}\,.$$

Answer (Ex. 2.5) — 1.First, the sample space is: $\Omega = \{\mathsf{B}, \mathsf{I}, \mathsf{N}, \mathsf{G}, \mathsf{O}\}$.

2.The probabilities of simple events are:

$$\mathbf{P}(\mathsf{B}) = \mathbf{P}(\mathsf{I}) = \mathbf{P}(\mathsf{N}) = \mathbf{P}(\mathsf{G}) = \mathbf{P}(\mathsf{O}) = \frac{15}{75} = \frac{1}{5}\ .$$

3.Using the addition rule for mutually exclusive events,

$$\begin{aligned}
\mathbf{P}(\Omega) &= \mathbf{P}(\{\mathsf{B}, \mathsf{I}, \mathsf{N}, \mathsf{G}, \mathsf{O}\}) \\
&= \mathbf{P}(\{\mathsf{B}\} \cup \{\mathsf{I}\} \cup \{\mathsf{N}\} \cup \{\mathsf{G}\} \cup \{\mathsf{O}\}) \\
&= \mathbf{P}(\mathsf{B}) + \mathbf{P}(\mathsf{I}) + \mathbf{P}(\mathsf{N}) + \mathbf{P}(\mathsf{G}) + \mathbf{P}(\mathsf{O}) \quad \text{simplifying notation} \\
&= \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} \\
&= 1
\end{aligned}$$

4.Since the events $\{\mathsf{B}\}$ and $\{\mathsf{I}\}$ are disjoint,

$$\mathbf{P}(\{\mathsf{B}\} \cup \{\mathsf{I}\}) = \mathbf{P}(\mathsf{B}) + \mathbf{P}(\mathsf{I}) = \frac{1}{5} + \frac{1}{5} = \frac{2}{5}\,.$$

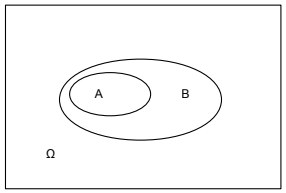
5.Using the addition rule for two arbitrary events we get,

$$\begin{aligned}
\mathbf{P}(C \cup D) &= \mathbf{P}(C) + \mathbf{P}(D) - \mathbf{P}(C \cap D) \\
&= \mathbf{P}(\{\mathsf{B}, \mathsf{I}, \mathsf{G}\}) + \mathbf{P}(\{\mathsf{G}, \mathsf{I}, \mathsf{N}\}) - \mathbf{P}(\{\mathsf{G}, \mathsf{I}\}) \\
&= \frac{3}{5} + \frac{3}{5} - \frac{2}{5} \\
&= \frac{4}{5}\ .
\end{aligned}$$

Answer (Ex. 2.6) — We can assume that the first shot is independent of the second shot so we can multiply the probabilities here.
For case A, there is only one shot so the probability of hitting at least once is $\frac{1}{2}$.
For case B, the probability of missing both shots is $\frac{2}{3}\frac{2}{3} = \frac{4}{9}$, so the probability hitting some target at least once is

$$1 - \mathbf{P}(\text{missing the target both times}) = 1 - \frac{4}{9} = \frac{5}{9}$$

Therefore, case B has the greater probability of hitting the target at least once.



Answer (Exercise 1.5) — We start by assuming that order does matter, that is, we have a permutation, so that the number of ways we can select the three class representatives is

$$^{50}P_3 = \frac{50!}{(50-3)!} = \frac{50!}{47!}$$

But, because order doesn't matter, all we have to do is to adjust our permutation formula by a factor representing the number of ways the objects could be in order. Here, three students can be placed in order 3! ways, so the required number of ways of choosing the class representatives is:

$$\frac{50!}{47!\,3!} = \frac{50\,.\,49\,.\,48}{3\,.\,2\,.\,1} = 19,600$$

Answer (Exercise 2.1) — This is an optional exercise. You will understand this as you progress through your mathematics programme. The explanation in the said item was (or will be explained in person again) in the lectures. This exercise was created to answer natural questions that were asked by students who wanted to know.

Answer (Ex. 2.2) — (a)$\mathbf{P}(\{\mathsf{Z}\}) = 0.1\% = \frac{0.1}{100} = 0.001$

(b)$\mathbf{P}$('picking any letter') $= \mathbf{P}(\Omega) = 1$

(c)$\mathbf{P}(\{\mathsf{E},\mathsf{Z}\}) = \mathbf{P}(\{\mathsf{E}\} \cup \{\mathsf{Z}\}) = \mathbf{P}(\{\mathsf{E}\}) + \mathbf{P}(\{\mathsf{Z}\}) = 0.13 + 0.001 = 0.131$, by Axiom (3)

(d)$\mathbf{P}$('picking a vowel') $= \mathbf{P}(\{\mathsf{A},\mathsf{E},\mathsf{I},\mathsf{O},\mathsf{U}\}) = (7.3\% + 13.0\% + 7.4\% + 7.4\% + 2.7\%) = 37.8\%$, by the addition rule for mutually exclusive events, rule (2).

(e)$\mathbf{P}$('picking any letter in the word WAZZZUP') $= \mathbf{P}(\{\mathsf{W},\mathsf{A},\mathsf{Z},\mathsf{U},\mathsf{P}\}) = 14.4\%$, by the addition rule for mutually exclusive events, rule (2).

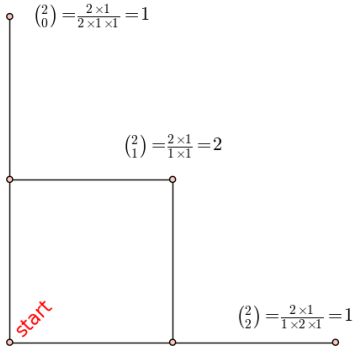
(f)$\mathbf{P}$('picking any letter in the word WAZZZUP or a vowel') = $\mathbf{P}(\{\mathsf{W},\mathsf{A},\mathsf{Z},\mathsf{U},\mathsf{P}\}) + \mathbf{P}(\{\mathsf{A},\mathsf{E},\mathsf{I},\mathsf{O},\mathsf{U}\}) - \mathbf{P}(\{\mathsf{A},\mathsf{U}\}) = 14.4\% + 37.8\% - 10\% = 42.2\%$, by the addition rule for two arbitrary events, rule (3).

Answer (Ex. 2.3) — 1.$\{\mathsf{BB},\mathsf{BW},\mathsf{WB},\mathsf{WW}\}$

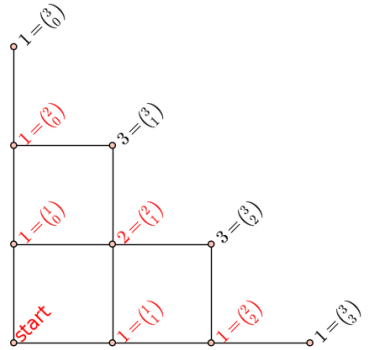
2. $\{\mathsf{RRRR},\mathsf{RRRL},\mathsf{RRLR},\mathsf{RLRR},\mathsf{LRRR},\mathsf{RLRL},\mathsf{RRLL},\mathsf{LLRR},$
$\mathsf{LRLR},\mathsf{LRRL},\mathsf{RLLR},\mathsf{LLLL},\mathsf{LLLR},\mathsf{LLRL},\mathsf{LRLL},\mathsf{RLLL}\}$

3.$\{6,16,26,36,46,56,116,126,136,146,156,216,226,236,246,256,\ldots\}$

Answer (Ex. 2.4) — 1.The sample space $\Omega = \{\mathsf{W},\mathsf{A},\mathsf{I},\mathsf{M},\mathsf{K},\mathsf{R}\}$.



(a) Walking two blocks north-easterly.



(b) Walking three blocks north-easterly.

Definition 1 (Permutations and Factorials) A **permutation** of n objects is an arrangement of n distinct objects in a row. For example, there are 2 permutations of the two objects $\{1,2\}$:

$$12, \qquad 21\ ,$$

and 6 permutations of the three objects $\{a,b,c\}$:

$$abc, \quad acb, \quad bac, \quad bca, \quad cab, \quad cba\ .$$

Let the number of ways to choose k objects out of n and to arrange them in a row be denoted by $p_{n,k}$. For example, we can choose two ($k=2$) objects out of three ($n=3$) objects, $\{a,b,c\}$, and arrange them in a row in six ways ($p_{3,2}$):

$$ab, \quad ac, \quad ba, \quad bc, \quad ca, \quad cb\ .$$

Given n objects, there are n ways to choose the left-most object, and once this choice has been made there are $n-1$ ways to select a different object to place next to the left-most one. Thus, there are $n(n-1)$ possible choices for the first two positions. Similarly, when $n>2$, there are $n-2$ choices for the third object that is distinct from the first two. Thus, there are $n(n-1)(n-2)$ possible ways to choose three distinct objects from a set of n objects and arrange them in a row. In general,

$$p_{n,k} = n(n-1)(n-2)\ldots(n-k+1)$$

and the total number of permutations called 'n **factorial**' and denoted by $n!$ is

$$n! := p_{n,n} = n(n-1)(n-2)\ldots(n-n+1) = n(n-1)(n-2)\ldots(3)\ (2)\ (1) =: \prod_{i=1}^{n} i\ .$$

Some factorials to bear in mind

$$0! := 1 \quad 1! = 1, \quad 2! = 2, \quad 3! = 6, \quad 4! = 24, \quad 5! = 120 \quad 10! = 3,628,800\ .$$

When n is large we can get a good idea of $n!$ without laboriously carrying out the $n-1$ multiplications via Stirling's approximation (*Methodus Differentialis (1730), p. 137*) :

$$n! \approxeq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n\ .$$

Definition 2 (Combinations) The combinations of n objects taken k at a time are the possible choices of k different elements from a collection of n objects, disregarding order. They are called the k-combinations of the collection. The combinations of the three objects $\{a, b, c\}$ taken two at a time, called the 2-combinations of $\{a, b, c\}$, are

$$ab, \quad ac, \quad bc \ ,$$

and the combinations of the five objects $\{1, 2, 3, 4, 5\}$ taken three at a time, called the 3-combinations of $\{1, 2, 3, 4, 5\}$ are

$$123, \quad 124, \quad 125, \quad 134, \quad 135, \quad 145, \quad 234, \quad 235, \quad 245, \quad 345 \ .$$

The total number of k-combination of n objects, called a **binomial coefficient**, denoted $\binom{n}{k}$ and read "n choose k," can be obtained from $p_{n,k} = n(n-1)(n-2)\dots(n-k+1)$ and $k! := p_{k,k}$. Recall that $p_{n,k}$ is the number of ways to choose the first k objects from the set of n objects and arrange them in a row with regard to order. Since we want to disregard order and each k-combination appears exactly $p_{k,k}$ or $k!$ times among the $p_{n,k}$ many permutations, we perform a division:

$$\binom{n}{k} := \frac{p_{n,k}}{p_{k,k}} = \frac{n(n-1)(n-2)\dots(n-k+1)}{k(k-1)(k-2)\dots 2\ 1} \ .$$

Binomial coefficients are often called "Pascal's Triangle" and attributed to Blaise Pascal's *Traité du Triangle Arithmétique* from 1653, but they have many "fathers". There are earlier treatises of the binomial coefficients including Szu-yüan Yü-chien ("The Precious Mirror of the Four Elements") by the Chinese mathematician Chu Shih-Chieh in 1303, and in an ancient Hindu classic, *Piṅgala's Chanda**d**hśāstra*, due to Halāyudha (10-th century AD).

1.7 Array, Sequence, Limit, ...

In this section we will study a basic data structure in MATLAB called an **array** of numbers. Arrays are finite sequences and they can be processed easily in MATLAB . The notion of infinite sequences lead to **limits**, one of the most fundamental concepts in mathematics.

For any natural number n, we write

$$\boxed{\langle x_{1:n} \rangle := x_1, x_2, \dots, x_{n-1}, x_n}$$

to represent the **finite sequence** of real numbers $x_1, x_2, \dots, x_{n-1}, x_n$. For two integers m and n such that $m \le n$, we write

$$\boxed{\langle x_{m:n} \rangle := x_m, x_{m+1}, \dots, x_{n-1}, x_n}$$

to represent the **finite sequence** of real numbers $x_m, x_{m+1}, \dots, x_{n-1}, x_n$. In mathematical analysis, finite sequences and their countably infinite counterparts play a fundamental role in limiting processes. Given an integer m, we denote an **infinite sequence** or simply a sequence as:

$$\boxed{\langle x_{m:\infty} \rangle := x_m, x_{m+1}, x_{m+2}, x_{m+3}, \dots \ .}$$

Given index set $\mathcal{I}$ which may be finite or infinite in size, a sequence can either be seen as a set of ordered pairs:

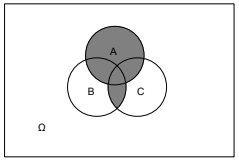
$$\{(i, x_i) : i \in \mathcal{I}\} \ ,$$

Answers to Selected Exercises

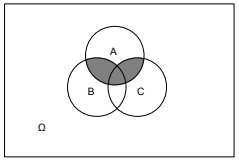
Answer (Ex. 1.1) — By operating with Ω, T, L and S we can obtain the answers as follows:

(a)$T \cap L = \{L_3\}$
(b)$T \cap S = \emptyset$
(c)$T \cup L = \{T_1, T_2, T_3, L_3, L_1, L_2\}$
(d)$T \cup L \cup S = \Omega$
(e)$S^c = \{T_1, T_2, T_3, L_3, L_1, L_2\}$
(f)$S \cap L = \emptyset$
(g)$S^c \cap L = \{L_1, L_2, L_3\} = L$
(h)$T^c = \{L_1, L_2, S_1, S_2, S_2, \dots S_{50}\}$
(i)$T^c \cap L = \{L_1, L_2\}$
(j)$T^c \cap T = \emptyset$

Answer (Ex. 1.3) — We can check $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ from the following sketch:



Answer (Ex. 1.3) — We can check $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ from the following sketch: We can check $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ from the following sketch:



Answer (Ex. 1.4) — To illustrate the idea that $A \subseteq B$ if and only if $A \cup B = B$, we need to illustrate two implications:

1.if $A \subseteq B$ then $A \cup B = B$ and

2.if $A \cup B = B$ then $A \subseteq B$.

The folowing Venn diagram illustrates the two implications clearly.

Symbol	Meaning
$\mathbb{1}_A(x)$	Indicator or set membership function that returns 1 if $x \in A$ and 0 otherwise
$\mathbb{R}^d := (-\infty, \infty)^d$	d-dimensional Real Space
$\mathrm{R}\vec{\mathrm{V}}$	random vector
$F_{X,Y}(x,y)$	Joint distribution function (JDF) of the $\mathrm{R}\vec{\mathrm{V}}$ (X,Y)
$F_{X,Y}(x,y)$	Joint cumulative distribution function (JCDF) of the $\mathrm{R}\vec{\mathrm{V}}$ (X,Y) — same as JDF
$f_{X,Y}(x,y)$	Joint probability mass function (JPMF) of the discrete $\mathrm{R}\vec{\mathrm{V}}$ (X,Y)
$\mathcal{S}_{X,Y}$ $= \{(x_i,y_j) : f_{X,Y}(x_i,x_j) > 0\}$	The support set of the discrete $\mathrm{R}\vec{\mathrm{V}}$ (X,Y)
$f_{X,Y}(x,y)$	Joint probability density function (JPDF) of the continuous $\mathrm{R}\vec{\mathrm{V}}$ (X,Y)
$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$	Marginal probability density/mass function (MPDF/MPMF) of X
$f_Y(y)$ $= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$	Marginal probability density/mass function (MPDF/MPMF) of Y
$E(g(X,Y))$ $= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)dxdy$	Expectation of a function $g(x,y)$ for continuous $\mathrm{R}\vec{\mathrm{V}}$
$E(g(X,Y))$ $= \sum_{(x,y)\in\mathcal{S}_{X,Y}} g(x,y)f_{X,Y}(x,y)$	Expectation of a function $g(x,y)$ for discrete $\mathrm{R}\vec{\mathrm{V}}$
$E(X^rY^s)$	Joint moment
$\mathbf{Cov}(X,Y) = E(XY) - E(X)E(Y)$	Covariance of X and Y, provided $E(X^2) < \infty$ and $E(Y^2) > \infty$
$F_{X,Y}(x,y) = F_X(x)F_Y(y)$, for every (x,y)	if and only if X and Y are said to be independent
$f_{X,Y}(x,y) = f_X(x)f_Y(y)$, for every (x,y)	if and only if X and Y are said to be independent
$F_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$	Joint (cumulative) distribution function (JDF/JCDF) of the discrete or continuous $\mathrm{R}\vec{\mathrm{V}}$ $(X_1,X_2,\ldots,X_n)$
$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$	Joint probability mass/density function (JPMF/JPDF) of the discrete/continuous $\mathrm{R}\vec{\mathrm{V}}$ $(X_1,X_2,\ldots,X_n)$
$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$ $= \prod_{i=1}^n f_{X_i}(x_i)$, for every $(x_1,x_2,\ldots,x_n)$	if and only if $X_1,X_2,\ldots,X_n$ are (mutually/jointly) independent

Table 6.4: Symbol Table: Probability and Statistics

or as a function that maps the index set to the set of real numbers:

$$x(i) = x_i : \mathcal{I} \to \{x_i : i \in \mathcal{I}\} \ ,$$

The finite sequence $\langle x_{m:n}\rangle$ has $\mathcal{I} = \{m, m+1, m+2, m+3, \ldots, n\}$ as its index set while an infinite sequence $\langle x_{m:\infty}\rangle$ has $\mathcal{I} = \{m, m+1, m+2, m+3, \ldots\}$ as its index set. A **sub-sequence** $\langle x_{j:k}\rangle$ of a finite sequence $\langle x_{m:n}\rangle$ or an infinite sequence $\langle x_{m:\infty}\rangle$ is:

$$\langle x_{j:k}\rangle = x_j, x_{j+1}, \ldots, x_{k-1}, x_k \quad \text{where,} \quad m \le j \le k \le n < \infty \ .$$

A rectangular arrangement of $m \cdot n$ real numbers in m rows and n columns is called an $m \times n$ **matrix**. The '$m \times n$' represents the **size** of the matrix. We use bold upper-case letters to denote matrices, for e.g:

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,n-1} & x_{1,n} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,n-1} & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m-1,1} & x_{m-1,2} & \ldots & x_{m-1,n-1} & x_{m-1,n} \\ x_{m,1} & x_{m,2} & \ldots & x_{m,n-1} & x_{m,n} \end{bmatrix}$$

Matrices with only one row or only one column are called **vectors**. An $1 \times n$ matrix is called a **row vector** since there is only one row and an $m \times 1$ matrix is called a **column vector** since there is only one column. We use bold-face lowercase letters to denote row and column vectors.

$$\text{A row vector } \boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix} = (x_1, x_2, \ldots, x_n)$$

$$\text{and a column vector } \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{m-1} \\ y_m \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & \ldots & y_m \end{bmatrix}' = (y_1, y_2, \ldots, y_m)' \ .$$

The superscripting by $'$ is the transpose operation and simply means that the rows and columns are exchanged. Thus the transpose of the matrix $\boldsymbol{X}$ is:

$$\boldsymbol{X}' = \begin{bmatrix} x_{1,1} & x_{2,1} & \ldots & x_{m-1,1} & x_{m,1} \\ x_{1,2} & x_{2,2} & \ldots & x_{m-1,2} & x_{m,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{1,n-1} & x_{2,n-1} & \ldots & x_{m-1,n-1} & x_{m,n-1} \\ x_{1,n} & x_{2,n} & \ldots & x_{m-1,n} & x_{m,n} \end{bmatrix}$$

In linear algebra and calculus, it is natural to think of vectors and matrices as points (ordered m-tuples) and ordered collection of points in Cartesian co-ordinates. We assume that the reader has heard of operations with matrices and vectors such as matrix multiplication, determinants, transposes, etc. Such concepts will be introduced as they are needed in the sequel.

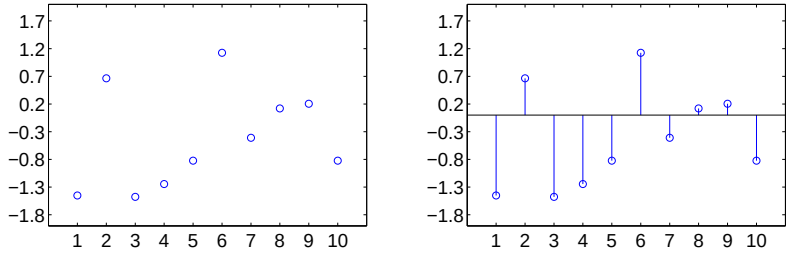
Finite sequences, vectors and matrices can be represented in a computer by an elementary data structure called an **array**.

Labwork 9 (Sequences as arrays) Let us learn to represent, visualise and operate finite sequences as MATLAB arrays. Try out the commands and read the comments for clarification.

```
>> a = [17]                        % Declare the sequence of one element 17 in array a
a =    17
>> % Declare the sequence of 10 numbers in array b
>> b=[-1.4508 0.6636 -1.4768 -1.2455 -0.8235 1.1254 -0.4093 0.1199 0.2043 -0.8236]
b =
   -1.4508    0.6636   -1.4768   -1.2455   -0.8235    1.1254   -0.4093    0.1199    0.2043   -0.8236
>> c = [1 2 3]                  % Declare the sequence of 3 consecutive numbers 1,2,3
c =     1     2     3
>> % linspace(x1, x2, n) generates n points linearly spaced between x1 and x2
>> r = linspace(1, 3, 3)               % Declare sequence r = c using linspace
r =     1     2     3
>> s1 = 1:10 % declare an array s1 starting at 1, ending by 10, in increments of 1
s =     1     2     3     4     5     6     7     8     9    10
>> s2 = 1:2:10  % declare an array s2 starting at 1, ending by 10, in increments of 2
s =     1     3     5     7     9
>> s2(3) % obtain the third element of the finite sequence s2
ans =     5
>> s2(2:4) % obtain the subsequence from second to fourth elements of the finite sequence s2
ans =     3     5     7
```

We may visualise (as per Figure 1.6) the finite sequences $\langle b_{1:n} \rangle$ stored in the array `b` as the set of ordered pairs $\{(1, b_1), (2, b_2), \ldots, (10, b_n)\}$ representing the function $b(i) = b_i : \{1, 2, \ldots, n\} \to \{b_1, b_2, \ldots, b_n\}$ via **point plot** and **stem plot** using `Matlab`'s `plot` and `stem` commands, respectively.

Figure 1.6: Point plot and stem plot of the finite sequence $\langle b_{1:10} \rangle$ declared as an array.




```
>> display(b) % display the array b in memory
b =
   -1.4508    0.6636   -1.4768   -1.2455   -0.8235    1.1254   -0.4093    0.1199    0.2043   -0.8236
>> plot(b,'o') % point plot of ordered pairs (1,b(1)), (2,b(2)), ..., (10,b(10))
>> stem(b) % stem plot of ordered pairs (1,b(1)), (2,b(2)), ..., (10,b(10))
>> plot(b,'-o') % point plot of ordered pairs (1,b(1)), (2,b(2)), ..., (10,b(10)) connected by lines
```

Labwork 10 (Vectors and matrices as arrays) Let us learn to represent, visualise and operate vectors as MATLAB arrays. Syntactically, a vector is stored in an array exactly in the same way we stored a finite sequence. However, mathematically, we think of a vector as an ordered m-tuple that can be visualised as a point in Cartesian co-ordinates. Try out the commands and read the comments for clarification.

```
>> a = [1 2]        % an 1 X 2 row vector
>> z = [1 2 3]        % Declare an 1 X 3 row vector z with three numbers
```


Model	PDF or PMF	Mean	Variance
Bernoulli(θ)	$\theta^x(1-\theta)^{1-x}\mathbb{1}_{\{0,1\}}(x)$	θ	$\theta(1-\theta)$
Binomial(n, θ)	$\binom{n}{\theta}\theta^x(1-\theta)^{n-x}\mathbb{1}_{\{0,1,\ldots,n\}}(x)$	$n\theta$	$n\theta(1-\theta)$
Geometric(θ)	$\theta(1-\theta)^x\mathbb{1}_{\mathbb{Z}_+}(x)$	$\frac{1}{\theta}-1$	$\frac{1-\theta}{\theta^2}$
Poisson(λ)	$\frac{\lambda^x e^{-\lambda}}{x!}\mathbb{1}_{\mathbb{Z}_+}(x)$	λ	λ
Uniform(θ_1, θ_2)	$\mathbb{1}_{[\theta_1,\theta_2]}(x)/(\theta_2-\theta_1)$	$\frac{\theta_1+\theta_2}{2}$	$\frac{(\theta_2-\theta_1)^2}{12}$
Exponential(λ)	$\lambda e^{-\lambda x}$	λ^{-1}	λ^{-2}
Normal(μ, σ^2)	$\frac{1}{\sigma\sqrt{2\pi}}e^{(x-\mu)^2/(2\sigma^2)}$	μ	σ^2

Table 6.2: Random Variables with PDF and PMF (using indicator function), Mean and Variance

Symbol	Meaning
$A = \{\star, \circ, \bullet\}$	A is a set containing the elements $\star$, $\circ$ and $\bullet$
$\circ \in A$	$\circ$ belongs to A or $\circ$ is an element of A
$A \ni \circ$	$\circ$ belongs to A or $\circ$ is an element of A
$\odot \notin A$	$\odot$ does not belong to A
$\#A$	Size of the set A, for e.g. $\#\{\star, \circ, \bullet, \odot\} = 4$
$\mathbb{N}$	The set of natural numbers $\{1, 2, 3, \ldots\}$
$\mathbb{Z}$	The set of integers $\{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$
$\mathbb{Z}_+$	The set of non-negative integers $\{0, 1, 2, 3, \ldots\}$
$\emptyset$	Empty set or the collection of nothing or $\{\}$
$A \subset B$	A is a subset of B or A is contained by B, e.g. $A = \{\circ\}, B = \{\bullet\}$
$A \supset B$	A is a superset of B or A contains B e.g. $A = \{\circ, \star, \bullet\}, B = \{\circ, \bullet\}$
$A = B$	A equals B, i.e. $A \subset B$ and $B \subset A$
$Q \implies R$	Statement Q implies statement R or If Q then R
$Q \iff R$	$Q \implies R$ and $R \implies Q$
$\{x : x \text{ satisfies property } R\}$	The set of all x such that x satisfies property R
$A \cup B$	A union B, i.e. $\{x : x \in A \text{ or } x \in B\}$
$A \cap B$	A intersection B, i.e. $\{x : x \in A \text{ and } x \in B\}$
$A \setminus B$	A minus B, i.e. $\{x : x \in A \text{ and } x \notin B\}$
$A := B$	A is equal to B by definition
$A =: B$	B is equal to A by definition
A^c	A complement, i.e. $\{x : x \in U, \text{ the universal set, but } x \notin A\}$
$A_1 \times A_2 \times \cdots \times A_m$	The m-product set $\{(a_1, a_2, \ldots, a_m) : a_1 \in A_1, a_2 \in A_2, \ldots, a_m \in A_m\}$
A^m	The m-product set $\{(a_1, a_2, \ldots, a_m) : a_1 \in A, a_2 \in A, \ldots, a_m \in A\}$
$f := f(x) = y : \mathbb{X} \to \mathbb{Y}$	A function f from domain $\mathbb{X}$ to range $\mathbb{Y}$
$f^{[-1]}(y)$	Inverse image of y
$f^{[-1]} := f^{[-1]}(y \in \mathbb{Y}) = X \subset \mathbb{X}$	Inverse of f
$a < b$ or $a \leq b$	a is less than b or a is less than or equal to b
$a > b$ or $a \geq b$	a is greater than b or a is greater than or equal to b
$\mathbb{Q}$	Rational numbers
(x, y)	the open interval (x, y), i.e. $\{r : x < r < y\}$
$[x, y]$	the closed interval (x, y), i.e. $\{r : x \leq r \leq y\}$
$(x, y]$	the half-open interval $(x, y]$, i.e. $\{r : x < r \leq y\}$
$[x, y)$	the half-open interval $[x, y)$, i.e. $\{r : x \leq r < y\}$
$\mathbb{R} := (-\infty, \infty)$	Real numbers, i.e. $\{r : -\infty < r < \infty\}$
$\mathbb{R}_+ := [0, \infty)$	Real numbers, i.e. $\{r : 0 \leq r < \infty\}$
$\mathbb{R}_{>0} := (0, \infty)$	Real numbers, i.e. $\{r : 0 < r < \infty\}$

Table 6.3: Symbol Table: Sets and Numbers

CONTINUOUS RANDOM VARIABLES: NOTATION

$f(x)$: Probability density function (PDF)

- $f(x) \geq 0$
- Areas underneath $f(x)$ measure probabilities.

$F(x)$: Distribution function (DF)

- $0 \leq F(x) \leq 1$
- $F(x) = P(X \leq x)$ is a probability
- $F'(x) = f(x)$ for every x where $f(x)$ is continuous
- $F(x) = \int_{-\infty}^{x} f(v)dv$
- $P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(v)dv$

Expectation of a function $g(X)$ of a random variable X is defined as:

$$E(g(X)) = \begin{cases} \sum_x g(x)f(x) & \text{if } X \text{ is a discrete RV} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{if } X \text{ is a continuous RV} \end{cases}$$

Some Common Expectations

$g(x)$	definition	also known as
x	$E(X)$	Expectation, Population Mean or First Moment of X
$(x - E(X))^2$	$V(X) := E((X - E(X))^2)$ $= E(X^2) - (E(X))^2$	Variance or Population Variance of X
e^{itx}	$\phi_X(t) := E\left(e^{itX}\right)$	Characteristic Function (CF) of X
x^k	$E(X^k) = \frac{1}{i^k}\left[\frac{d^k \phi_X(t)}{dt^k}\right]_{t=0}$	k-th Moment of X

Symbol	Meaning
$\mathbb{1}_A(x)$	Indicator or set membership function that returns 1 if $x \in A$ and 0 otherwise
$\mathbb{R}^d := (-\infty, \infty)^d$	d-dimensional Real Space

Table 6.1: Symbol Table: Probability and Statistics


```
z =     1     2     3
>> % linspace(x1, x2, n) generates n points linearly spaced between x1 and x2
>> r = linspace(1, 3, 3)               % Declare an 1 X 3 row vector r = z using linspace
r =     1     2     3
>> c = [1; 2; 3]          % Declare a 3 X 1 column vector c with three numbers.  Semicolons delineate columns
c =
     1
     2
     3
>> rT = r'        % The column vector (1,2,3)' by taking the transpose of r via r'
rT =
     1
     2
     3
>> y = [1 1 1]                         % y is a sequence or row vector of 3 1's
y =     1     1     1
>> ones(1,10)        % ones(m,n) is an m X n matrix of ones.  Useful when m or n is large.
ans =     1     1     1     1     1     1     1     1     1     1
```

We can use two dimensional arrays to represent matrices. Some useful built-in commands to generate standard matrices are:

```
>> Z=zeros(2,10) % the 2 X 10 matrix of zeros
Z =
     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0
>> O=ones(4,5) % the 4 X 5 matrix of ones
O =
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     1
>> E=eye(4) % the 4 X 4 identity matrix
E =
     1     0     0     0
     0     1     0     0
     0     0     1     0
     0     0     0     1
```

We can also perform operations with arrays representing vectors, finite sequences, or matrices.

```
>> y % the array y is
y =     1     1     1
>> z % the array z is
z =     1     2     3
>> x = y + z                           % x is the sum of vectors y and z (with same size 1 X 3)
x =     2     3     4
>> y = y * 2                            % y is updated to 2 * y (each term of y is multiplied by 2)
y =     2     2     2
>> p = z .* y                           % p is the vector obtained by term-by-term product of z and y
p =     2     4     6
>> d = z ./ y                            % d is the vector obtained by term-by-term division of z and y
d =    0.5000    1.0000    1.5000
>> t=linspace(-10,10,4)                % t has 4 numbers equally-spaced between -10 and 10
t =  -10.0000   -3.3333    3.3333   10.0000
>> s = sin(t)                          % s is a vector obtained from the term-wise sin of the vector t
s =    0.5440    0.1906   -0.1906   -0.5440
>> sSq = sin(t) .^ 2         % sSq is an array obtained from term-wise squaring ( .^ 2) of the sin(t) array
sSq =    0.2960    0.0363    0.0363    0.2960
>> cSq = cos(t) .^ 2 % cSq is an array obtained from term-wise squaring ( .^ 2) of the cos(t) array
cSq =    0.7040    0.9637    0.9637    0.7040
```

```
>> sSq + cSq % we can add the two arrays sSq and cSq to get the array of 1's
ans =     1     1     1     1
>> n = sin(t) .^2 + cos(t) .^2        % we can directly do term-wise operation sin^2(t) + cos^2(t) of t as well
n =     1     1     1     1
>> t2 = (-10:6.666665:10)        % t2 is similar to t above but with ':' syntax of (start:increment:stop)
t2 =  -10.0000   -3.3333    3.3333   10.0000
```

Similarly, operations can be performed with matrices.

```
>>  (0+0) .^ (1/2) % term-by-term square root of the matrix obtained by adding 0=ones(4,5) to itself
ans =
    1.4142    1.4142    1.4142    1.4142    1.4142
    1.4142    1.4142    1.4142    1.4142    1.4142
    1.4142    1.4142    1.4142    1.4142    1.4142
    1.4142    1.4142    1.4142    1.4142    1.4142
```

We can access specific rows or columns of a matrix as follows:

```
>> % declare a 3 X 3 array A of row vectors
>> A = [0.2760    0.4984    0.7513; 0.6797    0.9597    0.2551; 0.1626    0.5853    0.6991]
A =
    0.2760    0.4984    0.7513
    0.6797    0.9597    0.2551
    0.1626    0.5853    0.6991
>> A(2,:) % access the second row of A
ans =
    0.6797    0.9597    0.2551
>> B = A(2:3,:) % store the second and third rows of A in matrix B
B =
    0.6797    0.9597    0.2551
    0.1626    0.5853    0.6991
>> C = A(:,[1 3]) % store the first and third columns of A in matrix C
C =
    0.2760    0.7513
    0.6797    0.2551
```

Labwork 11 (Plotting a function as points of ordered pairs in two arrays) Next we plot the function $sin(x)$ from several ordered pairs $(x_i, sin(x_i))$. Here x_i's are from the domain $[-2\pi, 2\pi]$. We use the `plot` function in MATLAB . Create an M-file called `MySineWave.m` and copy the following commands in it. By entering `MySineWave` in the command window you should be able to run the script and see the figure in the figure window.

SineWave.m

```
x = linspace(-2*pi,2*pi,100);         % x has 100 points equally spaced in [-2*pi, 2*pi]
y = sin(x);                % y is the term-wise sin of x, ie sin of every number in x is in y, resp.
plot(x,y,'.');                  % plot x versus y as dots should appear in the Figure window
xlabel('x');                % label x-axis with the single quote enclosed string x
ylabel('sin(x)','FontSize',16);           % label y-axis with the single quote enclosed string
title('Sine Wave in [-2 pi, 2 pi]','FontSize',16);        % give a title; click Figure window to see changes
set(gca,'XTick',-8:1:8,'FontSize',16) % change the range and size of X-axis ticks
% you can go to the Figure window's File menu to print/save the plot
```

The plot was saved as an encapsulated postscript file from the File menu of the Figure window and is displayed below.

CONDITIONAL PROBABILITY SUMMARY

$P(A|B)$ means the probability that A occurs given that B has occurred.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)} \quad \text{if} \quad P(B) \neq 0$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B)P(A|B)}{P(A)} \quad \text{if} \quad P(A) \neq 0$$

Conditional probabilities obey the 4 axioms of probability.

DISCRETE RANDOM VARIABLE SUMMARY

Probability mass function

$$f(x) = P(X = x_i)$$

Distribution function

$$F(x) = \sum_{x_i \leq x} f(x_i)$$

Random Variable	Possible Values	Probabilities	Modeled situations
Discrete uniform	$\{x_1, x_2, \ldots, x_k\}$	$P(X = x_i) = \frac{1}{k}$	Situations with k equally likely values. Parameter: k.
Bernoulli(θ)	$\{0, 1\}$	$P(X = 0) = 1 - \theta$ $P(X = 1) = \theta$	Situations with only 2 outcomes, coded 1 for success and 0 for failure. Parameter: $\theta = P(\text{success}) \in (0, 1)$.
Geometric(θ)	$\{1, 2, 3, \ldots\}$	$P(X = x)$ $= (1 - \theta)^{x-1}\theta$	Situations where you count the number of trials until the first success in a sequence of independent trails with a constant probability of success. Parameter: $\theta = P(\text{success}) \in (0, 1)$.
Binomial(n, θ)	$\{0, 1, 2, \ldots, n\}$	$P(X = x)$ $= \binom{n}{x}\theta^x(1-\theta)^{n-x}$	Situations where you count the number of success in n trials where each trial is independent and there is a constant probability of success. Parameters: $n \in \{1, 2, \ldots\}$; $\theta = P(\text{success}) \in (0, 1)$.
Poisson(λ)	$\{0, 1, 2, \ldots\}$	$P(X = x)$ $= \frac{\lambda^x e^{-\lambda}}{x!}$	Situations where you count the number of events in a continuum where the events occur one at a time and are independent of one another. Parameter: λ= rate $\in (0, \infty)$.

Summary of Probability Theory I

SET SUMMARY

$\{a_1, a_2, \ldots, a_n\}$	–	a set containing the elements, $a_1, a_2, \ldots, a_n$.
$a \in A$	–	a is an element of the set A.
$A \subseteq B$	–	the set A is a subset of B.
$A \cup B$	–	"union", meaning the set of all elements which are in A or B, or both.
$A \cap B$	–	"intersection", meaning the set of all elements in both A and B.
{} or $\emptyset$	–	empty set.
Ω	–	universal set.
A^c	–	the complement of A, meaning the set of all elements in Ω, the universal set, which are not in A.

EXPERIMENT SUMMARY

Experiment	–	an activity producing distinct outcomes.
Ω	–	set of all outcomes of the experiment.
ω	–	an individual outcome in Ω, called a simple event.
$A \subseteq \Omega$	–	a subset A of Ω is an event.
Trial	–	one performance of an experiment resulting in 1 outcome.

PROBABILITY SUMMARY

Axioms:

1. If $A \subseteq \Omega$ then $0 \leq P(A) \leq 1$ and $P(\Omega) = 1$.
2. If A, B are disjoint events, then $P(A \cup B) = P(A) + P(B)$.
   [This is true only when A and B are disjoint.]
3. If $A_1, A_2, \ldots$ are disjoint then $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$

Rules:

$$P(A^c) \;=\; 1 - P(A)$$

$$P(A \cup B) \;=\; P(A) + P(B) - P(A \cap B) \qquad \text{[always true]}$$

Figure 1.7: A plot of the sine wave over $[-2\pi, 2\pi]$.

1.8 Elementary Real Analysis

1.8.1 Limits of Real Numbers – A Review

Let us first recall some elementary ideas from real analysis.

Definition 3 (Convergent sequence of real numbers) A sequence of real numbers $\langle x_i \rangle_{i=1}^{\infty} := x_1, x_2, \ldots$ is said to converge to a limit $a \in \mathbb{R}$ and denoted by:

$$\lim_{i \to \infty} x_i = a \ ,$$

if for every natural number $m \in \mathbb{N}$, a natural number $N_m \in \mathbb{N}$ exists such that for every $j \geq N_m$, $|x_j - a| \leq \frac{1}{m}$.

In words, $\lim_{i\to\infty} x_i = a$ means the following: no matter how small you make $\frac{1}{m}$ by picking as large an m as you wish, I can find an N_m, that may depend on m, such that every number in the sequence beyond the N_m-th element is within distance $\frac{1}{m}$ of the limit a.

Example 12 (Limit of a sequence of 17s) Let $\langle x_i \rangle_{i=1}^{\infty} = 17, 17, 17, \ldots$. Then $\lim_{i\to\infty} x_i = 17$. This is because for every $m \in \mathbb{N}$, we can take $N_m = 1$ and satisfy the definition of the limit, i.e.:

$$\text{for every } j \geq N_m = 1, \ |x_j - 17| = |17 - 17| = 0 \leq \frac{1}{m} \ .$$

Example 13 (Limit of $1/i$) Let $\langle x_i \rangle_{i=1}^{\infty} = \frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \ldots$, i.e. $x_i = \frac{1}{i}$, then $\lim_{i\to\infty} x_i = 0$. This is because for every $m \in \mathbb{N}$, we can take $N_m = m$ and satisfy the definition of the limit, i.e.:

$$\text{for every } j \geq N_m = m, \ |x_j - 0| = \left|\frac{1}{j} - 0\right| = \frac{1}{j} \leq \frac{1}{m} \ .$$

However, several other sequences also approach the limit 0. Some such sequences that approach the limit 0 from the right are:

$$\langle x_{1:\infty} \rangle = \frac{1}{1}, \frac{1}{4}, \frac{1}{9}, \ldots \qquad \text{and} \qquad \langle x_{1:\infty} \rangle = \frac{1}{1}, \frac{1}{8}, \frac{1}{27}, \cdots \ ,$$

and some that approach the limit 0 from the left are:

$$\langle x_{1:\infty} \rangle = -\frac{1}{1}, -\frac{1}{2}, -\frac{1}{3}, \ldots \qquad \text{and} \qquad \langle x_{1:\infty} \rangle = -\frac{1}{1}, -\frac{1}{4}, -\frac{1}{9}, \cdots \ ,$$

and finally some that approach 0 from either side are:

$$\langle x_{1:\infty} \rangle = -\frac{1}{1}, +\frac{1}{2}, -\frac{1}{3}, \ldots \qquad \text{and} \qquad \langle x_{1:\infty} \rangle = -\frac{1}{1}, +\frac{1}{4}, -\frac{1}{9} \cdots \ .$$

When we do not particularly care about the specifics of a sequence of real numbers $\langle x_{1:\infty} \rangle$, in terms of the exact values it takes for each i, but we are only interested that it converges to a limit a we write:

$$x \to a$$

and say that x approaches a. If we are only interested in those sequences that converge to the limit a from the right or left, we write:

$$x \to a^+ \qquad \text{or} \qquad x \to a^-$$

and say x approaches a from the right or left, respectively.

Definition 4 (Limits of Functions) We say a function $f(x) : \mathbb{R} \to \mathbb{R}$ has a **limit** $L \in \mathbb{R}$ as x approaches a and write:

$$\lim_{x \to a} f(x) = L \ ,$$

provided $f(x)$ is arbitrarily close to L for all values of x that are sufficiently close to, but not equal to, a. We say that f has a **right limit** L_R or **left limit** L_L as x approaches a from the left or right, and write:

$$\lim_{x \to a^+} f(x) = L_R \qquad \text{or} \qquad \lim_{x \to a^-} f(x) = L_L \ ,$$

provided $f(x)$ is arbitrarily close to L_R or L_L for all values of x that are sufficiently close to, but not equal to, a from the right of a or the left of a, respectively. When the limit is not an element of $\mathbb{R}$ or when the left and right limits are distinct, we say that the limit does not exist.

Example 14 (Limit of $1/x^2$) Consider the function $f(x) = \frac{1}{x^2}$. Then

$$\lim_{x \to 1} f(x) = \lim_{x \to 1} \frac{1}{x^2} = 1$$

exists since the limit $1 \in \mathbb{R}$, and the right and left limits are the same:

$$\lim_{x \to 1^+} f(x) = \lim_{x \to 1^+} \frac{1}{x^2} = 1 \qquad \text{and} \qquad \lim_{x \to 1^-} f(x) = \lim_{x \to 1^-} \frac{1}{x^2} = 1 \ .$$

However, the following limit does not exist:

$$\lim_{x \to 0} f(x) = \lim_{x \to 0} \frac{1}{x^2} = \infty$$

since $\infty \notin \mathbb{R}$.

Let us next look at some limits of functions that exist despite the function itself being undefined at the limit point.

Example 15 (Limit of $(1+x)^{\frac{1}{x}}$) The limit of $f(x) = (1+x)^{\frac{1}{x}}$ as x approaches 0 exists and it is

6.6 Standard normal distribution function table

For any given value z, its cumulative probability $\Phi(z)$.

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
0.01	0.5040	0.51	0.6950	1.01	0.8438	1.51	0.9345	2.01	0.9778	2.51	0.9940
0.02	0.5080	0.52	0.6985	1.02	0.8461	1.52	0.9357	2.02	0.9783	2.52	0.9941
0.03	0.5120	0.53	0.7019	1.03	0.8485	1.53	0.9370	2.03	0.9788	2.53	0.9943
0.04	0.5160	0.54	0.7054	1.04	0.8508	1.54	0.9382	2.04	0.9793	2.54	0.9945
0.05	0.5199	0.55	0.7088	1.05	0.8531	1.55	0.9394	2.05	0.9798	2.55	0.9946
0.06	0.5239	0.56	0.7123	1.06	0.8554	1.56	0.9406	2.06	0.9803	2.56	0.9948
0.07	0.5279	0.57	0.7157	1.07	0.8577	1.57	0.9418	2.07	0.9808	2.57	0.9949
0.08	0.5319	0.58	0.7190	1.08	0.8599	1.58	0.9429	2.08	0.9812	2.58	0.9951
0.09	0.5359	0.59	0.7224	1.09	0.8621	1.59	0.9441	2.09	0.9817	2.59	0.9952
0.10	0.5398	0.60	0.7257	1.10	0.8643	1.60	0.9452	2.10	0.9821	2.60	0.9953
0.11	0.5438	0.61	0.7291	1.11	0.8665	1.61	0.9463	2.11	0.9826	2.61	0.9955
0.12	0.5478	0.62	0.7324	1.12	0.8686	1.62	0.9474	2.12	0.9830	2.62	0.9956
0.13	0.5517	0.63	0.7357	1.13	0.8708	1.63	0.9484	2.13	0.9834	2.63	0.9957
0.14	0.5557	0.64	0.7389	1.14	0.8729	1.64	0.9495	2.14	0.9838	2.64	0.9959
0.15	0.5596	0.65	0.7422	1.15	0.8749	1.65	0.9505	2.15	0.9842	2.65	0.9960
0.16	0.5636	0.66	0.7454	1.16	0.8770	1.66	0.9515	2.16	0.9846	2.66	0.9961
0.17	0.5675	0.67	0.7486	1.17	0.8790	1.67	0.9525	2.17	0.9850	2.67	0.9962
0.18	0.5714	0.68	0.7517	1.18	0.8810	1.68	0.9535	2.18	0.9854	2.68	0.9963
0.19	0.5753	0.69	0.7549	1.19	0.8830	1.69	0.9545	2.19	0.9857	2.69	0.9964
0.20	0.5793	0.70	0.7580	1.20	0.8849	1.70	0.9554	2.20	0.9861	2.70	0.9965
0.21	0.5832	0.71	0.7611	1.21	0.8869	1.71	0.9564	2.21	0.9864	2.71	0.9966
0.22	0.5871	0.72	0.7642	1.22	0.8888	1.72	0.9573	2.22	0.9868	2.72	0.9967
0.23	0.5910	0.73	0.7673	1.23	0.8907	1.73	0.9582	2.23	0.9871	2.73	0.9968
0.24	0.5948	0.74	0.7704	1.24	0.8925	1.74	0.9591	2.24	0.9875	2.74	0.9969
0.25	0.5987	0.75	0.7734	1.25	0.8944	1.75	0.9599	2.25	0.9878	2.75	0.9970
0.26	0.6026	0.76	0.7764	1.26	0.8962	1.76	0.9608	2.26	0.9881	2.76	0.9971
0.27	0.6064	0.77	0.7794	1.27	0.8980	1.77	0.9616	2.27	0.9884	2.77	0.9972
0.28	0.6103	0.78	0.7823	1.28	0.8997	1.78	0.9625	2.28	0.9887	2.78	0.9973
0.29	0.6141	0.79	0.7852	1.29	0.9015	1.79	0.9633	2.29	0.9890	2.79	0.9974
0.30	0.6179	0.80	0.7881	1.30	0.9032	1.80	0.9641	2.30	0.9893	2.80	0.9974
0.31	0.6217	0.81	0.7910	1.31	0.9049	1.81	0.9649	2.31	0.9896	2.81	0.9975
0.32	0.6255	0.82	0.7939	1.32	0.9066	1.82	0.9656	2.32	0.9898	2.82	0.9976
0.33	0.6293	0.83	0.7967	1.33	0.9082	1.83	0.9664	2.33	0.9901	2.83	0.9977
0.34	0.6331	0.84	0.7995	1.34	0.9099	1.84	0.9671	2.34	0.9904	2.84	0.9977
0.35	0.6368	0.85	0.8023	1.35	0.9115	1.85	0.9678	2.35	0.9906	2.85	0.9978
0.36	0.6406	0.86	0.8051	1.36	0.9131	1.86	0.9686	2.36	0.9909	2.86	0.9979
0.37	0.6443	0.87	0.8078	1.37	0.9147	1.87	0.9693	2.37	0.9911	2.87	0.9979
0.38	0.6480	0.88	0.8106	1.38	0.9162	1.88	0.9699	2.38	0.9913	2.88	0.9980
0.39	0.6517	0.89	0.8133	1.39	0.9177	1.89	0.9706	2.39	0.9916	2.89	0.9981
0.40	0.6554	0.90	0.8159	1.40	0.9192	1.90	0.9713	2.40	0.9918	2.90	0.9981
0.41	0.6591	0.91	0.8186	1.41	0.9207	1.91	0.9719	2.41	0.9920	2.91	0.9982
0.42	0.6628	0.92	0.8212	1.42	0.9222	1.92	0.9726	2.42	0.9922	2.92	0.9982
0.43	0.6664	0.93	0.8238	1.43	0.9236	1.93	0.9732	2.43	0.9925	2.93	0.9983
0.44	0.6700	0.94	0.8264	1.44	0.9251	1.94	0.9738	2.44	0.9927	2.94	0.9984
0.45	0.6736	0.95	0.8289	1.45	0.9265	1.95	0.9744	2.45	0.9929	2.95	0.9984
0.46	0.6772	0.96	0.8315	1.46	0.9279	1.96	0.9750	2.46	0.9931	2.96	0.9985
0.47	0.6808	0.97	0.8340	1.47	0.9292	1.97	0.9756	2.47	0.9932	2.97	0.9985
0.48	0.6844	0.98	0.8365	1.48	0.9306	1.98	0.9761	2.48	0.9934	2.98	0.9986
0.49	0.6879	0.99	0.8389	1.49	0.9319	1.99	0.9767	2.49	0.9936	2.99	0.9986
0.50	0.6915	1.00	0.8413	1.50	0.9332	2.00	0.9772	2.50	0.9938	3.00	0.9987

Example 187 (Drunkard's biased walk around the block) Consider the Markov chain $(X_t)_{t\in\mathbb{Z}_+}$ on $\mathbb{X}=\{0,1,2,3\}$ with initial distribution $\mathbb{1}_{\{3\}}(x)$ and transition matrix

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

Draw the transition diagram for this Markov chain that corresponds to a drunkard who flips a biased coin to make his next move at each corner. The stationary distribution is $\pi=(1/4,1/4,1/4,1/4)$ (verify $\pi P=\pi$).

We will show that $(X_t)_{t\in\mathbb{Z}_+}$ is not a reversible Markov chain. Sine $(X_t)_{t\in\mathbb{Z}_+}$ is irreducible (aperiodicity is not necessary for uniqueness of π) π is the unique stationary distribution. Due to Proposition 97, π has to be a reversible distribution in order for $(X_t)_{t\in\mathbb{Z}_+}$ to be a reversible Markov chain. But reversibility fails for π since,

$$\pi(0)P(0,1)=\frac{1}{4}\times\frac{1}{3}=\frac{1}{12}<\frac{1}{6}=\frac{1}{4}\times\frac{2}{3}=\pi(1)P(1,0)\ .$$

Exercise 188 Find the stationary distribution of the Markov chain in Exercise 180.

Model 22 (Random Walk on a Directed Graph) A random walk on a directed graph $\mathbb{G}=(\mathbb{V},\mathbb{E})$ is a Markov chain with state space $\mathbb{V}:=\{v_1,v_2,\ldots,v_k\}$ and transition matrix given by:

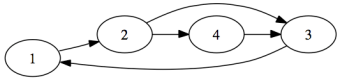
$$P(v_i,v_j)=\begin{cases}\frac{1}{\text{odeg}(v_i)} & \text{if } \langle v_i,v_j\rangle\in\mathbb{E}\\ 0 & \text{otherwise,}\end{cases}$$

Example 189 (Directed Triangulated Quadrangle) The random walk on the directed graph

$$\mathbb{G}=(\{1,2,3,4\},\{\langle1,2\rangle,\langle3,1\rangle,\langle2,3\rangle,\langle2,4\rangle,\langle4,3\rangle\})$$

depicted below with adjacency matrix A is a Markov chain on $\{1,2,3,4\}$ with transition matrix P:

$$A=\begin{array}{c} \\1\\2\\3\\4\end{array}\begin{array}{c}\begin{array}{cccc}1&2&3&4\end{array}\\\begin{pmatrix}0&1&0&0\\0&0&1&1\\1&0&0&0\\0&0&1&0\end{pmatrix}\end{array},\quad P=\begin{array}{c} \\1\\2\\3\\4\end{array}\begin{array}{c}\begin{array}{cccc}1&2&3&4\end{array}\\\begin{pmatrix}0&1&0&0\\0&0&\frac{1}{2}&\frac{1}{2}\\1&0&0&0\\0&0&1&0\end{pmatrix}\end{array},$$



.

Exercise 190 Show that the there is no reversible distibution for the Markov chain in Example 189.

Example 191 (Random surf on the word wide web) Consider the huge graph with vertices as webpages and hyper-links as undirected edges. Then Model 21 gives a random walk on this graph. However if a page has no links to other pages, it becomes a sink and therefore terminates the random walk. Let us modify this random walk into a **random surf** to avoid getting stuck. If the random surfer arrives at a sink page, she picks another page at random and continues surfing at random again. Google's PageRank formula uses a random surfer model who gets bored after several clicks and switches to a random page. The PageRank value of a page reflects the chance that the random surfer will land on that page by clicking on a link. The stationary distribution of the random surfer on the world wide web is a very successful model for ranking pages.

the Euler's constant e :

$$\begin{aligned}\lim_{x\to0}f(x) &= \lim_{x\to0}(1+x)^{\frac{1}{x}}\\ &= \lim_{x\to0}(x+1)^{(1/x)} \quad \text{Indeterminate form of type } 1^\infty.\\ &= \exp\left(\lim_{x\to0}\log((x+1)^{(1/x)})\right) \quad \text{Transformed using } \exp(\lim_{x\to0}\log((x+1)^{(1/x)}))\\ &= \exp\left(\lim_{x\to0}(\log(x+1))/x\right) \quad \text{Indeterminate form of type } 0/0.\\ &= \exp\left(\lim_{x\to0}\frac{d\log(x+1)/dx}{dx/dx}\right) \quad \text{Applying L'Hospital's rule}\\ &= \exp\left(\lim_{x\to0}1/(x+1)\right) \quad \text{limit of a quotient is the quotient of the limits}\\ &= \exp\left(1/(\lim_{x\to0}(x+1))\right) \quad \text{The limit of } x+1 \text{ as } x \text{ approaches 0 is 1}\\ &= \exp(1)=e\approxeq 2.71828\ .\end{aligned}$$

$$\lim_{x\to0}f(x)=\lim_{x\to0}(1+x)^{\frac{1}{x}}=e\approxeq 2.71828\ .$$

Notice that the above limit exists despite the fact that $f(0)=(1+0)^{\frac{1}{0}}$ itself is undefined and does not exist.

Example 16 (Limit of $\frac{x^3-1}{x-1}$) For $f(x)=\frac{x^3-1}{x-1}$, this limit exists:

$$\lim_{x\to1}f(x)=\lim_{x\to1}\frac{x^3-1}{x-1}=\lim_{x\to1}\frac{(x-1)(x^2+x+1)}{(x-1)}=\lim_{x\to1}x^2+x+1=3$$

despite the fact that $f(1)=\frac{1^3-1}{1-1}=\frac{0}{0}$ itself is undefined and does not exist.

Next we look at some examples of limits at infinity.

Example 17 (Limit of $(1-\frac{\lambda}{n})^n$) The limit of $f(n)=\left(1-\frac{\lambda}{n}\right)^n$ as n approaches ∞ exists and it is $e^{-\lambda}$:

$$\lim_{n\to\infty}f(n)=\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n=e^{-\lambda}\ .$$

Example 18 (Limit of $(1-\frac{\lambda}{n})^{-\alpha}$) The limit of $f(n)=\left(1-\frac{\lambda}{n}\right)^{-\alpha}$, for some $\alpha>0$, as n approaches ∞ exists and it is 1 :

$$\lim_{n\to\infty}f(n)=\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-\alpha}=1\ .$$

Definition 5 (Continuity of a function) We say a real-valued function $f(x):D\to\mathbb{R}$ with the domain $D\subset\mathbb{R}$ is **right continuous** or **left continuous** at a point $a\in D$, provided:

$$\lim_{x\to a^+}f(x)=f(a)\qquad\text{or}\qquad\lim_{x\to a^-}f(x)=f(a)\ ,$$

respectively. We say f is **continuous** at $a\in D$, provided:

$$\lim_{x\to a^+}f(x)=f(a)=\lim_{x\to a^-}f(x)\ .$$

Finally, f is said to be continuous if f is continuous at every $a\in D$.

Example 19 (Discontinuity of $f(x) = (1+x)^{\frac{1}{x}}$ at 0) Let us reconsider the function $f(x) = (1+x)^{\frac{1}{x}} : \mathbb{R} \to \mathbb{R}$. Clearly, $f(x)$ is continuous at 1, since:

$$\lim_{x\to 1} f(x) = \lim_{x\to 1}(1+x)^{\frac{1}{x}} = 2 = f(1) = (1+1)^{\frac{1}{1}} \ ,$$

but it is not continuous at 0, since:

$$\lim_{x\to 0} f(x) = \lim_{x\to 0}(1+x)^{\frac{1}{x}} = e \approxeq 2.71828 \neq f(0) = (1+0)^{\frac{1}{0}} \ .$$

Thus, $f(x)$ is not a continuous function over $\mathbb{R}$.

1.9 Elementary Number Theory

We introduce basic notions that we need from elementary number theory here. These notions include integer functions and modular arithmetic as they will be needed later on.

For any real number x:

$\lfloor x \rfloor := \max\{y : y \in \mathbb{Z} \text{ and } y \leq x\}$, i.e., the greatest integer less than or equal to x (the **floor** of x),

$\lceil x \rceil := \min\{y : y \in \mathbb{Z} \text{ and } y \geq x\}$, i.e., the least integer greater than or equal to x (the **ceiling** of x).

Example 20 (Floors and ceilings)

$$\lfloor 1 \rfloor = 1, \quad \lceil 1 \rceil = 1, \quad \lfloor 17.8 \rfloor = 17, \quad \lfloor -17.8 \rfloor = -18, \quad \lceil \sqrt{2} \rceil = 2, \quad \lfloor \pi \rfloor = 3, \quad \lceil \frac{1}{10^{100}} \rceil = 1.$$

Labwork 21 (Floors and ceilings in MATLAB) We can use MATLAB functions `floor` and `ceil` to compute $\lfloor x \rfloor$ and $\lceil x \rceil$, respectively. Also, the argument x to these functions can be an array.

```
>> sqrt(2) % the square root of 2 is
ans =    1.4142
>> ceil(sqrt(2)) % ceiling of square root of 2
ans =     2
>> floor(-17.8) % floor of -17.8
ans =   -18
>> ceil([1 sqrt(2) pi -17.8 1/(10^100)]) %the ceiling of each element of an array
ans =     1     2     4   -17     1
>> floor([1 sqrt(2) pi -17.8 1/(10^100)]) % the floor of each element of an array
ans =     1     1     3   -18     0
```

Classwork 22 (Relations between floors and ceilings) Convince yourself of the following formulae. Use examples, plots and/or formal arguments.

$$\begin{aligned} \lceil x \rceil = \lfloor x \rfloor &\iff x \in \mathbb{Z} \\ \lceil x \rceil = \lfloor x \rfloor + 1 &\iff x \notin \mathbb{Z} \\ \lfloor -x \rfloor &= -\lceil x \rceil \\ x - 1 < \lfloor x \rfloor \leq x \leq \lceil x \rceil &< x + 1 \end{aligned}$$

Let us define modular arithmetic next. Suppose x and y are any real numbers, i.e. $x, y \in \mathbb{R}$, we define the binary operation called "$x \mod y$" as:

$$x \mod y := \begin{cases} x - y\lfloor x/y \rfloor & \text{if } y \neq 0 \\ x & \text{if } y = 0 \end{cases}$$

Proposition 99 The random walk on an undirected graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$, with vertex set $\mathbb{V} := \{v_1, v_2, \ldots, v_k\}$ and degree sum $d = \sum_{i=1}^{k} \deg(v_i)$ is a reversible Markov chain with the reversible distribution π given by:

$$\pi = \left(\frac{\deg(v_1)}{d}, \frac{\deg(v_2)}{d}, \ldots, \frac{\deg(v_k)}{d} \right) \ .$$

Proof: First note that π is a probability distribution provided that $d > 0$. To show that π is reversible we need to verify Equation 6.12 for each $(v_i, v_j) \in \mathbb{V}^2$. Fix a pair of states $(v_i, v_j) \in \mathbb{V}^2$, then

$$\pi(v_i)P(v_i, v_j) = \begin{cases} \frac{\deg(v_i)}{d}\frac{1}{\deg(v_i)} = \frac{1}{d} = \frac{\deg(v_j)}{d}\frac{1}{\deg(v_j)} = \pi(v_j)P(v_j, v_i) & \text{if } \langle v_i, v_j \rangle \in \mathbb{E} \\ 0 = \pi(v_j)P(v_j, v_i) & \text{otherwise.} \end{cases}$$

By Proposition 97 π is also the stationary distribution.

Exercise 184 Prove Proposition 99 by directly showing that $\pi P = \pi$, i.e., for each $v_i \in \mathbb{V}$, $\sum_{i=1}^{k} \pi(v_i)P(v_i, v_j) = \pi(v_j)$.

Example 185 (Random Walk on a regular graph) A graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ is called regular if every vertex in $\mathbb{V} = \{v_1, v_2, \ldots, v_k\}$ has the same degree δ, i.e., $\deg(v_i) = \delta$ for every $v_i \in \mathbb{V}$. Consider the random walk on a regular graph with symmetric transition matrix

$$Q(v_i, v_j) = \begin{cases} \frac{1}{\delta} & \text{if } \langle v_i, v_j \rangle \in \mathbb{E} \\ 0 & \text{otherwise} \end{cases} \ .$$

By Proposition 99, the stationary distribution of the random walk on $\mathbb{G}$ is the uniform distribution on $\mathbb{V}$ given by

$$\pi = \left(\frac{\delta}{\delta \# \mathbb{V}}, \ldots, \frac{\delta}{\delta \# \mathbb{V}} \right) = \left(\frac{1}{\# \mathbb{V}}, \ldots, \frac{1}{\# \mathbb{V}} \right) \ .$$

Example 186 (Triangulated Quadrangle) The random walk on the undirected graph

$$\mathbb{G} = (\{1, 2, 3, 4\}, \{\langle 1, 2 \rangle, \langle 3, 1 \rangle, \langle 2, 3 \rangle, \langle 2, 4 \rangle, \langle 4, 3 \rangle\})$$

depicted below with adjacency matrix A is a Markov chain on $\{1, 2, 3, 4\}$ with transition matrix P:

$$A = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \end{array}, \quad P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \end{array},$$

.

By Proposition 99, the stationary distribution of the random walk on $\mathbb{G}$ is

$$\pi = \left(\frac{\deg(v_1)}{d}, \frac{\deg(v_2)}{d}, \frac{\deg(v_3)}{d}, \frac{\deg(v_4)}{d} \right) = \left(\frac{2}{10}, \frac{3}{10}, \frac{3}{10}, \frac{2}{10} \right) \ .$$

In words, $\pi(x)P(x,y) = \pi(y)P(y,x)$ says that if you start the chain at the reversible distribution π, i.e., $\mu_0 = \pi$, then the probability of going from x to y is the same as that of going from y to x.

Proposition 97 (A reversible π is a stationary π) Let $(X_t)_{t\in\mathbb{Z}_+}$ be a Markov chain on $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$ with transition matrix P. If π is a reversible distribution for $(X_t)_{t\in\mathbb{Z}_+}$ then π is a stationary distribution for $(X_t)_{t\in\mathbb{Z}_+}$.

Proof: Suppose π is a reversible distribution for $(X_t)_{t\in\mathbb{Z}_+}$ then π is a probability distribution on $\mathbb{X}$ and $\pi(x)P(x,y) = \pi(y)P(y,x)$ for each $(x,y) \in \mathbb{X}^2$. We need to show that for any $y \in \mathbb{X}$ we have

$$\pi(y) = \sum_{x\in\mathbb{X}} \pi(y)P(y,x) \ .$$

Fix a $y \in \mathbb{X}$,

$$\begin{aligned} LHS &= \pi(y) = \pi(y)\,1 = \pi(y) \sum_{x\in\mathbb{X}} P(y,x), \text{ since } P \text{ is a stochastic matrix} \\ &= \sum_{x\in\mathbb{X}} \pi(y)P(y,x) = \sum_{x\in\mathbb{X}} \pi(x)P(x,y), \text{ by reversibility} \\ &= RHS \ . \end{aligned}$$

Definition 98 (Graph) A **Graph** $\mathbb{G} := (\mathbb{V}, \mathbb{E})$ consists of a **vertex set** $\mathbb{V} := \{v_1, v_2, \ldots, v_k\}$ together with an **edge set** $\mathbb{E} := \{e_1, e_2, \ldots, e_l\}$. Each edge connects two of the vertices in $\mathbb{V}$. An edge e_h connecting vertices v_i and v_j is denoted by $\langle v_i, v_j\rangle$. Two vertices are **neighbours** if they share an edge. The **negihbourhood** of a vertex v_i denoted by $\text{nbhd}(v_i) := \{v_j : \langle v_i, v_j\rangle \in \mathbb{E}\}$ is the set of neighbouring vertices of v_i. The number of neighbours of a vertex v_i in an undirected graph is called its **degree** and is denoted by $\deg(v_i)$. Note that $\deg(v_i) = \#\,\text{nbhd}(v_i)$. In a graph we only allow one edge per pair of vertices but in a **multigraph** we allow more than one edge per pair of vertices. An edge can be **directed** to preserve the order of the pair of vertices they connect or they can be **undirected**. An edge can be **weighted** by being associated with a real number called its weight. We can represent a directed graph by its **adjacency matrix** given by:

$$A := (A(v_i, v_j))_{(v_i,v_j)\in\mathbb{V}\times\mathbb{V}}, \quad A(v_i, v_j) = \begin{cases} 1 & \text{if } \langle v_i, v_j\rangle \in \mathbb{E} \\ 0 & \text{otherwise} \ . \end{cases}$$

Thus the adjacency matrix of an undirected graph is symmetric. In a directed graph, each vertex v_i has **in-edges** that come into it and **out-edges** that go out of it. The number of in-edges and out-edges of v_i is denoted by $\text{ideg}(v_i)$ and $\text{odeg}(v_i)$ respectively. Note that a transition diagram of a Markov chain is a weighted directed graph and is represented by the transition probability matrix.

Model 21 (Random Walk on an Undirected Graph) A random walk on an undirected graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ is a Markov chain with state space $\mathbb{V} := \{v_1, v_2, \ldots, v_k\}$ and the following transition rules: if the chain is at vertex v_i at time t then it moves uniformly at random to one of the neighbours of v_i at time $t+1$. If $\deg(v_i)$ is the degree of v_i then the transition probabilities of this Markov chain is

$$P(v_i, v_j) = \begin{cases} \frac{1}{\deg(v_i)} & \text{if } \langle v_i, v_j\rangle \in \mathbb{E} \\ 0 & \text{otherwise,} \end{cases}$$

Chapter 2

Probability Model

2.1 Experiments

Ideas about chance events and random behaviour arose out of thousands of years of game playing, long before any attempt was made to use mathematical reasoning about them. Board and dice games were well known in Egyptian times, and Augustus Caesar gambled with dice. Calculations of odds for gamblers were put on a proper theoretical basis by Fermat and Pascal in the early 17th century.

Definition 6 An **experiment** is an activity or procedure that produces distinct, well-defined possibilities called **outcomes**. The set of all outcomes is called the **sample space**, and is denoted by Ω.

The subsets of Ω are called **events**. A single outcome, ω, when seen as a subset of Ω, as in $\{\omega\}$, is called a **simple event**.

Events, E_1, E_2 ... E_n, that cannot occur at the same time are called **mutually exclusive** events, or **pair-wise disjoint** events. This means that $E_i \cap E_j = \emptyset$ where $i \neq j$.

Example 23 Some standard examples of experiments are the following:

- $\Omega = \{\mathsf{Defective},\ \mathsf{Non\text{-}defective}\}$ if our experiment is to inspect a light bulb.

  There are only two outcomes here, so $\Omega = \{\omega_1, \omega_2\}$ where $\omega_1 = \mathsf{Defective}$ and $\omega_2 = \mathsf{Non\text{-}defective}$.

- $\Omega = \{\mathsf{Heads},\ \mathsf{Tails}\}$ if our experiment is to note the outcome of a coin toss.

  This time, $\Omega = \{\omega_1, \omega_2\}$ where $\omega_1 = \mathsf{Heads}$ and $\omega_2 = \mathsf{Tails}$.

- If our experiment is to roll a die then there are six outcomes corresponding to the number that shows on the top. For this experiment, $\Omega = \{1, 2, 3, 4, 5, 6\}$.

  Some examples of events are the set of odd numbered outcomes $A = \{1, 3, 5\}$, and the set of even numbered outcomes $B = \{2, 4, 6\}$.

  The simple events of Ω are $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}$, and $\{6\}$.

The outcome of a random experiment is uncertain until it is performed and observed. Note that sample spaces need to reflect the problem in hand. The example below is to convince you that an experiment's sample space is merely a collection of distinct elements called outcomes and these outcomes have to be *discernible in some well-specified sense* to the experimenter!

Example 24 Consider a generic die-tossing experiment by a human experimenter. Here $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots, \omega_6\}$, but the experiment might correspond to rolling a die whose faces are:

1. sprayed with six different scents (nose!), or
2. studded with six distinctly flavoured candies (tongue!), or
3. contoured with six distinct bumps and pits (touch!), or
4. acoustically discernible at six different frequencies (ears!), or
5. painted with six different colours (eyes!), or
6. marked with six different numbers $\mathsf{1}, \mathsf{2}, \mathsf{3}, \mathsf{4}, \mathsf{5}, \mathsf{6}$ (eyes!), or , ...

These six experiments are equivalent as far as probability goes.

Definition 7 A **trial** is a single performance of an experiment and it results in an outcome.

Example 25 Some standard examples of a trial are:

- A roll of a die.
- A toss of a coin.
- A release of a chaotic double pendulum.

An experimenter often performs more than one trial. Repeated trials of an experiment forms the basis of science and engineering as the experimenter learns about the phenomenon by repeatedly performing the same mother experiment with possibly different outcomes. This repetition of trials in fact provides the very motivation for the definition of probability.

Definition 8 An **n-product experiment** is obtained by repeatedly performing n trials of some experiment. The experiment that is repeated is called the "mother" experiment.

Example 26 (Toss a coin n times) Suppose our experiment entails tossing a coin n times and recording $\mathtt{H}$ for Heads and $\mathtt{T}$ for Tails. When $n = 3$, one possible outcome of this experiment is $\mathtt{HHT}$, ie. a Head followed by another Head and then a Tail. Seven other outcomes are possible.

The sample space for "toss a coin three times" experiment is:

$$\Omega = \{\mathtt{H}, \mathtt{T}\}^3 = \{\mathtt{HHH}, \mathtt{HHT}, \mathtt{HTH}, \mathtt{HTT}, \mathtt{THH}, \mathtt{THT}, \mathtt{TTH}, \mathtt{TTT}\} \ ,$$

with a particular sample point or outcome $\omega = \mathtt{HTH}$, and another distinct outcome $\omega' = \mathtt{HHH}$. An event, say A, that 'at least two Heads occur' is the following subset of Ω:

$$A = \{\mathtt{HHH}, \mathtt{HHT}, \mathtt{HTH}, \mathtt{THH}\} \ .$$

Another event, say B, that 'no Heads occur' is:

$$B = \{\mathtt{TTT}\}$$

Note that the event B is also an outcome or sample point. Another interesting event is the empty set $\emptyset \subset \Omega$. The event that 'nothing in the sample space occurs' is $\emptyset$.

Proposition 94 (Markov chain convergence theorem) Let $(X_t)_{t\in\mathbb{Z}_+}$ be an irreducible aperiodic Markov chain with state space $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$, transition matrix $P = (P(x,y))_{(x,y)\in\mathbb{X}^2}$ and initial distribution μ_0. Then for any distribution π which is stationary for the transition matrix P, we have

$$\mu_t \xrightarrow{\mathsf{TV}} \pi \ . \tag{6.11}$$

Proof: TBD

Proposition 95 (Uniqueness of stationary distribution) Any irreducible aperiodic Markov chain has a unique stationary distribution.

Proof: TBD

Exercise 181 Consider the Markov chain on $\{1,2,3,4,5,6\}$ with the following transition matrix:

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix} \end{array} .$$

Show that this chain is reducible and it has three stationary distributions:

$$(1/2, 1/2, 0, 0, 0, 0), \quad (0, 0, 1/2, 1/2, 0, 0), \quad (0, 0, 0, 0, 1/2, 1/2) \ .$$

Exercise 182 If there are two stationary distributions π and π' then show that there is a infinite family of stationary distributions $\{\pi_p : p \in [0,1]\}$, called the convex combinations of π and π'.

Exercise 183 Show that for a drunkard's walk chain started at state 0 around a polygonal block with k corners labelled $\{0, 1, 2, \ldots, k-1\}$, the state probability vector at time step t

$$\mu_t \xrightarrow{\mathsf{TV}} \pi$$

if and only if k is odd. Explain what happens to μ_t when k is even.

6.5 Reversibility

We introduce another specific property called reversibility. This property will assist in conjuring Markov chains with a desired stationary distibution.

Definition 96 (Reversible) A probability distribution π on $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$ is said to be a **reversible distribution** for a Markov chain $(X_t)_{t\in\mathbb{Z}}$ on $\mathbb{X}$ with transition matrix P if for every pair of states $(x, y) \in \mathbb{X}^2$:

$$\pi(x)P(x,y) = \pi(y)P(y,x) \ . \tag{6.12}$$

A Markov chain that has a reversible distribution is said to be a reversible Markov chain.

and let $T(x,y) = \min\{\} = \infty$ if the Markov chain never visits y after having started from x. Let the **mean hitting time**

$$\tau(x,y) := \mathbf{E}(T(x,y)),$$

be the expected time taken to reach y after having started at x. Note that $\tau(x,x)$ is the **mean return time** to state x.

Proposition 91 (Hitting times of irreducible aperiodic Markov chains) If $(X_t)_{t\in\mathbb{Z}_+}$ is an irreducible aperiodic Markov chain with state space $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$, transition matrix $P = (P(x,y))_{(x,y)\in\mathbb{X}^2}$ then for any pair of states $(x,y) \in \mathbb{X}^2$,

$$\mathbf{P}\left(T(x,y) < \infty\right) = 1 \ ,$$

and the mean hitting time is finite, i.e.,

$$\tau(x,y) < \infty \ .$$

Proposition 92 (Existence of Stationary distribution) For any irreducible and aperiodic Markov chain there exists at least one stationary distribution.

Proof: TBD

Definition 93 (Total variation distance) If $\nu_1 := (\nu_1(x))_{x\in\mathbb{X}}$ and $\nu_2 := (\nu_2(x))_{x\in\mathbb{X}}$ are elements of $\mathcal{P}(\mathbb{X})$, the set of all probability distributions on $\mathbb{X} := \{s_1, s_2, \ldots, s_k\}$, then we define the **total variation distance** between ν_1 and ν_2 as

$$\mathsf{d_{TV}}\left(\nu_1, \nu_2\right) := \frac{1}{2}\sum_{x\mathbb{X}} \mathrm{abs}\left(\nu_1(x) - \nu_2(x)\right), \quad \mathsf{d_{TV}} : \mathcal{P}(\mathbb{X}) \times \mathcal{P}(\mathbb{X}) \to [0,1] \ . \tag{6.10}$$

If $\nu_1, \nu_2, \ldots$ and ν are probability distributions on $\mathbb{X}$, then we say that ν_t **converges in total variation** to ν as $n \to \infty$ and write $\nu_t \xrightarrow{\mathsf{TV}} \nu$, if

$$\lim_{t\to\infty} \mathsf{d_{TV}}\left(\nu_t, \nu\right) = 0 \ .$$

Observe that if $\mathsf{d_{TV}}(\nu_1,\nu_2) = 0$ then $\nu_1 = \nu_2$. The constant $1/2$ in Equation 6.10 ensures that the range of $\mathsf{d_{TV}}$ is in $[0,1]$. If $\mathsf{d_{TV}}(\nu_1,\nu_2) = 1$ then ν_1 and ν_2 have disjoint supports, i.e., we can partition $\mathbb{X}$ into $\mathbb{X}_1$ and $\mathbb{X}_2$, i.e., $\mathbb{X} = \mathbb{X}_1 \cup \mathbb{X}_2$ and $\mathbb{X}_1 \cap \mathbb{X}_2 = \emptyset$, such that $\sum_{x\in\mathbb{X}_1} \nu_1(x) = 1$ and $\sum_{x\in\mathbb{X}_2} \nu_2(x) = 1$. The total variation distance gets its name from the following natural interpretation:

$$\mathsf{d_{TV}}\left(\nu_1, \nu_2\right) = \max_{A\subset\mathbb{X}} \mathrm{abs}\left(\nu_1(A) - \nu_2(A)\right) \ .$$

This interpretation means that the total variation distance between ν_1 and ν_2 is the maximal difference in probabilities that the two distributions assign to any one event $A \in \sigma(\mathbb{X}) = 2^{\mathbb{X}}$.

In words, Proposition 94 says that if you run the chain for a sufficiently long enough time t, then, regardless of the initial distribution μ_0, the distribution at time t will be close to the stationary distribution π. This is referred to as the Markov chain **approaching equilibrium** or **stationarity** as $t \to \infty$.

Figure 2.1: A binary tree whose leaves are all possible outcomes.

Classwork 27 (A thrice-bifurcating tree of outcomes) Can you think of a graphical way to enumerate the outcomes of the Experiment 26? Draw a diagram of this under the caption of Figure 2.1, using the caption as a hint (in other words, draw your own Figure 2.1).

EXPERIMENT SUMMARY

Experiment	–	an activity producing distinct outcomes.
Ω	–	set of all outcomes of the experiment.
ω	–	an individual outcome in Ω, called a simple event.
$A \subseteq \Omega$	–	a subset A of Ω is an event.
Trial	–	one performance of an experiment resulting in 1 outcome.

2.2 Probability

The mathematical model for probability or the probability model is an axiomatic system that may be motivated by the intuitive idea of 'long-term relative frequency'. If the axioms and definitions are intuitively motivated, the probability model simply follows from the application of logic to these axioms and definitions. No attempt to define probability in the real world is made. However, the application of probability models to real-world problems through statistical experiments has a fruitful track record. In fact, you are here for exactly this reason.

Idea 9 (The long-term relative frequency (LTRF) idea) Suppose we are interested in the fairness of a coin, i.e. if landing Heads has the same "probability" as landing Tails. We can toss it n times and call $N(\mathtt{H}, n)$ the fraction of times we observed Heads out of n tosses. Suppose that after conducting the tossing experiment 1000 times, we rarely observed Heads, e.g. 9 out of the 1000 tosses, then $N(\mathtt{H}, 1000) = 9/1000 = 0.009$. Suppose we continued the number of tosses to a million and found that this number approached closer to 0.1, or, more generally, $N(\mathtt{H}, n) \to 0.1$ as $n \to \infty$. We might, at least intuitively, think that the coin is unfair and has a lower "probability" of 0.1 of landing Heads. We might think that it is fair had we observed $N(\mathtt{H}, n) \to 0.5$ as $n \to \infty$. Other crucial assumptions that we have made here are:

1. **Something Happens**: Each time we toss a coin, we are certain to observe Heads **or** Tails, denoted by $\mathtt{H} \cup \mathtt{T}$. The probability that "something happens" is 1. More formally:

$$N(\mathtt{H} \cup \mathtt{T}, n) = \frac{n}{n} = 1.$$

This is an intuitively reasonable assumption that simply says that one of the possible outcomes is certain to occur, provided the coin is not so thick that it can land on or even roll along its circumference.

2. **Addition Rule**: Heads and Tails are mutually exclusive events in any given toss of a coin, i.e. they cannot occur simultaneously. The intersection of mutually exclusive events is the empty set and is denoted by $\mathtt{H} \cap \mathtt{T} = \emptyset$. The event $\mathtt{H} \cup \mathtt{T}$, namely that the event that "coin lands Heads **or** coin lands Tails" satisfies:

$$N(\mathtt{H} \cup \mathtt{T}, n) = N(\mathtt{H}, n) + N(\mathtt{T}, n).$$

3. The coin-tossing experiment is repeatedly performed in an **independent** manner, i.e. the outcome of any individual coin-toss does not affect that of another. This is an intuitively reasonable assumption since the coin has no memory and the coin is tossed identically each time.

We will use the LTRF idea more generally to motivate a mathematical model of probability called probability model. Suppose A is an event associated with some experiment $\mathcal{E}$, so that A either does or does not occur when the experiment is performed. We want the probability that event A occurs in a specific performance of $\mathcal{E}$, denoted by $\mathbf{P}(A)$, to intuitively mean the following: if one were to perform a super-experiment $\mathcal{E}^\infty$ by independently repeating the experiment $\mathcal{E}$ and recording $N(A, n)$, the fraction of times A occurs in the first n performances of $\mathcal{E}$ within the super-experiment $\mathcal{E}^\infty$. Then the LTRF idea suggests:

$$N(A, n) := \frac{\text{Number of times } A \text{ occurs}}{n = \text{Number of performances of } \mathcal{E}} \to \mathbf{P}(A), \ as \quad n \to \infty \tag{2.1}$$

Now, we are finally ready to define probability.

Definition 10 (Probability) Let $\mathcal{E}$ be an experiment with sample space Ω. Let $\mathcal{F}$ denote a suitable collection of events in Ω that satisfy the following conditions:

1. It (the collection) contains the sample space: $\boxed{\Omega \in \mathcal{F}}$.
2. It is closed under complementation: $\boxed{A \in \mathcal{F} \quad \Longrightarrow \quad A^c \in \mathcal{F}}$.
3. It is closed under countable unions: $\boxed{A_1, A_2, \ldots \in \mathcal{F} \quad \Longrightarrow \quad \bigcup_i A_i := A_1 \cup A_2 \cup \cdots \in \mathcal{F}}$.

Formally, this collection of events is called a **sigma field** or a **sigma algebra**. Our experiment $\mathcal{E}$ has a sample space Ω and a collection of events $\mathcal{F}$ that satisfy the three condition.

Given a double, e.g. $(\Omega, \mathcal{F})$, **probability** is just a function $\mathbf{P}$ which assigns each event $A \in \mathcal{F}$ a number $\mathbf{P}(A)$ in the real interval $[0, 1]$, i.e. $\boxed{\mathbf{P} : \mathcal{F} \to [0, 1]}$, such that:

1. The 'Something Happens' axiom holds, i.e. $\boxed{\mathbf{P}(\Omega) = 1}$.
2. The 'Addition Rule' axiom holds, i.e. for events A and B:

$$\boxed{A \cap B = \emptyset \quad \Longrightarrow \quad \mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)}\ .$$

Proof: TBD

Exercise 179 (King's random walk on a chessboard) Consider the squares in the chessboard as the state space $\mathbb{X} = \{0, 1, 2, \ldots, 7\}^2$ with a randomly walking black king, i.e., for each move from current state $(u, v) \in \mathbb{X}$ the king chooses one of his $k(u, v)$ possible moves uniformly at random. Is the Markov chain corresponding to the randomly walking black king on the chessboard irredicible and/or aperiodic?

Exercise 180 (King's random walk on a chesstorus) We can obtain a chesstorus from a pliable chessboard by identifying the eastern edge with the western edge (roll the chessboard into a cylinder) and then identifying the northern edge with the southern edge (gluing the top and bottom end of the cylinder together by turning into a doughnut or torus). Consider the squares in the chesstorus as the state space $\mathbb{X} = \{0, 1, 2, \ldots, 7\}^2$ with a randomly walking black king, i.e., for each move from current state $(x, y) \in \mathbb{X}$ the king chooses one of his 8 possible moves uniformly at random according to the scheme: $X_t \leftarrow X_{t-1} + W_t$, where W_t is independent and identically distributed as follows:

$$\mathbf{P}(W_t = w) = \begin{cases} \frac{1}{8} & \text{if } w \in \{(1,1),(1,0),(1,-1),(0,-1),(-1,-1),(-1,0),(-1,1),(0,1)\}, \\ 0 & \text{otherwise.} \end{cases}$$

Is the Markov chain corresponding to the randomly walking black king on the chesstorus irredicible and/or aperiodic? Write a MATLAB script to simulate a sequence of n states visited by the king if he started from $(0, 0)$ on the chesstorus.

6.4 Stationarity

We are interested in statements about a Markov chain that has been running for a long time. For any nontrivial Markov chain $(X_0, X_1, \ldots)$ the value of X_t will keep fluctuating in the state space $\mathbb{X}$ as $t \to \infty$ and we cannot hope for convergence to a fixed point state $x^* \in \mathbb{X}$ or to a k-cycle of states $\{x_1, x_2, \ldots, x_k\} \subset \mathbb{X}$. However, we can look one level up into the space of probability distributions over $\mathbb{X}$ that give the probability of the Markov chain visiting each state $x \in \mathbb{X}$ at time t, and hope that the distribution of X_t over $\mathbb{X}$ settles down as $t \to \infty$. The Markov chain convergence theorem indeed sattes that the distribution of X_t over $\mathbb{X}$ settles down as $t \to \infty$, provided the Markov chain is irreducible and aperiodic.

Definition 89 (Stationary distribution) Let $(X_t)_{t \in \mathbb{Z}_+}$ be a Markov chain with state space $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$ and transition matrix $P = (P(x, y))_{(x,y) \in \mathbb{X}^2}$. A row vector

$$\pi = (\pi(s_1), \pi(s_2), \ldots, \pi(s_k)) \in \mathbb{R}^{1 \times k}$$

is said to be a **stationary distribution** for the Markov chain, if it satisfies the conditions of being:

1. *a probability distribution*: $\pi(x) \geq 0$ for each $x \in \mathbb{X}$ and $\sum_{x \in \mathbb{X}} \pi(x) = 1$, and
2. *a fixed point*: $\pi P = \pi$, i.e., $\sum_{x \in \mathbb{X}} \pi(x) P(x, y) = \pi(y)$ for each $y \in \mathbb{X}$.

Definition 90 (Hitting times) If a Markov chain $(X_t)_{t \in \mathbb{Z}_+}$ with state space $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$ and transition matrix $P = (P(x, y))_{(x,y) \in \mathbb{X}^2}$ starts at state x, then we can define the **hitting time**

$$T(x, y) = \min\{t \geq 1 : X_t = y\}\ .$$

Definition 83 (Return times and period) Let $\mathbb{T}(x) := \{t \in \mathbb{N} : P^t(x,x) > 0\}$ be the set of **possible return times** to the starting state x. The **period** of state x is defined to be $\gcd(\mathbb{T}(x))$, the greatest common divisor of $\mathbb{T}(x)$. When the period of a state x is 1, i.e., $\gcd(\mathbb{T}(x)) = 1$, then x is said to be an **aperiodic state**.

Proposition 84 If the Markov chain $(X_t)_{t\in\mathbb{Z}_+}$ with transition matrix P on state space $\mathbb{X}$ is irreducible then $\gcd(\mathbb{T}(x)) = \gcd(\mathbb{T}(y))$ for any $(x,y) \in \mathbb{X}^2$.

Proof: Fix any pair of states $(x,y) \in \mathbb{X}^2$. Since, P is irreducible, $x \leftrightarrow y$ and therefore there exists natural numbers $\eta(x,y)$ and $\eta(y,x)$ such that $P^{\eta(x,y)}(x,y) > 0$ and $P^{\eta(y,x)}(y,x) > 0$. Let $\eta' = \eta(x,y) + \eta(y,x)$ and observe that $\eta' \in \mathbb{T}(x) \cap \mathbb{T}(y)$, $\mathbb{T}(x) \subset \mathbb{T}(y) - \eta' := \{t - \eta' : t \in \mathbb{T}(y)\}$ and $\gcd(\mathbb{T}(y))$ divides all elements in $\mathbb{T}(x)$. Thus, $\gcd(\mathbb{T}(y)) \leq \gcd(\mathbb{T}(x))$. By a similar argument we can also conclude that $\gcd(\mathbb{T}(x)) \leq \gcd(\mathbb{T}(y))$. Therefore $\gcd(\mathbb{T}(x)) = \gcd(\mathbb{T}(y))$.

Definition 85 (Aperiodic) A Markov chain $(X_t)_{t\in\mathbb{Z}_+}$ with transition matrix P on state space $\mathbb{X}$ is said to be aperiodic if all of its states are aperiodic, i.e., $\gcd(\mathbb{T}(x)) = 1$ for every $x \in \mathbb{X}$. If a chain is not aperiodic, we call it **periodic**.

We have already seen example of irreducible Markov chains that were either periodic or aperiodic. For instance, Freddy's Markov chain with $(p,q) \in (0,1)^2$ is aperiodic since the period of either of its two states is given by $\gcd(\{1,2,3,\ldots\}) = 1$. However, the Markov chain model for a drunkard's walk around a block over the state space $\{0,1,2,3\}$ is periodic because you can only return to the starting state in an even number of time steps and

$$\gcd(\mathbb{T}(0)) = \gcd(\mathbb{T}(1)) = \gcd(\mathbb{T}(2)) = \gcd(\mathbb{T}(3)) = \gcd(\{2,4,6,\ldots\}) = 2 \neq 1 \ .$$

Exercise 178 Show that the Markov chain corresponding to a drunkard's walk around a polygonal block with k corners is irreducible for any integer $k > 1$. Show that it is aperiodic only when k is odd and has period 2 when k is even.

Proposition 86 Let $A = \{a_1, a_2, \ldots\} \subset \mathbb{N}$ that satisfies the following two conditions:

1. A is a **nonlattice**, meaning that $\gcd(A) = 1$ and
2. A is closed undur addition, meaning that if $(a,a') \in A^2$ then $a + a' \in A$.

Then there exists a positive integer $\eta < \infty$ such that $n \in A$ for all $n \geq \eta$.

Proof: See Proofs of Lemma 1.1, Lemma 1.2 and Theorem 1.1 in Appendix of *Pierre Brémaud, Markov Chains, Gibbs Fields, Monte Carlo Simulation, and Queues, Springer, 1999.*

Proposition 87 If the Markov chain $(X_t)_{t\in\mathbb{Z}_+}$ with transition matrix P on state space $\mathbb{X}$ is irreducible and aperiodic then there is an integer τ such that $P^t(x,x) > 0$ for all $t \geq \tau$ and all $x \in \mathbb{X}$.

Proof: TBD

Proposition 88 If the Markov chain $(X_t)_{t\in\mathbb{Z}_+}$ with transition matrix P on state space $\mathbb{X}$ is irreducible and aperiodic then there is an integer τ such that $P^t(x,y) > 0$ for all $t \geq \tau$ and all $(x,y) \in \mathbb{X}^2$.

2.2.1 Consequences of our Definition of Probability

It is important to realize that we accept the 'addition rule' as an axiom in our mathematical definition of probability (or our probability model) and we do **not** prove this rule. However, the facts which are stated (with proofs) below, are logical consequences of our definition of probability:

1. For any event A, $\boxed{\mathbf{P}(A^c) = 1 - \mathbf{P}(A)}$.

   Proof: One line proof.

   $$\overbrace{\mathbf{P}(A) + \mathbf{P}(A^c)}^{LHS} \underbrace{=}_{\text{+ rule } \because A \cap A^c = \emptyset} \mathbf{P}(A \cup A^c) \underbrace{=}_{A \cup A^c = \Omega} \mathbf{P}(\Omega) \underbrace{=}_{\because \mathbf{P}(\Omega)=1} \overbrace{1}^{RHS} \quad \underbrace{\Longrightarrow}_{LHS-\mathbf{P}(A) \ \& \ RHS-\mathbf{P}(A)} \quad \mathbf{P}(A^c) = 1 - \mathbf{P}(A)$$

   - If $A = \Omega$ then $A^c = \Omega^c = \emptyset$ and $\boxed{\mathbf{P}(\emptyset) = 1 - \mathbf{P}(\Omega) = 1 - 1 = 0}$.

2. For any two events A and B, we have the **inclusion-exclusion principle**:

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

   Proof: Since:

   $$\begin{aligned} A &= (A \setminus B) \cup (A \cap B) \quad \text{and} \quad (A \setminus B) \cap (A \cap B) = \emptyset, \\ A \cup B &= (A \setminus B) \cup B \quad \text{and} \quad (A \setminus B) \cap B = \emptyset \end{aligned}$$

   the addition rule implies that:

   $$\begin{aligned} \mathbf{P}(A) &= \mathbf{P}(A \setminus B) + \mathbf{P}(A \cap B) \\ \mathbf{P}(A \cup B) &= \mathbf{P}(A \setminus B) + \mathbf{P}(B) \end{aligned}$$

   Substituting the first equality above into the second, we get:

   $$\mathbf{P}(A \cup B) = \mathbf{P}(A \setminus B) + \mathbf{P}(B) = \mathbf{P}(A) - \mathbf{P}(A \cap B) + \mathbf{P}(B)$$

3. From inclusion-exclusion principle we get **Boole's inequality**: for any two events A, B

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

4. The inclusion-exclusion principle extends similarly to any three events A_1, A_2, A_3 as follows:

   $$\mathbf{P}(A_1 \cup A_2 \cup A_3) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3) - \mathbf{P}(A_1 \cap A_2) - \mathbf{P}(A_1 \cap A_3) - \mathbf{P}(A_2 \cap A_3) + \mathbf{P}(A_1 \cap A_2 \cap A_3)$$

   and generalises to any n events $A_1, A_2, \ldots, A_n$ as folows:

   $$\mathbf{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbf{P}(A_i) - \sum_{i<j} \mathbf{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbf{P}(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n-1} \sum_{i<\cdots<n} \mathbf{P}\left(\bigcap_{i=1}^n A_i\right)$$

   Proof: See the counting argument in `https://en.wikipedia.org/wiki/Inclusion%E2%80%93exclusion_principle` if you are curious.

5. Once again by the inclusion-exclusion principle, the Boole's inequality generalises to any n events $A_1, A_2, \ldots, A_n$ as folows:

   $$\mathbf{P}\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathbf{P}(A_i)$$

6. For a sequence of mutually disjoint events $A_1, A_2, A_3, \ldots, A_n$:

$$\boxed{A_i \cap A_j = \emptyset \quad \text{for any } i \neq j \quad \Longrightarrow \quad \mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots + \mathbf{P}(A_n)}.$$

Proof: If A_1, A_2, A_3 are mutually disjoint events, then $A_1 \cup A_2$ is disjoint from A_3. Thus, two applications of the addition rule for disjoint events yields:

$$\mathbf{P}(A_1 \cup A_2 \cup A_3) = \mathbf{P}((A_1 \cup A_2) \cup A_3) \underbrace{=}_{+\text{ rule}} \mathbf{P}(A_1 \cup A_2) + \mathbf{P}(A_3) \underbrace{=}_{+\text{ rule}} \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3)$$

The n-event case follows by mathematical induction.

We have formally defined the **probability model** specified by the **probability triple** $(\Omega, \mathcal{F}, \mathbf{P})$ that can be used to model an **experiment** $\mathcal{E}$.

Example 28 (First Ball out of NZ Lotto) Let us observe the number on *the first ball that pops out in a New Zealand Lotto trial*. There are forty balls labelled 1 through 40 for this experiment and so the sample space is

$$\Omega \;=\; \{1, 2, 3, \ldots, 39, 40\}\,.$$

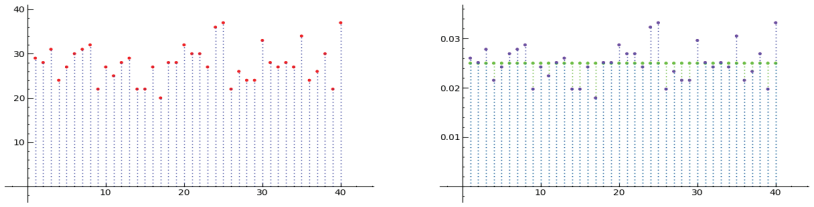
Because the balls are vigorously whirled around inside the Lotto machine, modelled as a well-stirrred urn, before the first one pops out, we can model each ball to pop out first with the same probability. So, we assign each outcome $\omega \in \Omega$ the same probability of $\frac{1}{40}$, i.e., our probability model for this experiment is:

$$\mathbf{P}(\omega) = \frac{1}{40}, \text{ for each } \; \omega \in \Omega = \{1, 2, 3, \ldots, 39, 40\} \;.$$

Note: We sometimes abuse notation and write $\mathbf{P}(\omega)$ instead of the more accurate but cumbersome $\mathbf{P}(\{\omega\})$ when writing down probabilities of simple events.

Crucially, by $\omega = 17$ for example, we mean all the detailed dynamics inside the Lotto machine that lead to the event that the ball labelled by the number 17 ends up popping out. So, Ω here is indeed a more complicated set although it only leads to 40 possible outcomes.

Figure 2.2 (a) shows the frequency of the first ball number in 1114 NZ Lotto draws. Figure 2.2 (b) shows the relative frequency, i.e., the frequency divided by 1114, the number of draws. Figure 2.2 (b) also shows the equal probabilities under our model.



(a) Frequency of first ball. (b) Relative frequency and probability of first ball.

Figure 2.2: First ball number in 1114 NZ Lotto draws from 1987 to 2008.

Next, let us take a detour into how one might interpret it in the real world. The following is an adaptation from Williams D, *Weighing the Odds: A Course in Probability and Statistics*, Cambridge University Press, 2001, which henceforth is abbreviated as WD2001.


```
>> mu1 = mu0 * Phot % distribution for tomorrow since today is hot
mu1 =    0.9500    0.0500
>> mu2 = mu1 * Pcold % distribution for day after tomorrow since tomorrow is supposed to be cold
mu2 =    0.6400    0.3600
>> mu2 = mu0 * Phot * Pcold % we can also get the distribution for day after tomorrow directly
mu2 =    0.6400    0.3600
```

Exercise 176 For the Markov chain in Example 175 compute the probability that the day after tomorrow is wet if today is dry and hot but tomorrow is supposed to be cold.

6.3 Irreducibility and Aperiodicity

The utility of our mathematical constructions with Markov chains depends on a delicate balance between generality and specificity. We introduce two specific conditions called irreducibility and aperiodicity that make Markov chains more useful to model real-word phenomena.

Definition 81 (Communication between states) Let $(X_t)_{t\in\mathbb{Z}_+}$ be a homogeneous Markov chain with transition matrix P on state space $\mathbb{X} := \{s_1, s_2, \ldots, s_k\}$. We say that a state s_i **communicates** with a state s_j and write $s_i \to s_j$ or $s_j \leftarrow s_i$ if there exists an $\eta(s_i, s_j) \in \mathbb{N}$ such that:

$$\mathbf{P}\left(X_{t+\eta(s_i,s_j)} = s_j | X_t = s_i\right) = P^{\eta(s_i,s_j)}(s_i, s_j) > 0 \;.$$

In words, s_i communicates with s_j if you can eventually reach s_j from s_i. If $P^\eta(s_i, s_j) = 0$ for every $\eta \in \mathbb{N}$ then we say that s_i **does not communicate** with s_j and write $s_i \nrightarrow s_j$ or $s_j \nleftarrow s_i$.

We say that two states s_i and s_j **intercommunicate** and write $s_i \leftrightarrow s_j$ if $s_i \to s_j$ and $s_j \to s_i$. In words, two states intercommunicate if you can eventually reach one from another and vice versa. When s_i and s_j do not intercommunicate we write $s_i \nleftrightarrow s_j$.

Definition 82 (Irreducible) A homogeneous Markov chain $(X_t)_{t\in\mathbb{Z}_+}$ with transition matrix P on state space $\mathbb{X} := \{s_1, s_2, \ldots, s_k\}$ is said to be **irreducible** if $s_i \leftrightarrow s_j$ for each $(s_i, s_j) \in \mathbb{X}^2$. Otherwise the chain is said to be **reducible**.

We have already seen examples of reducible and irreducible Markov chains. For example, Flippant Freddy's family of Markov chains with the (p, q)-parametric family of transition matrices, $\{P_{(p,q)} : (p, q) \in [0, 1]^2\}$, where each $P_{(p,q)}$ is given by Equation 6.3. If $(p, q) \in (0, 1)^2$, then the corresponding Markov chain is irreducible because we can go from rollopia to flippopia or vice versa in just one step with a positive probability. Thus, the Markov chains with transition matrices in $\{P_{(p,q)} : (p, q) \in (0, 1)^2\}$ are irreducible. But if p or q take probability values at the boundary of $[0, 1]$, i.e., $p \in \{0, 1\}$ or $q \in \{0, 1\}$ then we have to be more careful because we may never get from at least one state to the other and the corresponding Markov chains may be reducible. For instance, if $p = 0$ or $q = 0$ then we will be stuck in either rollopia or flippopia, respectively. However, if $p = 1$ and $q \neq 0$ or $q = 1$ and $p \neq 0$ then we can get from each state to the other. Therefore, only the transition matrices in $\left\{P_{(p,q)} : p \in \{0\} \text{ or } q \in \{0\}\right\}$ are reducible.

The simplest way to verify whether a Markov chain is irreducible is by looking at its transition diagram (without the positive edge probabilities) and checking that from each state there is a sequence of arrows leading to any other state.

Exercise 177 Revisit all the Markov chains we have considered up to now and determine whether they are reducible or irreducible by checking that from each state there is a sequence of arrows leading to any other state in their transition graphs.

Definition 79 (Inhomogeneous finite Markov chain) Let $P_1, P_2, \ldots$ be a sequence of $k \times k$ stochastic matrices satisfying the conditions in Equation 6.2. Then, the stochastic sequence $(X_t)_{t\in\boldsymbol{Z}_+} := (X_0, X_1, \ldots)$ with finite state space $\mathbb{X} := \{s_1, s_2, \ldots, s_k\}$ is called an inhomogeneous Markov chain with transition matrices $P_1, P_2, \ldots$, if for all pairs of states $(x, y) \in \mathbb{X}\times\mathbb{X}$, all integers $t \geq 1$, and all probable historical events $H_{t-1} := \bigcap_{n=0}^{t-1}\{X_n = x_n\}$ with $\mathbf{P}\left(H_{t-1} \cap \{X_t = x\}\right) > 0$, the following **Markov property** is satisfied:

$$\mathbf{P}\left(X_{t+1} = y | H_{t-1} \cap \{X_t = x\}\right) = \mathbf{P}\left(X_{t+1} = y | X_t = x\right) =: P_{t+1}(x, y) \ . \tag{6.8}$$

Proposition 80 For a finite inhomogeneous Markov chain $(X_t)_{t\in\mathbb{Z}_+}$ with state space $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$, initial distribution

$$\mu_0 := \left(\mu_0(s_1), \mu_0(s_2), \ldots, \mu_0(s_k)\right),$$

where $\mu_0(s_i) = \mathbf{P}(X_0 = s_i)$, and transition matrices

$$(P_1, P_2, \ldots), \quad P_t := \left(P_t(s_i, s_j)\right)_{(s_i,s_j)\in\mathbb{X}\times\mathbb{X}}, \ t \in \{1, 2, \ldots\}$$

we have for any $t \in \mathbb{Z}_+$ that the distribution at time t given by:

$$\mu_t := \left(\mu_t(s_1), \mu_t(s_2), \ldots, \mu_t(s_k)\right),$$

where $\mu_t(s_i) = \mathbf{P}(X_t = s_i)$, satisfies:

$$\mu_t = \mu_0 P_1 P_2 \cdots P_t \ . \tag{6.9}$$

Proof: Left as Exercise 174.

Exercise 174 Prove Proposition 80 using induction as done for Proposition 78.

Example 175 (a more sophisticated dry-wet chain) Let us make a more sophisticated version of the dry-wet chain of Example 170 with state space $\{d, w\}$. In order to take some seasonality into account in our weather model for dry and wet days in Christchurch, let us have two transition matrices for hot and cold days:

$$P_{\text{hot}} = \begin{array}{c} \\ d \\ w \end{array}\!\!\begin{array}{c} \begin{array}{cc} d & w \end{array} \\ \begin{pmatrix} 0.95 & 0.05 \\ 0.75 & 0.25 \end{pmatrix} \end{array}, \qquad P_{\text{cold}} = \begin{array}{c} \\ d \\ w \end{array}\!\!\begin{array}{c} \begin{array}{cc} d & w \end{array} \\ \begin{pmatrix} 0.65 & 0.35 \\ 0.45 & 0.55 \end{pmatrix} \end{array} \ .$$

We say that a day is hot if its maximum temperature is more than 20° Celsius, otherwise it is cold. We use the transition matrix for today to obtain the state probabilities for tomorrow. If today is dry and hot and tomorrow is supposed to be cold then what is the probability that the day after tomorrow will be wet? We can use (6.9) to obtain the answer as 0.36:

```
>> Phot = [0.95 0.05; 0.75 0.25] % Transition Probability Matrix for hot day
Phot =
    0.9500    0.0500
    0.7500    0.2500
>> Pcold = [0.65 0.35; 0.45 0.55] % Transition Probability Matrix for cold day
Pcold =
    0.6500    0.3500
    0.4500    0.5500
>> mu0 = [1 0] % today is dry
mu0 =     1     0
```


Probability Model	**Real-world Interpretation**
Sample space Ω	Set of all outcomes of an experiment
Sample point ω	Possible outcome of an experiment
(No counterpart)	Actual outcome $\omega^\star$ of an experiment
Event A, a (suitable) subset of Ω	The real-world event corresponding to A occurs if and only if $\omega^\star \in A$
$\mathbf{P}(A)$, a number between 0 and 1	Probability that A will occur for an experiment yet to be performed

Events in Probability Model	**Real-world Interpretation**
Sample space Ω	The certain even 'something happens'
The $\emptyset$ of Ω	The impossible event 'nothing happens'
The intersection $A \cap B$	'Both A and B occur'
$A_1 \cap A_2 \cap \cdots \cap A_n$	'All of the events $A_1, A_2, \ldots, A_n$ occur simultaneously'
The union $A \cup B$	'At least one of A and B occurs'
$A_1 \cup A_2 \cup \cdots \cup A_n$	'At least one of the events $A_1, A_2, \ldots, A_n$ occurs'
A^c, the complement of A	'A does not occur'
$A \setminus B$	'A occurs, but B does not occur'
$A \subset B$	'If A occurs, then B must occur'

In the probability model of Example 28, show that for any event $E \subset \Omega$,

$$\mathbf{P}(E) \ = \ \frac{1}{40} \ \times \ \text{number of elements in } E \ .$$

Let $E = \{\omega_1, \omega_2, \ldots, \omega_k\}$ be an event with k outcomes (simple events). Then by the addition rule for mutually exclusive events we get:

$$\mathbf{P}(E) \ = \ \mathbf{P}\left(\{\omega_1, \omega_2, \ldots, \omega_k\}\right) = \mathbf{P}\left(\bigcup_{i=1}^{k}\{\omega_i\}\right) \ = \ \sum_{i=1}^{k}\mathbf{P}\left(\{\omega_i\}\right) \ = \ \sum_{i=1}^{k}\frac{1}{40} \ = \ \frac{k}{40} \ .$$

2.2.2 Sigma Algebras of Typical Experiments*

Example 29 ('Toss a fair coin once') Consider the 'Toss a fair coin once' experiment. What is its sample space Ω and a reasonable collection of events $\mathcal{F}$ that underpin this experiment?

$$\Omega = \{\mathtt{H}, \mathtt{T}\}, \qquad \mathcal{F} = \{\mathtt{H}, \mathtt{T}, \Omega, \emptyset\} \ ,$$

A function that will satisfy the definition of probability for this collection of events $\mathcal{F}$ and assign $\mathbf{P}(\mathtt{H}) = \frac{1}{2}$ is summarized below. First check that the above $\mathcal{F}$ is a sigma-algebra. Draw a picture for $\mathbf{P}$ with arrows that map elements in the domain $\mathcal{F}$ given above to elements in its range.

Event $A \in \mathcal{F}$	$\mathbf{P} : \mathcal{F} \to [0, 1]$	$\mathbf{P}(A) \in [0, 1]$
$\Omega = \{\mathtt{H}, \mathtt{T}\}\bullet$	$\longrightarrow$	1
$\mathtt{T}\bullet$	$\longrightarrow$	$1 - \frac{1}{2}$
$\mathtt{H}\bullet$	$\longrightarrow$	$\frac{1}{2}$
$\emptyset\bullet$	$\longrightarrow$	0

Classwork 30 (The trivial sigma algebra) Note that $\mathcal{F}' = \{\Omega, \emptyset\}$ is also a sigma algebra of the sample space $\Omega = \{\mathtt{H}, \mathtt{T}\}$. Can you think of a probability for the collection $\mathcal{F}'$?

Event $A \in \mathcal{F}'$	$\mathbf{P} : \mathcal{F}' \to [0,1]$	$\mathbf{P}(A) \in [0,1]$
$\Omega = \{\mathtt{H}, \mathtt{T}\}$ •	$\longrightarrow$	
$\emptyset$ •	$\longrightarrow$	

Thus, $\mathcal{F}$ and $\mathcal{F}'$ are two distinct sigma algebras over our $\Omega = \{\mathtt{H}, \mathtt{T}\}$. Moreover, $\mathcal{F}' \subset \mathcal{F}$ and is called a sub sigma algebra. Try to show that $\{\Omega, \emptyset\}$ is the smallest possible sigma algebra over all possible sigma algebras over any given sample space Ω (think of intersecting an arbitrary family of sigma algebras)?

Generally one encounters four types of sigma algebras (you will understand the last two types after taking more advanced courses in mathematics, so it is fine to understand the ideas intuitively for now!) and they are:

1. When the sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ is a finite set with k outcomes and $\mathbf{P}(\omega_i)$, the probability for each outcome $\omega_i \in \Omega$ is known, then one typically takes the sigma-algebra $\mathcal{F}$ to be the set of all subsets of Ω called the **power set** and denoted by 2^Ω. The probability of each event $A \in 2^\Omega$ can be obtained by adding the probabilities of the outcomes in A, i.e., $\mathbf{P}(A) = \sum_{\omega_i \in A} \mathbf{P}(\omega_i)$. Clearly, 2^Ω is indeed a sigma-algebra and it contains $2^{\#\Omega}$ events in it.

2. When the sample space $\Omega = \{\omega_1, \omega_2, \ldots\}$ is a countable set then one typically takes the sigma-algebra $\mathcal{F}$ to be the set of all subsets of Ω. Note that this is very similar to the case with finite Ω except now $\mathcal{F} = 2^\Omega$ could have uncountably many events in it.

3. If $\Omega = \mathbb{R}^d$ for finite $d \in \{1, 2, 3, \ldots\}$ then the **Borel sigma-algebra** is the smallest sigma-algebra containing all **half-spaces**, i.e., sets of the form

   $$\{x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d : x_1 \le c_1, x_2 \le c_2, \ldots, x_d \le c_d\}, \quad \text{for any } c = (c_1, c_2, \ldots, c_d) \in \mathbb{R}^d \ ,$$

   When $d = 1$ the half-spaces are the half-lines $\{(-\infty, c] : c \in \mathbb{R}\}$ and when $d = 2$ the half-spaces are the south-west quadrants $\{(-\infty, c_1] \times (-\infty, c_2] : (c_1, c_2) \in \mathbb{R}^2\}$, etc. (Equivalently, the Borel sigma-algebra is the smallest sigma-algebra containing all open sets in $\mathbb{R}^d$).

4. Given a finite set $\mathbb{S} = \{s_1, s_2, \ldots, s_k\}$, let Ω be the sequence space $\mathbb{S}^\infty := \mathbb{S} \times \mathbb{S} \times \mathbb{S} \times \cdots$, i.e., the set of sequences of infinite length that are made up of elements from $\mathbb{S}$. A set of the form

   $$A_1 \times A_2 \times \cdots \times A_n \times \mathbb{S} \times \mathbb{S} \times \cdots \ , \quad A_k \subset \mathbb{S} \text{ for all } k \in \{1, 2, \ldots, n\} \ ,$$

   is called a **cylinder set**. The set of events in $\mathbb{S}^\infty$ is the smallest sigma-algebra containing the cylinder sets.

   - **A most primitive sigma-algebra for probability theory:** For example if $\mathbb{S} = \{0, 1\}$, then $\Omega = \{0, 1\}^\infty$ is the set of all infinite sequences made of 0's and 1's. To take advantage of arithmetic and analysis, Ω can be seen as the binary representation of all real numbers in the unit interval $[0, 1]$. We can take advantage of combinatorics and algebra if we further represent the dyadic partition of $[0, 1]$ by a binary tree (as drawn in lectures). Then, a cylinder set such as $1 \times 1 \times 0 \times \{0, 1\} \times \{0, 1\} \times \cdots$, an event here, can be interpreted as the finite binary sequence $(1, 1, 0)$ — corresponding to the third leaf of a finite binary tree with four leaves obtained by splitting the right-most leaf twice. This cylindrical event $(1, 1, 0)$ contains all real numbers in the interval $[\frac{3}{4}, \frac{7}{8}] \subset [0, 1] =: \Omega$.

Exercise 2.1 (Intuiting a most primitive sigma-algebra – this is optional) Try to carefully recollect and understand the most primitive sigma-algebra in the last item above as it was explained in lectures.

Suppose you sell \$100 of lemonade at a road-side stand on a hot day but only \$50 on a cold day. Then we can compute your expected sales tomorrow if today is dry as follows:

```
>> P=[0.75 0.25; 0.5 0.5] % Transition Probability Matrix
P =
    0.7500    0.2500
    0.5000    0.5000
>> f = [100; 50] % sales of lemonade in dollars on a dry and wet day
f =
   100
    50
>> P*f % expected sales tomorrow
ans =
   87.5000
   75.0000
>> mu0 = [1 0] % today is dry
mu0 =     1     0
>> mu0*P*f % expected sales tomorrow if today is dry
ans =   87.5000
```

Exercise 171 (Freddy discovers a gold coin) Flippant Freddy of Example 169 found a gold coin at the bottom of the pond. Since this discovery he jumps around differently in the enchanted pond. He can be found now in one of three states: flipopia, rollopia and hydropia (when he dives into the pond). His state space is $\mathbb{X} = \{r, f, h\}$ now and his transition mechanism is as follows: If he rolls an odd number with his fair die in rollopia he will jump to flipopia but if he rolls an even number then he will stay in rollopia only if the outcome is 2 otherwise he will dive into hydropia. If the fair gold coin toss at the bottom of hydropia is Heads then Freddy will swim to flipopia otherwise he will remain in hydropia. Finally, if he is in flipopia he will remain there if the silver coin lands Heads otherwise he will jump to rollopia.

Make a Markov chain model of the new jumping mechanism adopted by Freddy. Draw the transition diagram, produce the transition matrix P and compute using MATLAB the probability that Freddy will be in hydropia after one, two, three, four and five jumps given that he starts in hydropia.

Exercise 172 Let $(X_t)_{t \in \mathbb{Z}_+}$ be a Markov chain with state space $\{a, b, c\}$, initial distribution $\mu_0 = (1/3, 1/3, 1/3)$ and transition matrix

$$P = \begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \end{array} .$$

For each t, define $Y_t = \mathbb{1}_{\{b,c\}}(X_t)$. Show that $(Y_t)_{t \in \mathbb{Z}_+}$ is not a Markov chain.

Exercise 173 Let $(X_t)_{t \in \mathbb{Z}_+}$ be a (homogeneous) Markov chain on $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$ with transition matrix P and initial distribution μ_0. For a given $m \in \mathbb{N}$, let $(Y_t)_{t \in \mathbb{Z}_+}$ be a stochastic sequence with $Y_t = X_{mt}$. Show that $(Y_t)_{t \in \mathbb{Z}_+}$ is a Markov chain with transition matrix P^m. This establishes that Markov chains that are sampled at regular time steps are also Markov chains.

Until now our Markov chains have been **homogeneous** in time according to Definition 77, i.e., the transition matrix P does not change with time. We define inhomogeneous Markov chains that allow their transition matrices to possibly change with time. Such Markov chains are more realistic as models in some situations and more flexible as algorithms in the sequel.

Thus, multiplying a row vector μ_0 by P^t on the right takes you from current distribution over the state space to the distribution in t steps of the chain.

Since we will be interested in Markov chains on $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$ with the same transition matrix P but different initial distributions, we introduce $\mathbf{P}_\mu$ and $\mathbf{E}_\mu$ for probabilities and expectations given that the initial distribution is μ, respectively. When the initial distribution is concentrated at a single initial state x given by:

$$\mathbb{1}_{\{x\}}(y) := \begin{cases} 1 & \text{if } y = x \\ 0 & \text{if } y \neq x \end{cases}$$

we represent it by e_x, the $1 \times k$ ortho-normal basis row vector with a 1 in the x-th entry and a 0 elsewhere. We simply write $\mathbf{P}_x$ for $\mathbf{P}_{\mathbb{1}_{\{x\}}}$ or $\mathbf{P}_{e_x}$ and $\mathbf{E}_x$ for $\mathbf{E}_{\mathbb{1}_{\{x\}}}$ or $\mathbf{E}_{e_x}$. Thus, Proposition 78 along with our new notations means that:

$$\mathbf{P}_x(X_t = y) = (e_x P^t)(y) = P^t(x, y) \ .$$

In words, the probability of going to y from x in t steps is given by the (x, y)-th entry of P^t, the **t-step transition matrix**. We refer to the x-th row and the x-th column of P by $P(x, \cdot)$ and $P(\cdot, x)$, respectively.

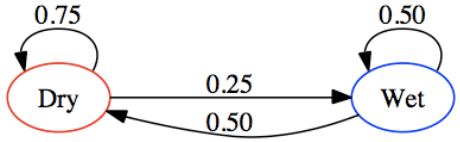
Let the function $f(x) : \mathbb{X} \to \mathbb{R}$ be represented by the column vector $f := (f(s_1), f(s_2), \ldots, f(s_k)) \in \mathbb{R}^{k \times 1}$. Then the x-th entry of $P^t f$ is:

$$P^t f(x) = \sum_y P^t(x, y) f(y) = \sum_y f(y) \mathbf{P}_x(X_t = y) = \mathbf{E}_x(f(X_t)) \ .$$

This is the expected value of f under the distribution of states in t steps given that we start at state x. Thus multiplying a column vector f by P^t from the left takes you from a function on the state space to its expected value in t steps of the chain.

Example 170 (Dry-Wet Christchurch Weather) Consider a toy weather model for dry or wet days in Christchurch using a Markov chain with state space $\{d, w\}$. Let the transition diagram in Figure 6.3 give the transition matrix P for our dry-wet Markov chain. Using (6.7) we can find

Figure 6.3: Transition Diagram of Dry and Wet Days in Christchurch.



that the probability of being dry on the day after tomorrow is 0.625 given that it is wet today as follows:

```
>> P=[0.75 0.25; 0.5 0.5] % Transition Probability Matrix
P =
    0.7500    0.2500
    0.5000    0.5000
>> mu0=[0 1] % it is wet today gives the initial distribution
mu0 =     0     1
>> mu0 * P^2 % the distribution in 2 days from today
ans =    0.6250    0.3750
```


PROBABILITY SUMMARY

Axioms:

1. If $A \subseteq \Omega$ then $0 \leq \mathbf{P}(A) \leq 1$ and $\mathbf{P}(\Omega) = 1$.
2. If A, B are disjoint events, then $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$.
   [This is true only when A and B are disjoint.]
3. If $A_1, A_2, \ldots$ are disjoint then $\mathbf{P}(A_1 \cup A_2 \cup \ldots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \ldots$

Rules:

$$\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$$

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B) \qquad \text{[always true]}$$

2.3 Exercises in Probability

Ex. 2.2 — In English language text, the twenty six letters in the alphabet occur with the following frequencies:

E	13%	R	7.7%	A	7.3%	H	3.5%	F	2.8%	M	2.5%	W	1.6%	X	0.5%	J	0.2%		
T	9.3%	O	7.4%	S	6.3%	L	3.5%	P	2.7%	Y	1.9%	V	1.3%	K	0.3%	Z	0.1%		
N	7.8%	I	7.4%	D	4.4%	C	3%	U	2.7%	G	1.6%	B	0.9%	Q	0.3%				

Suppose you pick one letter at random from a randomly chosen English book from our central library with $\Omega = \{\mathsf{A}, \mathsf{B}, \mathsf{C}, \ldots, \mathsf{Z}\}$ (ignoring upper/lower cases), then what is the probability of these events?

(a) $\mathbf{P}(\{\mathsf{Z}\})$
(b) $\mathbf{P}$('picking any letter')
(c) $\mathbf{P}(\{\mathsf{E}, \mathsf{Z}\})$
(d) $\mathbf{P}$('picking a vowel')
(e) $\mathbf{P}$('picking any letter in the word WAZZZUP')
(f) $\mathbf{P}$('picking any letter in the word WAZZZUP or a vowel').

Ex. 2.3 — Find the sample spaces for the following experiments:

1. Tossing 2 coins whose faces are sprayed with black paint denoted by B and white paint denoted by W.
2. Drawing 4 screws from a bucket of left-handed and right-handed screws denoted by L and R, respectively.
3. Rolling a die and recording the number on the upturned face until the first 6 appears.

Ex. 2.4 — Suppose we pick a letter at random from the word WAIMAKARIRI.

1. What is the sample space Ω?
2. What probabilities should be assigned to the outcomes?
3. What is the probability of *not* choosing the letter R?

Ex. 2.5 — There are seventy five balls in total inside the Bingo Machine. Each ball is labelled by one of the following five letters: B, I, N, G, and O. There are fifteen balls labelled by each letter. The letter on the first ball that comes out of a BINGO machine after it has been well-mixed is the outcome of our experiment.

(a) Write down the sample space of this experiment.

(b) Find the probabilities of each simple event.

(c) Show that $\mathbf{P}(\Omega)$ is indeed 1.

(d) Check that the addition rule for mutually exclusive events holds for the simple events $\{B\}$ and $\{I\}$.

(e) Consider the following events: $C = \{\mathsf{B}, \mathsf{I}, \mathsf{G}\}$ and $D = \{\mathsf{G}, \mathsf{I}, \mathsf{N}\}$. Using the addition rule for two arbitrary events, find $\mathbf{P}(C \cup D)$.

2.4 Conditional Probability

Conditional probabilities arise when we have partial information about the result of an experiment which restricts the sample space to a range of outcomes. For example, if there has been a lot of recent seismic activity in Christchurch, then the probability that an already damaged building will collapse tomorrow is clearly higher than if there had been no recent seismic activity.

Conditional probabilities are often expressed in English by phrases such as:

"If A happens, what is the probability that B happens?"

or

"What is the probability that A happens if B happens?"

or

" What is the probability that A occurs given that B occurs?"

Next, we define conditional probability and the notion of independence of events. We use the LTRF idea to motivate the definition.

Idea 11 (LTRF intuition for conditional probability) Let A and B be any two events associated with our experiment $\mathcal{E}$ with $\mathbf{P}(A) \neq 0$. The 'conditional probability that B occurs given that A occurs' denoted by $\mathbf{P}(B|A)$ is again intuitively underpinned by the super-experiment $\mathcal{E}^\infty$ which is the 'independent' repitition of our original experiment $\mathcal{E}$ 'infinitely' often. The LTRF idea is that $\mathbf{P}(B|A)$ is the long-term proportion of those experiments on which A occurs that B also occurs.

Recall that $N(A, n)$ as defined in (2.1) is the fraction of times A occurs out of n independent repetitions of our experiment $\mathcal{E}$ (ie. the experiment $\mathcal{E}^n$). If $A \cap B$ is the event that 'A and B occur simultaneously', then we intuitively want

$$\mathbf{P}(B|A) \quad \text{" } \to \text{ "} \quad \frac{N(A \cap B, n)}{N(A, n)} = \frac{N(A \cap B, n)/n}{N(A, n)/n} = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)}$$

as our $\mathcal{E}^n \to \mathcal{E}^\infty$. So, we **define** conditional probability as we want.

Now let us generalise the lessons learned from Example 169.

Proposition 78 For a finite Markov chain $(X_t)_{t \in \mathbb{Z}_+}$ with state space $\mathbb{X} = \{s_1, s_2, \ldots, s_k\}$, initial distribution

$$\mu_0 := (\mu_0(s_1), \mu_0(s_2), \ldots, \mu_0(s_k)) \,,$$

where $\mu_0(s_i) = \mathbf{P}(X_0 = s_i)$, and transition matrix

$$P := (P(s_i, s_j))_{(s_i, s_j) \in \mathbb{X}^2} \,,$$

we have for any $t \in \mathbb{Z}_+$ that the distribution at time t given by:

$$\mu_t := (\mu_t(s_1), \mu_t(s_2), \ldots, \mu_t(s_k)) \,,$$

where $\mu_t(s_i) = \mathbf{P}(X_t = s_i)$, satisfies:

$$\mu_t = \mu_0 P^t \ . \tag{6.7}$$

Proof: We will prove this by induction on $\boldsymbol{Z}_+ := \{0, 1, 2, \ldots\}$. First consider the case when $t = 0$. Since P^0 is the identity matrix I, we get the desired equality:

$$\mu_0 P^0 = \mu_0 I = \mu_0 \ .$$

Next consider the case when $t = 1$. We get for each $j \in \{1, 2, \ldots, k\}$, that

$$\begin{aligned}
\mu_1(s_j) = \mathbf{P}(X_1 = s_j) &= \sum_{i=1}^{k} \mathbf{P}(X_1 = s_j, X_0 = s_i) \\
&= \sum_{i=1}^{k} \mathbf{P}(X_1 = s_j | X_0 = s_i) \mathbf{P}(X_0 = s_i) \\
&= \sum_{i=1}^{k} P(s_i, s_j) \mu_0(s_i) \\
&= (\mu_0 P)(s_j), \quad \text{the } j\text{-th entry of the row vector } (\mu_0 P) \ .
\end{aligned}$$

Hence, $\mu_1 = \mu_0 P$. Now, we will fix m and suppose that (6.7) holds for $t = m$ and prove that (6.7) also holds for $t = m + 1$. For each $j \in \{1, 2, \ldots, k\}$, we get

$$\begin{aligned}
\mu_{m+1}(s_j) = \mathbf{P}(X_{m+1} = s_j) &= \sum_{i=1}^{k} \mathbf{P}(X_{m+1} = s_j, X_m = s_i) \\
&= \sum_{i=1}^{k} \mathbf{P}(X_{m+1} = s_j | X_m = s_i) \mathbf{P}(X_m = s_i) \\
&= \sum_{i=1}^{k} P(s_i, s_j) \mu_m(s_i) \\
&= (\mu_m P)(s_j), \quad \text{the } j\text{-th entry of the row vector } (\mu_m P) \ .
\end{aligned}$$

Hence, $\mu_{m+1} = \mu_m P$. But $\mu_m = \mu_0 P^m$ by the induction hypothesis, and therefore:

$$\mu_{m+1} = \mu_m P = \mu_0 P^m P = \mu_0 P^{m+1} \ .$$

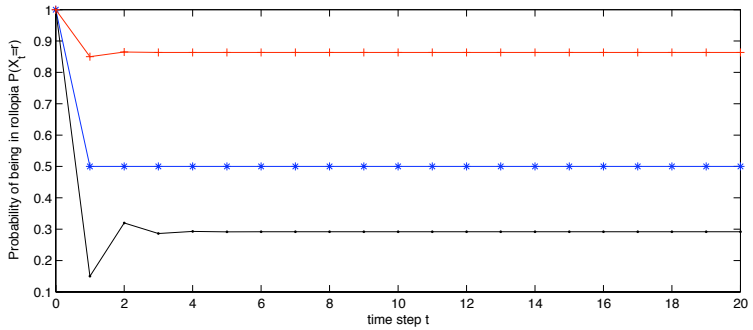
Thus by the principle of mathematical induction we have proved the proposition.

Now, let us compute and look at the probability of being in rollopia after having started there for three values of p and q according to the following script:

FlippantFreddyRollopiaProbs.m

```
p=0.5; q=0.5; P = [1-p p; q 1-q]; % assume a fair coin and a fair die
mu0 = [1, 0]; % inital state vector since Freddy started in rollopia
for t = 1: 1: 21, mut(t,:) = mu0*P^(t-1); end
t=0:1:20; % vector of time steps t
plot(t,mut(:,1)','b*-')
hold on;
p=0.85; q=0.35; P = [1-p p; q 1-q]; % assume an unfair coin and an unfair die
for t = 1: 1: 21, mut(t,:) = mu0*P^(t-1); end
t=0:1:20; % vector of time steps t
plot(t,mut(:,1)','k.-')
p=0.15; q=0.95; P = [1-p p; q 1-q]; % assume another unfair coin and another unfair die
for t = 1: 1: 21, mut(t,:) = mu0*P^(t-1); end
t=0:1:20; % vector of time steps t
plot(t,mut(:,1)','r+-')
xlabel('time step t'); ylabel('Probability of being in rollopia $P(X_t=r)$')
xlabel('time step t'); ylabel('Probability of being in rollopia P(X_t=r)')
```

Figure 6.2: The probability of being back in rollopia in t time steps after having started there under transition matrix P with (i) $p = q = 0.5$ (blue line with asterisks), (ii) $p = 0.85$, $q = 0.35$ (black line with dots) and (iii) $p = 0.15$, $q = 0.95$ (red line with pluses).



It is evident from Figure 6.2 that as $t \to \infty$, μ_t approaches a distribution, say π, that depends on p and q in P. Such a limit distribution is called the **stationary distribution** and must satisfy the fixed point condition:

$$\pi P = \pi \ ,$$

that gives the solution:

$$\pi(r) = \frac{q}{p+q}, \qquad \pi(f) = \frac{p}{p+q} \ .$$

In Figure 6.2 we see that $\mathbf{P}(X_t = r)$ approaches $\pi(r) = \frac{q}{p+q}$ for the three cases of p and q:

$$\text{(i) } p = 0.50, q = 0.50, \qquad \mathbf{P}(X_t = r) \to \pi(r) = \frac{q}{p+q} = \frac{0.50}{0.50+0.50} = 0.5000,$$

$$\text{(ii) } p = 0.85, q = 0.35, \qquad \mathbf{P}(X_t = r) \to \pi(r) = \frac{q}{p+q} = \frac{0.35}{0.85+0.35} = 0.2917,$$

$$\text{(iii) } p = 0.15, q = 0.95, \qquad \mathbf{P}(X_t = r) \to \pi(r) = \frac{q}{p+q} = \frac{0.95}{0.15+0.95} = 0.8636.$$

Definition 12 (Conditional Probability) Suppose we are given an experiment $\mathcal{E}$ with a triple $(\Omega, \mathcal{F}, \mathbf{P})$. Let A and B be events, ie. $A, B \in \mathcal{F}$, such that $\mathbf{P}(A) \neq 0$. Then, we define the **conditional probability** of B given A by,

$$\mathbf{P}(B|A) := \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} \ . \tag{2.2}$$

Note that A serves as the new reduced sample space so that conditional probabilities given A *are* indeed probabilities. Thus, for a **fixed** event $A \in \mathcal{F}$ with $\mathbf{P}(A) > 0$ and **any** event $B \in \mathcal{F}$, the conditional probability $\mathbf{P}(B|A)$ is a probability as in Definition 10, ie. a function:

$$\mathbf{P}(B|A) : \mathcal{F} \to [0, 1]$$

that assigns to each $B \in \mathcal{F}$ a number in the interval $[0, 1]$, such that, the axioms of probability are satisfied:

Axiom (1): For any event B, $0 \leq P(B|A) \leq 1$.

Axiom (2): $\mathbf{P}(\Omega|A) = 1$ Meaning 'Something Happens given the event A happens'

Axiom (3): The 'Addition Rule' axiom holds, ie. for events $B_1, B_2 \in \mathcal{F}$,

$$B_1 \cap B_2 = \emptyset \quad \text{implies} \quad \mathbf{P}(B_1 \cup B_2|A) = \mathbf{P}(B_1|A) + \mathbf{P}(B_2|A) \ .$$

Axiom (4): For mutually exclusive events, $B_1, B_2, \ldots$,

$$\mathbf{P}(B_1 \cup B_2 \cup \cdots |A) = \mathbf{P}(B_1|A) + \mathbf{P}(B_2|A) + \cdots \ .$$

From the definition of conditional probability we get the following properties or rules:

Complementation rule: $\mathbf{P}(B|A) = 1 - \mathbf{P}(B^c|A)$.

Addition rule for two arbitrary events B_1 and B_2:

$$\mathbf{P}(B_1 \cup B_2|A) = \mathbf{P}(B_1|A) + \mathbf{P}(B_2|A) - \mathbf{P}(B_1 \cap B_2|A) \, .$$

Solving for $P(A \cap B)$ with these definitions of conditional probability gives another rule:

Multiplication rule for two likely events:
If A and B are events, and if $\mathbf{P}(A) \neq 0$ and $\mathbf{P}(B) \neq 0$, then

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

Example 31 (Wasserman03, p. 11) A medical test for a disease D has outcomes $+$ and $-$. the probabilities are:

	Have Disease (D)	Don't have disease (D^c)
Test positive ($+$)	0.009	0.099
Test negative ($-$)	0.001	0.891

Using the definition of conditional probability, we can compute the conditional probability that you test positive given that you have the disease:

$$\mathbf{P}(+|D) = \frac{\mathbf{P}(+ \cap D)}{\mathbf{P}(D)} = \frac{0.009}{0.009 + 0.001} = 0.9 \ ,$$

and the conditional probability that you test negative given that you don't have the disease:

$$\mathbf{P}(-|D^c) = \frac{\mathbf{P}(- \cap D^c)}{\mathbf{P}(D^c)} = \frac{0.891}{0.099 + 0.891} \approxeq 0.9 \ .$$

Thus, the test is quite accurate since sick people test positive 90% of the time and healthy people test negative 90% of the time.

Now, suppose you go for a test and and test positive. What is the probability that you have the disease ?

$$\mathbf{P}(D|+) = \frac{\mathbf{P}(D \cap +)}{\mathbf{P}(+)} = \frac{0.009}{0.009 + 0.099} \approxeq 0.08$$

Most people who are not used to the definition of conditional probability would intuitively associate a number much bigger than 0.08 for the answer. Interpret conditional probability in terms of the meaning of the numbers that appear in the numerator and denominator of the above calculations.

2.4.1 Bayes' Theorem

Next we look at one of the most elegant applications of the definition of conditional probability along with the addition rule for a partition of Ω called *Bayes' Theorem*. We will present a two event case first called *Bayes' Rule* and then present the more general case of the Theorem.

This is useful because many problems involve reversing the order of conditional probabilities. Suppose we want to investigate some phenomenon A and have an observation B that is evidence about A: for example, A may be breast cancer and B may be a positive mammogram. Then Bayes' Theorem tells us how we should update our probability of A, given the new evidence B.

Or, put more simply, Bayes' Rule is useful when you know $P(B|A)$ but want $P(A|B)$!

Proposition 13 (Bayes' Rule)

$$P(A|B) \ = \ \frac{P(A)P(B|A)}{P(B)} \ . \tag{2.3}$$

Proof: From the definition of conditional probability and the multiplication rule for two likely events A and B we get:

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

Example 32 (Mammogram) Approximately 1% of women aged 40–50 have breast cancer. A woman with breast cancer has a 90% chance of a positive test from a mammogram, while a woman without breast cancer has a 10% chance of a false positive result from the test. What is the probability that a woman indeed has breast cancer given that she just had a positive test?

Definition of conditional probability and the Markov property, we see that,

$$\begin{aligned}
\mathbf{P}(X_2 = f|X_0 = r) &= \mathbf{P}(X_2 = f, X_1 = f|X_0 = r) + \mathbf{P}(X_2 = f, X_1 = r|X_0 = r) \\
&= \frac{\mathbf{P}(X_2 = f, X_1 = f, X_0 = r)}{\mathbf{P}(X_0 = r)} + \frac{\mathbf{P}(X_2 = f, X_1 = r, X_0 = r)}{\mathbf{P}(X_0 = r)} \\
&= \mathbf{P}(X_2 = f|X_1 = f, X_0 = r)\frac{\mathbf{P}(X_1 = f, X_0 = r)}{\mathbf{P}(X_0 = r)} \\
&\quad + \mathbf{P}(X_2 = f|X_1 = r, X_0 = r)\frac{\mathbf{P}(X_1 = r, X_0 = r)}{\mathbf{P}(X_0 = r)} \\
&= \mathbf{P}(X_2 = f|X_1 = f, X_0 = r)\mathbf{P}(X_1 = f|X_0 = r) \\
&\quad + \mathbf{P}(X_2 = f|X_1 = r, X_0 = r)\mathbf{P}(X_1 = r|X_0 = r) \\
&= \mathbf{P}(X_2 = f|X_1 = f)\mathbf{P}(X_1 = f|X_0 = r) \\
&\quad + \mathbf{P}(X_2 = f|X_1 = r)\mathbf{P}(X_1 = r|X_0 = r) \\
&= P(f, f)P(r, f) + P(r, f)P(r, r) \\
&= (1 - q)p + p(1 - p)
\end{aligned} \tag{6.5}$$

Similarly,

$$\mathbf{P}(X_2 = r|X_0 = r) = P(f, r)P(r, f) + P(r, r)P(r, r) = qp + (1 - p)(1 - p) \tag{6.6}$$

Instead of elaborate computations of the probabilities of being in a given state after Freddy's t-th restless moment, we can store the state probabilities at time t in a row vector:

$$\mu_t := (\mathbf{P}(X_t = r|X_0 = r), \mathbf{P}(X_t = f|X_0 = r)) \ \ ,$$

Now, we can conveniently represent Freddy starting in rollopia by the **initial distribution** $\mu_0 = (1, 0)$ and obtain the 1-step **state probability vector** in (6.4) from $\mu_1 = \mu_0 P$ and the 2-step state probabilities in (6.5) and (6.6) by $\mu_2 = \mu_1 P = \mu_0 PP = \mu_0 P^2$. In general, multiplying μ_t, the state probability vector at time t, by the transition matrix P on the right updates the state probabilities by another step:

$$\mu_t = \mu_{t-1} P \qquad \text{for all } t \geq 1 \ .$$

And for any initial distribution μ_0,

$$\mu_t = \mu_0 P^t \qquad \text{for all } t \geq 0 \ .$$

This can be easily implemented in MATLAB as follows:

```
>> p=0.85; q=0.35; P = [1-p p; q 1-q] % assume an unfair coin and an unfair die
P =
    0.1500    0.8500
    0.3500    0.6500
>> mu0 = [1, 0] % inital state vector since Freddy started in rollopia
mu0 =     1     0
>> mu0*P^0    % intial state distribution at t=0 is just mu0
ans =     1     0
>> mu0*P^1    % state distribution at t=1
ans =    0.1500    0.8500
>> mu0*P^2    % state distribution at t=2
ans =    0.3200    0.6800
>> mu0*P^3    % state distribution at t=3
ans =    0.2860    0.7140
```

$|\mathbb{X}| \times |\mathbb{X}|$ matrix P is enough to obtain the state transitions since the x-th row of P is the probability distribution $P(x,\cdot) := (P(x,y))_{y\in\mathbb{X}}$. For this reason P is called a **stochastic matrix**, i.e.,

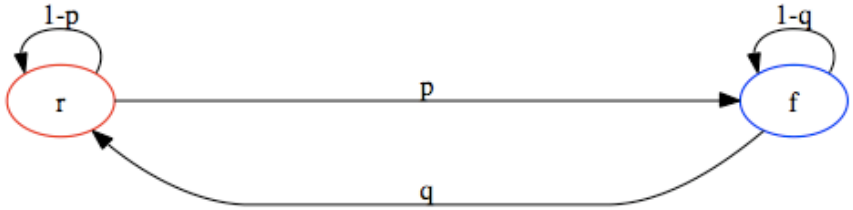
$$P(x,y) \geq 0 \quad \text{for all } (x,y) \in \mathbb{X}^2 \qquad \text{and} \qquad \sum_{y\in\mathbb{X}} P(x,y) = 1 \quad \text{for all } x \in \mathbb{X} \ . \tag{6.2}$$

Thus, for a Markov chain $(X_n)_{n\in\mathbb{Z}_+}$, the distribution of X_{t+1} given $X_0, \ldots, X_t$ depends on X_t alone. Because of this dependence on the previous state, the stochastic sequence, $(X_0, X_1, \ldots)$, are *not* independent. We introduce the most important concepts using a simple example.

Example 169 (Flippant Freddy) Freddy the flippant frog lives in an enchanted pond with only two lily pads, *rollopia* and *flipopia*. A wizard gave a die and a silver coin to help flippant Freddy decide where to jump next. Freddy left the die on rollopia and the coin on flipopia. When Freddy got restless in rollopia he would roll the die and if the die landed odd he would leave the die behind and jump to flipopia, otherwise he would stay put. When Freddy got restless in flipopia he would flip the coin and if it landed Heads he would leave the coin behind and jump to rollopia, otherwise he would stay put.

Let the state space $\mathbb{X} = \{r, f\}$, and let $(X_0, X_1, \ldots)$ be the sequence of lily pads occupied by Freddy after his restless moments. Say the die on rollopia r has probability p of turning up odd and the coin on flipopia f has probability q of turning up heads. We can visualise the rules of Freddy's jumps by the following **transition diagram**:

Figure 6.1: Transition Diagram of Flippant Freddy's Jumps.



Then Freddy's sequence of jumps $(X_0, X_1, \ldots)$ is a Markov chain on $\mathbb{X}$ with transition matrix:

$$P = \begin{array}{c} \\ r \\ f \end{array} \begin{array}{c} \begin{array}{cc} r & f \end{array} \\ \begin{pmatrix} P(r,r) & P(r,f) \\ P(f,r) & P(f,f) \end{pmatrix} \end{array} = \begin{array}{c} \\ r \\ f \end{array} \begin{array}{c} \begin{array}{cc} r & f \end{array} \\ \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \end{array} . \tag{6.3}$$

Suppose we first see Freddy in rollopia, i.e., $X_0 = r$. When he gets restless for the first time we know from the first row of P that he will leave to flippopia with probability p and stay with probability $1-p$, i.e.,

$$\mathbf{P}(X_1 = f | X_0 = r) = p, \quad \mathbf{P}(X_1 = r | X_0 = r) = 1 - p \ . \tag{6.4}$$

What happens when he is restless for the second time? By considering the two possibilities for X_1,

Solution:

Let $A =$"the woman has breast cancer", and $B =$"a positive test."

We want $P(A|B)$ but what we are given is $P(B|A) = 0.9$.

By the definition of conditional probability,

$$P(A|B) \ = \ P(A \cap B)/P(B)$$

To evaluate the numerator we use the multiplication rule

$$P(A \cap B) \ = \ P(A)P(B|A) \ = \ 0.01 \times 0.9 \ = \ 0.009$$

Similarly,

$$P(A^c \cap B) \ = \ P(A^c)P(B|A^c) \ = \ 0.99 \times 0.1 \ = \ 0.099$$

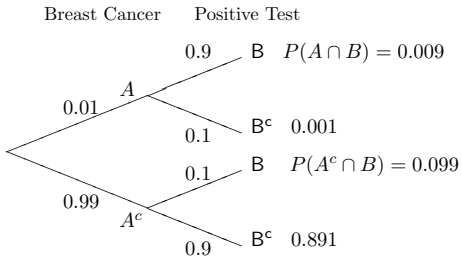
Now $P(B) \ = \ P(A \cap B) + P(A^c \cap B)$ so

$$P(A|B) \ = \ \frac{P(A \cap B)}{P(B)} \ = \ \frac{0.009}{0.009 + 0.099} \ = \ \frac{9}{108}$$

or a little less than 9%. This situation comes about because it is much easier to have a false positive for a healthy woman, which has probability 0.099, than to find a woman with breast cancer having a positive test, which has probability 0.009.

This answer is somewhat surprising. Indeed when ninety-five physicians were asked this question their average answer was 75%. The two statisticians who carried out this survey indicated that physicians were better able to see the answer when the data was presented in frequency format. 10 out of 1000 women have breast cancer. Of these 9 will have a positive mammogram. However of the remaining 990 women without breast cancer 99 will have a positive reaction, and again we arrive the answer $9/(9+99)$.

Alternative solution using a tree diagram:



So the probability that a woman has breast cancer given that she has just had a positive test is

$$P(A|B) \ = \ \frac{P(A \cap B)}{P(B)} \ = \ \frac{0.009}{0.009 + 0.099} \ = \ \frac{9}{108}$$

**In the exam, there won't be any need for electronic calculators and you may leave the answer in either of the last two numerical forms for full credit, provided you show the steps in your reasoning.*

Before we see the more general form of Bayes' Rule, let us make a simple observation called the *total probaility theorem.*

Proposition 14 (Total probability theorem) Suppose $A_1 \cup A_2 \ldots \cup A_k$ is a sequence of events with positive probability that partition the sample space, that is, $A_1 \cup A_2 \cdots \cup A_k = \Omega$ and $A_i \cap A_j = \emptyset$ for any $i \neq j$, then for some arbitrary event B.

$$P(B) \;=\; \sum_{h=1}^{k} P(B \cap A_h) \;=\; \sum_{h=1}^{k} P(B|A_h)P(A_h) \tag{2.4}$$

Proof: The first equality is due to the addition rule for mutually exclusive events,

$$B \cap A_1, B \cap A_2, \ldots, B \cap A_k$$

and the second equality is due to the multiplication rule for two likely events.

Reference to the Venn digram (you should draw below as done in lectures) will help you understand this idea for the four event case.

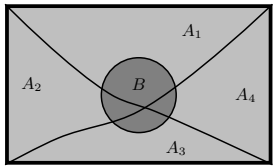


Figure 2.3: Reference to the Venn digram will help you understand this idea behind the proof of the total probability theorem in Proposition 14 for the four event case.

Example 33 (Urn with red and black balls) A well-mixed urn contains five red and ten black balls. We draw two balls from the urn without replacement. What is the probability that the second ball drawn is red?

This is easy to see if we draw a probability tree diagram. The first split in the tree is based on the outcome of the first draw and the second on the outcome of the last draw. The outcome of the first draw dictates the probabilities for the second one since we are sampling without replacement. We multiply the probabilities on the edges to get probabilities of the four endpoints, and then sum the ones that correspond to red in the second draw, that is

$$P(\text{second ball is red}) \;=\; 4/42 + 10/42 \;=\; 1/3 \;.$$

Definition 75 (Independent and Identically Distributed (IID) Process) The finite or infinite sequence of RVs or the stochastic process $X_1, X_2, \ldots$ is said to be independent and identically distributed or IID if :

- they are an idependently distributed according to Definition 43, and
- $F(X_1) = F(X_2) = \cdots$, ie. all the X_i's have the same DF $F(X_1)$.

This is perhaps the most elementary class of stochastic processes and we succinctly denote it by

$$(X_i)_{i\in[n]} := X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} F, \quad \text{or} \quad (X_i)_{i\in\mathbb{N}} := X_1, X_2, \ldots \overset{\text{IID}}{\sim} F \;.$$

We sometimes replace the DF F above by the name of the RV.

Definition 76 (Independently Distributed) The sequence of RVs or the stochastic process $(X_i)_{i\in\mathbb{N}} := X_1, X_2, \ldots$ is said to be independently distributed if :

- $X_1, X_2, \ldots$ is independently distributed according to Definition 43.

This is a class of stochastic processes that is more general than the IID class.

As an example of such a class consider the sequence of RVs that are independent but non-identically distributed with each $X_i \sim \text{Bernoulli}(\theta_i)$.

When a stochastic process $(X_\alpha)_{\alpha\in\mathbb{A}}$ is not independent it is said to be dependent. So far we have mostly concerned ourselves with independent processes. In this chapter we introduce finite Markov chains and their simulation methods. Finit Markov chains are among the simplest stochastic processes with a 'first-order' dependence called Markov dependence.

6.2 Introduction

A finite Markov chain is a stochastic process that moves among elements in a finite set $\mathbb{X}$ as follows: when at $x \in \mathbb{X}$ the next position is chosen at random according to a fixed probability distribution $P(\cdot|x)$. We define such a process more formally below.

Definition 77 (Finite Markov Chain) A stochastic sequence,

$$(X_n)_{n\in\mathbb{Z}_+} := (X_0, X_1, \ldots),$$

is a homogeneous **Markov chain** with **state space** $\mathbb{X}$ and **transition matrix** $P := (P(x,y))_{(x,y)\in\mathbb{X}^2}$ if for all pair of **states** $(x,y) \in \mathbb{X}^2 := \mathbb{X} \times \mathbb{X}$, all integers $t \geq 1$, and all probable historical events $H_{t-1} := \bigcap_{n=0}^{t-1}\{X_n = x_n\}$ with $\mathbf{P}\left(H_{t-1} \cap \{X_t = x\}\right) > 0$, the following **Markov property** is satisfied:

$$\mathbf{P}\left(X_{t+1} = y | H_{t-1} \cap \{X_t = x\}\right) = \mathbf{P}\left(X_{t+1} = y | X_t = x\right) =: P(x,y) \;. \tag{6.1}$$

The Markov property means that the conditional probability of going to state y at time $t+1$ from state x at current time t is always given by the (x,y)-th entry $P(x,y)$ of the transition matrix P, no matter what sequence of states $(x_0, x_1, \ldots, x_{t-1})$ preceded the current state x. Thus, the

Chapter 6

Finite Markov Chains

NOTE: No materials from this Chapter will be in Probability Theory I exam!

This topic is only introduced briefly in Probability Theory I to give concrete instances of dependent sequence of random variables. We will revisit these ideas in the sequel. You only need to understand the ideas behind:

- *Example 169 and*
- *Example 170.*

as done in lectures.

6.1 Stochastic Processes

Definition 74 (Stochastic Process) A collection of RVs

$$(X_\alpha)_{\alpha\in N} := (\ X_\alpha : \alpha \in \mathbb{A}\ )$$

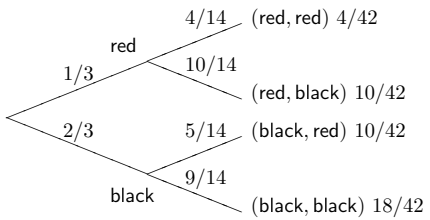
is called a **stochastic process**. Thus, for every $\alpha \in \mathbb{A}$, the index set of the stochastic process, X_α is a RV. If the index set $\mathbb{A} \subset \mathbb{Z}$ then we have a **discrete time stochastic process**, typically denoted by

$$(X_i)_{i\in\mathbb{Z}} := \ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots, \text{ or}$$
$$(X_i)_{i\in\mathbb{N}} := X_1, X_2, \ldots, \text{ or}$$
$$(X_i)_{i\in[n]} := X_1, X_2, \ldots, X_n, \text{ where, } [n] := \{1, 2, \ldots, n\}\ .$$

If $\mathbb{A} \subset \mathbb{R}$ then we have a **continuous time stochastic process**, typically denoted by $\{X_t\}_{t\in\mathbb{R}}$, etc.

Of course the above process is quite general and can allow for arbitrary dependence among the RVs. Generally, we cannot produce useful models without making simplifying assumptions. The absolutely simplest but extremely useful assumption is that of the **Independent and Identically Distributed or IID Process** or merely **IID Sequence of RVs** when the index set is a subset of $\mathbb{N}$.

This is exactly the sequence of RVs associated with our product experiment $\mathcal{E}^{\otimes\infty} := (\Omega, \mathcal{F}_\mathcal{X}, P_\theta)^{\otimes\infty}$.



Alternatively, use the total probability theorem to break the problem down into manageable pieces. Let $R_1 = \{(\mathsf{red}, \mathsf{red}), (\mathsf{red}, \mathsf{black})\}$ and $R_2 = \{(\mathsf{red}, \mathsf{red}), (\mathsf{black}, \mathsf{red})\}$ be the events corresponding to a red ball in the 1st and 2nd draws, respectively, and let $B_1 = \{(\mathsf{black}, \mathsf{red}), (\mathsf{black}, \mathsf{black})\}$ be the event of a black ball on the first draw.

Now R_1 and B_1 partition Ω so we can write:

$$\begin{aligned} P(R_2) &= \mathbf{P}(R_2 \cap R_1) + \mathbf{P}(R_2 \cap B_1) \\ &= \mathbf{P}\ (R_2|R_1)\mathbf{P}(R_1) + \mathbf{P}(R_2|B_1)\mathbf{P}(B_1) \\ &= (4/14)(1/3)\ +\ (5/14)(2/3) = 1/3\ . \end{aligned}$$

Proposition 15 (Bayes' Theorem, 1763) Suppose the events $A_1, A_2, \ldots, A_k \in \mathcal{F}$, with $\mathbf{P}(A_h) > 0$ for each $h \in \{1, 2, \ldots, k\}$, partition the sample space Ω, ie. they are mutually exclusive (disjoint) and exhaustive events with positive probability:

$$A_i \cap A_j = \emptyset, \text{ for any distinct } i, j \in \{1, 2, \ldots, k\}, \qquad \bigcup_{h=1}^{k} A_h = \Omega, \qquad \mathbf{P}(A_h) > 0$$

Thus, precisely one of the A_h's will occur on any performance of our experiment $\mathcal{E}$.

Let $B \in \mathcal{F}$ be some event with $\mathbf{P}(B) > 0$, then

$$\mathbf{P}(A_h|B) = \frac{\mathbf{P}(B|A_h)\mathbf{P}(A_h)}{\sum_{h=1}^{k}\mathbf{P}(B|A_h)\mathbf{P}(A_h)} \tag{2.5}$$

Proof: We apply elementary set theory, the definition of conditional probability $k+2$ times and the addition rule once:

$$\begin{aligned} \mathbf{P}(A_h|B) &= \frac{\mathbf{P}(A_h \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(B \cap A_h)}{\mathbf{P}(B)} = \frac{\mathbf{P}(B|A_h)\mathbf{P}(A_h)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(B|A_h)\mathbf{P}(A_h)}{\mathbf{P}\left(\bigcup_{h=1}^{k}(B \cap A_h)\right)} = \frac{\mathbf{P}(B|A_h)\mathbf{P}(A_h)}{\sum_{h=1}^{k}\mathbf{P}\ (B \cap A_h)} \\ &= \frac{\mathbf{P}(B|A_h)\mathbf{P}(A_h)}{\sum_{h=1}^{k}\mathbf{P}(B|A_h)\mathbf{P}(A_h)} \end{aligned}$$

The operations done to the denominator in the proof above is merely the total probability theorem:

$$\mathbf{P}(B) = \sum_{h=1}^{k}\mathbf{P}(B|A_h)\mathbf{P}(A_h)$$

We call $\mathbf{P}(A_h)$ the **prior probability of** A_h, i.e., before observing B or *a priori*, and $\mathbf{P}(A_h|B)$ the **posterior probability of** A_h, i.e., after observing B or *a posteriori*.

This theorem is at the heart of solving Bayesian *Decision Problems* which fall into several sub-problems called *inference*, *learning* and *control* problems. Let's see one of the simplest such *learning problems* called *prediction*, more specifically *classification*, where we need to choose between finitely many possible choices based on past information next.

Example 34 (Wasserman2003 p.12) Suppose Larry divides his email into three categories: A_1 = "spam", A_2 = "low priority", and A_3 = "high priority". From previous experience, he finds that $\mathbf{P}(A_1) = 0.7$, $\mathbf{P}(A_2) = 0.2$ and $\mathbf{P}(A_3) = 0.1$. Note that $\mathbf{P}(A_1 \cup A_2 \cup A_3) = \mathbf{P}(\Omega) = 0.7 + 0.2 + 0.1 = 1$. Let B be the event that the email contains the word "free." From previous experience, $\mathbf{P}(B|A_1) = 0.9$, $\mathbf{P}(B|A_2) = 0.01$ and $\mathbf{P}(B|A_3) = 0.01$. Note that $\mathbf{P}(B|A_1)+\mathbf{P}(B|A_2)+\mathbf{P}(B|A_3) = 0.9+0.01+0.01 \neq 1$. Now, suppose Larry receives an email with the word "free." What is the probability that it is "spam," "low priority," and "high priority" ? Solution:

$$\mathbf{P}(A_1|B) = \frac{\mathbf{P}(B|A_1)\mathbf{P}(A_1)}{\mathbf{P}(B|A_1)\mathbf{P}(A_1)+\mathbf{P}(B|A_2)\mathbf{P}(A_2)+\mathbf{P}(B|A_3)\mathbf{P}(A_3)} = \frac{0.9\times 0.7}{(0.9\times 0.7)+(0.01\times 0.2)+(0.01\times 0.1)} = \frac{0.63}{0.633} \approxeq 0.995$$

$$\mathbf{P}(A_2|B) = \frac{\mathbf{P}(B|A_2)\mathbf{P}(A_2)}{\mathbf{P}(B|A_1)\mathbf{P}(A_1)+\mathbf{P}(B|A_2)\mathbf{P}(A_2)+\mathbf{P}(B|A_3)\mathbf{P}(A_3)} = \frac{0.01\times 0.2}{(0.9\times 0.7)+(0.01\times 0.2)+(0.01\times 0.1)} = \frac{0.002}{0.633} \approxeq 0.003$$

$$\mathbf{P}(A_3|B) = \frac{\mathbf{P}(B|A_3)\mathbf{P}(A_3)}{\mathbf{P}(B|A_1)\mathbf{P}(A_1)+\mathbf{P}(B|A_2)\mathbf{P}(A_2)+\mathbf{P}(B|A_3)\mathbf{P}(A_3)} = \frac{0.01\times 0.1}{(0.9\times 0.7)+(0.01\times 0.2)+(0.01\times 0.1)} = \frac{0.001}{0.633} \approxeq 0.002$$

Note that $\mathbf{P}(A_1|B) + \mathbf{P}(A_2|B) + \mathbf{P}(A_3|B) = 0.995 + 0.003 + 0.002 = 1$.

This is essentially the idea behind *Bayes classifiers*, that are used to solve such *prediction* problems across different problem domains in *statistical machine learning*, where solutions are given from computer programs.

2.4.2 Independence and Dependence

In general, $P(A|B)$ and $P(A)$ are different, but sometimes the occurrence of B makes no difference, and gives no new information about the chances of A occurring. This is the idea behind independence. Events like "having blue eyes" and "having blond hair" are associated due to common genetic ancestry, but events like "my neighbour wins Lotto" and "I win Lotto" are not due to the Lotto machine being chaotically whirled around before ejection (as modelled by a well-stirred urn).

Definition 16 (Independence of two events) Any two events A and B are said to be **independent** if and only if

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) \ . \tag{2.6}$$

Let us make sense of this definition in terms of our previous definitions. When $\mathbf{P}(A) = 0$ or $\mathbf{P}(B) = 0$, both sides of the above equality are 0. If $\mathbf{P}(A) \neq 0$, then rearranging the above equation we get:

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

But, the LHS is $\mathbf{P}(B|A)$ by definition 2.2, and thus for independent events A and B, we get:

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

This says that information about the occurrence of A does not affect the occurrence of B. If $\mathbf{P}(B) \neq 0$, then an analogous argument:

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

Example 168 We model the tosses of a coin with unknown $E(X_1) = \theta^*$ as

$$X_1, X_2, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(\theta^*)$$

and observed the following data, sample mean, sample variance and sample standard deviation:

$$(x_1, x_2, \ldots, x_7) = (0, 1, 1, 0, 0, 1, 0),\ \overline{x}_7 = 0.4286,\ s_7^2 = 0.2857,\ s_7 = 0.5345 \ ,$$

respectively. Our point estimate and $1 - \alpha = 95\%$ confidence interval for $E(X_1)$ are:

$$\overline{x}_7 = 0.4286 \quad \text{and} \quad (\overline{x}_7 \pm z_{\alpha/2} s_7/\sqrt{7}) = (0.4286 \pm 1.96 \times 0.5345/\sqrt{7}) = (0.0326, 0.8246) \ ,$$

respectively. So with 95% probability the true population mean $E(X_1) = \theta^*$ is contained in $(0.0326, 0.8246)$ and since $1/2$ is contained in this interval of width 0.792 we cannot rule out that the flipped coin is not fair with $\theta^* = 1/2$.

Remark 73 The normal-based confidence interval for θ^* (as well as λ^* in the previous example) may not be a valid approximation here with just $n = 7$ samples. After all, the CLT only tells us that the point estimator $\widehat{\Theta}_n$ can be approximated by a normal distribution for large sample sizes. When the sample size n was increased from 7 to 100 by tossing the same coin another 93 times, a total of 57 trials landed as Heads. Thus the point estimate and confidence interval for $E(X_1) = \theta^*$ based on the sample mean and sample standard deviations are:

$$\widehat{\theta}_{100} = \frac{57}{100} = 0.57 \qquad \text{and} \qquad (0.57 \pm 1.96 \times 0.4975/\sqrt{100}) = (0.4725, 0.6675) \ .$$

Thus our confidence interval shrank considerably from a width of 0.792 to 0.195 after an additional 93 Bernoulli trials. Thus, we can make the width of the confidence interval as small as we want by making the number of observations or sample size n as large as we can.

5.4 Exercises in Limit Laws of Statistics

Ex. 5.2 — Suppose you plan to obtain a simple random sequence (SRS) — also known as independent and identically distributed (IID) sequence — of n measurements from an instrument. This instrument has been calibrated so that the distribution of measurements made with it have population variance of $1/4$. Your boss wants you to make a point estimate of the unknown population mean from a SRS of sample size n. He also insists that the tolerance for error has to be $1/10$ and the probability of meeting this tolerance should be just above 95%. Use CLT to find how large should n be to meet the specifications of your boss.

Ex. 5.3 — Suppose the collection of RVs $X_1, X_2, \ldots, X_n$ model the number of errors in n computer programs named $1, 2, \ldots, n$, respectively. Suppose that the RV X_i modeling the number of errors in the i-th program is the Poisson($\lambda = 5$) for any $i = 1, 2, \ldots, n$. Further suppose that they are independently distributed. Succinctly, we suppose that

$$X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} \text{Poisson}(\lambda = 5) \ .$$

Suppose we have $n = 125$ programs and want to make a probability statement about $\overline{X}_{125}$ which is the average error per program out of these 125 programs. Since $E(X_i) = \lambda = 5$ and $V(X_i) = \lambda = 5$, we want to know how often our sample mean $\overline{X}_{125}$ differs from the expectation of 5 errors per program. Using the CLT find the $P(\overline{X}_{125} < 5.5)$.

Ex. 5.4 — What is the distribution of $\sum_{i=1}^n X_i/n$ as $n \to \infty$ when $X_i \overset{IID}{\sim} \text{Uniform}(-10, 10)$?

Ex. 5.5 — What is the distribution of $\sum_{i=1}^n X_i/\sqrt{\mathbf{V}(X_i)n}$ as $n \to \infty$ when $X_i \overset{IID}{\sim} \text{Uniform}(-10, 10)$?

5.3.2 Application: Set Estimation of $E(X_1)$

A useful byproduct of the CLT is the $(\mathbf{1}-\alpha)$ **confidence interval**, a random interval (or bivariate R$\vec{\text{V}}$) that contains $E(X_1)$, the quantity of interest, with probability $1-\alpha$:

$$\left(\overline{X}_n \pm z_{\alpha/2}\sqrt{V(X_1)/n}\right) := \left(\overline{X}_n - z_{\alpha/2}\sqrt{V(X_1)/n}\,,\, \overline{X}_n + z_{\alpha/2}\sqrt{V(X_1)/n}\right) \quad . \tag{5.10}$$

We can easily see how Equation (5.10) is derived from CLT as follows:

$$\begin{aligned}
P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) &= 1-\alpha \\
P\left(-z_{\alpha/2} < \frac{\overline{X}_n - E(X_1)}{\sqrt{V(X_1)/n}} < z_{\alpha/2}\right) &= 1-\alpha \\
P\left(-\overline{X}_n - z_{\alpha/2}\sqrt{V(X_1)/n} < -E(X_1) < -\overline{X}_n + z_{\alpha/2}\sqrt{V(X_1)/n}\right) &= 1-\alpha \\
P\left(\overline{X}_n + z_{\alpha/2}\sqrt{V(X_1)/n} > E(X_1) > \overline{X}_n - z_{\alpha/2}\sqrt{V(X_1)/n}\right) &= 1-\alpha \\
P\left(\overline{X}_n - z_{\alpha/2}\sqrt{V(X_1)/n} < E(X_1) < \overline{X}_n + z_{\alpha/2}\sqrt{V(X_1)/n}\right) &= 1-\alpha \\
P\left(E(X_1) \in \left(\overline{X}_n - z_{\alpha/2}\sqrt{V(X_1)/n}, \overline{X}_n + z_{\alpha/2}\sqrt{V(X_1)/n}\right)\right) &= 1-\alpha \ .
\end{aligned}$$

Remark 72 (Heuristic interpretation of the $(1-\alpha)$ confidence interval) If we repeatedly produced samples of size n to contain $E(X_1)$ within a $\left(\overline{X}_n \pm z_{\alpha/2}\sqrt{V(X_1)/n}\right)$, say 100 times, then on average, $(1-\alpha)\times 100$ repetitions will actually contain $E(X_1)$ within the random interval and $\alpha \times 100$ repetitions will fail to contain $E(X_1)$.

So far, we have assumed we know the population variance $V(X_1)$ in an IID experiment with n samples and tried to estimate the population mean $E(X_1)$. But in general, we will not know $V(X_1)$. We can still get a point estimate of $E(X_1)$ from the sample mean due to LLN but we won't be able to get a confidence interval for $E(X_1)$. Fortunately, a more elaborate form of the CLT tells us that even when we substitute the sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ for the population variance $V(X_1)$ the following $1-\alpha$ confidence interval for $E(X_1)$ works!

$$\left(\overline{X}_n \pm z_{\alpha/2}S_n/\sqrt{n}\right) := \left(\overline{X}_n - z_{\alpha/2}S_n/\sqrt{n}\,,\, \overline{X}_n + z_{\alpha/2}S_n/\sqrt{n}\right) \quad , \tag{5.11}$$

where, $S_n = \sqrt{S_n^2}$ is the sample standard deviation.
Let's return to our two examples again.

Example 167 We model the waiting times between Orbiter buses with unknown $E(X_1) = 1/\lambda^*$ as

$$X_1, X_2, \ldots, X_n \overset{IID}{\sim} \text{Exponential}(\lambda^*)$$

and observed the following data, sample mean, sample variance and sample standard deviation:

$$(x_1, x_2, \ldots, x_7) = (2, 12, 8, 9, 14, 15, 11),\ \overline{x}_7 = 10.143,\ s_7^2 = 19.143,\ s_7 = 4.375 \ ,$$

respectively. Our point estimate and $1-\alpha = 95\%$ confidence interval for $E(X_1)$ are:

$$\overline{x}_7 = 10.143 \quad \text{and} \quad (\overline{x}_7 \pm z_{\alpha/2}s_7/\sqrt{7}) = (10.143 \pm 1.96 \times 4.375/\sqrt{7}) = (6.9016, 13.3841) \ ,$$

respectively. So with 95% probability the true population mean $E(X_1) = 1/\lambda^*$ is contained in $(6.9016, 13.3841)$ and since the mean waiting time of 10 minutes promised by the Orbiter bus company is also within $(6.9016, 13.3841)$ we can be fairly certain that the company sticks to its promise.

says that information about the occurrence of B does not affect the occurrence of A. Therefore, the probability of their joint occurence $\mathbf{P}(A \cap B)$ is simply the product of their individual probabilities $\mathbf{P}(A)\mathbf{P}(B)$.

Definition 17 (Independence of a sequence of events) We say that a finite or infinite sequence of events $A_1, A_2, \ldots$ are independent if whenever $i_1, i_2, \ldots, i_k$ are distinct elements from the set of indices $\mathbb{N}$, such that $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$ are defined (elements of $\mathcal{F}$), then

$$\mathbf{P}(A_{i_1} \cap A_{i_2} \ldots \cap A_{i_k}) = \mathbf{P}(A_{i_1})\mathbf{P}(A_{i_2}) \cdots \mathbf{P}(A_{i_k})$$

Example 35 (Some Standard Examples) A sequence of events in a sequence of independent trials is independent.

(a) Suppose you toss a fair coin twice such that the first toss is independent of the second. Then,

$$\mathbf{P}(\mathsf{Heads\ on\ the\ first\ toss} \cap \mathsf{Tails\ on\ the\ second\ toss}) = \mathbf{P}(\mathsf{H})\mathbf{P}(\mathsf{T}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \ .$$

(b) Suppose you independently toss a fair die three times. Let E_i be the event that the outcome is an even number on the i-th trial. The probability of getting an even number in all three trials is:

$$\begin{aligned}
\mathbf{P}(E_1 \cap E_2 \cap E_3) &= \mathbf{P}(E_1)\mathbf{P}(E_2)\mathbf{P}(E_3) \\
&= (\mathbf{P}(\{2,4,6\}))^3 \\
&= (\mathbf{P}(\{2\} \cup \{4\} \cup \{6\}))^3 \\
&= (\mathbf{P}(\{2\}) + \mathbf{P}(\{4\}) + \mathbf{P}(\{6\}))^3 \\
&= \left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right)^3 = \left(\frac{1}{2}\right)^3 = \frac{1}{8} \ .
\end{aligned}$$

(c) Suppose you toss a fair coin independently m times. Then each of the 2^m possible outcomes in the sample space Ω has equal probability of $\frac{1}{2^m}$ due to independence.

Example 36 (dependence and independence) Suppose we toss two fair dice. Let A denote the event that the sum of the dice is six and B denote the event that the first die equals four. The sample space encoding the thirty six ordered pairs of outcomes for the two dice is $\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(2,6),\ldots,(5,6),(6,6)\}$ and due to independence $\mathbf{P}(\omega) = 1/36$ for each $\omega \in \Omega$. Then

$$\mathbf{P}(A \cap B) = \mathbf{P}(\{(4,2)\}) = \frac{1}{36} \ ,$$

but

$$\begin{aligned}
\mathbf{P}(A)\mathbf{P}(B) &= \mathbf{P}(\{(1,5),(2,4),(3,3),(4,2),(5,1)\})\mathbf{P}(\{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\}) \\
&= \frac{5}{36} \times \frac{6}{36} = \frac{5}{36} \times \frac{1}{6} = \frac{5}{216} \ ,
\end{aligned}$$

and therefore A and B are not independent. The reason for the events A and B being dependent is clear because the chance of getting a total of six depends on the outcome of the first die (not being six).

Now, let C be the event that the sum of the two dice equals seven. Then

$$\mathbf{P}(C \cap B) = \mathbf{P}(\{(4,3)\}) = \frac{1}{36} \ ,$$

while

$$\begin{aligned} \mathbf{P}(C \cap B) &= \mathbf{P}(\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\})\mathbf{P}(\{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\}) \\ &= \frac{6}{36} \times \frac{6}{36} = \frac{1}{36} \ , \end{aligned}$$

and therefore C and B are independent events. Once again this is clear because the chance of getting a total of seven does not depend any more on the outcome of the first die (it is allowed to be any one of the six possible outcomes).

Example 37 (Pairwise independent events that are not jointly independent) Let a ball be drawn from an well-stirred urn containing four balls labelled 1,2,3,4. Consider the events $A = \{1,2\}$, $B = \{1,3\}$ and $C = \{1,4\}$. Then,

$$\begin{aligned} \mathbf{P}(A \cap B) &= \mathbf{P}(A)\mathbf{P}(B) = \frac{2}{4} \times \frac{2}{4} = \frac{1}{4}, \\ \mathbf{P}(A \cap C) &= \mathbf{P}(A)\mathbf{P}(C) = \frac{2}{4} \times \frac{2}{4} = \frac{1}{4}, \\ \mathbf{P}(B \cap C) &= \mathbf{P}(B)\mathbf{P}(C) = \frac{2}{4} \times \frac{2}{4} = \frac{1}{4}, \end{aligned}$$

but,

$$\frac{1}{4} = \mathbf{P}(\{1\}) = \mathbf{P}(A \cap B \cap C) \neq \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C) = \frac{2}{4} \times \frac{2}{4} \times \frac{2}{4} = \frac{1}{8} \ .$$

Therefore, inspite of being pairwise independent, the events A, B and C are not jointly independent.

CONDITIONAL PROBABILITY SUMMARY

$\mathbf{P}(A|B)$ means the probability that A occurs given that B has occurred.

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A)\mathbf{P}(B|A)}{\mathbf{P}(B)} \quad \text{if} \quad \mathbf{P}(B) \neq 0$$

$$\mathbf{P}(B|A) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} = \frac{\mathbf{P}(B)\mathbf{P}(A|B)}{\mathbf{P}(A)} \quad \text{if} \quad \mathbf{P}(A) \neq 0$$

Conditional probabilities obey the axioms and rules of probability.

2.5 Exercises in Conditional Probability

Ex. 2.6 — What gives the greater probability of hitting some target at least once:

1.hitting in a shot with probability $\frac{1}{2}$ and firing 1 shot, or

2.hitting in a shot with probability $\frac{1}{3}$ and firing 2 shots?

First guess. Then calculate.

Example 165 Suppose an IID sequence of observations $(x_1, x_2, \ldots, x_{80})$ was drawn from a distribution with variance $V(X_1) = 4$. What is the probability that the error in $\overline{x}_n$ used to estimate $E(X_1)$ is less than 0.1?

By CLT,

$$P(\mathsf{error} < 0.1) \approxeq P\left(-\frac{0.1}{\sqrt{4/80}} < Z < \frac{0.1}{\sqrt{4/80}}\right) = P(-0.447 < Z < 0.447) = 0.345 \ .$$

Suppose you want the error to be less than tolerance $= \epsilon$ with a certain probability $1-\alpha$. Then we can use CLT to do such **sample size calculations**. Recall the DF $\Phi(z) = P(Z < z)$ is tabulated in the standard normal table and now we want

$$P\left(-\frac{\epsilon}{\sqrt{V(X_1)/n}} < Z < \frac{\epsilon}{\sqrt{V(X_1)/n}}\right) = 1-\alpha \ .$$

We know,

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1-\alpha \ ,$$

make the picture here of $f_Z(z) = \Phi'(z)$ to recall what $z_{\alpha/2}$, $z_{-\alpha/2}$, and the various areas below $f_Z(\cdot)$ in terms of $\Phi(\cdot)$ from the table really mean... (See Example 59).

where, $\Phi(z_{\alpha/2}) = 1-\alpha/2$ and $\Phi(z_{-\alpha/2}) = 1-\Phi(z_{\alpha/2}) = \alpha/2$. So, we set

$$\frac{\epsilon}{\sqrt{V(X_1)/n}} = z_{\alpha/2}$$

and rearrange to get

$$n = \left(\frac{\sqrt{V(X_1)}z_{\alpha/2}}{\epsilon}\right)^2 \tag{5.9}$$

for the needed sample size that will ensure that our error is less than our tolerance $= \epsilon$ with probability $1-\alpha$. Of course, if n given by Equation (5.9) is not a natural number then we naturally round up to make it one!

A useful $z_{\alpha/2}$ value to remember: If $\alpha = 0.05$ when the probability of interest $1-\alpha = 0.95$ then $z_{\alpha/2} = z_{0.025} = 1.96$.

Example 166 How large a sample size is needed to make the error in our estimate of the population mean $E(X_1)$ to be less than 0.1 with probability $1-\alpha = 0.95$ if we are observing IID samples from a distribution with a population variance $V(X_1)$ of 4?

Using Equation (5.9) we see that the needed sample size is

$$n = \left(\frac{\sqrt{4} \times 1.96}{0.1}\right)^2 \approxeq 1537$$

Thus, it pays to check the sample size needed in advance of experimentation, provided you already know the population variance of the distribution whose population mean you are interested in estimating within a given tolerance and with a high probability.

Second,

$$U_n := \frac{\overline{X}_n - E(X_1)}{\sqrt{V(X_1)/n}} = \frac{\sum_{k=1}^n X_k - nE(X_1)}{\sqrt{nV(X_1)}} = \frac{1}{\sqrt{n}} \sum_{k=1}^n \left(\frac{X_k - E(X_1)}{\sqrt{V(X_1)}} \right) \ .$$

Therefore, the CF of U_n is

$$\varphi_{U_n}(t) = E\left(\exp\left(\imath t U_n\right)\right) = E\left(\exp\left(\imath \frac{t}{\sqrt{n}} \sum_{k=1}^n \frac{X_k - E(X_1)}{\sqrt{V(X_1)}}\right)\right) = \prod_{k=1}^n E\left(\exp\left(\imath \frac{t}{\sqrt{n}} \frac{X_k - E(X_1)}{\sqrt{V(X_1)}}\right)\right)$$
$$= \left(E\left(\exp\left(\imath \frac{t}{\sqrt{n}} \frac{X_1 - E(X_1)}{\sqrt{V(X_1)}}\right)\right)\right)^n \ .$$

Now, if we let

$$Y = \frac{X_1 - E(X_1)}{\sqrt{V(X_1)}}$$

then

$$E(Y) = 0\,, \quad E(Y^2) = 1\,, \text{and} \quad V(Y) = 1 \ .$$

So, the CF of U_n is

$$\varphi_{U_n}(t) = \left(\varphi_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n \ ,$$

and since we can Taylor expand $\varphi_Y(t)$ as follows:

$$\varphi_Y(t) = 1 + \imath t E(Y) + \imath^2 \frac{t^2}{2} E(Y^2) + o(t^2) \ ,$$

which implies

$$\varphi_Y\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{\imath t}{\sqrt{n}} E(Y) + \frac{\imath^2 t^2}{2n} E(Y^2) + o\left(\frac{t^2}{n}\right) \ ,$$

we finally get

$$\varphi_{U_n}(t) = \left(\varphi_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n = \left(1 + \frac{\imath t}{\sqrt{n}} \times 0 + \frac{\imath^2 t^2}{2n} \times 1 + o\left(\frac{t^2}{n}\right)\right)^n = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \to e^{-t^2/2} = \varphi_Z(t) \ .$$

For the last limit we have used $\left(1 + \frac{x}{n}\right)^n \to e^x$ as $n \to \infty$. Thus, we have proved Equation (5.8) which is equivalent to Equation (5.7) by a standardization argument that if $W \sim \text{Normal}(\mu, \sigma^2)$ then $Z = \frac{W-\mu}{\sigma} \sim \text{Normal}(0,1)$ through the linear transformation $W = \sigma Z + \mu$ of Example 67.

5.3.1 Application: Tolerating Errors in our estimate of $E(X_1)$

Recall that we wanted to ensure the $\mathsf{error} = |\overline{X}_n - E(X_1)|$ in our estimate of $E(X_1)$ is within a required $\mathsf{tolerance} = \epsilon$ and make the following probability statement:

$$P(\mathsf{error} < \mathsf{tolerance}) = P(|\overline{X}_n - E(X_1)| < \epsilon) = P(-\epsilon < \overline{X}_n - E(X_1) < \epsilon) = 1 - \alpha \ .$$

To be able to do this we needed to know the full distribution of $\overline{X}_n - E(X_1)$ itself.
Due to the Central Limit Theorem (CLT) we now know that (assuming n is large)

$$P(-\epsilon < \overline{X}_n - E(X_1) < \epsilon) \cong P\left(-\frac{\epsilon}{\sqrt{V(X_1)/n}} < \frac{\overline{X}_n - E(X_1)}{\sqrt{V(X_1)/n}} < \frac{\epsilon}{\sqrt{V(X_1)/n}}\right)$$
$$= P\left(-\frac{\epsilon}{\sqrt{V(X_1)/n}} < Z < \frac{\epsilon}{\sqrt{V(X_1)/n}}\right) \ ,$$

where $Z \sim \text{Normal}(0,1)$.

Ex. 2.7 — Suppose we independently roll two fair dice each of whose faces are marked by numbers 1,2,3,4, 5 and 6.

1. List the sample space for the experiment if we note the numbers on the 2 upturned faces.
2. What is the probability of obtaining a sum greater than 4 but less than 7?

Ex. 2.8 — Based on past experience, 70% of students in a certain course pass the midterm test. The final exam is passed by 80% of those who passed the midterm test, but only by 40% of those who fail the midterm test. What fraction of students pass the final exam?

Ex. 2.9 — A small brewery has two bottling machines. Machine 1 produces 75% of the bottles and machine 2 produces 25%. One out of every 20 bottles filled by machine 1 is rejected for some reason, while one out of every 30 bottles filled by machine 2 is rejected. What is the probability that a randomly selected bottle comes from machine 1 given that it is accepted?

Ex. 2.10 — A process producing micro-chips, produces 5% defective, at random. Each micro-chip is tested, and the test will correctly detect a defective one 4/5 of the time, and if a good micro-chip is tested the test will declare it is defective with probability 1/10.

(a) If a micro-chip is chosen at random, and tested to be good, what was the probability that it was defective anyway?

(b) If a micro-chip is chosen at random, and tested to be defective, what was the probability that it was good anyway?

(c) If 2 micro-chips are tested and determined to be good, what is the probability that at least one is in fact defective?

Ex. 2.11 — Suppose that $\frac{2}{3}$ of all gales are force 1, $\frac{1}{4}$ are force 2 and $\frac{1}{12}$ are force 3. Furthermore, the probability that force 1 gales cause damage is $\frac{1}{4}$, the probability that force 2 gales cause damage is $\frac{2}{3}$ and the probability that force 3 gales cause damage is $\frac{5}{6}$.

(a) If a gale is reported, what is the probability of it causing damage?

(b) If the gale caused damage, find the probabilities that it was of: force 1; force 2; force 3.

(c) If the gale did NOT cause damage, find the probabilities that it was of: force 1; force 2; force 3.

Ex. 2.12 — **The sensitivity and specificity of a medical diagnostic test for a disease are defined as follows:

$$\begin{aligned} \text{sensitivity} &= \mathbf{P}\left(\text{ test is positive } | \text{ patient has the disease }\right) \ , \\ \text{specificity} &= \mathbf{P}\left(\text{ test is negative } | \text{ patient does not have the disease }\right) \ . \end{aligned}$$

Suppose that a medical test has a sensitivity of 0.7 and a specificity of 0.95. If the prevalence of the disease in the general populaiton is 1%, find

(a) the probability that a patient who tests positive actually has the disease,

(b) the probability that a patient who tests negative is free from the disease.

Ex. 2.13 — **The detection rate and false alarm rate of an intrusion sensor are defined as

$$\begin{aligned} \text{detection rate} &= \mathbf{P}\left(\text{ detection declared } | \text{ intrusion }\right) \ , \\ \text{false alarm rate} &= \mathbf{P}\left(\text{ detection declared } | \text{ no intrusion }\right) \ . \end{aligned}$$

If the detection rate is 0.999 and the false alarm rate is 0.001, and the probability of an intrusion occuring is 0.01, find

(a)the probability that there is an intrusion when a detection is declared,

(b)the probability that there is no intrusion when no detection is declared.

Ex. 2.14 — **Let A and B be events such that $\mathbf{P}(A) \neq 0$ and $\mathbf{P}(B) \neq 0$. When A and B are disjoint, are they also independent? Explain clearly why or why not.

trajectories for simulated tosses of a fair coin from IID Bernoulli($\theta^* = 1/2$) RVs and the twenty red sample mean trajectories for simulated waiting times from IID Exponential($\lambda^* = 1/10$) RVs in Figure 5.4 with $n = 7$. Clearly, the point estimates for such a small sample size are fluctuating wildly! However, the fluctuations in the point estimates settles down for larger sample sizes.

The *next natural question is how large should the sample size be* in order to have a small interval of width, say 2ϵ, "contain" $E(X_1)$, the quantity of interest, with a high probability, say $1 - \alpha$? If we can answer this then we can make probability statements like the following:

$$P(\mathsf{error} < \mathsf{tolerance}) = P(|\overline{X}_n - E(X_1)| < \epsilon) = P(-\epsilon < \overline{X}_n - E(X_1) < \epsilon) = 1 - \alpha \ .$$

In order to ensure the $\mathsf{error} = |\overline{X}_n - E(X_1)|$ in our estimate of $E(X_1)$ is within a required $\mathsf{tolerance} = \epsilon$ we need to know the full distribution of $\overline{X}_n - E(X_1)$ itself. The Central Limit Theorem (CLT) helps us here.

5.3 Central Limit Theorem

What if we scale the sum of X_i's by $\sqrt{n}$ instead of n?

Exercise 5.1 (What if we scale by $\sqrt{n}$) After reading Sec. 5.1 up to now, think carefully about what you need to be able to show that $Z_n := 1/\sqrt{n}\sum_{i=1}^{n} X_i$ converges in distribution to the Normal$(0, 1/3)$ RV, where $X_i \overset{IID}{\sim}$ Uniform$(-1, 1)$. Hint: Characteristic functions

Proposition 71 (Central Limit Theorem (CLT)) If we are given a sequence of independently and identically distributed (IID) RVs, $X_1, X_2, \ldots \overset{\text{IID}}{\sim} X_1$ and if $E(X) < \infty$ and $V(X_1) < \infty$, then the sample mean $\overline{X}_n$ converges in distribution to the Normal RV with mean given by any one of the IID RVs, say $\mathbf{E}(X_1)$ by convention, and variance given by $\frac{1}{n}$ times the variance of any one of the IID RVs, say $V(X_1)$ by convention. More formally, we write:

$$\text{If} \quad X_1, X_2, \ldots \overset{\text{IID}}{\sim} X_1 \text{ and if } \mathbf{E}(X_1) < \infty, V(X_1) < \infty$$
$$\text{then} \quad \overline{X}_n \rightsquigarrow \text{Normal}\left(E(X_1), \frac{V(X_1)}{n}\right) \quad \text{as } n \to \infty \ , \quad (5.7)$$

or equivalently after standardization:

$$\text{If} \quad X_1, X_2, \ldots \overset{\text{IID}}{\sim} X_1 \text{ and if } \mathbf{E}(X_1) < \infty, V(X_1) < \infty$$
$$\text{then} \quad \frac{\overline{X}_n - E(X_1)}{\sqrt{V(X_1)/n}} \rightsquigarrow Z \sim \text{Normal}\,(0, 1) \quad \text{as } n \to \infty \ . \quad (5.8)$$

Proof: Our proof is based on the convergence of characteristic functions (CFs). We will prove the standardized form of the CLT in Equation (5.8) by showing that the CF of

$$U_n := \frac{\overline{X}_n - E(X_1)}{\sqrt{V(X_1)/n}}$$

converges to the CF of Z, the Normal$(0, 1)$ RV. First, note from Equation (3.72) that the CF of $Z \sim$ Normal$(0, 1)$ is:

$$\varphi_Z(t) = E\left(e^{\imath t Z}\right) = e^{-t^2/2} \ .$$

and therefore, we can use the sample mean $\overline{X}_n$ as a point estimator of $E(X_1) = 1/\lambda^*$.

Now, suppose you model seven waiting times in nearest minutes between Orbiter buses at Balgay street as follows:

$$X_1, X_2, \ldots, X_7 \overset{IID}{\sim} \text{Exponential}(\lambda^*) \ ,$$

and have the following realization as your observed data:

$$(x_1, x_2, \ldots, x_7) = (2, 12, 8, 9, 14, 15, 11) \ .$$

Then you can use the observed sample mean $\overline{x}_7 = (2+12+8+9+14+15+11)/7 = 71/7 \approxeq 10.14$ as a **point estimate** of the population mean $E(X_1) = 1/\lambda^*$. By the rearrangement $\lambda^* = 1/E(X_1)$, we can also obtain a point estimate of the "true" parameter λ^* from $1/\overline{x}_7 = 7/71 \approxeq 0.0986$.

Remark 70 (Point estimates are realizations of the Point Estimator) We say the statistic $\overline{X}_n$, which is a random variable that depends on the data $\text{R}\vec{\text{V}}$ $(X_1, X_2, \ldots, X_n)$, is a **point estimator** of $E(X_1)$. But once we have a realization of the data $\text{R}\vec{\text{V}}$, i.e., our observed data vector $(x_1, x_2, \ldots, x_n)$ and its corresponding realization as observed sample mean $\overline{x}_n$, we say $\overline{x}_n$ is a **point estimate** of $E(X_1)$. In other words, the point estimate $\overline{x}_n$ is a realization of the the random variable $\overline{X}_n$ called the point estimator of $E(X_1)$. Therefore, when we observe a new data vector $(x'_1, x'_2, \ldots, x'_n)$ that is different from our first data vector $(x_1, x_2, \ldots, x_n)$, our point estimator of $E(X_1)$ is still $\overline{X}_n$ but the point estimate $n^{-1}\sum_{i=1}^n x'_i$ may be different from the first point estimate $n^{-1}\sum_{i=1}^n x_i$. The sample means from n samples for 20 replications (repeats of the experiment) are typically distinct especially for small n as shown in Figure 5.4.

Example 164 Let $X_1, X_2, \ldots, X_n \overset{IID}{\sim} X_1$, where X_1 is an Bernoulli(θ^*) RV, i.e., let

$$X_1, X_2, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(\theta^*) \ .$$

Typically, we do not know the "true" parameter $\theta^* \in \Theta = [0, 1]$, which is the same as the population mean $E(X_1) = \theta^*$. But by LLN, we know that

$$\overline{X}_n \rightsquigarrow \text{Point Mass}(E(X_1)) \ ,$$

and therefore, we can use the sample mean $\overline{X}_n$ as a point estimator of $E(X_1) = \theta^*$.

Now, suppose you model seven coin tosses (encoding Heads as 1 with probability θ^* and Tails as 0 with probability $1 - \theta^*$) as follows:

$$X_1, X_2, \ldots, X_7 \overset{IID}{\sim} \text{Bernoulli}(\theta^*) \ ,$$

and have the following realization as your observed data:

$$(x_1, x_2, \ldots, x_7) = (0, 1, 1, 0, 0, 1, 0) \ .$$

Then you can use the observed sample mean $\overline{x}_7 = (0+1+1+0+0+1+0)/7 = 3/7 \approxeq 0.4286$ as a **point estimate** of the population mean $E(X_1) = \theta^*$. Thus, our "single best guess" for $E(X_1)$ which is the same as the probability of Heads is $\overline{x}_7 = 3/7$.

Of course, if we tossed the same coin in the same IID manner another seven times or if we observed another seven waiting times of orbiter buses at a different bus-stop or on a different day we may get a different point estimate for $E(X_1)$. See the intersection of the twenty magenta sample mean

Chapter 3

Random Variables

We are used to classical variables such as x as an "unknown" in the equation: $x + 3 = 7$.

We also use classical variables to represent geometric objects such as a line:

$$y = 3x - 2 \ ,$$

where the variable y for the y-axis is determined by the value taken by the variable x, as x varies over the real line $\mathbb{R} = (-\infty, \infty)$.

Yet another example is the use of variables to represent sequences such as:

$$\{a_n\}_{n=1}^{\infty} = a_1, a_2, a_3, \ldots \ .$$

What these *classical variables* have in common is that they *take a* fixed or *deterministic value* when we can solve for them.

We need a different kind of variable to deal with real-world situations where the same variable may take different values in a non-deterministic manner. **Random variables** do this job for us. Random variables, unlike classical deterministic variables, can take a bunch of different values.

Crucially, it can become inconvenient to work with a set of outcomes Ω upon which arithmetic is not possible. We are often measuring our outcomes with subsets of real numbers. Some examples include:

Experiment	Possible measured outcomes
Counting the number of typos up to now	$\mathbb{Z}_+ := \{0, 1, 2, \ldots\} \subset \mathbb{R}$
Length in centi-meters of some shells on New Brighton beach	$(0, +\infty) \subset \mathbb{R}$
Waiting time in minutes for the next Orbiter bus to arrive	$\mathbb{R}_+ := [0, \infty) \subset \mathbb{R}$
Vertical displacement from current position of a pollen on water	$\mathbb{R}$

Thus, we want a **random variable** to be a function from the sample space Ω to the set of real numbers $\mathbb{R}$, that is, $X : \Omega \to \mathbb{R}$ that should satisfy certain conditions to keep the meaning of the underlying probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Let us go through some examples before giving the formal definition of such a real-valued or $\mathbb{R}$-valued random variable.

Example 38 (Rain or Shine) Suppose our experiment is to observe whether it will rain or not rain tomorrow. The sample space of this experiment is $\Omega = \{\text{rain}, \text{not rain}\}$. We can associate a random variable X with this experiment as follows:

$$X(\omega) = \begin{cases} 1, & \text{if } \omega = \text{rain} \\ 0, & \text{if } \omega = \text{not rain} \end{cases}$$

Thus, X will take the value 1 if it will rain tomorrow and 0 otherwise. Note that another equally valid (though possibly not so useful) random variable, say Y, for this experiment is:

$$Y(\omega) = \begin{cases} \pi, & \text{if } \omega = \text{rain} \\ \sqrt{2}, & \text{if } \omega = \text{not rain} \end{cases}$$

Example 39 (Rain Fall on Angstrom) Suppose our experiment instead is to measure the volume of rain that falls into a large funnel stuck on top of a graduated cylinder that is placed on top of the middle of House 1 of Angstrom Laboratory. Suppose the cylinder is graduated in millimeters then our random variable $X(\omega)$ can report a non-negative real number given by the lower miniscus of the water column, if any, in the cylinder tomorrow. Thus, $X(\omega)$ will measure the volume of rain in millilitres that will fall into our funnel tomorrow.

Example 40 (Counting Seedlings) Suppose ten seeds are planted. Perhaps fewer than ten will actually germinate. The number which do germinate, say X, must be one of the integer numbers in $\mathbb{R}$ given by the set:

$$\mathbb{X} := \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \ .$$

But until the seeds are actually planted and allowed to germinate it is impossible to say which number $X(\omega) : \Omega \to \mathbb{X}$ will take. The number of seeds which germinate is a variable, but it is not necessarily the same for each group of ten seeds planted, but takes values from the same set $\mathbb{X}$. As X is not known in advance it is called a **random variable**. Its value cannot be known until we actually perform the experiment, i.e., plant the seeds.

Certain things can be said about the value a random variable might take. In the case of these ten seeds we can be sure the number that germinate is less than eleven, and not less than zero! It may also be known that that the probability of seven seeds germinating is greater than the probability of one seed; or perhaps that the number of seeds germinating averages eight. These statements are based on probabilities unlike the sort of statements made about deterministic variables.

Discrete versus continuous random variables.

A **discrete** random variable is one in which the set of possible values of the random variable is finite or at most countably infinite, whereas a **continuous** random variable may take on any value in some range, and its value may be any real value in that range (Think: uncountably infinite). Examples 38 and 40 are about discrete random variables and Example 39 is about a continuous random variable.

Discrete random variables are usually generated from experiments where things are "counted" rather than "measured" such as the seed planting experiment in Example 40. Continuous random variables appear in experiments in which we measure, such as the amount of rain, in millilitres in Example 39.

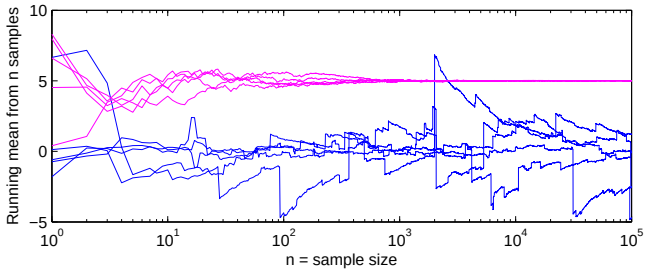
Random variables as functions.

In fact, random variables are actually functions, more formally measurable maps from $\mathcal{F}$ to certain subsets of $\mathbb{R}$ that you will learn carefully in more advanced courses. They take you from the "world of random processes and phenomena" to the world of real numbers. In other words, a random variable is a numerical value determined by the outcome of the experiment.

We said that a random variable can take one of many values, but we cannot be certain of which value it will take. However, *we can make probabilistic statements about the value x the random variable X will take.* A question like,


```
u=rand(1,N);            % draw N IID samples from Uniform(0,1)
x=tan(pi * u);     % draw N IID samples from Standard cauchy RV using inverse CDF
n=1:N;              % make a vector n of current sample size [1 2 3 ... N-1 N]
CSx=cumsum(x); % CSx is the cumulative sum of the array x (type 'help cumsum')
% Runnign Means <- vector division of cumulative sum of samples by n
RunningMeanStdCauchy = CSx ./ n; % Running Mean for Standard Cauchy samples
RunningMeanUnif010 = cumsum(u*10.0) ./ n; % Running Mean for Uniform(0,10) samples
semilogx(n, RunningMeanStdCauchy) %
hold on;
semilogx(n, RunningMeanUnif010, 'm')
end
xlabel('n = sample size');
ylabel('Running mean from n samples')
```

Figure 5.5: Unending fluctuations of the running means based on n IID samples from the Standard Cauchy RV X in each of five replicate simulations (blue lines). The running means, based on n IID samples from the Uniform$(0, 10)$ RV, for each of five replicate simulations (magenta lines).



The resulting plot is shown in Figure 5.5. Notice that the running means or the sample mean of n samples as a function of n, for each of the five replicate simulations, never settles down to a particular value. This is because of the "thick tails" of the density function for this RV which produces extreme observations. Compare them with the running means, based on n IID samples from the Uniform$(0, 10)$ RV, for each of five replicate simulations (magenta lines). The latter sample means have settled down stably to the mean value of 5 after about 700 samples.

5.2.1 Application: Point Estimation of $E(X_1)$

LLN gives us a method to obtain a **point estimator** that gives "the single best guess" for the possibly unknown population mean $E(X_1)$ based on $\overline{X}_n$, the sample mean, of a simple random sequence (SRS) or independent and identically distributed (IID) sequence of n RVs $X_1, X_2, \ldots, X_n \overset{IID}{\sim} X_1$.

Example 163 Let $X_1, X_2, \ldots, X_n \overset{IID}{\sim} X_1$, where X_1 is an Exponential(λ^*) RV, i.e., let

$$X_1, X_2, \ldots, X_n \overset{IID}{\sim} \text{Exponential}(\lambda^*) \ .$$

Typically, we do not know the "true" parameter $\lambda^* \in \mathbf{\Lambda} = (0, \infty)$ or the population mean $E(X_1) = 1/\lambda^*$. But by LLN, we know that

$$\overline{X}_n \rightsquigarrow \text{Point Mass}(E(X_1)) \ ,$$

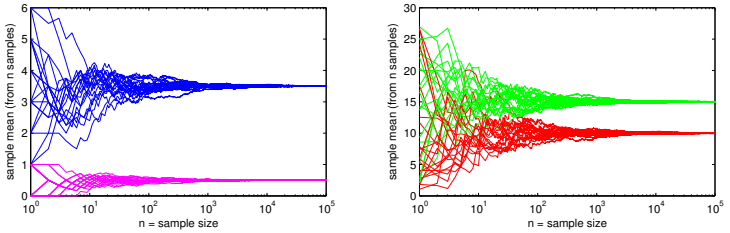
For the last limit we have used $\left(1+\frac{x}{n}\right)^n \to e^x$ as $n \to \infty$.

Finally, we have shown that $E\left(e^{tt\overline{X}_n}\right)$, the CF of the n-sample mean RV $\overline{X}_n$, converges to $\mathbf{E}(e^{tt\mathbf{E}(X_1)}) = e^{tt\mathbf{E}(X_1)}$, the CF of the Point Mass($\mathbf{E}(X_1)$) RV, as the sample size n tends to infinity.

Heuristic Interpretation of LLN

The distribution of the sample mean RV $\overline{X}_n$ obtained from an independent and identically distributed sequence of RVs $X_1, X_2, \ldots$ [i.e. all the RVs X_i's are independent of one another and have the same distribution function, and thereby the same expectation, variance and higher moments], concentrates around the expectation of any one of the RVs in the sequence, say that of the first one $\mathbf{E}(X_1)$ [without loss of generality], as n approaches infinity. See Figure 5.4 for examples of 20 replicates of the sample mean of IID sequences from four RVs. All the sample mean trajectories converge to the corresponding population mean.

Figure 5.4: Sample mean $\overline{X}_n$ as a function of sample size n for 20 replications from independent realizations of a fair die (blue), fair coin (magenta), Uniform$(0, 30)$ RV (green) and Exponential(0.1) RV (red) with population means $(1+2+3+4+5+6)/6 = 21/6 = 3.5$, $(0+1)/2 = 0.5$, $(30-0)/2 = 15$ and $1/0.1 = 10$, respectively.



Example 161 (Bernoulli WLLN and Galton's Quincunx) We can appreciate the WLLN for $\overline{X}_n = n^{-1}S_n = \sum_{i=1}^n X_i$, where $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} Bernoulli(p)$ using the paths of balls dropped into Galton's Quincunx of Sec. 3.2.3.

Cauchy whose expectations does not exist has no Law of Large Numbers

Recall that the mean of the Cauchy RV X does not exist since $\int |x|\, dF(x) = \infty$ (3.55). We will investigate this in Labwork 162.

Labwork 162 (Running mean of the Standard Cauchy RV) Let us see what happens when we plot the running sample mean for an increasing sequence of IID samples from the Standard Cauchy RV X by implementing the following script file:

PlotStandardCauchyRunningMean.m

```
% script to plot the oscillating running mean of Std Cauchy samples
% relative to those for the Uniform(0,10) samples
rand('twister',25567);        % initialize the fundamental sampler
for i=1:5
N = 10^5;                     % maximum sample size
```


"What is the probability of it raining tomorrow?"

in the rain/not experiment of Example 38 becomes

"What is $P(\{\omega : X(\omega) = 1\})$?"

or, more simply,

"What is $P(X = 1)$?"

With this motivation we are ready to formally define such a random variable.

3.1 Basic Definitions

To take advantage of our measurements over the real numbers, in terms of its metric structure and arithmetic, we need to formally define this measurement process using the notion of a random variable.

Definition 18 (Random Variable) Let $(\Omega, \mathcal{F}, P)$ be some probability triple. Then, a **Random Variable (RV)**, say X, is a function from the sample space Ω to the set of real numbers $\mathbb{R}$

$$X : \Omega \to \mathbb{R}$$

such that for every $x \in \mathbb{R}$, the inverse image of the half-open real interval $(-\infty, x]$ is an element of the collection of events $\mathcal{F}$, i.e.:

$$\text{for every } x \in \mathbb{R}, \qquad X^{[-1]}(\ (-\infty, x]\ ) := \{\omega : X(\omega) \leq x\} \in \mathcal{F}\ .$$

This definition can be summarised by the statement that a RV is an $\mathcal{F}$-measurable map. We assign probability to the RV X as follows:

$$\mathbf{P}(X \leq x) = \mathbf{P}(\ X^{[-1]}(\ (-\infty, x]\ )\ ) := \mathbf{P}(\ \{\omega : X(\omega) \leq x\}\ )\ . \tag{3.1}$$

Definition 19 (Distribution Function) The **Distribution Function (DF)** or **Cumulative Distribution Function (CDF)** of any RV X, over a probability triple $(\Omega, \mathcal{F}, P)$, denoted by F is:

$$F(x) := \mathbf{P}(X \leq x) = \mathbf{P}(\ \{\omega : X(\omega) \leq x\}\ ), \qquad \text{for any} \quad x \in \mathbb{R}\ . \tag{3.2}$$

Thus, $F(x)$ or simply F is a non-decreasing, right continuous, $[0, 1]$-valued function over $\mathbb{R}$. When a RV X has DF F we write $X \sim F$.

Remark 20 (Notation) It is enough to understand the idea of random variables as explained above, and work with random variables using simplified notation like

$$P(2 \leq X \leq 3)$$

rather than

$$P(\{\omega : 2 \leq X(\omega) \leq 3\})$$

but note that when learning or doing more advanced work this sample space notation is usually needed to clarify the true meaning of the simplified notation at least to yourself! But in the exam you can use the simpler notation as done in the solutions to exercises.

From the idea of a distribution function, we get:

Proposition 21 The probability that the random variable X takes a value x in the half-open interval $(a, b]$, i.e., $a < x \leq b$, is:

$$P(a < X \leq b) \;=\; F(b) - F(a) \;. \tag{3.3}$$

Proof: Since $(X \leq a)$ and $(a < X \leq b)$ are disjoint events whose union is the event $(X \leq b)$,

$$F(b) = P(X \leq b) = P(X \leq a) + P(a < X \leq b) = F(a) + P(a < X \leq b) \;.$$

Subtraction of $F(a)$ from both sides of the above equation yields Equation 3.3.

A special RV that often plays the role of 'building-block' in Probability and Statistics is the indicator function of an event A that tells us whether the event A has occurred or not. Recall that an event belongs to the collection of possible events $\mathcal{F}$ for our experiment.

Definition 22 (Indicator Function) Given a probability triple $(\Omega, \mathcal{F}, \mathbf{P})$, the **Indicator Function** of an event $A \in \{$ which is denoted $\mathbb{1}_A$ is defined as follows:

$$\mathbb{1}_A(\omega) := \begin{cases} 1 & \text{if} \quad \omega \in A \\ 0 & \text{if} \quad \omega \notin A \end{cases} \tag{3.4}$$

Model 1 (Indicator of an event as Bernoulli **RV)** This is the most primitive RV from which all others are obtained. Let us convince ourselves that $\mathbb{1}_A$ is really a RV. For $\mathbb{1}_A$ to be a RV, we need to verify that for any real number $x \in \mathbb{R}$, the inverse image $\mathbb{1}_A^{[-1]}(\ (-\infty, x]\ )$ is an event, ie :

$$\mathbb{1}_A^{[-1]}(\ (-\infty, x]\ ) := \{\omega : \mathbb{1}_A(\omega) \leq x\} \in \mathcal{F} \;.$$

All we can assume about the collection of events $\mathcal{F}$ is that it contains the event A and that it is a sigma algebra. A careful look at the Figure 3.1 yields:

$$\mathbb{1}_A^{[-1]}(\ (-\infty, x]\ ) := \{\omega : \mathbb{1}_A(\omega) \leq x\} = \begin{cases} \emptyset & \text{if} \quad x < 0 \\ A^c & \text{if} \quad 0 \leq x < 1 \\ A \cup A^c = \Omega & \text{if} \quad 1 \leq x \end{cases}$$

Thus, $\mathbb{1}_A^{[-1]}(\ (-\infty, x]\ )$ is one of the following three sets that belong to $\mathcal{F}$; (1) $\emptyset$, (2) A^c and (3) Ω depending on the value taken by x relative to the interval $[0, 1]$. We have proved that $\mathbb{1}_A$ is indeed a RV.

Model 1 is called the Bernoulli RV for event A with a known probability $\mathbf{P}(A)$. We will define as our next model the Bernoulli(θ) RV by introducing a paramater $\theta \in [0, 1]$ for the typically unknown probability $\mathbf{P}(A)$.

Therefore, for any given $\epsilon > 0$,

$$\begin{aligned} \mathbf{P}(|\overline{X}_n - \mathbf{E}(X_1)| \geq \epsilon) &= \mathbf{P}(|\overline{X}_n - \mathbf{E}(\overline{X}_n)| \geq \epsilon) \qquad \text{[by the IID assumption of } X_1, X_2, \ldots, \mathbf{E}(\overline{X}_n) = \mathbf{E}(X_1), \text{ as per (3.76)]} \\ &= \frac{\frac{1}{n}\mathbf{V}(X_1)}{\epsilon^2} \to 0, \quad \text{as} \quad n \to \infty \;, \end{aligned}$$

or equivalently, $\lim_{n\to\infty} \mathbf{P}(|\overline{X}_n - \mathbf{E}(X_1)| \geq \epsilon) = 0$. And the last statement is the definition of the claim made by the law of large numbers (LLN), namely that $\overline{X}_n \xrightarrow{\mathbf{P}} \mathbf{E}(X_1)$.

Proposition 69 (Weak Law of Large Numbers (WLLN): $\overline{X}_n \rightsquigarrow$ Point Mass($\mathbf{E}(X_1)$)**)** If we are given a sequence of independently and identically distributed (IID) RVs, $X_1, X_2, \ldots \overset{\text{IID}}{\sim} X_1$ and if $\mathbf{E}(X_1)$ exists, i.e. $\mathbf{E}(\text{abs}(X)) < \infty$, then the sample mean $\overline{X}_n$ converges in distribution to the expectation of any one of the IID RVs, say Point Mass($\mathbf{E}(X_1)$) by convention. More formally, we write:

$$\text{If} \quad X_1, X_2, \ldots \overset{\text{IID}}{\sim} X_1 \text{ and if } \mathbf{E}(X_1) \text{ exists, then } \overline{X}_n \rightsquigarrow \text{Point Mass}(\mathbf{E}(X_1)) \text{ as } n \to \infty \;.$$

Proof: Our proof now is based on the convergence of characteristic functions (CFs) pointwise to the CF of the limiting RV, as this implies, by Lévy's Continuity Theorem on CFs [7], the convergence of the corresponding distribution functions (DFs).

First, the CF of Point Mass($\mathbf{E}(X_1)$) is

$$\mathbf{E}(e^{\imath t \mathbf{E}(X_1)}) = e^{\imath t \mathbf{E}(X_1)} \;,$$

since $\mathbf{E}(X_1)$ is just a constant, i.e., a Point Mass RV that puts all of its probability mass at $\mathbf{E}(X_1)$.

Second, the CF of $\overline{X}_n$ is

$$\begin{aligned} \mathbf{E}\left(e^{\imath t \overline{X}_n}\right) = \mathbf{E}\left(e^{\imath t \frac{1}{n}\sum_{k=1}^n X_k}\right) = \mathbf{E}\left(\prod_{k=1}^n e^{\imath t X_k/n}\right) = \prod_{k=1}^n \mathbf{E}\left(e^{\imath t X_k/n}\right) = \prod_{k=1}^n \varphi_{X_k}(t/n) \\ = \prod_{k=1}^n \varphi_{X_1}(t/n) = (\varphi_{X_1}(t/n))^n \;. \end{aligned}$$

Let us recall Landau's "small o" notation for the relation between two functions. We say, $f(x)$ is **small o** of $g(x)$ if f is dominated by g as $x \to \infty$, i.e., $\frac{|f(x)|}{|g(x)|} \to 0$ as $x \to \infty$. More formally, for every $\epsilon > 0$, there exists an x_ϵ such that for all $x > x_\epsilon$ $|f(x)| < \epsilon |g(x)|$. For example, $\log(x)$ is $o(x)$, x^2 is $o(x^3)$ and x^m is $o(x^{m+1})$ for $m \geq 1$.

Third, we can expand any CF whose expectation exists as a Taylor series with a remainder term that is $o(t)$ as follows:

$$\varphi_X(t) = 1 + \imath t \mathbf{E}(X) + o(t) \;.$$

Hence,

$$\varphi_{X_1}(t/n) = 1 + \imath \frac{t}{n}\mathbf{E}(X_1) + o\left(\frac{t}{n}\right)$$

and

$$E\left(e^{\imath t \overline{X}_n}\right) = \left(1 + \imath \frac{t}{n}\mathbf{E}(X_1) + o\left(\frac{t}{n}\right)\right)^n \to e^{\imath t \mathbf{E}(X_1)} \text{ as } n \to \infty \;.$$

[7] https://en.wikipedia.org/wiki/L%C3%A9vy%27s_continuity_theorem

5.1.1 Properties of Convergence of RVs**

We will merely state some properties (without proofs that are hyper-linked for the curious student as they are advanced for this course) and relations between the three notions of convergence with some examples to better appreciate the subtleties among them. You will study the proofs of these statements in Probability Theory II. Just remember that subtle implication relations exist between the three notions.

- Convergence almost surely implies convergence in probability[2]

$$\boxed{X_n \overset{a.s.}{\to} X \implies X_n \overset{\mathbf{P}}{\to} X \ .}$$

- By the Borel-Cantelli Lemma [3], convergence in probability does not imply almost sure convergence in the discrete case [4]
- Convergence in probability implies convergence in distribution [5]

$$\boxed{X_n \overset{\mathbf{P}}{\to} X \implies X_n \rightsquigarrow X \ .}$$

- Convergence in distribution to a constant θ implies convergence in probability to θ: [6]

$$X_n \rightsquigarrow \text{Point Mass}(\theta) \implies X_n \overset{\mathbf{P}}{\to} \text{Point Mass}(\theta) \ .$$

- In general, convergence in distribution does not imply convergence in probability.

5.2 Law of Large Numbers

Proposition 68 (Law of Large Numbers (LLN): $\overline{X}_n \xrightarrow{\mathbf{P}} \mathbf{E}(X_1)$) If we are given a sequence if independent and identically distributed RVs, $X_1, X_2, \ldots \overset{\text{IID}}{\sim} X_1$ and if $\mathbf{E}(X_1)$ exists, as per (3.45), i.e., $\mathbf{E}(\text{abs}(X_1)) < \infty$, and the variance is finite, i.e., $\mathbf{V}(X_1) < \infty$, then the sample mean $\overline{X}_n$ converges in probability to the expectation of any one of the IID RVs, say $\mathbf{E}(X_1)$ by convention. More formally, we write:

$$\text{If} \quad X_1, X_2, \ldots \overset{\text{IID}}{\sim} X_1 \text{ and if } \mathbf{E}(X_1) \text{ exists, then } \overline{X}_n \xrightarrow{\mathbf{P}} \mathbf{E}(X_1) \ .$$

Proof: Because $\mathbf{V}(X_1) < \infty$, we have:

$$\begin{aligned} \mathbf{P}(|\overline{X}_n - \mathbf{E}(\overline{X}_n)| \geq \epsilon) &= \frac{\mathbf{V}(\overline{X}_n)}{\epsilon^2} && \text{[by applying Chebychev's inequality (5.6) to the RV } \overline{X}_n\text{]} \\ &= \frac{\frac{1}{n}\mathbf{V}(X_1)}{\epsilon^2} && \text{[by the IID assumption of } X_1, X_2, \ldots \text{ we can apply (3.77)]} \end{aligned}$$

[2] https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_almost_surely_implies_convergence_in_probability

[3] https://en.wikipedia.org/wiki/Borel%E2%80%93Cantelli_lemma

[4] https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_in_probability_does_not_imply_almost_sure_convergence_in_the_discrete_case

[5] https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_in_probability_implies_convergence_in_distribution

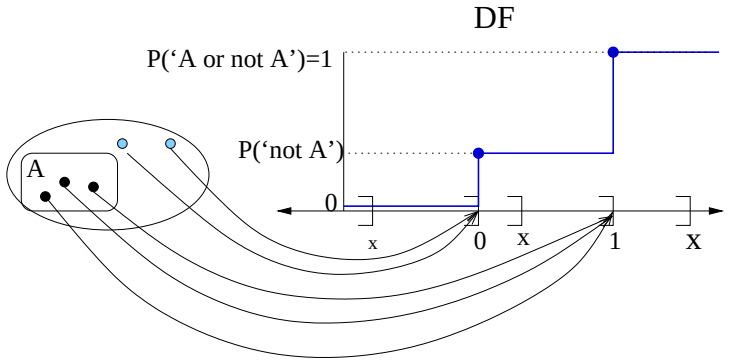
[6] https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_in_distribution_to_a_constant_implies_convergence_in_probability

Some useful properties of the Indicator Function are:

$$\mathbb{1}_{A^c} = 1 - \mathbb{1}_A, \qquad \mathbb{1}_{A \cap B} = \mathbb{1}_A \mathbb{1}_B, \qquad \mathbb{1}_{A \cup B} = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_A \mathbb{1}_B$$

We slightly abuse notation when A is a single element set by ignoring the curly braces.

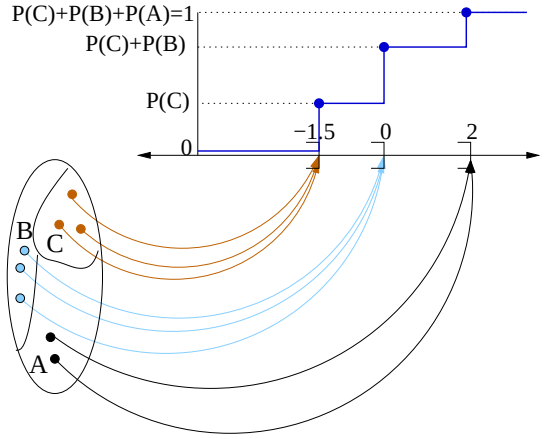
Figure 3.1: The Indicator function of event $A \in \mathcal{F}$ is a RV $\mathbb{1}_A$ with DF F



Exercise 3.1 (Drawing discontinuous functions) Identify the mistakes in how the $\mathbb{1}_A$ is drawn as a discontinous function in Figure 3.1.

Classwork 41 (A random variable with three values and eight sample points) Consider the RV X of Figure 3.2. First draw this properly as done in Ex. 3.1. Let the events $A = \{\omega_1, \omega_2\}$, $B = \{\omega_3, \omega_4, \omega_5\}$ and $C = \{\omega_6, \omega_7, \omega_8\}$. Define the RV X formally. What sets should $\mathcal{F}$ minimally include? What do you need to do to make sure that $\mathcal{F}$ is a sigma algebra?

Figure 3.2: A RV X from a sample space Ω with 8 elements to $\mathbb{R}$ and its DF F.



Exercise 3.2 (Fair coin toss RV) Consider the *fair coin toss experiment* with $\Omega = \{\mathsf{H}, \mathsf{T}\}$ and $P(\mathsf{H}) = P(\mathsf{T}) = 1/2$.

We can associate a Bernoulli random variable X (in Model 1) for the event that the coin lands as H, with this experiment as follows:

$$X(\omega) = \begin{cases} 1, & \text{if } \omega = \mathsf{H} \\ 0, & \text{if } \omega = \mathsf{T} \end{cases}$$

Find the distribution function for X.

3.2 Discrete Random Variables

When a RV takes at most countably many values from a discrete set $\mathbb{X}$, we call it a **discrete** RV. Recall that a set $\mathbb{X}$ is said to be discrete if we can enumerate its elements, i.e., find an enumerating or counting function $\mathbb{X} \ni x \mapsto i \in \mathbb{N}$ that associates each element $x \in \mathbb{X}$ to a natural number $i \in \mathbb{N}$. So, $\mathbb{X}$ is either finite with k elements in $\mathbb{X} = \{x_1, x_2, \ldots, x_k\}$ or countably infinite with the same cardinality as $\mathbb{N}$ with $\mathbb{X} = \{x_1, x_2, \ldots\}$. When $\mathbb{X} \subset \mathbb{R}$, we have a real-valued or $\mathbb{R}$-valued discrete random variable.

Definition 23 (probability mass function (PMF)) Let X be a $\mathbb{R}$-valued discrete RV over a probability triple $(\Omega, \mathcal{F}, \mathbf{P})$. We define the **probability mass function** (PMF) f of X to be the function $f : \mathbb{R} \to [0, 1]$ defined as follows:

$$f(x) := \mathbf{P}(X = x) = \mathbf{P}(\ \{\omega : X(\omega) = x\}\ ) = \begin{cases} \theta_i & \text{if } x = x_i \in \mathbb{X}. \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

The DF F and PMF f for a discrete RV X satisfy the following:

1. For any $x \in \mathbb{R}$,
$$\boxed{\mathbf{P}(X \leq x) = F(x) = \sum_{x_i \leq x} f(x_i) = \sum_{x_i \leq x} \theta_i\ .} \tag{3.6}$$
2. For any $a, b \in \mathbb{R}$ with $a < b$,
$$\boxed{\mathbf{P}(a < X \leq b) = F(b) - F(a) = \sum_{a < x_i \leq b} \theta_i\ .} \tag{3.7}$$
This is just the sum of all probabilities θ_i for which x_i satisfies $a < x_i \leq b$.
3. From the fact that $\mathbf{P}(\Omega) = 1$, we get that the sum of all the probabilties is 1:
$$\boxed{\sum_i \theta_i = 1\ .} \tag{3.8}$$

Proposition 66 (Chebychev's Inequality) For **any** RV X and any $\epsilon > 0$,

$$\mathbf{P}(|X| > \epsilon) \leq \frac{\mathbf{E}(|X|)}{\epsilon} \tag{5.4}$$

$$\mathbf{P}(|X| > \epsilon) = \mathbf{P}(X^2 \geq \epsilon^2) \leq \frac{\mathbf{E}(X^2)}{\epsilon^2} \tag{5.5}$$

$$\mathbf{P}(|X - \mathbf{E}(X)| \geq \epsilon) = \mathbf{P}((X - \mathbf{E}(X))^2 \geq \epsilon^2) \leq \frac{\mathbf{E}(X - \mathbf{E}(X))^2}{\epsilon^2} = \frac{\mathbf{V}(X)}{\epsilon^2} \tag{5.6}$$

Proof: All three forms of Chebychev's inequality are mere corollaries (careful reapplications) of Markov's inequality.

Armed with Markov's inequality we next enquire the convergence in probability for the sequence of RVs in Classwork 157 and Example 158.

Example 160 (Convergence in probability) Does the the sequence of RVs $\{X_n\}_{n=1}^{\infty}$, where $X_n \sim \text{Normal}(0, 1/n)$, converge in probability to $X \sim \text{Point Mass}(0)$, i.e. does $X_n \xrightarrow{\mathbf{P}} X$?

To find out if $X_n \xrightarrow{\mathbf{P}} X$, we need to show that for any $\epsilon > 0$, $\lim_{n\to\infty} \mathbf{P}(|X_n - X| > \epsilon) = 0$.

Let ϵ be any real number greater than 0, then

$$\begin{aligned} \mathbf{P}(|X_n| > \epsilon) &= \mathbf{P}(|X_n|^2 > \epsilon^2) \\ &\leq \frac{\mathbf{E}(X_n^2)}{\epsilon^2} \qquad \text{[by Markov's Inequality (5.2)]} \\ &= \frac{\frac{1}{n}}{\epsilon^2} \to 0, \quad \text{as} \quad n \to \infty \qquad \text{[in the sense of Definition 4].} \end{aligned}$$

Hence, we have shown that for any $\epsilon > 0$, $\lim_{n\to\infty} \mathbf{P}(|X_n - X| > \epsilon) = 0$ and therefore by Definition 64, $X_n \xrightarrow{\mathbf{P}} X$ or $X_n \xrightarrow{\mathbf{P}} 0$.

Convention: When X has a Point Mass(θ) distribution and $X_n \xrightarrow{\mathbf{P}} X$, we simply write $X_n \xrightarrow{\mathbf{P}} \theta$.

Definition 67 (Convergence Almost Surely (or with Probability 1)) To say that the sequence of RVs $\{X_n\}_{n=1}^{\infty}$ converges almost surely (or with probability 1 or strongly) towards another RV X on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$, as denoted by

$$X_n \overset{a.s.}{\to} X$$

means that

$$\mathbf{P}\left(\{\lim_{n\to\infty} X_n = X\}\right) = 1 \quad \iff \quad \mathbf{P}\left(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right) = 1.$$

This means that the values of X_n approach the value of X, in the sense that events for which X_n does not converge to X have probability 0.

Other notions of convergence are termed sure convergence or pointwise convergence, such as convergence in mean. But the above three types of convergence are elementary.

As $n \to \infty$, the expression below the first overbrace $\to 1$, while that below the second overbrace, being independent of n remains the same. By the elementary examples of limits 17 and 18, as $n \to \infty$, the expression over the first underbrace approaches $e^{-\lambda}$ while that over the second underbrace approaches 1. Finally, we get the desired limit:

$$\lim_{n\to\infty} \mathbf{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \ .$$

The second notion of convergence of RVs is convergence in probability.

Definition 64 (Convergence in Probability) Let $X_1, X_2, \ldots,$ be a sequence of RVs and let X be another RV. Let F_n denote the DF of X_n and F denote the DF of X. The we say that X_n converges to X in probability, and write:

$$X_n \xrightarrow{\mathbf{P}} X$$

if for every real number $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}(|X_n - X| > \epsilon) = 0 \qquad \text{[in the sense of Definition 4]}.$$

Once again, the above limit, by (3.1) in our Definition 18 of a RV, can be equivalently expressed as follows:

$$\lim_{n\to\infty} \mathbf{P}(\ \{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}\ ) = 0, \qquad \text{ie,} \qquad \mathbf{P}(\ \{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}\ ) \to 0, \quad \text{as} \quad n \to \infty \ .$$

For the same sequence of RVs in Classwork 157 and Example 158 we are tempted to ask whether $X_n \sim \text{Normal}(0, 1/n)$ converges in probability to $X \sim \text{Point Mass}(0)$, i.e. whether $X_n \xrightarrow{\mathbf{P}} X$. We need some elementary inequalities in Probability to help us answer this question. We visit these inequalities next.

Proposition 65 (Markov's Inequality) Let $(\Omega, \mathcal{F}, P)$ be a probability triple and let $X = X(\omega)$ be a non-negative RV. Then,

$$\mathbf{P}(X \geq \epsilon) \leq \frac{\mathbf{E}(X)}{\epsilon}, \qquad \text{for any} \quad \epsilon > 0 \ . \tag{5.2}$$

Proof:

$$\begin{aligned} X &= X\mathbb{1}_{\{y:y\geq\epsilon\}}(x) + X\mathbb{1}_{\{y:y<\epsilon\}}(x) \\ &\geq X\mathbb{1}_{\{y:y\geq\epsilon\}}(x) \\ &\geq \epsilon\mathbb{1}_{\{y:y\geq\epsilon\}}(x) \end{aligned} \tag{5.3}$$

Finally, taking expectations on both sides of the above inequality and then using the fact that the expectation of an indicator function of an event is simply the probability of that event (3.46), we get the desired result:

$$\mathbf{E}(X) \geq \epsilon\mathbf{E}(\mathbb{1}_{\{y:y\geq\epsilon\}}(x)) = \epsilon\mathbf{P}(X \geq \epsilon) \ .$$

Let us look at some immediate consequences of Markov's inequality.

4. When X only has finitely many possibilities, say k with $\mathbb{X} = \{x_1, x_2, \ldots, x_k\}$, then we may think of the probability $\mathbf{P}$ specified by $(\theta_1, \theta_2, \ldots, \theta_k)$ as a point in the **unit** $(k-1)$ **simplex**:

$$\Delta^{k-1} := \{(\theta_1, \theta_2, \ldots, \theta_k) \in \mathbb{R}^k : \sum_i \theta_i = 1 \text{ and } \theta_i \geq 0, \text{ for all } i\} \tag{3.9}$$

In particular when X has only two possible values with $\mathbb{X} = \{x_1, x_2\}$ then $\theta_2 = 1 - \theta_1$, so we can avoid subscripts and take $\theta := \theta_1$ and realize that the probability $\mathbf{P}$ is now specified by the point $(\theta, 1-\theta)$ in the **unit** 1 **simplex**:

$$\Delta^1 := \{(\theta, 1-\theta) \in \mathbb{R}^2 : 0 \leq \theta \leq 1\} \ . \tag{3.10}$$

See `https://en.wikipedia.org/wiki/Simplex` for the images scribed on the board.

DISCRETE RANDOM VARIABLES - SIMPLIFIED NOTATION

Notice that in equations (3.5), (3.6) and (3.7) the use of the "$\omega \in \Omega$" notation, where random variables are defined as functions, is much reduced. The reason is that in straightforward examples it is convenient to associate the possible values $x_1, x_2, \ldots$ with the outcomes $\omega_1, \omega_2, \ldots$ Hence, we can describe a discrete random variable by the table:

Possible values: x_i	x_1	x_2	x_3	$\ldots$
Probability: $\mathbf{P}(X = x_i) = \theta_i$	θ_1	θ_2	θ_3	$\ldots$

It is customary to use p_i instead of θ_i for the probabilities. But we try to avoid it as it will hurt us when we start doing Inference Theory soon!
Note that this table hides the more complex notation but it is still there, under the surface. In Probability Theory I, you should be able to work with and manipulate discrete random variables using the simplified notation given above. The same comment applies to the continuous random variables discussed later. But you are students of mathematics and should know more about what is "under the hood".

Out of the class of discrete random variables we will define specific kinds as they arise often in applications. We classify discrete random variables into three types for convenience as follows:

- Discrete uniform random variables with finitely many possibilities
- Discrete non-uniform random variables with finitely many possibilities
- Discrete non-uniform random variables with (countably) infinitely many possibilities

Model 2 (Discrete Uniform) We say that a discrete random variable X is uniformly distributed over k possible values in $\mathbb{X} = \{x_1, x_2, \ldots, x_k\}$ if its probability mass function is:

$$f(x) = \begin{cases} \theta_i = \frac{1}{k} & \text{if } x = x_i, \quad \text{where } i = 1, 2, \ldots, k \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{3.11}$$

The distribution function for the discrete uniform random variable X is:

$$F(x) = \sum_{x_i \le x} f(x_i) = \sum_{x_i \le x} \theta_i = \begin{cases} 0 & \text{if } -\infty < x < x_1 \ , \\ \frac{1}{k} & \text{if } x_1 \le x < x_2 \ , \\ \frac{2}{k} & \text{if } x_2 \le x < x_3 \ , \\ \vdots & \\ \frac{k-1}{k} & \text{if } x_{k-1} \le x < x_k \ , \\ 1 & \text{if } x_k \le x < \infty \ . \end{cases} \tag{3.12}$$

The discrete uniform RV with values in $\mathbb{X} = \{1, 2, \ldots, k\}$ is called the equi-probable de Moivre(k) RV as we will see in the sequel.

Example 42 The *fair coin toss experiment* of Exercise 3.2 is an example of a discrete uniform random variable with finitely many possibilities. Its probability mass function is given by

$$f(x) \ = \ \mathbf{P}(X = x) \ = \ \begin{cases} \frac{1}{2} & \text{if } x = 0 \\ \frac{1}{2} & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

and its distribution function is given by

$$F(x) \ = \ \mathbf{P}(X \le x) \ = \ \begin{cases} 0, & \text{if } -\infty < x < 0 \\ \frac{1}{2}, & \text{if } 0 \le x < 1 \\ 1, & \text{if } 1 \le x < \infty \end{cases}$$

Let us sketch the probability mass function and distribution function for X below.

Figure 3.3: $f(x)$ and $F(x)$ of *the fair coin toss* random variable X, a discrete uniform RV on $\{0, 1\}$.

Example 43 (Fair dice RV) Now consider the *toss a fair die* experiment and define X to be the number that shows up on the top face. Note that here Ω is the set of numerical symbols $\{\mathsf{1}, \mathsf{2}, \mathsf{3}, \mathsf{4}, \mathsf{5}, \mathsf{6}\}$ that label each face while each of these symbols are associated with the real number $x \in \{1, 2, 3, 4, 5, 6\}$. We can describe this random variable by the table

Possible values, x_i	1	2	3	4	5	6
Probability, θ_i	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

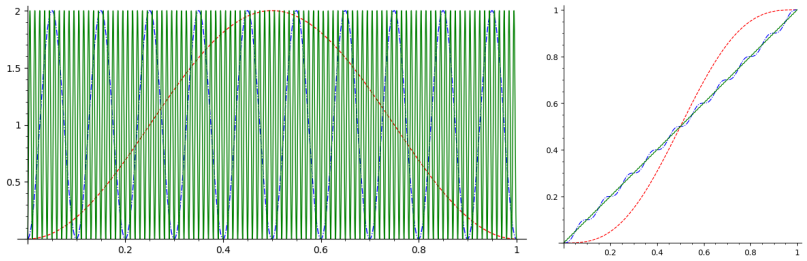


Figure 5.3: PDF $f_{X_n}(x) := \mathbb{1}_{(0,1)}(x)(1 - \cos(2\pi n x))$ of the RV X_n [the left sub-figure] and its DF $F_n(x) := \int_{-\infty}^{x} \mathbb{1}_{(0,1)}(v)(1 - \cos(2\pi n v))dv$ [the right sub-figure], for $n = 1$ [red '- -'], $n = 10$ [blue '-.'], and $n = 100$ [green '-'], respectively. One can see clear convergence of the DFs F_n to $\mathbb{1}_{(0,1)}(x)x$, the DF of the Uniform$(0, 1)$ RV, while the corresponding PDFs $f_n(x)$ keep oscillating wildly with n across $[0, 2]$ about $\mathbb{1}_{(0,1)}(x)$, the PDF of the Uniform$(0, 1)$ RV X. Thus giving a counter-example to the claim that convergence in DFs does not imply convergence in PDFs.

Since $F(x) = \mathbf{P}(X \le x)$, convergence in distribution means that the probability for X_n to be in a given range is approximately equal to the probability that the value of the limiting RV X is in that range, provided n is sufficiently large.

Thus, for a discrete sequence of RVs X_n'n to converge in distribution to another discrete RV X taking values in $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$, it is sufficient to show that $\lim_{n\to\infty} \mathbf{P}(X_n = x) = \mathbf{P}(X = x)$ for each $x \in \mathbb{Z}_+$. We will use this fact to prove why we can approximate Binomial RVs by a Poisson under some limiting conditions.

Example 159 (Binomial$(n, \lambda/n) \rightsquigarrow$Poisson$(\lambda)$**)** In several situations, as we saw already, it becomes cumbersome to model the events using the Binomial(n, θ) RV, especially when when the parameter $\theta \propto 1/n$ and the events become rare.

Binomial$(n, \lambda/n)$ converges in distribution to Poisson(λ) as $n \to \infty$, $\theta = \lambda/n \to 0$

However, for some real parameter $\lambda > 0$, the Binomial$(n, \lambda/n)$ RV with probability of the number of successes in n trials, with per-trial success probability λ/n, approaches the Poisson distribution with expectation λ, as n approaches ∞ (actually, it converges in distribution). The Poisson(λ) RV is much simpler to work with than the combinatorially laden Binomial$(n, \theta = \lambda/n)$ RV. We sketch the details of this next.

Let $X_n \sim$ Binomial$(n, \theta = \lambda/n)$ and $Y \sim$ Poisson(λ) and let $\lambda = n\theta$ remain constant as $n \to \infty$, $\theta \to 0$. We need to show that $\lim_{n\to\infty} \mathbf{P}(X_n = x) = \mathbf{P}(Y = x) = e^{-\lambda}\lambda^x/x!$ for any $x \in \{0, 1, 2, 3, \ldots, n\}$.

$$\begin{aligned} \mathbf{P}(X = x) &= \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \frac{n(n-1)(n-2)\cdots(n-x+1)}{x(x-1)(x-2)\cdots(2)(1)} \left(\frac{\lambda^x}{n^x}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \\ &= \underbrace{\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-x+1}{n}\right)\left(\frac{\lambda^x}{x!}\right)}\underbrace{\left(1 - \frac{\lambda}{n}\right)^n}\underbrace{\left(1 - \frac{\lambda}{n}\right)^{-x}} \end{aligned} \tag{5.1}$$

We can formalize our observation in Classwork 157 that X_n is concentrating about 0 as $n \to \infty$ by the statement:

$$X_n \text{ is converging in distribution to } X, \text{ ie}, \qquad X_n \rightsquigarrow X \ .$$

Proof: To check that the above statement is true we need to verify that the definition of convergence in distribution is satisfied for our sequence of RVs $X_1, X_2, \ldots$ and the limiting RV X. Thus, we need to verify that for any continuity point t of the Point Mass(0) DF F, $\lim_{n\to\infty} F_n(t) = F(t)$. First note that

$$X_n \sim \text{Normal}(0, 1/n) \implies Z := \sqrt{n}X_n \sim \text{Normal}(0, 1) \ ,$$

and thus

$$F_n(t) = \mathbf{P}(X_n < t) = \mathbf{P}(\sqrt{n}X_n < \sqrt{n}t) = \mathbf{P}(Z < \sqrt{n}t) \ .$$

The only discontinuous point of F is 0 where F jump from 0 to 1.

When $t < 0$, $F(t)$, being the constant 0 function over the interval $(-\infty, 0)$, is continuous at t. Since $\sqrt{n}t \to -\infty$, as $n \to \infty$,

$$\lim_{n\to\infty} F_n(t) = \lim_{n\to\infty} \mathbf{P}(Z < \sqrt{n}t) = 0 = F(t) \ .$$

And, when $t > 0$, $F(t)$, being the constant 1 function over the interval $(0, \infty)$, is again continuous at t. Since $\sqrt{n}t \to \infty$, as $n \to \infty$,

$$\lim_{n\to\infty} F_n(t) = \lim_{n\to\infty} \mathbf{P}(Z < \sqrt{n}t) = 1 = F(t) \ .$$

Thus, we have proved that $X_n \rightsquigarrow X$ by verifying that for any t at which the Point Mass(0) DF F is continuous, we also have the desired equality: $\lim_{n\to\infty} F_n(t) = F(t)$.

However, note that

$$F_n(0) = \frac{1}{2} \neq F(0) = 1 \ ,$$

and so convergence fails at 0, i.e. $\lim_{n\to\infty} F_n(t) \neq F(t)$ at $t = 0$. But, $t = 0$ is not a continuity point of F and the definition of convergence in distribution only requires the convergence to hold at continuity points of F.

Convergence in distribution does not in general imply that the sequence of corresponsing probability density functions will also converge. Consider for example RV X_n with density $\mathbb{1}_{(0,1)}(x)(1 - \cos(2\pi n x))$. These RVs converge in distribution to $X \sim \text{Uniform}(0, 1)$, but their densities (PDFs) do not converge at all as evident in Figure 5.3.

Proposition 63 (Scheffé's Theorem) According to **Scheffé's Theorem** convergence of the probability density function (for a continuous RV) or probability mass function (for a discrete RV) implies convergence in distribution.

Proof: We will state this without a Proof here as Proof of the Theorem requires measure theory in generality [1]. However, you should be able to see why convergence of PMFs $f_n(x)$ for discrete RVs X_n, to $f(x)$, the PMF of another discrete RV X, implies convergence in their corresponding DFs, i.e., $F_n(x) \to F(x)$ for each x as $n \to \infty$.

[1] See https://en.wikipedia.org/wiki/Scheff%C3%A9%27s_lemma.

Find the probability mass function and distribution function for this random variable, and sketch their graphs.

Solution:

The probability mass function of this random variable is:

$$f(x) \ = \ \mathbf{P}(X = x) \ = \ \begin{cases} \frac{1}{6} & \text{if } x = 1 \\ \frac{1}{6} & \text{if } x = 2 \\ \frac{1}{6} & \text{if } x = 3 \\ \frac{1}{6} & \text{if } x = 4 \\ \frac{1}{6} & \text{if } x = 5 \\ \frac{1}{6} & \text{if } x = 6 \\ 0 & \text{otherwise} \end{cases}$$

and the distribution function is:

$$F(x) \ = \ \mathbf{P}(X \leq x) \ = \ \begin{cases} 0, & \text{if } -\infty < x < 1 \\ \frac{1}{6}, & \text{if } 1 \leq x < 2 \\ \frac{1}{3}, & \text{if } 2 \leq x < 3 \\ \frac{1}{2}, & \text{if } 3 \leq x < 4 \\ \frac{2}{3}, & \text{if } 4 \leq x < 5 \\ \frac{5}{6}, & \text{if } 5 \leq x < 6 \\ 1, & \text{if } 6 \leq x < \infty \end{cases}$$

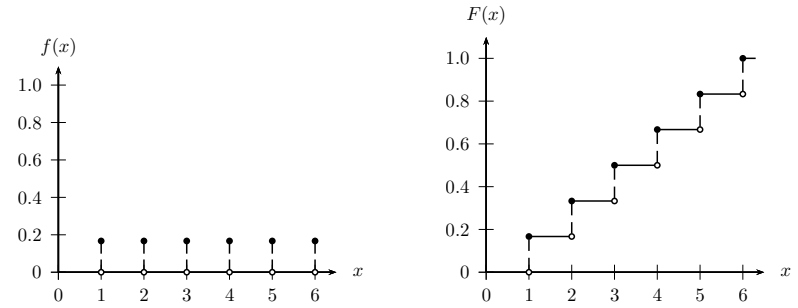


Figure 3.4: $f(x)$ and $F(x)$ of *the fair die toss* random variable X, a discrete uniform RV on $\{1, 2, 3, 4, 5, 6\}$.

Example 44 (Astragali with a Kiwi sheep ankle bone) Astragali. Board games involving chance were known in Egypt, 3000 years before Christ. The element of chance needed for these games was at first provided by tossing astragali, the ankle bones of sheep. These bones could come to rest on only four sides, the other two sides being rounded. The upper side of the bone, broad

and slightly convex counted four; the opposite side broad and slightly concave counted three; the lateral side flat and narrow, one, and the opposite narrow lateral side, which is slightly hollow, six. You may examine an astragali of a kiwi sheep.

This is an example of a discrete non-uniform random variable with finitely many possibilities. A surmised probability mass function with $f(4) = \frac{4}{10}$, $f(3) = \frac{3}{10}$, $f(1) = \frac{2}{10}$, $f(6) = \frac{1}{10}$ and distribution function are shown below.

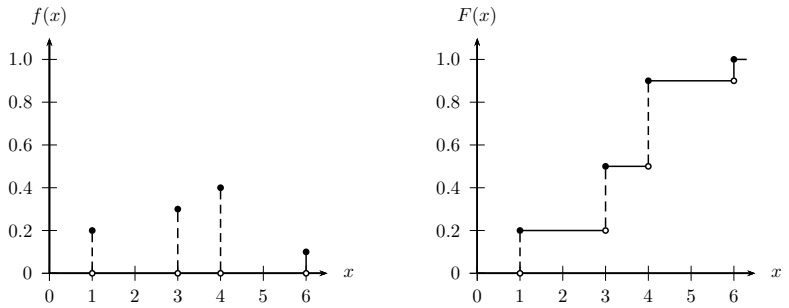


Figure 3.5: $f(x)$ and $F(x)$ of surmised *astragali toss* random variable X, a discrete (non-uniform) RV on $\{1, 2, 3, 4\}$.

3.2.1 An Elementary Family of Bernoulli Random Variables

In many experiments there are only two outcomes. For instance:

- Flip a coin to see whether it is defective.
- Roll a die and determine whether it is a 6 or not.
- Determine whether it will be below 0 degrees Celsius at 0600 hours in Uppsala tomorrow or not.

Performing such an experiment $\mathcal{E}$ once to see if an event of interest A occurs is called a **Bernoulli trial** and its probability model over a triple $(\Omega, \mathcal{F}, \mathbf{P})$, with $A \in \mathcal{F}$, given by the Indicator Function $\mathbb{1}_A$ in Model 1 is called the Bernoulli RV.

If we do not know the probability θ that 'A occurs', i.e., the Bernoulli RV will equal 1, then we can define a whole family of Bernoulli RVs for each $\theta \in [0, 1]$ or more precisely for each $(\theta, 1-\theta) \in \Delta^1$, the unit 1-Simplex. Note that this family includes the fair Bernoulli trial of Example 42 when $\theta = 0.5$. Let us formalise this as the Bernoulli(θ) RV for each $\theta \in [0, 1]$ next.

Model 3 (Bernoulli(θ) **RV)** Given a parameter $\theta \in [0, 1]$, the probability mass function (PMF) for the Bernoulli(θ) RV X is:

$$f(x;\theta) = \theta^x (1-\theta)^{1-x} \mathbb{1}_{\{0,1\}}(x) = \begin{cases} \theta & \text{if } x = 1, \\ 1-\theta & \text{if } x = 0, \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

Figure 5.1: Sequence of $\{\text{Point Mass}(17)\}_{i=1}^{\infty}$ RVs (left panel) and $\{\text{Point Mass}(1/i)\}_{i=1}^{\infty}$ RVs (only the first seven are shown on right panel) and their limiting RVs in red.

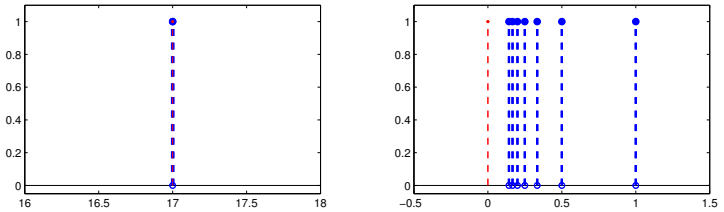
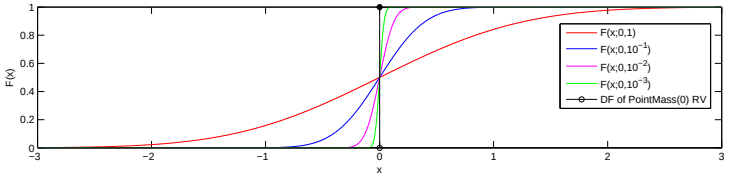


Figure 5.2: Distribution functions of several Normal(μ, σ^2) RVs for $\sigma^2 = 1, \frac{1}{10}, \frac{1}{100}, \frac{1}{1000}$.



Thus, we need more sophisticated notions of convergence for sequences of RVs. Two such notions are formalized next as they are minimal prerequisites for a clear understanding of two basic propositions in Statistics :

1. Law of Large Numbers,
2. Central Limit Theorem,

Definition 62 (Convergence in Distribution (or Weakly, or in Law)) Let $X_1, X_2, \ldots,$ be a sequence of RVs and let X be another RV. Let F_n denote the DF of X_n and F denote the DF of X. The we say that X_n converges to X in distribution, and write:

$$X_n \rightsquigarrow X$$

if for any real number t at which F is continuous,

$$\lim_{n\to\infty} F_n(t) = F(t) \qquad \text{[in the sense of Definition 4]}.$$

The above limit, by (3.2) in our Definition 19 of a DF, can be equivalently expressed as follows:

$$\lim_{n\to\infty} \mathbf{P}(\ \{\omega : X_n(\omega) \le t\}\ ) = \mathbf{P}(\ \{\omega : X(\omega) \le t\}\ ),$$
$$\text{i.e.} \quad \mathbf{P}(\ \{\omega : X_n(\omega) \le t\}\ ) \to \mathbf{P}(\ \{\omega : X(\omega) \le t\}\ ), \quad \text{as} \quad n \to \infty\ .$$

Let us revisit the problem of convergence in Classwork 157 armed with our new notions of convergence.

Example 158 (Convergence in distribution) Suppose you are given an independent sequence of RVs $\{X_i\}_{i=1}^n$, where $X_i \sim \text{Normal}(0, 1/i)$ with DF F_n and let $X \sim$ Point Mass(0) with DF F.

Chapter 5

Limit Laws of Statistics

5.1 Convergence of Random Variables

This important topic is concerned with the limiting behavior of sequences of RVs. We want to understand what it means for a sequence of random variables $\{X_n\}_{n=1}^{\infty} := X_1, X_2, \ldots$ to converge to another random variable X, when all RVs are defined on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

$$\{X_i\}_{i=1}^{n} := X_1, X_2, X_3, \ldots X_{n-1}, X_n \qquad \text{as } n \to \infty \ .$$

From a statistical or decision-making viewpoint, as you will see in Inference Theory I course, $n \to \infty$ is associated with the amount of data or information $\to \infty$. More abstractly, we are interested in what happens to the limiting RV $X := \lim_{n\to\infty} X_n$ when given the DFs $F_n(x)$ for each X_n.

We need different notions of convergence to characterize such a behavior: two simplest behaviors are that the sequence eventually takes a constant value θ, i.e. X_n approaches $X \sim$ Point Mass(θ) RV, or that values in the sequence continue to change but can be described by an unchanging probability distribution, i.e., X_n approaches $X \sim F(x)$. See `https://en.wikipedia.org/wiki/Convergence_of_random_variables`.

Let us first refresh ourselves with notions of convergence, limits and continuity in the real line (Sec. 1.8.1) before proceeding further.

Can the sequences of $\{\text{Point Mass}(\theta_i = 17)\}_{i=1}^{\infty}$ and $\{\text{Point Mass}(\theta_i = 1/i)\}_{i=1}^{\infty}$ RVs be the same as the two sequences of real numbers $\{x_i\}_{i=1}^{\infty} = 17, 17, 17, \ldots$ and $\{x_i\}_{i=1}^{\infty} = \frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \ldots$ we saw in Examples 12 and 13?

Yes why not – just move to space of distributions over the reals! See Figure 5.1.

Classwork 157 (Convergence of $X_i \sim \text{Normal}(0, 1/i)$) Suppose you are given an independent sequence of RVs $\{X_i\}_{i=1}^{n}$, where $X_i \sim \text{Normal}(0, 1/i)$. How would you talk about the convergence of $X_n \sim \text{Normal}(0, 1/n)$ as n approaches ∞ ? Take a look at Figure 5.2 for insight. The probability mass of X_n increasingly concentrates about 0 as n approaches ∞ and the variance $1/n$ approaches 0, as depicted in Figure 5.2. Based on this observation, can we expect $\lim_{n\to\infty} X_n = X$, where the limiting RV $X \sim$ Point Mass(0) ?

The answer is **no**. This is because $\mathbf{P}(X_n = X) = 0$ for any n, since $X \sim$ Point Mass(0) is a discrete RV with exactly one outcome 0 and $X_n \sim \text{Normal}(0, 1/n)$ is a continuous RV for every n, however large. In other words, a continuous RV, such as X_n, has 0 probability of realizing any single real number in its support, such as 0.

and its DF is:

$$F(x;\theta) = \begin{cases} 1 & \text{if } 1 \le x, \\ 1-\theta & \text{if } 0 \le x < 1, \\ 0 & \text{otherwise} \end{cases} \tag{3.14}$$

We emphasise the dependence of the probabilities on the parameter θ by specifying it following the semicolon in the argument for f and F and by subscripting the probabilities, i.e. $\mathbf{P}_\theta(X = 1) = \theta$ and $\mathbf{P}_\theta(X = 0) = 1 - \theta$.

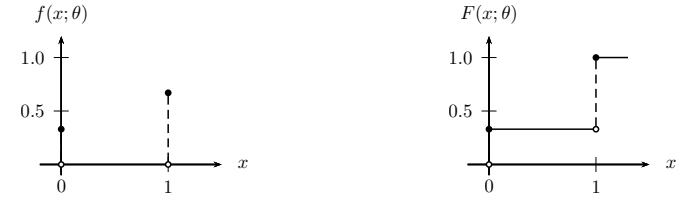


Figure 3.6: PMF $f(x;\theta)$ and DF $f(x;\theta)$ with $\theta = 0.33$. You should see how PMF and DF change as θ goes from 0 to 1

3.2.2 Independent Bernoulli Trials

Random variables make sense for a series of trials as well as just a single trial of an experiment. We now look at what happens when we perform a sequence of independent Bernoulli trials. For instance:

- Flip a coin 10 times; count the number of heads
  - by possibly allowing for the coin's $\mathbf{P}(\mathsf{H})$ to change each time because each of them are manufactured in a terrible mint.
- Test 50 randomly selected circuits from an assembly line; count the number of defective circuits.
- Roll a die 100 times; count the number of sixes you throw.
- Provide a property near a particular bridge in our archipelago with flood insurance for 20 years; count the number of years, during the 20-year period, during which the property is flooded. Note: we assume that flooding is independent from year to year, and that the probability of flooding is the same each year.

Since the Bernoulli(θ) RV has only two outcomes, i.e., simple events, we know how to obtain the probability of each of the two outcomes in a given Bernoulli trial with the probability given by the deterministic variable or parameter θ. Now consider doing more than one trial so we have sequence of Bernoulli(θ_i) trials, say,

$$X_i \sim \text{Bernoulli}(\theta_i) \text{ with } i \in \mathbb{N} \ ,$$

with each $\theta_i \in [0, 1]$ being possibly unkown but fixed as a parameter. Now, if we assume independence across trials, so one trial's outcome does not affect the outcome of any of the other trials, in the sense of Definition 17 about *independence of a sequence of events*, then we can obtain the

probability of the entire sequence of outcomes for this sequence of **independently distributed** Bernoulli(θ_i) **trails** which can be any infinite sequence of 0's and 1's, i.e., any element of $\{0,1\}^\infty$, by simply multiplying the corresponding probabilities given by θ_i's in $(\theta_1, \theta_2, \ldots) \in [0,1]^\infty$, an infinite dimensional parameter space, as follows:

$$\mathbf{P}(x; (\theta_1, \theta_2, \ldots)) = \prod_i f(x_i; \theta_i) = \prod_i \theta_i^{x_i} (1-\theta_i)^{1-x_i} \mathbb{1}_{\{0,1\}}(x_i), \quad (3.15)$$
$$\text{where } x := (x_1, x_2, \ldots) \in \mathbb{X}_\infty = \{0,1\}^\infty := \{0,1\} \times \{0,1\} \times \cdots$$

By further assuming that all the θ_i's are identical, say $\theta = \theta_1 = \theta_2 = \cdots$, with $\theta \in [0,1]$, a one-dimensional parameter space, we get the much simpler expression for the **independent and identically distributed (IID)** Bernoulli(θ) **trails** as follows:

$$\begin{aligned} \mathbf{P}(x; \theta) &= \prod_i f(x_i; \theta) = \prod_i \theta^{x_i} (1-\theta)^{1-x_i} \mathbb{1}_{\{0,1\}}(x_i) = \mathbb{1}_{\{0,1\}^\infty}(x)\, \theta^{\sum_i x_i} (1-\theta)^{\sum_i (1-x_i)} \\ &= \begin{cases} \theta^{\sum_i x_i} (1-\theta)^{n - \sum_i x_i} & \text{if } x := (x_1, x_2, \ldots) \in \mathbb{X}_\infty = \{0,1\}^\infty \\ 0 & \text{otherwise .} \end{cases} \end{aligned} \quad (3.16)$$

Remembering that all other RVs can be derived from such IID Bernoulli(θ) trials using $\theta = 1/2$, as we will see in the sequel, we are ready to take a tour through some common discrete and continuous random variables that are useful in many applications.

3.2.3 Some Common Discrete Random Variables

Let us start with the simplest example to fix ideas carefully.

Example 45 (Waiting For the First Heads) Suppose our experiment is to toss a fair coin independently and identically (that is, the same coin is tossed in essentially the same manner independent of the other tosses in each trial) as often as necessary until we have a head, H. Let the random variable X denote the *Number of trials until the first* H *appears.*

Let's first find the probability mass function of X.

Now X can take on the values $\{1, 2, 3, \ldots\}$, so we have a non-uniform random variable with infinitely many possibilities. Since

$$\begin{aligned} f(1) &= \mathbf{P}(X=1) = P(\mathsf{H}) = \frac{1}{2}\ , \\ f(2) &= \mathbf{P}(X=2) = P(\mathsf{TH}) = \frac{1}{2}\cdot\frac{1}{2} = \left(\frac{1}{2}\right)^2\ , \\ f(3) &= \mathbf{P}(X=3) = P(\mathsf{TTH}) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \left(\frac{1}{2}\right)^3\ , \qquad \text{etc.} \end{aligned}$$

the probability mass function of X is:

$$f(x) = \mathbf{P}(X=x) = \left(\frac{1}{2}\right)^x, \quad x = 1, 2, \ldots\ .$$

In the previous Example, noting that we have independent trials, we get:

4. A related distribution, denoted by $N_-(\mu, \tau, \sigma^2)$, is the right-truncated normal distribution truncated on the right at τ. Describe how samples from $N_-(\mu, \tau, \sigma^2)$ can be obtained by simulating from an appropriate left-truncated normal distribution.

5. Write a MATLAB function that provides samples from a truncated normal distribution. The function should have the following inputs: number of samples required, left or right truncation, μ, σ^2 and τ.

4.4 Exercises in Simulation

Ex. 4.1 — Suppose the continuous RV X has PDF:

$$f_X(x) = \left(\pi(1+x^2)\right)^{-1}$$

Devise an algorithm to transform samples from Uniform$(0,1)$ RV to those from X. Present your answer as pseudo-code.

```
>> randn('state',67678); % initialise the seed at 67678 and method as Ziggurat -- TYPE help randn
>> randn % produce 1 sample from Normal(0,1) RV
ans =    1.5587
>> randn(2,8) % produce an 2 X 8 array of samples from Normal(0,1) RV
ans =
    1.2558    0.7834    0.6612    0.3247    0.1407    1.0562    0.8034    1.2970
   -0.5317    0.0417   -0.3454    0.6182   -1.4162    0.4796   -1.5015    0.3718
```

If we want to produce samples from $X \sim \text{Normal}(\mu, \sigma^2)$ with some user-specified μ and σ, then we can use the following relationship between X and $Z \sim \text{Normal}(0,1)$:

$$X \leftarrow \mu + \sigma Z, \qquad Z \sim \text{Normal}(0,1)\ .$$

Suppose we want samples from $X \sim \text{Normal}(\mu = \pi, \sigma^2 = 2)$, then we can do the following:

```
>> randn('state',679); % initialise the seed at 679 and method as Ziggurat -- TYPE help randn
>> mu=pi % set the desired mean parameter mu
mu =    3.1416
>> sigma=sqrt(2) % set the desired standard deviation parameter sigma
sigma =    1.4142
>> mu + sigma * randn(2,8) % produces a 2 X 8 array of samples from Normal(3.1416,1.4.42)
ans =
    1.3955    1.7107    3.9572    3.2618    6.1652    2.6971    2.4940    4.5928
    0.8442    4.7617    3.5397    5.0282    1.6139    5.0977    2.0477    2.3286
```

Labwork 156 (Sampling from truncated normal distributions) [Christian P. Robert, Simulation of truncated normal variables, Statistics and Computing (1995) 5, 121-125] Let $N_+(\mu, \tau, \sigma^2)$ denote the left-truncated normal distribution with truncation point τ and density given by

$$f(x|\mu, \tau, \sigma^2) = \frac{exp(-(x-\mu)^2/2\sigma^2)}{\sqrt{2\pi}\sigma[1 - \Phi((\tau-\mu)/\sigma)]} \mathbb{1}_{x \geq \tau}.$$

When $\tau < \mu$, the rejection sampler can readily be used to simulate from $N_+(\mu, \tau, \sigma^2)$ by simulating from Normal(μ, σ^2) until a number larger than τ is obtained. When $\tau > \mu$, however, this can be inefficient and increasingly so as τ gets further out into the right tail. In this case, a more efficient approach is to use the rejection sampler with the following translated exponential distribution as the proposal distribution:

$$g(y|\lambda, \tau) = \lambda exp(-\lambda(y-\tau)) \mathbb{1}_{y \geq \tau}.$$

1. Show that for simulating from $N_+(\mu = 0, \tau, \sigma^2 = 1)$ when $\tau \geq 0$, the best choice of λ that maximizes the expected acceptance probability for the rejection sampler is given by
$$\lambda = \frac{\tau + \sqrt{\tau^2 + 4}}{2}$$
2. Find the maximum expected acceptance probabilities for the following truncation points, $\tau = 0, 0.5, 1, 1.5, 2, 2.5$ and 3. What can you conclude about efficiency as τ gets further out into the right tail?
3. Describe how samples from $N_+(\mu, \tau, \sigma^2)$ can be obtained by simulating from $N_+(\mu = 0, \tau, \sigma^2 = 1)$ and using location-scale transformation.

$$f(x) \;=\; \mathbf{P}(X = x) \;=\; \mathbf{P}(\underbrace{\mathsf{TT}\ldots\mathsf{T}}_{\mathsf{n-1}}\mathsf{H}) \;=\; \mathbf{P}(\mathsf{T})^{x-1}\,\mathbf{P}(\mathsf{H}) \;=\; \left(\frac{1}{2}\right)^{x-1}\frac{1}{2}\,.$$

More generally, let there be two possibilities, success (S) or failure (F), with $\mathbf{P}(\mathsf{S}) = \theta$ and $\mathbf{P}(\mathsf{F}) = 1 - \theta$ so that:

$$\mathbf{P}(X = x) \;=\; \mathbf{P}(\underbrace{\mathsf{FF}\ldots\mathsf{F}}_{x-1}\mathsf{S}) \;=\; (1-\theta)^{x-1}\,\theta\,.$$

This is called a **geometric random variable** with "success probability" parameter θ. We can spot a geometric distribution because there will be *a sequence of independent trials with a constant probability of success. We are counting the number of trials until the first success appears.* Let us define this random variable formally next.

Model 4 (Geometric(θ) **RV)** Given a parameter $\theta \in (0,1)$, the PMF of the Geometric(θ) RV X is

$$f(x;\theta) = \begin{cases} \theta(1-\theta)^x & \text{if } x \in \mathbb{Z}_+ := \{0,1,2,\ldots\} \\ 0 & \text{otherwise} \end{cases} \tag{3.17}$$

It is straightforward to verify that $f(x;\theta)$ is indeed a PMF :

$$\sum_{x=0}^{\infty} f(x;\theta) = \sum_{x=0}^{\infty} \theta(1-\theta)^x = \theta\left(\frac{1}{1-(1-\theta)}\right) = \theta\left(\frac{1}{\theta}\right) = 1$$

The above equality is a consequence of the geometric series identity (3.18) with $a = \theta$ and $\vartheta := 1 - \theta$:

$$\sum_{x=0}^{\infty} a\vartheta^x = a\left(\frac{1}{1-\vartheta}\right), \text{ provided, } 0 < \vartheta < 1\ . \tag{3.18}$$

Proof:

$$a + a\vartheta + a\vartheta^2 + \cdots + a\vartheta^n = \sum_{0\leq x\leq n} a\vartheta^x = a + \sum_{1\leq x\leq n} a\vartheta^x = a + \vartheta \sum_{1\leq x\leq n} a\vartheta^{x-1} = a + \vartheta \sum_{0\leq x\leq n-1} a\vartheta^x = a + \vartheta \sum_{0\leq x\leq n} a\vartheta^x - a\vartheta^{n+1}$$

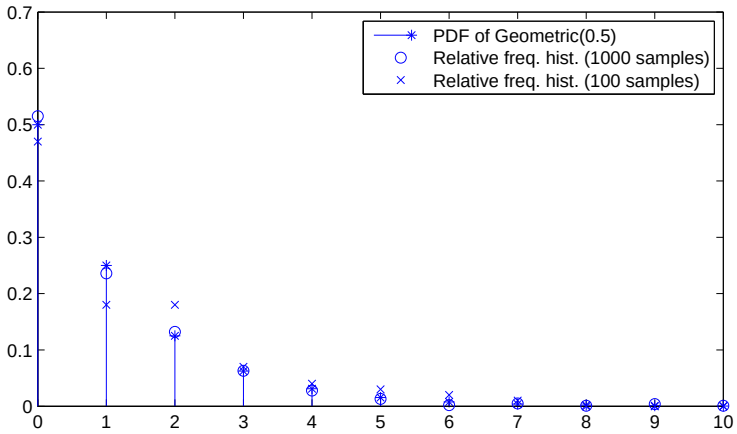
Therefore,

$$\begin{aligned}
\sum_{0\leq x\leq n} a\vartheta^x &= a + \vartheta \sum_{0\leq x\leq n} a\vartheta^x - a\vartheta^{n+1} \\
\left(\sum_{0\leq x\leq n} a\vartheta^x\right) - \left(\vartheta \sum_{0\leq x\leq n} a\vartheta^x\right) &= a - a\vartheta^{n+1} \\
\left(\sum_{0\leq x\leq n} a\vartheta^x\right)(1-\vartheta) &= a(1-\vartheta^{n+1}) \\
\sum_{0\leq x\leq n} a\vartheta^x &= a\left(\frac{1-\vartheta^{n+1}}{1-\vartheta}\right) \\
\sum_{x=0}^{\infty} a\vartheta^x := \lim_{n\to\infty} \sum_{0\leq x\leq n} a\vartheta^x &= a\left(\frac{1}{1-\vartheta}\right), \text{ provided, } 0 < \vartheta < 1
\end{aligned}$$

The outcome of a Geometric(θ) RV can be thought of as "the number of tosses needed before the appearance of the first 'Head' when tossing a coin with probability of 'Heads' equal to θ in a independent and identical manner."

Exercise 3.3 (Coupon Collector's Probem) Recall the Coupon Collector's Problem from lectures.

Figure 3.7: PMF of $X \sim$ Geometric($\theta = 0.5$) and the relative frequency histogram based on 100 and 1000 samples from X according to Simulation 144 and Labwork 145 you will see in the sequel.



Example 46 Suppose we flip a coin 10 times and count the number of heads. Let's consider the probability of getting three heads, say. The probability that the first three flips are heads and the last seven flips are tails, *in order*, is

$$\underbrace{\frac{1}{2}\frac{1}{2}\frac{1}{2}}_{\text{3 successes}}\ \underbrace{\frac{1}{2}\frac{1}{2}\cdots\frac{1}{2}}_{\text{7 failures}}.$$

But there are

$$\binom{10}{3} = \frac{10!}{7!\,3!} = 120$$

ways of ordering three heads and seven tails, so the probability of getting three heads and seven tails *in any order*, is

$$\mathbf{P}(\text{`3 heads'}) = \binom{10}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^7 \approx 0.117$$

We can describe this sort of situation by considering a random variable X which counts the number of successes, as follows:

Model 5 (Binomial(n, θ) **RV)** Let the RV $X = \sum_{i=1}^{n} X_i$ be the sum of n independent and identically distributed Bernoulli(θ) RVs, i.e.:

$$X = \sum_{i=1}^{n} X_i, \qquad X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta)\ .$$

Given two parameters n and θ, the PMF of the Binomial(n, θ) RV X is:

$$f(x; n, \theta) = \begin{cases} \binom{n}{x}\theta^x(1-\theta)^{n-x} & \text{if } x \in \{0, 1, 2, 3, \ldots, n\}\ , \\ 0 & \text{otherwise} \end{cases} \tag{3.19}$$

Proof: For the continuous case:

$$\mathbf{P}(\text{`accept } y\text{'}) = \mathbf{P}\left(u \le \frac{f(y)}{ag(y)}\right) = \int_{-\infty}^{\infty}\left(\int_0^{f(y)/ag(y)} du\right) g(y)dy = \int_{-\infty}^{\infty}\frac{f(y)}{ag(y)}g(y)dy = \frac{1}{a}.$$

And the number of proposals before acceptance is the Geometric$(1/a)$ RV with expectation $\frac{1}{1/a} = a$.

The closer $ag(x)$ is to $f(x)$, especially in the tails, the closer a will be to 1, and hence the more efficient the rejection method will be.

The rejection method can still be used only if the un-normalised form of f or g (or both) is known. In other words, if we use:

$$f(x) = \frac{\tilde{f}(x)}{\int \tilde{f}(x)dx} \text{ and } g(x) = \frac{\tilde{g}(x)}{\int \tilde{g}(x)dx}$$

we know only $\tilde{f}(x)$ and/or $\tilde{g}(x)$ in closed-form. Suppose the following are satisfied:

(a) we can generate random variables from g;

(b) the support of g contains the support of f, i.e. $\mathbb{Y} \supset \mathbb{X}$;

(c) a constant $\tilde{a} > 0$ exists, such that:

$$\tilde{f}(x) \le \tilde{a}\tilde{g}(x), \tag{4.7}$$

for any $x \in \mathbb{X}$, the support of X. Then x can be generated from Algorithm 8.

Algorithm 8 Rejection Sampler (RS) of von Neumann – target shape

1: *input:*
   (1) shape of a target density $\tilde{f}(x) = \left(\int \tilde{f}(x)dx\right) f(x)$,
   (2) a proposal density $g(x)$ satisfying (a), (b) and (c) above.
2: *output:* a sample x from RV X with density f
3: **repeat**
4: Generate $y \sim g$ and $u \sim$ Uniform$(0, 1)$
5: **until** $u \le \frac{\tilde{f}(y)}{\tilde{a}\tilde{g}(y)}$
6: *return:* $x \leftarrow y$

Now, the expected number of iterations to get an x is no longer $\tilde{a}$ but rather the integral ratio:

$$\left(\frac{\int_{\mathbb{X}} \tilde{f}(x)dx}{\int_{\mathbb{Y}} \tilde{a}\tilde{g}(y)dy}\right)^{-1}.$$

The **Ziggurat Method** [G. Marsaglia and W. W. Tsang, SIAM Journal of Scientific and Statistical Programming, volume 5, 1984] is a rejection sampler that can efficiently draw samples from the $Z \sim$ Normal$(0, 1)$ RV. The MATLAB function `randn` uses this method to produce samples from Z.[1]

Labwork 155 (Gaussian Sampling with `randn`) We can use MATLAB function `randn` that implements the Ziggurat method to draw samples from an RV $Z \sim$ Normal$(0, 1)$ as follows:

[1] See `http://en.wikipedia.org/wiki/Ziggurat_algorithm` for more details.

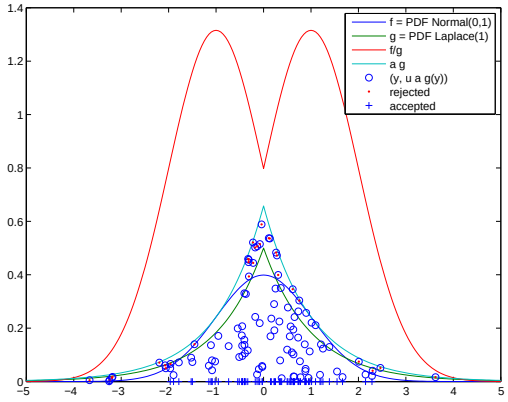
RejectionNormalLaplace.m

```
function x = RejectionNormalLaplace()
Accept = 0; % a binary variable to indicate whether a proposed point is accepted
while ~Accept % ~ is the logical NOT operation
    y = LaplaceInvCDF(rand(),1); % sample Laplace(1) RV
    Bound = exp( abs(y) -  (y*y+1)/2 );
    u = rand();
    if u <= Bound
        x = y;
        Accept = 1;
    end % if
end % while
```

We may obtain a large number of samples an plot them as a histogram using the following commands:

```
>> % use funarray to convert 1000 zeros into samples from the Normal(0,1)
>> y=arrayfun(@(x)(RejectionNormalLaplace()),zeros(1,1000));
>> hist(y,20) % histogram with 20 bins
```

Figure 4.12: Rejection Sampling from $X \sim \text{Normal}(0,1)$ with PDF f based on 100 proposals from $Y \sim \text{Laplace}(1)$ with PDF g.



Classwork 154 (A note on the proposal's tail in rejection sampling) The condition $f(x) \leq ag(x)$ is equivalent to $f(x)/g(x) \leq a$, which says that $f(x)/g(x)$ must be bounded; therefore, g must have higher tails than f. The rejection method cannot be used to generate from a Cauchy distribution using a normal distribution, because the latter has lower tails than the former.

The next result tells us how many iterations of the algorithm are needed, on average, to get a sample value from a RV with PDF f.

Proposition 61 (Acceptance Probability of RS) The expected number of iterations of the rejection algorithm to get a sample x is the constant a.

where, $\binom{n}{x}$ is:

$$\binom{n}{x} = \frac{n(n-1)(n-2)\dots(n-x+1)}{x(x-1)(x-2)\cdots(2)(1)} = \frac{n!}{x!(n-x)!}\,.$$

$\binom{n}{x}$ is read as "n choose x."

A Quick Justification: The argument from Example 46 generalises as follows. Since the trials are independent and identical, the probability of x successes followed by $n-x$ failures, *in order*, is given by

$$\underbrace{\mathsf{SS}\dots\mathsf{S}}_{x}\underbrace{\mathsf{FF}\dots\mathsf{F}}_{n-x} = \theta^x(1-\theta)^{n-x}\,.$$

Since the n symbols $\mathsf{SS}\dots\mathsf{S}\,\mathsf{FF}\dots\mathsf{F}$ may be arranged in

$$\binom{n}{x} = \frac{n!}{(n-x)!x!}$$

ways, the probability of x successes and $n-x$ failures, *in any order*, is given by

$$\binom{n}{x}\theta^x(1-\theta)^{n-x}\,.$$

Proof: This is only a sketch. A formal proof should start with the mathematical induction for the very formula for the binomial coefficient.

Observe that for the Binomial(n,θ) RV X, $\mathbf{P}(X=x) = f(x;n,\theta)$ is the probability that x of the n Bernoulli(θ) trials result in an outcome of 1's. Next note that if all n X_i's are 0's, then $X=0$, and if all n X_i's are 1's, then $X=n$. In general, if some of the n X_i's are 1's and the others are 0, then X can only take values in $\{0,1,2,\dots,n\}$ and therefore $f(x;n,\theta)=0$ if $x \notin \{0,1,2,\dots,n\}$.

Now, let us compute $f(x;n,\theta)$ when $x \in \{0,1,2,\dots,n\}$. Consider the set of indices $\{1,2,3,\dots,n\}$ for the n IID Bernoulli(θ) RVs $\{X_1,X_2,\dots,X_n\}$. Now choose x indices from $\{1,2,\dots,n\}$ to mark those trials in a particular realization of $\{x_1,x_2,\dots,x_n\}$ with the Bernoulli outcome of 1. The probability of each such event is $\theta^x(1-\theta)^{n-x}$ due to the IID assumption. For each realization $\{x_1,x_2,\dots,x_n\} \in \{0,1\}^n := \{\text{all binary } (0-1) \text{ strings of length } n\}$, specified by a choice of x trial indices with Bernoulli outcome 1, the binomial RV $X = \sum_{i=1}^n X_i$ takes the value x. Since there are exactly $\binom{n}{x}$ many ways in which we can choose x trial indices (with outcome 1) from the set of n trial indices $\{1,2,\dots,n\}$, we get the desired product for $f(x;n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ when $x \in \{0,1,\dots,n\}$.

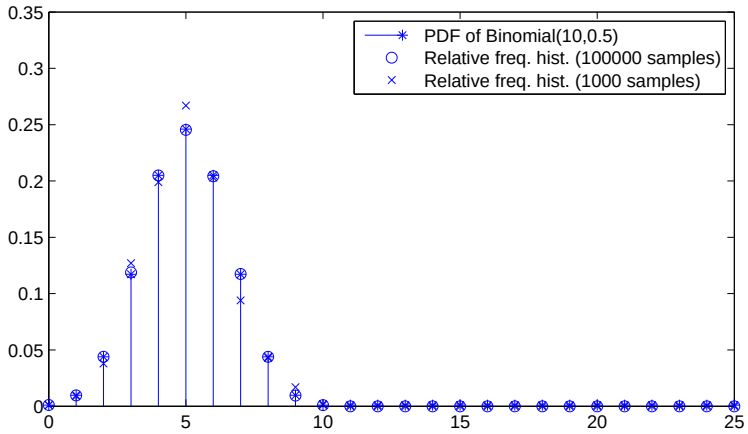
Example 47 Find the probability that seven of ten persons will recover from a tropical disease where the probability is identically 0.80 that any one of them will recover from the disease.

Solution:

We can assume independence here, so we have a binomial situation with $x=7$, $n=10$, and $\theta=0.8$. Substituting these into the formula for the probability mass function for Binomial$(10,0.8)$ random variable, we get:

$$\begin{aligned} f(7;10,0.8) &= \binom{10}{7} \times (0.8)^7 \times (1-0.8)^{10-7} \\ &= \frac{10!}{(10-7)!7!} \times (0.8)^7 \times (1-0.8)^{10-7} \\ &= 120 \times (0.8)^7 \times (1-0.8)^{10-7} \qquad {}^*\text{In the exam you can give your answer as such an expression.} \\ &\approx 0.20 \end{aligned}$$

Figure 3.8: PDF of $X \sim \text{Binomial}(n = 10, \theta = 0.5)$ and the relative frequency histogram based on $100,000$ samples from X obtained according to Simulation 148.



Example 48 Compute the probability of obtaining *at least two* 6*'s* in rolling a fair die independently and identically four times.

Solution:

In any given toss let $\theta = \mathbf{P}(\{6\}) = 1/6$, $1 - \theta = 5/6$, $n = 4$.

The event *at least two* 6*'s* occurs if we obtain two or three or four 6's. Hence the answer is:

$$\begin{aligned}
\mathbf{P}(\textit{at least two } 6\textit{'s}) &= f\left(2; 4, \frac{1}{6}\right) + f\left(3; 4, \frac{1}{6}\right) + f\left(4; 4, \frac{1}{6}\right) \\
&= \binom{4}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^{4-2} + \binom{4}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^{4-3} + \binom{4}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^{4-4} \\
&= \frac{1}{6^4}\left(6 \cdot 25 + 4 \cdot 5 + 1\right) \\
&\approx 0.132
\end{aligned}$$

To make concrete sense of the Binomial(n, θ) and other more sophisticated concepts in the sequel, let us take a historical detour into some origins of statistical thinking in 19th century England.

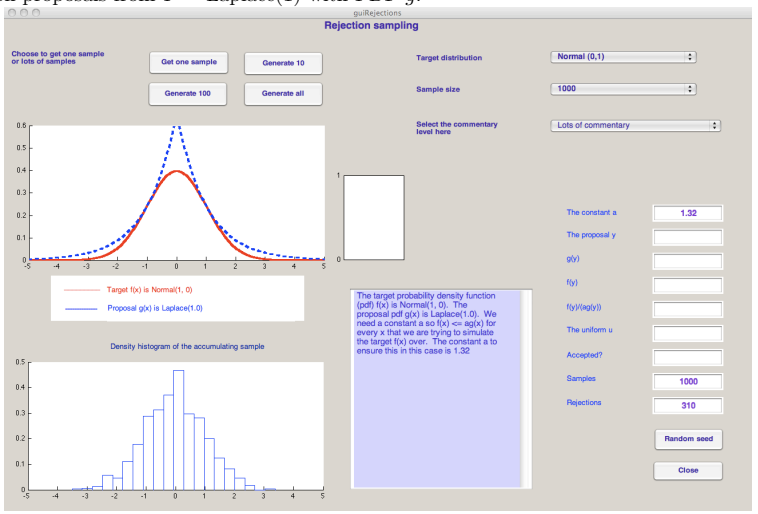
Sir Francis Galton's Quincunx

This section is introduced to provide some forms for a kinesthetic (hands-on) and visual understanding of some elementary statistical distributions and laws. The following words are from Sir Francis Galton, F.R.S., *Natural Inheritance*, pp. 62-65, Macmillan, 1889. In here you will already find the kernels behind the construction of Binomial(θ) RV as sum of IID Bernoulli(θ) RVs, Weak Law of Large Numbers, Central Limit Theorem, and more. We will mathematically present these concepts


```
>> guiRejections
```

The M-file `guiRejections.m` will bring a graphical user interface (GUI) as shown in Figure 4.11. Try various buttons and see how the output changes with explanations. Try switching the "Target distribution" to "Mywavy4" and generate several rejection samples and see the density histogram of the accumulating samples.

Figure 4.11: Visual Cognitive Tool GUI: Rejection Sampling from $X \sim \text{Normal}(0, 1)$ with PDF f based on proposals from $Y \sim \text{Laplace}(1)$ with PDF g.



Simulation 153 (Rejection Sampling Normal$(0, 1)$ **with** Laplace(1) **proposals)** Suppose we wish to generate from $X \sim \text{Normal}(0, 1)$. Consider using the rejection sampler with proposals from $Y \sim \text{Laplace}(1)$ (using inversion sampler of Simulation 137). The support of both RVs is $(-\infty, \infty)$. Next:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \text{ and } g(x) = \frac{1}{2}\exp(-|x|)$$

and therefore:

$$\frac{f(x)}{g(x)} = \sqrt{\frac{2}{\pi}} \exp\left(|x| - \frac{x^2}{2}\right) \leq \sqrt{\frac{2}{\pi}} \exp\left(\frac{1}{2}\right) = a \approx 1.3155 \, .$$

Hence, we can use the rejection method with:

$$\frac{f(y)}{ag(y)} = \frac{f(y)}{g(y)}\frac{1}{a} = \sqrt{\frac{2}{\pi}} \exp\left(|y| - \frac{y^2}{2}\right) \frac{1}{\sqrt{\frac{2}{\pi}} \exp\left(\frac{1}{2}\right)} = \exp\left(|y| - \frac{y^2}{2} - \frac{1}{2}\right)$$

Let us implement a rejection sampler as a function in the M-file `RejectionNormalLaplace.m` by reusing the function in `LaplaceInvCDF.m`.

(b) Set the variable `Mytheta=rand`.

(c) Draw 100 samples from the Geometric(`Mytheta`) RV and report the sample mean. [Note: the inputs θ_0 and $C(i)$ for the Geometric(θ) RV should be derived and the workings shown].

4.3.3 von Neumann Rejection Sampler (RS)

Rejection sampling [John von Neumann, 1947, in *Stanislaw Ulam 1909-1984*, a special issue of Los Alamos Science, Los Alamos National Lab., 1987, p. 135-136] is a Monte Carlo method to draw independent samples from a target RV X with probability density $f(x)$, where $x \in \mathbb{X} \subset \mathbb{R}^k$. Typically, the target density f is only known up to a constant and therefore the (normalised) density f itself may be unknown and it is difficult to generate samples directly from X.

Suppose we have another density or mass function g for which the following are true:

(a) we can generate random variables from g;

(b) the support of g contains the support of f, i.e. $\mathbb{Y} \supset \mathbb{X}$;

(c) a constant $a > 1$ exists, such that:

$$f(x) \leq ag(x). \tag{4.6}$$

for any $x \in \mathbb{X}$, the support of X. Then x can be generated from Algorithm 7.

Algorithm 7 Rejection Sampler (RS) of von Neumann

1: *input:*

(1) a target density $f(x)$,

(2) a proposal density $g(x)$ satisfying (a), (b) and (c) above.

2: *output:* a sample x from RV X with density f
3: **repeat**
4: Generate $y \sim g$ and $u \sim \text{Uniform}(0,1)$
5: **until** $u \leq \frac{f(y)}{ag(y)}$
6: *return:* $x \leftarrow y$

Proposition 60 (Fundamental Theorem of Simulation) The von Neumann rejection sampler of Algorithm 7 produces a sample x from the random variable X with density $f(x)$.

Proof: We shall prove the result for the continuous case. For any real number t:

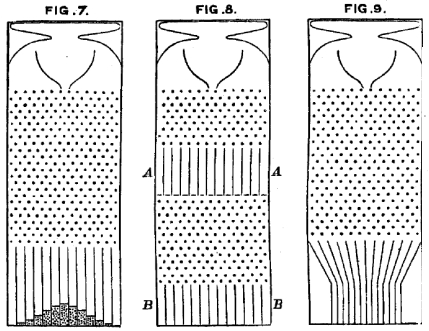
$$\begin{aligned} F(t) &= \mathbf{P}(X \leq t) = \mathbf{P}\left(Y \leq t \,|\, U \leq \tfrac{f(Y)}{ag(Y)}\right) = \frac{\mathbf{P}\left(Y \leq t, U \leq \frac{f(Y)}{ag(Y)}\right)}{\mathbf{P}\left(U \leq \frac{f(Y)}{ag(Y)}\right)} \\ &= \frac{\int_{-\infty}^{t}\left(\int_0^{f(y)/ag(y)} 1du\right)g(y)dy}{\int_{-\infty}^{\infty}\left(\int_0^{f(y)/ag(y)} 1du\right)g(y)dy} = \frac{\int_{-\infty}^{t}\left(\frac{f(y)}{ag(y)}\right)g(y)dy}{\int_{-\infty}^{\infty}\left(\frac{f(y)}{ag(y)}\right)g(y)dy} \\ &= \int_{-\infty}^{t} f(y)dy \end{aligned}$$

Labwork 152 (Rejection Sampler Demo) Let us understand the rejection sampler by calling the interactive visual cognitive tool:

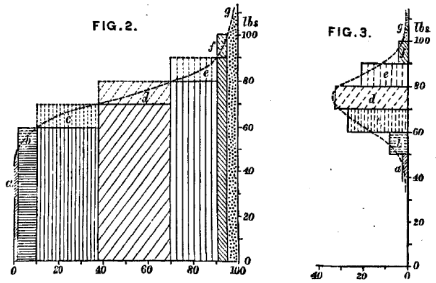
in the sequel as a way of giving precise meanings to Galton's observations with his Quincunx. "The Charms of Statistics.–*It is difficult to understand why statisticians commonly limit their inquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, it its mountains could be thrown into its lakes, two nuances would be got rid of at once. An Average is but a solitary fact, whereas if a single other fact be added to it, an entire Normal Scheme, which nearly corresponds to the observed one, starts potentially into existence.*

Some people hate the very name of statistics, but I find them full of beauty and interest. Whenever they are not brutalised, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomenon is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of man.

Figure 3.9: Figures from Sir Francis Galton, F.R.S., *Natural Inheritance*, , Macmillan, 1889.



(a) FIG. 7, FIG. 8, and FIG. 9 (p. 63)



(b) FIG. 2 and FIG. 3 (p. 38)

Mechanical Illustration of the Cause of the Curve of Frequency.–*The Curve of Frequency, and that of Distribution, are convertible : therefore if the genesis of either of them can be made clear, that of the other also becomes intelligible. I shall now illustrate the origin of the Curve of Frequency, by means of an apparatus shown in Fig. 7, that mimics in a very pretty way the conditions on which Deviation depends. It is a frame glazed in front, leaving a depth of about a quarter of an inch behind the glass. Strips are placed in the upper part to act as a funnel. Below the outlet of the funnel stand a succession of rows of pins stuck squarely into the backboard, and below these again are a series of vertical compartments. A charge of small shot is inclosed. When the frame is held topsy-turvy, all the shot runs to the upper end; then, when it is turned back into its working position, the desired action commences. Lateral strips, shown in the diagram, have the effect of directing all the shot that had collected at the upper end of the frame to run into the wide moutn of the funnel. The shot passes through the funnel and issuing from its narrow end, scampers deviously down through the pins in a curious and interesting way; each of them darting a step to the right or left, as the case may be, every time it strikes a pin. The pins are disposed in a quincunx fashion, so that every descending shot strikes against a pin in each successive row. The cascade issuing from the funnel broadens as it descends, and, at length, every shot finds itself caught in a compartment immediately after freeing itself from the last row of pins. The outline of the columns of shot that accumulate in the successive compartments approximates to the Curve of Frequency (Fig. 3, p. 38), and is closely of the same shape however often the experiment is repeated. The outline of the columns would*

become more nearly identical with the Normal Curve of Frequency, if the rows of pins were much more numerous, the shot smaller, and the compartments narrower; also if a larger quantity of shot was used.

The principle on which the action of the apparatus depends is, that a number of small and independent accidents befall each shot in its career. In rare cases, a long run of luck continues to favour the course of a particular shot towards either outside place, but in the large majority of instances the number of accidents that cause Deviation to the right, balance in a greater or less degree those that cause Deviation to the left. Therefore most of the shot finds its way into the compartments that are situated near to a perpendicular line drawn from the outlet of the funnel, and the Frequency with which shots stray to different distances to the right or left of that line diminishes in a much faster ratio than those distances increase. This illustrates and explains the reason why mediocrity is so common."

We now consider the last of our common discrete random variables for now, the **Poisson** case. A Poisson random variable counts the number of times an event occurs.

We might, for example, ask:

- How many customers visit Cafe Angstrom each day?
- How many sixes are scored in a cricket season? Cricket is a game played in the English-speaking worlds.
- How many bombs hit a city block in south London during World War II?

A Poisson experiment has the following characteristics:

- The average rate of an event occurring is known. This rate is constant.
- The probability that an event will occur during a short continuum is proportional to the size of the continuum.
- Events occur independently.

The number of events occurring in a Poisson experiment is referred to as a **Poisson random variable**.

Model 6 (Poisson(λ) **RV)** Given a real parameter $\lambda > 0$, the discrete RV X is said to be Poisson(λ) distributed if X has PDF:

$$f(x;\lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{if } x \in \mathbb{Z}_+ := \{0,1,2,\ldots\} \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{3.20}$$

Note that the PDF integrates to 1:

$$\sum_{x=0}^{\infty} f(x;\lambda) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1 \ ,$$

where we exploit the Taylor series of e^{λ} to obtain the second-last equality above.

We interpret X as the number of times an event occurs during a specified continuum given that the average value in the continuum is λ.

Algorithm 6 allows us to simulate from any member of the class of non-negative discrete RVs as specified by the probabilities $(\theta_0, \theta_1, \ldots)$. When an RV X takes values in another countable set $\mathbb{X} \neq \mathbb{Z}_+$, then we can still use the above algorithm provided we have a one-to-one and onto mapping $D(i) = x : \mathbb{Z}_+ \to \mathbb{X}$ that allows us to think of $(0, 1, 2, \ldots)$ as indices of an array D giving $\mathbb{X} = (D(0), D(1), \ldots)$.

Algorithm 6 Inversion Sampler for $GD(\theta_0, \theta_1, \ldots)$ RV X

1: *input:*

   1. θ_0 and $\{C(i) = \theta_i/\theta_{i-1}\}$ for any $i \in \{1, 2, 3, \ldots\}$.
   2. $u \sim$ Uniform$(0, 1)$

2: *output:* a sample from X
3: *initialise:* $p \leftarrow \theta_0$, $q \leftarrow \theta_0$, $i \leftarrow 0$
4: **while** $u > q$ **do**
5: $\quad i \leftarrow i + 1$, $p \leftarrow p\, C(i)$, $q \leftarrow q + p$
6: **end while**
7: *return:* $x = i$

Simulation 150 (Binomial(n, θ)**)** To simulate from a Binomial(n, θ) RV X, we can use Algorithm 6 with:

$$\theta_0 = (1-\theta)^n, \qquad C(x+1) = \frac{\theta(n-x)}{(1-\theta)(x+1)}, \qquad \text{Mean Efficiency: } O(1+n\theta) \ .$$

Similarly, with the appropriate θ_0 and $C(x+1)$, we can also simulate from the Geometric(θ) and Poisson(λ) RVs.

Labwork 151 This is a challenging exercise for the student who is finding the other Labworks too easy. So those who are novice to MATLAB may skip this Labwork.

1. Implement Algorithm 6 via a function named `MyGenDiscInvSampler` in `MATLAB`. Hand in the `M-file` named `MyGenDiscInvSampler.m` giving detailed comments explaining your understanding of each step of the code. [Hint: $C(i)$ should be implemented as a function (use function handles via `@`) that can be passed as a parameter to the function `MyGenDiscInvSampler`].
2. Show that your code works for drawing samples from a Binomial(n, p) RV by doing the following:
   (a) Seed the fundamental sampler by your Student ID (if your ID is `11424620` then type `rand('twister', 11424620);`)
   (b) Draw 100 samples from the Binomial$(n = 20, p = 0.5)$ RV and report the results in an 2×2 table with column headings `x` and No. of observations. [Hint: the inputs θ_0 and $C(i)$ for the Binomial(n, p) RV is given above].
3. Show that your code works for drawing samples from a Geometric(p) RV by doing the following:
   (a) Seed the fundamental sampler by your Student ID.

Here is a call to simulate 12 samples from Binomial($n = 10, \theta = 0.5$) RV:

```
>> theta=0.5; % declare theta
>> n=10; % say n=10
>> SampleSize=12;% say you want to simulate 12 samples
>> rand('twister',10001) % seed the fundamental sampler
>> Samples=arrayfun(@(T)Sim1BinomByGeoms(n,T),theta*ones(1,SampleSize))
Samples =     7     5     8     8     4     1     4     8     2     4     6     5
```

Figure 3.8 depicts a comparison of the PDF of Binomial($n = 10, \theta = 0.5$) RV and a relative frequency histogram based on 100,000 simulations from it.

Let us simulate from the Poisson(λ) RV of Model 6 as shown in Figure 3.18.

Simulation 149 (Poisson(λ) **from IID** Exponential(λ) **RVs)** By this principle, we can simulate from the Poisson(λ) X by Step 1: generating IID Exponential(λ) RVs $Y_1, Y_2, \ldots$, Step 2: stopping as soon as $\sum_{i=1}^{k} Y_i \geq 1$ and Step 3: setting $x \leftarrow k - 1$.

We implement the above algorithm via the following M-file:

Sim1Poisson.m

```
function x = Sim1Poisson(lambda)
% Simulate one sample from Poisson(lambda) via Exponentials
YSum=0; k=0; % initialise
while (YSum < 1),
    YSum = YSum + -(1/lambda) * log(rand);
    k=k+1;
end
x=k-1; % return x
```

Here is a call to simulate 10 samples from Poisson($\lambda = 10.0$) and Poisson($\lambda = 0.1$) RVs:

```
>> arrayfun(@(lambda)Sim1Poisson(lambda),10.0*ones(1,10)) % lambda=10.0
ans =    14     7    10    13    11     3     6     5     8     5
>> arrayfun(@(lambda)Sim1Poisson(lambda),0.1*ones(1,10)) % lambda=0.1
ans =     2     0     0     0     0     0     0     0     0     0
```

Figure 3.18 depicts a comparison of the PDF of Poisson($\lambda = 10$) RV and a relative frequency histogram based on 1000 simulations from it.

Simulating from a Poisson(λ) RV is also a special case of simulating from the following more general RV.

Model 20 ($GD(\theta_0, \theta_1, \ldots)$**)** We say X is a General Discrete($\theta_0, \theta_1, \ldots$) or GD($\theta_0, \theta_1, \ldots$) RV over the countable discrete state space $\mathbb{Z}_+ := \{0, 1, 2, \ldots\}$ with parameters $(\theta_0, \theta_1, \ldots)$ if the PMF of X is defined as follows:

$$f(X = x; \theta_0, \theta_1, \ldots) = \begin{cases} 0, & if \quad x \notin \{0, 1, 2, \ldots\} \\ \theta_0, & if \quad x = 0 \\ \theta_1, & if \quad x = 1 \\ \vdots \end{cases}$$

Example 49 If on the average, 2 cars enter a certain parking lot per minute, what is the probability that during any given minute three cars or fewer will enter the lot?

Think: Why are the assumptions for a Poisson random variable likely to be correct here?

Note: Use calculators, or Excel or Maple, etc. In an exam you may be given needed values from Poisson tables.

Let the random variable X denote the number of cars arriving per minute. Note that the continuum is 1 minute here. Then X can be considered to have a Poisson distribution with $\lambda = 2$ because 2 cars enter on average.

The probability that three cars or fewer enter the lot is:

$$\begin{aligned} \mathbf{P}(X \leq 3) &= f(0;2) + f(1;2) + f(2;2) + f(3;2) \\ &= e^{-2}\left(\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!}\right) \qquad {}^{*}\text{This is a perfectly fine answer in the exam.} \\ &= 0.857 \qquad (3 \text{ sig. fig.}) \end{aligned}$$

Example 50 (Arrivals at a Service Station) The proprietor of a service station finds that, on average, 8 cars arrive *per hour* on Saturdays. What is the probability that during a randomly chosen 15 *minute period* on a Saturday:

(a) No cars arrive?

(b) At least three cars arrive?

Solution:

Let the random variable X denote the number of cars arriving in a 15 minute interval. The continuum is 15 minutes here so we need the average number of cars that arrive in a 15 minute period, or $\frac{1}{4}$ of an hour. We know that 8 cars arrive per hour, so X has a Poisson distribution with

$$\lambda = \frac{8}{4} = 2\,.$$

(a)

$$\mathbf{P}(X = 0) = f(0;2) = \frac{e^{-2}2^0}{0!} = 0.135 \quad (3 \text{ sig. fig})$$

(b)

$$\begin{aligned} \mathbf{P}(X \geq 3) &= 1 - \mathbf{P}(X < 3) \\ &= 1 - \mathbf{P}(X = 0) - \mathbf{P}(X = 1) - \mathbf{P}(X = 2) \\ &= 1 - f(0;2) - f(1;2) - f(2;2) \\ &= 1 - 0.1353 - 0.2707 - 0.2707 \\ &= 0.323 \quad (3 \text{ sig. fig.}) \end{aligned}$$

Remark 24 In the binomial case where θ is small and n is large, it can be shown that the binomial distribution with parameters n and θ is closely approximated by the Poisson distribution having $\lambda = n\theta$. The smaller the value of θ and larger the value of n, the better the approximation.

In the sequel we will see more formally, after understanding notions of convergence of RVs, that the sum of a sequence of n IID Bernoulli(θ) RVs with $\lambda = n\theta$ converges to the Poisson(λ) RV as $n \to \infty$ and $\theta \to 0$ in a specific sense.

Example 51 (Still-born Babies) About 0.01% of babies are stillborn in a certain hospital. We find the probability that of the next 5000 babies born, there will be no more than 1 stillborn baby.

Let the random variable X denote the number of stillborn babies. Then X has a binomial distribution with parameters $n = 5000$ and $\theta = 0.0001$. Since θ is so small and n is large, this binomial distribution may be approximated by a Poisson distribution with parameter

$$\lambda \;=\; n\,\theta \;=\; 5000 \times 0.0001 \;=\; 0.5\,.$$

Hence

$$\mathbf{P}(X \leq 1) \;=\; \mathbf{P}(X = 0) + \mathbf{P}(X = 1) \;=\; f(0; 0.5) \;+\; f(1; 0.5) \;=\; 0.910 \quad \text{(3 sig. fig.)}$$

Exercise 3.4 (Nazi Bombs on London) Feller discusses the probability and statistics of flying bomb hits in an area of southern London during II world war. The area in question was partitioned into $24 \times 24 = 576$ small squares. The total number of hits was 537. There were 229 squares with 0 hits, 211 with 1 hit, 93 with 2 hits, 35 with 3 hits, 7 with 4 hits and 1 with 5 or more hits. Assuming the hits were purely random, use the Poisson approximation to find the probability that a particular square would have exactly k hits. Compute the expected number of squares that would have 0, 1, 2, 3, 4, and 5 or more hits and compare this with the observed results (Snell 9.2.14).

THINKING POISSON

The Poisson distribution has been described as a limiting version of the Binomial. In particular, Exercise 49 thinks of a Poisson distribution as a model for the number of events (cars) that occur in a period of time (1 minute) when in each little chunk of time one car arrives with constant probability, independently of the other time intervals. This leads to the general view of the Poisson distribution as a good model when:

You count the number of events in a continuum when the events occur at constant rate, one at a time and independent of each other.

DISCRETE RANDOM VARIABLE SUMMARY

Probability mass function

$$f(x) = \mathbf{P}(X = x_i)$$

Distribution function

$$F(x) = \sum_{x_i \leq x} f(x_i)$$


```
>> x=0:1:8
x =     0     1     2     3     4     5     6     7     8
>> BinomialPdf(x,8,0.5)
ans =    0.0039    0.0312    0.1094    0.2188    0.2734    0.2188    0.1094    0.0312    0.0039
```

Simulation 147 (Binomial(n, θ) **as** $\sum_{i=1}^{n}$ Bernoulli(θ)**)** Since the Binomial(n, θ) RV X is the sum of n IID Bernoulli(θ) RVs we can also simulate from X by first simulating n IID Bernoulli(θ) RVs and then adding them up as follows:

```
>> rand('twister',17678);
>> theta=0.5; % give some desired theta value, say 0.5
>> n=5; % give the parameter n for Binomial(n,theta) RV X, say n=5
>> xis=floor(rand(1,n)+theta) % produce n IID samples from Bernoulli(theta=0.5) RVs X1,X2,...Xn
xis =     1     1     0     0     0
>> x=sum(xis) % sum up the xis to get a sample from Binomial(n=5,theta=0.5) RV X
x =     2
```

It is straightforward to produce more than one sample from X by exploiting the column-wise summing property of MATLAB 's `sum` function when applied to a two-dimensional array:

```
>> rand('twister',17);
>> theta=0.25; % give some desired theta value, say 0.25 this time
>> n=3; % give the parameter n for Binomial(n,theta) RV X, say n=3 this time
>> xis10 = floor(rand(n,10)+theta) % produce an n by 10 array of IID samples from Bernoulli(theta=0.25) RVs
xis10 =
     0     0     0     0     1     0     0     0     0     0
     0     1     0     1     1     0     0     0     0     0
     0     0     0     0     0     0     0     1     0     0
>> x=sum(xis10) % sum up the array column-wise to get 10 samples from Binomial(n=3,theta=0.25) RV X
x =     0     1     0     1     2     0     0     1     0     0
```

In Simulation 147, the number of IID Bernoulli(θ) RVs needed to simulate one sample from the Binomial(n, θ) RV is exactly n. Thus, as n increases, the amount of time needed to simulate from Binomial(n, θ) is $O(n)$, i.e. linear in n. We can simulate more efficiently by exploiting a simple relationship between the Geometric(θ) RV and the Binomial(n, θ) RV.

The Binomial(n, θ) RV X is related to the IID Geometric(θ) RV $Y_1, Y_2, \ldots$: X is the number of successful Bernoulli(θ) outcomes (outcome is 1) that occur in a total of n Bernoulli(θ) trials, with the number of trials between consecutive successes distributed according to IID Geometric(θ) RV.

Simulation 148 (Binomial(θ) **from IID** Geometric(θ) **RVs)** By this principle, we can simulate from the Binomial(θ) X by Step 1: generating IID Geometric(θ) RVs $Y_1, Y_2, \ldots$, Step 2: stopping as soon as $\sum_{i=1}^{k}(Y_i + 1) > n$ and Step 3: setting $x \leftarrow k - 1$.

We implement the above algorithm via the following M-file:

Sim1BinomByGeoms.m

```
function x = Sim1BinomByGeoms(n,theta)
% Simulate one sample from Binomial(n,theta) via Geometric(theta) RVs
YSum=0; k=0; % initialise
while (YSum <= n),
    TrialsToSuccess=floor(log(rand)/log (1-theta)) + 1; % sample from Geometric(theta)+1
    YSum = YSum + TrialsToSuccess; % total number of trials
    k=k+1; % number of Bernoulli successes
end
x=k-1; % return x
```

Labwork 145 (PMF versus relative frequency histogram of simulated Geometric(θ) **RV)** It is a good idea to make a relative frequency histogram of a simulation algorithm and compare that to the PDF of the discrete RV we are simulating from. We use the following script to create Figure 3.7:

PlotPdfSimGeometric.m

```
theta=0.5;
SampleSize=1000;
% simulate 1000 samples from Geometric(theta) RV
Samples=floor(log(rand(1,SampleSize))/ log (1-theta));
Xs = 0:10; % get some values for x
RelFreqs=hist(Samples,Xs)/SampleSize; % relative frequencies of Samples
stem(Xs,theta*((1-theta) .^ Xs),'*')% PDF of Geometric(theta) over Xs
hold on;
plot(Xs,RelFreqs,'o')% relative frequence histogram
RelFreqs100=hist(Samples(1:100),Xs)/100; % Relative Frequencies of first 100 samples
plot(Xs,RelFreqs100,'x')
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (1000 samples)', ...
    'Relative freq. hist. (100 samples)')
```

Let us simulate from the Binomial(n, θ) RV of Model 5.

Labwork 146 (Binomial coefficient) The MATLAB function `BinomialCoefficient` can be used to compute:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} = \frac{n(n-1)(n-2)\dots(n-x+1)}{x(x-1)(x-2)\cdots(2)(1)} = \frac{\prod_{i=(n-x+1)}^{n} i}{\prod_{i=2}^{x}} ,$$

with the following M-file:

BinomialCoefficient.m

```
function BC = BinomialCoefficient(n,x)
% returns the binomial coefficient of n choose x
% i.e. the combination of n objects taken x at a time
% x and n are scalar integers and 0 <= x <= n
NminusX = n-x;
NumeratorPostCancel = prod(n:-1:(max([NminusX,x])+1)) ;
DenominatorPostCancel = prod(2:min([NminusX, x]));
BC = NumeratorPostCancel/DenominatorPostCancel;
```

and call `BinomialCoefficient` in the function `BinomialPdf` to compute the PDF $f(x; n, \theta)$ of the Binomial(n, θ) RV X as follows:

BinomialPdf.m

```
function fx = BinomialPdf(x,n,theta)
%  Binomial probability mass function. Needs BinomialCoefficient(n,x)
%   f = BinomialPdf(x,n,theta)
%   f is the prob mass function for the Binomial(x;n,theta) RV
%   and x can be array of samples.
%   Values of x are integers in [0,n] and theta is a number in [0,1]
fx = zeros(size(x));
fx = arrayfun(@(xi)(BinomialCoefficient(n,xi)),x);
fx = fx .* (theta .^ x) .* (1-theta) .^ (n-x);
```

For example, we can compute the desired PDF for an array of samples `x` from Binomial($8, 0.5$) RV X, as follows:

Random Variable	Possible Values	Probabilities	Modelled situations
Discrete uniform	$\{x_1, x_2, \dots, x_k\}$	$\mathbf{P}(X = x_i) = \frac{1}{k}$	Situations with k equally likely values. Parameter: k.
Bernoulli(θ)	$\{0, 1\}$	$\mathbf{P}(X = 0) = 1 - \theta$ $\mathbf{P}(X = 1) = \theta$	Situations with only 2 outcomes, coded 1 for success and 0 for failure. Parameter: $\theta = \mathbf{P}(\text{success}) \in (0, 1)$.
Geometric(θ)	$\{1, 2, 3, \dots\}$	$\mathbf{P}(X = x)$ $= (1-\theta)^{x-1}\theta$	Situations where you count the number of trials until the first success in a sequence of independent trails with a constant probability of success. Parameter: $\theta = \mathbf{P}(\text{success}) \in (0, 1)$.
Binomial(n, θ)	$\{0, 1, 2, \dots, n\}$	$\mathbf{P}(X = x)$ $= \binom{n}{x}\theta^x(1-\theta)^{n-x}$	Situations where you count the number of success in n trials where each trial is independent and there is a constant probability of success. Parameters: $n \in \{1, 2, \dots\}$; $\theta = \mathbf{P}(\text{success}) \in (0, 1)$.
Poisson(λ)	$\{0, 1, 2, \dots\}$	$\mathbf{P}(X = x)$ $= \frac{\lambda^x e^{-\lambda}}{x!}$	Situations where you count the number of events in a continuum where the events occur one at a time and are independent of one another. Parameter: $\lambda =$ rate $\in (0, \infty)$.

3.3 Exercises in Discrete Random Variables

Ex. 3.5 — One number in the following table for the probability function of a random variable X is incorrect. Which is it, and what should the correct value be?

x	1	2	3	4	5
$\mathbf{P}(X = x)$	0.07	0.10	1.10	0.32	0.40

Ex. 3.6 — Let X be the number of years before a particular type of machine will need replacement. Assume that X has the probability function $f(1) = 0.1$, $f(2) = 0.2$, $f(3) = 0.2$, $f(4) = 0.2$, $f(5) = 0.3$.

1. Find the distribution function, F, for X, and graph both f and F.
2. Find the probability that the machine needs to be replaced during the first 3 years.
3. Find the probability that the machine needs no replacement during the first 3 years.

Ex. 3.7 — Of 200 adults, 176 own one TV set, 22 own two TV sets, and 2 own three TV sets. A person is chosen at random. What is the probability mass function of X, the number of TV sets owned by that person?

Ex. 3.8 — Suppose a discrete random variable X has probability function give by

x	3	4	5	6	7	8	9	10	11	12	13
$\mathbf{P}(X = x)$	0.07	0.01	0.09	0.01	0.16	0.25	0.20	0.03	0.02	0.11	0.05

(a)Construct a row of cumulative probabilities for this table, that is, find the distribution function of X.

(b)Find the following probabilities.

(i)$\mathbf{P}(X \leq 5)$	(iii)$\mathbf{P}(X > 9)$	(v)$\mathbf{P}(4 < X \leq 9)$
(ii)$\mathbf{P}(X < 12)$	(iv)$\mathbf{P}(X \geq 9)$	(vi)$\mathbf{P}(4 < X < 11)$

Ex. 3.9 — A box contains 4 right-handed and 6 left-handed screws. Two screws are drawn at random without replacement. Let X be the number of left-handed screws drawn. Find the probability mass function for X, and then calculate the following probabilities:

1.$\mathbf{P}(X \leq 1)$

2.$\mathbf{P}(X \geq 1)$

3.$\mathbf{P}(X > 1)$

Ex. 3.10 — Suppose that a random variable X has geometric probability mass function,

$$f(x) \;=\; \frac{k}{2^x} \quad (x = 0, 1, 2, \dots)\,.$$

1.Find the value of k.

2.What is $\mathbf{P}(X \geq 4)$?

Ex. 3.11 — Four fair coins are tossed simultaneously. If we count the number of heads that appear then we have a binomial random variable, $X =$ *the number of heads*.

1.Find the probability mass function of X.

2.Compute the probabilities of obtaining no heads, precisely 1 head, at least 1 head, not more than 3 heads.

Ex. 3.12 — The distribution of blood types in a certain population is as follows:

Blood type	Type O	Type A	Type B	Type AB
Proportion	0.45	0.40	0.10	0.05

A random sample of 15 blood donors is observed from this population. Find the probabilities of the following events.

1.Only one type AB donor is included.

2.At least three of the donors are type B.

3.More than ten of the donors are *either* type O *or* type A.

4.Fewer that five of the donors are *not* type A.

Ex. 3.13 — If the probability of hitting a target in a single shot is 10% and 10 shots are fired independently, what is the probability that the target will be hit at least once?

Ex. 3.14 — Suppose that a certain type of magnetic tape contains, on the average, 2 defects per 100 meters. What is the probability that a roll of tape 300 meters long will contain no defects?

Ex. 3.15 — In 1910, E. Rutherford and H. Geiger showed experimentally that the number of alpha particles emitted per second in a radioactive process is a random variable X having a Poisson distribution. If the average number of particles emitted per second is 0.5, what is the probability of observing two or more particles during any given second?


```
>> rand('twister',1777); % initialise the fundamental sampler
>> f2=rand(1,10); % create an arbitrary array
>> f2=f2/sum(f2); % normalize to make a probability mass function
>> disp(f2); % display the weights of our 10-faced die
    0.0073    0.0188    0.1515    0.1311    0.1760    0.1121    ...
    0.1718    0.1213    0.0377    0.0723
>> disp(sum(f2)); % the weights sum to 1
    1.0000
>> disp(arrayfun(@(u)(SimdeMoivreOnce(u,f2)),rand(5,5))) % the samples from f2 are
     4     3     4     7     3
     6     7     4     5     3
     5     8     7    10     6
     2     3     5     7     7
     6     5     9     5     7
```

Note that the principal work here is the sequential search, in which the mean number of comparisons until success is:

$$1\theta_1 + 2\theta_2 + 3\theta_3 + \ldots + k\theta_k = \sum_{i=1}^{k} i\theta_i$$

For the de Moivre$(1/k, 1/k, \ldots, 1/k)$ RV, the right-hand side of the above expression is:

$$\sum_{i=1}^{k} i\frac{1}{k} = \frac{1}{k}\sum_{i=1}^{k} i = \frac{1}{k}\frac{k(k+1)}{2} = \frac{k+1}{2}\;,$$

indicating that the average-case efficiency is linear in k. This linear dependence on k is denoted by $O(k)$. In other words, as the number of faces k increases, one has to work linearly harder to get samples from de Moivre$(1/k, 1/k, \ldots, 1/k)$ RV using Algorithm 5. Using the simpler Algorithm 4, which exploits the fact that all values of θ_i are equal, we generated samples in constant time, which is denoted by $O(1)$.

Simulation 144 (Geometric(θ)**)** We can simulate a sample x from a Geometric(θ) RV X using the following simple algorithm:

$$x \leftarrow \lfloor \log(u)/\log(1-\theta) \rfloor, \qquad \text{where,}\;\; u \sim \text{Uniform}(0,1)\;.$$

To verify that the above procedure is valid, note that:

$$\begin{aligned} \lfloor \log(U)/\log(1-\theta) \rfloor = x \;&\iff\; x \leq \log(U)/\log(1-\theta) < x+1 \\ &\iff\; x \leq \log_{1-\theta}(U) < x+1 \\ &\iff\; (1-\theta)^x \geq U > (1-\theta)^{x+1} \end{aligned}$$

The inequalities are reversed since the base being exponentiated is $1-\theta \leq 1$. The uniform event $(1-\theta)^x \geq U > (1-\theta)^{x+1}$ happens with the desired probability:

$$(1-\theta)^x - (1-\theta)^{x+1} = (1-\theta)^x(1-(1-\theta)) = \theta(1-\theta)^x =: f(x;\theta), \quad X \sim \text{Geometric}(\theta)\;.$$

We implement the sampler to generate samples from Geometric(θ) RV with $\theta = 0.5$, for instance:

```
>> theta=0.5; u=rand(); % choose some theta and uniform(0,1) variate
>> % Simulate from a Geomertic(theta) RV
>> floor(log (u) / log (1 - theta))
ans =     0
>> floor(log ( rand(1,10) ) / log (1 - 0.5)) % theta=0.5, 10 samples
ans =     0     0     1     0     2     1     0     0     0     0
```

Simulation 143 (de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$**)** We can generate samples from a de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ RV X when $(\theta_1, \theta_2, \ldots, \theta_k)$ are specifiable as an input vector via the following algorithm.

Algorithm 5 Inversion Sampler for de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ RV X

1: *input:*
   1. parameter vector $(\theta_1, \theta_2, \ldots, \theta_k)$ of de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ RV X.
   2. $u \sim$ Uniform$(0, 1)$
2: *output:* a sample from X
3: *initialise:* $F \leftarrow \theta_1$, $i \leftarrow 1$
4: **while** $u > F$ **do**
5: $\quad i \leftarrow i + 1$
6: $\quad F \leftarrow F + \theta_i$
7: **end while**
8: *return:* $x \leftarrow i$

The M-file implementing Algorithm 5 is:

SimdeMoivreOnce.m

```
function  x = SimdeMoivreOnce(u,thetas)
% Returns a sample from the de Moivre(thetas=(theta_1,...,theta_k)) RV X
% Call Syntax:  x = SimdeMoivreOnce(u,thetas);
%               deMoivreEqui(u,thetas);
% Input      : u = a uniform random number eg. rand
%              thetas = an array of probabilities thetas=[theta_1 ... theta_k]
% Output     : x = sample from X
x=1; % initial index is 1
cum_theta=thetas(x);
while u > cum_theta;
    x=x+1;
    cum_theta = cum_theta + thetas(x);
end
```

Let us use the function `deMoivreEqui` to draw five samples from a fair seven-faced dice.

```
>> k=7; % number of faces of the fair dice
>> n=5; % number of trials
>> rand('twister',78657); % initialise the fundamental sampler
>> Us=rand(1,n); % draw n samples from Uniform(0,1)
>> disp(Us);
    0.8330    0.6819    0.6468    0.6674    0.2577
>> % inverse transform samples from Uniform(0,1) to samples
>> % from de Moivre(1/7,1/7,1/7,1/7,1/7,1/7,1/7)
>> f=[1/7 1/7 1/7 1/7 1/7 1/7 1/7];
>> disp(f);
    0.1429    0.1429    0.1429    0.1429    0.1429    0.1429    0.1429
>> % use funarray to apply function-handled SimdeMoivreOnce to
>> % each element of array Us and save it in array outcomes2
>> outcomes2=arrayfun(@(u)(SimdeMoivreOnce(u,f)),Us);
>> disp(outcomes2);
     6     5     5     5     2
>> disp(SimdeMoivreEqui(u,k)); % same result using the previous algorithm
     6     5     5     5     2
```

Clearly, Algorithm 5 may be used to sample from any de Moivre$(\theta_1, \ldots, \theta_k)$ RV X. We demostrate this by producing five samples from a randomly generated PMF `f2`.

Ex. 3.16 — The number of lacunae (surface pits) on specimens of steel, polished and examined in a metallurgical laboratory, is thought to have a Poisson distribution.

1. Write down the formula for the probability that a specimen has x defects, explaining the meanings of the symbols you use.
2. Simplify the formula in the case $x = 0$.
3. In a large homogeneous collection of specimens, 10% have one or more lacunae. Find (approximately) the percentage having exactly two.
4. Why might the Poisson distribution not apply in this situation?

   [HINT: Recall the *emphasised sentence* in THINKING POISSON and what the continuum on which the number of events occur is for the problem, and what could possibly go wrong in your imagination of the manufacturing process of the steel specimens (normally you need to melt and manipulate iron with other elements and cast them in moulds and this needs energy and raw materials of possibly varying quality and the machines used in the process could break down, etc.) to violate the Poisson assumption about the occurrence of pits on the surface of the specimens.]

3.4 Continuous Random Variables

If X is a measurement of a continuous quantity, such as,

- the maximum diameter in millimeters of a venus shell I picked up at New Brighton beach,
- the distance you transported yourself to lectures today in meters,
- the volume of rain that fell on the roof of this building over the past 365 days in litres,
- the vertical position (in micro meters above sea-level) since the release of a pollen grain at a location in Lake Rogen in Härjedalen, as it traces through Göta älvKlarälven, the longest river of Sweden before discharging in a delta into Vänern at Karlstad.
- the volume of water (in cubic meters) that fell on the southern Alps of the South Island of New Zealand throughout last year.
- etc.,

then X is a continuous random variable. Continuous random variables are based on measurements in a continuous scale of a given precision as opposed to discrete random variables that are based on counting.

Example 52 Suppose that X is the time, in minutes, before the next student leaves the lecture room. This is an example of a continuous random variable that takes one of (uncountably) infinitely many values. When a student leaves, X will take on the value x and this x could be 2.1 minutes, or 2.1000000001 minutes, or 2.9999999 minutes, etc., depending the measurement precision of the clock being used to measure time.

Finding $\mathbf{P}(X = 2)$, for example, doesn't make sense because how can it ever be *exactly* $2.00000\cdots$ minutes? It is more sensible to consider probabilities like $\mathbf{P}(X > x)$ or $\mathbf{P}(X < x)$ or $\mathbf{P}(a < X < b)$ with $a < b$, up to measurement precision of the time-measuring clock rather than the discrete approach of trying to compute $\mathbf{P}(X = x)$.

The characteristics of continuous random variables are:

- The outcomes are measured, not counted.
- Geometrically, *the probability of an outcome is equal to an area* under a mathematical curve.
- Each individual value has zero probability of occurring. So we find the probability that the value is between two endpoints of an interval, or a set of intervals, including half-lines in $\mathbb{R}$.

Definition 25 (probability density function (PDF)) A RV X with distribution function (DF) given by F is said to be **continuous** if there exists a piecewise-continuous function f, called the **probability density function (PDF)** of X, such that

$$F(x) \;=\; \mathbf{P}(X \leq x) \;=\; \int_{-\infty}^{x} f(v)\, dv \tag{3.21}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a non-negative function, i.e., $f(x) \geq 0$. We write v because x is needed as the upper limit of the integral. Piecewise-continuity of f means f is continuous, perhaps possibly at the x-values where f is discontinuous between the continuous pieces (see `https://en.wikipedia.org/wiki/Piecewise`).

The following hold for a continuous RV X with PDF f:

1. For any $x \in \mathbb{R}$, $\mathbf{P}(X = x) = \mathbf{P}(X \in [x, x]) = \int_x^x f(v)dv = 0$.
2. By the fundamental theorem of calculus:
$$f(x) = \frac{d}{dx}F(x) =: F'(x), \tag{3.22}$$
for every x at which $f(x)$ is continuous.
3. Consequentially, for any $a, b \in \mathbb{R}$ with $a < b$,
$$\mathbf{P}(a < X < b) \;=\; \mathbf{P}(a < X \leq b) = \mathbf{P}(a \leq X \leq b) = \mathbf{P}(a \leq X < b) \tag{3.23}$$
$$\;=\; F(b) - F(a) \;=\; \int_a^b f(v)dv \;. \tag{3.24}$$
4. And $P(\Omega) = 1$ implies that:
$$\int_{-\infty}^{\infty} f(x)\, dx = \mathbf{P}(-\infty < X < \infty) = 1 \;.$$

The next set of examples illustrate notation and typical applications of the formulae above.

Example 53 Consider the continuous random variable, X, whose probability density function is:

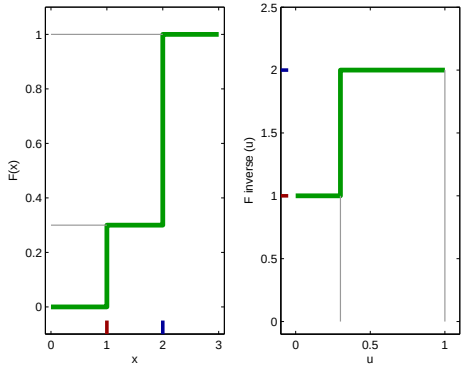
$$f(x) \;=\; \begin{cases} 3x^2 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the distribution function, $F(x)$.

(b) Find $P(\frac{1}{3} \leq X \leq \frac{2}{3})$.

Solution

Figure 4.10: The DF $F(x; 0.3, 0.7)$ of the de Moivre$(0.3, 0.7)$ RV and its inverse $F^{[-1]}(u; 0.3, 0.7)$.



Algorithm 4 Inversion Sampler for de Moivre$(1/k, 1/k, \ldots, 1/k)$ RV

1: *input:*
   1. k in de Moivre$(1/k, 1/k, \ldots, 1/k)$ RV X
   2. $u \sim$ Uniform$(0, 1)$
2: *output:* a sample from X
3: *return:* $x \leftarrow \lceil ku \rceil$

SimdeMoivreEqui.m

```
function x = SimdeMoivreEqui(u,k);
% return samples from de Moivre(1/k,1/k,...,1/k) RV X
% Call Syntax:  x = SimdeMoivreEqui(u,k);
% Input      : u = array of uniform random numbers eg. rand
%              k = number of equi-probabble outcomes of X
% Output     : x = samples from X
x = ceil(k * u) ; % ceil(y) is the smallest integer larger than y
%x = floor(k * u); if outcomes are in {0,1,...,k-1}
```

Let us use the function `SimdeMoivreEqui` to draw five samples from a fair seven-faced cylindrical dice.

```
>> k=7; % number of faces of the fair dice
>> n=5; % number of trials
>> rand('twister',78657); % initialise the fundamental sampler
>> u=rand(1,n); % draw n samples from Uniform(0,1)
>> % inverse transform samples from Uniform(0,1) to samples
>> % from de Moivre(1/7,1/7,1/7,1/7,1/7,1/7,1/7)
>> outcomes=SimdeMoivreEqui(u,k); % save the outcomes in an array
>> disp(outcomes);
     6     5     5     5     2
```

Now, let us consider the more general problem of implementing a sampler for an arbitrary but specified de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ RV. That is, the values of θ_i need not be equal to $1/k$.

Simulation 141 (Point Mass(θ)**)** Let us simulate a sample from the Point Mass(θ) RV X. Since this RV produces the same realisation θ we can implement it via the following M-file:

Sim1PointMass.m

```
function  x = Sim1PointMass(u,theta)
% Returns one sample from the Point Mass(theta) RV X
% Call Syntax:  x = SimPointMass(u,theta);
% Input      : u = one uniform random number eg. rand()
%              theta = a real number (scalar)
% Output     : x = sample from X
x=theta;
```

Here is call to the function.

```
>> Sim1PointMass(rand(),2)
ans =     2
>> % % we can use arrayfun to apply Sim1Pointmass to any array of Uniform(0,1) samples
>> arrayfun(@(u)(Sim1PointMass(u,17)),rand(2,10))
ans =
    17    17    17    17    17    17    17    17    17    17
    17    17    17    17    17    17    17    17    17    17
```

Note that it is not necessary to have input IID samples from Uniform$(0,1)$ RV via `rand` in order to draw samples from the Point Mass(θ) RV. For instance, an input matrix of zeros can do the job:

```
>> arrayfun(@(u)(Sim1PointMass(u,17)),zeros(2,8))
ans =
    17    17    17    17    17    17    17    17
    17    17    17    17    17    17    17    17
```

Next we simulate from de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ RV X of Model 14 via its inverse DF

$$F^{[-1]} : [0,1] \to [k] := \{1, 2, \ldots, k\} \,,$$

given by:

$$F^{[-1]}(u; \theta_1, \theta_2, \ldots, \theta_k) = \begin{cases} 1 & \text{if } 0 \le u < \theta_1 \\ 2 & \text{if } \theta_1 \le u < \theta_1 + \theta_2 \\ 3 & \text{if } \theta_1 + \theta_2 \le u < \theta_1 + \theta_2 + \theta_3 \\ \vdots & \\ k & \text{if } \theta_1 + \theta_2 + \cdots + \theta_{k-1} \le u < 1 \end{cases} \tag{4.5}$$

When $k = 2$ in the de Moivre(θ_1, θ_2) model, we have an RV that is similar to the Bernoulli$(p = \theta_1)$ RV. The DF F and its inverse $F^{[-1]}$ for a specific $\theta_1 = 0.3$ are depicted in Figure 4.10.

First we simulate from an equi-probable special case of the de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ RV, with $\theta_1 = \theta_2 = \cdots = \theta_k = 1/k$.

Simulation 142 (de Moivre$(1/k, 1/k, \ldots, 1/k)$**)** The equi-probable *de Moivre*$(1/k, 1/k, \ldots, 1/k)$ RV X with a discrete uniform distribution over $[k] = \{1, 2, \ldots k\}$ can be efficiently sampled using the ceiling function. Recall that $\lceil y \rceil$ is the smallest integer larger than or equal to y, eg. $\lceil 13.1 \rceil = 14$. Algorithm 4 produces samples from the *de Moivre*$(1/k, 1/k, \ldots, 1/k)$ RV X.

The M-file implementing Algorithm 4 is:

(a) First note that if $x \le 0$, then

$$F(x) \;=\; \int_{-\infty}^{x} 0 dv \;=\; 0 \,.$$

If $0 < x < 1$, then

$$\begin{aligned} F(x) &= \int_{-\infty}^{0} 0 dv \;+\; \int_{0}^{x} 3v^2 dv \\ &= 0 \;+\; \left[v^3\right]_0^x \\ &= x^3 \end{aligned}$$

If $x \ge 1$, then

$$\begin{aligned} F(x) &= \int_{-\infty}^{0} 0 dv \;+\; \int_{0}^{1} 3v^2 dv + \int_{1}^{x} 0 dv \\ &= 0 \;+\; \left[v^3\right]_0^1 \;+\; 0 \\ &= 1 \end{aligned}$$

Hence

$$F(x) \;=\; \begin{cases} 0 & x \le 0 \\ x^3 & 0 < x < 1 \\ 1 & x \ge 1 \end{cases}$$

(b)

$$\begin{aligned} P\left(\frac{1}{3} \le X \le \frac{2}{3}\right) &= F\left(\frac{2}{3}\right) - F\left(\frac{1}{3}\right) \\ &= \left(\frac{2}{3}\right)^3 - \left(\frac{1}{3}\right)^3 \\ &= \frac{7}{27} \end{aligned}$$

Example 54 Consider the continuous random variable, X, whose distribution function is:

$$F(x) \;=\; \begin{cases} 0 & x \le 0 \\ \sin(x) & 0 < x < \frac{\pi}{2} \\ 1 & x \ge \frac{\pi}{2} \end{cases} .$$

(a) Find the probability density function, $f(x)$.

(b) Find $P\left(X > \frac{\pi}{4}\right)$

Solution

(a) The probability density function, $f(x)$ is given by

$$f(x) \;=\; F'(x) \;=\; \begin{cases} 0 & x < 0 \\ \cos x & 0 < x < \frac{\pi}{2} \\ 0 & x \ge \frac{\pi}{2} \end{cases}$$

(b)

$$P\left(X > \frac{\pi}{4}\right) \;=\; 1 - P\left(X \leq \frac{\pi}{4}\right) \;=\; 1 \;-\; F\left(\frac{\pi}{4}\right) \;=\; 1 - \sin\left(\frac{\pi}{4}\right) \;=\; 0.293 \text{ (3 sig. fig.)}$$

* You may stop at $1 - \sin\left(\frac{\pi}{4}\right)$ for full credit in the exam.

Note: $f(x)$ is not defined at $x = 0$ as $F(x)$ is not differentiable at $x = 0$. There is a "kink" in the distribution function at $x = 0$ causing this problem. It is standard to define $f(0) = 0$ in such situations, as $f(x) = 0$ for $x < 0$. This choice is arbitrary but it simplifies things and makes no difference to the calculated probability.

Now that we have warmed-up with two examples of continuous RVs, let us define the most elementary continuous RV next.

3.4.1 An Elementary Continuous Random Variable

An elementary and fundamental example of a continuous RV is the Uniform$(0, 1)$ RV of Model 7. It forms the foundation for all non-uniform random variate generation and simulation as we will see in Chapter 4. In fact, it is appropriate to call this the fundamental model since every other probability model can be obtained from this one!

Model 7 (The Fundamental Model) The probability density function (PDF) of the fundamental model or the Uniform$(0, 1)$ RV is

$$f(x) = \mathbb{1}_{[0,1]}(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise} \end{cases} \tag{3.25}$$

and its distribution function (DF) or cumulative distribution function (CDF) is:

$$F(x) := \int_{-\infty}^{x} f(y)\, dy = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x > 1 \end{cases} \tag{3.26}$$

Note that the DF is the identity map in $[0, 1]$. The PDF and DF are depicted in Figure 3.11.

Let us draw the PDF and DF for Uniform$(0, 1)$ RV next by hand.

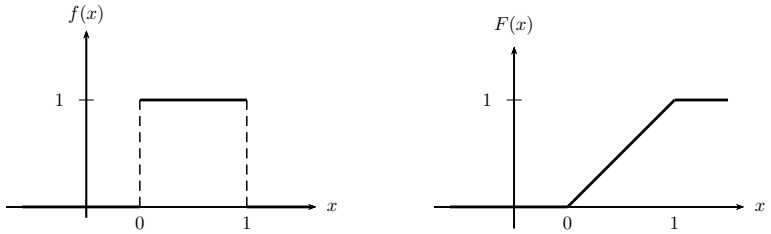


Figure 3.10: $f(x)$ and $F(x)$ of the Uniform$(0, 1)$ random variable X.

Simulation 139 (Cauchy**)** We can draw n IID samples from the Cauchy RV X by transforming n IID samples from Uniform$(0, 1)$ RV U using the inverse DF as follows:

```
>> rand('twister',2435567);       % initialise the fundamental sampler
>> u=rand(1,5);                       % draw 5 IID samples from Uniform(0,1) RV
>> disp(u);                           % display the samples in u
    0.7176    0.6655    0.9405    0.9198    0.2598
>> x=tan(pi * u);     % draw 5 samples from Standard cauchy RV using inverse CDF
>> disp(x);  % display the samples in x
   -1.2272   -1.7470   -0.1892   -0.2575    1.0634
```

4.3.2 Inversion Sampler for Discrete Random Variables

Next, consider the problem of **sampling from a random variable X with a discontinuous or discrete DF** using the inversion sampler. We need to define the inverse more carefully here.

Proposition 59 (Inversion sampler with compact support) Let the support of the RV X be over some real interval $[a, b]$ and let its inverse DF be defined as follows:

$$F^{[-1]}(u) := \inf\{x \in [a, b] : F(x) \geq u,\ 0 \leq u \leq 1\}\ .$$

If $U \sim$ Uniform$(0, 1)$ then $F^{[-1]}(U)$ has the DF F, i.e. $F^{[-1]}(U) \sim F \sim X$.

Proof: The proof is a consequence of the following equalities:

$$\mathbf{P}(F^{[-1]}(U) \leq x) = \mathbf{P}(U \leq F(x)) = F(x) := \mathbf{P}(X \leq x)$$

Simulation 140 (Bernoulli(θ)**)** Consider the problem of simulating from a Bernoulli(θ) RV based on an input from a Uniform$(0, 1)$ RV. Recall that $\lfloor x \rfloor$ (called the 'floor of x') is the largest integer that is smaller than or equal to x, e.g. $\lfloor 3.8 \rfloor = 3$. Using the floor function, we can simulate a Bernoulli(θ) RV X as follows:

```
>>  theta = 0.3;                % set theta = Prob(X=1)
  % return x  -- floor(y) is the largest integer less than or equal to y
>> x = floor(rand + theta)        % rand is the Fundamental Sampler
>> disp(x) % display the outcome of the simulation
     0
>> n=10;  % set the number of IID Bernoulli(theta=0.3) trials you want to simulate
>> x = floor(rand(1,10)+theta); % vectorize the operation
>> disp(x) % display the outcomes of the simulation
     0     0     1     0     0     0     0     0     1     1
```

Again, it is straightforward to do replicate experiments, e.g. to demonstrate the Central Limit Theorem for a sequence of n IID Bernoulli(θ) trials.

```
>> % a demonstration of Central Limit Theorem --
>> % the sample mean of a sequence of n IID Bernoulli(theta) RVs is Gaussian(theta,theta(1-theta)/n)
>> theta=0.5; Reps=10000; n=10; hist(sum(floor(rand(n,Reps)+theta))/n)
>> theta=0.5; Reps=10000; n=100; hist(sum(floor(rand(n,Reps)+theta))/n,20)
>> theta=0.5; Reps=10000; n=1000; hist(sum(floor(rand(n,Reps)+theta))/n,30)
```

Recall the Point Mass(θ) RV. Formally, we can simulate from it trivially as follows.

We can simply call the function to draw a sample from, say the Laplace($\lambda = 1.0$) RV by

```
>> lambda=1.0;               % some value for lambda
>> rand('twister',6567);       % initialize the fundamental sampler
>> u=rand(1,5);               % draw 5 IID samples from Uniform(0,1) RV
>> disp(u);                 % display the samples in u
    0.6487    0.9003    0.3481    0.6524    0.8152

>> x=LaplaceInvCDF(u,lambda); % draw 5 samples from Laplace(1) RV using inverse CDF
>> disp(x);                   % display the samples
    0.3530    1.6127   -0.3621    0.3637    0.9953
```

Labwork 138 (Inversion Sampler Demo – Cauchy) Let us comprehend the inversion sampler by calling the interactive visual cognitive tool:

```
>> guiInversionSampler
```

The M-file `guiInversionSampler.m` will bring a graphical user interface (GUI) as shown in Figure 4.9. Using the drop-down menu change from the default target distribution Uniform($-5, 5$) to Cauchy RV of Model 13. Now repeatedly push the "Draw one sample" button several times and comprehend the simulation process. You can also press "Draw 10 samples" several times and see the density histogram of the generated samples. Next try changing the numbers in the "Scale parameter" and "Location Parameter" boxes from the default values of 1.00 and 0.00, respectively. Although our formulation of Cauchy RV is also called *Standard Cauchy* as it implicitly had a location parameter of 0.00 and scale parameter of 1. With a pencil and paper (in conjunction with a wikipedia search if you have to) try to rewrite the PDF in (3.53) with an additional location parameter μ and scale parameter σ.

Figure 4.9: Visual Cognitive Tool GUI: Inversion Sampling from $X \sim$ Cauchy.

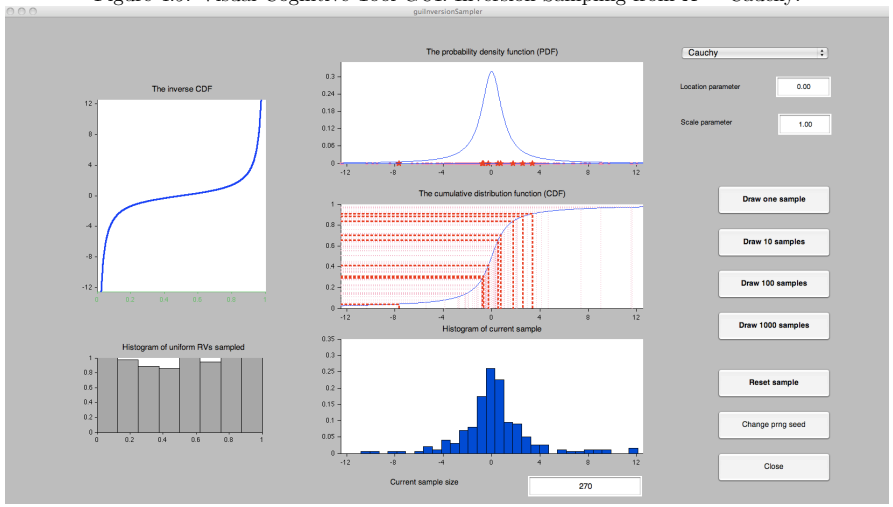
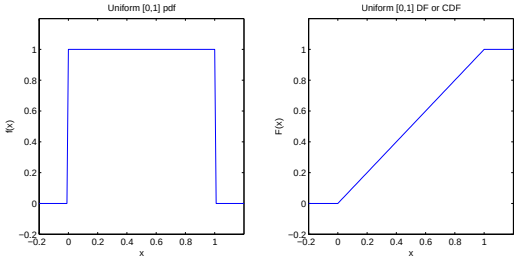


Figure 3.11: A convenient but mathematically imprecise Matlab plot from a polyline interpolation for the PDF and DF or CDF of the Uniform$(0, 1)$ continuous RV X.



****tossing a fair coin infinitely often, i.e., IID sequence of** Bernoulli$(1/2)$ **trials, and the fundamental model**

— The fundamental model is equivalent to infinite tosses of a fair coin (see using binary expansion of any $x \in (0, 1)$ if you want as suggested in optional Exercise 2.1 on intuiting a most primitive sigma-algebra)

— The fundamental model has infinitely many copies of itself within it! You can see this since its DF F is the identity function on $[0, 1]$ or equivalently how the dyadic binary tree is identical below a given node in the tree no matter which node in the tree you choose.

****universality of the fundamental model**

— one can obtain any other random variable from the fundamental model whose unique DF is its own inverse, i.e., $F(x) = F^{[-1]}(x)$, as you will See from von Neumann's Fundamental Theorem of Simulation in Chapter 4.

3.4.2 Some Common Continuous Random Variables

Let us warm-up with an example.

Example 55 Let X have density function $f(x) = e^{-x}$, if $x \geq 0$, and zero otherwise.

(a) Find the distribution function.

(b) Find the probabilities, $P(\frac{1}{4} \leq X \leq 2)$ and $P\left(-\frac{1}{2} \leq X \leq \frac{1}{2}\right)$.

(c) Find x such that $P(X \leq x) = 0.95$.

Solution:

(a)

$$F(x) = \int_0^x e^{-v} dv = -e^{-v}\Big]_0^x = -e^{-x} + 1 = 1 - e^{-x} \qquad \text{if } x \geq 0$$

Therefore,

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

(b)

$$P\left(\frac{1}{4} \le X \le 2\right) = F(2) - F\left(\frac{1}{4}\right) = 0.634 \text{ (3 sig. fig.)}$$
$$P\left(-\frac{1}{2} \le X \le \frac{1}{2}\right) = F\left(\frac{1}{2}\right) - F\left(-\frac{1}{2}\right) = 0.394 \text{ (3 sig. fig.)}$$

(c)

$$P(X \le x) = F(x) = 1 - e^{-x} = 0.95$$

Therefore,

$$x = -\log(1 - 0.95) = 3.00 \text{ (3 sig. fig.)} .$$

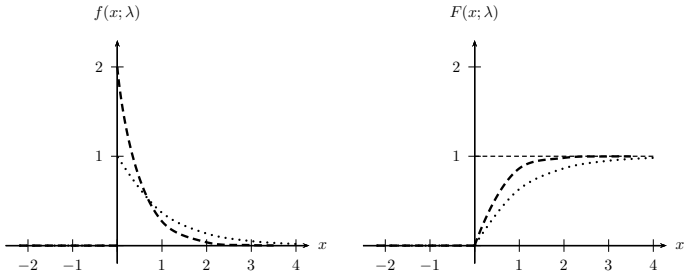


Figure 3.12: $f(x;\lambda)$ and $F(x;\lambda)$ of an exponential random variable where $\lambda = 1$ (dotted) and $\lambda = 2$ (dashed).

The previous example is a special case of the following parametric family of random variables.

Model 8 (Exponential(λ)) For a given $\lambda > 0$, an Exponential(λ) RV has the following PDF f and DF F and its complementary distribution function denoted by $\overline{F}(x;\lambda) := \mathbf{P}(X > x) = 1 - F(x;\lambda)$:

$$f(x;\lambda) = \mathbb{1}_{(0,\infty)}\lambda e^{-\lambda x} = \begin{cases} \lambda \exp(-\lambda x) & x > 0 , \\ 0 & \text{otherwise} , \end{cases} \quad (3.27)$$
$$F(x;\lambda) = 1 - e^{-\lambda x} , \quad (3.28)$$
$$\overline{F}(x;\lambda) = e^{-\lambda x} . \quad (3.29)$$

The last two equations are derived from definitions as follows:

$$F(x;\lambda) = \int_{-\infty}^{x} \mathbb{1}_{(0,\infty)}\lambda e^{-\lambda v} dv = \lambda \int_{0}^{x} e^{-\lambda v} dv = \lambda \left(-\frac{1}{\lambda} e^{-\lambda v}\right]_0^x = \left(-e^{-\lambda v}\right]_0^x$$
$$= -e^{-\lambda x} - (-e^{-0}) = -e^{-\lambda x} - (-1/e^{0}) = -e^{-\lambda x} - (-1/1) = -e^{-\lambda x} - (-1) = -e^{-\lambda x} + 1$$

$$\mathbf{P}(X > x) = 1 - \mathbf{P}(X \le x) = 1 - F(x;\lambda) = 1 - \left(1 - e^{-\lambda x}\right) = e^{-\lambda x}$$

.

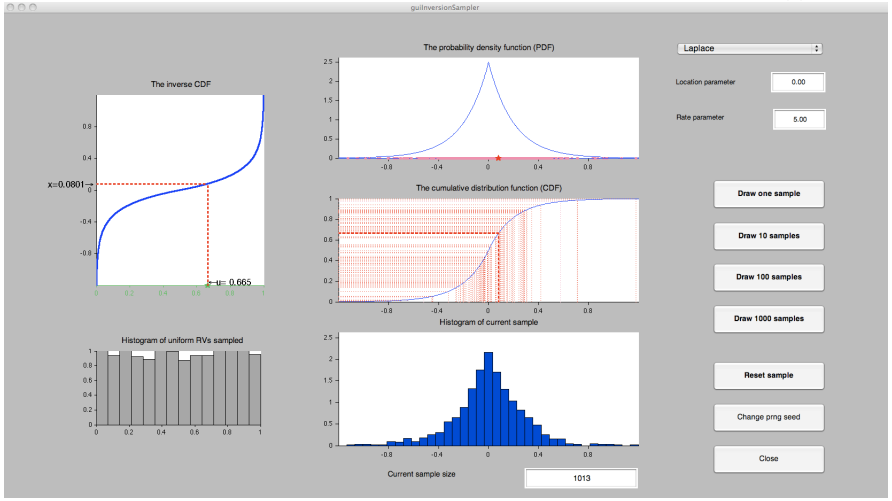
This distribution is unique because of its property of **memorylessness**, i.e., $\mathbf{P}(X > x+y | X > y) = e^{-\lambda x}$, and plays a fundamental role in modeling continuous time processes, such as time between occurrence of events of interest, as we will see in the sequel.

Labwork 136 (Rejection Sampler Demo – Laplace(5)**)** Let us comprehend the rejection sampler by calling the interactive visual cognitive tool:

```
>> guiInversionSampler
```

The M-file `guiInversionSampler.m` will bring a graphical user interface (GUI) as shown in Figure 4.8. Using the drop-down menu change from the default target distribution Uniform$(-5, 5)$ to Laplace(5). Now repeatedly push the "Draw one sample" button several times and comprehend the simulation process. You can also press "Draw 1000 samples" and see the density histogram of the generated samples. Next try changing the numbers in the "Rate parameter" box from 5.00 to 1.00 in order to alter the parameter λ of Laplace(λ) RV. If you are more adventurous then try to alter the number in the "Location parameter" box from 0.00 to some thing else, say 10.00. Although our formulation of Laplace(λ) implicitly had a location parameter of 0.00, we can easily introduce a location parameter μ into the PDF. With a pencil and paper try to rewrite the PDF in (4.2) with an additional location parameter μ.

Figure 4.8: Visual Cognitive Tool GUI: Inversion Sampling from $X \sim$ Laplace(5).



Simulation 137 (Laplace(λ)) Here is an implementation of an inversion sampler to draw IID samples from a Laplace(λ) RV X by transforming IID samples from the Uniform$(0, 1)$ RV U:

LaplaceInvCDF.m

```
function x = LaplaceInvCDF(u,lambda);
% Call Syntax: x = LaplaceInvCDF(u,lambda);
%              or LaplaceInvCDF(u,lambda);
%
% Input      : lambda = rate parameter > 0,
%              u = an 1 X n array of IID samples from Uniform[0,1] RV
% Output     : an 1Xn array x of IID samples from Laplace(lambda) RV
%                                or Inverse CDF of Laplace(lambda) RV
x=-(1/lambda)*sign(u-0.5) .* log(1-2*abs(u-0.5));
```

```
    0.4204    0.1219    2.2182    3.6692    0.9654    0.0093    1.7126
    2.1427    0.1281    0.8500    1.4065    0.1160    0.1324    0.2635
    0.6620    1.1729    0.6301    0.6375    0.3793    0.6525    0.8330
>> %sum up the outcomes of the sequence of 5 draws in each replicate
>> s=sum(a); disp(s);
    5.2216    2.1856    6.6378    7.6490    2.0168    1.3073    7.9990
```

Labwork 135 (Next seven buses at your bus-stop) Consider the problem of modelling the arrival of buses at a bus stop. Suppose that the time between arrivals is an Exponential($\lambda = 0.1$) RV X with a mean inter-arrival time of $1/\lambda = 10$ minutes. Suppose you go to your bus stop and zero a stop-watch. Simulate the times of arrival for the next seven buses as indicated by your stop-watch. Seed the fundamental sampler by your Student ID (eg. if your ID is `11424620` then type `rand('twister', 11424620);` just before the simulation). Hand in the code with the arrival times of the next seven buses at your ID-seeded bus stop.

The support of the Exponential(λ) RV is $\mathbb{R}_+ := [0, \infty)$. Let us consider a RV built by mirroring the Exponential(λ) RV about the origin with the entire real line as its support.

Model 19 (Laplace(λ) **or** Double Exponential(λ) **RV)** If a RV X is equally likely to be either positive or negative with an exponential density, then the Laplace(λ) or Double Exponential(λ) RV, with the rate parameter $\lambda > 0, \lambda \in \mathbb{R}$, may be used to model it. The density function for the Laplace(λ) RV given by $f(x; \lambda)$ is

$$f(x; \lambda) = \frac{\lambda}{2} e^{-\lambda|x|} = \begin{cases} \frac{\lambda}{2} e^{\lambda x} & \text{if } x < 0 \\ \frac{\lambda}{2} e^{-\lambda x} & \text{if } x \geq 0 \end{cases} . \tag{4.2}$$

Let us define the sign of a real number x by

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \, . \end{cases}$$

Then, the DF of the Laplace(λ) RV X is

$$F(x; \lambda) = \int_{-\infty}^{x} f(y; \lambda)\, dy = \frac{1}{2}\left(1 + \text{sign}(x)\left(1 - e^{-\lambda|x|}\right)\right) , \tag{4.3}$$

and its inverse DF is

$$F^{[-1]}(u; \lambda) = -\frac{1}{\lambda}\, \text{sign}\left(u - \frac{1}{2}\right) \log\left(1 - 2\left|u - \frac{1}{2}\right|\right) , \; u \in [0, 1] \tag{4.4}$$

Mean and Variance of Laplace(λ) **RV** X**:** Show that the mean of a Laplace(λ) RV X is

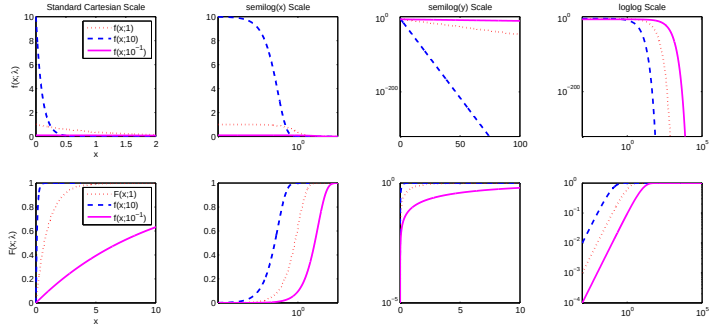
$$\mathbf{E}(X) = \int_0^\infty x f(x; \lambda)\, dx = \int_0^\infty x \frac{\lambda}{2} e^{-\lambda|x|}\, dx = 0 ,$$

and the variance is

$$\mathbf{V}(X) = \left(\frac{1}{\lambda}\right)^2 + \left(\frac{1}{\lambda}\right)^2 = 2\left(\frac{1}{\lambda}\right)^2 .$$

Note that the mean is 0 due to the symmetry of the density about 0 and the variance is twice that of the Exponential(λ) RV.

Figure 3.13: Density and distribution functions of Exponential(λ) RVs, for $\lambda = 1, 10, 10^{-1}$, in four different axes scales.



Example 56 (On a dark desert highway) At a certain location on a dark desert highway, the time in minutes between arrival of cars that exceed the speed limit is an Exponential($\lambda = 1/60$) random variable. If you just saw a car that exceeded the speed limit then what is the probability of waiting less than 5 minutes before seeing another car that will exceed the speed limit?

Solution:

The waiting time in minutes is simply given by the Exponential($\lambda = 1/60$) random variable. Thus, the desired probability is

$$P(0 \leq X < 5) \;=\; \int_0^5 \frac{1}{60} e^{-\frac{1}{60}x} dx \;=\; -e^{-\frac{1}{60}x}\Big]_0^5 \;=\; -e^{-\frac{1}{12}} + 1 \;\approx\; 0.07996.$$

In exam you can stop at the expression $-e^{-\frac{1}{12}} + 1$ for full credit. You may need a calculator for the last step (with answer 0.07996).

Note: We could use the distribution function directly:

$$P(0 \leq X < 5) \;=\; F\left(5; \frac{1}{60}\right) - F\left(0; \frac{1}{60}\right) \;=\; F\left(5; \frac{1}{60}\right) \;=\; 1 - e^{-\frac{1}{60}5} \;=\; 1 - -e^{-\frac{1}{12}} \;\approx\; 0.07996$$

Proposition 26 (Memorylessness of Exponential(λ) **RV)** If $X \sim$ Exponential(λ), then X has the property of **memorylessness**, i.e.,

$$\boxed{\mathbf{P}(X > x + y | X > y) = \mathbf{P}(X > x) \; .} \tag{3.30}$$

Proof: By the definition of conditional probability,

$$\mathbf{P}(X > x + y | X > y) = \mathbf{P}(\{X > x + y\} | \{X > y\}) = \frac{\mathbf{P}(\{X > x + y\} \cap \{X > y\})}{\mathbf{P}(\{X > y\})}$$

Due to redundancy, i.e., $\{X > x + y\} \subset \{X > y\} \implies \{X > x + y\} \cap \{X > y\} = \{X > x + y\}$, so

$$\begin{aligned} \mathbf{P}(X > x + y | X > y) &= \frac{\mathbf{P}(\{X > x + y\})}{\mathbf{P}(\{X > y\})} = \frac{\overline{F}(x + y; \lambda)}{\overline{F}(y; \lambda)} = \frac{e^{-\lambda(x+y)}}{e^{-\lambda y}} = \frac{e^{-\lambda x} e^{-\lambda y}}{e^{-\lambda y}} = e^{-\lambda x} = \overline{F}(x; \lambda) \\ &= \mathbf{P}(X > x) \end{aligned}$$

Exercise 3.17 (Memoryless Server Times) Suppose customers in a Queue are served one at a time by a server whose service time is an independent and identical Exponential(λ) RV, with $\lambda = 1/10$. The server is immediately free to serve the next customer once the current customer being served is done. Suppose you just arrive and are the first in the queue and know that the server is busy serving another customer. You do not know how long the customer has already been in service. What is the probability that the server will be free after 2 units of time?

Let us introduce parameters for the lower and upper bounds of the interval upon which a continuous RV is uniformly distributed using the following probability model.

Model 9 (Uniform(θ_1, θ_2)**)** Given two real parameters $\theta_1, \theta_2 \in \mathbb{R}$, such that $\theta_1 < \theta_2$, the PDF of the $Uniform(\theta_1, \theta_2)$ RV X is:

$$f(x; \theta_1, \theta_2) = \begin{cases} \frac{1}{\theta_2 - \theta_1} & \text{if } \theta_1 \leq x \leq \theta_2, \\ 0 & \text{otherwise} \end{cases} \tag{3.31}$$

and its DF given by $F(x; \theta_1, \theta_2) = \int_{-\infty}^{x} f(y; \theta_1, \theta_2)\, dy$ is:

$$F(x; \theta_1, \theta_2) = \begin{cases} 0 & \text{if } x < \theta_1 \\ \frac{x - \theta_1}{\theta_2 - \theta_1} & if\ \theta_1 \leq x \leq \theta_2, \\ 1 & \text{if } x > \theta_2 \end{cases} \tag{3.32}$$

Recall that we emphasise the dependence of the probabilities on the two parameters θ_1 and θ_2 by specifying them following the semicolon in the argument for f and F.

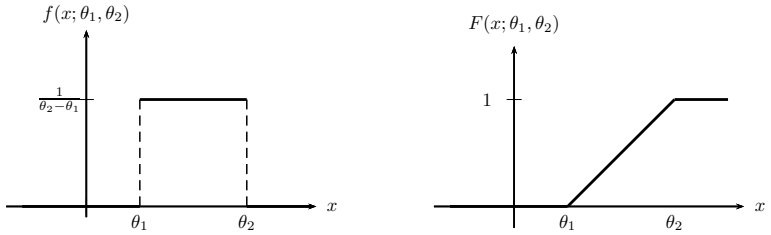


Figure 3.14: $f(x)$ and $F(x)$ of the Uniform(θ_1, θ_2) random variable X.

Exercise 3.18 Consider a random variable with a probability density function

$$f(x) = \begin{cases} k & \text{if } \quad 2 \leq x \leq 6 \\ 0 & \text{otherwise} \end{cases},$$

(a) Find the value of k.

(b) Sketch the graphs of $f(x)$ and $F(x)$.

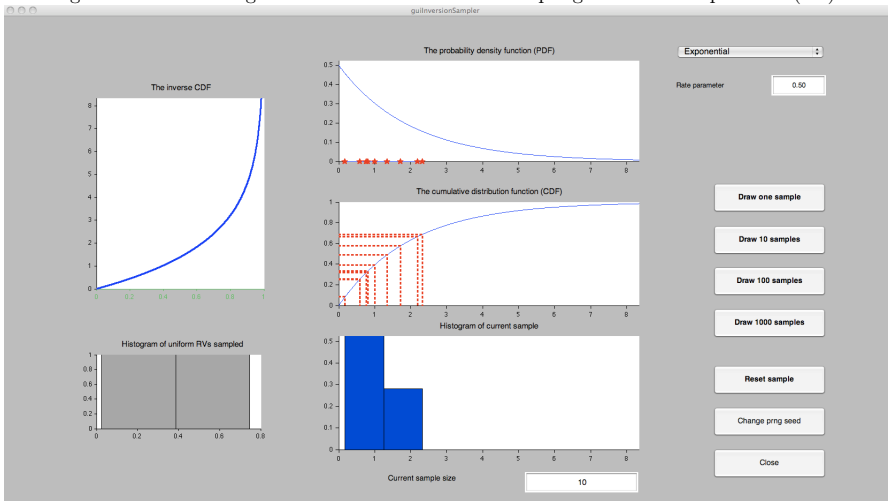

```
>> rand('twister',46678); % initialise the fundamental sampler
>> Lambda=1.0;  % declare Lambda=1.0
>> x=ExpInvSam(rand(1,5),Lambda); % pass an array of 5 Uniform(0,1) samles from rand
>> disp(x); % display the Exponential(1.0) distributed samples
    0.5945    2.5956    0.9441    1.9015    1.3973
```

Labwork 134 (Inversion Sampler Demo – Exponential(0.5)**)** Let us understand the inversion sampler by calling the interactive visual cognitive tool:

```
>> guiInversionSampler
```

The M-file `guiInversionSampler.m` will bring a graphical user interface (GUI) as shown in Figure 4.7. First change the target distribution from the default Uniform$(-5, 5)$ to Exponential(0.5) from the drop-down menu. Now push the "Draw 10 samples" button and comprehend the simulation process. Next try changing the "Rate Parameter" from 0.5 to 10.0 for example and generate several inversion samples and see the density histogram of the accumulating samples. You can press "Draw one sample" to really comprehend the inversion sampler in action one step at a time.

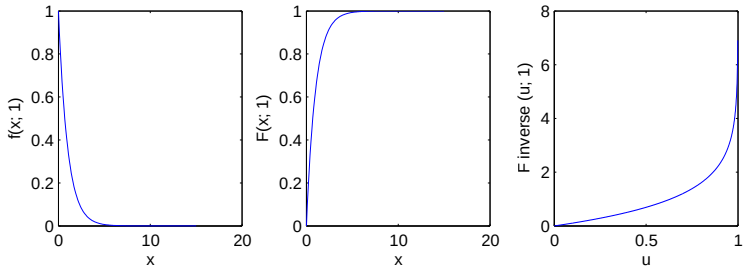
Figure 4.7: Visual Cognitive Tool GUI: Inversion Sampling from $X \sim$ Exponential(0.5).



It is straightforward to do replicate experiments. Consider the experiment of drawing five independent samples from the Exponential($\lambda = 1.0$) RV. Suppose we want to repeat or replicate this experiment seven times and find the sum of the five outcomes in each of these replicates. Then we may do the following:

```
>> rand('twister',1973); % initialise the fundamental sampler
>> % store 7 replications of 5 IID draws from Exponential(1.0) RV in array a
>> lambda=1.0; a= -1/lambda * log(rand(5,7)); disp(a);
    0.7267    0.3226    1.2649    0.4786    0.3774    0.0394    1.8210
    1.2698    0.4401    1.6745    1.4571    0.1786    0.4738    3.3690
```

Figure 4.6: The PDF f, DF F, and inverse DF $F^{[-1]}$ of the the Exponential($\lambda = 1.0$) RV.



Simulation 133 (Exponential(λ)) For a given $\lambda > 0$, an Exponential(λ) RV has the following PDF f, DF F and inverse DF $F^{[-1]}$:

$$f(x;\lambda) = \lambda e^{-\lambda x} \qquad F(x;\lambda) = 1 - e^{-\lambda x} \qquad F^{[-1]}(u;\lambda) = \frac{-1}{\lambda}\log_e(1-u) \tag{4.1}$$

We write the natural logarithm $\log_e$ as log for notational simplicity. An implementation of the Inversion Sampler for Exponential(λ) as a function in the M-file:

ExpInvCDF.m

```
function x = ExpInvCDF(u,lambda);
% Return the Inverse CDF of Exponential(lambda) RV X
% Call Syntax: x = ExpInvCDF(u,lambda);
%              ExpInvCDF(u,lambda);
% Input      : lambda = rate parameter,
%              u = array of numbers in [0,1]
% Output     : x
x=-(1/lambda) * log(1-u);
```

We can simply call the function to draw a sample from, say the Exponential($\lambda = 1.0$) RV by:

```
lambda=1.0;                    % some value for lambda
u=rand;                      % rand is the Fundamental Sampler
ExpInvCDF(u,lambda)       % sample from Exponential(1) RV via function in ExpInvCDF.m
```

Because of the following (recall Example 65):

$$U \sim \text{Uniform}(0,1) \implies -U \sim \text{Uniform}(-1,0) \implies 1-U \sim \text{Uniform}(0,1)\ ,$$

we could save a subtraction operation in the above algorithm by replacing `-(1/lambda) * log(1-u)` by `-(1/lambda) * log(u)`. Recall that the transformation of $U \sim \text{Uniform}(0,1)$ by $X = -(1/\lambda)\log(U)$ is exactly how we defined X as the Exponential(λ) RV in Model 8. This is implemented as the following function.

ExpInvSam.m

```
function x = ExpInvSam(u,lambda);
% Return the Inverse CDF based Sample from Exponential(lambda) RV X
% Call Syntax: x = ExpInvSam(u,lambda);
%              or ExpInvSam(u,lambda);
% Input      : lambda = rate parameter,
%              u = array of numbers in [0,1] from Uniform[0,1] RV
% Output     : x
x=-(1/lambda) * log(u);
```


The standard normal distribution is the most important continuous probability distribution. It was first described by De Moivre in 1733 and subsequently by C. F. Gauss (1777 - 1885). Many random variables have a normal distribution, or they are approximately normal, or can be transformed into normal random variables in a relatively simple fashion. Furthermore, the normal distribution is a useful approximation of more complicated distributions.

Model 10 (Normal$(0,1)$ **or standard normal or Gaussian RV)** A continuous random variable Z is called **standard normal** or **standard Gaussian** if its probability density function is

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)\ . \tag{3.33}$$

An exercise in calculus yields the first two derivatives of ϕ as follows:

$$\frac{d\phi}{dz} = -\frac{1}{\sqrt{2\pi}}z\exp\left(-\frac{z^2}{2}\right) = -z\phi(z), \quad \frac{d^2\phi}{dz^2} = \frac{1}{\sqrt{2\pi}}(z^2-1)\exp\left(-\frac{z^2}{2}\right) = (z^2-1)\phi(z)\ .$$

Thus, ϕ has a global maximum at 0, it is concave down if $z \in (-1,1)$ and concave up if $z \in (-\infty,-1)\cup(1,\infty)$. This shows that the graph of ϕ is shaped like a smooth symmetric bell centred at the origin over the real line.

Classwork 57 From the above exercise in calculus let us draw the graph of ϕ by hand now! Do it step by step: z^2, $-z^2$, $-z^2/2$, $\exp(-z^2/2)$, $\phi(z) = \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)$ now!

The distribution function of Z is given by

$$\Phi(z) \ = \ \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-v^2/2}\,dv\ . \tag{3.34}$$

Remark 27 The integral for $\Phi(z)$ has no closed form expression and cannot be evaluated exactly by standard methods of calculus, but its values can be obtained numerically and tabulated. Values of $\Phi(z)$ are tabulated in the "Standard Normal Distribution Function Table" in Sec. 6.6.

We can express $\Phi(z)$ in terms of the error function (erf) as follows:

$$\Phi(z) = \frac{1}{2}\ \text{erf}\left(\frac{z}{\sqrt{2}}\right) + \frac{1}{2} \tag{3.35}$$

And use MATLAB 's `erf` function to get $\Phi(z)$ numerically instead of looking up the Table.

Classwork 58 Note that the curve of $\Phi(z)$ is S-shaped, increasing in a strictly monotone way from 0 at $-\infty$ to 1 at ∞, and intersects the vertical axis at 1/2. Draw this by hand too.

just do it!

Example 59 Find the probabilities, using normal tables, that a random variable having the standard normal distribution will take on a value:

(a) less that 1.72

(b) less than -0.88

(c) between 1.30 and 1.75

(d) between -0.25 and 0.45

(a)

$$P(Z < 1.72) = \Phi(1.72) = 0.9573$$

(b) First note that $P(Z < 0.88) = 0.8106$, so that

$$\begin{aligned} P(Z < -0.88) &= P(Z > 0.88) \\ &= 1 - P(Z < 0.88) \\ &= 1 - \Phi(0.88) \\ &= 1 - 0.8106 = 0.1894 \end{aligned}$$

(c) $P(1.30 < Z < 1.75) = \Phi(1.75) - \Phi(1.30) = 0.9599 - 0.9032 = 0.0567$

(d)

$$\begin{aligned} P(-0.25 < Z < 0.45) &= P(Z < 0.45) - P(Z < -0.25) \\ &= P(Z < 0.45) - (1 - P(Z < 0.25)) \\ &= \Phi(0.45) - (1 - \Phi(0.25)) \\ &= (0.6736) - (1 - 0.5987) \\ &= 0.2723 \end{aligned}$$

CONTINUOUS RANDOM VARIABLES: NOTATION

$f(x)$: Probability density function (PDF)

- $f(x) \geq 0$
- Areas underneath $f(x)$ measure probabilities.

$F(x)$: Distribution function (DF)

- $0 \leq F(x) \leq 1$
- $F(x) = P(X \leq x)$ is a probability
- $F'(x) = f(x)$ for every x where $f(x)$ is continuous
- $F(x) = \int_{-\infty}^{x} f(v)dv$
- $P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(v)dv$

It is just as easy to draw n IID samples from Uniform(θ_1, θ_2) RV X by transforming n IID samples from the Uniform$(0, 1)$ RV as follows:

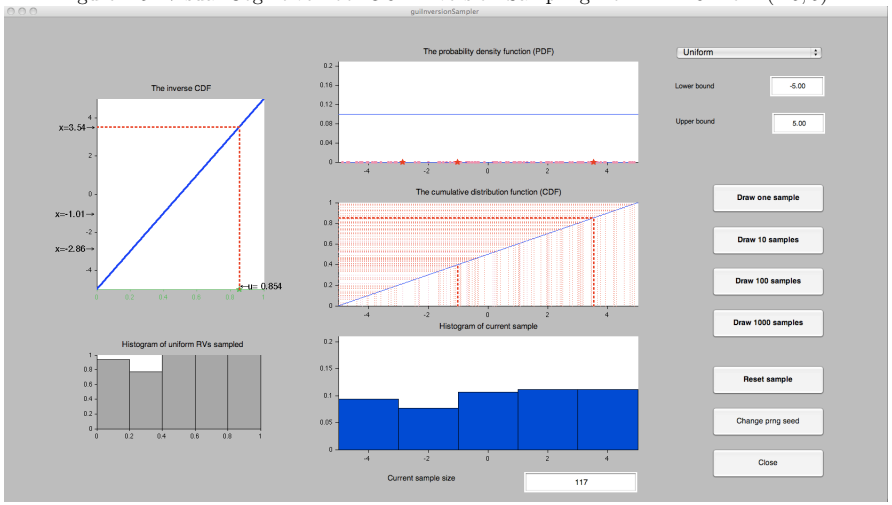
```
>> rand('twister',786543); % initialise the fundamental sampler
>> theta1=-83; theta2=1004; % declare values for parameters a and b
>> u=rand(1,5); % now u is an array of 5 samples from Uniform(0,1)
>> x=theta1+(theta2 - theta1)*u; % x is an array of 5 samples from Uniform(-83,1004]) RV
>> disp(x); % display the 5 samples just drawn from Uniform(-83,1004) RV
  465.3065  111.4994   14.3535  724.8881  254.0168
```

Labwork 132 (Inversion Sampler Demo – Uniform$(-5, 5)$**)** Let us comprehend the inversion sampler by calling the interactive visual cognitive tool built by Jennifer Harlow under a grant from University of Canterbury's Centre for Teaching and Learning (UCTL):

```
>> guiInversionSampler
```

The M-file `guiInversionSampler.m` will bring a graphical user interface (GUI) as shown in Figure 4.5. The default target distribution is Uniform$(-5, 5)$. Now repeatedly push the "Draw one sample" button several times and comprehend the simulation process. You can press "Draw 100 samples" to really comprehend the inversion sampler in action after 100 samples are drawn and depicted in the density histogram of the accumulating samples. Next try changing the numbers in the "Lower bound" and "Upper bound" boxes in order to alter the parameters θ_1 and θ_2 of Uniform(θ_1, θ_2) RV.

Figure 4.5: Visual Cognitive Tool GUI: Inversion Sampling from $X \sim$ Uniform$(-5, 5)$.



Recall the Exponential(λ) RV of Model 8. Let us simulate from it using the inversion sampler.

Let us consider the problem of simulating from an Exponential(λ) RV with realisations in $\mathbb{R}_+ := [0, \infty) := \{x : x \geq 0, x \in \mathbb{R}\}$ to model the waiting time for a bus at a bus stop.

Proof: The "one-line proof" of the proposition is due to the following equalities:

$$\mathbf{P}(F^{[-1]}(U) \leq x) = \mathbf{P}(\inf\{y : F(y) = U)\} \leq x) = \mathbf{P}(U \leq F(x)) = F(x), \quad \textit{for all } x \in \mathbb{R}.$$

This yields the inversion sampler or the inverse (C)DF sampler, where we (i) *generate* $u \sim$ Uniform$(0, 1)$ and (ii) *return* $x = F^{[-1]}(u)$, as formalised by the following algorithm.

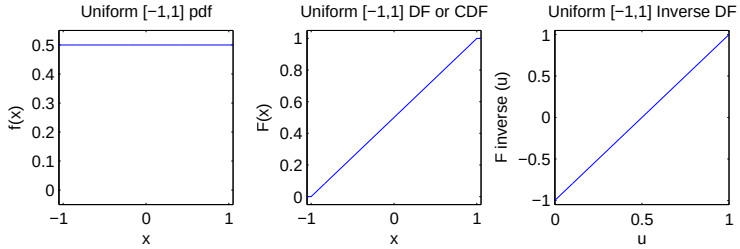
Algorithm 3 Inversion Sampler or Inverse (C)DF Sampler

1: *input:* (1) $F^{[-1]}(x)$, inverse of the DF of the target RV X, (2) the fundamental sampler
2: *initialise:* set the seed, if any, for the fundamental sampler
3: *output:* a sample from X distributed according to F
4: *draw* $u \sim$ Uniform$(0, 1)$
5: *return:* $x = F^{[-1]}(u)$

This algorithm emphasises the fundamental sampler's availability in an *input* step, and its set-up needs in an *initialise* step. In the following sections, we will not mention these universal steps; they will be taken for granted. The direct applicability of Algorithm 3 is limited to univariate densities for which the inverse of the cumulative distribution function is explicitly known. The next section will consider some examples.

Recall the Uniform(θ_1, θ_2) RV of Model 9 with the following PDF, DFand inverse DF. Let us simulate from it using the inversion sampler.

Figure 4.4: A plot of the PDF, DF or CDF and inverse DF of the Uniform$(-1, 1)$ RV X.



Simulation 131 (Uniform(θ_1, θ_2)**)** To simulate from Uniform(θ_1, θ_2) RV X using the Inversion Sampler, we first need to find $F^{[-1]}(u)$ by solving for x in terms of $u = F(x; \theta_1, \theta_2)$:

$$u = \frac{x - \theta_1}{\theta_2 - \theta_1} \quad \Longleftrightarrow \quad x = (\theta_2 - \theta_1)u + \theta_1 \quad \Longleftrightarrow \quad F^{[-1]}(u; \theta_1, \theta_2) = \theta_1 + (\theta_2 - \theta_1)u$$

Here is a simple implementation of the Inversion Sampler for the Uniform(θ_1, θ_2) RV in MATLAB :

```
>> rand('twister',786); % initialise the fundamental sampler for Uniform(0,1)
>> theta1=-1; theta2=1; % declare values for parameters theta1 and theta2
>> u=rand; % rand is the Fundamental Sampler and u is a sample from it
>> x=theta1+(theta2 - theta1)*u; % sample from Uniform(-1,1]) RV
>> disp(x); % display the sample from Uniform[-1,,1] RV
    0.5134
```


3.5 Exercises in Continuous Random Variables

Ex. 3.19 — Consider the probability density function

$$f(x) = \begin{cases} k & -4 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases} .$$

1.Find the value of k.
2.Find the distribution function, F.
3.Graph f and F.

Ex. 3.20 — Assume that a new light bulb will burn out at time t hours according to the probability density function given by

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

In this context, λ is often called the failure rate of the bulb.

(a)Assume that $\lambda = 0.01$, and find the probability that the bulb will not burn out before τ hours. This τ-specific probability is often called the reliability of the bulb.
Hint: Use the distribution function for an Exponential(λ) random variable (recall, $F(\tau; \lambda) = \int_{-\infty}^{\tau} f(t)dt$)!

(b)For what value of τ is the reliability of the bulb exactly $\frac{1}{2}$?

Ex. 3.21 — Let the random variable X be the time after which certain ball bearings wear out, with density

$$f(x) = \begin{cases} ke^{-x} & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases} .$$

Note: X is measured in years.

1.Find k.
2.Find the probability that a bearing will last at least 1 year.

3.6 Transformations of random variables

Suppose we know the distribution of a random variable X. How do we find the distribution of a transformation of X, say $g(X)$? Before we answer this question let us ask a motivational question. Why are we interested in functions of random variables?

Example 60 Consider a simple financial example where an individual sells X items per day, the profit per item is \$5 and the overhead costs are \$500 per day. The original random variable is X, but the random variable Y which gives the daily profit is of more interest, where

$$Y = 5X - 500 .$$

Example 61 In a cell-phone system a mobile signal may have a signal-to-noise-ratio of X, but engineers prefer to express such ratios in decibels, i.e.,

$$Y = 10 \log_{10}(X) .$$

3.6.1 A Review of Inverse Images

Hence in a great many situations we are more interested in functions of random variables. Let us return to our original question of determining the distribution of a transformation or function of X. First note that this transformation of X is itself another random variable, say $Y = g(X)$, where g is a function from a subset $\mathbb{X}$ of $\mathbb{R}$ to a subset $\mathbb{Y}$ of $\mathbb{R}$, i.e., $g : \mathbb{X} \to \mathbb{Y}$, $\mathbb{X} \subset \mathbb{R}$ and $\mathbb{Y} \subset \mathbb{R}$.

The **inverse image** of a set A is the set of all real numbers in $\mathbb{X}$ whose image is in A, i.e.,

$$g^{[-1]}(A) = \{x \in \mathbb{X} : g(x) \in A\} \ .$$

In other words,

$$x \in g^{[-1]}(A) \text{ if and only if } g(x) \in A \ .$$

For example,

- if $g(x) = 2x$ then $g^{[-1]}([4, 6]) = [2, 3]$
- if $g(x) = 2x + 1$ then $g^{[-1]}([5, 7]) = [2, 3]$
- if $g(x) = x^3$ then $g^{[-1]}([1, 8]) = [1, 2]$
- if $g(x) = x^2$ then $g^{[-1]}([1, 4]) = [-2, -1] \cup [1, 2]$
- if $g(x) = \sin(x)$ then $g^{[-1]}([-1, 1]) = \mathbb{R}$
- if ...

For the singleton set $A = \{y\}$, we write $g^{[-1]}(y)$ instead of $g^{[-1]}(\{y\})$. For example,

- if $g(x) = 2x$ then $g^{[-1]}(4) = \{2\}$
- if $g(x) = 2x + 1$ then $g^{[-1]}(7) = \{3\}$
- if $g(x) = x^3$ then $g^{[-1]}(8) = \{2\}$
- if $g(x) = x^2$ then $g^{[-1]}(4) = \{-2, 2\}$
- if $g(x) = \sin(x)$ then $g^{[-1]}(0) = \{k\pi : k \in \mathbb{Z}\} = \{\ldots, -3\pi, -2\pi, -\pi, 0, \pi, 2\pi, 3\pi, \ldots\}$
- if ...

If $g : \mathbb{X} \to \mathbb{Y}$ is one-to-one (injective) and onto (surjective), then the inverse image of a singleton set is itself a singleton set. Thus, the inverse image of such a function g becomes itself a function and is called the **inverse function**. One can find the inverse function, if it exists by the following steps:

Step 1; write $y = g(x)$

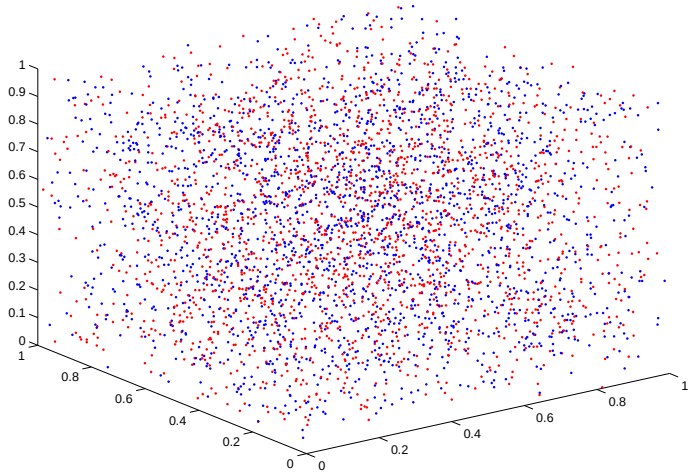
Step 2; solve for x in terms of y

Step 3; set $g^{-1}(y)$ to be this solution

We write g^{-1} whenever the inverse image $g^{[-1]}$ exists as an inverse function of g. Thus, the inverse function g^{-1} is a specific type of inverse image $g^{[-1]}$. For example,

- if $g(x) = 2x$ then $g : \mathbb{R} \to \mathbb{R}$ is injective and surjective and therefore its inverse function is: Step 1; $y = 2x$, Step 2; $x = \frac{y}{2}$, Step 3; $g^{-1}(y) = \frac{y}{2}$

Figure 4.3: Triplet point clouds from the "Mersenne Twister" with two different seeds (see Labwork 130). .



4.3 Simulation of non-Uniform$(0, 1)$ Random Variables

The Uniform$(0, 1)$ RV of Model 7 forms the foundation for random variate generation and simulation. This is appropriately called the fundamental model or experiment, since every other experiment can be obtained from this one.

Next, we simulate or generate samples from other RVs by making the following two assumptions:

1. independent samples from the Uniform$(0, 1)$ RV can be generated, and
2. real arithmetic can be performed exactly in a computer.

Both these assumptions are, in fact, not true and require a more careful treatment of the subject. We may return to these careful treatments later on.

4.3.1 Inversion Sampler for Continuous Random Variables

Proposition 58 (Inversion sampler) Let $F(x) := \int_{-\infty}^{x} f(y)\, dy : \mathbb{R} \to [0, 1]$ be a continuous DF with density f, and let its inverse $F^{[-1]} : [0, 1] \to \mathbb{R}$ be:

$$F^{[-1]}(u) := \inf\{x : F(x) = u\} \ .$$

Then, $F^{[-1]}(U)$ has the distribution function F, provided U is a Uniform$(0, 1)$ RV. Recall $\inf(A)$ or infimum of a set A of real numbers is the greatest lower bound of every element of A.

```
>> rand(1,10) % generate a 1 X 10 array of PRNs
ans =
    0.8147    0.9058    0.1270    0.9134    0.6324    0.0975    0.2785    0.5469    0.9575    0.9649
>> rand(1,10) % generate another 1 X 10 array of PRNs
ans =
    0.1576    0.9706    0.9572    0.4854    0.8003    0.1419    0.4218    0.9157    0.7922    0.9595
>> rand('twister',5489) % reset the PRNG to default state Mersenne Twister with seed=5489
>> rand(1,10) % reproduce the first array
ans =
    0.8147    0.9058    0.1270    0.9134    0.6324    0.0975    0.2785    0.5469    0.9575    0.9649
>> rand(1,10) % reproduce the second array
ans =
    0.1576    0.9706    0.9572    0.4854    0.8003    0.1419    0.4218    0.9157    0.7922    0.9595
```

In general, you can use any seed value to initiate your PRNG. You may use the `clock` command to set the seed:

```
>> SeedFromClock=sum(100*clock); % save the seed from clock
>> rand('twister',SeedFromClock) % initialize the PRNG
>> rand(1,10)
ans =
    0.3696    0.3974    0.6428    0.6651    0.6961    0.7311    0.8982    0.6656    0.6991    0.8606
>> rand(2,10)
ans =
    0.3432    0.9511    0.3477    0.1007    0.8880    0.0853    0.6067    0.6976    0.4756    0.1523
    0.5827    0.5685    0.0125    0.1555    0.5551    0.8994    0.2502    0.5955    0.5960    0.5700
>> rand('twister',SeedFromClock) % initialize the PRNG to same SeedFromClock
>> rand(1,10)
ans =
    0.3696    0.3974    0.6428    0.6651    0.6961    0.7311    0.8982    0.6656    0.6991    0.8606
```

Labwork 130 (3D plots of triplets generated by the "Mersenne Twister") Try to find any correlation between triplets generated by the "Mersenne Twister" by rotating the 3D plot generated by the following code:

```
>> rand('twister',1234)
>> x=rand(3,2000); % store PRNs in a 3X2000 matrix named x
>> plot3(x(1,:),x(2,:),x(3,:),'.')
```

Compare this with the 3D plot of triplets from `RANDU` of Labwork 127. Which of these two PRNGs do you think is "more random" looking? and why?

Change the seed value to the recommended default by the authors and look at the point cloud (in red) relative to the previous point cloud (in blue). Rotate the plots to visualise from multiple angles. Are they still random looking?

```
>> rand('twister',1234)% same seed as before
>> x=rand(3,2000); % store PRNs in a 3X2000 matrix named x
>> rand('twister',5489)% the recommended default seed
>> y=rand(3,2000);% store PRNs seeded by 5489 in a 3X2000 matrix named y
>> plot3(x(1,:),x(2,:),x(3,:),'b.') % plot triplets as blue dots
>> hold on;
>> plot3(y(1,:),y(2,:),y(3,:),'r.') % plot triplets as red dots
```


- if $g(x) = 2x + 1$ then $g : \mathbb{R} \to \mathbb{R}$ is injective and surjective and therefore its inverse function is:
  Step 1; $y = 2x + 1$, Step 2; $x = \frac{y-1}{2}$, Step 3; $g^{-1}(y) = \frac{y-1}{2}$
- if $g(x) = x^3$ then $g : \mathbb{R} \to \mathbb{R}$ is injective and surjective and therefore its inverse function is:
  Step 1; $y = x^3$, Step 2; $x = y^{\frac{1}{3}}$, Step 3; $g^{-1}(y) = y^{\frac{1}{3}}$

However, you need to be careful by limiting the domain to obtain the inverse function for the following examples:

- if $g(x) = x^2$ and domain of g is $[0, +\infty)$ then its inverse function is $g^{-1}(y) = \sqrt{y}$, i.e., if $g(x) = x^2 : [0, +\infty) \to [0, +\infty)$ then the inverse image $g^{[-1]}(y)$ for $y \in [0, +\infty)$ is given by the inverse function $g^{-1}(y) = \sqrt{y} : [0, +\infty) \to [0, +\infty)$.
- if $g(x) = x^2$ and domain of g is $(-\infty, 0]$ then its inverse function is $g^{-1}(y) = -\sqrt{y}$, i.e., if $g(x) = x^2 : (-\infty, 0] \to [0, +\infty)$ then the inverse image $g^{[-1]}(y)$ for $y \in [0, +\infty)$ is given by the inverse function $g^{-1}(y) = -\sqrt{y} : [0, +\infty) \to (-\infty, 0]$.
- if $g(x) = \sin(x)$ and domain of g is $[0, \frac{\pi}{2}]$ then its inverse function $g^{-1}(y) = \arcsin(y)$, i.e., if $g(x) = \sin(x) : [0, \frac{\pi}{2}] \to [0, 1]$ then the inverse image $g^{[-1]}(y)$ for $y \in [0, 1]$ is given by the inverse function $g^{-1}(y) = \arcsin(y) : [0, 1] \to [0, \frac{\pi}{2}]$.
- if $g(x) = \sin(x)$ and domain of g is $[-\frac{\pi}{2}, \frac{\pi}{2}]$ then its inverse function $g^{-1}(y) = \arcsin(y)$, i.e., if $g(x) = \sin(x) : [-\frac{\pi}{2}, \frac{\pi}{2}] \to [-1, 1]$ then the inverse image $g^{[-1]}(y)$ for $y \in [-1, 1]$ is given by the inverse function $g^{-1}(y) = \arcsin(y) : [-1, 1] \to [-\frac{\pi}{2}, \frac{\pi}{2}]$.
- if ...

Now, let us return to our question of determining the distribution of the transformation $g(X)$. To answer this question we must first observe that the inverse image $g^{[-1]}$ satisfies the following properties:

- $g^{[-1]}(\mathbb{Y}) = \mathbb{X}$
- For any set A, $g^{[-1]}(A^c) = \left(g^{[-1]}(A)\right)^c$
- For any collection of sets $\{A_1, A_2, \ldots\}$,
$$g^{[-1]}\left(A_1 \cup A_2 \cup \cdots\right) = g^{[-1]}(A_1) \cup g^{[-1]}(A_2) \cup \cdots \ .$$

Consequentially,

$$\boxed{P\left(g(X) \in A\right) = P\left(X \in g^{[-1]}(A)\right)} \tag{3.36}$$

satisfies the axioms of probability and gives the desired probability of the event A from the transformation $Y = g(X)$ in terms of the probability of the event given by the inverse image of A underpinned by the random variable X. It is crucial to understand this from the sample space Ω of the underlying experiment in the sense that Equation (3.36) is just short-hand for its actual meaning:

$$P\left(\{\omega \in \Omega : g(X(\omega)) \in A\}\right) = P\left(\left\{\omega \in \Omega : X(\omega) \in g^{[-1]}(A)\right\}\right) \ .$$

Because we have more than one random variable to consider, namely, X and its transformation $Y = g(X)$ we will subscript the probability density or mass function and the distribution function by the random varaible itself. For example we denote the distribution function of X by $F_X(x)$ and that of Y by $F_Y(y)$.

3.6.2 Transformations of discrete random variables

For a discrete random variable X with probability mass function f_X we can obtain the probability mass function f_Y of $Y = g(X)$ using Equation (3.36) as follows:

$$\begin{aligned} f_Y(y) &= \mathbf{P}(Y = y) = \mathbf{P}(Y \in \{y\}) \\ &= P\left(g(X) \in \{y\}\right) = P\left(X \in g^{[-1]}(\{y\})\right) \\ &= P\left(X \in g^{[-1]}(y)\right) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{x : g(x) = y\}} f_X(x) \ . \end{aligned}$$

This gives the formula:

$$\boxed{f_Y(y) = \mathbf{P}(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{x : g(x) = y\}} f_X(x) \ .} \tag{3.37}$$

Example 62 Let X be the discrete random variable with probability mass function f_X as tabulated below:

x	-1	0	1
$f_X(x) = \mathbf{P}(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

If $Y = 2X$ then the transformation $g(X) = 2X$ has inverse image $g^{[-1]}(y) = \{y/2\}$. Then, by Equation (3.37) the probability mass function of Y is expressed in terms of the known probabilities of X as:

$$f_Y(y) = \mathbf{P}(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{y/2\}} f_X(x) = f_X(y/2) \ ,$$

and tabulated below:

y	-2	0	2
$f_Y(y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

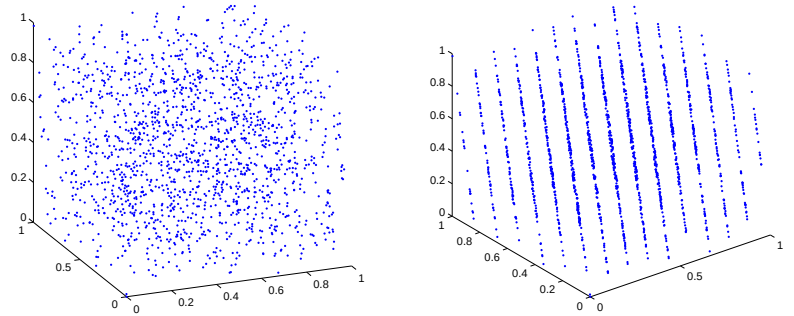
Example 63 If X is the random variable in the previous Example then what is the probability mass function of $Y = 2X + 1$? Once again,

$$f_Y(y) = \mathbf{P}(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{(y-1)/2\}} f_X(x) = f_X((y-1)/2) \ ,$$

and tabulated below:

y	-1	1	3
$f_Y(y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Figure 4.2: The LCG called `RANDU` with $(m, a, c) = (2147483648, 65539, 0)$ has strong correlation between three consecutive points as: $x_{i+2} = 6x_{k+1} - 9x_k$. The two plots are showing (x_i, x_{i+1}, x_{i+2}) from two different view points. .



The number of random numbers n should at most be about $m/1000$ in order to avoid the future sequence from behaving like the past. Thus, if $m = 2^{32}$ then a new generator, with a new suitable set of (m, a, c, x_0, n) should be adopted after the consumption of every few million pseudo-random numbers.

The LCGs are the least sophisticated type of PRNGs. They are easier to understand but are not recommended for intensive simulation purposes. The next section briefly introduces a more sophisticated PRNG we will be using in this course. Moreover our implementation of LCGs using the variable precision integer package is extremely slow in MATLAB and is only of pedagogical interest.

4.2.2 Generalized Feedback Shift Register and the "Mersenne Twister" PRNG

The following generator termed `twister` in MATLAB is recommended for use in simulation. It has extremely long periods, low correlation and passes most statistical tests (the DIEHARD statistical tests). The `twister` random number generator of Makoto Matsumoto and Takuji Nishimura is a variant of the twisted generalized feedback shift-register algorithm, and is known as the "Mersenne Twister" generator [Makoto Matsumoto and Takuji Nishimura, *Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator*, ACM Transactions on Modeling and Computer Simulation, Vol. 8, No. 1 (Jan. 1998), Pages 3–30]. It has a Mersenne prime period of $2^{19937} - 1$ (about 10^{6000}) and is **equi-distributed** in 623 dimensions. It uses 624 words of state per generator and is comparable in speed to the other generators. The recommended default seed is 5489. See `http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html` and `http://en.wikipedia.org/wiki/Mersenne_twister` for details.

Let us learn to implement the MATLAB function that generates PRNs. In MATLAB the function `rand` produces a deterministic PRN sequence. First, read `help rand`. We can generate PRNs as follows.

Labwork 129 (Calling PRNG in MATLAB) In MATLAB `rand` is basic PRNG command.

Since f maps a the finite set $\{1, 2, \ldots, m-1\}$ into itself, such systems are bound to have a repeating cycle of numbers called the **period**. In Labwork 124, the generator `LinConGen(13,12,11,10,12)` has period $(10, 1)$ of length 2, the generator `LinConGen (13,10,9,8,12)` has period $(8, 11, 2, 3, 0, 9)$ of length 6 and the generator `LinConGen(256,137,0,123,257)` has a period of length 32. All these generators have a non-maximal period length less than their modulus m. A good generator should have a maximal period of m. Let us try to implement a generator with a maximal period of $m = 256$.

The period of a general LCG is at most m, and for some choices of a the period can be much less than m as shown in the examples considered earlier. The LCG will have a full period if and only if:

1. c and m are relatively prime,
2. $a - 1$ is divisible by all prime factors of m,
3. $a - 1$ is a multiple of 4 if m is a multiple of 4

Labwork 126 (LCG with maximal period length of 256**)** Consider the linear congruential sequence with $(m, a, c, x_0, n) = (256, 137, 123, 13, 256)$. First check that these parameters do indeed satisfy the three condition above and therefore can produce the maximal period length of only $m = 256$. Modify the input parameter to `LinConGen` and repeat Labwork 125 in order to first produce a sequence of length 257. Do you see that the period is of maximal length of 256 as opposed to the generator of Labwork 125? Next produce a Figure to visualise the sequence as done in Figure 4.1.

A useful sequence should clearly have a relatively long period, say at least 2^{30}. Therefore, the **modulus m has to be rather large** because the **period** cannot have more than m elements. Moreover, the quality of pseudo-random numbers of a LCG is extremely sensitive to the choice of m, a and c even if the maximal period is attained. The next example illustrates this point.

Labwork 127 (The infamous `RANDU`) `RANDU` is an infamous LCG, which has been used since the 1960s. It is widely considered to be one of the most ill-conceived random number generators designed. Notably, it fails the **spectral test** badly for dimensions greater than 2. The following commands help visualise the sequence of first 5001/3 triplets (x_i, x_{i+1}, x_{i+2}) seeded from $x_0 = 1$ (Figure 4.2). Read `help reshape` and `help plot3`.

```
>> x=reshape( (LinConGen(2147483648,65539,0,1,5001)./ 2147483648) ,3,[]);
>> plot3(x(1,:),x(2,:),x(3,:),'.')
```

Labwork 128 (Fishman20 and Lecuyer21 LCGs) The following two LCGs are recommended in Knuth's Art of Computer Programming, vol. 2, for generating pseudo-random numbers for simple simulation tasks.

```
>> LinConGen(2147483647,48271,0,08787458,10) ./ 2147483647
ans =    0.0041    0.5239    0.0755    0.7624    0.6496    0.0769    0.9030    0.4259    0.9948    0.8868

>> LinConGen(2147483399,40692,0,01234567,10) ./ 2147483399
ans =    0.0006    0.3934    0.4117    0.7893    0.3913    0.6942    0.6790    0.3337    0.2192    0.1883
```


In fact, obtaining the probability of a one-to-one transformation of a discrete random variable as in Examples 62 and 63 is merely a matter of looking up the probability at the image of the inverse function. This is because there is only one term in the sum that appears in Equation (3.37). When the transformation is not one-to-one the number of terms in the sum can be more than one as shown in the next Example.

Example 64 Reconsider the random variable X of the last two Examples and let $Y = X^2$. Recall that $g(x) = x^2$ does not have an inverse function unless the domain is restricted to the positive or the negative parts of the real line. Since our random variable X takes values on both sides of the real line, namely $\{-1, 0, 1\}$, let us note that the transformation $g(X) = X^2$ is no longer a one-to-one function. Then, by Equation (3.37) the probability mass function of Y is expressed in terms of the known probabilities of X as:

$$f_Y(y) = \mathbf{P}(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{\{x:g(x)=y\}} f_X(x) = \sum_{\{x:x^2=y\}} f_X(x) \ ,$$

computed for each $y \in \{0, 1\}$ as follows:

$$\begin{aligned} f_Y(0) &= \sum_{\{x:x^2=0\}} f_X(x) = f_X(0) = \frac{1}{2} \ , \\ f_Y(1) &= \sum_{\{x:x^2=1\}} f_X(x) = f_X(-1) + f_X(1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \ , \end{aligned}$$

and finally tabulated below:

y	0	1
$f_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$

.

3.6.3 Transformations of continuous random variables

Suppose we know F_X and/or f_X of a continuous random variable X. Let $Y = g(X)$ be a transformation of X. Our objective is to obtain F_Y and/or f_Y of Y from F_X and/or f_X.

One-to-one transformations

The easiest case for transformations of continuous random variables is when g is **one-to-one and monotone**.

- First, let us consider the case when g is **monotone and increasing** on the range of the random variable X. In this case g^{-1} is also an increasing function and we can obtain the distribution function of $Y = g(X)$ in terms of the distribution function of X as

$$F_Y(y) = P\left(Y \leq y\right) = P\left(g(X) \leq y\right) = P\left(X \leq g^{-1}(y)\right) = F_X(g^{-1}(y)) \ .$$

  Now, let us use a form of chainrule to compute the density of Y as follows:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(g^{-1}(y)\right) = f_X\left(g^{-1}(y)\right) \frac{d}{dy}\left(g^{-1}(y)\right) \ .$$

- Second, let us consider the case when g is **monotone and decreasing** on the range of the random variable X. In this case g^{-1} is also a decreasing function and we can obtain the distribution function of $Y = g(X)$ in terms of the distribution function of X as

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P\left(X \geq g^{-1}(y)\right) = 1 - F_X(g^{-1}(y)) \ ,$$

and the density of Y as

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\left(1 - F_X\left(g^{-1}(y)\right)\right) = -f_X\left(g^{-1}(y)\right)\frac{d}{dy}\left(g^{-1}(y)\right) \ .$$

For a monotonic and decreasing g, its inverse function g^{-1} is also decreasing and consequently the density f_Y is indeed positive because $\frac{d}{dy}\left(g^{-1}(y)\right)$ is negative.

We can combine the above two cases and obtain the following **change of variable formula** for the probability density of $Y = g(X)$ when g is one-to-one and monotone on the range of X.

$$\boxed{f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| \ .} \tag{3.38}$$

The steps involved in finding the density of $Y = g(X)$ for a one-to-one and monotone g are:

1. Write $y = g(x)$ for x in range of x and check that $g(x)$ is monotone over the required range to apply the change of variable formula.
2. Write $x = g^{-1}(y)$ for y in range of y.
3. Obtain $\left|\frac{d}{dy}g^{-1}(y)\right|$ for y in range of y.
4. Finally, from Equation (3.38) get $f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|$ for y in range of y.

Let us use these four steps to obtain the density of monotone transformations of continuous random variables.

Example 65 Let X be Uniform$(0,1)$ random variable and let $Y = g(X) = 1 - X$. We are interested in the density of the tranformed random variable Y. Let us follow the four steps and use the change of variable formula to obtain f_Y from f_X and g.

1. $y = g(x) = 1 - x$ is a monotone decreasing function over $0 \leq x \leq 1$, the range of X. So, we can apply the change of variable formula.
2. $x = g^{-1}(y) = 1 - y$ is a monotone decreasing function over $1 - 0 \geq 1 - x \geq 1 - 1$, i.e., $0 \leq y \leq 1$.
3. For $0 \leq y \leq 1$,
$$\left|\frac{d}{dy}g^{-1}(y)\right| = \left|\frac{d}{dy}(1-y)\right| = |-1| = 1 \ .$$
4. we can use Equation (3.38) to find the density of Y as follows:
$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| = f_X(1-y)\,1 = 1 \ ,$$
for $0 \leq y \leq 1$


```
    x(i) = double(X); % convert to double
end
```

We can call it for some arbitrary input arguments as follows:

```
>> LinConGen(13,12,11,10,12)
ans =    10     1    10     1    10     1    10     1    10     1    10     1
>> LinConGen(13,10,9,8,12)
ans =     8    11     2     3     0     9     8    11     2     3     0     9
```

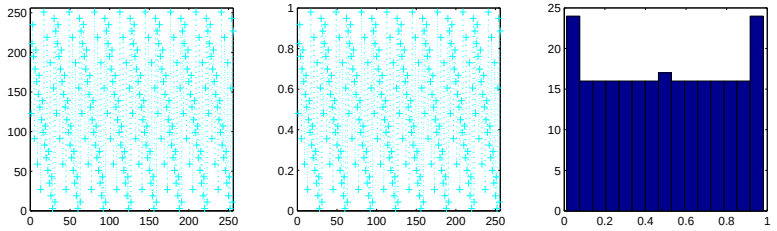
and observe that the generated sequences are not "random" for input values of (m, a, c, x_0, n) equalling $(13, 12, 11, 10, 12)$ or $(13, 10, 9, 8, 12)$. Thus, we need to do some work to determine the *suitable* input integers (m, a, c, x_0, n).

Labwork 125 (LCG with period length of 32) Consider the linear congruential sequence with $(m, a, c, x_0, n) = (256, 137, 0, 123, 257)$ with period length of only $32 < m = 256$. We can visualise the sequence as plots in Figure 4.1 after calling the following M-file.

LinConGenKnuth334T1L5Plots.m

```
LCGSeq=LinConGen(256,137,0,123,257)
subplot(1,3,1)
plot(LCGSeq,'c+:')
axis([0 256 0 256]); axis square
LCGSeqIn01=LCGSeq ./ 256
subplot(1,3,2)
plot(LCGSeqIn01,'c+:')
axis([0 256 0 1]); axis square
subplot(1,3,3)
hist(LCGSeqIn01,15)
axis square
```

Figure 4.1: The linear congruential sequence of `LinConGen(256,137,0,123,257)` with non-maximal period length of 32 as a line plot over $\{0, 1, \ldots, 256\}$, scaled over $[0, 1]$ by a division by 256 and a histogram of the 256 points in $[0, 1]$ with 15 bins.



Choosing the *suitable* magic input (m, a, c, x_0, n)

The linear congruential generator is a special case of a *discrete dynamical system*:

$$x_i = f(x_{i-1}), \ f : \{0, 1, 2, \ldots, m-1\} \to \{0, 1, 2, \ldots, m-1\} \text{ and } f(x_{i-1}) = (ax_{i-1} + c) \mod m \ .$$

In the following chapters, we focus on transforming IID Uniform$(0,1)$ variates to other non-uniform variates. In this chapter, we focus on the art of imitating IID Uniform$(0,1)$ variates using simple deterministic rules.

4.2.1 Linear Congruential Generators

The following procedure introduced by D. H. Lehmer in 1949 [*Proc. 2nd Symp. on Large-Scale Digital Calculating Machinery, Harvard Univ. Press, Cambridge, Mass., 1951, 141–146*] gives the simplest popular PRNG that can be useful in many statistical situations if used wisely.

Algorithm 2 Linear Congruential Generator (LCG)

1: *input:* five *suitable* integers:
   1. m, the modulus; $0 < m$
   2. a, the multiplier; $0 \leq a < m$
   3. c, the increment; $0 \leq c < m$
   4. x_0, the seed; $0 \leq x_0 < m$
   5. n, the number of desired pseudo-random numbers
2: *output:* $(x_0, x_1, \ldots, x_{n-1})$, the linear congruential sequence of length n
3: **for** $i = 1$ to $n-1$ **do**
4: $\quad x_i \leftarrow (ax_{i-1} + c) \mod m$
5: **end for**
6: *return:* $(x_1, x_2, \ldots, x_n)$

In order to implement LCGs we need to be able to do high precision exact integer arithmetic in MATLAB. We employ the Module `vpi` to implement variable precision integer arithmetic. You need to download this module for the next Labwork.

Labwork 124 (Generic Linear Congruential Sequence) Let us implement Algorithm 2 in MATLAB as follows.

LinConGen.m

```
function x = LinConGen(m,a,c,x0,n)
% Returns the linear congruential sequence
% Needs variable precision integer arithmetic in MATLAB!!!
% Usage: x = LinConGen(m,a,c,x0,n)
% Tested:
% Knuth3.3.4Table1.Line1: LinConGen(100000001,23,0,01234,10)
% Knuth3.3.4Table1.Line5: LinConGen(256,137,0,01234,10)
% Knuth3.3.4Table1.Line20: LinConGen(2147483647,48271,0,0123456,10)
% Knuth3.3.4Table1.Line21: LinConGen(2147483399,40692,0,0123456,10)

x=zeros(1,n); % initialize an array of zeros
X=vpi(x0); % X is a variable precision integer seed
x(1) = double(X); % convert to double
A=vpi(a); M=vpi(m); C=vpi(c); % A,M,C as variable precision integers
for i = 2:n % loop to generate the Linear congruential sequence
    % the linear congruential operation in variable precision integer
    % arithmetic
    % comment out the next ';' to get integer output
    X=mod(A * X + C, M);
```


Thus, we have shown that if X is a Uniform$(0,1)$ random variable then $Y = 1 - X$ is also a Uniform$(0,1)$ random variable.

Example 66 Let X be a Uniform$(0,1)$ random variable and let $Y = g(X) = -\log(X)$. We are interested in the density of the tranformed random variable Y. Once again, since g is a one-to-one monotone function let us follow the four steps and use the change of variable formula to obtain f_Y from f_X and g.

1. $y = g(x) = -\log(x)$ is a monotone decreasing function over $0 < x < 1$, the range of X. So, we can apply the change of variable formula.
2. $x = g^{-1}(y) = \exp(-y)$ is a monotone decreasing function over $0 < y < \infty$.
3. For $0 < y < \infty$,
$$\left|\frac{d}{dy} g^{-1}(y)\right| = \left|\frac{d}{dy}\left(\exp(-y)\right)\right| = |-\exp(-y)| = \exp(-y) \ .$$
4. We can use Equation (3.38) to find the density of Y as follows:
$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy} g^{-1}(y)\right| = f_X\left(\exp(-y)\right)\exp(-y) = 1\exp(-y) = \exp(-y) \ .$$
Note that $0 < \exp(-y) < 1$ for $0 < y < \infty$.

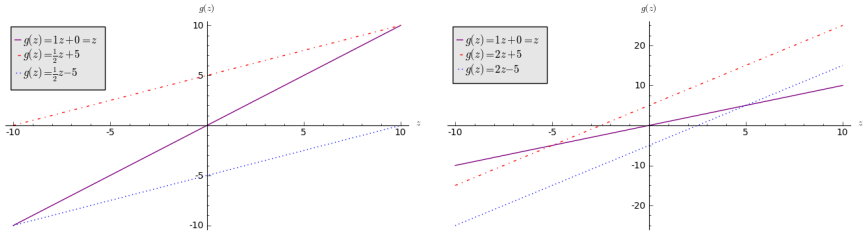
Thus, we have shown that if X is a Uniform$(0,1)$ random variable then $Y = -\log(X)$ is an random variable with PDF $f_Y(y) = \mathbb{1}_{(0,\infty)}(y)\exp(-y)$. We can similarly show that for a parameter $\lambda > 0$, if $X \sim$ Uniform$(0,1)$ then $Y = -\lambda^{-1}\log(X)$ yields a probability model of RVs that are parameterized by λ and extremely useful in applications. This is noting but our Exponential(λ) RV.

The next example yields the *l*ocation-scale family of normal random variables via a family of linear transformations of the standard normal random variable.

Example 67 Let Z be the standard Gaussian or standard normal random variable with probability density function $\phi(z)$ given by Equation (3.33). For real numbers $\sigma > 0$ and μ consider the linear transformation of Z given by
$$Y = g(Z) = \sigma Z + \mu \ .$$
Some graphs of such linear transformations of Z are shown in Figures (a) and (b).

We are interested in the density of the tranformed random variable $Y = g(Z) = \sigma Z + \mu$. Once again, since g is a one-to-one monotone function let us follow the four steps and use the change of variable formula to obtain f_Y from $f_Z = \phi$ and g.

1. $y = g(z) = \sigma z + \mu$ is a monotone increasing function over $-\infty < z < \infty$, the range of Z. So, we can apply the change of variable formula.
2. $z = g^{-1}(y) = (y - \mu)/\sigma$ is a monotone increasing function over the range of y given by, $-\infty < y < \infty$.
3. For $-\infty < y < \infty$,
$$\left|\frac{d}{dy} g^{-1}(y)\right| = \left|\frac{d}{dy}\left(\frac{y-\mu}{\sigma}\right)\right| = \left|\frac{1}{\sigma}\right| = \frac{1}{\sigma} \ .$$



(a) $g(z) = z$, $g(z) = \frac{1}{2}z + 5$ and $g(z) = \frac{1}{2}z - 5$. (b) $g(z) = z$, $g(z) = (z + 5)/0.5$ and $g(z) = (z - 5)/0.5$.

4. we can use Equation (3.38) and Equation (3.33) which gives

$$f_Z(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) ,$$

to find the density of Y as follows:

$$f_Y(y) = f_Z\left(g^{-1}(y)\right) \left|\frac{d}{dy} g^{-1}(y)\right| = \phi\left(\frac{y-\mu}{\sigma}\right) \frac{1}{\sigma} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right] ,$$

for $-\infty < y < \infty$.

Thus, we have obtained the expression for the probability density function of the linear transformation $\sigma Z + \mu$ of the standard normal random variable Z. This analysis leads to the following definition.

Model 11 (Normal(μ, σ^2) **RV)** Given a location parameter $\mu \in (-\infty, +\infty)$ and a scale parameter $\sigma^2 > 0$, the Normal(μ, σ^2) or Gaussian(μ, σ^2) random variable X has probability density function:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad (\sigma > 0) \ . \tag{3.39}$$

This is simpler than it may at first look. $f(x; \mu, \sigma^2)$ has the following features.

- μ is the expected value or mean parameter and σ^2 is the variance parameter. These concepts, mean and variance, are described in more detail in the next section on expectations.
- $1/(\sigma\sqrt{2\pi})$ is a constant factor that makes the area under the curve of $f(x)$ from $-\infty$ to ∞ equal to 1, as it must be.
- The curve of $f(x)$ is symmetric with respect to $x = \mu$ because the exponent is quadratic. Hence for $\mu = 0$ it is symmetric with respect to the y-axis $x = 0$.
- The exponential function decays to zero very fast — the faster the decay, the smaller the value of σ.

Chapter 4

Simulation

> "Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin." — John von Neumann (1951)

4.1 Physical Random Number Generators

Physical devices such as the BINGO machine demonstrated in class can be used to produce an integer uniformly at random from a finite set of possibilities. Such "ball bouncing machines" used in the British national lottery as well as the New Zealand LOTTO are complex nonlinear systems that are extremely sensitive to initial conditions ("chaotic" systems) and are physical approximations of the probability model called a "well-stirred urn" or an equi-probable de Moivre$(1/k, \ldots, 1/k)$ random variable.

Let us look at the New Zealand LOTTO draws at `http://lotto.nzpages.co.nz/statistics.html` and convince ourselves that all fourty numbers $\{1, 2, \ldots, 39, 40\}$ seem to be drawn uniformly at random. The British lottery animation at `http://understandinguncertainty.org/node/39` shows how often each of the 49 numbers came up in the first 1240 draws. Are these draws really random? We will answer these questions in the sequel (see `http://understandinguncertainty.org/node/40` if you can't wait).

4.2 Pseudo-Random Number Generators

Our probability model and the elementary continuous Uniform$(0, 1)$ RV are built from the abstract concept of a random variable over a probability triple. A direct implementation of these ideas on a computing machine is not possible. In practice, random variables are typically simulated by **deterministic** methods or algorithms. Such algorithms generate sequences of numbers whose behavior is virtually indistinguishable from that of truly random sequences. In computational statistics, simulating realisations from a given RV is usually done in two distinct steps. First, sequences of numbers that imitate independent and identically distributed (IID) Uniform$(0, 1)$ RVs are generated. Second, appropriate transformations are made to these imitations of IID Uniform$(0, 1)$ random variates in order to imitate IID random variates from other random variables or other random structures. These two steps are essentially independent and are studied by two non-overlapping groups of researchers. The formal term **pseudo-random number generator** (PRNG) or simply **random number generator** (RNG) usually refers to some deterministic algorithm used in the first step to produce pseudo-random numbers (PRNs) that imitate IID Uniform$(0, 1)$ random variates.

3.15 Exercises in Statistics

Ex. 3.56 — What is the sample mean and sample variance of the following dataset:

$$1, 3, 2, 1, 2, 3, 3$$

The normal distribution has the **distribution function**

$$F(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right] dv \ . \tag{3.40}$$

Here we need x as the upper limit of integration and so we write v in the integrand.

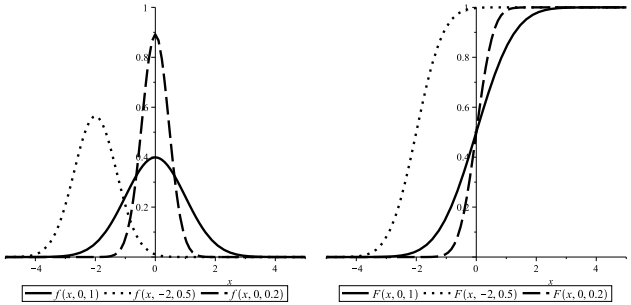


Figure 3.15: PDF and DF of a Normal(μ, σ^2) RV for different values of μ and σ^2

Using the direct method's Equation 3.41, we can obtain the distribution function of the Normal(μ, σ^2) random variable from that of the tabulated distribution function of the Normal$(0, 1)$ in the Standard normal distribution function table in Sec. 6.6.

Proposition 28 (One Table to Rule Them All Gaussians) The distribution function $F_X(x; \mu, \sigma^2)$ of the Normal(μ, σ^2) random variable X and the distribution function $F_Z(z) = \Phi(z)$ of the standard normal random variable Z are related by:

$$F_X(x; \mu, \sigma^2) \ = \ F_Z\left(\frac{x-\mu}{\sigma}\right) \ = \ \Phi\left(\frac{x-\mu}{\sigma}\right) \ .$$

Proof: Let Z be a Normal$(0, 1)$ random variable with distribution function $\Phi(z) = P(Z \leq z)$. We know that if $X = g(Z) = \sigma Z + \mu$ then X is the Normal(μ, σ^2) random variable. Therefore,

$$\begin{aligned} F_X(x; \mu, \sigma^2) &= P(X \leq x) = P\left(g(Z) \leq x\right) = P(\sigma Z + \mu \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right) \ . \end{aligned}$$

Hence we often transform a general Normal(μ, σ^2) random variable, X, to a standardised Normal$(0, 1)$ random variable, Z, by the substitution:

$$Z \ = \ \frac{X-\mu}{\sigma} \, .$$

Example 68 Suppose that the amount of cosmic radiation to which a person is exposed when flying by jet across the United States is a random variable, X, having a normal distribution with a mean of 4.35 mrem and a standard deviation of 0.59 mrem. What is the probability that a person will be exposed to more than 5.20 mrem of cosmic radiation on such a flight?

Solution:

$$\begin{aligned} P(X > 5.20) &= 1 - P(X \leq 5.20) \\ &= 1 - F(5.20) \\ &= 1 - \Phi\left(\frac{5.20 - 4.35}{0.59}\right) \\ &= 1 - \Phi(1.44) \\ &= 1 - 0.9251 \\ &= 0.0749 \end{aligned}$$

After some more notions you will see that Normal$(0, 1)$ RV can actually be obtained from an IID process of Bernoulli(θ) RVs. This is an instance of the central limit theorem. To appreciate this we first need to understand what we mean by statistics and then familiarise ourselves with notions of convergence of random variables.

Direct method

If the transformation g in $Y = g(X)$ is not necessarily one-to-one then special care is needed to obtain the distribution function or density of Y. For a continuous random variable X with a known distribution function F_X we can obtain the distribution function F_Y of $Y = g(X)$ using Equation (3.36) as follows:

$$\begin{aligned} F_Y(y) &= P\left(Y \leq y\right) = P\left(Y \in (-\infty, y]\right) \\ &= P\left(g(X) \in (-\infty, y]\right) = P\left(X \in g^{[-1]}((-\infty, y])\right) = P\left(X \in \{x : g(x) \in (-\infty, y]\}\right) \end{aligned} \quad (3.41)$$

In words, the above equalities just mean that the probability that $Y \leq y$ is the probability that X takes a value x that satisfies $g(x) \leq y$. We can use this approach if it is reasonably easy to find the set $g^{[-1]}((-\infty, y]) = \{x : g(x) = (-\infty, y]\}$.

Example 69 Let X be any random variable with distribution function F_X. Let $Y = g(X) = X^2$. Then we can find F_Y, the distribution function of Y from F_X as follows:

- Since $Y = X^2 \geq 0$, if $y < 0$ then $F_Y(y) = P\left(X \in \{x : x^2 < y\}\right) = \mathbf{P}(X \in \emptyset) = 0$.
- If $y \geq 0$ then

$$\begin{aligned} F_Y(y) = P\left(Y \leq y\right) &= P\left(X^2 \leq y\right) \\ &= P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \ . \end{aligned}$$

By differentiation we get:

Labwork 122 (favourite word cloud) This is just for fun. Produce a "word cloud" of your honours thesis or summer project or any other document that fancies your interest by using *wordle* from `http://www.wordle.net/`. Play with the aesthetic features to change colour, shapes, etc.

3.14.10 Machine Sensor Data

Instrumentation of modern machines, such as planes, rockets and cars allow the sensors in the machines to collect live data and dynamically take *decisions* and subsequent *actions* by executing algorithms to drive their devices in response to the data that is streaming into their sensors. For example, a rocket may have to adjust its boosters to compensate for the prevailing directional changes in wind in order to keep going up and launch a satellite. These types of decisions and actions, theorised by *controlled Markov processes*, typically arise in various fields of engineering such as, aerospace, civil, electrical, mechanical, robotics, etc.

In an observational setting, without an associated control problem, one can use machine sensor data to get information about some state of the system or phenomenon, i.e., what is it doing? or where is it?, etc. Sometimes sensors are attached to a sample of individuals from a wild population, say Emperor Penguins in Antarctica where the phenomenon of interest may be the diving habits of this species after the eggs hatch. As an other example we can attach sensors to a double pendulum and find what it is doing when we give it a spin.

Based on such observational data the experimenter typically tries to learn about the behaviour of the system from the sensor data to estimate parameters, test hypotheses, etc. Such types of experiments are typically performed by scientists in various fields of science, such as, astronomy, biology, chemistry, geology, physics, etc.

Chaotic Time Series of a Double Pendulum

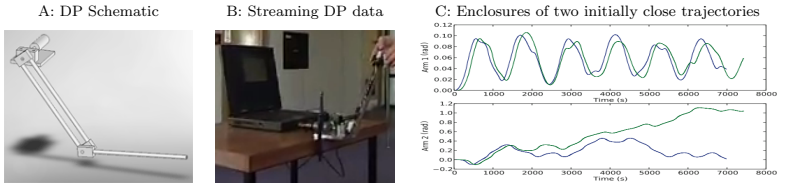


Figure 3.37: Double Pendulum

Sensors called *optical encoders* have been attached to the top end of each arm of a chaotic double pendulum in order to obtain the angular position of each arm through time as shown in Figure 3.37. Time series of the angular position of each arm for two trajectories that were initialized very similarly, say the angles of each arm of the double pendulum are almost the same at the initial time of release. Note how quickly the two trajectories diverge! System with such a sensitivity to initial conditions are said to be *chaotic*.

Labwork 123 (A Challenging Task) Try this if you are interested. Read any of the needed details about the design and fabrication of the double pendulum at `http://www.math.canterbury.ac.nz/~r.sainudiin/lmse/double-pendulum/`. Then use MATLAB to generate a plot similar to Figure 3.37(C) using time series data of trajectory 1 and trajectory 2 linked from the bottom of the above URL.

3.14.9 Textual Data

Processing and analysing textual data to make a decision is another important computational statistical experiment. An obvious example is machine translation and a less obvious one is exploratory data analysis of the textual content of

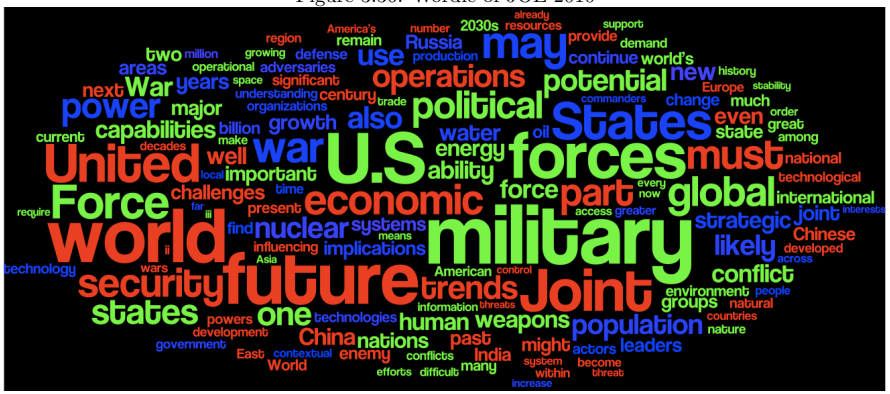
- a large document
- twitter messages within an online social network of interest
- etc.

An interesting document with a current affairs projection is the Joint Operating Environment 2010 Report by the US Department of Defense. This document was downloaded from `http://www.jfcom.mil/newslink/storyarchive/2010/JOE_2010_o.pdf`. The first paragraph of this 74 page document (JOE 2010 Reprort) reads:

ABOUT THIS STUDY The Joint Operating Environment is intended to inform joint concept development and experimentation throughout the Department of Defense. It provides a perspective on future trends, shocks, contexts, and implications for future joint force commanders and other leaders and professionals in the national security field. This document is speculative in nature and does not suppose to predict what will happen in the next twenty-five years. Rather, it is intended to serve as a starting point for discussions about the future security environment at the operational level of war. Inquiries about the Joint Operating Environment should be directed to USJFCOM Public Affairs, 1562 Mitscher Avenue, Suite 200, Norfolk, VA 23551-2488, (757) 836-6555.

Distribution Statement A: Approved for Public Release

Figure 3.36: Wordle of JOE 2010



We can try to produce a statistic of this document by recording the frequency of words in its textual content. Then we can produce a "word histogram" or "word cloud" to explore the document visually at one of the coarsest possible resolutions of the textual content in the JOE 2010 Report. The "word cloud" shown in Figure 3.36 was produced by Phillip Wilson using *wordle* from `http://www.wordle.net/`. A description from the wordle URL says:

Wordle is a toy for generating word clouds from text that you provide. The clouds give greater prominence to words that appear more frequently in the source text. You can tweak your clouds with different fonts, layouts, and color schemes. The images you create with Wordle are yours to use however you like. You can print them out, or save them to the Wordle gallery to share with your friends.

- If $y < 0$ then $f_Y(y) = \frac{d}{dy}(F_Y(y)) = \frac{d}{dy}0 = 0$.
- If $y \geq 0$ then

$$\begin{aligned} f_Y(y) = \frac{d}{dy}\left(F_Y(y)\right) &= \frac{d}{dy}\left(F_X(\sqrt{y}) - F_X(-\sqrt{y})\right) \\ &= \frac{d}{dy}\left(F_X(\sqrt{y})\right) - \frac{d}{dy}\left(F_X(-\sqrt{y})\right) \\ &= \frac{1}{2}y^{-\frac{1}{2}}f_X(\sqrt{y}) - \left(-\frac{1}{2}y^{-\frac{1}{2}}f_X(-\sqrt{y})\right) \\ &= \frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right) \ . \end{aligned}$$

Therefore, the distribution function of $Y = X^2$ is:

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \text{if } y \geq 0 \ . \end{cases} \tag{3.42}$$

and the probability density function of $Y = X^2$ is:

$$f_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right) & \text{if } y \geq 0 \ . \end{cases} \tag{3.43}$$

Example 70 If X is the standard normal random variable with density

$$f_X(x) = \phi(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-x^2/2\right)$$

then by Equation (3.43) the density of $Y = X^2$ is:

$$f_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right) = \frac{1}{\sqrt{2\pi y}}\exp\left(-\frac{y}{2}\right) & \text{if } y \geq 0 \ . \end{cases}$$

Y is called the **chi-square** random variable with one degree of freedom. This distribution plays a fundamental role in hypothesis testing as we will see in Inference Theory and was derived at the beginning of last century to settle "supposedly evidence-based disputes" among scientists using mathematics.

3.7 Exercises in Transformations of Random Variables

Ex. 3.22 — Let X be the outcome of a fair die roll with probability mass function given by

$$f_X(x) = \begin{cases} \frac{1}{6} & \text{if } x \in \{1,2,3,4,5,6\} \\ 0 & \text{otherwise}. \end{cases}$$

If $Y = (X-3)^2$ then find the probability mass function of Y, $f_Y(y)$.

Ex. 3.23 — Given a natural number n as a parameter, i.e., given a parameter $n \in \{1,2,3,\ldots\}$, let X be a discrete uniform random variable on the finite set

$$\mathbb{X} = \{-n, -n+1, \ldots, -1, 0, 1, \ldots, n-1, n\}$$

i.e. the probability mass function of X is:

$$f_X(x;n) = \begin{cases} \frac{1}{2n+1} & \text{if } x \in \mathbb{X} \\ 0 & \text{otherwise}. \end{cases}$$

Find the probability mass function $f_Y(y;n)$ for $Y = |X|$, the absolute value of X.

Ex. 3.24 — If X is a Geometric(θ) random variable and $Y = \left(\frac{1}{2}\right)^X$ then find an expression for $f_Y(y)$.

Ex. 3.25 — If X is a Poisson(λ) random variable find the probability mass function, $f_Y(y)$, of

$$Y = \frac{1}{(X+1)^2} .$$

Ex. 3.26 — If X is a continuous random variable with probability density function

$$f_X(x) = \begin{cases} xe^{-x} & x \geqslant 0 \\ 0 & x < 0 \end{cases},$$

find the probability density function of $Y = e^X$.

Ex. 3.27 — If X, the received power at an antenna is an Exponential(λ) random variable then find the probability density function of the amplitude $Y = \sqrt{X}$.

Ex. 3.28 — If X is a Uniform(a,b) random variable where $0 < a < b$, find the probability density function, $f_Y(y)$, of

$$Y = \log_e(X) .$$

3.8 Expectations

Expectation is perhaps the most fundamental concept in probability theory. In fact, probability is itself an expectation as you will soon see!

Expectation is one of the fundamental concepts in probability. The expected value of a real-valued random variable gives the population mean, a measure of the centre of the distribution of the variable in some sense. Its variance measures its spread and so on.


```
clf % clears all current figures

% Daily max and min temperature in the 100 days with good data
% before last date in this data, i.e, March 27 2011 in Christchurch NZ
H365Days = T(end-365:end,2);
L365Days = T(end-365:end,3);
F365Days = H365Days-L365Days; % assign the maximal fluctuation, i.e. max-min
plot(H365Days,'r*') % plot daily high or maximum temperature = Tmax
hold on;     % hold the Figure so that we can overlay more plots on it
plot(L365Days,'b*') % plot daily low or minimum temperature = Tmin
plot(F365Days, 'g*') % plot daily Fluctuation = Tmax - Tmin
% filter for running means
windowSize = 7;
WeeklyHighs = filter(ones(1,windowSize)/windowSize,1,H365Days);
plot(WeeklyHighs,'r.-')
WeeklyLows = filter(ones(1,windowSize)/windowSize,1,L365Days);
plot(WeeklyLows,'b.-')
WeeklyFlucs = filter(ones(1,windowSize)/windowSize,1,F365Days);
plot(WeeklyFlucs,'g.-')
xlabel('Number of days since March 27 2010 in Christchurch, NZ','FontSize',20);
ylabel('Temperature in Celsius','FontSize',20)
MyLeg = legend('Daily High','Daily Low',' Daily Fluc.','Weekly High','Weekly Low',...
      'Weekly Fluc.','Location','NorthEast')
% Create legend
% legend1 = legend(axes1,'show');
set(MyLeg,'FontSize',20);
xlim([0 365]); % set the limits or boundary on the x-axis of the plots
hold off % turn off holding so we stop overlaying new plots on this Figure
```

Figure 3.35: Daily temperatures in Christchurch for one year since March 27 2010

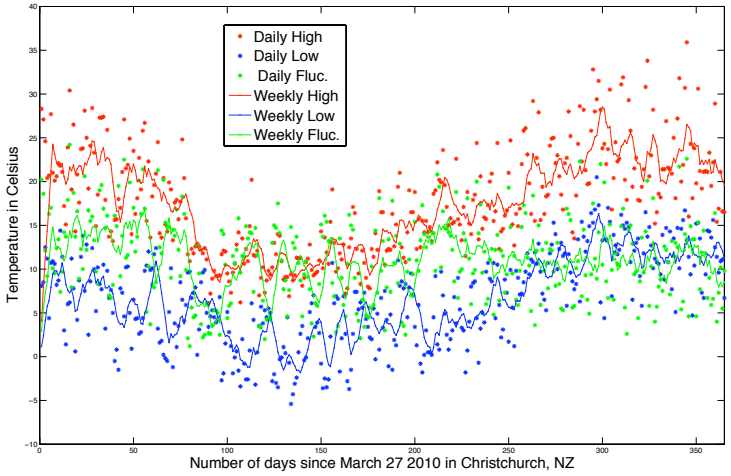
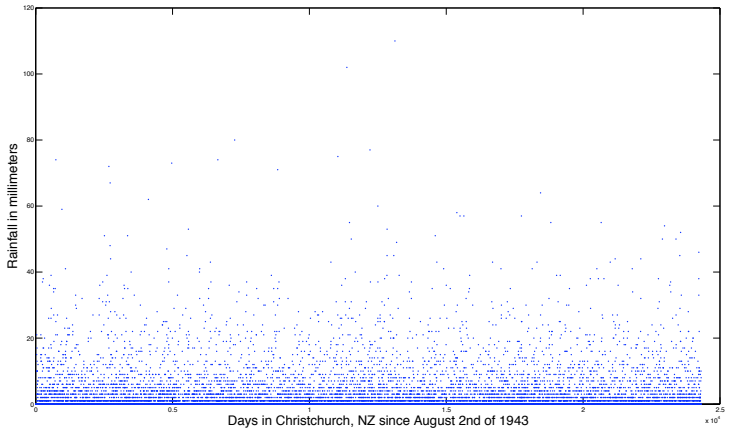


Figure 3.34: Daily rainfalls in Christchurch since March 27 2010



Daily Temperatures in Christchurch

Labwork 121 Understand how Figure 3.35 is being generated by following the comments in the script file `ChchTempsLoad.m` by typing:

```
>> ChchTempsLoad
```

ChchTempsLoad.m

```
%% Load the data from the comma delimited text file 'NIWACliFloChchAeroClubStationTemps.txt'
%% with the following column IDs
%% Max_min: Daily Temperature in Christchurch New Zealand
%% Stationate(NZST),Tmax(C),Period(Hrs),Tmin(C),Period(Hrs),Tgmin(C),Period(Hrs),Tmean(C),RHmean(%),Period(Hrs)

% the matrix T is about to be assigned the data as a matrix; the option [27,1,20904,5] to
% specify the upper-left and lower-right corners of an imaginary rectangle
% over the text file 'NIWACliFloChchAeroClubStationTemps.txt'.
% here we start from line number 27 and end at the last line number 20904
% and we read only columns NZST,Tmax(C),Period(Hrs),Tmin(C),Period(Hrs)

T = dlmread('NIWACliFloChchAeroClubStationTemps.txt',',',[27,1,20904,5]);
% just keep column 1,2 and 4 named NZST,Tmax(C),Period(Hrs),Tmin(C),
% i.e. date in YYYYMMDD foramt, maximum temperature, minimum temperature
T = T(:,[1,2,4]); % just pull the time
% print size before removing missig data rows are removed
size(T) % report the dimensions or size of the matrix T


% This file has a lot of missing data points and they were replaced with
% NaN values - see the file for various manipulations that were done to the
% raw text file from NIWA (Copyright NIWA 2011 Subject to NIWA's Terms and
% Conditions.  See: http://cliflo.niwa.co.nz/pls/niwp/doc/terms.html)
T(any(isnan(T),2),:) = [];% Remove any rows containing NaNs from a matrix

size(T) % if the matrix has a different size now then the data-less days now!
```


Definition 29 (Expectation of a RV) The **expectation**, or **expected value**, or **mean**, or **first moment**, of a random variable X, with distribution function F and density f, is defined to be

$$\mathbf{E}(X) := \int x\, dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x)\, dx & \text{if } X \text{ is continuous} , \end{cases} \tag{3.44}$$

provided the sum or integral is well-defined. We say the expectation exists if

$$\int |x|\ dF(x) < \infty . \tag{3.45}$$

Sometimes, we denote $\mathbf{E}(X)$ by $\mathbf{E}X$ for brevity. Thus, the expectation is a single-number summary of the RV X and may be thought of as the average. We subscript E to specify the parameter $\theta \in \mathbf{\Theta}$ with respect to which the integration is undertaken.

$$\mathbf{E}_\theta X := \int x\, dF(x;\theta)$$

Definition 30 (Variance of a RV) Let X be a RV with mean or expectation $\mathbf{E}(X)$. Variance of X denoted by $\mathbf{V}(X)$ or VX is

$$\mathbf{V}(X) := \mathbf{E}\left((X - \mathbf{E}(X))^2\right) = \int (x - \mathbf{E}(X))^2\, dF(x) ,$$

provided this expectation exists. The **standard deviation** denoted by $\mathrm{sd}(X) := \sqrt{\mathbf{V}(X)}$. Thus variance is a measure of "spread" of a distribution.

Definition 31 (k-th moment of a RV) We call

$$\mathbf{E}(X^k) = \int x^k\, dF(x)$$

as the k-th moment of the RV X and say that the k-th moment exists when $\mathbf{E}(|X|^k) < \infty$. We call the following expectation as the k-th central moment:

$$\mathbf{E}\left((X - \mathbf{E}(X))^k\right) .$$

3.8.1 Expectations of functions of random variables

More generally, by taking the expected value of various functions of a random variable, we can measure many interesting features of its distribution, including spread and correlation.

Definition 32 (Expectation of a function of a RV) The **Expectation** of a function $g(X)$ of a random variable X is defined as:

$$\mathbf{E}(g(X)) \ := \int g(x) dF(x) = \begin{cases} \sum\limits_x g(x) f(x) & \text{if } X \text{ is a discrete RV} \\ \int_{-\infty}^{\infty} g(x) f(x) dx & \text{if } X \text{ is a continuous RV} \end{cases}$$

provided $\mathbf{E}(g(X))$ exists, i.e., $\int |g(x)| dF(x) < \infty$.

The **mean** which characterises the central location of the random variable X is merely the expectation of the identity function $g(x) = x$:

$$\mathbf{E}(X) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is a discrete RV} \\ \int_{-\infty}^{\infty} x f(x) dx & \text{if } X \text{ is a continuous RV} \end{cases}$$

Often, mean is denoted by μ.

The **variance** which characterises the spread or the variability of the random variable X is also the expectation of the function $g(x) = (x - \mathbf{E}(X))^2$:

$$\mathbf{V}(X) = \mathbf{E}\left((X - \mathbf{E}(X))^2\right) = \begin{cases} \sum_x (x - \mathbf{E}(X))^2 f(x) & \text{if } X \text{ is a discrete RV} \\ \int_{-\infty}^{\infty} (x - \mathbf{E}(X))^2 f(x) dx & \text{if } X \text{ is a continuous RV} \end{cases}$$

Often, variance is denoted by σ^2.

INTUITIVELY, WHAT IS EXPECTATION?

Definition 32 gives expectation as a "weighted average" of the possible values. This is true but some intuitive idea of expectation is also helpful.

- Expectation is what you expect.

  Consider tossing a fair coin. If it is heads you lose $10. If it is tails you win $10. What do you expect to win? Nothing. If X is the amount you win then

  $$\mathbf{E}(X) = -10 \times \frac{1}{2} + 10 \times \frac{1}{2} = 0\,.$$

  So what you expect (nothing) and the weighted average ($\mathbf{E}(X) = 0$) agree.

- Expectation is a long run average.

  Suppose you are able to repeat an experiment independently, over and over again. Each experiment produces one value x of a random variable X. If you take the average of the x values for a large number of trials, then this average converges to $\mathbf{E}(X)$ as the number of trials grows. In fact, this is called the **law of large numbers**.

We can concretize the above two intuitive insights by the following two examples.

Example 71 (Winnings on Average) Let $Y = r(X)$. Then

$$\mathbf{E}(Y) = \mathbf{E}(r(X)) = \int r(x)\, dF(x)\ .$$

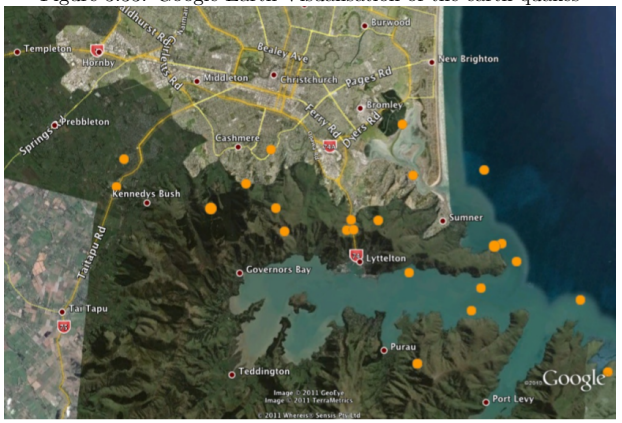
Think of playing a game where we draw $x \sim X$ and then I pay you $y = r(x)$. Then your average income is $r(x)$ times the chance that $X = x$, summed (or integrated) over all values of x.

Example 72 (Probability is an Expectation) Let A be an event and let $r(X) = \mathbb{1}_A(x)$. Recall $\mathbb{1}_A(x)$ is 1 if $x \in A$ and $\mathbb{1}_A(x) = 0$ if $x \notin A$. Then

$$\mathbf{E}(\mathbb{1}_A(X)) = \int \mathbb{1}_A(x)\, dF(x) = \int_A dF(x) = \mathbf{P}(X \in A) = \mathbf{P}(A) \tag{3.46}$$

Thus, probability is a special case of expectation. Recall our LTRF motivation for the definition of probability and make the connection.

Figure 3.33: Google Earth Visualisation of the earth quakes



We will explore some data of rainfall and temperatures from NIWA.

Daily Rainfalls in Christchurch

Automagic downloading of the data by Method B can be done if the data provider allows automated queries. It can be accomplished by `urlread` for instance.

Paul Brouwers has a basic CliFlo datafeed on `http://www.math.canterbury.ac.nz/php/lib/cliflo/rainfall.php`. This returns the date and rainfall in milli meters as measured from the CHCH aeroclub station. It is assumed that days without readings would not be listed. The data doesn't go back much before 1944.

Labwork 120 Understand how Figure 3.34 is obtained by the script file `RainFallsInChch.m` by typing and following the comments:

```
>> RainFallsInChch
RainFallsChch =     [24312x1 int32]    [24312x1 double]
ans =       24312           2
FirstDayOfData =    19430802
LastDayOfData =     20100721
```

RainFallsInChch.m

```
%% How to download data from an URL directly without having to manually
%% fill out forms
% first make a string of the data using urlread (read help urlread if you want details)
StringData = urlread('http://www.math.canterbury.ac.nz/php/lib/cliflo/rainfall.php');
RainFallsChch = textscan(StringData, '%d %f', 'delimiter', ',')
RC = [RainFallsChch{1} RainFallsChch{2}]; % assign Matlab cells as a matrix
size(RC) % find the size of the matrix

FirstDayOfData = min(RC(:,1))
LastDayOfData = max(RC(:,1))

plot(RC(:,2),'.')
xlabel('Days in Christchurch, NZ since August 2nd of 1943','FontSize',20);
ylabel('Rainfall in millimeters','FontSize',20)
```

NZEQChCch20110222EDA.m

```
%% Load the data from the comma delimited text file 'NZ20110222earthquakes.csv'
% using the script M-file NZEQChCch20110222.m
NZEQChCch20110222
%% working with time stamps is tricky
%% time is encoded by columns 6 through 11
%% as origin of earthquake in year, month, day, hour, minute, sec:
%% ORI_YEAR,ORI_MONTH,ORI_DAY,ORI_HOUR,ORI_MINUTE,ORI_SECOND
%% datenum is Matlab's date encoding function see help datenum
TimeData=datenum(EQ(:,6:11)); % assign origin times of earth quakes in datenum coordinates
MaxD=max(TimeData); % get the latest time of observation in the data
MinD=min(TimeData); % % get the earliest time of observation in the data
datestr(MinD) % a nice way to conver to calendar time!
datestr(MaxD) % ditto

% recall that there four variables were assigned in NZEQChCch20110222.m
% LatData=EQ(:,2); LonData=EQ(:,3); MagData=EQ(:,12); DepData=EQ(:,13);

%clear any existing Figure windows
clf
plot(TimeData,MagData,'o-') % plot origin time against magnitude of each earth quake

figure % tell matlab you are about to make another figure
plotmatrix([LatData,LonData,MagData,DepData],'r.');

figure % tell matlab you are about to make another figure
scatter(LonData,LatData,'.') % plot the LONGitude Vs. LATtitude

figure % tell matlab you are about to make another figure
% relative frequency histogram of magnitudes from 0 to 12 on Richter Scale with 15 bins
hist(MagData,15)

%max(MagData)

figure % tell matlab you are about to make another figure
semilogx(DepData,MagData,'.') % see the depth in log scale

%%%%%%%
% more advanced topic - uncomment and read help if bored
%tri = delaunay(LatData,LonData);
%triplot(tri,LatData,LonData,DepData);
```

Geostatistical exploratory data analysis with Google Earth

A global search at `http://neic.usgs.gov/cgi-bin/epic/epic.cgi` with the following parameters:

```
Date Range: 2011 2 22 to 2011 2 22
Catalog: USGS/NEIC (PDE-Q)
```

produced 43 earth quakes world-wide, including those in Christchurch as shown in Figure 3.33. One can do a lot more than a mere visualisation with the USGS/NEIC database of earth-quakes wolrd-wide, the freely available `Google earth` software bundle `http://www.google.com/earth/index.html` and the freely available MATLAB package `googleearth` from `http://www.mathworks.com/matlabcentral/fx_files/12954/4/content/googleearth/html/html_product_page.html`.

3.14.8 Metereological Data

New Zealand's meteorological service NIWA provides weather data under it TERMS AND CONDITIONS FOR ACCESS TO DATA (See `http://cliflo.niwa.co.nz/doc/terms_print.html`).

Expectations of functions of $\mathbb{R}^2$-valued random variables

In the case of a single random variable we saw that its expectation gives the population mean, a measure of the center of the distribution of the variable in some sense. Similarly, by taking the expected value of various functions of a $\mathbb{R}^2$-valued random variable, we can measure many interesting features of its joint distribution.

Definition 33 The **Expectation** of a function $g(X,Y)$ of the $\mathbb{R}^2$-valued RV (X,Y) is defined as:

$$\mathbf{E}(g(X,Y)) = \begin{cases} \sum_{(x,y)} g(x,y) f_{X,Y}(x,y) & \text{if } (X,Y) \text{ is a discrete R}\vec{\text{V}} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)dxdy & \text{if } (X,Y) \text{ is a continuous R}\vec{\text{V}} \end{cases}$$

Some typical expectations for $\mathbb{R}^2$-valued random variables are:

1. Joint Moments
   $$\mathbf{E}(X^r Y^s)$$
   When $r = s = 1$, we have $\mathbf{E}(XY)$, the expectation of the product of two RVs.
2. We need a new notion for the variance of two RVs.
   If $\mathbf{E}(X^2) < \infty$ and $\mathbf{E}(Y^2) < \infty$ then $\mathbf{E}(|XY|) < \infty$ and $\mathbf{E}(|(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))|) < \infty$. This allows the definition of **covariance** of X and Y as
   $$\mathbf{Cov}(X,Y) := \mathbf{E}\left((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))\right) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$$

The same ideas naturally extend, via multiple sums and integrals, to define the expectation of functions of $\mathbb{R}^k$-valued random variables with $k > 2$.

Viewing a deterministic real variable as a random variable

Consider the class of discrete RVs with distributions that place all probability mass on a single real number. This is the probability model for the deterministic real variable, which is often thought of as an unknown constant $\theta \in \mathbb{R}$.

Model 12 (Point Mass(θ)**)** Given a specific point $\theta \in \mathbb{R}$, we say an RV X has point mass at θ or is Point Mass(θ) distributed if the DF is:

$$F(x;\theta) = \begin{cases} 0 & \text{if } x < \theta \\ 1 & \text{if } x \geq \theta \end{cases} \tag{3.47}$$

and the PMF is:

$$f(x;\theta) = \begin{cases} 0 & \text{if } x \neq \theta \\ 1 & \text{if } x = \theta \end{cases} \tag{3.48}$$

Thus, Point Mass(θ) RV X is deterministic in the sense that every realisation of X is exactly equal to $\theta \in \mathbb{R}$. We will see that this distribution plays a central limiting role in asymptotic statistics.

Example 73 (Mean and variance of Point Mass(θ) **RV)** Let $X \sim$ Point Mass(θ). Then:

$$\mathbf{E}(X) = \sum_x x f(x) = \theta \times 1 = \theta\ , \qquad \mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \theta^2 - \theta^2 = 0\ .$$

3.8.2 Properties of expectations

The following results, where a is a constant, may easily be proved using the properties of summations and integrals:

$$\boxed{\mathbf{E}(a) = a}$$

$$\boxed{\mathbf{E}(a\,g(X)) = a\,\mathbf{E}(g(X))}$$

$$\boxed{\mathbf{E}(g(X)+h(X)) = \mathbf{E}(g(X)) + \mathbf{E}(h(X))}$$

Note that here $g(X)$ and $h(X)$ are functions of the random variable X: e.g. $g(X) = X^2$.

Using these results we can obtain the following useful formula for variance:

$$\begin{aligned} V(X) &= E\left((X-\mathbf{E}(X))^2\right) \\ &= E\left(X^2 - 2X\mathbf{E}(X) + (\mathbf{E}(X))^2\right) \\ &= E(X^2) - E\left(2X\mathbf{E}(X)\right) + E\left((\mathbf{E}(X))^2\right) \\ &= E(X^2) - 2\mathbf{E}(X)E\left(X\right) + (\mathbf{E}(X))^2 \\ &= E(X^2) - 2(\mathbf{E}(X))^2 + (\mathbf{E}(X))^2 \\ &= \mathbf{E}(X^2) - (\mathbf{E}(X))^2 \ . \end{aligned}$$

That is,

$$\boxed{V(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2}.$$

The above properties of expectations imply that for constants a and b,

$$\boxed{\mathbf{V}(aX+b) = a^2\mathbf{V}(X)\ .} \tag{3.49}$$

More generally, for random variables $X_1, X_2, \ldots, X_n$ and constants $a_1, a_2, \ldots, a_n$

- $$\mathbf{E}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i \mathbf{E}(X_i)\ . \tag{3.50}$$
- $$\mathbf{V}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \mathbf{V}(X_i), \text{ provided } X_1, X_2, \ldots, X_n \text{ are independent}\ . \tag{3.51}$$
- Let $X_1, X_2, \ldots, X_n$ be independent RVs, then

$$\mathbf{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} \mathbf{E}(X_i), \text{ provided } X_1, X_2, \ldots, X_n \text{ are independent}\ . \tag{3.52}$$

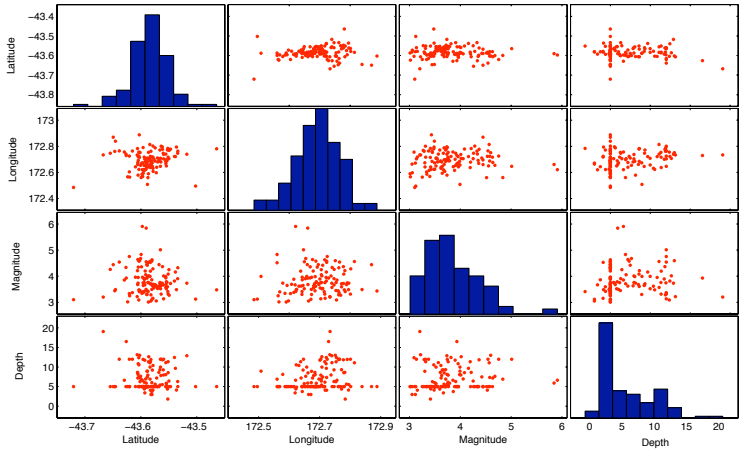

```
>> LatData=EQ(:,2); LonData=EQ(:,3); MagData=EQ(:,12); DepData=EQ(:,13);
>> % finally make a plot matrix of these 124 4-tuples as red points
>> plotmatrix([LatData,LonData,MagData,DepData], 'r.');
```

All of these commands have been put in a script M-file `NZEQChCch20110222.m` and you can simply call it from the command window to automatically load the data and assign it to the variables `EQAll EQ`, `LatData`, `LonData`, `MagData` and `DepData`, instead of retyping each command above every time you need these matrices in MATLAB , as follows:

```
>> NZEQChCch20110222
ans =   145    14
ans =   145    14
ans =   124    14
```

In fact, we will do exactly this to conduct more exploratory data analysis with these earth quake measurements in Labwork 119.

Figure 3.32: Matrix of Scatter Plots of the latitude, longitude, magnitude and depth of the 22-02-2011 earth quakes in Christchurch, New Zealand.



Labwork 119 Try to understand how to manipulate time stamps of events in MATLAB and the Figures being output by following the comments in the script file `NZEQChCch20110222EDA.m`.

```
>> NZEQChCch20110222
ans =   145    14
ans =   145    14
ans =   124    14
ans =   145    14
ans =   145    14
ans =   124    14
ans = 22-Feb-2011 00:00:31
ans = 22-Feb-2011 23:50:01
```

```
CUSP_ID,LAT,LONG,NZMGE,NZMGN,ORI_YEAR,ORI_MONTH,ORI_DAY,ORI_HOUR,ORI_MINUTE,ORI_SECOND,MAG,DEPTH
3481751,-43.55432,172.68898,2484890,5739375,2011,2,22,0,0,31.27814,3.79,5.8559,
3481760,-43.56579,172.70621,2486287,5738106,2011,2,22,0,0,43.70276,3.76,5.4045,
.
.
.
3469114,-43.58007,172.67126,2483470,5736509,2011,2,22,23,28,11.1014,3.117,3,
3469122,-43.55949,172.70396,2486103,5738805,2011,2,22,23,50,1.06171,3.136,12,
```

The thirteen columns correspond to fairly self-descriptive features of each measured earth quake given in the first line or row. They will become clear in the sequel. Note that the comma character (‘,’) separates each unit or measurement or descpiption in any CSV file.

3. The next set of commands show you how to load, manipulate and visually explore this data.

```
%% Load the data from the comma delimited text file 'NZ20110222earthquakes.csv' with
%% the following column IDs
%% CUSP_ID,LAT,LONG,NZMGE,NZMGN,ORI_YEAR,ORI_MONTH,ORI_DAY,ORI_HOUR,ORI_MINUTE,ORI_SECOND,MAG,DEPTH
%% Using MATLAB's dlmread command we can assign the data as a matrix to EQ;
%% note that the option 1,0 to dlmread skips first row of column descriptors
%
% the variable EQall is about to be assigned the data as a matrix
EQall = dlmread('NZ20110222earthquakes.csv', ',' , 1, 0);
size(EQall) % report the dimensions or size of the matrix EQall
ans =
   145    14
```

4. In order to understand the syntax in detail get `help` from MATLAB !

```
>> help dlmread
 DLMREAD Read ASCII delimited file.
 .
 .
 .
```

5. When there are units in the CSV file that can't be converted to floating-point numbers, it is customary to load them as a `NaN` or *Not-a-Number* value in MATLAB . So, let's check if there are any rows with `NaN` values and remove them from our analysis. Note that this is not the only way to deal with missing data! After that let's remove any locations outside Christchurch and its suburbs (we can find the latitude and longitude bounds from online resources easily) and finally view the 4-tuples of (latitude, longitude, magnitude, depth) for each measured earth quake in Christchurch on February 22 of 2011 as a scatter plot shown in Figure 3.32 (the axes labels were subsequently added from clicking `<Edit>` and `<Figure Properties...>` tabs of the output Figure Window).

```
>> EQall(any(isnan(EQall),2),:) = []; %Remove any rows containing NaNs from the matrix EQall
>> % report the size of EQall and see if it is different from before we removed and NaN containing rows
>> size(EQall)
ans =   145    14
>> % remove locations outside Chch and assign it to a new variable called EQ
>> EQ = EQall(-43.75<EQall(:,2) & EQall(:,2)<-43.45 ...
             & 172.45<EQall(:,3) & EQall(:,3)<172.9 & EQall(:,12)>3, :);
>> % now report the size of the earthquakes in Christchurch in variable EQ
>> size(EQ)
ans =   124    14
>> % assign the four variables of interest
```


3.8.3 Expectation of Common Random Variables

Let us compute the mean and variance of our familiar RVs.

Example 74 (Mean and variance of Bernoulli(θ) **RV)** Let $X \sim \text{Bernoulli}(\theta)$. Then,

$$\mathbf{E}(X) = \sum_{x=0}^{1} x f(x) = (0 \times (1-\theta)) + (1 \times \theta) = 0 + \theta = \theta \ ,$$

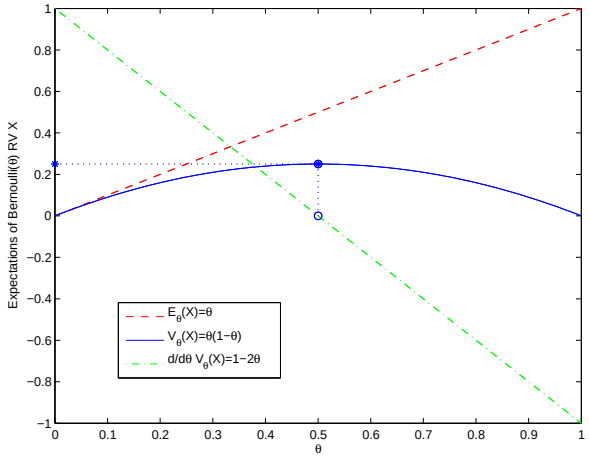
$$\mathbf{E}(X^2) = \sum_{x=0}^{1} x^2 f(x) = (0^2 \times (1-\theta)) + (1^2 \times \theta) = 0 + \theta = \theta \ ,$$

$$\mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \theta - \theta^2 = \theta(1-\theta) \ .$$

Parameter specifically,

$$\mathbf{E}_\theta(X) = \theta \qquad \text{and} \qquad \mathbf{V}_\theta(X) = \theta(1-\theta) \ .$$

Figure 3.16: Mean ($\mathbf{E}_\theta(X)$), variance ($\mathbf{V}_\theta(X)$) and the rate of change of variance ($\frac{d}{d\theta}\mathbf{V}_\theta(X)$) of a Bernoulli($\theta$) RV X as a function of the parameter θ.



Maximum of the variance $\mathbf{V}_\theta(X)$ is found by setting the derivative to zero, solving for θ and showing the second derivative is locally negative, i.e. $\mathbf{V}_\theta(X)$ is concave down:

$$\mathbf{V}'_\theta(X) := \frac{d}{d\theta}\mathbf{V}_\theta(X) = 1 - 2\theta = 0 \iff \theta = \frac{1}{2} \ , \qquad \mathbf{V}''_\theta(X) := \frac{d}{d\theta}\left(\frac{d}{d\theta}\mathbf{V}_\theta(X)\right) = -2 < 0 \ ,$$

$$\max_{\theta \in [0,1]} \mathbf{V}_\theta(X) = \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{4} \ , \text{since } \mathbf{V}_\theta(X) \text{ is maximized at } \theta = \frac{1}{2}$$

The plot depicting these expectations as well as the rate of change of the variance are depicted in Figure 3.16. Note from this Figure that $\mathbf{V}_\theta(X)$ attains its maximum value of $1/4$ at $\theta = 0.5$ where $\frac{d}{d\theta}\mathbf{V}_\theta(X) = 0$. Furthermore, we know that we don't have a minimum at $\theta = 0.5$ since the second derivative $\mathbf{V}''_\theta(X) = -2$ is negative for any $\theta \in [0,1]$. This confirms that $\mathbf{V}_\theta(X)$ is concave down and therefore we have a maximum of $\mathbf{V}_\theta(X)$ at $\theta = 0.5$. We will revisit this example when we employ a numerical approach called Newton-Raphson method to solve for the maximum of a differentiable function by setting its derivative equal to zero.

Example 75 (Mean and variance of Uniform$(0,1)$ **RV)** Let $X \sim$ Uniform$(0,1)$. Then,

$$\mathbf{E}(X) = \int_{x=0}^{1} x f(x)\,dx = \int_{x=0}^{1} x\ 1\,dx = \frac{1}{2}\left(x^2\right]_{x=0}^{x=1} = \frac{1}{2}(1-0) = \frac{1}{2}\ ,$$

$$\mathbf{E}(X^2) = \int_{x=0}^{1} x^2 f(x)\,dx = \int_{x=0}^{1} x^2\ 1\,dx = \frac{1}{3}\left(x^3\right]_{x=0}^{x=1} = \frac{1}{3}(1-0) = \frac{1}{3}\ ,$$

$$\mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}\ .$$

Exercise 3.29 (Mean and variance of Uniform(θ_1,θ_2) **RV)** Let $X \sim$ Uniform(θ_1,θ_2) of Model 9. Derive expressions for $\mathbf{E}(X)$ and $\mathbf{V}(X)$ in terms of the parameters θ_1 and θ_2. Make sure that when $\theta_2 = 1$ and $\theta_1 = 0$ you recover the expectation and variance of the Uniform$(0,1)$ RV in Example 75.

Example 76 (Expected Exponential of the Uniform$(0,1)$ **RV)** Let $X \sim$ Uniform$(0,1)$ and $Y = r(X) = e^X$. Compute $\mathbf{E}(Y)$.

We can simply apply the definition of $\mathbf{E}(r(X))$, since $Y = r(X)$, is just a function of X, as follows:

$$\mathbf{E}(Y) = \int_0^1 e^x f(x)dx = \int_0^1 e^x 1\ dx = e - 1\ .$$

Example 77 (Mean and variance of Exponential(λ)**)** Show that the mean of an Exponential(λ) RV X is:

$$\mathbf{E}_\lambda(X) = \int_0^\infty x f(x;\lambda)\,dx = \int_0^\infty x\lambda e^{-\lambda x}\,dx = \frac{1}{\lambda}\ ,$$

and the variance is:

$$\mathbf{V}_\lambda(X) = \left(\frac{1}{\lambda}\right)^2\ .$$

Example 78 (Mean and variance of Geometric(θ) **RV)** Let $X \sim$ Geometric(θ) RV. Then,

$$\mathbf{E}(X) = \sum_{x=0}^{\infty} x\theta(1-\theta)^x = \theta\sum_{x=0}^{\infty} x(1-\theta)^x$$

In order to simplify the RHS above, let us employ differentiation with respect to θ:

$$\frac{-1}{\theta^2} = \frac{d}{d\theta}\left(\frac{1}{\theta}\right) = \frac{d}{d\theta}\sum_{x=0}^{\infty}(1-\theta)^x = \sum_{x=0}^{\infty} -x(1-\theta)^{x-1}$$

Multiplying the LHS and RHS above by $-(1-\theta)$ and substituting in $\mathbf{E}(X) = \theta\sum_{x=0}^{\infty} x(1-\theta)^x$, we get a much simpler expression for $\mathbf{E}(X)$:

$$\frac{1-\theta}{\theta^2} = \sum_{x=0}^{\infty} x(1-\theta)^x \implies \mathbf{E}(X) = \theta\left(\frac{1-\theta}{\theta^2}\right) = \frac{1-\theta}{\theta}\ .$$

Labwork 117 Let us make matrix plots from a uniformly generated sequence of 100 points in 5D unit cube $[0,1]^5$ as shown in Figure 3.31.

```
>> rand('twister',5489);
>> % generate a sequence of 1000 points uniformly distributed in 5D unit cube [0,1]X[0,1]X[0,1]X[0,1]X[0,1]
>> x=rand(1000,5);
>> x(1:6,:) % first six points in our 5D unit cube, i.e., the first six rows of x
ans =
    0.8147    0.6312    0.7449    0.3796    0.4271
    0.9058    0.3551    0.8923    0.3191    0.9554
    0.1270    0.9970    0.2426    0.9861    0.7242
    0.9134    0.2242    0.1296    0.7182    0.5809
    0.6324    0.6525    0.2251    0.4132    0.5403
    0.0975    0.6050    0.3500    0.0986    0.7054
>> plotmatrix(x(1:5,:),'r*') % make a plot matrix
>> plotmatrix(x) % make a plot matrix of all 1000 points
```

3.14.6 Loading and Exploring Real-world Data

All of the data we have played with so far were computer-generated. It is time to get our hands dirty with real-world data. The first step is to obtain the data. Often, publicly-funded institutions allow the public to access their databases. Such data can be fetched from appropriate URLs in one of the two following ways:

Method A: Manually download by filling the appropriate fields in an online request form.

Method B: Automagically download directly from your MATLAB session.

Then we want to inspect it for inconsistencies, missing values and replace them with `NaN` values in MATLAB that stand for not-any-number. Finally, we can visually explore, transform and interact with the data to discover interesting patterns that are hidden in the data. This process is called *exploratory data analysis* and is the foundational first step towards subsequent computational statistical experiments [*John W. Tukey, Exploratory Data Analysis, Addison-Wesely, New York, 1977*].

3.14.7 Geological Data

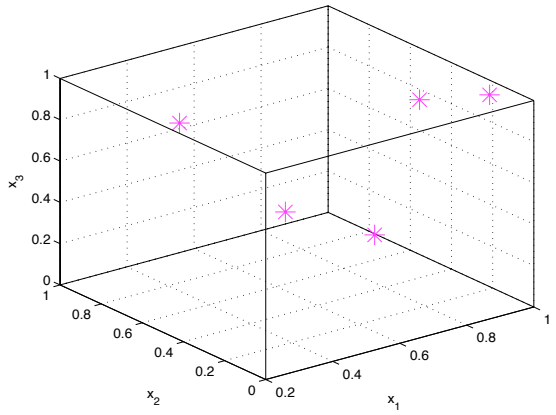
Let us focus on the data of earth quakes that heavily damaged Christchurch on February 22 2011. This data can be fetched from the URL `http://magma.geonet.org.nz/resources/quakesearch/` by Method A and loaded into MATLAB for exploratory data analysis as done in Labwork 118.

Labwork 118 Let us go through the process one step at a time using Method A.

1. Download the data as a CSV or *comma separated variable* file in plain ASCII text (this has been done for this data already for you and saved as `NZ20110222earthquakes.csv` in the `CSEMatlabScripts` directory).

2. Open the file in a simple text editor such as `Note Pad` in Windows or one of the following editors in OS X, Unix, Solaris, Linux/GNU variants such as Ubuntu, SUSE, etc: `vi`, `vim`, `emacs`, `geany`, etc. The first three and last two lines of this file look as follows:

Figure 3.30: 3D Scatter Plot

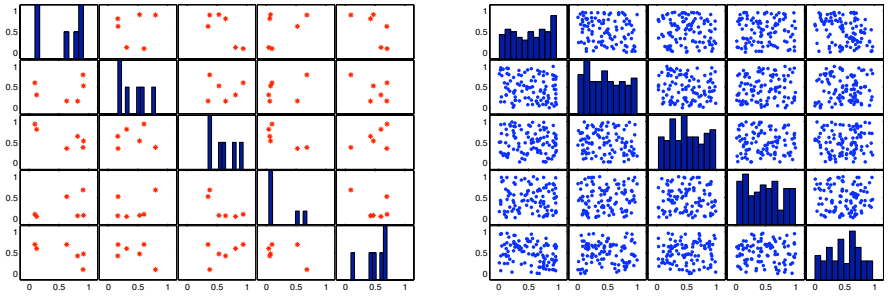


There are several other techniques for visualising trivariate data, including, iso-surface plots, moving surface or heat plots, and you will encounter some of them in the future.

3.14.5 Multivariate Data

For high-dimensional data in d-dimensional space $\mathbb{R}^d$ with $d \geq 3$ you have to look at several lower dimensional projections of the data. We can simultaneously look at 2D scatter plots for every pair of co-ordinates $\{(i,j) \in \{1,2,\ldots,d\}^2 : i \neq j\}$ and at histograms for every co-ordinate $i \in \{1,2,\ldots,d\}$ of the n data points in $\mathbb{R}^d$. Such a set of low-dimensional projections can be conveniently represented in a $d \times d$ matrix of plots called a **matrix plot**.

Figure 3.31: Plot Matrix of uniformly generated data in $[0,1]^5$



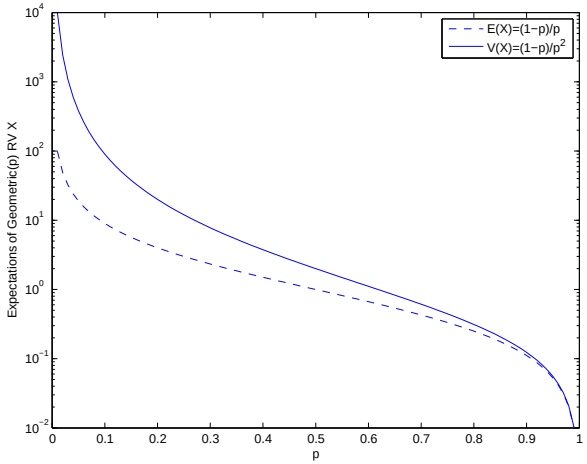
(a) First six samples

(b) All thousand samples

Similarly, it can be shown that

$$\mathbf{V}(X) = \frac{1-\theta}{\theta^2} \ .$$

Figure 3.17: Mean and variance of a Geometric(θ) RV X as a function of the parameter θ.



Example 79 (Mean and variance of Binomial(n,θ) **RV)** Let $X \sim \text{Binomial}(n,\theta)$. Based on the definition of expectation:

$$\mathbf{E}(X) = \int x \, dF(x;n,\theta) = \sum_x x f(x;n,\theta) = \sum_{x=0}^{n} x \binom{n}{x} \theta^x (1-\theta)^{n-x} \ .$$

However, this is a nontrivial sum to evaluate. Instead, we may use (3.50) and (3.51) by noting that $X = \sum_{i=1}^n X_i$, where the $\{X_1, X_2, \ldots, X_n\} \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta)$, $\mathbf{E}(X_i) = \theta$ and $\mathbf{V}(X_i) = \theta(1-\theta)$:

$$\mathbf{E}(X) = \mathbf{E}(X_1 + X_2, \cdots, X_n) = \mathbf{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbf{E}(X_i) = n\theta \ ,$$

$$\mathbf{V}(X) = \mathbf{V}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbf{V}(X_i) = \sum_{i=1}^n \theta(1-\theta) = n\theta(1-\theta) \ .$$

Example 80 (Mean and variance of Poisson(λ) **RV)** Let $X \sim \text{Poisson}(\lambda)$. Then:

$$\mathbf{E}(X) = \sum_{x=0}^{\infty} x f(x;\lambda) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x-1=0}^{\infty} \frac{\lambda \lambda^{x-1}}{(x-1)!} = e^{-\lambda}\lambda e^{\lambda} = \lambda \ .$$

Similarly,

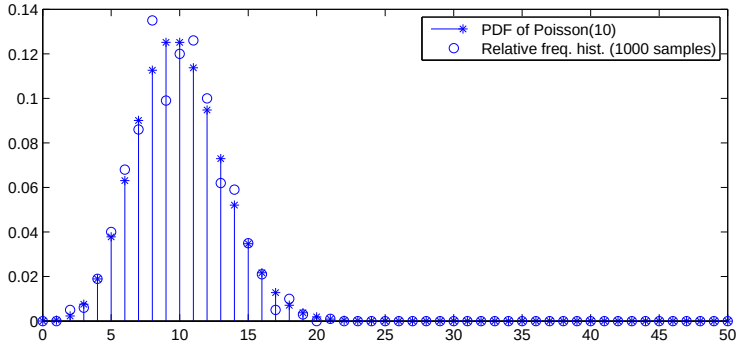
$$\mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \lambda + \lambda^2 - \lambda^2 = \lambda \ .$$

since

$$\begin{aligned}\mathbf{E}(X^2) &= \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{x\,\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda}\left(1 + \frac{2\lambda}{1} + \frac{3\lambda^2}{2!} + \frac{4\lambda^3}{3!} + ...\right) \\ &= \lambda e^{-\lambda}\left(\left(1 + \frac{\lambda}{1} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + ...\right) + \left[\frac{\lambda}{1} + \frac{2\lambda^2}{2!} + \frac{3\lambda^3}{3!} + ...\right]\right) \\ &= \lambda e^{-\lambda}\left(\left(e^{\lambda}\right) + \lambda\left(1 + \frac{2\lambda}{2!} + \frac{3\lambda^2}{3!} + ...\right)\right) = \lambda e^{-\lambda}\left(e^{\lambda} + \lambda\left(1 + \lambda + \frac{\lambda^2}{2!} + ...\right)\right) \\ &= \lambda e^{-\lambda}\left(e^{\lambda} + \lambda\left(e^{\lambda}\right)\right) = \lambda e^{-\lambda}\left(e^{\lambda} + \lambda e^{\lambda}\right) = \lambda(1+\lambda) = \lambda + \lambda^2\end{aligned}$$

Note that Poisson(λ) distribution is one whose mean and variance are the same, namely λ.

Figure 3.18: PDF of $X \sim$ Poisson($\lambda = 10$) and the relative frequency histogram based on 1000 samples from X according to Simulation 149.



The Poisson(λ) RV X is also related to the IID Exponential(λ) RV $Y_1, Y_2, \ldots$: X is the number of occurrences, per unit time, of an instantaneous event whose inter-occurrence time is the IID Exponential(λ) RV. For example, the number of buses arriving at our bus-stop in the next minute, with exponentially distributed inter-arrival times, has a Poisson distribution.

Example 81 (Mean and variance of Normal(μ, σ^2) **RV)** The location-scale family of RVs is indeed parameterised by its mean and variance, i.e., if $X \sim$ Normal(μ, σ^2) where $X = g(Z) = \sigma Z + \mu$ and $Z \sim$ Normal(0, 1) then $\mathbf{E}(X) = \mu$ and $\mathbf{V}(X) = \sigma^2$ follows directly from the properties of Expectations, provided $\mathbf{E}(Z) = 0$ and $\mathbf{V}(Z) = \mathbf{E}(Z^2) - (\mathbf{E}(Z))^2 = \mathbf{E}(Z^2) = 1$.

The mean of a Normal(0, 1) RV Z is:

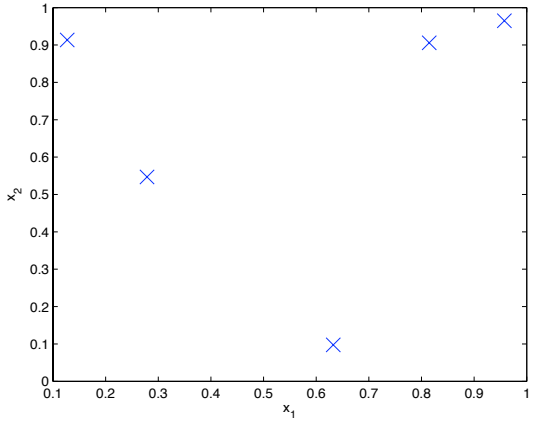
$$\mathbf{E}(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \exp\left(-\frac{1}{2}z^2\right) dz = \frac{1}{\sqrt{2\pi}} \left(-\exp\left(-\frac{1}{2}z^2\right)\right]_{-\infty}^{\infty} = 0\ ,$$

and the variance is:

$$\mathbf{V}(Z) = \mathbf{E}(Z^2) - (\mathbf{E}(Z))^2 = \mathbf{E}(Z^2) - 0 = \mathbf{E}(Z^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz.$$


```
    0.9058    0.9134    0.0975    0.5469    0.9649
>> plot(x(1,:),x(2,:),'x') % a 2D scatter plot with marker cross or 'x'
>> plot(x(1,:),x(2,:),'x', 'MarkerSize',15) % a 2D scatter plot with marker cross or 'x' and larger Marker size
>> xlabel('x_1'); ylabel('x_2'); % label the axes
```

Figure 3.29: 2D Scatter Plot



There are several other techniques for visualising bivariate data, including, 2D histograms, surface plots, heat plots, and we will encounter some of them in the sequel.

3.14.4 Trivariate Data

Trivariate data is more difficult to visualise on paper but playing around with the rotate 3D feature in MATLAB 's Figure window can help bring a lot more perspective.

Labwork 116 (Visualising trivariate data) We can make **3D scatter plots** as shown in Figure 3.30 as follows:

```
>> rand('twister',5489);
>> x=rand(3,5)% create a sequence of 5 ordered triples uniformly from unit cube [0,1]X[0,1]X[0,1]
x =
    0.8147    0.9134    0.2785    0.9649    0.9572
    0.9058    0.6324    0.5469    0.1576    0.4854
    0.1270    0.0975    0.9575    0.9706    0.8003
>> plot3(x(1,:),x(2,:),x(3,:),'x') % a simple 3D scatter plot with marker 'x'
>>% a more interesting one with options that control marker type, line-style,
>>% colour in [Red Green Blue] values and marker size - read help plot3 for more options
>> plot3(x(1,:),x(2,:),x(3,:),'Marker','*','LineStyle','none','Color',[1 0 1],'MarkerSize',15)
>> plot3(x(1,:),x(2,:),x(3,:),'m*','MarkerSize',15) % makes same  figure as before but shorter to write
>> box on % turn on the box and see the effect on the Figure
>> grid on % turn on the grid and see the effect on the Figure
>> xlabel('x_1'); ylabel('x_2'); zlabel('x_3'); % assign labels to x,y and z axes
```

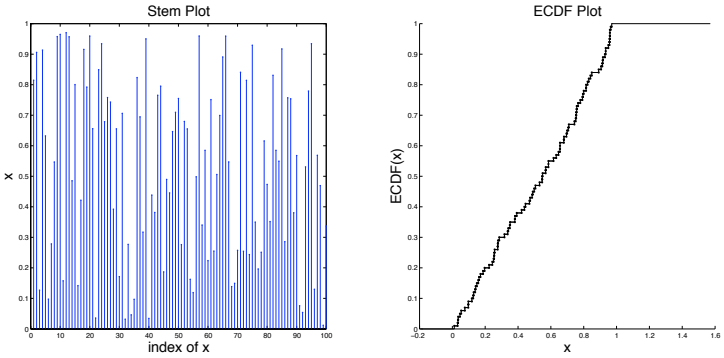
Repeat the visualisation below with a larger array, say `x=rand(3,1000)`, and use the rotate 3D feature in the Figure window to visually explore the samples in the unit cube. Do they seem to be uniformly distributed inside the unit cube?

Try making a density histogram with 1000 samples from `rand` with 15 bins. You can specify the number of bins by adding an extra argument to `hist`, for e.g. `[Fs, Cs] = hist(x,15)` will produce 15 bins of equal width over the data range $\mathcal{R}(x)$.

Labwork 114 (Stem plots and ECDF plots for univariate data) We can also visualise the 100 data points in the array `x` using stem plot and ECDF plot as shown in Figure 3.28 as follows:

```
>> rand('twister',5489);
>> x=rand(1,100); % produce 100 samples with rand
>> stem(x,'.') % make a stem plot of the 100 data points in x (the option '.' gives solid circles for x)
>>% ECDF (type help ECDF) plot is extended to left and right by .2 and .6, respectively
>>% (second parameter 6 makes the dots in the plot smaller).
>> ECDF(x,6,.2,.6);
```

Figure 3.28: Frequency, Relative Frequency and Density Histograms



We can also visually summarise univariate data using the **box plot** or **box-whisker plot** available in the Stats Toolbox of MATLAB. These family of plots display a set of sample quantiles, typically they are include, the median, the first and third quartiles and the minimum and maximum values of our data array x.

3.14.3 Bivariate Data

By bivariate data array x we mean a $2 \times n$ matrix of real numbers or equivalently n ordered pairs of points $(x_{1,i}, x_{2,i})$ as $i = 1, 2, \ldots, n$. The most elementary visualisation of these n ordered pairs is in orthogonal Cartesian co-ordinates. Such plots are termed 2D **scatter plots** in statistics.

Labwork 115 (Visualising bivariate data) Let us generate a 2×5 array representing samples of 5 ordered pairs sampled uniformly at random over the unit square $[0, 1] \times [0, 1]$. We can make 2D scatter plot as shown in Figure 3.29 as follows:

```
>> rand('twister',5489);
>> x=rand(2,5)% create a sequence of 5 ordered pairs uniformly from unit square [0,1]X[0,1]
x =
    0.8147    0.1270    0.6324    0.2785    0.9575
```


Using integration by parts with $u = z, dv = ze^{-z^2/2} \implies du = 1, v = -e^{-z^2/2}$, $\int udv = uv - \int vdu$

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \left(-ze^{-z^2/2} \right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz = 0 + 1 = 1$$

The first term after the first equality above equals 0 because the exponential goes to 0 much faster than z grows to $\pm\infty$. The second term equals 1 because it is exactly the total probability integral of the PDF of the Normal$(0, 1)$ RV.

Next, let us become familiar with an RV for which the expectation does not exist.

Model 13 (Cauchy**)** The density of the Cauchy RV Y is:

$$f(y) = \frac{1}{\pi(1 + y^2)}, \qquad -\infty < y < \infty \ , \tag{3.53}$$

and its DF is:

$$F(y) = \frac{1}{\pi} \tan^{-1}(y) + \frac{1}{2} \ . \tag{3.54}$$

Randomly spinning a LASER emitting improvisation of "Darth Maul's double edged lightsaber" that is centered at $(1, 0)$ in the plane $\mathbb{R}^2$ and recording its intersection with the y-axis, in terms of the y coordinates of the point $(0, y)$, gives rise to the *Standard Cauchy* RV.

The Cauchy RV Y can be derived from a RV $X \sim \text{Uniform}(-\pi/2, \pi/2)$ by the simple transformation $Y = \tan(X)$ for the above construction. Since $\tan(x)$ is one-to-one and monotone on the range of X given by $(-\pi/2, \pi/2)$, we can use the change of variable formula in Equation 3.38 to obtain the PDF $f_Y(y)$ from the PDF $f_X(x) = \frac{1}{\pi}\mathbb{1}_{(-\pi/2,\pi/2)}(x)$ as follows:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| = f_X(\tan^{-1}(y)) \left| \frac{d}{dy} \tan^{-1}(y) \right| = \frac{1}{\pi} \left| \frac{1}{1 + y^2} \right|$$

Note that the construction is valid even if we sample X uniformly from $(0, \pi)$ and take its $\tan(X)$.

Example 82 (Mean of Cauchy **RV)** The expectation of the Cauchy RV X, obtained via integration by parts (set $u = x$ and $v = \tan^{-1}(x)$) does not exist , since:

$$\int |x| \ dF(x) = \frac{2}{\pi} \int_0^{\infty} \frac{x}{1 + x^2} \, dx = \left(x \tan^{-1}(x) \right]_0^{\infty} - \int_0^{\infty} \tan^{-1}(x) \, dx = \infty \ . \tag{3.55}$$

Note that we consider symmetry of integral about the origin and take twice the integral over $(0, \infty)$ above. Variance and higher moments cannot be defined when the expectation itself is undefined.

Next let us consider a natural generalization of the Bernoulli(θ) RV with more than two outcomes but in the set $\{1, 2, \ldots, k\}$.

Model 14 (de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$**)** Given a specific point $(\theta_1, \theta_2, \ldots, \theta_k)$ in the unit $k-1$-Simplex:

$$\triangle^{k-1} := \{ (\theta_1, \theta_2, \ldots, \theta_k) : \theta_1 \geq 0, \theta_2 \geq 0, \ldots, \theta_k \geq 0, \sum_{i=1}^{k} \theta_i = 1 \} \ ,$$

we say that an RV X is de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ distributed if its PMF is:

$$f(x; \theta_1, \theta_2, \ldots, \theta_k) = \begin{cases} 0 & \text{if } x \notin [k] := \{1, 2, \ldots, k\}, \\ \theta_x & \text{if } x \in [k]. \end{cases}$$

The DF for de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ RV X is:

$$F(x;\theta_1,\theta_2,\ldots,\theta_k) = \begin{cases} 0 & \text{if } -\infty < x < 1 \\ \theta_1 & \text{if } 1 \le x < 2 \\ \theta_1+\theta_2 & \text{if } 2 \le x < 3 \\ \vdots & \\ \theta_1+\theta_2+\cdots+\theta_{k-1} & \text{if } k-1 \le x < k \\ \theta_1+\theta_2+\cdots+\theta_{k-1}+\theta_k = 1 & \text{if } k \le x < \infty \end{cases} \tag{3.56}$$

The de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ RV can be thought of as a probability model for "the outcome of rolling a polygonal cylindrical die with k rectangular faces that are marked with $1, 2, \ldots, k$". The parameters $\theta_1, \theta_2, \ldots, \theta_k$ specify how the die is loaded and may be idealised as specifying the cylinder's centre of mass with respect to the respective faces. Thus, when $\theta_1 = \theta_2 = \cdots = \theta_k = 1/k$, we have a probability model for the outcomes of a fair die.

Mean and variance of de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ **RV:** The not too useful expressions for the first two moments of $X \sim$ de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ are,

$$\mathbf{E}(X) = \sum_{x=1}^{k} x\theta(x) = \theta_1 + 2\theta_2 + \cdots + k\theta_k \text{ , and}$$

$$\mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \left(\theta_1 + 2^2\theta_2 + \cdots + k^2\theta_k\right) - (\theta_1 + 2\theta_2 + \cdots + k\theta_k)^2 \text{ .}$$

However, if $X \sim$ de Moivre$(1/k, 1/k, \ldots, 1/k)$, then the mean and variance for the fair k-faced die based on Faulhaber's formula for $\sum_{i=1}^{k} i^m$, with $m \in \{1, 2\}$, are,

$$\mathbf{E}(X) = \frac{1}{k}\left(1 + 2 + \cdots + k\right) = \frac{1}{k}\frac{k(k+1)}{2} = \frac{k+1}{2} \text{ ,}$$

$$\mathbf{E}(X^2) = \frac{1}{k}\left(1^2 + 2^2 + \cdots + k^2\right) = \frac{1}{k}\frac{k(k+1)(2k+1)}{6} = \frac{2k^2+3k+1}{6} \text{ ,}$$

$$\mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \frac{2k^2+3k+1}{6} - \left(\frac{k+1}{2}\right)^2 = \frac{2k^2+3k+1}{6} - \left(\frac{k^2+2k+1}{4}\right)$$
$$= \frac{8k^2+12k+4-6k^2-12k-6}{24} = \frac{2k^2-2}{24} = \frac{k^2-1}{12} \text{ .}$$

3.9 Exercises in Expectations of Random Variables

Ex. 3.30 — Let X be the number of air conditioners a store sells each day, and assume that X has probability mass function $f(10) = 0.1$, $f(11) = 0.3$, $f(12) = 0.4$, $f(13) = 0.2$.

1. Find the expected number of conditioners that the store sells each day.
2. If the profit per conditioner is \$55, what is the expected daily profit?

Ex. 3.31 — A small petrol station is supplied with fuel every Saturday afternoon. Assume that its volume of sales X, in ten thousands of litres, has density

$$f(x) = \begin{cases} 6x(1-x) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \text{ .}$$

Determine the mean and variance of X.

For a given partition of the data range $\mathcal{R}(x)$ or some superset of $\mathcal{R}(x)$, three types of histograms are possible: frequency histogram, relative frequency histogram and density histogram. Typically, the partition b is assumed to be composed of m overlapping intervals of the same width $w = \bar{b}_i - \underline{b}_i$ for all $i = 1, 2, \ldots, m$. Thus, a histogram can be obtained by a set of bins along with their corresponding **heights**:

$$h = (h_1, h_2, \ldots, h_m) \text{ , where } h_k := g(\#\{x_i : x_i \in b_k\})$$

Thus, h_k, the height of the k-th bin, is some function g of the number of data points that fall in the bin b_k Formally, a histogram is a sequence of ordered pairs:

$$((b_1, h_1), (b_2, h_2), \ldots, (b_m, h_m)) \text{ .}$$

Given a partition b, a **frequency histogram** is the histogram:

$$((b_1, h_1), (b_2, h_2), \ldots, (b_m, h_m)) \text{ , where } h_k := \#\{x_i : x_i \in b_k\} \text{ ,}$$

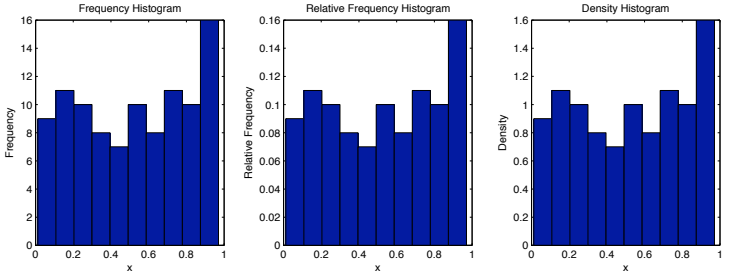
a **relative frequency histogram** is the histogram:

$$((b_1, h_1), (b_2, h_2), \ldots, (b_m, h_m)) \text{ , where } h_k := n^{-1}\#\{x_i : x_i \in b_k\} \text{ ,}$$

and a **density histogram** is the histogram:

$$((b_1, h_1), (b_2, h_2), \ldots, (b_m, h_m)) \text{ , where } h_k := (w_k n)^{-1}\#\{x_i : x_i \in b_k\} \text{ , } w_k := \bar{b}_k - \underline{b}_k \text{ .}$$

Figure 3.27: Frequency, Relative Frequency and Density Histograms



Labwork 113 (Histograms with specified number of bins for univariate data) Let us use samples from the `rand('twister',5489)` as our data set x and plot various histograms. Let us use `hist` function (read `help hist`) to make a default histogram with ten bins. Then we can make three types of histogarms as shown in Figure 3.27 as follows:

```
>> rand('twister',5489);
>> x=rand(1,100); % generate 100 PRNs
>> hist(x) % see what default hist does in Figure Window
>> % Now let us look deeper into the last hist call
>> [Fs, Cs] = hist(x) % Cs is the bin centers and Fs is the frequencies of data set x
Fs =
     9    11    10     8     7    10     8    11    10    16
Cs =
    0.0598    0.1557    0.2516    0.3474    0.4433    0.5392    0.6351    0.7309    0.8268    0.9227
>> % produce a histogram plot the last argument 1 is the width value for immediately adjacent bars -- help bar
>> bar(Cs,Fs,1) % create a frequency histogram
>> bar(Cs,Fs/100,1) % create a relative frequency histogram
>> bar(Cs,Fs/(0.1*100),1) % create a density histogram (area of bars sum to 1)
>> sum(Fs/(0.1*100) .* ones(1,10)*0.1) % checking if area does sum to 1
>> ans = 1
```

Algorithm 1 q^{th} Sample Quantile of Order Statistics

input:

1. q in the q^{th} sample quantile, i.e. the argument q of $\widehat{F}_n^{[-1]}(q)$,
2. order statistic $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$, i.e. the sorted $(x_1, x_2, \ldots, x_n)$, where $n > 0$.

output: $\widehat{F}_n^{[-1]}(q)$, the q^{th} sample quantile
$i \leftarrow \lfloor (n-1)q \rfloor$
$\delta \leftarrow (n-1)q - i$
if $i = n-1$ **then**
  $\widehat{F}_n^{[-1]}(q) \leftarrow x_{(i+1)}$
else
  $\widehat{F}_n^{[-1]}(q) \leftarrow (1-\delta)x_{(i+1)} + \delta x_{(i+2)}$
end if
return: $\widehat{F}_n^{[-1]}(q)$

```
>> disp(SortedXsFromUni01Twstr101)
    0.0285    0.1715    0.5164    0.5707    0.6853
>> rand('twister',420);
>> SortedXs=sort(rand(1,7));
>> disp(SortedXs)
    0.1089    0.2670    0.3156    0.3525    0.4530    0.6297    0.8682
>> for q=[0, 0.25, 0.5, 0.75, 1.0]
       disp([q, qthSampleQuantile(q,SortedXsFromUni01Twstr101) ...
                qthSampleQuantile(q,SortedXs)])
   end
         0    0.0285    0.1089
    0.2500    0.1715    0.2913
    0.5000    0.5164    0.3525
    0.7500    0.5707    0.5414
    1.0000    0.6853    0.8682
```

3.14.2 Univariate Data

A **histogram** is a graphical representation of the frequency with which elements of a data array:

$$x = (x_1, x_2, \ldots, x_n) \ ,$$

of real numbers fall within each of the m intervals or **bins** of some **interval partition**:

$$b := (b_1, b_2, \ldots, b_m) := ([\underline{b}_1, \overline{b}_1], [\underline{b}_2, \overline{b}_2], \ldots, [\underline{b}_m, \overline{b}_m])$$

of the **data range** of x given by the closed interval:

$$\mathcal{R}(x) := [\min\{x_1, x_2, \ldots, x_n\}, \max\{x_1, x_2, \ldots, x_n\}] \ .$$

Elements of this partition b are called bins, their mid-points are called **bin centres**:

$$c := (c_1, c_2, \ldots, c_m) := ((\underline{b}_1 + \overline{b}_1)/2, (\underline{b}_2 + \overline{b}_2)/2, \ldots, (\underline{b}_m + \overline{b}_m)/2)$$

and their overlapping boundaries, i.e. $\overline{b}_i = \underline{b}_{i+1}$ for $1 \leq i < m$, are called **bin edges**:

$$d := (d_1, d_2, \ldots, d_{m+1}) := (\underline{b}_1, \underline{b}_2, \ldots, \underline{b}_{m-1}, \underline{b}_m, \overline{b}_m) \ .$$

Ex. 3.32 — Starting from the definition of the variance of a random variable (Definition 30) show that

$$\mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 \ .$$

Ex. 3.33 — Show that $V(aX + b) = a^2 V(X)$ for constants a and b and a random variable X.

Ex. 3.34 — **Let X be a discrete random variable with PMF given by

$$f(x) = \begin{cases} \frac{x}{10} & \text{if } x \in \{1,2,3,4\}, \\ 0 & \text{otherwise.} \end{cases}$$

(a)Find:
  (i)$\mathbf{P}(X = 0)$
  (ii)$\mathbf{P}(2.5 < X < 5)$
  (iii)$\mathbf{E}(X)$
  (iv)$\mathbf{V}(X)$
(b)Write down the DF (or CDF) of X.
(c)Plot the PMF and CDF of X.

Ex. 3.35 — Find the mean and the variance of the following random variables.

1. X a discrete uniform random variable on $\{1,2,3,4,5,6\}$, i.e., *the number a fair die turns up.*
2. X is a Uniform$(0,8)$ random variable, i.e., *a continuous uniform random variable from the interval* $[0,8]$.
3. X has a density function

$$f(x) = \begin{cases} 2e^{-2x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \, . \end{cases}$$

3.10 Multivariate Random Variables

Often, in experiments we are measuring two or more aspects simultaneously. For example, we may be measuring the diameters and lengths of cylindrical shafts manufactured in a plant or heights, weights and blood-sugar levels of individuals in a clinical trial. Thus, the underlying outcome $\omega \in \Omega$ needs to be mapped to measurements as realizations of random vectors in the real plane $\mathbb{R}^2 = (-\infty, \infty) \times (-\infty, \infty)$ or the real space $\mathbb{R}^3 = (-\infty, \infty) \times (-\infty, \infty) \times (-\infty, \infty)$:

$$\omega \mapsto (X(\omega), Y(\omega)) : \Omega \to \mathbb{R}^2 \qquad \omega \mapsto (X(\omega), Y(\omega), Z(\omega)) : \Omega \to \mathbb{R}^3$$

More generally, we may be interested in heights, weights, blood-sugar levels, family medical history, known allergies, etc. of individuals in the clinical trial and thus need to make m measurements of the outcome in $\mathbb{R}^m$ using a "measurable mapping" from $\Omega \to \mathbb{R}^m$. To deal with such multivariate measurements we need the notion of **random vectors** (R$\vec{\text{V}}$s), i.e. ordered pairs of random variables (X, Y), ordered triples of random variables (X, Y, Z), or more generally ordered m-tuples of random variables $(X_1, X_2, \ldots, X_m)$.

3.10.1 $\mathbb{R}^2$-valued Random Variables

We first focus on understanding (X, Y), a bivariate R$\vec{\text{V}}$ or $\mathbb{R}^2$-valued RV that is obtained from a pair of discrete or continuous RVs. We then generalize to $\mathbb{R}^m$-valued RVs with $m > 2$ in the next section.

Definition 34 (JDF) The **joint distribution function (JDF)** or **joint cumulative distribution function (JCDF)**, $F_{X,Y}(x, y) : \mathbb{R}^2 \to [0, 1]$, of the bivariate random vector (X, Y) is

$$\begin{aligned} F_{X,Y}(x, y) &= \mathbf{P}(X \leq x \cap Y \leq y) = \mathbf{P}(X \leq x, Y \leq y) \\ &= P(\{\omega : X(\omega) \leq x, Y(\omega) \leq y\}), \text{ for any } (x, y) \in \mathbb{R}^2 , \end{aligned} \tag{3.57}$$

where the right-hand side represents the probability that the random vector (X, Y) takes on a value in $\{(x', y') : x' \leq x, y' \leq y\}$, the set of points in the plane that are south-west of the point (x, y).

The JDF $F_{X,Y}(x, y) : \mathbb{R}^2 \to \mathbb{R}$ satisfies the following conditions to remain a probability:

1. $0 \leq F_{X,Y}(x, y) \leq 1$
2. $F_{X,Y}(x, y)$ is an non-decreasing function of both x and y
3. $F_{X,Y}(x, y) \to 1$ as $x \to \infty$ and $y \to \infty$
4. $F_{X,Y}(x, y) \to 0$ as $x \to -\infty$ and $y \to -\infty$

Definition 35 (JPMF) If (X, Y) is a **discrete random vector** that takes values in a discrete support set $\mathcal{S}_{X,Y} = \{(x_i, y_j) : i = 1, 2, \ldots, j = 1, 2, \ldots\} \subset \mathbb{R}^2$ with probabilities $p_{i,j} = \mathbf{P}(X = x_i, Y = y_j) > 0$, then its **joint probability mass function** (or JPMF) is:

$$f_{X,Y}(x_i, y_j) = \mathbf{P}(X = x_i, Y = y_j) = \begin{cases} p_{i,j} & \text{if } (x_i, y_j) \in \mathcal{S}_{X,Y} \\ 0 & \text{otherwise} \end{cases} . \tag{3.58}$$

Since $\mathbf{P}(\Omega) = 1$, $\sum_{(x_i,y_j)\in\mathcal{S}_{X,Y}} f_{X,Y}(x_i, y_j) = 1$.

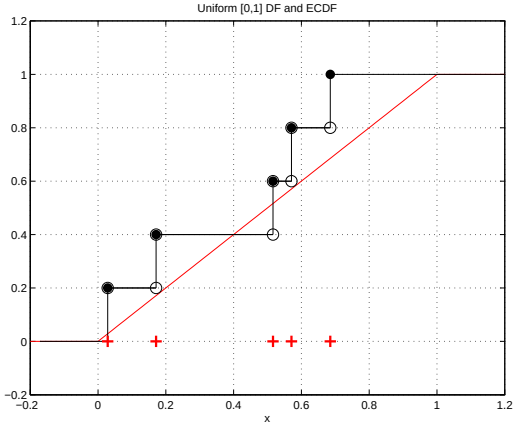
From JPMF $f_{X,Y}$ we can get the values of the JDF $F_{X,Y}(x, y)$ and the probability of any event B by simply taking sums,

$$\boxed{F_{X,Y}(x, y) = \sum_{x_i \leq x, y_j \leq y} f_{X,Y}(x_i, y_j)} , \qquad \boxed{\mathbf{P}(B) = \sum_{(x_i,y_j)\in B\cap\mathcal{S}_{X,Y}} f_{X,Y}(x_i, y_j)} , \tag{3.59}$$

Example 83 Let (X, Y) be a discrete bivariate R$\vec{\text{V}}$with the following joint probability mass function (JPMF):

$$f_{X,Y}(x, y) := P(X = x, Y = y) = \begin{cases} 0.1 & \text{if } (x, y) = (0, 0) \\ 0.3 & \text{if } (x, y) = (0, 1) \\ 0.2 & \text{if } (x, y) = (1, 0) \\ 0.4 & \text{if } (x, y) = (1, 1) \\ 0.0 & \text{otherwise.} \end{cases}$$

Figure 3.26: Plot of the DF of Uniform$(0, 1)$, five IID samples from it, and the ECDF $\widehat{F}_5$ for these five data points $x = (x_1, x_2, x_3, x_4, x_5) = (0.5164, 0.5707, 0.0285, 0.1715, 0.6853)$ that jumps by $1/5 = 0.20$ at each of the five samples.



Definition 57 (q^{th} Sample Quantile) For some $q \in [0, 1]$ and n IID RVs $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} F$, we can obtain the ECDF $\widehat{F}_n$ using (3.82). The **q^{th} sample quantile** is defined as the statistic (statistical functional):

$$T(\widehat{F}_n) = \widehat{F}_n^{[-1]}(q) := \inf\{x : \widehat{F}_n^{[-1]}(x) \geq q\} . \tag{3.83}$$

By replacing q in this definition of the q^{th} sample quantile by 0.25, 0.5 or 0.75, we obtain the first, second (**sample median**) or third **sample quartile**, respectively.

The following algorithm can be used to obtain the q^{th} sample quantile of n IID samples $(x_1, x_2, \ldots, x_n)$ on the basis of their order statistics $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$.

The q^{th} sample quantile, $\widehat{F}_n^{[-1]}(q)$, is found by interpolation from the order statistics $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ of the n data points $(x_1, x_2, \ldots, x_n)$, using the formula:

$$\widehat{F}_n^{[-1]}(q) = (1 - \delta)x_{(i+1)} + \delta x_{(i+2)}, \quad \text{where,} \quad i = \lfloor (n-1)q \rfloor \quad \text{and} \quad \delta = (n-1)q - \lfloor (n-1)q \rfloor .$$

Thus, the **sample minimum** of the data points $(x_1, x_2, \ldots, x_n)$ is given by $\widehat{F}_n^{[-1]}(0)$, the **sample maximum** is given by $\widehat{F}_n^{[-1]}(1)$ and the **sample median** is given by $\widehat{F}_n^{[-1]}(0.5)$, etc.

Labwork 112 (The q^{th} sample quantile) Use the implementation of Algorithm 1 as the MATLAB function `qthSampleQuantile` to find the q^{th} sample quantile of two simulated data arrays:

1. `SortedXsFromUni01Twstr101`, the order statistics that was constructed in Labwork 110 and
2. Another sorted array of 7 samples called `SortedXs`

1. The **mean** of RV $X \sim F$ is a function of the DF F:
$$T(F) = \mathbf{E}(X) = \int x \, dF(x) \ .$$
2. The **variance** of RV $X \sim F$ is a function of the DF F:
$$T(F) = \mathbf{E}(X - \mathbf{E}(X))^2 = \int (x - \mathbf{E}(X))^2 \, dF(x) \ .$$
3. The **value of DF at a given** $x \in \mathbb{R}$ of RV $X \sim F$ is also a function of DF F:
$$T(F) = F(x) \ .$$

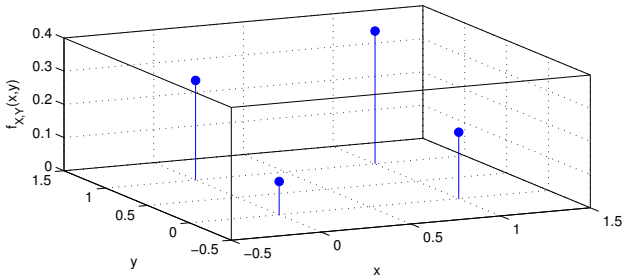
Other functionals of F that depend on the quantile function $F^{[-1]}$ are:

1. The q^{th} **quantile** of RV $X \sim F$:
$$T(F) = F^{[-1]}(q) \ \text{ where } q \in [0,1] \ .$$
2. The **first quartile** or the 0.25^{th} **quantile** of the RV $X \sim F$:
$$T(F) = F^{[-1]}(0.25) \ .$$
3. The **median** or the **second quartile** or the 0.50^{th} **quantile** of the RV $X \sim F$:
$$T(F) = F^{[-1]}(0.50) \ .$$
4. The **third quartile** or the 0.75^{th} **quantile** of the RV $X \sim F$:
$$T(F) = F^{[-1]}(0.75) \ .$$

Definition 56 (Empirical Distribution Function (EDF or ECDF)) Suppose we have n IID RVs, $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} F$, where F is a DF from the set of all DFs over the real line. Then, the n-sample empirical distribution function (EDF or ECDF) is the discrete distribution function $\widehat{F}_n$ that puts a probability mass of $1/n$ at each sample or data point x_i:
$$\widehat{F}_n(x) = \frac{\sum_{i=1}^n \mathbb{1}(X_i \leq x)}{n} \ , \quad \textit{where} \quad \mathbb{1}(X_i \leq x) := \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > x \end{cases} \tag{3.82}$$

Labwork 111 (Plot of empirical CDF) Let us plot the ECDF for the five samples drawn from the $Uniform(0,1)$ RV in Labwork 108 using the MATLAB function `ECDF`. Let us super-impose the samples and the true DF as depicted in Figure 3.26 with the following script:

plotunifecdf.m
```
xs = -1:0.01:2; % vector xs from -1 to 2 with increment .05 for x values
% get the [0,1] uniform DF or cdf of xs in vector cdf
cdf=zeros(size(xs));% initialise cdf as zero
indices = find(xs>=1); cdf(indices) = 1; % set cdf as 1 when xs >= 1
indices = find(xs>=0 & xs<=1); cdf(indices)=xs(indices); % cdf=xs when 0 <= xs <= 1
plot(xs,cdf,'r') % plot the DF
hold on; title('Uniform [0,1] DF and ECDF'); xlabel('x'); axis([-0.2 1.2 -0.2 1.2])
x=[0.5164, 0.5707, 0.0285, 0.1715, 0.6853]; % five samples
plot(x,zeros(1,5),'r+','LineWidth',2,'MarkerSize',10)% plot the data as red + marks
hold on; grid on; % turn on grid
ECDF(x,1,.2,.6);% ECDF (type help ECDF) plot is extended to left and right by .2 and .4, respectively.
```

It is helpful to write down the JPMF $f_{X,Y}(x,y)$ in a tabular form:

	$Y=0$	$Y=1$
$X=0$	0.1	0.3
$X=1$	0.2	0.4

From the above Table we can read for instance that the joint probability $f_{X,Y}(0,0) = 0.1$.

Find $\mathbf{P}(B)$ for the event $B = \{(0,0),(1,1)\}$, $F_{X,Y}(1/2,1/2)$, $F_{X,Y}(3/2,1/2)$, $F_{X,Y}(4,5)$ and $F_{X,Y}(-4,-1)$.

1. $\mathbf{P}(B) = \sum_{(x,y)\in\{(0,0),(1,1)\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) + f_{X,Y}(1,1) = 0.1 + 0.4$
2. $F_{X,Y}(1/2,1/2) = \sum_{\{(x,y):x\leq 1/2, y\leq 1/2\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) = 0.1$
3. $F_{X,Y}(3/2,1/2) = \sum_{\{(x,y):x\leq 3/2, y\leq 1/2\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) + f_{X,Y}(1,0) = 0.1 + 0.2 = 0.3$
4. $F_{X,Y}(4,5) = \sum_{\{(x,y):x\leq 4, y\leq 5\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) + f_{X,Y}(0,1) + f_{X,Y}(1,0) + f_{X,Y}(1,1) = 1$
5. $F_{X,Y}(-4,-1) = \sum_{\{(x,y):x\leq -4, y\leq -1\}} f_{X,Y}(x,y) = 0$

Definition 36 (JPDF) We say (X,Y) is a **continuous $\mathbb{R}^2$-valued random variable** if its JDF $F_{X,Y}(x,y)$ is differentiable and its **joint probability density function (JPDF)** is given by:
$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) \ .$$

For notational convenience, we sometimes suppress the subscripting when the random variables are clear from the context and write $f(x,y)$ and $F(x,y)$ instead of $f_{X,Y}(x,y)$ and $F_{X,Y}(x,y)$, respectively.

From JPDF $f_{X,Y}$ we can compute the JDF $F_{X,Y}$ at any point $(x,y) \in \mathbb{R}^2$ and more generally we can compute the probability of any event B, that can be cast as a region in $\mathbb{R}^2$, by simply taking two-dimensional integrals:
$$\boxed{F_{X,Y}(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u,v) du dv} \ , \tag{3.60}$$
and
$$\boxed{\mathbf{P}(B) \ = \ \int \int_B f_{X,Y}(x,y) dx dy} \ . \tag{3.61}$$

In particular, if $\mathbb{B}_\delta(x,y)$ denotes a square of a small area $\delta > 0$ that is centered at (x,y), then the following approximate equality holds and improves as $\delta \to 0$:

$$\mathbf{P}\left((X,Y) \in \mathbb{B}_\delta(x,y)\right) \approxeq \delta f_{X,Y}(x,y) \ . \tag{3.62}$$

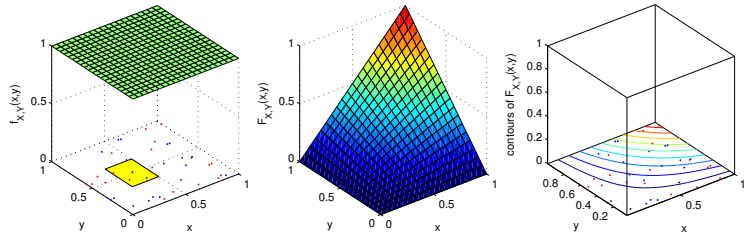
The JPDF satisfies the following two properties:

1. integrates to 1, i.e., $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$
2. is a non-negative function, i.e., $f_{X,Y}(x,y) \geq 0$ for every $(x,y) \in \mathbb{R}^2$.

Example 84 Let (X,Y) be a continuous R$\vec{\text{V}}$ that is uniformly distributed on the unit square $[0,1]^2 := [0,1]\times[0,1]$ with following JPDF:

$$f(x,y) = \mathbb{1}_{[0,1]^2}(x) \begin{cases} 1 & \text{if } (x,y) \in [0,1]^2 \\ 0 & \text{otherwise.} \end{cases}$$

Find explicit expressions for the following: (1) DF $F(x,y)$ for any $(x,y) \in [0,1]^2$, (2) $\mathbf{P}(X \leq 1/3, Y \leq 1/2)$, (3) $P\left((X,Y) \in [1/4,1/2]\times[1/3,2/3]\right)$.



Let us begin to find the needed expressions.

1. Let $(x,y) \in [0,1]^2$ then by Equation (3.60):
$$F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)dudv = \int_0^y\int_0^x 1 dudv = \int_0^y [u]_{u=0}^x dv = \int_0^y xdv \quad = [xv]_{v=0}^y = xy$$
2. We can obtain $\mathbf{P}(X \leq 1/3, Y \leq 1/2)$ by evaluating $F_{X,Y}$ at $(1/3, 1/2)$:
$$\mathbf{P}(X \leq 1/3, Y \leq 1/2) = F_{X,Y}(1/3,1/2) = \frac{1}{3}\frac{1}{2} = \frac{1}{6}$$
We can also find $\mathbf{P}(X \leq 1/3, Y \leq 1/2)$ by integrating the JPDF over the rectangular event $A = \{X < 1/3, Y < 1/2\} \subset [0,1]^2$ according to Equation (3.61). This amounts here to finding the area of A, we compute $\mathbf{P}(A) = (1/3)(1/2) = 1/6$.
3. We can find $P\left((X,Y) \in [1/4,1/2]\times[1/3,2/3]\right)$ by integrating the JPDF over the rectangular event $B = [1/4,1/2]\times[1/3,2/3]$ according to Equation (3.61):
$$\begin{aligned} P\left((X,Y) \in [1/4,1/2]\times[1/3,2/3]\right) &= \int\int_B f_{X,Y}(x,y)dxdy = \int_{1/3}^{2/3}\int_{1/4}^{1/2} 1 dxdy \\ &= \int_{1/3}^{2/3} [x]_{1/4}^{1/2} dy = \int_{1/3}^{2/3} \left[\frac{1}{2} - \frac{1}{4}\right] dy = \left(\frac{1}{2} - \frac{1}{4}\right)[y]_{1/3}^{2/3} \\ &= \left(\frac{1}{2} - \frac{1}{4}\right)\left(\frac{2}{3} - \frac{1}{3}\right) = \frac{1}{4}\left(\frac{1}{3}\right) = \frac{1}{12} \end{aligned}$$

Labwork 109 (Sample variance and sample standard deviation) We can compute the sample variance and sample standard deviation for the five samples stored in the array `XsFromUni01Twstr101` from Labwork 108 using MATLAB's functions `var` and `std`, respectively.

```
>> disp(XsFromUni01Twstr101); % The data-points x_1,x_2,x_3,x_4,x_5 are :
    0.5164    0.5707    0.0285    0.1715    0.6853
>> SampleVar=var(XsFromUni01Twstr101);% find sample variance
>> SampleStd=std(XsFromUni01Twstr101);% find sample standard deviation
>> disp(SampleVar) % The sample variance is:
    0.0785
>> disp(SampleStd) % The sample standard deviation is:
    0.2802
```

It is important to bear in mind that the statistics such as sample mean and sample variance are random variables and have an underlying distribution.

Definition 54 (Order Statistics) Suppose $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} F$, where F is the DF from the set of all DFs over the real line. Then, the n-sample **order statistics** $X_{([n])}$ is:

$$X_{([n])}(\ (X_1, X_2, \ldots, X_n)\ ) := \left(X_{(1)}, X_{(2)}, \ldots X_{(n)}\right), \text{ such that, } X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)} \ . \tag{3.80}$$

For brevity, we write $X_{([n])}(\ (X_1, X_2, \ldots, X_n)\ )$ as $X_{([n])}$ and its realisation $X_{([n])}(\ (x_1, x_2, \ldots, x_n)\ )$ as $x_{([n])} = (x_{(1)}, x_{(2)}, \ldots x_{(n)})$.

Without going into the details of how to sort the data in ascending order to obtain the order statistics (an elementary topic of an Introductory Computer Science course), we simply use MATLAB's function `sort` to obtain the order statistics, as illustrated in the following example.

Labwork 110 (Order statistics and sorting) The order statistics for the five samples stored in `XsFromUni01Twstr101` from Labwork 108 can be computed using `sort` as follows:

```
>> disp(XsFromUni01Twstr101); % display the sample points
    0.5164    0.5707    0.0285    0.1715    0.6853
>> SortedXsFromUni01Twstr101=sort(XsFromUni01Twstr101); % sort data
>> disp(SortedXsFromUni01Twstr101); % display the order statistics
    0.0285    0.1715    0.5164    0.5707    0.6853
```

Therefore, we can use `sort` to obtain our order statistics $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ from n sample points $x_1, x_2, \ldots, x_n$.

Next, we will introduce a family of common statistics, called the q^{th} quantile, by first defining the function:

Definition 55 (Inverse DF or Inverse CDF or Quantile Function) Let X be an RV with DF F. The **inverse DF** or **inverse CDF** or **quantile function** is:

$$F^{[-1]}(q) := \inf\{x : F(x) > q\}, \quad \text{for some } q \in [0,1] \ . \tag{3.81}$$

If F is strictly increasing and continuous then $F^{[-1]}(q)$ is the unique $x \in \mathbb{R}$ such that $F(x) = q$.

A **functional** is merely a function of another function. Thus, $T(F) : \{\text{All DFs }\} \to \mathbb{T}$, being a map or function from the space of DFs to its range $\mathbb{T}$, is a functional. Some specific examples of functionals we have already seen include:

Labwork 108 (Sample mean) After initializing the fundamental sampler, we draw five samples and then obtain the sample mean using the MATLAB function `mean`. In the following, we will reuse the samples stored in the array `XsFromUni01Twstr101`.

```
>> rand('twister',101); % initialise the fundamental Uniform(0,1) sampler
>> XsFromUni01Twstr101=rand(1,5); % simulate n=5 IID samples from Uniform(0,1) RV
>> SampleMean=mean(XsFromUni01Twstr101);% find sample mean
>> disp(XsFromUni01Twstr101); % The data-points x_1,x_2,x_3,x_4,x_5 are:
    0.5164    0.5707    0.0285    0.1715    0.6853
>> disp(SampleMean); % The Sample mean is :
    0.3945
```

We can thus use `mean` to obtain the sample mean $\overline{x}_n$ of n sample points $x_1, x_2, \ldots, x_n$.

We may also obtain the sample mean using the `sum` function and a division by sample size:

```
>> sum(XsFromUni01Twstr101) % take the sum of the elements of the XsFromUni01Twstr101 array
ans =    1.9723
>> sum(XsFromUni01Twstr101) / 5 % divide the sum by the sample size 5
ans =    0.3945
```

We can also obtain the sample mean via matrix product or multiplication as follows:

```
>> size(XsFromUni01Twstr101) % size(SomeArray) gives the size or dimensions of the arrar SomeArray
ans =     1     5
>> ones(5,1) % here ones(5,1) is an array of 1's with size or dimension 5 X 1
ans =
     1
     1
     1
     1
     1
>> XsFromUni01Twstr101 * ones(5,1) % multiplying an 1 X 5 matrix with a 5 X 1 matrix of Ones
ans =    1.9723
>> XsFromUni01Twstr101 * ( ones(5,1) * 1/5) % multiplying an 1 X 5 matrix with a 5 X 1 matrix of 1/5 's
ans =    0.3945
```

Definition 53 (Sample Variance & Standard Deviation) From a given a sequence of random variables $X_1, X_2, \ldots, X_n$, we may obtain another statistic called the n-samples variance or simply the sample variance :

$$T_n(\ (X_1, X_2, \ldots, X_n)\ ) = S_n^2(\ (X_1, X_2, \ldots, X_n)\ ) := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\ . \tag{3.78}$$

For brevity, we write $S_n^2(\ (X_1, X_2, \ldots, X_n)\ )$ as S_n^2 and its realisation $S_n^2(\ (x_1, x_2, \ldots, x_n)\ )$ as s_n^2.

Sample standard deviation is simply the square root of sample variance:

$$S_n(\ (X_1, X_2, \ldots, X_n)\ ) = \sqrt{S_n^2(\ (X_1, X_2, \ldots, X_n)\ )} \tag{3.79}$$

For brevity, we write $S_n(\ (X_1, X_2, \ldots, X_n)\ )$ as S_n and its realisation $S_n(\ (x_1, x_2, \ldots, x_n)\ )$ as s_n.

Once again, if $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} X_1$, the expectation of the sample variance is:

$$\mathbf{E}(S_n^2) = \mathbf{V}(X_1)\ .$$

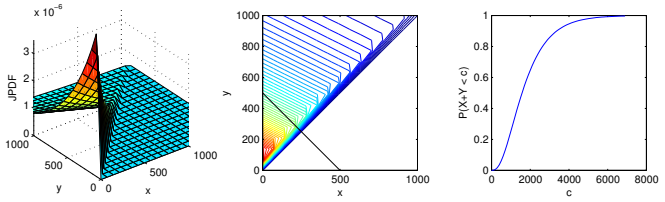
In general, for a bivariate uniform R$\vec{\text{V}}$ on the unit square the $\mathbf{P}([a,b]\times[c,d]) = (b-a)(d-c)$ for any event given by the rectangular region $[a,b]\times[c,d]$ inside the unit square $[0,1]\times[0,1]$. Thus any two events with the same rectangular area have the same probability (imagine sliding a small rectangle inside the unit square... no matter where you slide this rectangle to while remaining in the unit square, the probability of $\omega \mapsto (X(\omega), Y(\omega)) = (x, y)$ falling inside this "slidable" rectangle is the same...).

Example 85 Let the RV X denote the time until a web server connects to your computer, and let the RV Y denote the time until the server authorizes you as a valid user. Each of these RVs measures the waiting time from a common starting time (in milliseconds) and $X < Y$. From past response times of the web server we know that a good approximation for the JPDF of the R$\vec{\text{V}}$ (X, Y) is

$$f_{X,Y}(x,y) = \begin{cases} \frac{6}{10^6}\exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right) & \text{if } x > 0, y > 0, x < y \\ 0 & \text{otherwise.} \end{cases}$$

Answer the following:

1. identify the support of (X, Y), i.e., the region in the plane where $f_{X,Y}$ takes positive values
2. check that $f_{X,Y}$ indeed integrates to 1 as it should
3. Find $\mathbf{P}(X \leq 400, Y \leq 800)$
4. It is known that humans prefer a response time of under 1/10 seconds (10^2 milliseconds) from the web server before they get impatient. What is $\mathbf{P}(X + Y < 10^2)$?



Let us answer the questions.

1. The support is the intersection of the positive quadrant with the $y > x$ half-plane.

2.

$$\begin{aligned}
\int_{y=-\infty}^{\infty}\int_{x=-\infty}^{\infty} f_{X,Y}(x,y)dxdy &= \int_{x=0}^{\infty}\int_{y=x}^{\infty} f_{X,Y}(x,y)dydx \\
&= \int_{x=0}^{\infty}\int_{y=x}^{\infty} \frac{6}{10^6}\exp\left(-\frac{1}{1000}x-\frac{2}{1000}y\right)dydx \\
&= \frac{6}{10^6}\int_{x=0}^{\infty}\left(\int_{y=x}^{\infty}\exp\left(-\frac{2}{1000}y\right)dy\right)\exp\left(-\frac{1}{1000}x\right)dx \\
&= \frac{6}{10^6}\int_{x=0}^{\infty}\left[-\frac{1000}{2}\exp\left(-\frac{2}{1000}y\right)\right]_{y=x}^{\infty}\exp\left(-\frac{1}{1000}x\right)dx \\
&= \frac{6}{10^6}\int_{x=0}^{\infty}\left[0+\frac{1000}{2}\exp\left(-\frac{2}{1000}x\right)\right]\exp\left(-\frac{1}{1000}x\right)dx \\
&= \frac{6}{10^6}\int_{x=0}^{\infty}\frac{1000}{2}\exp\left(-\frac{2}{1000}x-\frac{1}{1000}x\right)dx \\
&= \frac{6}{10^6}\frac{1000}{2}\left[-\frac{1000}{3}\exp\left(-\frac{3}{1000}x\right)\right]_{x=0}^{\infty} \\
&= \frac{6}{10^6}\frac{1000}{2}\left[0+\frac{1000}{3}\right] \\
&= 1
\end{aligned}$$

3. First, identify the region with positive JPDF for the event $(X \leq 400, Y \leq 800)$

$$\begin{aligned}
&\mathbf{P}(X \leq 400, Y \leq 800) \\
&= \int_{x=0}^{400}\int_{y=x}^{800} f_{X,Y}(x,y)dydx \\
&= \int_{x=0}^{400}\int_{y=x}^{800} \frac{6}{10^6}\exp\left(-\frac{1}{1000}x-\frac{2}{1000}y\right)dydx \\
&= \frac{6}{10^6}\int_{x=0}^{400}\left[-\frac{1000}{2}\exp\left(-\frac{2}{1000}y\right)\right]_{y=x}^{800}\exp\left(-\frac{1}{1000}x\right)dx \\
&= \frac{6}{10^6}\frac{1000}{2}\int_{x=0}^{400}\left(-\exp\left(-\frac{1600}{1000}\right)+\exp\left(-\frac{2}{1000}x\right)\right)\exp\left(-\frac{1}{1000}x\right)dx \\
&= \frac{6}{10^6}\frac{1000}{2}\int_{x=0}^{400}\left(\exp\left(-\frac{3}{1000}x\right)-e^{-8/5}\exp\left(-\frac{1}{1000}x\right)\right)dx \\
&= \frac{6}{10^6}\frac{1000}{2}\left(\left(-\frac{1000}{3}\exp\left(-\frac{3}{1000}x\right)\right)_{x=0}^{400}-e^{-8/5}\left(-1000\exp\left(-\frac{1}{1000}x\right)\right)_{x=0}^{400}\right) \\
&= \frac{6}{10^6}\frac{1000}{2}1000\left(\frac{1}{3}\left(1-e^{-6/5}\right)-e^{-8/5}\left(1-e^{-2/5}\right)\right) \\
&= 3\left(\frac{1}{3}\left(1-e^{-6/5}\right)-e^{-8/5}\left(1-e^{-2/5}\right)\right) \\
&\approxeq 0.499\ .
\end{aligned}$$

4. First, identify the region with positive JPDF for the event $(X+Y \leq c)$, say $c = 500$ (but generally c can be any positive number). This is the triangular region at the intersection of the four half-planes: $x > 0$, $x < c$, $y > x$ and $y < c - x$. (*Draw picture here*) Let's integrate

data $X(\omega) = x$, a statistic? In other words, is the data a statistic? [Hint: consider the identity map $T(x) = x : \mathbb{X} \to \mathbb{T} = \mathbb{X}$.]

Next, we define two important statistics called the **sample mean** and **sample variance**. Since they are obtained from the sample data, they are called **sample moments**, as opposed to the **population moments**. The corresponding population moments are $\mathbf{E}(X_1)$ and $\mathbf{V}(X_1)$, respectively.

Definition 52 (Sample Mean) From a given a sequence of RVs $X_1, X_2, \ldots, X_n$, we may obtain another RV called the n-samples mean or simply the sample mean:

$$T_n(\ (X_1, X_2, \ldots, X_n)\ ) = \overline{X}_n(\ (X_1, X_2, \ldots, X_n)\ ) := \frac{1}{n}\sum_{i=1}^{n} X_i\ . \tag{3.75}$$

For brevity, we write

$$\overline{X}_n(\ (X_1, X_2, \ldots, X_n)\ ) \quad \text{as} \quad \overline{X}_n\ ,$$

and its realisation

$$\overline{X}_n(\ (x_1, x_2, \ldots, x_n)\ ) \quad \text{as} \quad \overline{x}_n\ .$$

Note that the expectation and variance of $\overline{X}_n$ are:

$$\begin{aligned}
\mathbf{E}(\overline{X}_n) &= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) && \text{[by definition (3.75)]} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(X_i\right) && \text{[by property (3.50)]}
\end{aligned}$$

Furthermore, if every X_i in the original sequence of RVs $X_1, X_2, \ldots$ is **identically** distributed with the same expectation, by convention $\mathbf{E}(X_1)$, then:

$$\mathbf{E}(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(X_1\right) = \frac{1}{n}\ n\ \mathbf{E}\left(X_1\right) = \mathbf{E}\left(X_1\right)\ . \tag{3.76}$$

Similarly, we can show that:

$$\begin{aligned}
\mathbf{V}(\overline{X}_n) &= \mathbf{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) && \text{[by definition (3.75)]} \\
&= \left(\frac{1}{n}\right)^2\mathbf{V}\left(\sum_{i=1}^{n} X_i\right) && \text{[by property (3.49)]}
\end{aligned}$$

Furthermore, if the original sequence of RVs $X_1, X_2, \ldots$ is **independently** distributed then:

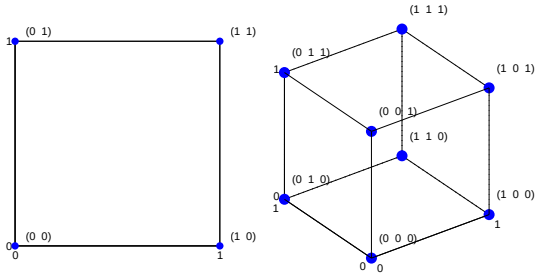
$$\mathbf{V}(\overline{X}_n) = \left(\frac{1}{n}\right)^2\mathbf{V}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\ \sum_{i=1}^{n}\mathbf{V}\left(X_i\right) \qquad \text{[by property (3.51)]}$$

Finally, if the original sequence of RVs $X_1, X_2, \ldots$ is **independently and identically** distributed with the same variance ($\mathbf{V}(X_1)$ by convention) then:

$$\mathbf{V}(\overline{X}_n) = \frac{1}{n^2}\ \sum_{i=1}^{n}\mathbf{V}\left(X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{V}\left(X_1\right) = \frac{1}{n^2}\ n\ \mathbf{V}\left(X_1\right) = \frac{1}{n}\ \mathbf{V}\left(X_1\right)\ . \tag{3.77}$$

Figure 3.24: Sample Space, Random Variable, Realisation, Data, and Data Space.

Figure 3.25: Data Spaces $\mathbb{X} = \{0,1\}^2$ and $\mathbb{X} = \{0,1\}^3$ for two and three Bernoulli trials, respectively.



Example 106 (Tossing a coin n times) For some given parameter $\theta \in \Theta := [0,1]$, consider n IID Bernoulli(θ) trials, i.e. $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta)$. Then the random vector $X = (X_1, X_2, \ldots, X_n)$, which takes values in the data space $\mathbb{X} = \{0,1\}^n := \{(x_1, x_2, \ldots, x_n) : x_i \in \{0,1\},\ i = 1,2,\ldots,n\}$, made up of vertices of the n-dimensional hyper-cube, measures the outcomes of this experiment. A particular realisation of X, upon performance of this experiment, is the observation, data or data vector $(x_1, x_2, \ldots, x_n)$. For instance, if we observed $n-1$ tails and 1 heads, in that order, then our data vector $(x_1, x_2, \ldots, x_{n-1}, x_n) = (0, 0, \ldots, 0, 1)$.

Definition 51 (Statistic) A **statistic** T is any function of the data:

$$T(x) : \mathbb{X} \to \mathbb{T} \,.$$

Thus, a statistic T is also an RV that takes values in the space $\mathbb{T}$. When $x \in \mathbb{X}$ is the realisation of an experiment, we let $T(x) = t$ denote the corresponding realisation of the statistic T. Sometimes we use $T_n(X)$ and $\mathbb{T}_n$ to emphasise that X is an n-dimensional random vector, i.e. $\mathbb{X} \subset \mathbb{R}^n$

Classwork 107 (Is data a statistic?) Is the RV X, for which the realisation is the observed

the JPDF over our triangular event as follows:

$$\begin{aligned}
\mathbf{P}(X+Y \le c) &= \int_{x=0}^{c/2}\int_{y=x}^{c-x} f_{X,Y}(x,y)dydx \\
&= \int_{x=0}^{c/2}\int_{y=x}^{c-x} \frac{6}{10^6}\exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right)dydx \\
&= \frac{6}{10^6}\int_{x=0}^{c/2}\int_{y=x}^{c-x} \exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right)dydx \\
&= \frac{6}{10^6}\frac{1000}{2}\int_{x=0}^{c/2}\left[-\exp\left(-\frac{2}{1000}y\right)\right]_{y=x}^{c-x}\exp\left(-\frac{1}{1000}x\right)dx \\
&= \frac{3}{10^3}\int_{x=0}^{c/2}\left[-\exp\left(-\frac{2c-2x}{1000}\right) + \exp\left(-\frac{2x}{1000}\right)\right]\exp\left(-\frac{x}{1000}\right)dx \\
&= \frac{3}{10^3}\int_{x=0}^{c/2}\left(\exp\left(-\frac{3x}{1000}\right) - \exp\left(\frac{x-2c}{1000}\right)\right)dx \\
&= 3\left(\left[-\frac{1}{3}\exp\left(-\frac{3x}{1000}\right)\right]_{x=0}^{c/2} - \left[e^{-2c/1000}\exp\left(\frac{x}{1000}\right)\right]_{x=0}^{c/2}\right) \\
&= 3\left(\frac{1}{3}(1-e^{-3c/2000}) - e^{-2c/1000}(e^{c/2000}-1)\right) \\
&= 1 - e^{-3c/2000} + 3e^{-2c/1000} - 3e^{-3c/2000} \\
&= 1 - 4e^{-3c/2000} + 3e^{-c/500}
\end{aligned}$$

5. $\mathbf{P}(X+Y<100) = 1 - 4e^{-300/2000} + 3e^{-100/500} \approxeq 0.134$. This means only about one in one hundred requests to this server will be processed within 100 milliseconds.

We can obtain $\mathbf{P}(X+Y<c)$ for several values of c using MATLAB and note that about 96% of requests are processed in less than 3000 milliseconds or 3 seconds.

```
>> c = [100 1000 2000 3000 4000]
c =  100         1000         2000         3000         4000

>> p = 1 - 4 * exp(-3*c/2000) + 3 * exp(-c/500)

p =  0.0134      0.5135       0.8558       0.9630       0.9911
```

Definition 37 (Marginal PDF or PMF) If the R$\vec{\text{V}}$ (X,Y) has $f_{X,Y}(x,y)$ as its joint PDF or joint PMF, then the **marginal PDF or PMF** of a random vector (X,Y) is defined by :

$$f_X(x) = \begin{cases} \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy & \text{if } (X,Y) \text{ is a continuous R}\vec{\text{V}} \\ \sum_y f_{X,Y}(x,y) & \text{if } (X,Y) \text{ is a discrete R}\vec{\text{V}} \end{cases}$$

and the **marginal PDF or PMF** of Y is defined by:

$$f_Y(y) = \begin{cases} \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx & \text{if } (X,Y) \text{ is a continuous R}\vec{\text{V}} \\ \sum_x f_{X,Y}(x,y) & \text{if } (X,Y) \text{ is a discrete R}\vec{\text{V}} \end{cases}$$

Example 86 Obtain the marginal PMFs $f_Y(y)$ and $f_X(x)$ from the joint PMF $f_{X,Y}(x,y)$ of the discrete R$\vec{\text{V}}$ in Example 83. Just sum $f_{X,Y}(x,y)$ over x's and y's (reported in a tabular form):

	$Y=0$	$Y=1$
$X=0$	0.1	0.3
$X=1$	0.2	0.4

From the above Table we can find:

$$f_X(x) = \mathbf{P}(X = x) = \sum_y f_{X,Y}(x, y)$$

$$= f_{X,Y}(x, 0) + f_{X,Y}(x, 1) = \begin{cases} f_{X,Y}(0,0) + f_{X,Y}(0,1) = 0.1 + 0.3 = 0.4 & \text{if } x = 0 \\ f_{X,Y}(1,0) + f_{X,Y}(1,1) = 0.2 + 0.4 = 0.6 & \text{if } x = 1 \end{cases}$$

Similarly,

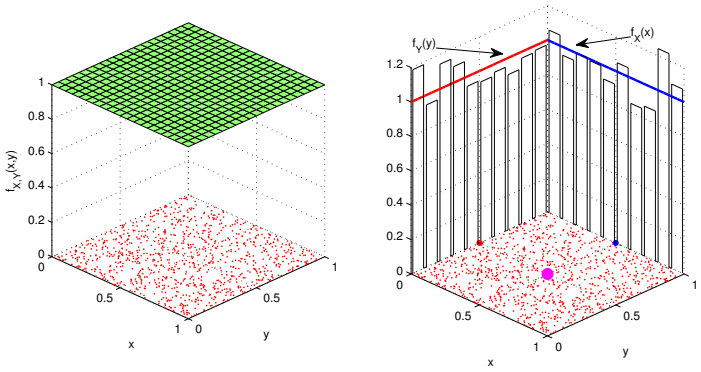
$$f_Y(y) = \mathbf{P}(Y = y) = \sum_x f_{X,Y}(x, y)$$

$$= f_{X,Y}(0, y) + f_{X,Y}(1, y) = \begin{cases} f_{X,Y}(0,0) + f_{X,Y}(1,0) = 0.1 + 0.2 = 0.3 & \text{if } y = 0 \\ f_{X,Y}(0,1) + f_{X,Y}(1,1) = 0.3 + 0.4 = 0.7 & \text{if } y = 1 \end{cases}$$

Just report the marginal probabilities as row and column sums of the JPDF table.

Thus marginal PMF gives us the probability of a specific RV, within a R$\vec{\text{V}}$, taking a value irrespective of the value taken by the other RV in this R$\vec{\text{V}}$.

Example 87 Obtain the marginal PDFs $f_Y(y)$ and $f_X(x)$ from the joint PDF $f_{X,Y}(x, y)$ of the continuous R$\vec{\text{V}}$ in Example 84 (the bivariate uniform R$\vec{\text{V}}$ on $[0, 1]^2$).



Let us suppose $(x, y) \in [0, 1]^2$ and note that $f_{X,Y} = 0$ if $(x, y) \notin [0, 1]^2$. We can obtain marginal PMFs $f_X(x)$ and $f_Y(y)$ by integrating the JPDF $f_{X,Y} = 1$ along y and x, respectively.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy = \int_0^1 f_{X,Y}(x, y)dy = \int_0^1 1dy = [y]_0^1 = 1 - 0 = 1$$

Similarly,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx = \int_0^1 f_{X,Y}(x, y)dx = \int_0^1 1dx = [x]_0^1 = 1 - 0 = 1$$

We are seeing a histogram of the **marginal samples** and their marginal PDFs in the Figure.

1. Find the characteristic function (CF) of X
2. Using the CF find $V(X)$, the variance of X. Hint: $V(X) = E(X^2) - (E(X))^2$

Ex. 3.51 — Recall that the Geometric(θ) RV X has the following PMF

$$f_X(x; \theta) = \begin{cases} \theta(1-\theta)^x & \text{if } x \in \{0, 1, 2, 3, \ldots\} \\ 0 & \text{otherwise} \end{cases}$$

1. Find the CF of X. (Hint: the sum of the infinite geometric series $\sum_{x=0}^{\infty} ar^x = \frac{a}{1-r}$.)
2. Using the CF find $E(X)$.

Ex. 3.52 — Let X be the Uniform(a, b) RV with the following probability density function (PDF)

$$f_X(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise } . \end{cases}$$

Find the CF of X.

Ex. 3.53 — Recall that the Poisson(λ) RV has the following PMF

$$f_X(x; \lambda) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{if } x \in \{0, 1, 2, 3, \ldots\} \\ 0 & \text{otherwise } . \end{cases}$$

Hint: the power series of $e^{\alpha} = \sum_{x=0}^{\infty} \frac{\alpha^x}{x!}$.

1. Find the CF of X.
2. Find the variance of X using its CF.

Ex. 3.54 — Let X be a Poisson(λ) RV and Y be another Poisson(μ) RV. Suppose X and Y are independent. Use Eqn. (3.74) to first find the CF of the RV $W = X + Y$. From the CF of W try to identify what RV it is.

Ex. 3.55 — Recall from lecture that if $Y = a + bX$ for some constants a and b with $b \neq 0$ then $\varphi_Y(t) = e^{\imath at}\varphi_X(bt)$ and that $\varphi_Z(t) = e^{-t^2/2}$ if Z is the Normal(0, 1) RV. Using these facts find the CF of $-Z$, the RV obtained from Z by simply switching its sign. From the CF of $-Z$ identify what RV it is.

3.14 Statistics

3.14.1 Data and Statistics

Definition 50 (Data) The function X measures the outcome ω of an experiment with sample space Ω [Often, the sample space is also denoted by S]. Formally, X is a random variable [or a random vector $X = (X_1, X_2, \ldots, X_n)$, i.e. a vector of random variables] taking values in the **data space** $\mathbb{X}$:

$$X(\omega) : \Omega \to \mathbb{X} .$$

The realisation of the RV X when an experiment is performed is the observation or data $x \in \mathbb{X}$. That is, when the experiment is performed once and it yields a specific $\omega \in \Omega$, the data $X(\omega) = x \in \mathbb{X}$ is the corresponding realisation of the RV X.

Example 105 Let Z_1 and Z_2 be independent Normal$(0,1)$ RVs.

1. Use Eqn. (3.74) to find the CF of $Z_1 + Z_2$.
2. From the CF of $Z_1 + Z_2$ identify what RV it is.
3. Use Eqn. (3.74) to find the CF of $2Z_1$.
4. From the CF of $2Z_1$ identify what RV it is.
5. Try to understand the difference between the distributions of $Z_1 + Z_2$ and $2Z_1$ inspite of Z_1 and Z_2 having the same distibution.

Hint: from lectures we know that $\varphi_X(t) = e^{\imath\mu t - (\sigma^2 t^2)/2}$ for a Normal(μ, σ^2) RV X.

Solution:

1. By Eqn. (3.74) we just multiply the characteristic functions of Z_1 and Z_2, both of which are $e^{-t^2/2}$,
$$\varphi_{Z_1+Z_2}(t) = \varphi_{Z_1}(t) \times \varphi_{Z_2}(t) = e^{-t^2/2} \times e^{-t^2/2} = e^{-2t^2/2} = e^{-t^2} \ .$$
2. The CF of $Z_1 + Z_2$ is that of the Normal(μ, σ^2) RV with $\mu = 0$ and $\sigma^2 = 2$. Thus $Z_1 + Z_2$ is the Normal$(0, 2)$ RV with mean parameter $\mu = 0$ and variance paramter $\sigma^2 = 2$.
3. We can again use Eqn. (3.74) to find the CF of $2Z_1$ as follows
$$\varphi_{2Z_1} = \varphi_{Z_1}(2t) = e^{-2^2 t^2/2} \ .$$
4. The CF of $2Z_1$ is that of the Normal(μ, σ^2) RV with $\mu = 0$ and $\sigma^2 = 2^2 = 4$. Thus $2Z_1$ is the Normal$(0, 4)$ RV with mean parameter $\mu = 0$ and variance paramter $\sigma^2 = 4$.
5. $2Z_1$ has a bigger variance from multiplying the standard normal RV by 2 while $Z_1 + Z_2$ has a smaller variance from adding two independent standard normal RVs. Thus, the result of adding the same RV twice does not have the same distribution as that of multiplying it by 2. In other words $2 \times Z$ is not equal to $Z + Z$ in terms of its probability distribution!

3.12.2 Moment Generating Function

Moment generating functions are special cases of characteristic functions and we won't be explicitly using them here as it is more convenient to work in the complex plane.

3.13 Exercises in Characteristic Functions

Ex. 3.50 — Let X be a discrete random variable (RV) with probability mass function (PMF)

$$f_X(x) = \begin{cases} \frac{1}{3} & \text{if } x = 0 \\ \frac{1}{3} & \text{if } x = 1 \\ \frac{1}{3} & \text{if } x = 2 \\ 0 & \text{otherwise} \ . \end{cases}$$

Thus marginal PDF gives us the probability density of a specific RV in a R$\vec{\text{V}}$, irrespective of the value taken by the other RV in this R$\vec{\text{V}}$.

Example 88 Obtain the marginal PDF $f_Y(y)$ from the joint PDF $f_{X,Y}(x,y)$ of the continuous R$\vec{\text{V}}$ in Example 85 that gave the response times of a web server.

$$f_{X,Y}(x,y) = \begin{cases} \frac{6}{10^6} \exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right) & \text{if } x > 0, y > 0, x < y \\ 0 & \text{otherwise.} \end{cases}$$

Use $f_Y(y)$ to compute the probability that Y exceeds 2000 milliseconds.
First we need to obtain an expression for $f_Y(y)$. For $y > 0$,

$$\begin{aligned} f_Y(y) &= \int_{x=-\infty}^{\infty} f_{X,Y}(x,y)dx \\ &= \int_{x=-\infty}^{\infty} 6 \times 10^{-6} e^{-0.001x - 0.002y} dx \\ &= 6 \times 10^{-6} \int_{x=0}^{y} e^{-0.001x - 0.002y} dx \\ &= 6 \times 10^{-6} e^{-0.002y} \int_{x=0}^{y} e^{-0.001x} dx \\ &= 6 \times 10^{-6} e^{-0.002y} \left[\frac{e^{-0.001x}}{-0.001}\right]_{x=0}^{x=y} \\ &= 6 \times 10^{-6} e^{-0.002y} \left(\frac{e^{-0.001y}}{-0.001} - \frac{e^{-0.001\times 0}}{-0.001}\right) \\ &= 6 \times 10^{-6} e^{-0.002y} \left(\frac{1 - e^{-0.001y}}{0.001}\right) \\ &= 6 \times 10^{-3} e^{-0.002y} \left(1 - e^{-0.001y}\right) \end{aligned}$$

We have the marginal PDF of Y and from this we can obtain

$$\begin{aligned} \mathbf{P}(Y > 2000) &= \int_{2000}^{\infty} f_Y(y)dy \\ &= \int_{2000}^{\infty} 6 \times 10^{-3} e^{-0.002y} \left(1 - e^{-0.001y}\right) dy \\ &= 6 \times 10^{-3} \int_{2000}^{\infty} e^{-0.002y} dy - \int_{2000}^{\infty} e^{-0.003y} dy \\ &= 6 \times 10^{-3} \left(\left(\left[\frac{e^{-0.002y}}{-0.002}\right]_{2000}^{\infty} - \left(\left[\frac{e^{-0.003y}}{-0.003}\right]_{2000}^{\infty}\right)\right)\right) \\ &= 6 \times 10^{-3} \left(\frac{e^{-4}}{0.002} - \frac{e^{-6}}{0.003}\right) \\ &= 0.05 \end{aligned}$$

Alternatively, you can obtain $\mathbf{P}(Y > 2000)$ by directly integrating the joint PDF $f_{X,Y}(x,y)$ over the appropriate region (but you may now have to integrate two pieces: rectangular infinite strip $(x,y) : 0 < x < 2000, y > 2000$ and a triangular infinite piece $\{(x,y) : y > x, y > 2000, x > 2000\}$)... more involved but we get the same answer.

$$\begin{aligned} \mathbf{P}(Y > 2000) = &\int_{x=0}^{2000} \left(\int_{y=2000}^{\infty} 6 \times 10^{-6} e^{-0.001x - 0.002y} dy\right) dx + \\ &\int_{x=2000}^{\infty} \left(\int_{y=x}^{\infty} 6 \times 10^{-6} e^{-0.001x - 0.002y} dy\right) dx \end{aligned}$$

$$\vdots \text{(try as a tutorial problem)}$$

$$\mathbf{P}(Y > 2000) = 0.0475 + 0.0025 = 0.05$$

We have seen the notion of independnece of two events in Definition 16 or of a sequence of events in Definition 17. Recall that independence amounts to having the probability of the joint occurrence of the events to be given by the product of the probabilities of each of the events.

We can use the definition of independence of two events to define the independence of two random variables using their distribution functions.

Definition 38 (Independence of Two RVs) Consider an $\mathbb{R}^2$-valued RV $X := (X_1, X_2)$. Then the $\mathbb{R}$-valued RVs X_1 and X_2 are said to be independent or independently distributed if and only if

$$\mathbf{P}(X_1 \leq x_1, X_2 \leq x_2) = \mathbf{P}(X_1 \leq x_1)\mathbf{P}(X_2 \leq x_2)$$

or equivalently,

$$F_{X_1,X_2}(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2) \ ,$$

for any pair of real numbers $(x_1, x_2) \in \mathbb{R}^2$.

By the above definition, for **discrete** RVs X_1, X_2 that are independent, the following equality is satisfied between the joint and marginal PMFs:

$$f_{X_1,X_2}(x_1, x_2) = \mathbf{P}(X_1 = x_1, X_2 = x_2) = \mathbf{P}(X_1 = x_1)\mathbf{P}(X_2 = x_2) = f_{X_1}(x_1)f_{X_2}(x_2) \text{ for any} (x_1, x_2) \in \mathbb{R}^2 \ ,$$

and for **continuous** RVs X_1, X_2 that are independent, the following equality is satisfied between the joint and marginal PDFs:

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2) \text{ for any} (x_1, x_2) \in \mathbb{R}^2 \ .$$

In summary, two RVs X and Y are said to be **independent** if and only if for every (x, y)

$$\boxed{F_{X,Y}(x, y) = F_X(x) \times F_Y(y) \qquad \text{or } f_{X,Y}(x, y) = f_X(x) \times f_Y(y)}$$

Let us confirm that our familiar experiment of tossing a fair coin twice independently when encoded by a pair of independent Bernoulli(1/2) RVs satisfies the above definition.

Example 89 (Pair of independent Bernoulli(1/2) **RVs)** Let X_1 and X_2 be a pair of independent Bernoulli(1/2) RVs each taking values in the set $\{0, 1\}$ with the following tabulated probabilities. Verify that the JPMF $f_{X_1,X_2}(x_1, x_2) = 1/4$ for each $(x_1, x_2) \in \{0, 1\}^2$ is indeed given by the marginal PMF $f_{X_i}(x_i) = 1/2$ for each $i \in \{1, 2\}$ and each $x_i \in \{0, 1\}$.

	$X_2 = 0$	$X_2 = 1$	
$X_1 = 0$	1/4	1/4	1/2
$X_1 = 1$	1/4	1/4	1/2
	1/2	1/2	1

From the above Table we can read for instance that the *joint probability* that $\mathbb{R}^2$-valued RV (X_1, X_2) takes the value or realization $(0, 0)$ is 1/4 from the first entry of the inner-most tabulated rectangle, i.e., $\mathbf{P}((X_1, X_2) = (0, 0)) = 1/4$, and that the *marginal probability* that the RV X_1 takes the value or relaization 0 is 1/2, i.e., $\mathbf{P}(X_1 = 0) = 1/2$. Clearly, $1/4 = 1/2 \times 1/2$, and so our familiar experiment when seen as an $\mathbb{R}^2$-valued RV is indeed composed of two independent$\mathbb{R}$-valued Bernoulli(1/2) RVs.

Thus the CF of the standard normal RV Z is

$$\boxed{\varphi_Z(t) = e^{-t^2/2}} \tag{3.72}$$

Let X be a RV with CF $\varphi_X(t)$. Let Y be a linear transformation of X

$$Y = a + bX$$

where a and b are two constant real numbers and $b \neq 0$. Then the CF of Y is

$$\boxed{\varphi_Y(t) = \exp(\imath at)\varphi_X(bt)} \tag{3.73}$$

Proof: This is easy to prove using the definition of CF as follows:

$$\begin{aligned}\varphi_Y(t) &= E\left(\exp(\imath tY)\right) = E\left(\exp(\imath t(a + bX))\right) = E\left(\exp(\imath ta + \imath tbX)\right) \\ &= E\left(\exp(\imath ta)\exp(\imath tbX)\right) = \exp(\imath ta)E\left(\exp(\imath tbX)\right) = \exp(\imath ta)\varphi_X(bt)\end{aligned}$$

Example 103 Let Y be a Normal(μ, σ^2) RV. Recall that Y is a linear transformation of Z, i.e., $Y = \mu + \sigma Z$ where Z is a Normal$(0, 1)$ RV. Using Equations (3.72) and (3.73) find the CF of Y.

Solution:

$$\begin{aligned}\varphi_Y(t) &= \exp(\imath\mu t)\varphi_Z(\sigma t), && \text{since } Y = \mu + \sigma Z \\ &= e^{\imath\mu t}e^{(-\sigma^2t^2)/2}, && \text{since } \varphi_Z(t) = e^{-t^2/2} \\ &= e^{\imath\mu t - (\sigma^2t^2)/2}\end{aligned}$$

A generalization of (3.73) is the following. If $X_1, X_2, \ldots, X_n$ are independent RVs and $a_1, a_2, \ldots, a_n$ are some constants, then the CF of the linear combination $Y = \sum_{i=1}^n a_i X_i$ is

$$\boxed{\varphi_Y(t) = \varphi_{X_1}(a_1t) \times \varphi_{X_2}(a_2t) \times \cdots \times \varphi_{X_n}(a_nt) = \prod_{i=1}^n \varphi_{X_i}(a_it) \ .} \tag{3.74}$$

Example 104 Using the following three facts:

- Eqn. (3.74)
- the Binomial(n, θ) RV Y is the sum of n independent Bernoulli(θ) RVs (from Probability Course)
- the CF of Bernoulli(θ) RV (from lecture notes for Inference Course)

find the CF of the Binomial(n, θ) RV Y.

Solution:

Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli(θ) RVs with CF $\left(1 - \theta + \theta e^{\imath t}\right)$ then $Y = \sum_{i=1}^n X_i$ is the Binomial(n, θ) RV and by Eqn. (3.74) with $a_1 = a_2 = \cdots = 1$, we get

$$\varphi_Y(t) = \varphi_{X_1}(t) \times \varphi_{X_2}(t) \cdots \varphi_{X_n}(t) = \prod_{i=1}^n \varphi_{X_i}(t) = \prod_{i=1}^n \left(1 - \theta + \theta e^{\imath t}\right) = \left(1 - \theta + \theta e^{\imath t}\right)^n \ .$$

Let us differentiate the CF to get moments using Equation (3.71) (CF has to be once and twice differentiable at $t = 0$ to get the first and second moments).

$$\begin{aligned}\frac{d}{dt}\varphi_X(t) &= \frac{d}{dt}\left(\frac{\lambda}{\lambda - \imath t}\right) = \lambda\left(-1 \times (\lambda - \imath t)^{-2} \times \frac{d}{dt}(\lambda - \imath t)\right) \\ &= \lambda\left(\frac{-1}{(\lambda - \imath t)^2} \times (-\imath)\right) = \frac{\lambda\imath}{(\lambda - \imath t)^2}\end{aligned}$$

We get $E(X)$ by evaluating $\frac{d}{dt}\varphi_X(t)$ at $t = 0$ and dividing by $\imath$ according to Equation (3.71) as follows:

$$E(X) = \frac{1}{\imath}\left[\frac{d}{dt}\varphi_X(t)\right]_{t=0} = \frac{1}{\imath}\left[\frac{\lambda\imath}{(\lambda - \imath t)^2}\right]_{t=0} = \frac{1}{\imath}\left(\frac{\lambda\imath}{\lambda^2}\right) = \frac{1}{\imath}\left(\frac{\imath}{\lambda}\right) = \frac{1}{\lambda}$$

Let's pause and see if this makes sense.... Yes, because the expected value of Exponential(λ) RV is indeed $1/\lambda$ (recall from when we introduced this RV).

Similarly from Equation (3.71) we can get $E(X^2)$ as follows:

$$\begin{aligned}E(X^2) &= \frac{1}{\imath^2}\left[\frac{d^2}{dt^2}\varphi_X(t)\right]_{t=0} = \frac{1}{\imath^2}\left[\frac{d}{dt}\frac{d}{dt}\varphi_X(t)\right]_{t=0} = \frac{1}{\imath^2}\left[\frac{d}{dt}\frac{\lambda\imath}{(\lambda - \imath t)^2}\right]_{t=0} \\ &= \frac{1}{\imath^2}\left[\lambda\imath \times \frac{d}{dt}(\lambda - \imath t)^{-2}\right]_{t=0} = \frac{1}{\imath^2}\left[\lambda\imath\left(-2(\lambda - \imath t)^{-3}\frac{d}{dt}(\lambda - \imath t)\right)\right]_{t=0} \\ &= \frac{1}{\imath^2}\left[\lambda\imath\left(-2(\lambda - \imath t)^{-3} \times (-\imath)\right)\right]_{t=0} = \frac{1}{\imath^2}\left[\frac{2\lambda\imath^2}{(\lambda - \imath t)^3}\right]_{t=0} = \frac{1}{\imath^2}\left(\frac{2\lambda\imath^2}{\lambda^3}\right) = \frac{2}{\lambda^2}\ .\end{aligned}$$

Finally, from the first and second moments we can get the variance as follows:

$$V(X) = E(X^2) - (E(X))^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{2-1}{\lambda^2} = \frac{1}{\lambda^2}\ .$$

Let's check that this is what we had as variance for the Exponential(λ) RV when we first introduced it and directly computed using integrals for definition of expectation.

Characteristic functions can be used to characterize the distribution of a random variable.

Two RVs X and Y have the same DFs , i.e., $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$, if and only if they have the same characteristic functions, i.e. $\varphi_X(t) = \varphi_Y(t)$ for all $t \in \mathbb{R}$ (for proof see Resnick, S. I. (1999) A Probability Path, Birkhauser).

Thus, if we can show that two RVs have the same CF then we know they are the same. This can be much more challenging or impossible to do directly with their DFs.

Let Z be Normal$(0, 1)$, the standard normal RV. We can find the CF for Z using couple of tricks as follows

$$\begin{aligned}\varphi_Z(t) &= E\left(e^{\imath tZ}\right) \\ &= \int_{-\infty}^{\infty} e^{\imath tz} f_Z(z)dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\imath tz}e^{-z^2/2}dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\imath tz - z^2/2}dz \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(t^2 + (z - \imath t)^2)/2}dz = e^{-t^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(z-\imath t)^2/2}dz \\ &= e^{-t^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}dy \qquad \text{substituting } y = z - \imath t, dy = dz \\ &= e^{-t^2/2}\frac{1}{\sqrt{2\pi}}\sqrt{2\pi} \qquad \text{using the normalizing constant in PDF of Normal}(0, 1) \text{ RV} \\ &= e^{-t^2/2}\end{aligned}$$

Example 90 Recall the $\mathbb{R}^2$-valued continuous RV (X, Y) of Example 84 that is uniformly distributed on the unit square $[0, 1]^2$. First show that X and Y independent. Then show that both X and Y are identically distributed according to the Uniform$(0, 1)$ RV.

Solution:

This can be shown by checking that the joint PDF is indeed equal to the product of the marginal PDFs of Uniform$(0, 1)$ RVs as follows:

$$\begin{cases} 1 = f_{X,Y}(x, y) = f_X(x) \times f_Y(y) = 1 \times 1 = 1 & \text{if } (x, y) \in [0, 1]^2 \\ 0 = f_{X,Y}(x, y) = f_X(x) \times f_Y(y) = 0 \times 0 = 0 & \text{if } (x, y) \notin [0, 1]^2 \end{cases}$$

Are X and Y independent in the server times R$\vec{\text{V}}$ from Example 85?

We can compute $f_X(x)$ and use the already computed $f_Y(y)$ to mechanically check if the JPDF is the product of the marginal PDFs. But intuitively, we know that these RVs (connection time and authentication time) are dependent – one is strictly greater than the other. Also the JPDF has zero density when $x > y$, but the product of the marginal densities won't.

Now, let us take advantage of independent random variables and solve some problems.

Example 91 (distance between random faults in a manufactured line) Suppose two points are tossed independently and uniformly at random onto a line segment of unit length. What is the probability that the distance between the two points does not exceed a given length l?

done in lectures...

Example 92 (Buffon's Needle Experiment to Physically Estimate π) Suppose a needle is tossed at random onto a plane ruled with parallel lines a distance L apart. By a "needle" we mean a line segment of length $l \leq L$.

What is the probability that the needle intersects one of the parallel lines? Can you use repeated trials of this experiment to find an approximation to π?

Solution:

Let X_1 be the angle between the needle and the direction of the rulings, and let X_2 be the distance between the bottom point of the needle and the nearest line above this point (see left sub-figure of Figure 3.19). Then the conditions of the "needle tossing at random" experiment are such that the RV X_1 is uniformly distributed in the interval $[0, \pi]$, while the RV X_2 is uniformly distributed in the interval $[0, L]$. Hence *assuming that the RVs X_1 and X_2 are independent*, we find that their joint probability density function (JPDF) is:

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{\pi}\mathbb{1}_{[0,\pi]}(x_1) \times \frac{1}{L}\mathbb{1}_{[0,L]}(x_2) = \frac{1}{\pi L}\mathbb{1}_{[0,\pi]}(x_1)\mathbb{1}_{[0,L]}(x_2) = \frac{1}{\pi L}\mathbb{1}_{[0,\pi]\times[0,L]}(x_1, x_2)\ .$$

Figure 3.19: Diagrams done on the board!

The event A that the needle intersects one of the parallel ruled lines occurs if and only if

$$X_2 \leq l\sin(X_1) \ ,$$

i.e., if and only if the corresponding point $X := (X_1, X_2)$ falls in the region B, where B is part of the rectangle $[0,\pi] \times [0, L]$ lying between the x_1-axis and the curve $x_2 = \sin(x_1)$ (area under the curve in right-subfigure of Figure 3.19). Hence, we can integrate the JPDF to get the probability of the event A of interest:

$$\mathbf{P}(A) = \mathbf{P}\left((X_1, X_2) \in B\right) = \int\int_B \frac{dx_1 dx_2}{\pi L} = \frac{2l}{\pi L}$$

where,

$$l\int_0^{\pi} \sin(x_1)dx_1 = l\left(-\cos(x_1)\right]_0^{\pi} = l(1-(-1)) = l(1+1) = 2l \ ,$$

is the area of B.

Thus, if the needle is repeatedly tossed onto the ruled plane and $n(A)$ is the number of times A occurs out of n trials, then the relative frequency of the event A should approach $\mathbf{P}(A)$ as $n \to \infty$ (we will see this as the Law of Large Numbers in the sequel, but recall that this is also how we motivated the LTRF or long-term relative frequency idea of probability):

$$\frac{n(A)}{n} \to \frac{2l}{\pi L}$$

Hence, for large n,

$$\frac{2l}{L}\frac{n}{n(A)}$$

should be a good approximation to $\pi = 3.14\ldots$. This is indeed the case.

3.10.2 Conditional Random Variables

Often we will have a condition where one of the two random variables that make up a random vector (X_1, X_2) already occurs and takes a value. And we might want to compute the probability of the occurrence of the other random variable given this conditional information. For this all we need to do is extend the idea of conditional probabiliies to $\mathbb{R}^2$-valued random variables as defined below.

Definition 39 (Conditional PDF or PMF) Let (X_1, X_2) be a discrete bivariate R$\vec{\text{V}}$. The conditional PMF of $X_1|X_2 = x_2$, where $f_{X_2}(x_2) := \mathbf{P}(X_2 = x_2) > 0$ is:

$$f_{X_1|X_2}(x_1|x_2) := \mathbf{P}(X_1 = x_1|X_2 = x_2) = \frac{\mathbf{P}(X_1 = x_1, X_2 = x_2)}{\mathbf{P}(X_2 = x_2)} = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} \ .$$

Part 2:

Let's differentiate CF

$$\frac{d}{dt}\varphi_X(t) = \frac{d}{dt}\left(1-\theta+\theta e^{\imath t}\right) = \theta\imath\exp(\imath t)$$

We get $E(X)$ by evaluating $\frac{d}{dt}\varphi_X(t)$ at $t = 0$ and dividing by $\imath$ according to Equation (3.71) as follows:

$$E(X) = \frac{1}{\imath}\left[\frac{d}{dt}\varphi_X(t)\right]_{t=0} = \frac{1}{\imath}\left[\theta\imath\exp(\imath t)\right]_{t=0} = \frac{1}{\imath}\left(\theta\imath\exp(\imath 0)\right) = \theta \ .$$

Similarly from Equation (3.71) we can get $E(X^2)$ as follows:

$$\begin{aligned} E(X^2) &= \frac{1}{\imath^2}\left[\frac{d^2}{dt^2}\varphi_X(t)\right]_{t=0} = \frac{1}{\imath^2}\left[\frac{d}{dt}\frac{d}{dt}\varphi_X(t)\right]_{t=0} = \frac{1}{\imath^2}\left[\frac{d}{dt}\theta\imath\exp(\imath t)\right]_{t=0} \\ &= \frac{1}{\imath^2}\left[\theta\imath^2\exp(\imath t)\right]_{t=0} = \frac{1}{\imath^2}\left(\theta\imath^2\exp(\imath 0)\right) = \theta \ . \end{aligned}$$

Finally, from the first and second moments we can get the variance as follows:

$$V(X) = E(X^2) - (E(X))^2 = \theta - \theta^2 = \theta(1-\theta) \ .$$

Let's check that this is what we have as variance for the Bernoulli(θ) RV if we directly computed it using weighted sums in the definition of expectations: $E(X) = 1\times\theta + 0\times(1-\theta) = \theta$, $E(X^2) = 1^2\times\theta + 0^2\times(1-\theta) = \theta$and thus giving the same $V(X) = E(X^2) - (E(X))^2 = \theta - \theta^2 = \theta(1-\theta)$.

Example 102 Let X be an Exponential(λ) RV. First show that its CF is $\lambda/(\lambda - \imath t)$. Then use CF to find $E(X)$, $E(X^2)$ and from this obtain the variance $V(X) = E(X^2) - (E(X))^2$.

Solution:

Recall that the PDF of an Exponential(λ) RV for a given parameter $\lambda \in (0,\infty)$ is $\lambda e^{-\lambda x}$ if $x \in [0,\infty)$ and 0 if $x \notin [0,\infty)$.

Part 1: Find the CF.

We will use the fact that

$$\int_0^{\infty} \alpha e^{-\alpha x}dx = \left[-e^{-\alpha x}\right]_0^{\infty} = 1$$

$$\begin{aligned} \varphi_X(t) = E(\exp(\imath Xt)) = E\left(e^{\imath Xt}\right) &= \int_{-\infty}^{\infty} e^{\imath xt}\lambda e^{-\lambda x}dx = \lambda\int_0^{\infty} e^{-(\lambda-\imath t)x}dx \\ &= \frac{\lambda}{\lambda-\imath t}\int_0^{\infty}(\lambda-\imath t)e^{-(\lambda-\imath t)x}dx = \frac{\lambda}{\lambda-\imath t}\int_0^{\infty}\alpha e^{-\alpha x}dx = \frac{\lambda}{\lambda-\imath t} \ , \end{aligned}$$

where $\alpha = \lambda - \imath t$ with $\lambda > 0$.

Alternatively, you can use $e^{\imath tx} = \cos(tx) + \imath\sin(tx)$ and do integration by parts to arrive at the same answer starting from:

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{\imath xt}\lambda e^{-\lambda x}dx = \int_{-\infty}^{\infty}\cos(tx)e^{-\lambda x}dx + \imath\int_{-\infty}^{\infty}\sin(tx)e^{-\lambda x}dx\vdots = \frac{\lambda}{\lambda-\imath t} \ .$$

Part 2:

Proof: The proper proof is very messy so we just give a sketch of the ideas in the proof. Due to the linearity of the expectation (integral) and the derivative operators, we can change the order of operations:

$$\frac{d^k\varphi_X(t)}{dt^k} = \frac{d^k}{dt^k}E(\exp(\imath tX)) = E\left(\frac{d^k}{dt^k}\exp(\imath tX)\right) = E\left((\imath X)^k\exp(\imath tX)\right) = \imath^k E\left(X^k\exp(\imath tX)\right)$$

The RHS evaluated at $t=0$ is

$$\left[\frac{d^k\varphi_X(t)}{dt^k}\right]_{t=0} = \left[\imath^k E\left(X^k\exp(\imath tX)\right)\right]_{t=0} = \imath^k E\left(X^k\right)$$

This completes the sketch of the proof.

The above Theorem gives us the relationship between the moments and the derivatives of the CF if we already know that the moment exists. When one wants to compute a moment of a random variable, what we need is the following Theorem.

Proposition 49 (Moments from CF.) Let X be a random variable and $\varphi_X(t)$ be its CF. If $\varphi_X(t)$ is k times differentiable at the point $t=0$, then

1. if k is even, the n-th moment of X exists and is finite for any $0 \le n \le k$;
2. if k is odd, the n-th moment of X exists and is finite for any $0 \le n \le k-1$.

In both cases,

$$E(X^k) = \frac{1}{\imath^k}\left[\frac{d^k\varphi_X(t)}{dt^k}\right]_{t=0} . \tag{3.71}$$

where $\left[\frac{d^k\varphi_X(t)}{dt^k}\right]_{t=0}$ is the k-th derivative of $\varphi_X(t)$ with respect to t, evaluated at the point $t=0$.

Proof: For proof see e.g., Ushakov, N. G. (1999) Selected topics in characteristic functions, VSP (p. 39).

Example 101 Let X be the Bernoulli(θ) RV. Find the CF of X. Then use CF to find $E(X)$, $E(X^2)$ and from this obtain the variance $V(X) = E(X^2) - (E(X))^2$.

Solution:

Part 1

Recall the PMF for this discrete RV with parameter $\theta \in (0,1)$ is

$$f_X(x;\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Let's first find the CF of X

$$\varphi_X(t) = E\left(\exp(\imath tX)\right) = \sum_x \exp(\imath tx) f_X(x;\theta) \qquad \text{By Defn. in Equation (3.70)}$$
$$= \exp(\imath t\times 0)(1-\theta) + \exp(\imath t\times 1)\theta = \exp(0)(1-\theta) + \exp(\imath t)\theta = 1-\theta+\theta\exp(\imath t)$$

Similarly, if $f_{X_1}(x_1) := \mathbf{P}(X_1 = x_1) > 0$, then the conditional PMF of $X_2|X_1 = x_1$ is:

$$f_{X_2|X_1}(x_2|x_1) := \mathbf{P}(X_2 = x_2|X_1 = x_1) = \frac{\mathbf{P}(X_1 = x_1, X_2 = x_2)}{\mathbf{P}(X_1 = x_1)} = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)} .$$

If (X_1, X_2) are continuous RVs such that the marginal PDF $f_{X_2}(x_2) > 0$, then the conditional PDF of $X_1|X_2 = x_2$ is:

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)}, \qquad \mathbf{P}(X_1 \in A|X_2 = x_2) = \int_A f_{X_1|X_2}(x_1|x_2)dx_1 .$$

Similarly, if $f_{X_1}(x_1) > 0$, then the conditional PDF of $X_2|X_1 = x_1$ is:

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)}, \qquad \mathbf{P}(X_2 \in A|X_1 = x_1) = \int_A f_{X_2|X_1}(x_2|x_1)dx_2 .$$

Let us consider a few discrete RVs for the simple coin tossing experiment $\mathcal{E}_\theta^3$ that build on the Bernoulli(θ) RV X_i for the i-th toss in an **independent and identically distibuted (IID.)** manner.

Table 3.1: The 8 ω's in the sample space Ω of the experiment $\mathcal{E}_\theta^3$ are given in the first row above. The RV Y is the number of 'Heads' in the 3 tosses and the RV Z is the number of 'Tails' in the 3 tosses. Finally, the RVs Y' and Z' are the indicator functions of the event that 'all three tosses were Heads' and the event that 'all three tosses were Tails', respectively.

ω:	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT	RV Definitions / Model
$\mathbf{P}(\omega)$:	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$X_i \overset{\text{IID}}{\sim} \text{Bernoulli}(\frac{1}{2})$
$Y(\omega)$:	3	2	2	1	2	1	1	0	$Y := X_1 + X_2 + X_3$
$Z(\omega)$:	0	1	1	2	1	2	2	3	$Z := (1-X_1) + (1-X_2) + (1-X_3)$
$Y'(\omega)$:	1	0	0	0	0	0	0	0	$Y' := X_1X_2X_3$
$Z'(\omega)$:	0	0	0	0	0	0	0	1	$Y' := (1-X_1)(1-X_2)(1-X_3)$

Classwork 93 (Two random variables of 'toss a coin thrice' experiment) Describe the probability of the RV Y and Y' of Table 3.1 in terms of its PMF. Repeat the process for the RV Z in your spare time.

$$\mathbf{P}(Y = y) = \begin{cases} \end{cases} \qquad \mathbf{P}(Y' = y') = \begin{cases}$$

Classwork 94 (The number of 'Heads' given there is at least one 'Tails') Consider the following two questions.

1. What is conditional probability $\mathbf{P}(Y|Y'=0)$?

$\mathbf{P}(Y=y\|Y'=0)$	$=\frac{\mathbf{P}(Y=y,Y'=0)}{\mathbf{P}(Y'=0)}$	$=\frac{\mathbf{P}(\{\omega:Y(\omega)=y\ \cap\ Y'(\omega)=0\})}{\mathbf{P}(\{\omega:Y'(\omega)=0\})}$	$=?$
$\mathbf{P}(Y=0\|Y'=0)$	$\frac{\mathbf{P}(Y=0,Y'=0)}{\mathbf{P}(Y'=0)}$	$\frac{\frac{1}{8}}{\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}}$	$\frac{1}{7}$
$\mathbf{P}(Y=1\|Y'=0)$	$\frac{\mathbf{P}(Y=1,Y'=0)}{\mathbf{P}(Y'=0)}$	$\frac{\frac{1}{8}+\frac{1}{8}+\frac{1}{8}}{\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}}$	$\frac{3}{7}$
$\mathbf{P}(Y=2\|Y'=0)$	$\frac{\mathbf{P}(Y=2,Y'=0)}{\mathbf{P}(Y'=0)}$	$\frac{\frac{1}{8}+\frac{1}{8}+\frac{1}{8}}{\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}}$	$\frac{3}{7}$
$\mathbf{P}(Y=3\|Y'=0)$	$\frac{\mathbf{P}(Y=3,Y'=0)}{\mathbf{P}(Y'=0)}$	$\frac{\mathbf{P}(\emptyset)}{\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}}$	0
$\mathbf{P}(Y\in\{0,1,2,3\}\|Y'=0)$	$\frac{\sum_{y=0}^{3}\mathbf{P}(Y=y,Y'=0)}{\mathbf{P}(Y'=0)}$	$\frac{\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}}{\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}}$	1

2. What is $\mathbf{P}(Y|Y'=1)$?

$$\mathbf{P}(Y=y|Y'=1)=\begin{cases}1 & \text{if } y=3\\ 0 & \text{otherwise}\end{cases}$$

3.10.3 $\mathbb{R}^m$-valued Random Variables

Consider the $\vec{\text{RV}}$ X whose components are the RVs $X_1, X_2, \ldots, X_m$, i.e., $X := (X_1, X_2, \ldots, X_m)$, where $m \geq 2$. A particular realization of this RV is a point $(x_1, x_2, \ldots, x_m)$ in $\mathbb{R}^m$. Now, let us extend the notions of JCDF, JPMF and JPDF to $\mathbb{R}^m$.

Definition 40 (multivariate JDF) The **joint distribution function (JDF)** or **joint cumulative distribution function (JCDF)**, $F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) : \mathbb{R}^m \to [0, 1]$, of the multivariate random vector $(X_1, X_2, \ldots, X_m)$ is

$$\begin{aligned} F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) &= \mathbf{P}(X \leq x_1 \cap X_2 \leq x_2 \cap \cdots \cap X_m \leq x_m) \\ &= \mathbf{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_m \leq x_m) \\ &= P\left(\{\omega : X_1(\omega) \leq x_1, X_2(\omega) \leq x_2, \ldots, X_m(\omega) \leq x_m\}\right), \end{aligned} \tag{3.63}$$

for any $(x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$, where the right-hand side represents the probability that the random vector $(X_1, X_2, \ldots, X_m)$ takes on a value in $\{(x'_1, x'_2, \ldots, x'_m) : x'_1 \leq x_1, x'_2 \leq x_2, \ldots, x'_m \leq x_m\}$, the set of points in $\mathbb{R}^m$ that are less than the point $(x_1, x_2, \ldots, x_m)$ in each coordinate $1, 2, \ldots, m$.

The JDF $F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) : \mathbb{R}^m \to \mathbb{R}$ satisfies the following conditions to remain a probability:

1. $0 \leq F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) \leq 1$
2. $F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m)$ is an increasing function of x_1, x_2, ... and x_m
3. $F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) \to 1$ as $x_1 \to \infty$, $x_2 \to \infty$, ... and $x_m \to \infty$

3.12 Characteristic Functions

The characteristic function (CF) of a random variable gives another way to specify its distribution. Thus CF is a powerful tool for analytical results involving random variables (more).

Definition 47 (Characteristic Function (CF)) Let X be a RV and $\imath = \sqrt{-1}$. The function $\varphi_X(t) : \mathbb{R} \to \mathbb{C}$ defined by

$$\boxed{\varphi_X(t) := E\left(\exp\left(\imath t X\right)\right) = \begin{cases}\sum_x \exp\left(\imath t x\right) f_X(x) & \text{if } X \text{ is discrete RV}\\ \int_{-\infty}^{\infty} \exp\left(\imath t x\right) f_X(x) dx & \text{if } X \text{ is continuous RV}\end{cases}} \tag{3.70}$$

is called the **characteristic function** of X.

NOTE: $\varphi_X(t)$ exists for any $t \in \mathbb{R}$, because

$$\begin{aligned} \varphi_X(t) &= E\left(\exp\left(\imath t X\right)\right) \\ &= E\left(\cos(tX) + \imath \sin(tX)\right) \\ &= E\left(\cos(tX)\right) + \imath E\left(\sin(tX)\right) \end{aligned}$$

and the last two expected values are well-defined, because the sine and cosine functions are bounded by $[-1, 1]$.

For a continuous RV, $\int_{-\infty}^{\infty} \exp\left(-\imath t x\right) f_X(x) dx$ is called the *Fourier transform* of f_X. This is the CF but with t replaced by $-t$. You will also encounter Fourier transforms when solving differential equations.

3.12.1 Obtaining Moments from Characteristic Function

Recall that the k-th moment of X is $E(X^k)$ for any $k \in \mathbb{N} := \{1, 2, 3, \ldots\}$ is

$$\boxed{E(X^k) \;=\; \begin{cases}\sum\limits_x x^k f_X(x) & \text{if } X \text{ is a discrete RV}\\ \int\limits_{-\infty}^{\infty} x^2 f_X(x) dx & \text{if } X \text{ is a continuous RV}\end{cases}}$$

The characteristic function can be used to derive the moments of X due to the following nice relationship between the the k-th moment of X and the k-th derivative of the CF of X.

Proposition 48 (Moment & CF.) Let X be a random variable and $\varphi_X(t)$ be its CF. If $E(X^k)$ exists and is finite, then $\varphi_X(t)$ is k times continuously differentiable and

$$E(X^k) = \frac{1}{\imath^k}\left[\frac{d^k \varphi_X(t)}{dt^k}\right]_{t=0} .$$

where $\left[\frac{d^k \varphi_X(t)}{dt^k}\right]_{t=0}$ is the k-th derivative of $\varphi_X(t)$ with respect to t, evaluated at the point $t = 0$.

Ex. 3.45 — Let (X, Y) be a continuous R$\vec{\text{V}}$ with joint probability density function (JPDF):

$$f_{X,Y}(x,y) = \begin{cases} e^{-x} & \text{if } x \in [0,\infty) \text{ and } y \in [2,3] \\ 0 & \text{otherwise .} \end{cases}$$

Are X and Y independent?

Ex. 3.46 — In an electronic assembly, let the RVs X_1, X_2, X_3, X_4 denote the lifetimes of four components in hours. Suppose that the JPDF of these variables is

$$f_{X_1,X_2,X_3,X_4}(x_1,x_2,x_3,x_4) = \begin{cases} 9\times 10^{-12} e^{-0.001x_1 - 0.002x_2 - 0.0015x_3 - 0.003x_4} & \text{if } x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0 \\ 0 & \text{otherwise .} \end{cases}$$

What is the probability that the device operates for more than 1000 hours without any failures? [Hint: The requested probability is $\mathbf{P}(X_1 > 1000, X_2 > 1000, X_3 > 1000, X_4 > 1000)$ since each one of the four components of the device must not fail before 1000 hours.]

Ex. 3.47 — Suppose the RVs Y_1, Y_2 and Y_3 represent the thickness in micrometers of a substrate, an active layer, and a coating layer of a chemical product. Assume Y_1, Y_2 and Y_3 are Normal$(10000, 250^2)$, Normal$(1000, 20^2)$ and Normal$(80, 4^2)$ RVs, respectively. Further suppose that they are independent. The required specifications for the thickness of the substrate, active layer and coating layer are $[9500, 10500]$, $[950, 1050]$ and $[75, 85]$, respectively. What proportion of chemical products meets all thickness specifications? [Hint: this is just $\mathbf{P}(9500 < Y_1 < 10500, 950 < Y_2 < 1050, 75 < Y_3 < 85)$] Which one of the three thicknesses has the least probability of meeting specifications?

Ex. 3.48 — Soft drink cans are filled by an automated filling machine. Assume the fill volumes of the cans are independent Normal$(12.1, 0.01)$ RVs. What is the probability that the average volume of ten cans selected from this process is less than 12.01 fluid ounces?

Ex. 3.49 — Let X_1, X_2, X_3, X_4 be RVs that denote the number of bits received in a digital channel that are classified as *excellent*, *good*, *fair* and *poor*, respectively. In a transmission of 10 bits, what is the probability that 6 of the bits received are *excellent*, 2 are *good*, 2 are *fair* and none are *poor* under the assumption that the classification of bits are independent events and that the probabilities of each bit being *excellent*, *good*, *fair* and *poor* are 0.6, 0.3, 0.08 and 0.02, respectively. [Hint: Think of Multinomial$(n = 10, \theta_1 = 0.6, \theta_2 = 0.3, \theta_3 = 0.08, \theta_4 = 0.02)$ as a model for bit classification in this digital channel.]

4. $F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) \to 0$ as $x_1 \to -\infty$, $x_2 \to -\infty$, … and $x_m \to -\infty$

Definition 41 (Multivariate JPMF) If $(X_1, X_2, \ldots, X_m)$ is a **discrete random vector** that takes values in a discrete support set $\mathcal{S}_{X_1,X_2,\ldots,X_m}$, then its **joint probability mass function** (or JPMF) is:

$$f_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m) = \mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) \ . \tag{3.64}$$

Since $\mathbf{P}(\Omega) = 1$, $\sum_{(x_1,x_2,\ldots,x_m)\in\mathcal{S}_{X_1,X_2,\ldots,X_m}} f_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m) = 1$.

From JPMF $f_{X_1,X_2,\ldots,X_m}$ we can get the JCDF $F_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m)$ and the probability of any event B by simply taking sums as in Equation (3.59) but now over all m coordinates.

Definition 42 (Multivariate JPDF) $(X_1, X_2, \ldots, X_m)$ is a **continuous random vector** if its JDF $F_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m)$ is differentiable and the **joint probability density function (JPDF)** is given by:

$$f_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m) = \frac{\partial^m}{\partial x_1 \partial x_2 \cdots \partial x_m} F_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m) \ ,$$

From JPDF $f_{X_1,X_2,\ldots,X_m}$ we can compute the JDF $F_{X_1,X_2,\ldots,X_m}$ at any point $(x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ and more generally we can compute the probability of any event B, that can be cast as a region in $\mathbb{R}^m$, by "simply" taking m-dimensional integrals (you have done such iterated integrals when $m = 3$):

$$\boxed{F_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m) = \int_{-\infty}^{x_m} \cdots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m) dx_1 dx_2 \ldots dx_m} \ , \tag{3.65}$$

and

$$\boxed{\mathbf{P}(B) \ = \ \int \cdots \int_B f_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m) dx_1 dx_2 \ldots dx_m} \ . \tag{3.66}$$

The JPDF satisfies the following two properties:

1. integrates to 1, i.e., $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m) dx_1 dx_2 \ldots dx_m = 1$
2. is a non-negative function, i.e., $f_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m) \geq 0$.

The marginal PDF (marginal PMF) is obtained by integrating (summing) the JPDF (JPMF) over all other random variables. For example, the marginal PDF of X_1 is

$$f_{X_1}(x_1) = \int_{x_2=-\infty}^{\infty} \cdots \int_{x_m=-\infty}^{\infty} f_{X_1,X_2,\ldots,X_m}(x_1,x_2,\ldots,x_m) dx_2 \ldots dx_m$$

Definition 43 (Independence of Sequence of RVs) A finite or infinite sequence of RVs $X_1, X_2, \ldots$ is said to be independent or independently distributed if and only if

$$\mathbf{P}(X_{i_1} \leq x_{i_1}, X_{i_2} \leq x_{i_2}, \ldots, X_{i_k} \leq x_{i_k}) = \mathbf{P}(X_{i_1} \leq x_{i_1}) \mathbf{P}(X_{i_2} \leq x_{i_2}) \cdots, \mathbf{P}(X_{i_k} \leq x_{i_k})$$

or equivalently,

$$F_{X_{i_1},X_{i_2},\ldots,X_{i_m}}(x_{i_1},x_{i_2},\ldots,x_{i_m}) = F_{X_{i_1}}(x_{i_1}) F_{X_{i_2}}(x_{i_2}) \cdots F_{X_{i_m}}(x_{i_m}) \ ,$$

for any distinct subset of indices $\{i_1, i_2, \ldots, i_m\}$ of $\{1, 2, \ldots\}$, the index set of the sequence of RVs and any sequence of real numbers $x_{i_1}, x_{i_2}, \ldots, x_{i_m}$.

By the above definition, the sequence of **discrete** RVs $X_1, X_2, \ldots$ taking values in an at most countable set $\mathbb{D}$ are said to be independently distributed if for any distinct subset of indices $\{i_1, i_2, \ldots, i_k\}$ such that the corresponding RVs $X_{i_1}, X_{i_2}, \ldots, X_{i_k}$ exists as a distinct subset of our original sequence of RVs $X_1, X_2, \ldots$ and for any elements $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ in $\mathbb{D}$, the following equality is satisfied:

$$\mathbf{P}(X_{i_1} = x_{i_1}, X_{i_2} = x_{i_2}, \ldots, X_{i_k} = x_{i_k}) = \mathbf{P}(X_{i_1} = x_{i_1})\mathbf{P}(X_{i_2} = x_{i_2}) \cdots \mathbf{P}(X_{i_k} = x_{i_k}) \ ,$$

or equivalently,

$$f_{X_{i_1}, X_{i_2}, \ldots, X_{i_k}}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = f_{X_{i_1}}(x_{i_1}) f_{X_{i_2}}(x_{i_2}) \cdots f_{X_{i_k}}(x_{i_k}) \ .$$

From Definition 43, we say m random variables $X_1, X_2, \ldots, X_m$ are jointly independent or mutually independent if and only if for every $(x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$

$$F_{X_1, X_2, \ldots, X_m}(x_1, x_2, \ldots, x_m) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_m}(x_m) \ , \tag{3.67}$$

$$f_{X_1, X_2, \ldots, X_m}(x_1, x_2, \ldots, x_m) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_m}(x_m) \ . \tag{3.68}$$

Proposition 44 (Conditional probability of independent sequence of RVs) For an independent sequence of RVs $\{X_1, X_2, \ldots\}$, we have

$$\mathbf{P}(X_{i+1} \leq x_{i+1} | X_i \leq x_i, X_{i-1} \leq x_{i-1}, \ldots, X_1 \leq x_1) = \mathbf{P}(X_{i+1} \leq x_{i+1}) \tag{3.69}$$

Proof:

$$\begin{aligned}
& \mathbf{P}(X_{i+1} \leq x_{i+1} | X_i \leq x_i, X_{i-1} \leq x_{i-1}, \ldots, X_1 \leq x_1) \\
= \ & \frac{\mathbf{P}(X_{i+1} \leq x_{i+1}, X_i \leq x_i, X_{i-1} \leq x_{i-1}, \ldots, X_1 \leq x_1)}{\mathbf{P}(X_i \leq x_i, X_{i-1} \leq x_{i-1}, \ldots, X_1 \leq x_1)} \\
= \ & \frac{\mathbf{P}(X_{i+1} \leq x_{i+1}) \mathbf{P}(X_i \leq x_i) \mathbf{P}(X_{i-1} \leq x_{i-1}) \cdots \mathbf{P}(X_1 \leq x_1)}{\mathbf{P}(X_i \leq x_i) \mathbf{P}(X_{i-1} \leq x_{i-1}) \cdots \mathbf{P}(X_1 \leq x_1)} \\
= \ & \mathbf{P}(X_{i+1} \leq x_{i+1})
\end{aligned}$$

Equation (3.69) simply says that the conditional distribution of the RV X_{i+1} given all previous RVs $X_i, X_{i-1}, \ldots, X_1$ is simply determined by the distribution of X_{i+1}.

Example 95 If X_1 and X_2 are independent random variables then what is their covariance $\mathbf{Cov}(X_1, X_2)$?

Solution:

We know for independent RVs from the properties of expectations that

$$E(X_1 X_2) = E(X_1) E(X_2)$$

From the formula for covariance

$$\begin{aligned}
\mathbf{Cov}(X_1, X_2) &= E(X_1 X_2) - E(X_1) E(X_2) \\
&= E(X_1) E(X_2) - E(X_1) E(X_2) \qquad \text{due to independence} \\
&= 0
\end{aligned}$$

Remark 45 The converse is not true: two random variables that have zero covariance are not necessarily independent.

7. $E(X)$, the expectation of X or the first moment of X
8. $E(Y)$, the expectation of Y or the first moment of Y
9. $E(XY)$, the expectation of XY
10. $\mathbf{Cov}(X, Y) = E(XY) - E(X)E(Y)$, the covariance of X and Y.

Ex. 3.40 — Logs are milled to have a width of μ. The actual width of a randomly selected item is X. If X is a Normal(μ, σ^2) random variable then find the probability density function of the *squared-error* of the milling process,

$$Y = (X - \mu)^2 \, .$$

Ex. 3.41 — Let (X, Y) be a discrete random vector (R$\vec{\text{V}}$) with support:

$$\mathcal{S}_{X,Y} = \{(0,0), (0,1), (1,0), (1,1)\} \ .$$

Let its joint probability mass function (JPMF) be:

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{4} & \text{if } (x,y) = (0,0) \\ \frac{1}{4} & \text{if } (x,y) = (0,1) \\ \frac{1}{4} & \text{if } (x,y) = (1,0) \\ \frac{1}{4} & \text{if } (x,y) = (1,1) \\ 0 & \text{otherwise} \ . \end{cases}$$

Are X and Y independent?

Ex. 3.42 — A semiconductor product consists of three layers that are fabricated independently. If the variances in thickness of the first, second and third third layers are 25, 40 and 30 nanometers squared, what is the variance of the thickness of the final product?

Ex. 3.43 — Find the covariance for the discrete R$\vec{\text{V}}$ (X, Y) with joint probability mass function

$$f_{X,Y}(x,y) = \begin{cases} 0.2 & \text{if } (x,y) = (0,0) \\ 0.1 & \text{if } (x,y) = (1,1) \\ 0.1 & \text{if } (x,y) = (1,2) \\ 0.1 & \text{if } (x,y) = (2,1) \\ 0.1 & \text{if } (x,y) = (2,2) \\ 0.4 & \text{if } (x,y) = (3,3) \\ 0 & \text{otherwise} \ . \end{cases}$$

[Hint: Recall that $\mathbf{Cov}(X, Y) = E(XY) - E(X)E(Y)$]

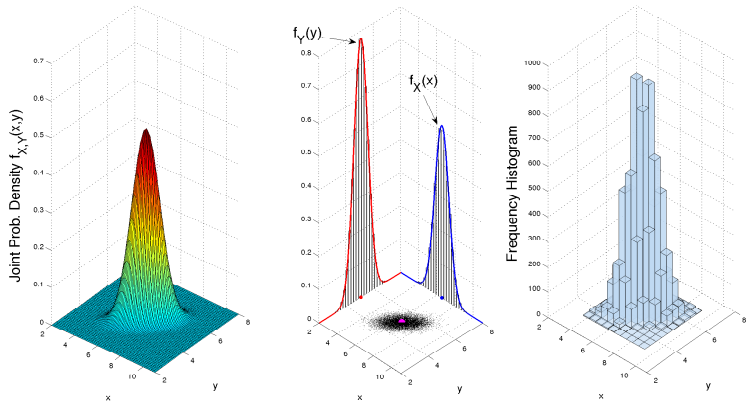
Ex. 3.44 — Consider two random variables (RVs) X and Y having marginal distribution functions

$$F_X(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{1}{2} & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1 - \frac{1}{2}e^{-y} - \frac{1}{2}e^{-2y} & \text{if } y \geq 0 \end{cases}$$

If X and Y are independent, what is their joint distribution function $F_{X,Y}(x,y)$? [Hint: you need to express $F_{X,Y}(x,y)$ for any $(x,y) \in \mathbb{R}^2$.]

Figure 3.23: JPDF, Marginal PDFs and Frequency Histogram of a Bivariate Normal R$\vec{\text{V}}$ for lengths of girths of cylindrical shafts in a manufacturing process (in cm).



3.11 Exercises in Multivariate Random Variables

Ex. 3.38 — Find the probability that none of the three bulbs in a traffic signal, that are assumed to have independent life-times (i.e., the time during which they are operational), need to be replaced during the first 1200 hours of operation if the length of time before a single bulb needs to be replaced is a continuous random variable X with density

$$f(x) = \begin{cases} 6\left(0.25 - (x-1.5)^2\right) & 1 < x < 2 \\ 0 & \text{otherwise} \end{cases} .$$

Note: X is measured in multiples of 1000 hours.

Ex. 3.39 — Let (X, Y) be a continuous R$\vec{\text{V}}$ with joint probability density function (JPDF)

$$f_{X,Y}(x,y) = \begin{cases} a(x^2+y) & \text{if } 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{otherwise} \end{cases} .$$

Find the following:

1. the normalizing constant a which will ensure $\mathbf{P}(\Omega) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$
2. $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$ called the marginal probability density function (MPDF) of X
3. $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$ called the marginal probability density function (MPDF) of Y
4. Check if $f_X(x)f_Y(y) = f_{X,Y}(x,y)$ for every (x,y) and decide whether X and Y are independent random variables. Hint: X and Y are said to be independent if $f_X(x)f_Y(y) = f_{X,Y}(x,y)$ for every (x,y).
5. $F_{X,Y}(x,y)$, the joint cumulative distribution function (JCDF) of (X,Y) for any $(x,y) \in (0,1)\times(0,1)$
6. the probability that $X > 0.5$ and $Y < 0.6$, i.e., $\mathbf{P}(X > 0.5, Y < 0.6)$

Linear Combination of Independent Normal RVs is a Normal RV

We can get the following special property of normal RVs using Eqn. (3.74). If $X_1, X_2, \ldots, X_m$ be jointly independent RVs, where X_i is Normal(μ_i, σ_i^2), for $i = 1, 2, \ldots, m$ then $Y = c + \sum_{i=1}^m a_i X_i$ for some constants $c, a_1, a_2, \ldots, a_m$ is the Normal $\left(c + \sum_{i=1}^m a_i\mu_i, \sum_{i=1}^m a_i^2\sigma_i^2\right)$ RV.

Example 96 Let X be Normal$(2,4)$, Y be Normal$(-1,2)$ and Z be Normal$(0,1)$ RVs that are jointly independent. Obtain the following:

1. $E(3X - 2Y + 4Z)$
2. $V(2Y - 3Z)$
3. the distribution of $6 - 2Z + X - Y$
4. the probability that $6 - 2Z + X - Y > 0$
5. **Cov**(X, W), where $W = X - Y$.

Solution

1. $$E(3X - 2Y + 4Z) = 3E(X) - 2E(Y) + 4(Z) = (3\times 2) + (-2\times(-1)) + 4\times 0 = 6 + 2 + 0 = 8$$

2. $$V(2Y - 3Z) = 2^2V(Y) + (-3)^2V(Z) = (4\times 2) + (9\times 1) = 8 + 9 = 17$$

3. From the special property of normal RVs, the distribution of $6 - 2Z + X - Y$ is
$$\begin{aligned} &\text{Normal}\left(6 + (-2\times 0) + (1\times 2) + (-1\times -1), ((-2)^2\times 1) + (1^2\times 4) + ((-1)^2\times 2)\right) \\ &= \text{Normal}\left(6 + 0 + 2 + 1, 4 + 4 + 2\right) \\ &= \text{Normal}(9, 10) \end{aligned}$$

4. Let $U = 6 - 2Z + X - Y$ and we know U is Normal$(9,10)$ RV.
$$\begin{aligned} P(6 - 2Z + X - Y > 0) = P(U > 0) = P(U - 9 > 0 - 9) &= P\left(\frac{U-9}{\sqrt{10}} > \frac{-9}{\sqrt{10}}\right) \\ &= P\left(Z > \frac{-9}{\sqrt{10}}\right) \\ &= P\left(Z < \frac{9}{\sqrt{10}}\right) \\ &\approxeq P(Z < 2.85) = 0.9978 \end{aligned}$$

5. $$\begin{aligned} \textbf{Cov}(X,W) &= E(XW) - E(X)E(W) = E(X(X-Y)) - E(X)E(X-Y) \\ &= E(X^2 - XY) - E(X)(E(X) - E(Y)) = E(X^2) - E(XY) - 2\times(2-(-1)) \\ &= E(X^2) - E(X)E(Y) - 6 = E(X^2) - (2\times(-1)) - 6 \\ &= (V(X) + (E(X))^2) + 2 - 6 = (4 + 2^2) - 4 = 4 \end{aligned}$$

3.10.4 Some Common $\mathbb{R}^m$-valued RVs

So far, we have treated our random vectors as random points in $\mathbb{R}^m$ and not been explicit about whether they are row or column vectors. We need to be more explicit now in order to perform arithmetic operations and transformations with them.

Let $X = (X_1, X_2, \ldots, X_{m_X})$ be a R$\vec{\text{V}}$ in $\mathbb{R}^{1\times m_X}$, i.e., X is a random row vector with 1 row and m_X columns, with JCDF $F_{X_1,X_2,\ldots,X_{m_X}}$ and JPDF $f_{X_1,X_2,\ldots,X_{m_X}}$. Similarly, let $Y = (Y_1, Y_2, \ldots, Y_{m_Y})$ be a R$\vec{\text{V}}$ in $\mathbb{R}^{1\times m_Y}$, i.e., Y is a random row vector with 1 row and m_Y columns, with JCDF $F_{Y_1,Y_2,\ldots,Y_{m_Y}}$ and JPDF $f_{Y_1,Y_2,\ldots,Y_{m_Y}}$. Let the JCDF of the random vectors X and Y together be $F_{X_1,X_2,\ldots,X_{m_X},Y_1,Y_2,\ldots,Y_{m_Y}}$ and JPDF be $f_{X_1,X_2,\ldots,X_{m_X},Y_1,Y_2,\ldots,Y_{m_Y}}$.

Independent Random Vectors and their sums

The notion of mutual independence or joint independence of n random vectors is obtained similarly from ensuring the independence of any subset of the n vectors in terms of their JCDFs (JPMFs or JPDFs) being equal to the product of their marginal CDFs (PMFs or PDFs).

Thus, for a given $m_X < \infty$ and $m_Y < \infty$, two **random vectors are independent** if and only if for any $(x_1, x_2, \ldots, x_{m_X}) \in \mathbb{R}^{1\times m_X}$ and any $(y_1, y_2, \ldots, y_{m_Y}) \in \mathbb{R}^{1\times m_Y}$

$$F_{X_1,X_2,\ldots,X_{m_X},Y_1,Y_2,\ldots,Y_{m_Y}}(x_1, x_2, \ldots, x_{m_X}, y_1, y_2, \ldots, y_{m_Y}) = F_{X_1,X_2,\ldots,X_{m_X}}(x_1, x_2, \ldots, x_{m_X}) \times F_{Y_1,Y_2,\ldots,Y_{m_Y}}(y_1, y_2, \ldots, y_{m_Y})$$

or, equivalently

$$f_{X_1,X_2,\ldots,X_{m_X},Y_1,Y_2,\ldots,Y_{m_Y}}(x_1, x_2, \ldots, x_{m_X}, y_1, y_2, \ldots, y_{m_Y}) = f_{X_1,X_2,\ldots,X_{m_X}}(x_1, x_2, \ldots, x_{m_X}) \times f_{Y_1,Y_2,\ldots,Y_{m_Y}}(y_1, y_2, \ldots, y_{m_Y})$$

Let us consider the natural two-dimensional analogue of the Bernoulli(θ) RV in the real plane $\mathbb{R}^2 := (-\infty, \infty)^2 := (-\infty, \infty) \times (-\infty, \infty)$. A natural possibility is to use the **ortho-normal basis vectors** in $\mathbb{R}^2$:

$$\boxed{e_1 := (1, 0), \qquad e_2 := (0, 1)}\,.$$

Recall that vector addition and subtraction are done component-wise, i.e. $(x_1, x_2) \pm (y_1, y_2) = (x_1 \pm y_1, x_2 \pm y_2)$. We introduce a useful function called the indicator function of a set, say A.

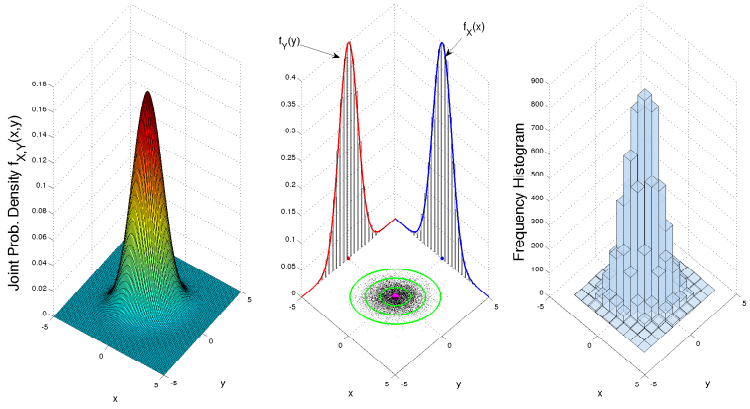
$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbb{1}_A(x)$ returns 1 if x belongs to A and 0 otherwise.

Example 97 Let us recall the geometry and arithmetic of vector addition in the plane.

1. What is $(1, 0) + (1, 0)$, $(1, 0) + (0, 1)$, $(0, 1) + (0, 1)$?
2. What is the relationship between $(1, 0)$, $(0, 1)$ and $(1, 1)$ geometrically?
3. How does the diagonal of the parallelogram relate the its two sides in the geometry of addition in the plane?
4. What is $(1, 0) + (0, 1) + (1, 0)$?

Figure 3.22: JPDF, Marginal PDFs and Frequency Histogram of Bivariate Standard Normal R$\vec{\text{V}}$.



When we have a non-zero mean vector

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} = \begin{pmatrix} 6.49 \\ 5.07 \end{pmatrix}$$

for the mean lengths and girths of cylindrical shafts from a manufacturing process with variance-covariance matrix

$$\Sigma = \begin{pmatrix} \mathbf{Cov}(X,X) & \mathbf{Cov}(X,Y) \\ \mathbf{Cov}(Y,X) & \mathbf{Cov}(Y,Y) \end{pmatrix} = \begin{pmatrix} V(X) & \mathbf{Cov}(X,Y) \\ \mathbf{Cov}(X,Y) & V(Y) \end{pmatrix} = \begin{pmatrix} 0.59 & 0.24 \\ 0.24 & 0.26 \end{pmatrix}$$

then the Normal(μ, Σ) R$\vec{\text{V}}$ has JPDF, marginal PDFs and samples with frequency histograms as shown in Figure 3.23.

We can use MATLAB to compute for instance the probability that a cylinder has length and girth below 6.0 cms as follows:

```
>> mvncdf([6.0 6.0],[6.49 5.07],[0.59  0.24; 0.24 0.26])
ans =    0.2615
```

Or find the probability (with numerical error tolerance) that the cylinders are within the rectangular specifications of 6 ± 1.0 along x and y as follows:

```
>> [F err] = mvncdf([5.0 5.0], [7.0 7.0], [6.49 5.07],[0.59  0.24; 0.24 0.26])

F =    0.3352

err =   1.0000e-08
```

3.10.5 Dependent Random Variables

When a sequence of RVs are not independent they are said to be **dependent**. The simplest form of dependence is *Markov dependence* that we will briefly see via a couple examples in Chapter 6.

where, the multinomial coefficient:

$$\binom{n}{y_1, y_2, \ldots, y_k} := \frac{n!}{y_1! y_2! \cdots y_k!} .$$

Note that the marginal PMF of Y_j is Binomial(n, θ_j) for any $j = 1, 2, \ldots, k$.

We can visualize the Multinomial$(n, \theta_1, \theta_2, \theta_3)$ process as a sum of n IID de Moivre$(\theta_1, \theta_2, \theta_3)$ R$\vec{\text{V}}$s via a three dimensional extension of the Quincunx called the "Septcunx" and relate the number of paths that lead to a given trivariate sum (y_1, y_2, y_3) with $\sum_{i=1}^{3} y_i = n$ as the multinomial coefficient $\frac{n!}{y_1! y_2! y_3!}$. In the Septcunx, balls choose from one of three paths along e_1, e_2 and e_3 with probabilities θ_1, θ_2 and θ_3, respectively, in an IID manner at each of the n levels, before they collect at buckets placed at the integral points in the 3-simplex, $\mathbb{Y} = \{(y_1, y_2, y_3) \in \mathbb{Z}_+^3 : \sum_{i=1}^{3} y_i = n\}$. Once again, we can visualize that the sum of n IID de Moivre$(\theta_1, \theta_2, \theta_3)$ R$\vec{\text{V}}$s constitute the Multinomial$(n, \theta_1, \theta_2, \theta_3)$ R$\vec{\text{V}}$.

Labwork 100 (Septcunx Sampler Demo – Sum of n IID de Moivre$(1/3, 1/3, 1/3)$ **R$\vec{\text{V}}$s)** Let us understand the Septcunx construction of the Multinomial$(n, 1/3, 1/3, 1/3)$ R$\vec{\text{V}}X$ as the sum of n independent and identical de Moivre$(1/3, 1/3, 13/)$ R$\vec{\text{V}}$s by calling the interactive visual cognitive tool as follows:

```
>> guiMultinomial
```

Multinomial distributions are at the very foundations of various machine learning algorithms, including, filtering junk email, learning from large knowledge-based resources like www, Wikipedia, word-net, etc.

Model 18 (Normal(μ, Σ) **R$\vec{\text{V}}$)** The univariate Normal(μ, σ^2) RV has two parameters, $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$. In the multivariate version $\mu \in \mathbb{R}^{m \times 1}$ is a column vector and σ^2 is replaced by a matrix Σ. To begin, let

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_m \end{pmatrix}$$

where, $Z_1, Z_2, \ldots, Z_m$ are jointly independent Normal$(0, 1)$ RVs. Then the JPDF of Z is

$$f_Z(z) = f_{Z_1, Z_2, \ldots, Z_m}(z_1, z_2, \ldots, z_m) = \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{m} z_j^2\right) = \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2} z^T z\right)$$

We say that Z has a standard multivariate normal distribution and write $Z \sim \text{Normal}(0, I)$, where it is understood that 0 represents the vector of m zeros and I is the $m \times m$ identity matrix (with 1 along the diagonal entries and 0 on all off-diagonal entries).

More generally, a vector X has a multivariate normal distribution denoted by $X \sim \text{Normal}(\mu, \Sigma)$, if it has joint probability density function

$$f_X(x; \mu, \Sigma) = f_{X_1, X_2, \ldots, X_m}(x_1, x_2, \ldots, x_m; \mu, \Sigma) = \frac{1}{(2\pi)^{m/2} |(\Sigma)|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

where $|\Sigma|$ denotes the determinant of Σ, μ is a vector of length m and Σ is a $m \times m$ symmetric, positive definite matrix. Setting $\mu = 0$ and $\Sigma = I$ gives back the standard multivariate normal R$\vec{\text{V}}$.

Solution:

1. addition is component-wise

$$(1, 0) + (1, 0) = (1 + 1, 0 + 0) = (2, 0)$$
$$(1, 0) + (0, 1) = (1 + 0, 0 + 1) = (1, 1)$$
$$(0, 1) + (0, 1) = (0 + 0, 1 + 1) = (0, 2)$$

2. $(1, 0)$ and $(0, 1)$ are vectors for the two sides of unit square and $(1, 1)$ is its diagonal.

3. Generally, the diagonal of the parallelogram is the resultant or sum of the vectors representing its two sides

4.
$$(1, 0) + (0, 1) + (1, 0) = (1 + 0 + 1, 0 + 1 + 0) = (2, 1)$$

Model 15 (Bernoulli(θ) **R$\vec{\text{V}}$)** Given a parameter $(\theta, 1 - \theta) \in \Delta^1$, the unit 1-Simplex, we say that $X := (X_1, X_2)$ is a Bernoulli(θ) random vector (R$\vec{\text{V}}$) if it has only two possible outcomes in the set $\{e_1, e_2\} \subset \mathbb{R}^2$, i.e. $x := (x_1, x_2) \in \{(1, 0), (0, 1)\}$. The PMF of the R$\vec{\text{V}}$ $X := (X_1, X_2)$ with realization $x := (x_1, x_2)$ is:

$$f(x; \theta) := P(X = x) = \theta \mathbb{1}_{\{e_1\}}(x) + (1 - \theta)\mathbb{1}_{\{e_2\}}(x) = \begin{cases} \theta & \text{if} \quad x = e_1 := (1, 0) \\ 1 - \theta & \text{if} \quad x = e_2 := (0, 1) \\ 0 & \text{otherwise} \end{cases}$$

Example 98 Let us find the Expectation of Bernoulli(θ) R$\vec{\text{V}}$in Model 15.

$$\mathbf{E}_\theta(X) = \mathbf{E}_\theta((X_1, X_2)) = \sum_{(x_1, x_2) \in \{e_1, e_2\}} (x_1, x_2) f((x_1, x_2); \theta) = (1, 0)\theta + (0, 1)(1 - \theta) = (\theta, 1 - \theta) .$$

Remark 46 We can write the Binomial(n, θ) RV Y as a Binomial(n, θ) R$\vec{\text{V}}$ $X := (Y, n - Y)$. In fact, this is the underlying model and the **bi** in the Binomial(n, θ) does refer to two in Latin. In the coin-tossing context this can be thought of keeping track of the number of Heads and Tails out of an IID sequence of n tosses of a coin with probability θ of observing Heads. In the Quincunx context, this amounts to keeping track of the number of right and left turns made by the ball as it

drops through n levels of pegs where the probability of a right turn at each peg is independently and identically θ. In other words, the Binomial(n, θ) R$\vec{\text{V}}$ $(Y, n-Y)$ is the sum of n IID Bernoulli(θ) R$\vec{\text{V}}$s $X_1 := (X_{1,1}, X_{1,2}), X_2 := (X_{2,1}, X_{2,2}), \ldots, X_n := (X_{n,1}, X_{n,2})$:

$$(Y, n-Y) = X_1 + X_2 + \cdots + X_n = (X_{1,1}, X_{1,2}) + (X_{2,1}, X_{2,2}) + \cdots + (X_{n,1}, X_{n,2})$$

Exercise 3.36 (Random walk in the first Quadrant) Consider an independent and identical random walk starting from $(0, 0)$ in the first quadrant where you go east, i.e., add $(1, 0)$ to your current position with probability θ, and go north, i.e., add $(0, 1)$ to your current position with probability $1 - \theta$. Suppose you take n such IID steps according to the Bernoulli(θ) R$\vec{\text{V}}$. Answer the following questions:

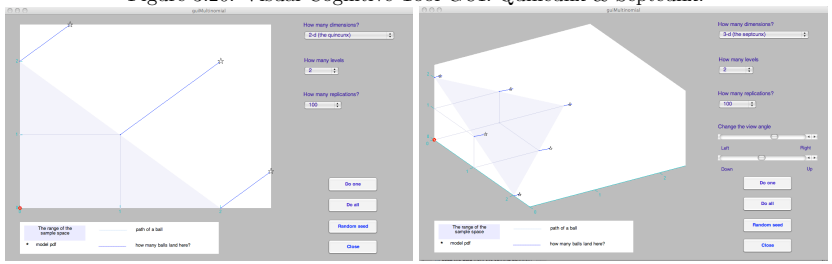
1. How does the number of paths that lead to a (x_1, x_2) with $x_1 + x_2 = n$ relate to the binomial coefficient $\binom{n}{x_1}$?
2. What is the probability of taking x_1 steps east and x_2 steps north?

Exercise 3.37 (Random walks in the first Quadrant and Galton's Quincunx) Compare the probability models for the Random walk in the first quadrant and Galton's Quincunx and explain how they are related.

Labwork 99 (Quincunx Sampler Demo – Sum of n IID Bernoulli(1/2) R$\vec{\text{V}}$s) Let us understand the Quincunx construction of the Binomial$(n, 1/2)$ R$\vec{\text{V}}X$ as the sum of n independent and identical Bernoulli$(1/2)$ R$\vec{\text{V}}$s by calling the interactive visual cognitive tool as follows:

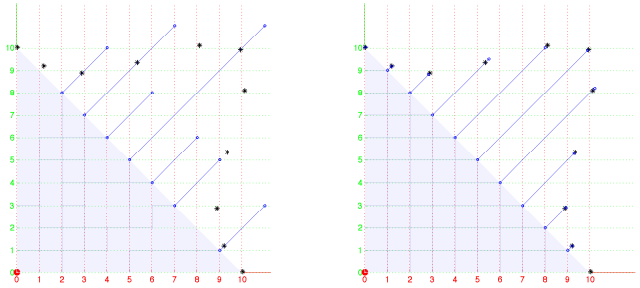
```
>> guiMultinomial
```

Figure 3.20: Visual Cognitive Tool GUI: Quincunx & Septcunx.



We are now ready to extend the Binomial(n, θ) RV or R$\vec{\text{V}}$ to its multivariate version called the Multinomial$(n, \theta_1, \theta_2, \ldots, \theta_k)$ R$\vec{\text{V}}$. We develop this R$\vec{\text{V}}$ as the sum of n IID de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ R$\vec{\text{V}}$ that is defined next by extending de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ RV taking values in $\{1, 2, \ldots, k\}$ of Model 14 to its vector-valued cousin taking values in $\{e_1, e_2, \ldots, e_k\}$, the ortho-normal basis vectors in $\mathbb{R}^k$.

Figure 3.21: Quincunx on the Cartesian plane. Simulations of Binomial$(n = 10, \theta = 0.5)$ RV as the x-coordinate of the ordered pair resulting from the culmination of sample trajectories formed by the accumulating sum of $n = 10$ IID Bernoulli$(\theta = 0.5)$ random vectors over $\{(1, 0), (0, 1)\}$ with probabilities $\{\theta, 1 - \theta\}$, respectively. The blue lines and black asterisks perpendicular to and above the diagonal line, i.e. the line connecting $(0, 10)$ and $(10, 0)$, are the density histogram of the samples and the PMF of our Binomial$(n = 10, \theta = 0.5)$ RV, respectively.



(a) Ten samples (b) Thousand samples

Model 16 (de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ **R$\vec{\text{V}}$)** The PMF of the de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ R$\vec{\text{V}}$ $X := (X_1, X_2, \ldots, X_k)$ taking value $x := (x_1, x_2, \ldots, x_k) \in \{e_1, e_2, \ldots, e_k\}$, where the e_i's are ortho-normal basis vectors in $\mathbb{R}^k$ is:

$$f(x; \theta_1, \theta_2, \ldots, \theta_k) := P(X = x) = \sum_{i=1}^{k} \theta_i \mathbb{1}_{\{e_i\}}(x) = \begin{cases} \theta_1 & \text{if} \quad x = e_1 := (1, 0, \ldots, 0) \in \mathbb{R}^k \\ \theta_2 & \text{if} \quad x = e_2 := (0, 1, \ldots, 0) \in \mathbb{R}^k \\ \vdots & \\ \theta_k & \text{if} \quad x = e_k := (0, 0, \ldots, 1) \in \mathbb{R}^k \\ 0 & \text{otherwise} \end{cases}$$

Of course, $\sum_{i=1}^{k} \theta_i = 1$.

When we add n IID de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ R$\vec{\text{V}}$ together, we get the Multinomial$(n, \theta_1, \theta_2, \ldots, \theta_k)$ R$\vec{\text{V}}$ as defined below.

Model 17 (Multinomial$(n, \theta_1, \theta_2, \ldots, \theta_k)$ **R$\vec{\text{V}}$)** We say that a R$\vec{\text{V}}$ $Y := (Y_1, Y_2, \ldots, Y_k)$ obtained from the sum of n IID de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ R$\vec{\text{V}}$s with realizations

$$y := (y_1, y_2, \ldots, y_k) \in \mathbb{Y} := \{(y_1, y_2, \ldots, y_k) \in \mathbb{Z}_+^k : \sum_{i=1}^{k} y_i = n\}$$

has the PMF given by:

$$f(y; n, \theta) := f(y; n, \theta_1, \theta_2, \ldots, \theta_k) := P(Y = y; n, \theta_1, \theta_2, \ldots, \theta_k) = \binom{n}{y_1, y_2, \ldots, y_k} \prod_{i=1}^{k} \theta_i^{y_i} ,$$