

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: September 4, 2019

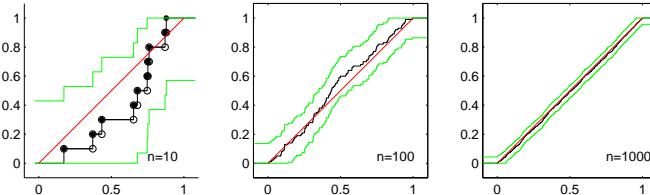
©2007–2019 Raazesh Sainudiin. ©2008–2013 Dominic Lee.

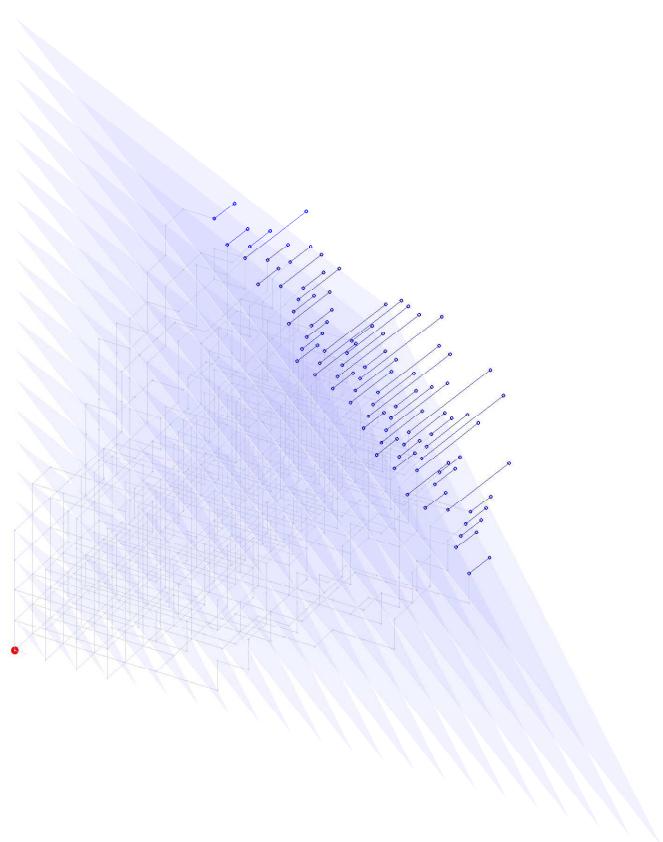
This work is licensed under the Creative Commons Attribution-Noncommercial-Share Alike 4.0

International License. To view a copy of this license, visit

<https://creativecommons.org/licenses/by-nc-sa/4.0/>.

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





Contents

1 Preliminaries	10
1.1 Elementary Set Theory	10
1.2 Natural Numbers, Integers and Rational Numbers	13
1.3 Real Numbers	17
1.4 Introduction to MATLAB	21
1.5 Permutations, Factorials and Combinations	24
1.6 Array, Sequence, Limit,	26
1.7 Elementary Number Theory	32
2 Probability Model	34
2.1 Experiments	34
2.2 Probability	36
2.2.1 Consequences of our Definition of Probability	38
2.2.2 Sigma Algebras of Typical Experiments*	40
2.3 Conditional Probability	42
2.3.1 Independence and Dependence	45
3 Random Variables	50
3.1 Basic Definitions	50
3.2 An Elementary Discrete Random Variable	52
3.3 An Elementary Continuous Random Variable	53
3.4 Expectations	56
3.5 Stochastic Processes	59
4 Random Numbers	61
4.1 Physical Random Number Generators	61
4.2 Pseudo-Random Number Generators	61
4.2.1 Linear Congruential Generators	62
4.2.2 Generalized Feedback Shift Register and the “Mersenne Twister” PRNG . . .	65

CONTENTS	4
5 Statistics	68
5.1 Data and Statistics	68
5.2 Exploring Data and Statistics	75
5.2.1 Univariate Data	75
5.2.2 Bivariate Data	77
5.2.3 Trivariate Data	78
5.2.4 Multivariate Data	78
5.3 Loading and Exploring Real-world Data	80
5.3.1 Geological Data	80
5.3.2 Metereological Data	84
5.3.3 Textual Data	87
5.3.4 Machine Sensor Data	88
6 Common Random Variables	89
6.1 Inversion Sampler for Continuous Random Variables	89
6.2 Some Simulations of Continuous Random Variables	90
6.3 Continuous Random Variables	90
6.4 Discrete Random Variables	104
6.5 Inversion Sampler for Discrete Random Variables	104
6.6 Some Simulations of Discrete Random Variables	104
6.7 Sir Francis Galton's Quincunx	120
6.8 Random Vectors	125
6.9 von Neumann Rejection Sampler (RS)	132
6.10 Importance Resampler	137
6.11 Other Continuous Random Variables	138
6.12 Other Random Vectors	139
6.13 Problems	140
7 Statistical Experiments	142
7.1 Introduction	142
7.2 Some Common Experiments	142
7.3 Typical Decision Problems with Experiments	144
8 Limits of Random Variables	145
8.1 Convergence of Random Variables	145
8.2 Some Basic Limit Laws of Statistics	148

<i>CONTENTS</i>	5
9 Finite Markov Chains	151
9.1 Introduction	151
9.2 Random Mapping Representation and Simulation	159
9.3 Irreducibility and Aperiodicity	165
9.4 Stationarity	168
9.5 Reversibility	170
9.6 Metropolis-Hastings Markov chain	174
9.7 Glauber Dynamics	177
9.7.1 Random Walks on \mathbb{Z} and the reflection principle	182
9.8 Coupling from the past	182
9.8.1 <i>Algorithm – Coupling from the past.</i>	183
10 Answers to Selected Exercises	186

List of Tables

1.1	Symbol Table: Sets and Numbers	20
6.1	Some continuous RVs that can be simulated from using Algorithm 3.	99
6.2	Random Variables with PDF, Mean and Variance	121

List of Figures

1.1	Union and intersection of sets shown by Venn diagrams	11
1.2	These Venn diagram illustrate De Morgan’s Laws.	12
1.3	A function f (“father of”) from \mathbb{X} (a set of children) to \mathbb{Y} (their fathers) and its inverse (“children of”).	15
1.4	A pictorial depiction of addition and its inverse. The domain is plotted in orthogonal Cartesian coordinates	16
1.5	A depiction of the real line segment $[-10, 10]$	18
1.6	Point plot and stem plot of the finite sequence $\langle b_{1:10} \rangle$ declared as an array.	28
1.7	A plot of the sine wave over $[-2\pi, 2\pi]$	30
2.1	A binary tree whose leaves are all possible outcomes.	36
2.2	First ball number in 1114 NZ Lotto draws from 1987 to 2008.	39
3.1	The Indicator function of event $A \in \mathcal{F}$ is a RV $\mathbf{1}_A$ with DF F	51
3.2	A RV X from a sample space Ω with 8 elements to \mathbb{R} and its DF F	52
3.3	The Indicator Function $\mathbf{1}_{\mathbb{H}}$ of the event ‘Heads occurs’, for the experiment ‘Toss 1 times,’ \mathcal{E}_{θ}^1 , as the RV X from the sample space $\Omega = \{\mathbb{H}, \mathbb{T}\}$ to \mathbb{R} and its DF F . The probability that ‘Heads occurs’ and that ‘Tails occurs’ are $f(1; \theta) = \mathbf{P}_{\theta}(X = 1) = \mathbf{P}_{\theta}(\mathbb{H}) = \theta$ and $f(0; \theta) = \mathbf{P}_{\theta}(X = 0) = \mathbf{P}_{\theta}(\mathbb{T}) = 1 - \theta$, respectively.	54
3.4	A plot of the PDF and DF or CDF of the Uniform(0, 1) continuous RV X	55
3.5	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(θ) RV X as a function of the parameter θ	58
4.1	The linear congruential sequence of <code>LinConGen(256, 137, 0, 123, 257)</code> with non-maximal period length of 32 as a line plot over $\{0, 1, \dots, 256\}$, scaled over $[0, 1]$ by a division by 256 and a histogram of the 256 points in $[0, 1]$ with 15 bins.	63
4.2	The LCG called <code>RANDU</code> 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.	65
4.3	Triplet point clouds from the “Mersenne Twister” with two different seeds (see Lab-work 42).	67
5.1	Sample Space, Random Variable, Realisation, Data, and Data Space.	68
5.2	Data Spaces $\mathbb{X} = \{0, 1\}^2$ and $\mathbb{X} = \{0, 1\}^3$ for two and three Bernoulli trials, respectively.	69

5.3	Plot of the DF of Uniform(0, 1), five IID samples from it, and the ECDF \hat{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.	74
5.4	Frequency, Relative Frequency and Density Histograms	76
5.5	Frequency, Relative Frequency and Density Histograms	77
5.6	2D Scatter Plot	78
5.7	3D Scatter Plot	79
5.8	Plot Matrix of uniformly generated data in $[0, 1]^5$	79
5.9	Matrix of Scatter Plots of the latitude, longitude, magnitude and depth of the 22-02-2011 earth quakes in Christchurch, New Zealand.	82
5.10	Google Earth Visualisation of the earth quakes	83
5.11	Daily rainfalls in Christchurch since March 27 2010	85
5.12	Daily temperatures in Christchurch for one year since March 27 2010	86
5.13	Wordle of JOE 2010	87
5.14	Double Pendulum	88
6.1	A plot of the PDF, DF or CDF and inverse DF of the Uniform($-1, 1$) RV X	90
6.2	Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Uniform}(-5, 5)$	91
6.3	Density and distribution functions of $\text{Exponential}(\lambda)$ RVs, for $\lambda = 1, 10, 10^{-1}$, in four different axes scales.	92
6.4	Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Exponential}(0.5)$	93
6.5	The PDF f , DF F , and inverse DF $F^{[-1]}$ of the the Exponential($\lambda = 1.0$) RV.	94
6.6	Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Laplace}(5)$	96
6.7	Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Cauchy}$	98
6.8	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).	99
6.9	Density and distribution function of several $\text{Normal}(\mu, \sigma^2)$ RVs.	100
6.10	PDF and CDF of $X \sim \text{Gamma}(\beta = 0.1, \alpha)$ with $\alpha \in \{1, 2, 3, 4, 5\}$	104
6.11	The DF $F(x; 0.3, 0.7)$ of the de Moivre($0.3, 0.7$) RV and its inverse $F^{[-1]}(u; 0.3, 0.7)$	107
6.12	Mean and variance of a $\text{Geometric}(\theta)$ RV X as a function of the parameter θ	111
6.13	PDF of $X \sim \text{Geometric}(\theta = 0.5)$ and the relative frequency histogram based on 100 and 1000 samples from X	112
6.14	PDF of $X \sim \text{Binomial}(n = 10, \theta = 0.5)$ and the relative frequency histogram based on 100,000 samples from X	115
6.15	PDF of $X \sim \text{Poisson}(\lambda = 10)$ and the relative frequency histogram based on 1000 samples from X	118
6.16	Figures from Sir Francis Galton, F.R.S., <i>Natural Inheritance</i> , , Macmillan, 1889.	120

6.17 Quincunx on the Cartesian plane. Simulations of $\text{Binomial}(n = 10, \theta = 0.5)$ 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 $\text{Bernoulli}(\theta = 0.5)$ 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 PDF of our $\text{Binomial}(n = 10, \theta = 0.5)$ RV, respectively.	127
6.18 Visual Cognitive Tool GUI: Quincunx.	129
6.19 Septcunx on the Cartesian co-ordinates. Simulations of $\text{Multinomial}(n = 2, \theta_1 = 1/3, \theta_2 = 1/3, \theta_3 = 1/3)$ $\vec{R}\vec{V}$ as the sum of n IID $\text{de Moivre}(\theta_1 = 1/3, \theta_2 = 1/3, \theta_3 = 1/3)$ $\vec{R}\vec{V}$ s over $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ with probabilities $\{\theta_1, \theta_2, \theta_3\}$, respectively. The blue lines perpendicular to the sample space of the $\text{Multinomial}(3, \theta_1, \theta_2, \theta_3)$ $\vec{R}\vec{V}$, i.e. the plane in \mathbb{R}^3 connecting $(n, 0, 0)$, $(0, n, 0)$ and $(0, 0, n)$, are the density histogram of the samples.	130
6.20 Visual Cognitive Tool GUI: Septcunx.	131
6.21 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	133
6.22 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	134
7.1 Geometry of the Θ 's for de Moivre[k] Experiments with $k \in \{1, 2, 3, 4\}$	143
8.1 Distribution functions of several $\text{Normal}(\mu, \sigma^2)$ RVs for $\sigma^2 = 1, \frac{1}{10}, \frac{1}{100}, \frac{1}{1000}$	145
9.1 Transition Diagram of Flippant Freddy's Jumps.	152
9.2 The probability of being back in rollovia 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).	154
9.3 Transition Diagram of Dry and Wet Days in Christchurch.	156
9.4 Transition diagram over six lounges (without edge probabilities).	164
9.5 Stochastic Optimization with Metropolis chain.	176
9.6 The sample at time step 10^6 from the Glauber dynamics for the hardcore model on 100×100 regular torus grid. A red site is occupied while a blue site is vacant.	180

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:

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

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

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

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, \dots). 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:

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

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:

$$LHS := RHS,$$

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**:

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

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:

$$b \in B.$$

When b does not belong to B , we write:

$$b \notin B.$$

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:

$$C \subset D$$

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:

$$C = D \iff C \subset D, D \subset C .$$

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

$$C \neq D .$$

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:

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

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:

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

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:

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

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.:

$$B^c := U \setminus B .$$

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**:

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

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

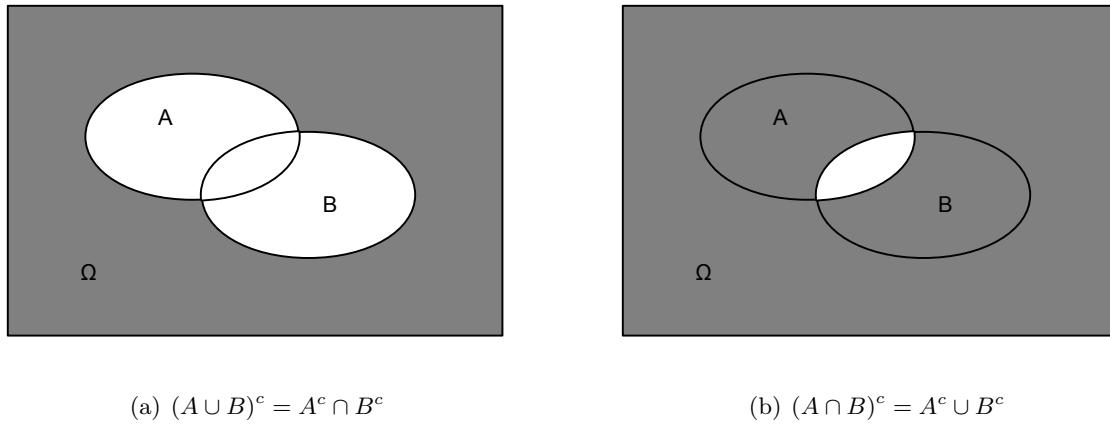


Figure 1.2: These Venn diagram illustrate De Morgan's Laws.

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 $A = B = C = U$ |

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, $S = \{S_1, S_2, S_3, \dots, S_{50}\}$.
- 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:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

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

$\{a_1, a_2, \dots, a_n\}$	— a set containing the elements, a_1, a_2, \dots, 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 .
$\{\} \text{ or } \emptyset$	— 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 .

1.2 Natural Numbers, Integers and Rational Numbers

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

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

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**:

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

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

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

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

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

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

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

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

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, \dots, A_m is:

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

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

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

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 $\#A_1 = s_1, \#A_2 = s_2, \dots, \#A_m = s_m$. 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:

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

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

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

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

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

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**:

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

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 $a \leq b$ 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. $a \geq b$, if $b \leq a$. 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 $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$. 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 ?

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

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:

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

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**

$$a + b = b + a, \quad a \cdot b = b \cdot a ,$$

and associative

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

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

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

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:

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

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.

1.3 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:

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

and the sequence $0.d_1d_2d_3 \dots$ 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:

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

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}, \bar{x} and \underline{y}, \bar{y} .

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

$$\sqrt{2} = 1.41421356237309 \dots \text{the side length of a square with area of 2 units}$$

$$\pi = 3.14159265358979 \dots \text{the ratio of the circumference to diameter of a circle}$$

$$e = 2.71828182845904 \dots \text{Euler's constant}$$

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

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

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:

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

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:

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

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:

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

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

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

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 $x = n + 0.d_1d_2d_3\dots \in \mathbb{R}$ can be arbitrarily precisely enclosed by the rational numbers $\underline{x}_k := n + \frac{d_1}{10} + \frac{d_2}{100} + \dots + \frac{d_k}{10^k}$ and $\bar{x}_k := n + \frac{d_1}{10} + \frac{d_2}{100} + \dots + \frac{d_k}{10^k} + \frac{1}{10^k}$ by increasing k . Suppose first that $b > 1$. Then, using rational powers, we can enclose b^x ,

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

within an interval of width $b^{n+\frac{d_1}{10}+\frac{d_2}{100}+\dots+\frac{d_k}{10^k}} \left(b^{\frac{1}{10^k}} - 1 \right) < b^{n+1}(b-1)/10^k$. 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 $b = 0$, $b^x := 1$.

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:

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

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 $x \leq a$ 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
$\circ \notin A$	\circ 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 $\{1, 2, 3, \dots\}$
\mathbb{Z}	The set of integers $\{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$
\mathbb{D}_+	The set of non-negative integers $\{0, 1, 2, 3, \dots\}$
\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. $\{x : x \in U, \text{ the universal set, but } x \notin A\}$
$A_1 \times A_2 \times \dots \times A_m$	The m -product set $\{(a_1, a_2, \dots, a_m) : a_1 \in A_1, a_2 \in A_2, \dots, a_m \in A_m\}$
A^m	The m -product set $\{(a_1, a_2, \dots, a_m) : a_1 \in A, a_2 \in A, \dots, a_m \in A\}$
$f := f(x) = y : \mathbb{X} \rightarrow \mathbb{Y}$	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

1.4 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/comdocs/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 :

$$\begin{aligned} p &= 45.89 * 1.00009 \\ m &= 5376.0 - 6.00 \end{aligned}$$

$$\begin{aligned} d &= 89.0 / 23.3454 \\ p &= 2^{0.5} \end{aligned}$$

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.

```
>> 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. brackets or parentheses; 2. exponents (powers and roots); 3. division and multiplication; 4. addition and subtraction. 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 **clearing** 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 .

1.5 Permutations, Factorials and Combinations

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 $\{1, 2\}$:

$$12, \quad 21,$$

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,

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

and the total number of permutations called ‘ n factorial’ and denoted by $n!$ is

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

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) :

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

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, *Pingala’s Chandasāstra*, due to Halāyudha (10-th century AD).

1.6 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

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

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 $m \leq n$, we write

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

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:

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

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:

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

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

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

The finite sequence $\langle x_{m:n} \rangle$ has $\mathcal{I} = \{m, m+1, m+2, m+3, \dots, n\}$ as its index set while an infinite sequence $\langle x_{m:\infty} \rangle$ has $\mathcal{I} = \{m, m+1, m+2, m+3, \dots\}$ 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:

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

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:

$$\mathbf{B}X = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n-1} & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n-1} & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m-1,1} & x_{m-1,2} & \dots & x_{m-1,n-1} & x_{m-1,n} \\ x_{m,1} & x_{m,2} & \dots & x_{m,n-1} & x_{m,n} \end{bmatrix}$$

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.

$$\text{A row vector } \mathbf{B}x = [x_1 \ x_2 \ \dots \ x_n] = (x_1, x_2, \dots, x_n)$$

$$\text{and a column vector } \mathbf{B}y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{m-1} \\ y_m \end{bmatrix} = [y_1 \ y_2 \ \dots \ y_m]' = (y_1, y_2, \dots, y_m)'.$$

The superscripting by $'$ is the transpose operation and simply means that the rows and columns are exchanged. Thus the transpose of the matrix BX is:

$$BX' = \begin{bmatrix} x_{1,1} & x_{2,1} & \dots & x_{m-1,1} & x_{m,1} \\ x_{1,2} & x_{2,2} & \dots & x_{m-1,2} & x_{m,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{1,n-1} & x_{2,n-1} & \dots & x_{m-1,n-1} & x_{m,n-1} \\ x_{1,n} & x_{2,n} & \dots & x_{m-1,n} & x_{m,n} \end{bmatrix}$$

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**.

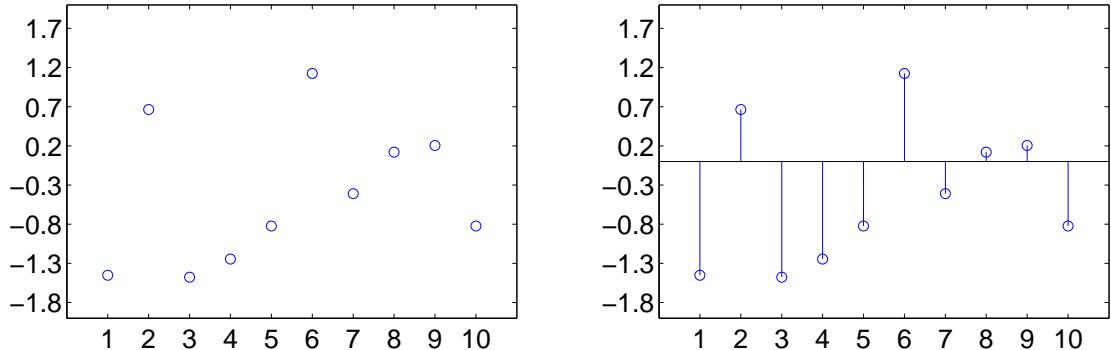
Labwork 6 (Sequences as arrays) Let us learn to represent, visualise and operate finite sequences as MATLAB arrays. Try out the commands and read the comments for clarification.

```
>> 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 $\{(1, b_1), (2, b_2), \dots, (10, b_n)\}$ representing the function $b(i) = b_i : \{1, 2, \dots, n\} \rightarrow \{b_1, b_2, \dots, b_n\}$ via **point plot** and **stem plot** using Matlab's **plot** and **stem** commands, respectively.

```
>> 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 7 (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.

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

```

>> a = [1 2]           % an 1 X 2 row vector
>> z = [1 2 3]         % Declare an 1 X 3 row vector z with three numbers
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 =      2      3      4
                                % x is the sum of vectors y and z (with same size 1 X 3)
>> y = y * 2
y =      2      2      2
                                % y is updated to 2 * y (each term of y is multiplied by 2)
>> p = z .* y
p =      2      4      6
                                % p is the vector obtained by term-by-term product of z and y
>> d = z ./ y
d = 0.5000    1.0000    1.5000
                                % d is the vector obtained by term-by-term division of z and y
>> t=linspace(-10,10,4)
t = -10.0000   -3.3333   3.3333   10.0000
                                % t has 4 numbers equally-spaced between -10 and 10
>> s = sin(t)
s = 0.5440    0.1906   -0.1906   -0.5440
                                % s is a vector obtained from the term-wise sin of the vector t
>> sSq = sin(t) .^ 2
sSq = 0.2960    0.0363    0.0363    0.2960
                                % sSq is an array obtained from term-wise squaring (. ^ 2) of the sin(t) array
>> cSq = cos(t) .^ 2
cSq = 0.7040    0.9637    0.9637    0.7040
                                % cSq is an array obtained from term-wise squaring (. ^ 2) of the cos(t) array
>> 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
n =      1      1      1      1
                                % we can directly do term-wise operation sin^2(t) + cos^2(t) of t as well
>> t2 = (-10:6.666665:10)
t2 = -10.0000   -3.3333   3.3333   10.0000
                                % t2 is similar to t above but with ':' syntax of (start:increment:stop)

```

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,:)
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 8 (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.

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

Let us first recall some elementary ideas from real analysis.

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

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

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

Example 9 (Limit of a sequence of 17s) Let $\langle x_i \rangle_{i=1}^{\infty} = 17, 17, 17, \dots$. Then $\lim_{i \rightarrow \infty} x_i = 17$. 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 10 (Limit of $1/i$) Let $\langle x_i \rangle_{i=1}^{\infty} = \frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \dots$, i.e. $x_i = \frac{1}{i}$, then $\lim_{i \rightarrow \infty} x_i = 0$. 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:

$$\langle x_{1:\infty} \rangle = \frac{1}{1}, \frac{1}{4}, \frac{1}{9}, \dots \quad \text{and} \quad \langle x_{1:\infty} \rangle = \frac{1}{1}, \frac{1}{8}, \frac{1}{27}, \dots ,$$

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

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

and finally some that approach 0 from either side are:

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

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:

$$x \rightarrow a$$

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:

$$x \rightarrow a^+ \quad \text{or} \quad x \rightarrow a^-$$

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

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

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

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:

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

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

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

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

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

However, the following limit does not exist:

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

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

Example 12 (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 the Euler's constant e :

$$\lim_{x \rightarrow 0} f(x) = \lim_{x \rightarrow 0} (1+x)^{\frac{1}{x}} = e \approx 2.71828 .$$

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

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

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

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

Example 15 (Limit of $(1 - \frac{\lambda}{n})^{-\alpha}$) The limit of $f(n) = (1 - \frac{\lambda}{n})^{-\alpha}$, for some $\alpha > 0$, as n approaches ∞ exists and it is 1 :

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

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

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

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

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

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

Example 16 (Discontinuity of $f(x) = (1+x)^{\frac{1}{x}}$ at 0) Let us reconsider the function $f(x) = (1+x)^{\frac{1}{x}} : \mathbb{R} \rightarrow \mathbb{R}$. Clearly, $f(x)$ is continuous at 1, since:

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

but it is not continuous at 0, since:

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

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

1.7 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 17 (Floors and ceilings)

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

Labwork 18 (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 19 (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 “ $x \bmod y$ ” as:

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

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, $E_1, E_2 \dots E_n$, 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 20 Some standard examples of experiments are the following:

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

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

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

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

- 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 21 Consider a generic die-tossing experiment by a human experimenter. Here $\Omega = \{\omega_1, \omega_2, \omega_3, \dots, \omega_6\}$, 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 1, 2, 3, 4, 5, 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 22 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.

Experiment 1 (The Bernoulli Product Experiment; Toss a coin n times) Suppose our experiment entails tossing a coin n times and recording H for Heads and T for Tails. When $n = 3$, one possible outcome of this experiment is 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:

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

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

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

Another event, say B , that ‘no Heads occur’ is:

$$B = TTT$$

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 .

Figure 2.1: A binary tree whose leaves are all possible outcomes.

Classwork 23 (A thrice-bifurcating tree of outcomes) Can you think of a graphical way to enumerate the outcomes of the Experiment 1? 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 $N(H, n)$ 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 $N(H, 1000) = 9/1000 = 0.009$. Suppose we continued the number of tosses to a million and found that this number approached closer to 0.1, or, more generally, $N(H, n) \rightarrow 0.1$ as $n \rightarrow \infty$. 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 $N(H, n) \rightarrow 0.5$ as $n \rightarrow \infty$. 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 $H \cup T$. The probability that “something happens” is 1. More formally:

$$N(H \cup T, n) = \frac{n}{n} = 1.$$

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 $H \cap T = \emptyset$. The event $H \cup T$, namely that the event that “coin lands Heads **or** coin lands Tails” satisfies:

$$N(H \cup T, n) = N(H, n) + N(T, n).$$

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}^∞ 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}^∞ . Then the LTRF idea suggests:

$$N(A, n) := \frac{\text{Number of times } A \text{ occurs}}{n = \text{Number of performances of } \mathcal{E}} \rightarrow \mathbf{P}(A), \text{ as } n \rightarrow \infty \quad (2.1)$$

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: $\boxed{A \in \mathcal{F} \implies A^c \in \mathcal{F}}$.
3. It is closed under countable unions: $\boxed{A_1, A_2, \dots \in \mathcal{F} \implies \bigcup_i A_i := A_1 \cup A_2 \cup \dots \in \mathcal{F}}$.

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. (Ω, \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. $\boxed{\mathbf{P} : \mathcal{F} \rightarrow [0, 1]}$, 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 :

$$\boxed{A \cap B = \emptyset \implies \mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)}.$$

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.

$$\overbrace{\mathbf{P}(A) + \mathbf{P}(A^c)}^{LHS} = \underbrace{\mathbf{P}(A \cup A^c)}_{+ \text{ rule } \because A \cap A^c = \emptyset} = \underbrace{\mathbf{P}(\Omega)}_{A \cup A^c = \Omega} = \underbrace{1}_{\because \mathbf{P}(\Omega) = 1} \stackrel{RHS}{\Rightarrow} \mathbf{P}(A^c) = 1 - \mathbf{P}(A)$$

- 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:

$$\begin{aligned} A &= (A \setminus B) \cup (A \cap B) && \text{and} && (A \setminus B) \cap (A \cap B) = \emptyset, \\ A \cup B &= (A \setminus B) \cup B && \text{and} && (A \setminus B) \cap B = \emptyset \end{aligned}$$

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. For a sequence of mutually disjoint events $A_1, A_2, A_3, \dots, A_n$:

$$\boxed{A_i \cap A_j = \emptyset \text{ for any } i, j \implies \mathbf{P}(A_1 \cup A_2 \cup \dots \cup A_n) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \dots + \mathbf{P}(A_n).}$$

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:

$$\mathbf{P}(A_1 \cup A_2 \cup A_3) = \mathbf{P}((A_1 \cup A_2) \cup A_3) \underset{+ \text{ rule}}{=} \mathbf{P}(A_1 \cup A_2) + \mathbf{P}(A_3) \underset{+ \text{ rule}}{=} \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3)$$

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 24 (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

$$\Omega = \{1, 2, 3, \dots, 39, 40\}.$$

Because the balls are vigorously whirled around inside the Lotto machine 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:

$$\mathbf{P}(\omega) = \frac{1}{40}, \text{ for each } \omega \in \Omega = \{1, 2, 3, \dots, 39, 40\} .$$

(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.)

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.

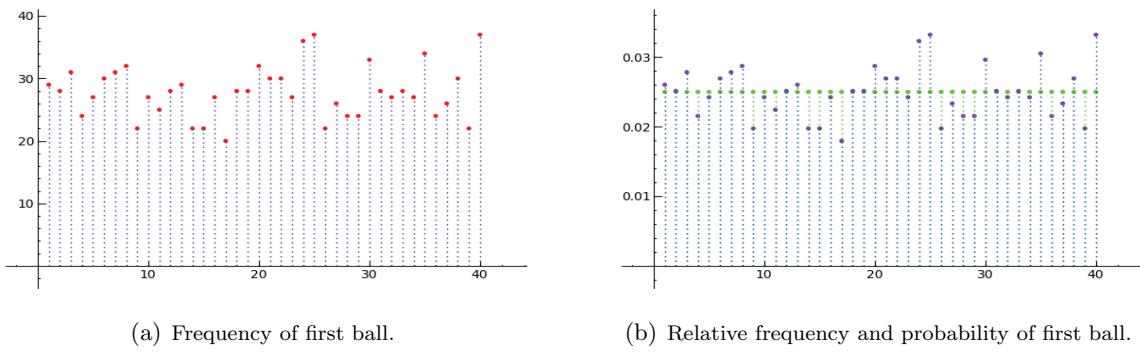


Figure 2.2: First ball number in 1114 NZ Lotto draws from 1987 to 2008.

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.

Probability Model

Sample space Ω

Sample point ω

(No counterpart)

Event A , a (suitable) subset of Ω

$\mathbf{P}(A)$, a number between 0 and 1

Real-world Interpretation

Set of all outcomes of an experiment

Possible outcome of an experiment

Actual outcome ω^* of an experiment

The real-world event corresponding to A

occurs if and only if $\omega^* \in A$

Probability that A will occur for an experiment yet to be performed

Events in Probability Model

Sample space Ω

The \emptyset of Ω

The intersection $A \cap B$

$A_1 \cap A_2 \cap \dots \cap A_n$

The union $A \cup B$

$A_1 \cup A_2 \cup \dots \cup A_n$

A^c , the complement of A

$A \setminus B$

$A \subset B$

Real-world Interpretation

The certain even ‘something happens’

The impossible event ‘nothing happens’

‘Both A and B occur’

‘All of the events A_1, A_2, \dots, A_n occur simultaneously’

‘At least one of A and B occurs’

‘At least one of the events A_1, A_2, \dots, A_n occurs’

‘ A does not occur’

‘ A occurs, but B does not occur’

‘If A occurs, then B must occur’

In the probability model of Example 24, show that for any event $E \subset \Omega$,

$$\mathbf{P}(E) = \frac{1}{40} \times \text{number of elements in } E .$$

Let $E = \{\omega_1, \omega_2, \dots, \omega_k\}$ be an event with k outcomes (simple events). Then by the addition rule for mutually exclusive events we get:

$$\mathbf{P}(E) = \mathbf{P}(\{\omega_1, \omega_2, \dots, \omega_k\}) = \mathbf{P}\left(\bigcup_{i=1}^k \{\omega_i\}\right) = \sum_{i=1}^k \mathbf{P}(\{\omega_i\}) = \sum_{i=1}^k \frac{1}{40} = \frac{k}{40} .$$

2.2.2 Sigma Algebras of Typical Experiments*

Example 25 ('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?

$$\Omega = \{\text{H}, \text{T}\}, \quad \mathcal{F} = \{\Omega, \emptyset, \{\text{H}\}, \{\text{T}\}\} ,$$

A function that will satisfy the definition of probability for this collection of events \mathcal{F} and assign $\mathbf{P}(\text{H}) = \frac{1}{2}$ 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 $A \in \mathcal{F}$	$\mathbf{P} : \mathcal{F} \rightarrow [0, 1]$	$\mathbf{P}(A) \in [0, 1]$
$\Omega = \{\text{H}, \text{T}\} \bullet$	→	1
$\text{T} \bullet$	→	$1 - \frac{1}{2}$
$\text{H} \bullet$	→	$\frac{1}{2}$
$\emptyset \bullet$	→	0

Classwork 26 (The trivial sigma algebra) Note that $\mathcal{F}' = \{\Omega, \emptyset\}$ is also a sigma algebra of the sample space $\Omega = \{\text{H}, \text{T}\}$. Can you think of a probability for the collection \mathcal{F}' ?

Event $A \in \mathcal{F}'$	$\mathbf{P} : \mathcal{F}' \rightarrow [0, 1]$	$\mathbf{P}(A) \in [0, 1]$
$\Omega = \{\text{H}, \text{T}\} \bullet$	→	
$\emptyset \bullet$	→	

Thus, \mathcal{F} and \mathcal{F}' are two distinct sigma algebras over our $\Omega = \{\text{H}, \text{T}\}$. 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 and they are:

- When the sample space $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ 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.
- When the sample space $\Omega = \{\omega_1, \omega_2, \dots\}$ 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 $d \in \{1, 2, 3, \dots\}$ then the **Borel sigma-algebra** is the smallest sigma-algebra containing all **half-spaces**, i.e., sets of the form

$$\{x = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d : x_1 \leq c_1, x_2 \leq c_2, \dots, x_d \leq c_d\}, \quad \text{for any } c = (c_1, c_2, \dots, c_d) \in \mathbb{R}^d,$$

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 $\mathbb{S} = \{s_1, s_2, \dots, s_k\}$, let Ω be the sequence space $\mathbb{S}^\infty := \mathbb{S} \times \mathbb{S} \times \mathbb{S} \times \dots$, i.e., the set of sequences of infinite length that are made up of elements from \mathbb{S} . A set of the form

$$A_1 \times A_2 \times \dots \times A_n \times \mathbb{S} \times \mathbb{S} \times \dots, \quad A_k \subset \mathbb{S} \text{ for all } k \in \{1, 2, \dots, n\},$$

is called a **cylinder set**. The set of events in \mathbb{S}^∞ is the smallest sigma-algebra containing the cylinder sets.

Exercises

Ex. 2.1 — 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 $\Omega = \{A, B, C, \dots, Z\}$ (ignoring upper/lower cases), then what is the probability of these events?

- (a) $\mathbf{P}(\{Z\})$
- (b) $\mathbf{P}(\text{'picking any letter'})$
- (c) $\mathbf{P}(\{E, Z\})$
- (d) $\mathbf{P}(\text{'picking a vowel'})$
- (e) $\mathbf{P}(\text{'picking any letter in the word WAZZZUP'})$
- (f) $\mathbf{P}(\text{'picking any letter in the word WAZZZUP or a vowel'})$.

Ex. 2.2 — 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. 2.3 — 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. 2.4 — 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: $C = \{B, I, G\}$ and $D = \{G, I, N\}$. Using the addition rule for two arbitrary events, find $\mathbf{P}(C \cup D)$.

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, \dots are disjoint then $\mathbf{P}(A_1 \cup A_2 \cup \dots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \dots$

Rules:

$$\begin{aligned}\mathbf{P}(A^c) &= 1 - \mathbf{P}(A) \\ \mathbf{P}(A \cup B) &= \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B) \quad [\text{always true}]\end{aligned}$$

2.3 Conditional Probability

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}^∞ which is the ‘independent’ repetition 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

$$\mathbf{P}(B|A) \quad “\rightarrow” \quad \frac{N(A \cap B, n)}{N(A, n)} = \frac{N(A \cap B, n)/n}{N(A, n)/n} = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)}$$

as our $\mathcal{E}^n \rightarrow \mathcal{E}^\infty$. So, we **define** conditional probability as we want.

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,

$$\mathbf{P}(B|A) := \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} . \quad (2.2)$$

Note that 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:

$$\mathbf{P}(B|A) : \mathcal{F} \rightarrow [0, 1]$$

that assigns to each $B \in \mathcal{F}$ a number in the interval $[0, 1]$, such that,

1. $\mathbf{P}(\Omega|A) = 1$ Meaning ‘Something Happens given the event A happens’
2. The ‘Addition Rule’ axiom holds, ie. for events $B_1, B_2 \in \mathcal{F}$,

$$B_1 \cap B_2 = \emptyset \text{ implies } \mathbf{P}(B_1 \cup B_2|A) = \mathbf{P}(B_1|A) + \mathbf{P}(B_2|A) .$$

3. For mutually exclusive or pairwise-disjoint events, B_1, B_2, \dots ,

$$\mathbf{P}(B_1 \cup B_2 \cup \dots |A) = \mathbf{P}(B_1|A) + \mathbf{P}(B_2|A) + \dots .$$

From the definition of conditional probability we get the following rules:

1. Complementation rule: $\mathbf{P}(B|A) = 1 - \mathbf{P}(B^c|A)$.
2. Addition rule for two arbitrary events B_1 and B_2 :

$$\mathbf{P}(B_1 \cup B_2|A) = \mathbf{P}(B_1|A) + \mathbf{P}(B_2|A) - \mathbf{P}(B_1 \cap B_2|A) .$$

3. 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 27 (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:

$$\mathbf{P}(-|D^c) = \frac{\mathbf{P}(- \cap D^c)}{\mathbf{P}(D^c)} = \frac{0.891}{0.099 + 0.891} \approx 0.9 .$$

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 test positive. What is the probability that you have the disease ?

$$\mathbf{P}(D|+) = \frac{\mathbf{P}(D \cap +)}{\mathbf{P}(+)} = \frac{0.009}{0.009 + 0.099} \approx 0.08$$

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.

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 .

Proposition 13 (Bayes' Theorem, 1763) Suppose the events $A_1, A_2, \dots, A_k \in \mathcal{F}$, with $\mathbf{P}(A_h) > 0$ for each $h \in \{1, 2, \dots, k\}$, partition the sample space Ω , ie. they are mutually exclusive (disjoint) and exhaustive events with positive probability:

$$A_i \cap A_j = \emptyset, \text{ for any distinct } i, j \in \{1, 2, \dots, k\}, \quad \bigcup_{h=1}^k A_h = \Omega, \quad \mathbf{P}(A_h) > 0$$

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

$$\mathbf{P}(A_h|B) = \frac{\mathbf{P}(B|A_h)\mathbf{P}(A_h)}{\sum_{h=1}^k \mathbf{P}(B|A_h)\mathbf{P}(A_h)} \quad (2.3)$$

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:

$$\mathbf{P}(B) = \sum_{h=1}^k \mathbf{P}(B|A_h)\mathbf{P}(A_h) \quad (2.4)$$

is also called ‘the law of total probability’ or ‘the total probability theorem’. We call $\mathbf{P}(A_h)$ the **prior probability of A_h** and $\mathbf{P}(A_h|B)$ the **posterior probability of A_h** .

Example 28 (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” ?

$$\mathbf{P}(A_1|B) = \frac{\mathbf{P}(B|A_1)\mathbf{P}(A_1)}{\mathbf{P}(B|A_1)\mathbf{P}(A_1) + \mathbf{P}(B|A_2)\mathbf{P}(A_2) + \mathbf{P}(B|A_3)\mathbf{P}(A_3)} = \frac{0.9 \times 0.7}{(0.9 \times 0.7) + (0.01 \times 0.2) + (0.01 \times 0.1)} = \frac{0.63}{0.633} \approx 0.995$$

$$\mathbf{P}(A_2|B) = \frac{\mathbf{P}(B|A_2)\mathbf{P}(A_2)}{\mathbf{P}(B|A_1)\mathbf{P}(A_1) + \mathbf{P}(B|A_2)\mathbf{P}(A_2) + \mathbf{P}(B|A_3)\mathbf{P}(A_3)} = \frac{0.01 \times 0.2}{(0.9 \times 0.7) + (0.01 \times 0.2) + (0.01 \times 0.1)} = \frac{0.002}{0.633} \approx 0.003$$

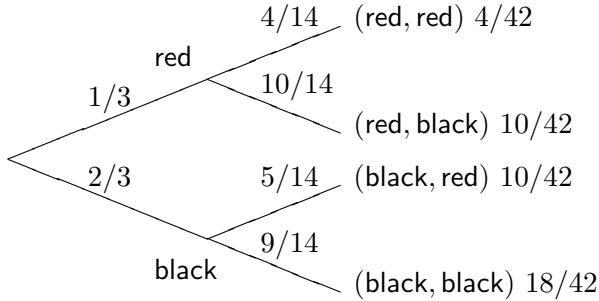
$$\mathbf{P}(A_3|B) = \frac{\mathbf{P}(B|A_3)\mathbf{P}(A_3)}{\mathbf{P}(B|A_1)\mathbf{P}(A_1) + \mathbf{P}(B|A_2)\mathbf{P}(A_2) + \mathbf{P}(B|A_3)\mathbf{P}(A_3)} = \frac{0.01 \times 0.1}{(0.9 \times 0.7) + (0.01 \times 0.2) + (0.01 \times 0.1)} = \frac{0.001}{0.633} \approx 0.002$$

Note that $\mathbf{P}(A_1|B) + \mathbf{P}(A_2|B) + \mathbf{P}(A_3|B) = 0.995 + 0.003 + 0.002 = 1$.

Example 29 (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

$$P(\text{second ball is red}) = 4/42 + 10/42 = 1/3 .$$



Alternatively, use the total probability theorem to break the problem down into manageable pieces. Let $R_1 = \{(\text{red}, \text{red}), (\text{red}, \text{black})\}$ and $R_2 = \{(\text{red}, \text{red}), (\text{black}, \text{red})\}$ be the events corresponding to a red ball in the 1st and 2nd draws, respectively, and let $B_1 = \{(\text{black}, \text{red}), (\text{black}, \text{black})\}$ be the event of a black ball on the first draw.

Now R_1 and B_1 partition Ω so we can write:

$$\begin{aligned} P(R_2) &= \mathbf{P}(R_2 \cap R_1) + \mathbf{P}(R_2 \cap B_1) \\ &= \mathbf{P}(R_2|R_1)\mathbf{P}(R_1) + \mathbf{P}(R_2|B_1)\mathbf{P}(B_1) \\ &= (4/14)(1/3) + (5/14)(2/3) = 1/3 . \end{aligned}$$

2.3.1 Independence and Dependence

Definition 14 (Independence of two events) Any two events A and B are said to be **independent** if and only if

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) . \quad (2.5)$$

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:

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

says that information about the occurrence of B does not affect the occurrence of A . Therefore, the probability of their joint occurrence $\mathbf{P}(A \cap B)$ is simply the product of their individual probabilities $\mathbf{P}(A)\mathbf{P}(B)$.

Definition 15 (Independence of a sequence of events) We say that a finite or infinite sequence of events A_1, A_2, \dots are independent if whenever i_1, i_2, \dots, i_k are distinct elements from the set of indices \mathbb{N} , such that $A_{i_1}, A_{i_2}, \dots, A_{i_k}$ are defined (elements of \mathcal{F}), then

$$\mathbf{P}(A_{i_1} \cap A_{i_2} \dots \cap A_{i_k}) = \mathbf{P}(A_{i_1})\mathbf{P}(A_{i_2}) \cdots \mathbf{P}(A_{i_k})$$

Example 30 (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,

$$\mathbf{P}(\text{Heads on the first toss} \cap \text{Tails on the second toss}) = \mathbf{P}(\mathsf{H})\mathbf{P}(\mathsf{T}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} .$$

(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 31 (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 $\Omega =$

$\{(1,1), (1,2), \dots, (1,6), (2,1), \dots, (2,6), \dots, (5,6), (6,6)\}$ 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 32 (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.

Exercises

Ex. 2.5 — 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.

Ex. 2.6 — 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. 2.7 — 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. 2.8 — 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. 2.9 — 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. 2.10 — 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. 2.11 — **The sensitivity and specificity of a medical diagnostic test for a disease are defined as follows:

$$\begin{aligned}\text{sensitivity} &= \mathbf{P}(\text{test is positive} \mid \text{patient has the disease}) , \\ \text{specificity} &= \mathbf{P}(\text{test is negative} \mid \text{patient does not have the disease}) .\end{aligned}$$

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 population 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. 2.12 — **The detection rate and false alarm rate of an intrusion sensor are defined as

$$\begin{aligned}\text{detection rate} &= \mathbf{P}(\text{detection declared} \mid \text{intrusion}) , \\ \text{false alarm rate} &= \mathbf{P}(\text{detection declared} \mid \text{no intrusion}) .\end{aligned}$$

If the detection rate is 0.999 and the false alarm rate is 0.001, and the probability of an intrusion occurring 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. 2.13 — **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.

CONDITIONAL PROBABILITY SUMMARY

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

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

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

Conditional probabilities obey the axioms and rules of probability.

Chapter 3

Random Variables

It can be 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	$\mathbb{Z}_+ := \{0, 1, 2, \dots\} \subset \mathbb{R}$
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}

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 16 (Random Variable) Let (Ω, \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}

$$X : \Omega \rightarrow \mathbb{R}$$

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.:

$$\text{for every } x \in \mathbb{R}, \quad X^{[-1]}((-\infty, x]) := \{\omega : X(\omega) \leq x\} \in \mathcal{F}.$$

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:

$$\mathbf{P}(X \leq x) = \mathbf{P}(X^{[-1]}((-\infty, x])) := \mathbf{P}(\{\omega : X(\omega) \leq x\}). \quad (3.1)$$

Definition 17 (Distribution Function) The **Distribution Function (DF)** or **Cumulative Distribution Function (CDF)** of any RV X , over a probability triple (Ω, \mathcal{F}, P) , denoted by F is:

$$F(x) := \mathbf{P}(X \leq x) = \mathbf{P}(\{\omega : X(\omega) \leq x\}), \quad \text{for any } x \in \mathbb{R}. \quad (3.2)$$

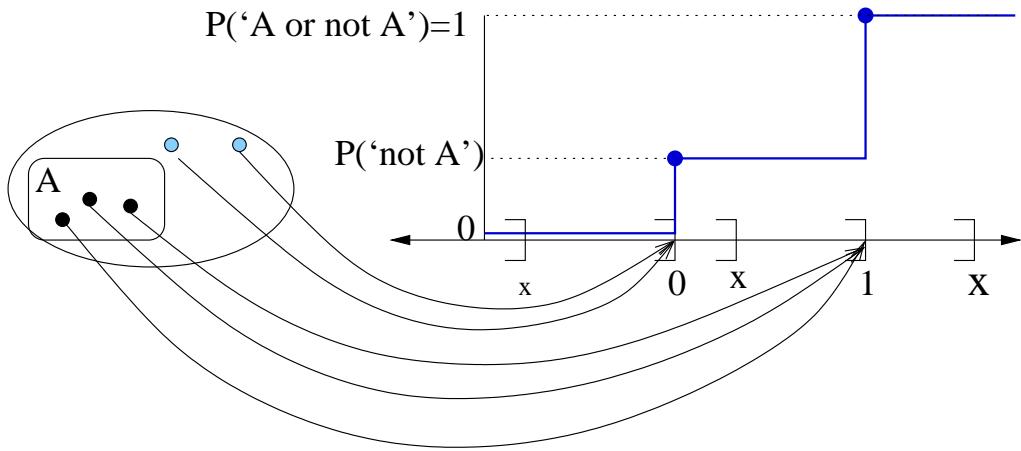
Thus, $F(x)$ or simply F is a non-decreasing, right continuous, $[0, 1]$ -valued function over \mathbb{R} . When a RV X has DF F we write $X \sim F$.

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 18 (Indicator Function) The **Indicator Function** of an event A denoted $\mathbb{1}_A$ is defined as follows:

$$\mathbb{1}_A(\omega) := \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} \quad (3.3)$$

Figure 3.1: The Indicator function of event $A \in \mathcal{F}$ is a RV $\mathbb{1}_A$ with DF F



Classwork 33 (Indicator function is a random variable) 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:

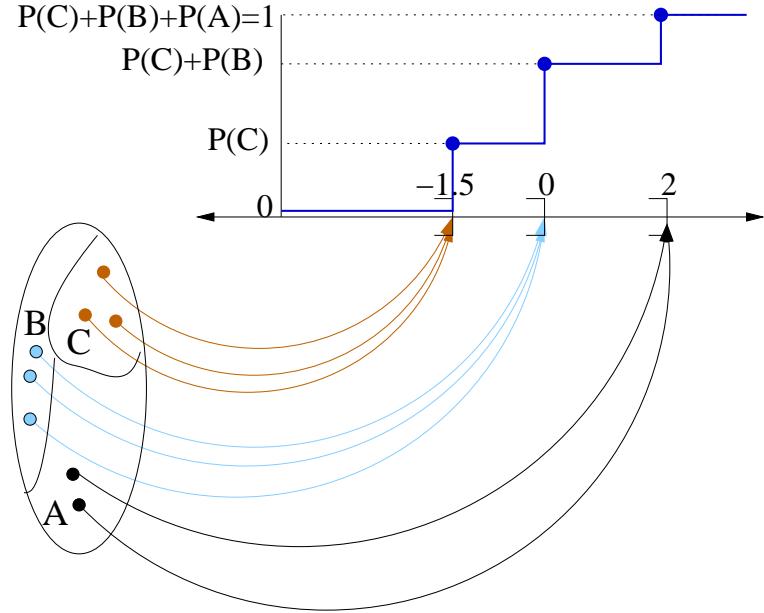
$$\mathbb{1}_A^{[-1]}((-\infty, x]) := \{\omega : \mathbb{1}_A(\omega) \leq x\} = \begin{cases} \emptyset & \text{if } x < 0 \\ A^c & \text{if } 0 \leq x < 1 \\ A \cup A^c = \Omega & \text{if } 1 \leq x \end{cases}$$

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.

Some useful properties of the Indicator Function are:

$$\mathbb{1}_{A^c} = 1 - \mathbb{1}_A, \quad \mathbb{1}_{A \cap B} = \mathbb{1}_A \mathbb{1}_B, \quad \mathbb{1}_{A \cup B} = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_A \mathbb{1}_B$$

We slightly abuse notation when A is a single element set by ignoring the curly braces.

Figure 3.2: A RV X from a sample space Ω with 8 elements to \mathbb{R} and its DF F 

Classwork 34 (A random variable with three values and eight sample points) Consider the RV X of Figure 3.2. 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?

3.2 An Elementary Discrete Random Variable

When a RV takes at most countably many values from a discrete set $\mathbb{D} \subset \mathbb{R}$, we call it a **discrete** RV. Often, \mathbb{D} is the set of integers \mathbb{Z} .

Definition 19 (probability mass function (PMF)) Let X be a discrete RV over a probability triple (Ω, \mathcal{F}, P) . We define the **probability mass function** (PMF) f of X to be the function $f : \mathbb{D} \rightarrow [0, 1]$ defined as follows:

$$f(x) := \mathbf{P}(X = x) = \mathbf{P}(\{\omega : X(\omega) = x\}), \quad \text{where } x \in \mathbb{D}.$$

The DF F and PMF f for a discrete RV X satisfy the following:

1. For any $x \in \mathbb{R}$,

$$\mathbf{P}(X \leq x) = F(x) = \sum_{\mathbb{D} \ni y \leq x} f(y) := \sum_{y \in \mathbb{D} \cap (-\infty, x]} f(y).$$

2. For any $a, b \in \mathbb{D}$ with $a < b$,

$$\mathbf{P}(a < X \leq b) = F(b) - F(a) = \sum_{y \in \mathbb{D} \cap (a, b]} f(y) .$$

In particular, when $\mathbb{D} = \mathbb{Z}$ and $a = b - 1$,

$$\mathbf{P}(b - 1 < X \leq b) = F(b) - F(b - 1) = f(b) = \mathbf{P}(\{\omega : X(\omega) = b\}) .$$

3. And of course

$$\sum_{x \in \mathbb{D}} f(x) = 1$$

The Indicator Function $\mathbf{1}_A$ of the event that ‘ A occurs’ for the θ -specific experiment \mathcal{E} over some probability triple $(\Omega, \mathcal{F}, \mathbf{P}_\theta)$, with $A \in \mathcal{F}$, is the Bernoulli(θ) RV. The parameter θ denotes the probability that ‘ A occurs’ (see Figure 3.3 when A is the event that ‘ H occurs’). This is our first example of a discrete RV.

Model 2 (Bernoulli(θ)) Given a parameter $\theta \in [0, 1]$, the probability mass function (PMF) for the Bernoulli(θ) RV X is:

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x} \mathbf{1}_{\{0,1\}}(x) = \begin{cases} \theta & \text{if } x = 1, \\ 1 - \theta & \text{if } x = 0, \\ 0 & \text{otherwise} \end{cases} \quad (3.4)$$

and its DF is:

$$F(x; \theta) = \begin{cases} 1 & \text{if } 1 \leq x, \\ 1 - \theta & \text{if } 0 \leq x < 1, \\ 0 & \text{otherwise} \end{cases} \quad (3.5)$$

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 $\mathbf{P}_\theta(X = 0) = 1 - \theta$.

3.3 An Elementary Continuous Random Variable

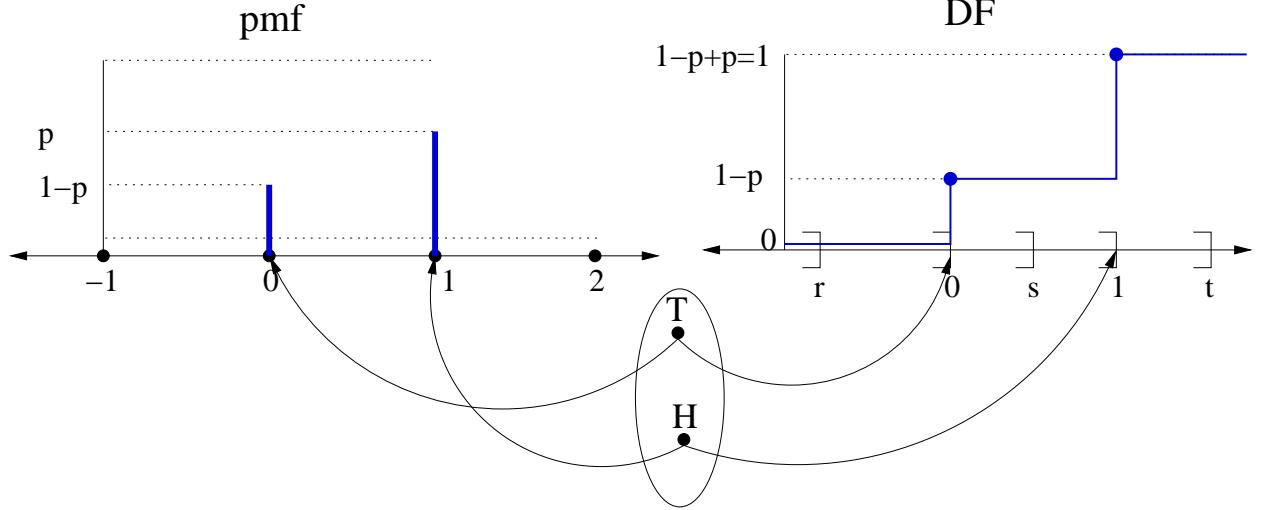
When a RV takes values in the continuum we call it a **continuous** RV. An example of such a RV is the vertical position (in micro meters) since the original release of a pollen grain on water. Another example of a continuous RV is the volume of water (in cubic meters) that fell on the southern Alps last year.

Definition 20 (probability density function (PDF)) A RV X is said to be ‘continuous’ if there exists a piecewise-continuous function f , called the probability density function (PDF) of X , such that for any $a, b \in \mathbb{R}$ with $a < b$,

$$\mathbf{P}(a < X \leq b) = F(b) - F(a) = \int_a^b f(x) dx .$$

The following hold for a continuous RV X with PDF f :

Figure 3.3: The Indicator Function $\mathbb{1}_H$ of the event ‘Heads occurs’, for the experiment ‘Toss 1 times,’ \mathcal{E}_θ^1 , as the RV X from the sample space $\Omega = \{H, T\}$ to \mathbb{R} and its DF F . The probability that ‘Heads occurs’ and that ‘Tails occurs’ are $f(1; \theta) = \mathbf{P}_\theta(X = 1) = \mathbf{P}_\theta(H) = \theta$ and $f(0; \theta) = \mathbf{P}_\theta(X = 0) = \mathbf{P}_\theta(T) = 1 - \theta$, respectively.



1. For any $x \in \mathbb{R}$, $\mathbf{P}(X = x) = 0$.
2. Consequentially, for any $a, b \in \mathbb{R}$ with $a \leq b$,

$$\mathbf{P}(a < X < b) = \mathbf{P}(a < X \leq b) = \mathbf{P}(a \leq X \leq b) = \mathbf{P}(a \leq X < b) .$$

3. By the fundamental theorem of calculus, except possibly at finitely many points (where the continuous pieces come together in the piecewise-continuous f):

$$f(x) = \frac{d}{dx} F(x)$$

4. And of course f must satisfy:

$$\int_{-\infty}^{\infty} f(x) dx = \mathbf{P}(-\infty < X < \infty) = 1 .$$

An elementary and fundamental example of a continuous RV is the Uniform(0, 1) RV of Model 3. It forms the foundation for random variate generation and simulation. In fact, it is appropriate to call this the fundamental model since every other experiment can be obtained from this one.

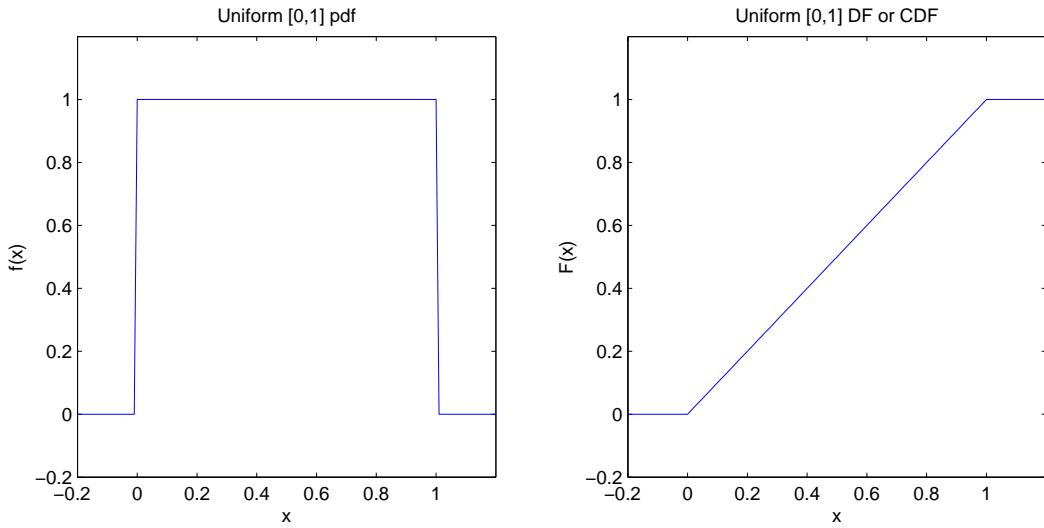
Model 3 (The Fundamental Model) The probability density function (PDF) of the fundamental model or the Uniform(0, 1) RV is

$$f(x) = \mathbb{1}_{[0,1]}(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise} \end{cases} \quad (3.6)$$

and its distribution function (DF) or cumulative distribution function (CDF) is:

$$F(x) := \int_{-\infty}^x f(y) dy = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x > 1 \end{cases} \quad (3.7)$$

Note that the DF is the identity map in $[0, 1]$. The PDF and DF are depicted in Figure 3.4.

Figure 3.4: A plot of the PDF and DF or CDF of the Uniform(0, 1) continuous RV X .****tossing a fair coin infinitely often and the fundamental model**

- The fundamental model is equivalent to infinite tosses of a fair coin (see using binary expansion of any $x \in (0, 1)$)
- The fundamental model has infinitely many copies of itself within it!

****universality of the fundamental model**

- one can obtain any other random object from the fundamental model!

3.4 Expectations

It is convenient to summarise a RV by a single number. This single number can be made to represent some average or expected feature of the RV via an integral with respect to the density of the RV.

Definition 21 (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

$$\mathbf{E}(X) := \int x dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous,} \end{cases} \quad (3.8)$$

provided the sum or integral is well-defined. We say the expectation exists if

$$\int |x| dF(x) < \infty. \quad (3.9)$$

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 $\theta \in \Theta$ with respect to which the integration is undertaken.

$$\mathbf{E}_\theta X := \int x dF(x; \theta)$$

Definition 22 (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

$$\mathbf{V}(X) := \mathbf{E}((X - \mathbf{E}(X))^2) = \int (x - \mathbf{E}(X))^2 dF(x),$$

provided this expectation exists. The **standard deviation** denoted by $\text{sd}(X) := \sqrt{\mathbf{V}(X)}$. Thus variance is a measure of “spread” of a distribution.

Definition 23 (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:

$$\mathbf{E}((X - \mathbf{E}(X))^k).$$

Properties of Expectations

1. If the k -th moment exists and if $j < k$ then the j -th moment exists.
2. If X_1, X_2, \dots, X_n are RVs and a_1, a_2, \dots, a_n are constants, then

$$\mathbf{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mathbf{E}(X_i). \quad (3.10)$$

3. Let X_1, X_2, \dots, X_n be independent RVs, then

$$\mathbf{E} \left(\prod_{i=1}^n X_i \right) = \prod_{i=1}^n \mathbf{E}(X_i) . \quad (3.11)$$

4. $\mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2$. [prove by completing the square and applying (3.10)]

5. If a and b are constants then:

$$\mathbf{V}(aX + b) = a^2 \mathbf{V}(X) . \quad (3.12)$$

6. If X_1, X_2, \dots, X_n are independent and a_1, a_2, \dots, a_n are constants, then:

$$\mathbf{V} \left(\sum_{i=1}^n a_i X_i \right) = \sum_{i=1}^n a_i^2 \mathbf{V}(X_i) . \quad (3.13)$$

Mean and variance of Bernoulli(θ) RV: Let $X \sim \text{Bernoulli}(\theta)$. Then,

$$\begin{aligned} \mathbf{E}(X) &= \sum_{x=0}^1 x f(x) = (0 \times (1 - \theta)) + (1 \times \theta) = 0 + \theta = \theta , \\ \mathbf{E}(X^2) &= \sum_{x=0}^1 x^2 f(x) = (0^2 \times (1 - \theta)) + (1^2 \times \theta) = 0 + \theta = \theta , \\ \mathbf{V}(X) &= \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \theta - \theta^2 = \theta(1 - \theta) . \end{aligned}$$

Parameter specifically,

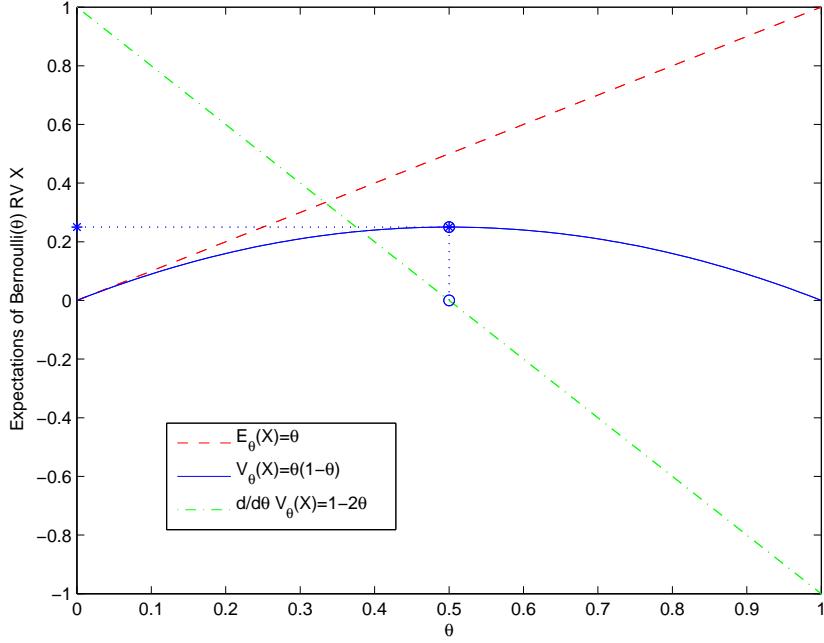
$$\mathbf{E}_\theta(X) = \theta \quad \text{and} \quad \mathbf{V}_\theta(X) = \theta(1 - \theta) .$$

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:

$$\begin{aligned} \mathbf{V}'_\theta(X) &:= \frac{d}{d\theta} \mathbf{V}_\theta(X) = 1 - 2\theta = 0 \iff \theta = \frac{1}{2} , & \mathbf{V}''_\theta(X) &:= \frac{d}{d\theta} \left(\frac{d}{d\theta} \mathbf{V}_\theta(X) \right) = -2 < 0 , \\ \max_{\theta \in [0,1]} \mathbf{V}_\theta(X) &= \frac{1}{2} \left(1 - \frac{1}{2} \right) = \frac{1}{4} , \text{ since } \mathbf{V}_\theta(X) \text{ is maximized at } \theta = \frac{1}{2} \end{aligned}$$

The plot depicting these expectations as well as the rate of change of the variance are depicted in Figure 3.5. 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.

Figure 3.5: 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(θ) RV X as a function of the parameter θ .



Mean and variance of Uniform(0, 1) RV: Let $X \sim \text{Uniform}(0, 1)$. Then,

$$\begin{aligned} \mathbf{E}(X) &= \int_{x=0}^1 x f(x) dx = \int_{x=0}^1 x 1 dx = \frac{1}{2} (x^2) \Big|_{x=0}^{x=1} = \frac{1}{2} (1 - 0) = \frac{1}{2}, \\ \mathbf{E}(X^2) &= \int_{x=0}^1 x^2 f(x) dx = \int_{x=0}^1 x^2 1 dx = \frac{1}{3} (x^3) \Big|_{x=0}^{x=1} = \frac{1}{3} (1 - 0) = \frac{1}{3}, \\ \mathbf{V}(X) &= \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}. \end{aligned}$$

Proposition 24 (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 35 (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

$$\mathbf{E}(\mathbb{1}_A(X)) = \int \mathbb{1}_A(x) dF(x) = \int_A f(x) dx = \mathbf{P}(X \in A) = \mathbf{P}(A) \quad (3.14)$$

Thus, probability is a special case of expectation. Recall our LTRF motivation for the definition of probability and make the connection.

3.5 Stochastic Processes

Definition 25 (Independence of RVs) A finite or infinite sequence of RVs X_1, X_2, \dots is said to be independent or independently distributed if

$$\mathbf{P}(X_{i_1} \leq x_{i_1}, X_{i_2} \leq x_{i_2}, \dots, X_{i_k} \leq x_{i_k}) = \mathbf{P}(X_{i_1} \leq x_{i_1})\mathbf{P}(X_{i_2} \leq x_{i_2}) \cdots \mathbf{P}(X_{i_k} \leq x_{i_k})$$

for any distinct subset $\{i_1, i_2, \dots, i_l\}$ of indices of the sequence of RVs and any sequence of real numbers $x_{i_1}, x_{i_2}, \dots, x_{i_k}$.

By the above definition, the sequence of **discrete** RVs X_1, X_2, \dots taking values in an at most countable set \mathbb{D} are said to be independently distributed if for any distinct subset of indices $\{i_1, i_2, \dots, i_k\}$ such that the corresponding RVs $X_{i_1}, X_{i_2}, \dots, X_{i_k}$ exists as a distinct subset of our original sequence of RVs X_1, X_2, \dots and for any elements $x_{i_1}, x_{i_2}, \dots, x_{i_k}$ in \mathbb{D} , the following equality is satisfied:

$$\mathbf{P}(X_{i_1} = x_{i_1}, X_{i_2} = x_{i_2}, \dots, X_{i_k} = x_{i_k}) = \mathbf{P}(X_{i_1} = x_{i_1})\mathbf{P}(X_{i_2} = x_{i_2}) \cdots \mathbf{P}(X_{i_k} = x_{i_k})$$

For an independent sequence of RVs $\{X_1, X_2, \dots\}$, we have

$$\begin{aligned} & \mathbf{P}(X_{i+1} \leq x_{i+1} | X_i \leq x_i, X_{i-1} \leq x_{i-1}, \dots, X_1 \leq x_1) \\ &= \frac{\mathbf{P}(X_{i+1} \leq x_{i+1}, X_i \leq x_i, X_{i-1} \leq x_{i-1}, \dots, X_1 \leq x_1)}{\mathbf{P}(X_i \leq x_i, X_{i-1} \leq x_{i-1}, \dots, X_1 \leq x_1)} \\ &= \frac{\mathbf{P}(X_{i+1} \leq x_{i+1})\mathbf{P}(X_i \leq x_i)\mathbf{P}(X_{i-1} \leq x_{i-1}) \cdots \mathbf{P}(X_1 \leq x_1)}{\mathbf{P}(X_i \leq x_i)\mathbf{P}(X_{i-1} \leq x_{i-1}) \cdots \mathbf{P}(X_1 \leq x_1)} \\ &= \mathbf{P}(X_{i+1} \leq x_{i+1}) \end{aligned}$$

The above equality that

$$\mathbf{P}(X_{i+1} \leq x_{i+1} | X_i \leq x_i, X_{i-1} \leq x_{i-1}, \dots, X_1 \leq x_1) = \mathbf{P}(X_{i+1} \leq x_{i+1})$$

simply says that the conditional distribution of the RV X_{i+1} given all previous RVs X_i, X_{i-1}, \dots, X_1 is simply determined by the distribution of X_{i+1} .

When a sequence of RVs are not independent they are said to be **dependent**.

Definition 26 (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

$$(X_i)_{i \in \mathbb{Z}} := \dots, X_{-2}, X_{-1}, X_0, X_1, X_2, \dots, \text{ or}$$

$$(X_i)_{i \in \mathbb{N}} := X_1, X_2, \dots, \text{ or}$$

$$(X_i)_{i \in [n]} := X_1, X_2, \dots, X_n, \text{ where, } [n] := \{1, 2, \dots, n\} .$$

If $\mathbb{A} \subset \mathbb{R}$ then we have a **continuous time stochastic process**, typically denoted by $\{X_t\}_{t \in \mathbb{R}}$, etc.

Definition 27 (Independent and Identically Distributed (IID)) The finite or infinite sequence of RVs or the stochastic process X_1, X_2, \dots is said to be independent and identically distributed or IID if :

- they are independently distributed according to Definition 25, and
- $F(X_1) = F(X_2) = \dots$, 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

$$(X_i)_{i \in [n]} := X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} F, \quad \text{or} \quad (X_i)_{i \in \mathbb{N}} := X_1, X_2, \dots \stackrel{\text{IID}}{\sim} F.$$

We sometimes replace the DF F above by the name of the RV.

Definition 28 (Independently Distributed) The sequence of RVs or the stochastic process $(X_i)_{i \in \mathbb{N}} := X_1, X_2, \dots$ is said to be independently distributed if :

- X_1, X_2, \dots is independently distributed according to Definition 25.

This is a class of stochastic processes that is more general than the IID class.

Chapter 4

Random Numbers

“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($1/k, \dots, 1/k$) 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 $\{1, 2, \dots, 39, 40\}$ 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.

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 1 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:* $(x_0, x_1, \dots, x_{n-1})$, the linear congruential sequence of length n

3: **for** $i = 1$ to $n - 1$ **do**

4: $x_i \leftarrow (ax_{i-1} + c) \bmod m$

5: **end for**

6: *return:* (x_1, x_2, \dots, 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 36 (Generic Linear Congruential Sequence) Let us implement Algorithm 1 in MATLAB as follows.

```
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);
```

```
x(i) = double(X); % convert to double
end
```

We can call it for some arbitrary input arguments as follows:

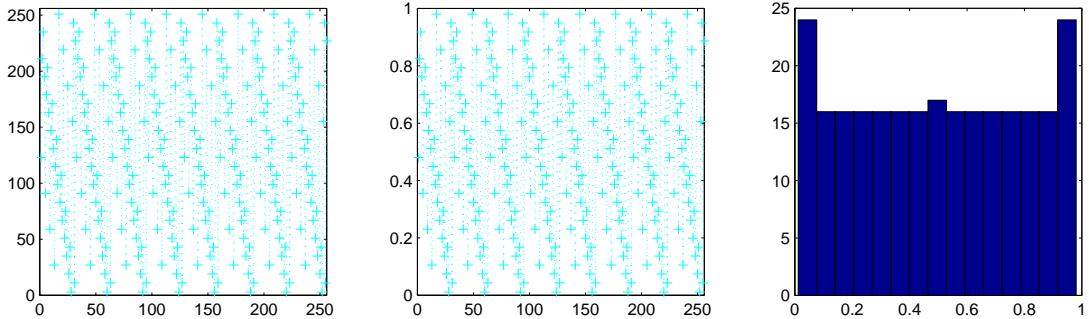
```
>> LinConGen(13,12,11,10,12)
ans =
    10     1     10     1     10     1     10     1     10     1     10     1
>> LinConGen(13,10,9,8,12)
ans =
     8     11     2     3     0     9     8     11     2     3     0     9
```

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 37 (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,'+')
axis([0 256 0 256]); axis square
LCGSeqIn01=LCGSeq ./ 256
subplot(1,3,2)
plot(LCGSeqIn01,'+')
axis([0 256 0 1]); axis square
subplot(1,3,3)
hist(LCGSeqIn01,15)
axis square
```

Figure 4.1: The linear congruential sequence of $\text{LinConGen}(256, 137, 0, 123, 257)$ with non-maximal period length of 32 as a line plot over $\{0, 1, \dots, 256\}$, 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*:

$$x_i = f(x_{i-1}), \quad f : \{0, 1, 2, \dots, m-1\} \rightarrow \{0, 1, 2, \dots, m-1\} \text{ and } f(x_{i-1}) = (ax_{i-1} + c) \pmod{m}.$$

Since f maps a the finite set $\{1, 2, \dots, m-1\}$ into itself, such systems are bound to have a repeating cycle of numbers called the **period**. In Labwork 36, 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 38 (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 37 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 37? 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 39 (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`.

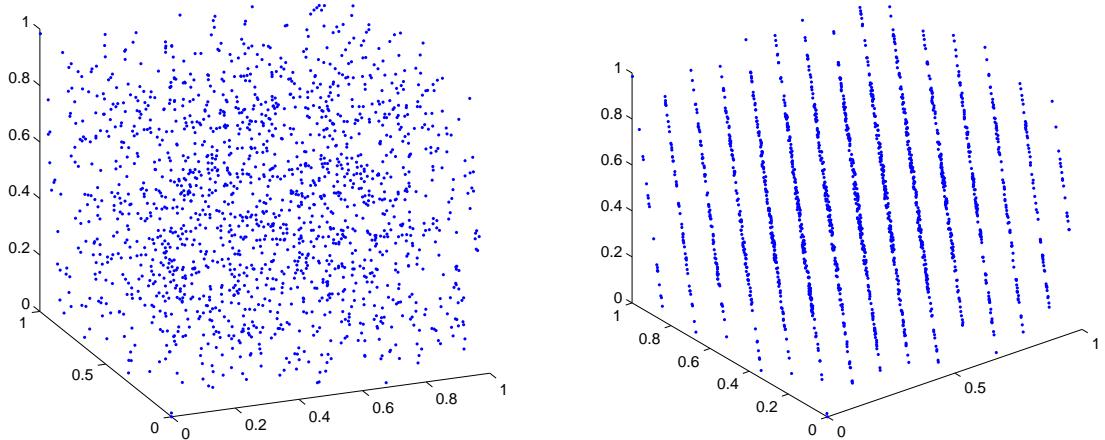
```
>> x=reshape( (LinConGen(2147483648,65539,0,1,5001)./ 2147483648) ,3,[]);
>> plot3(x(1,:),x(2,:),x(3,:),'.'
```

Labwork 40 (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
```

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 41 (Calling PRNG in MATLAB) In MATLAB `rand` is basic PRNG command.

```
>> 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 42 (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:

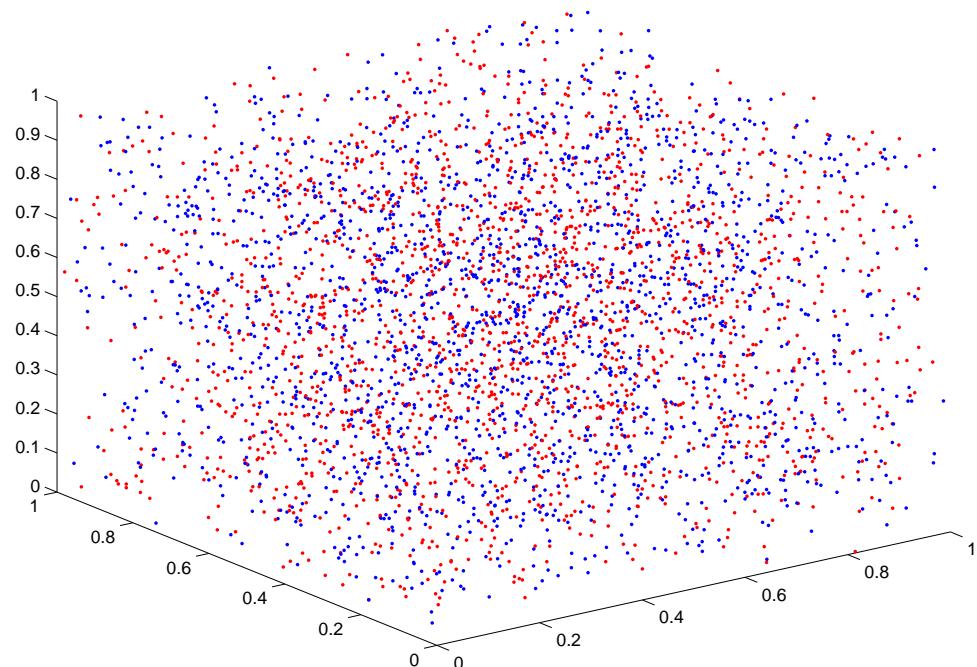
```
>> rand('twister',1234)
>> x=rand(3,2000); % store PRNs in a 3X2000 matrix named x
>> plot3(x(1,:),x(2,:),x(3,:),'.'
```

Compare this with the 3D plot of triplets from RANDU of Labwork 39. 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
```

Figure 4.3: Triplet point clouds from the “Mersenne Twister” with two different seeds (see Lab-work 42). .



Chapter 5

Statistics

5.1 Data and Statistics

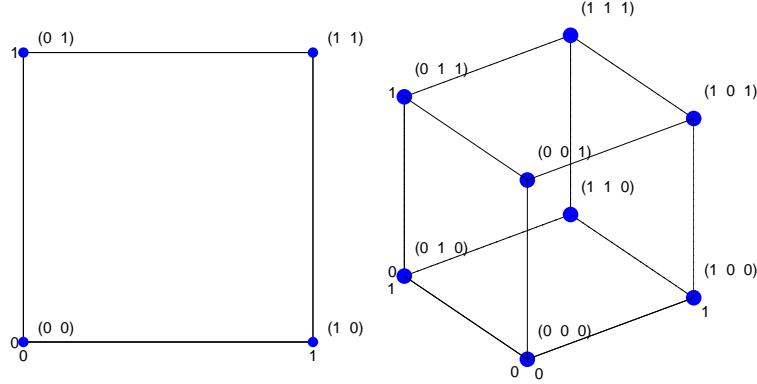
Definition 29 (Data) The function X measures the outcome ω of an experiment with sample space Ω [Often, the sample space is also denoted by S]. Formally, X is a random variable [or a random vector $X = (X_1, X_2, \dots, X_n)$, i.e. a vector of random variables] taking values in the **data space** \mathbb{X} :

$$X(\omega) : \Omega \rightarrow \mathbb{X}.$$

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 .

Figure 5.1: Sample Space, Random Variable, Realisation, Data, and Data Space.

Example 43 (Tossing a coin n times) For some given parameter $\theta \in \Theta := [0, 1]$, consider n IID Bernoulli(θ) trials, i.e. $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta)$. Then the random vector $X = (X_1, X_2, \dots, X_n)$, which takes values in the data space $\mathbb{X} = \{0, 1\}^n := \{(x_1, x_2, \dots, x_n) : x_i \in \{0, 1\}, i = 1, 2, \dots, n\}$, 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, \dots, x_n) . For instance, if we observed $n - 1$ tails and 1 heads, in that order, then our data vector $(x_1, x_2, \dots, x_{n-1}, x_n) = (0, 0, \dots, 0, 1)$.

Figure 5.2: Data Spaces $\mathbb{X} = \{0, 1\}^2$ and $\mathbb{X} = \{0, 1\}^3$ for two and three Bernoulli trials, respectively.

Definition 30 (Statistic) A **statistic** T is any function of the data:

$$T(x) : \mathbb{X} \rightarrow \mathbb{T} .$$

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 44 (Is data a statistic?) Is the RV X , for which the realisation is the observed data $X(\omega) = x$, a statistic? In other words, is the data a statistic? [Hint: consider the identity map $T(x) = x : \mathbb{X} \rightarrow \mathbb{T} = \mathbb{X}$.]

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 31 (Sample Mean) From a given a sequence of RVs X_1, X_2, \dots, X_n , we may obtain another RV called the n -samples mean or simply the sample mean:

$$T_n((X_1, X_2, \dots, X_n)) = \bar{X}_n((X_1, X_2, \dots, X_n)) := \frac{1}{n} \sum_{i=1}^n X_i . \quad (5.1)$$

For brevity, we write

$$\bar{X}_n((X_1, X_2, \dots, X_n)) \quad \text{as} \quad \bar{X}_n ,$$

and its realisation

$$\bar{X}_n((x_1, x_2, \dots, x_n)) \quad \text{as} \quad \bar{x}_n .$$

Note that the expectation and variance of \bar{X}_n are:

$$\begin{aligned} \mathbf{E}(\bar{X}_n) &= \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) && [\text{by definition (5.1)}] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{E}(X_i) && [\text{by property (3.10)}] \end{aligned}$$

Furthermore, if every X_i in the original sequence of RVs X_1, X_2, \dots is **identically** distributed with the same expectation, by convention $\mathbf{E}(X_1)$, then:

$$\mathbf{E}(\bar{X}_n) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}(X_i) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}(X_1) = \frac{1}{n} n \mathbf{E}(X_1) = \mathbf{E}(X_1) . \quad (5.2)$$

Similarly, we can show that:

$$\begin{aligned} \mathbf{V}(\bar{X}_n) &= \mathbf{V}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) && [\text{by definition (5.1)}] \\ &= \left(\frac{1}{n}\right)^2 \mathbf{V}\left(\sum_{i=1}^n X_i\right) && [\text{by property (3.12)}] \end{aligned}$$

Furthermore, if the original sequence of RVs X_1, X_2, \dots is **independently** distributed then:

$$\mathbf{V}(\bar{X}_n) = \left(\frac{1}{n}\right)^2 \mathbf{V}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \mathbf{V}(X_i) \quad [\text{by property (3.13)}]$$

Finally, if the original sequence of RVs X_1, X_2, \dots is **independently and identically** distributed with the same variance ($\mathbf{V}(X_1)$ by convention) then:

$$\mathbf{V}(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^n \mathbf{V}(X_i) = \frac{1}{n^2} \sum_{i=1}^n \mathbf{V}(X_1) = \frac{1}{n^2} n \mathbf{V}(X_1) = \frac{1}{n} \mathbf{V}(X_1) . \quad (5.3)$$

Labwork 45 (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 \bar{x}_n of n sample points x_1, x_2, \dots, 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 32 (Sample Variance & Standard Deviation) From a given a sequence of random variables X_1, X_2, \dots, X_n , we may obtain another statistic called the n -samples variance or simply the sample variance :

$$T_n((X_1, X_2, \dots, X_n)) = S_n^2((X_1, X_2, \dots, X_n)) := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 . \quad (5.4)$$

For brevity, we write $S_n^2((X_1, X_2, \dots, X_n))$ as S_n^2 and its realisation $S_n^2((x_1, x_2, \dots, x_n))$ as s_n^2 .

Sample standard deviation is simply the square root of sample variance:

$$S_n((X_1, X_2, \dots, X_n)) = \sqrt{S_n^2((X_1, X_2, \dots, X_n))} \quad (5.5)$$

For brevity, we write $S_n((X_1, X_2, \dots, X_n))$ as S_n and its realisation $S_n((x_1, x_2, \dots, x_n))$ as s_n .

Once again, if $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} X_1$, the expectation of the sample variance is:

$$\mathbf{E}(S_n^2) = \mathbf{V}(X_1) .$$

Labwork 46 (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 45 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 33 (Order Statistics) Suppose $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} F$, where F is the DF from the set of all DFs over the real line. Then, the n -sample **order statistics** $X_{([n])}$ is:

$$X_{([n])}((X_1, X_2, \dots, X_n)) := (X_{(1)}, X_{(2)}, \dots, X_{(n)}) , \text{ such that, } X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)} . \quad (5.6)$$

For brevity, we write $X_{([n])}((X_1, X_2, \dots, X_n))$ as $X_{([n])}$ and its realisation $X_{([n])}((x_1, x_2, \dots, x_n))$ as $x_{([n])} = (x_{(1)}, x_{(2)}, \dots, x_{(n)})$.

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 47 (Order statistics and sorting) The order statistics for the five samples stored in `XsFromUni01Twstr101` from Labwork 45 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 $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ from n sample points x_1, x_2, \dots, x_n .

Next, we will introduce a family of common statistics, called the q^{th} quantile, by first defining the function:

Definition 34 (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:

$$F^{[-1]}(q) := \inf \{x : F(x) > q\}, \quad \text{for some } q \in [0, 1] . \quad (5.7)$$

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, $T(F) : \{\text{All DFs}\} \rightarrow \mathbb{T}$, 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:

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 $X \sim F$:

$$T(F) = F^{[-1]}(q) \quad \text{where } q \in [0, 1] .$$

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 35 (Empirical Distribution Function (EDF or ECDF)) Suppose we have n IID RVs, $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} F$, 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 \hat{F}_n that puts a probability mass of $1/n$ at each sample or data point x_i :

$$\hat{F}_n(x) = \frac{\sum_{i=1}^n \mathbf{1}(X_i \leq x)}{n} , \quad \text{where} \quad \mathbf{1}(X_i \leq x) := \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > x \end{cases} \quad (5.8)$$

Labwork 48 (Plot of empirical CDF) Let us plot the ECDF for the five samples drawn from the $\text{Uniform}(0, 1)$ RV in Labwork 45 using the MATLAB function ECDF (given in Labwork ??). Let us super-impose the samples and the true DF as depicted in Figure 5.3 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.
```

Definition 36 (q^{th} Sample Quantile) For some $q \in [0, 1]$ and n IID RVs $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} F$, we can obtain the ECDF \hat{F}_n using (5.8). The q^{th} **sample quantile** is defined as the statistic (statistical functional):

$$T(\hat{F}_n) = \hat{F}_n^{[-1]}(q) := \inf \{x : \hat{F}_n^{[-1]}(x) \geq q\} . \quad (5.9)$$

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.

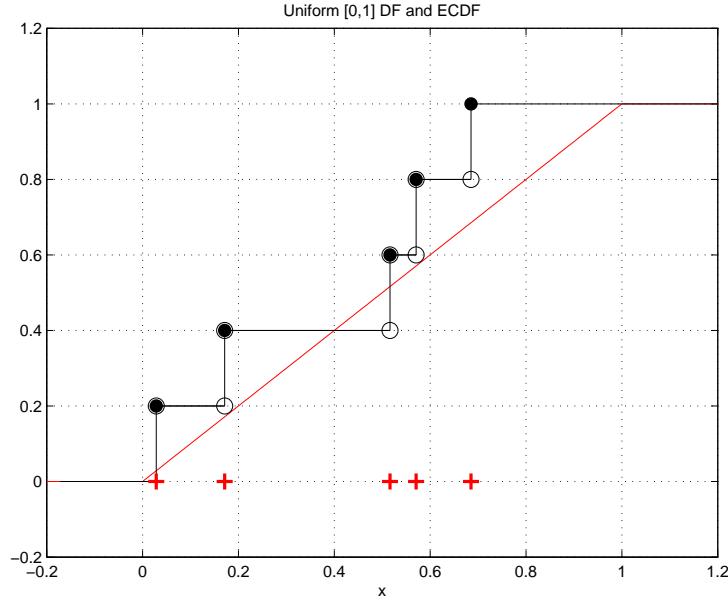
The following algorithm can be used to obtain the q^{th} sample quantile of n IID samples (x_1, x_2, \dots, x_n) on the basis of their order statistics $(x_{(1)}, x_{(2)}, \dots, x_{(n)})$.

The q^{th} sample quantile, $\hat{F}_n^{[-1]}(q)$, is found by interpolation from the order statistics $(x_{(1)}, x_{(2)}, \dots, x_{(n)})$ of the n data points (x_1, x_2, \dots, x_n) , using the formula:

$$\hat{F}_n^{[-1]}(q) = (1 - \delta)x_{(i+1)} + \delta x_{(i+2)}, \quad \text{where,} \quad i = \lfloor (n-1)q \rfloor \quad \text{and} \quad \delta = (n-1)q - \lfloor (n-1)q \rfloor .$$

Thus, the **sample minimum** of the data points (x_1, x_2, \dots, x_n) is given by $\hat{F}_n^{[-1]}(0)$, the **sample maximum** is given by $\hat{F}_n^{[-1]}(1)$ and the **sample median** is given by $\hat{F}_n^{[-1]}(0.5)$, etc.

Figure 5.3: Plot of the DF of Uniform(0, 1), five IID samples from it, and the ECDF \hat{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.



Algorithm 2 q^{th} Sample Quantile of Order Statistics

1: *input:*

1. q in the q^{th} sample quantile, i.e. the argument q of $\hat{F}_n^{[-1]}(q)$,
 2. order statistic $(x_{(1)}, x_{(2)}, \dots, x_{(n)})$, i.e. the sorted (x_1, x_2, \dots, x_n) , where $n > 0$.
- 2: *output:* $\hat{F}_n^{[-1]}(q)$, the q^{th} sample quantile
- 3: $i \leftarrow \lfloor (n-1)q \rfloor$
 - 4: $\delta \leftarrow (n-1)q - i$
 - 5: **if** $i = n-1$ **then**
 - 6: $\hat{F}_n^{[-1]}(q) \leftarrow x_{(i+1)}$
 - 7: **else**
 - 8: $\hat{F}_n^{[-1]}(q) \leftarrow (1-\delta)x_{(i+1)} + \delta x_{(i+2)}$
 - 9: **end if**
- 10: *return:* $\hat{F}_n^{[-1]}(q)$
-

Labwork 49 (The q^{th} sample quantile) Use the implementation of Algorithm 2 in Labwork ?? 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 47 and
2. Another sorted array of 7 samples called `SortedXs`

```
>> 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
```

5.2 Exploring Data and Statistics

5.2.1 Univariate Data

A **histogram** is a graphical representation of the frequency with which elements of a data array:

$$x = (x_1, x_2, \dots, x_n) ,$$

of real numbers fall within each of the m intervals or **bins** of some **interval partition**:

$$b := (b_1, b_2, \dots, b_m) := ([\underline{b}_1, \bar{b}_1], [\underline{b}_2, \bar{b}_2], \dots, [\underline{b}_m, \bar{b}_m])$$

of the **data range** of x given by the closed interval:

$$\mathcal{R}(x) := [\min\{x_1, x_2, \dots, x_n\}, \max\{x_1, x_2, \dots, x_n\}] .$$

Elements of this partition b are called bins, their mid-points are called **bin centres**:

$$c := (c_1, c_2, \dots, c_m) := ((\underline{b}_1 + \bar{b}_1)/2, (\underline{b}_2 + \bar{b}_2)/2, \dots, (\underline{b}_m + \bar{b}_m)/2)$$

and their overlapping boundaries, i.e. $\bar{b}_i = \underline{b}_{i+1}$ for $1 \leq i < m$, are called **bin edges**:

$$d := (d_1, d_2, \dots, d_{m+1}) := (\underline{b}_1, \underline{b}_2, \dots, \underline{b}_{m-1}, \underline{b}_m, \bar{b}_m) .$$

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 $i = 1, 2, \dots, m$. Thus, a histogram can be obtained by a set of bins along with their corresponding **heights**:

$$h = (h_1, h_2, \dots, h_m) , \text{ where } h_k := g(\#\{x_i : x_i \in b_k\})$$

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:

$$((b_1, h_1), (b_2, h_2), \dots, (b_m, h_m)) .$$

Given a partition b , a **frequency histogram** is the histogram:

$$((b_1, h_1), (b_2, h_2), \dots, (b_m, h_m)) , \text{ where } h_k := \#\{x_i : x_i \in b_k\} ,$$

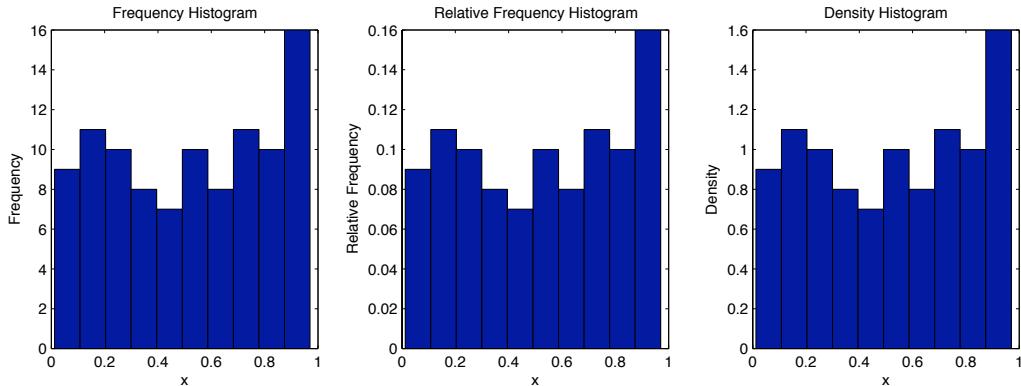
a **relative frequency histogram** is the histogram:

$$((b_1, h_1), (b_2, h_2), \dots, (b_m, h_m)) , \text{ where } h_k := n^{-1} \#\{x_i : x_i \in b_k\} ,$$

and a **density histogram** is the histogram:

$$((b_1, h_1), (b_2, h_2), \dots, (b_m, h_m)) , \text{ where } h_k := (w_k n)^{-1} \#\{x_i : x_i \in b_k\} , w_k := \bar{b}_k - \underline{b}_k .$$

Figure 5.4: Frequency, Relative Frequency and Density Histograms



Labwork 50 (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 histograms as shown in Figure 5.4 as follows:

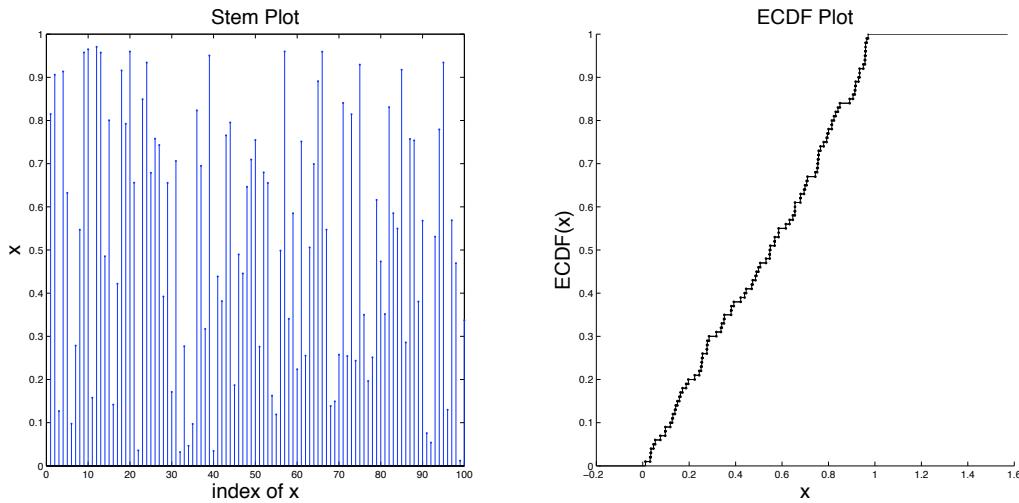
```
>> rand('twister',5489);
>> x=rand(1,100); % generate 100 PRNs
>> hist(x) % see what default hist does in Figure Window
>> % Now let us look deeper into the last hist call
>> [Fs, Cs] = hist(x) % Cs is the bin centers and Fs is the frequencies of data set x
Fs =
    9     11     10      8      7     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
```

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 51 (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 5.5 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 5.5: 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 .

5.2.2 Bivariate Data

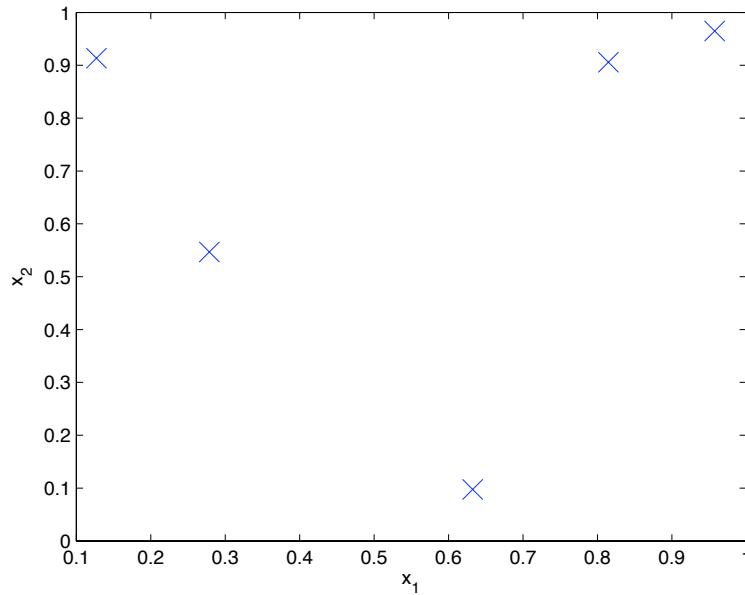
By bivariate data array x we mean a $2 \times n$ matrix of real numbers or equivalently n ordered pairs of points $(x_{1,i}, x_{2,i})$ as $i = 1, 2, \dots, n$. The most elementary visualisation of these n ordered pairs is in orthogonal Cartesian co-ordinates. Such plots are termed **2D scatter plots** in statistics.

Labwork 52 (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 5.6 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
    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
```

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.

Figure 5.6: 2D Scatter Plot



5.2.3 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 53 (Visualising trivariate data) We can make **3D scatter plots** as shown in Figure 5.7 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
```

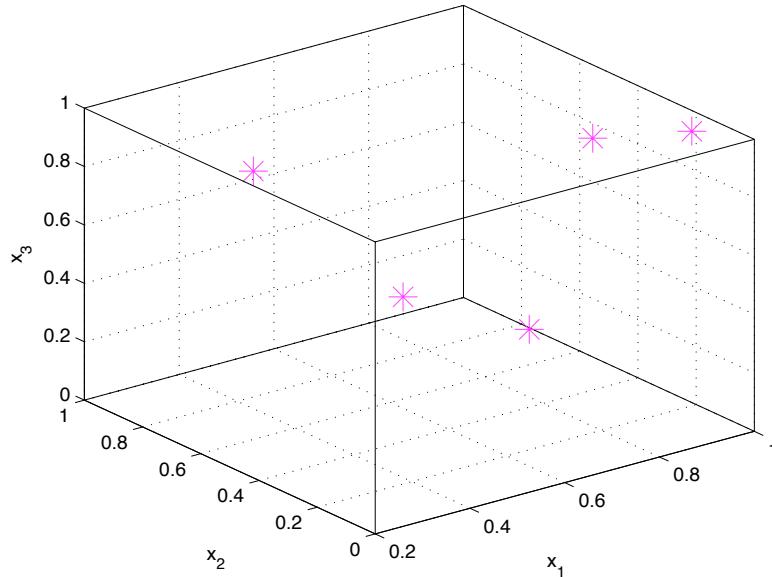
Repeat the visualisation below with a larger array, say $x=\text{rand}(3,1000)$, and use the rotate 3D feature in the Figure window to visually explore the samples in the unit cube. Do they seem to be uniformly distributed inside the unit cube?

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.

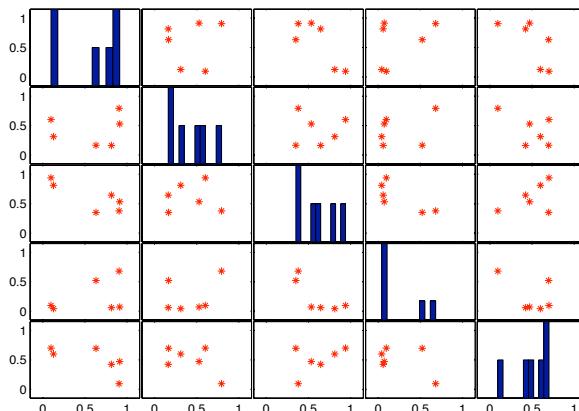
5.2.4 Multivariate Data

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

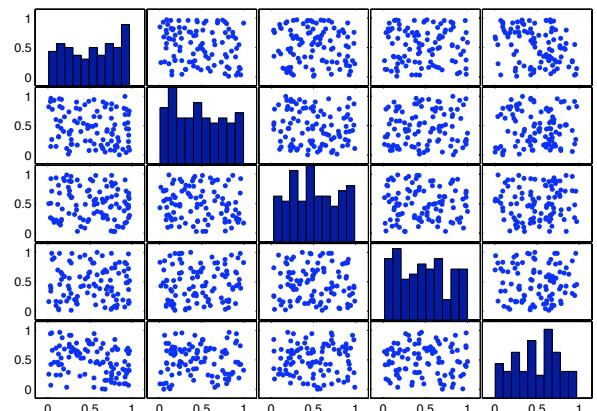
Figure 5.7: 3D Scatter Plot



co-ordinates $\{(i, j) \in \{1, 2, \dots, d\}^2 : i \neq j\}$ and at histograms for every co-ordinate $i \in \{1, 2, \dots, d\}$ 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 5.8: Plot Matrix of uniformly generated data in $[0, 1]^5$ 

(a) First six samples



(b) All thousand samples

Labwork 54 Let us make matrix plots from a uniformly generated sequence of 100 points in 5D unit cube $[0, 1]^5$ as shown in Figure 5.8.

```
>> 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

```

5.3 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*].

5.3.1 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 55.

Labwork 55 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:

```

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 description 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 5.9 (the axes labels were subsequently added from clicking <Edit> and <Figure Properties...> tabs of the output Figure Window).

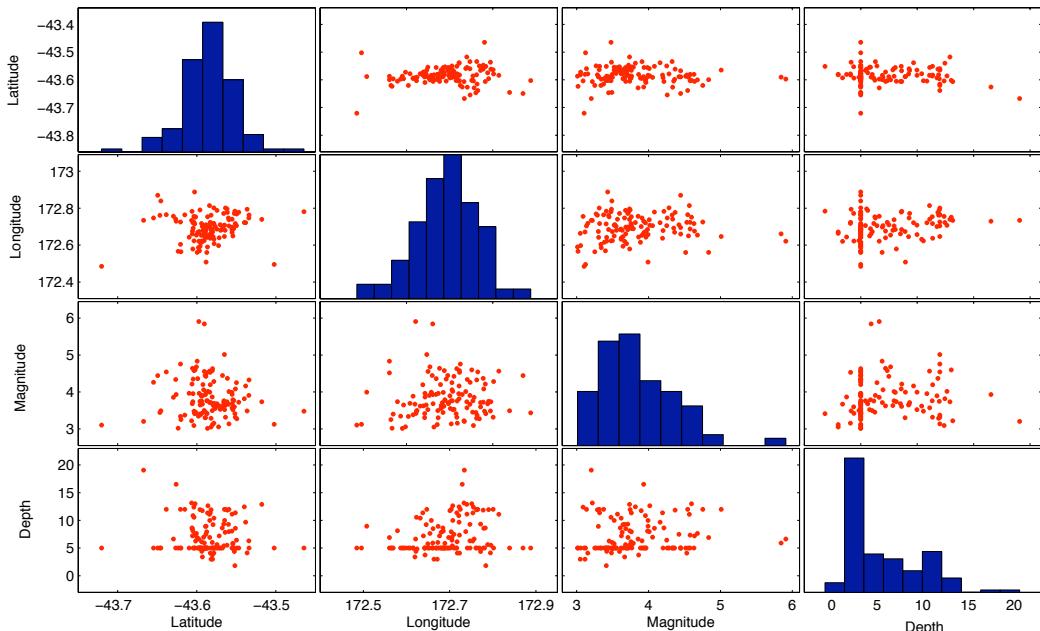
```
>> EQall(any(isnan(EQall),2),:) = []; %Remove any rows containing NaNs from the matrix EQall
>> % report the size of EQall and see if it is different from before we removed and NaN containing rows
>> size(EQall)
ans = 145 14
>> % remove locations outside Chch and assign it to a new variable called 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
>> 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` in Labwork ?? 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 56.

Figure 5.9: Matrix of Scatter Plots of the latitude, longitude, magnitude and depth of the 22-02-2011 earth quakes in Christchurch, New Zealand.



Labwork 56 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
```

```
----- 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)

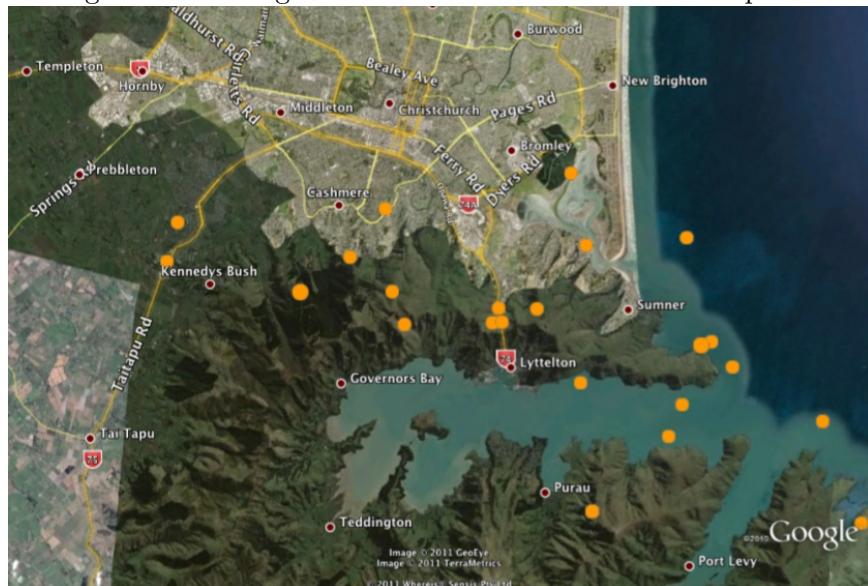
figure % tell matlab you are about to make another figure
semilogx(DepData,MagData,'.') % see the depth in log scale

%%%%%
% 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

Figure 5.10: Google Earth Visualisation of the earth quakes



A global search at <http://neic.usgs.gov/cgi-bin/epic/epic.cgi> with the following parame-

ters:

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 5.10. One can do a lot more than a mere visualisation with the USGS/NEIC database of earth-quakes worldwide, 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.

5.3.2 Metereological Data

New Zealand's meteorological service NIWA provides weather data under its TERMS AND CONDITIONS FOR ACCESS TO DATA (See http://cliflo.niwa.co.nz/doc/terms_print.html). 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 57 Understand how Figure 5.11 is obtained by the script file `RainFallsInChch.m` by typing and following the comments:

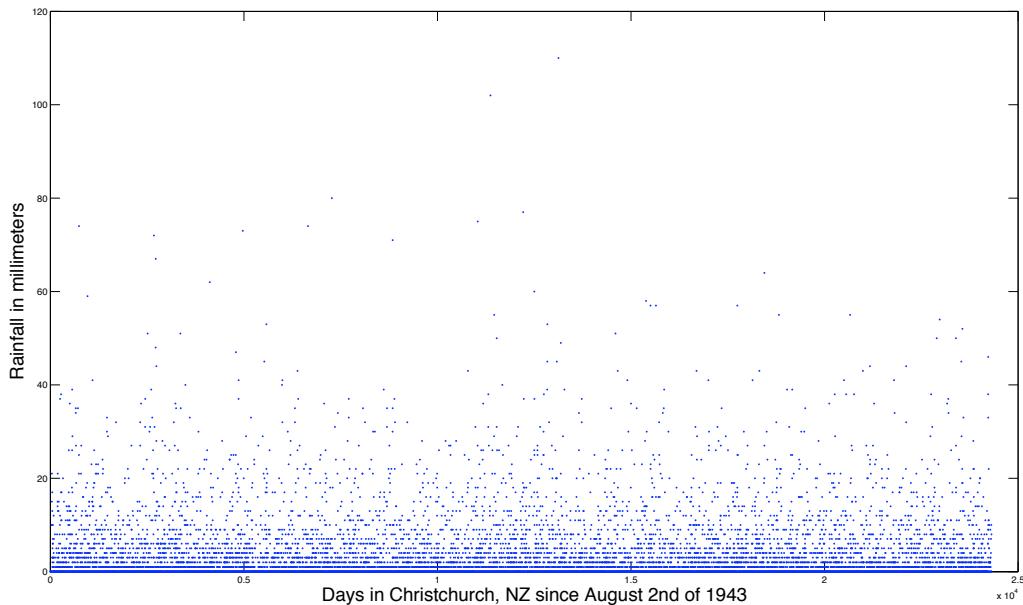
```
>> RainFallsInChch
RainFallsChch = [24312x1 int32] [24312x1 double]
ans = 24312 2
FirstDayOfData = 19430802
LastDayOfData = 20100721
```

```
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)
```

Figure 5.11: Daily rainfalls in Christchurch since March 27 2010



Daily Temperatures in Christchurch

Labwork 58 Understand how Figure 5.12 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!
```

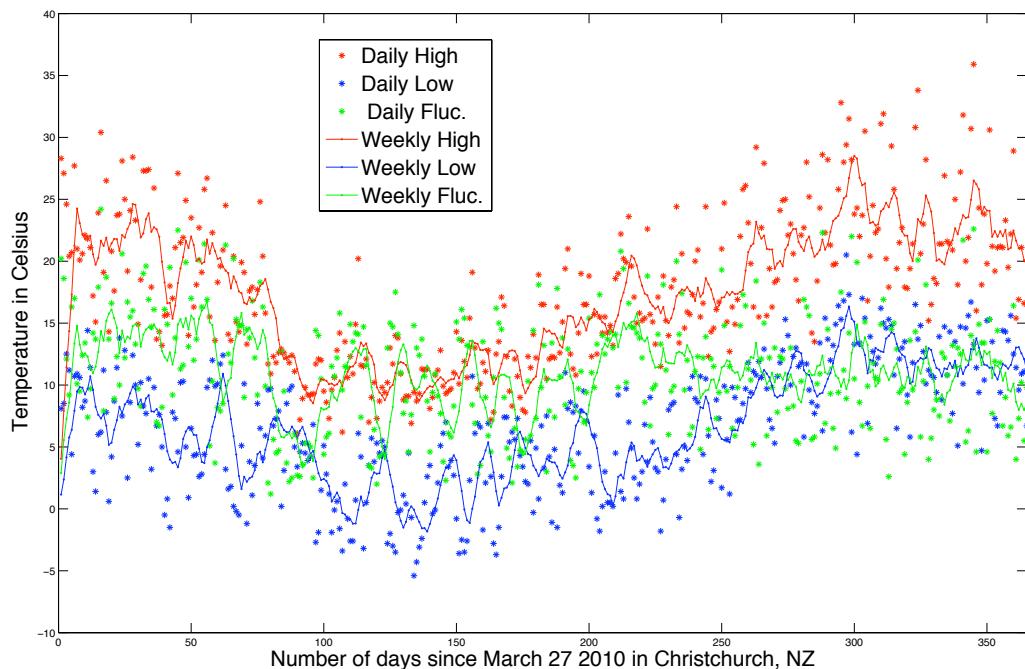
```

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 5.12: Daily temperatures in Christchurch for one year since March 27 2010



5.3.3 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

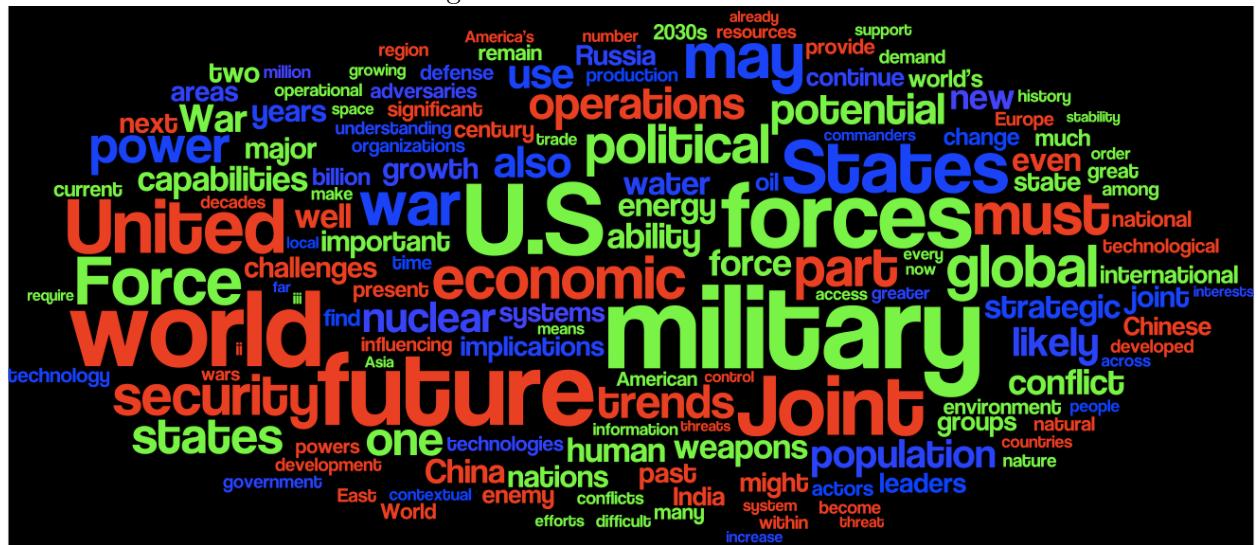
- a large document
 - twitter messages within an online social network of interest
 - etc.

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 Report) 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 5.13: 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 5.13 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.

Labwork 59 (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.

5.3.4 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

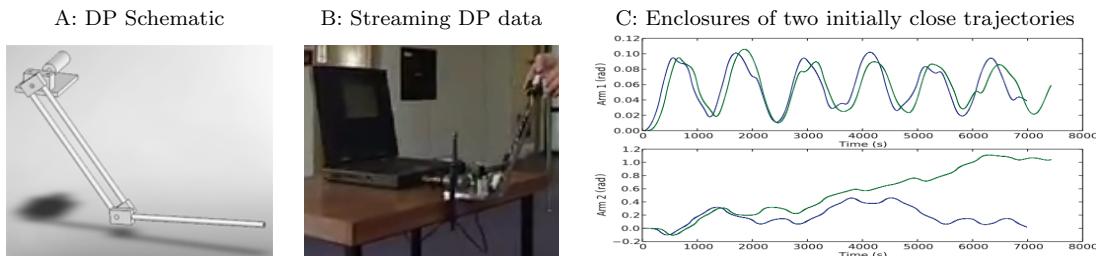


Figure 5.14: Double Pendulum

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 5.14. 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 60 (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 5.14(C) using time series data of trajectory 1 and trajectory 2 linked from the bottom of the above URL.

Chapter 6

Common Random Variables

The Uniform(0, 1) RV of Model 3 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.

6.1 Inversion Sampler for Continuous Random Variables

Proposition 37 (Inversion sampler) Let $F(x) := \int_{-\infty}^x f(y) dy : \mathbb{R} \rightarrow [0, 1]$ be a continuous DF with density f , and let its inverse $F^{[-1]} : [0, 1] \rightarrow \mathbb{R}$ 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 .

Proof: The “one-line proof” of the proposition is due to the following equalities:

$$\mathbf{P}(F^{[-1]}(U) \leq x) = \mathbf{P}(\inf\{y : F(y) = U\} \leq x) = \mathbf{P}(U \leq F(x)) = F(x), \quad \text{for all } x \in \mathbb{R}.$$

This yields the inversion sampler or the inverse (C)DF sampler, where we (i) *generate* $u \sim \text{Uniform}(0, 1)$ and (ii) *return* $x = F^{[-1]}(u)$, as formalised by the following algorithm.

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.

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* $u \sim \text{Uniform}(0, 1)$
 - 5: *return*: $x = F^{[-1]}(u)$
-

6.2 Some Simulations of Continuous Random Variables

6.3 Continuous Random Variables

Model 4 ($\text{Uniform}(\theta_1, \theta_2)$) Given two real parameters $\theta_1, \theta_2 \in \mathbb{R}$, such that $\theta_1 < \theta_2$, the PDF of the $\text{Uniform}(\theta_1, \theta_2)$ RV X is:

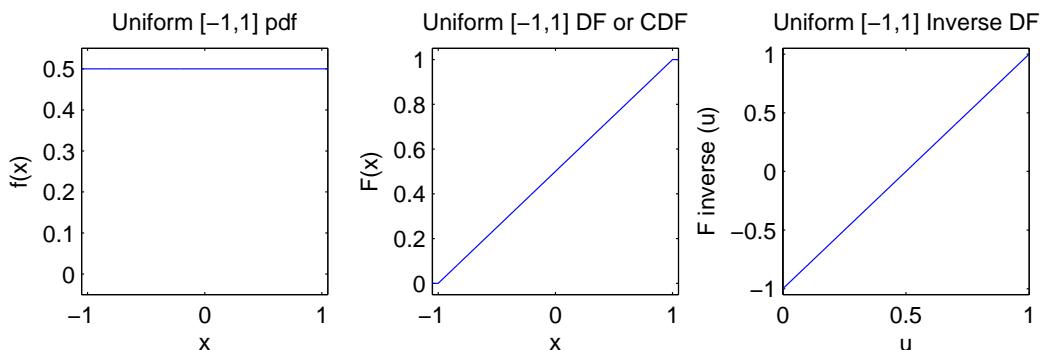
$$f(x; \theta_1, \theta_2) = \begin{cases} \frac{1}{\theta_2 - \theta_1} & \text{if } \theta_1 \leq x \leq \theta_2, \\ 0 & \text{otherwise} \end{cases} \quad (6.1)$$

and its DF given by $F(x; \theta_1, \theta_2) = \int_{-\infty}^x f(y; \theta_1, \theta_2) dy$ is:

$$F(x; \theta_1, \theta_2) = \begin{cases} 0 & \text{if } x < \theta_1 \\ \frac{x - \theta_1}{\theta_2 - \theta_1} & \text{if } \theta_1 \leq x \leq \theta_2, \\ 1 & \text{if } x > \theta_2 \end{cases} \quad (6.2)$$

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 6.1: A plot of the PDF, DF or CDF and inverse DF of the $\text{Uniform}(-1, 1)$ RV X .

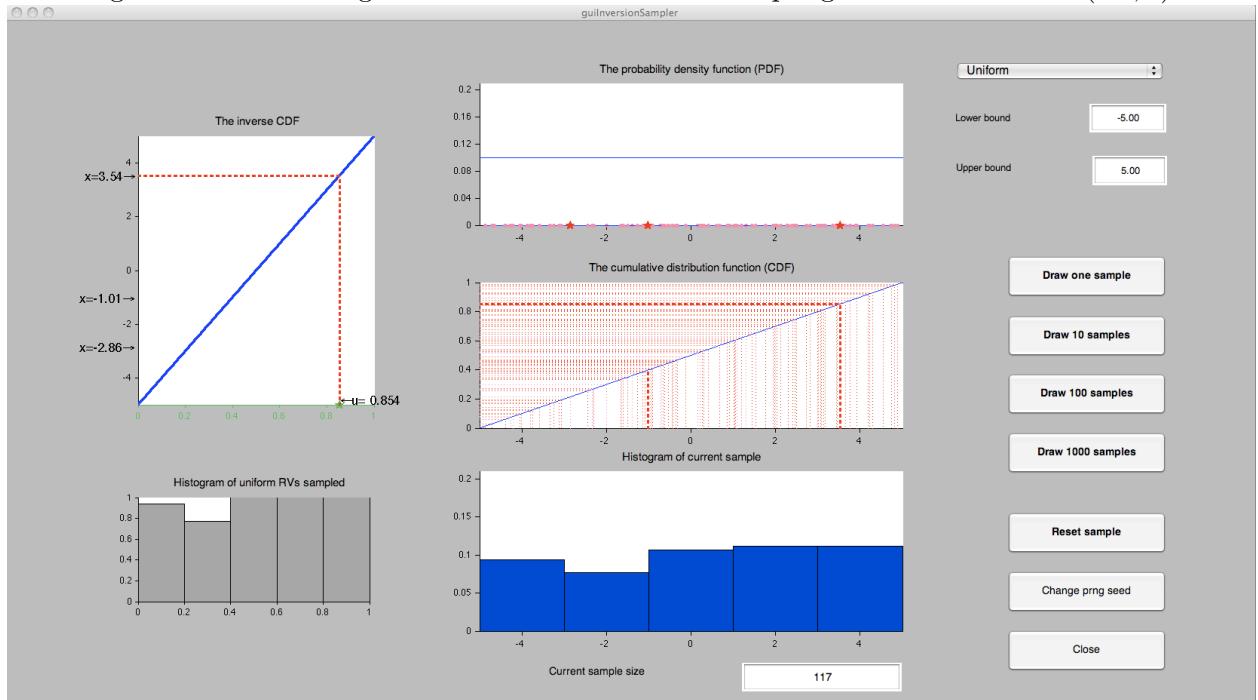


Labwork 61 (Inversion Sampler Demo – $\text{Uniform}(-5, 5)$) Let us comprehend the inversion sampler by calling the interactive visual cognitive tool built by Jennifer Harlow under a grant from University of Canterbury's Centre for Teaching and Learning (UCTL):

```
>> guiInversionSampler
```

The M-file `guiInversionSampler.m` will bring a graphical user interface (GUI) as shown in Figure 6.2. The default target distribution is $\text{Uniform}(-5, 5)$. 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 $\text{Uniform}(\theta_1, \theta_2)$ RV.

Figure 6.2: Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Uniform}(-5, 5)$.



Simulation 62 ($\text{Uniform}(\theta_1, \theta_2)$) To simulate from $\text{Uniform}(\theta_1, \theta_2)$ RV X using the Inversion Sampler, we first need to find $F^{[-1]}(u)$ by solving for x in terms of $u = F(x; \theta_1, \theta_2)$:

$$u = \frac{x - \theta_1}{\theta_2 - \theta_1} \iff x = (\theta_2 - \theta_1)u + \theta_1 \iff F^{[-1]}(u; \theta_1, \theta_2) = \theta_1 + (\theta_2 - \theta_1)u$$

Here is a simple implementation of the Inversion Sampler for the $\text{Uniform}(\theta_1, \theta_2)$ 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
```

It is just as easy to draw n IID samples from $\text{Uniform}(\theta_1, \theta_2)$ RV X by transforming n IID samples from the $\text{Uniform}(0, 1)$ 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
```

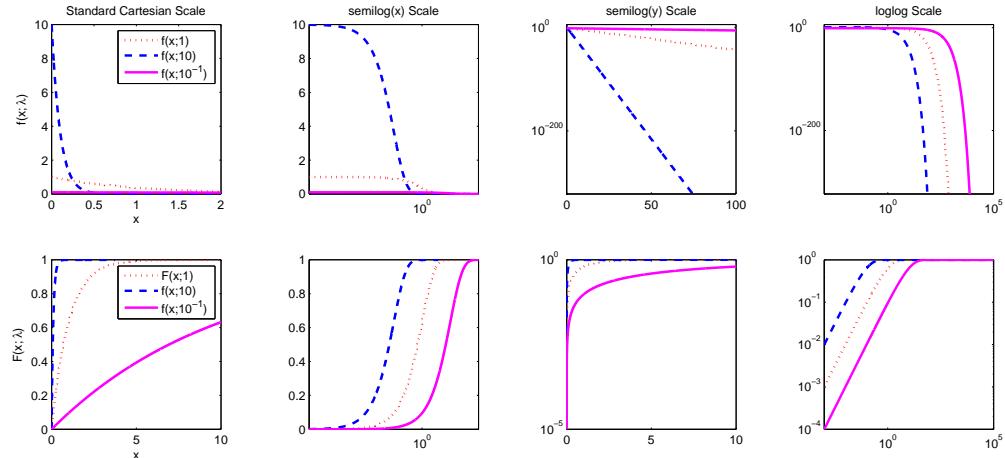
Model 5 ($\text{Exponential}(\lambda)$) For a given $\lambda > 0$, an $\text{Exponential}(\lambda)$ RV has the following PDF f and DF F :

$$f(x; \lambda) = \lambda e^{-\lambda x} \quad F(x; \lambda) = 1 - e^{-\lambda x}. \quad (6.3)$$

This distribution is fundamental because of its property of **memorylessness** and plays a fundamental role in continuous time processes as we will see later.

We encode the PDF and DF of the $\text{Exponential}(\lambda)$ RV as MATLAB functions `ExponentialPdf` and `ExponentialCdf` and use them to produce Figure 6.3 in Labwork ??.

Figure 6.3: Density and distribution functions of $\text{Exponential}(\lambda)$ RVs, for $\lambda = 1, 10, 10^{-1}$, in four different axes scales.



Mean and Variance of $\text{Exponential}(\lambda)$: Show that the mean of an $\text{Exponential}(\lambda)$ RV X is:

$$\mathbf{E}_\lambda(X) = \int_0^\infty x f(x; \lambda) dx = \int_0^\infty x \lambda e^{-\lambda x} dx = \frac{1}{\lambda},$$

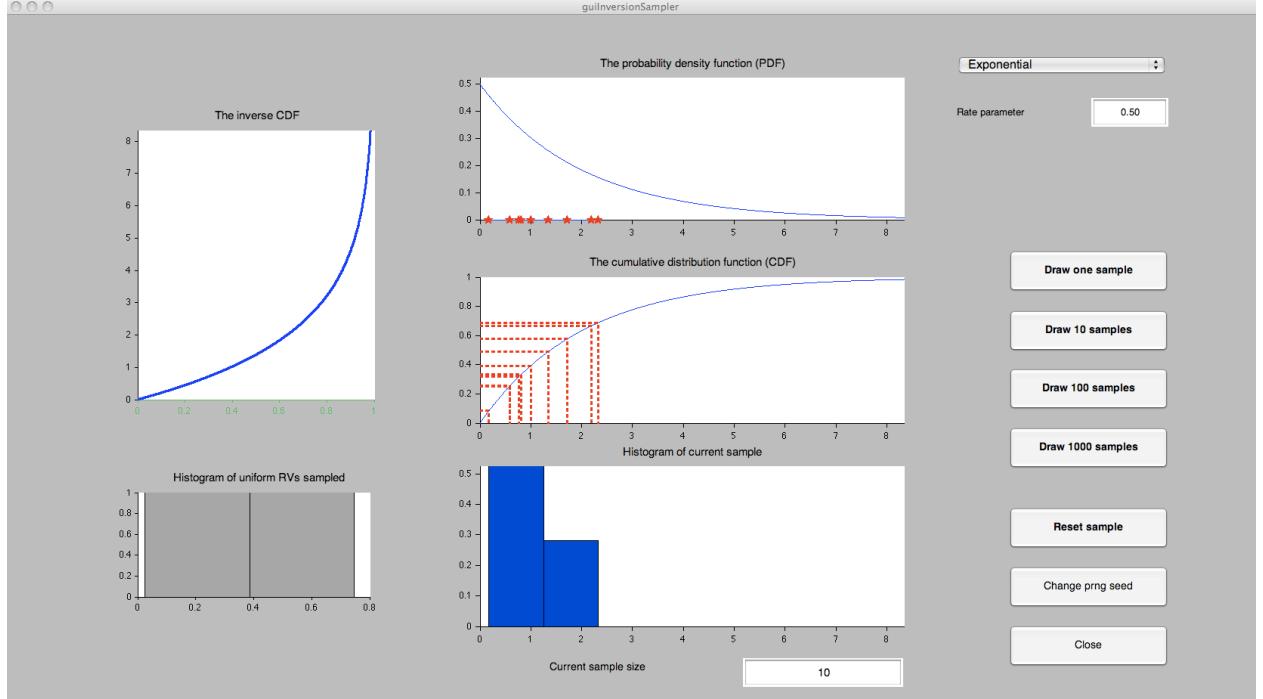
and the variance is:

$$\mathbf{V}_\lambda(X) = \left(\frac{1}{\lambda}\right)^2.$$

Labwork 63 (Inversion Sampler Demo – $\text{Exponential}(0.5)$) 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 6.4. First change the target distribution from the default $\text{Uniform}(-5, 5)$ to $\text{Exponential}(0.5)$ 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 6.4: Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Exponential}(0.5)$.

Let us consider the problem of simulating from an $\text{Exponential}(\lambda)$ 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.

Simulation 64 ($\text{Exponential}(\lambda)$) For a given $\lambda > 0$, an $\text{Exponential}(\lambda)$ RV has the following PDF f , DF F and inverse DF $F^{[-1]}$:

$$f(x; \lambda) = \lambda e^{-\lambda x} \quad F(x; \lambda) = 1 - e^{-\lambda x} \quad F^{[-1]}(u; \lambda) = \frac{-1}{\lambda} \log_e(1 - u) \quad (6.4)$$

We write the natural logarithm \log_e as log for notational simplicity. An implementation of the Inversion Sampler for $\text{Exponential}(\lambda)$ as a function in the M-file:

```
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 $\text{Exponential}(\lambda = 1.0)$ 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:

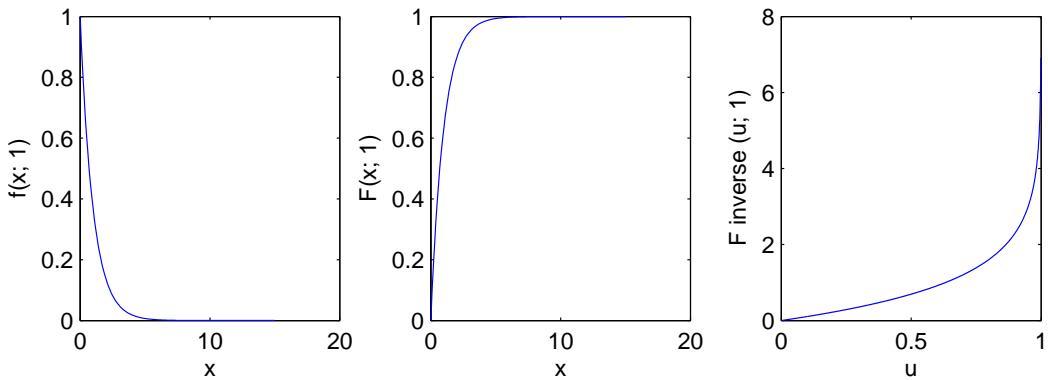
$$U \sim \text{Uniform}(0, 1) \implies -U \sim \text{Uniform}(-1, 0) \implies 1 - U \sim \text{Uniform}(0, 1),$$

we could save a subtraction operation in the above algorithm by replacing $-(1/\lambda) * \log(1-u)$ by $-(1/\lambda) * \log(u)$. This is implemented as the following function.

```
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);
```

```
>> 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) samples from rand
>> disp(x); % display the Exponential(1.0) distributed samples
    0.5945    2.5956    0.9441    1.9015    1.3973
```

Figure 6.5: The PDF f , DF F , and inverse DF $F^{[-1]}$ of the the Exponential($\lambda = 1.0$) RV.



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
    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 65 (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 $\text{Exponential}(\lambda)$ RV is $\mathbb{R}_+ := [0, \infty)$. Let us consider a RV built by mirroring the $\text{Exponential}(\lambda)$ RV about the origin with the entire real line as its support.

Model 6 (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

$$f(x; \lambda) = \frac{\lambda}{2} e^{-\lambda|x|} = \begin{cases} \frac{\lambda}{2} e^{\lambda x} & \text{if } x < 0 \\ \frac{\lambda}{2} e^{-\lambda x} & \text{if } x \geq 0 \end{cases}. \quad (6.5)$$

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

$$F(x; \lambda) = \int_{-\infty}^x f(y; \lambda) dy = \frac{1}{2} \left(1 + \text{sign}(x) \left(1 - e^{-\lambda|x|} \right) \right), \quad (6.6)$$

and its inverse DF is

$$F^{[-1]}(u; \lambda) = -\frac{1}{\lambda} \text{sign} \left(u - \frac{1}{2} \right) \log \left(1 - 2 \left| u - \frac{1}{2} \right| \right), \quad u \in [0, 1] \quad (6.7)$$

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

$$\mathbf{V}(X) = \left(\frac{1}{\lambda} \right)^2 + \left(\frac{1}{\lambda} \right)^2 = 2 \left(\frac{1}{\lambda} \right)^2.$$

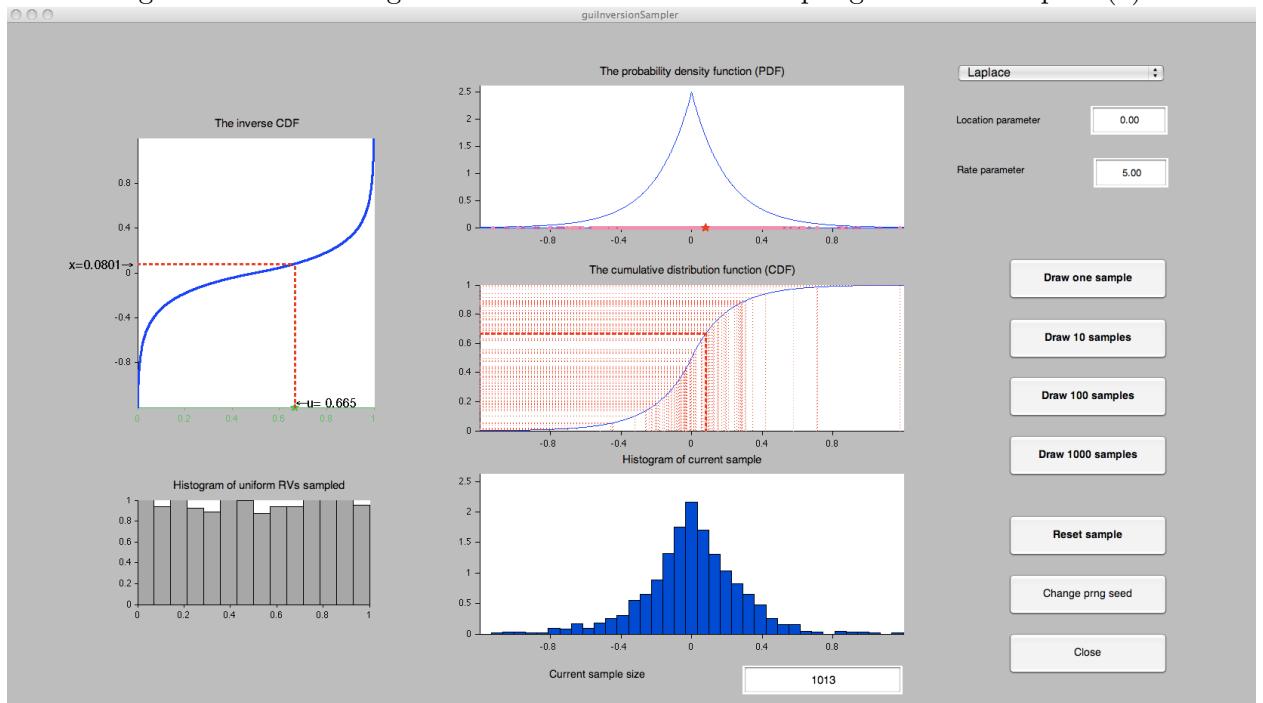
Note that the mean is 0 due to the symmetry of the density about 0 and the variance is twice that of the $\text{Exponential}(\lambda)$ RV.

Labwork 66 (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 6.6. Using the drop-down menu change from the default target distribution $\text{Uniform}(-5, 5)$ to $\text{Laplace}(5)$. 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 $\text{Laplace}(\lambda)$ 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 $\text{Laplace}(\lambda)$ 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 (6.5) with an additional location parameter μ .

Figure 6.6: Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Laplace}(5)$.



Simulation 67 ($\text{Laplace}(\lambda)$) Here is an implementation of an inversion sampler to draw IID samples from a $\text{Laplace}(\lambda)$ RV X by transforming IID samples from the $\text{Uniform}(0, 1)$ 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));
```

We can simply call the function to draw a sample from, say the $\text{Laplace}(\lambda = 1.0)$ 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);
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);
0.3530 1.6127 -0.3621 0.3637 0.9953

```

Next, let us become familiar with an RV for which the expectation does not exist. This will help us appreciate the phrase “none of which is dominant” in the informal statement of the CLT later.

Model 7 (Cauchy) The density of the Cauchy RV X is:

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty , \quad (6.8)$$

and its DF is:

$$F(x) = \frac{1}{\pi} \tan^{-1}(x) + \frac{1}{2} . \quad (6.9)$$

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, gives rise to the *Standard Cauchy* RV.

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:

$$\int |x| dF(x) = \frac{2}{\pi} \int_0^\infty \frac{x}{1+x^2} dx = (x \tan^{-1}(x)]_0^\infty - \int_0^\infty \tan^{-1}(x) dx = \infty . \quad (6.10)$$

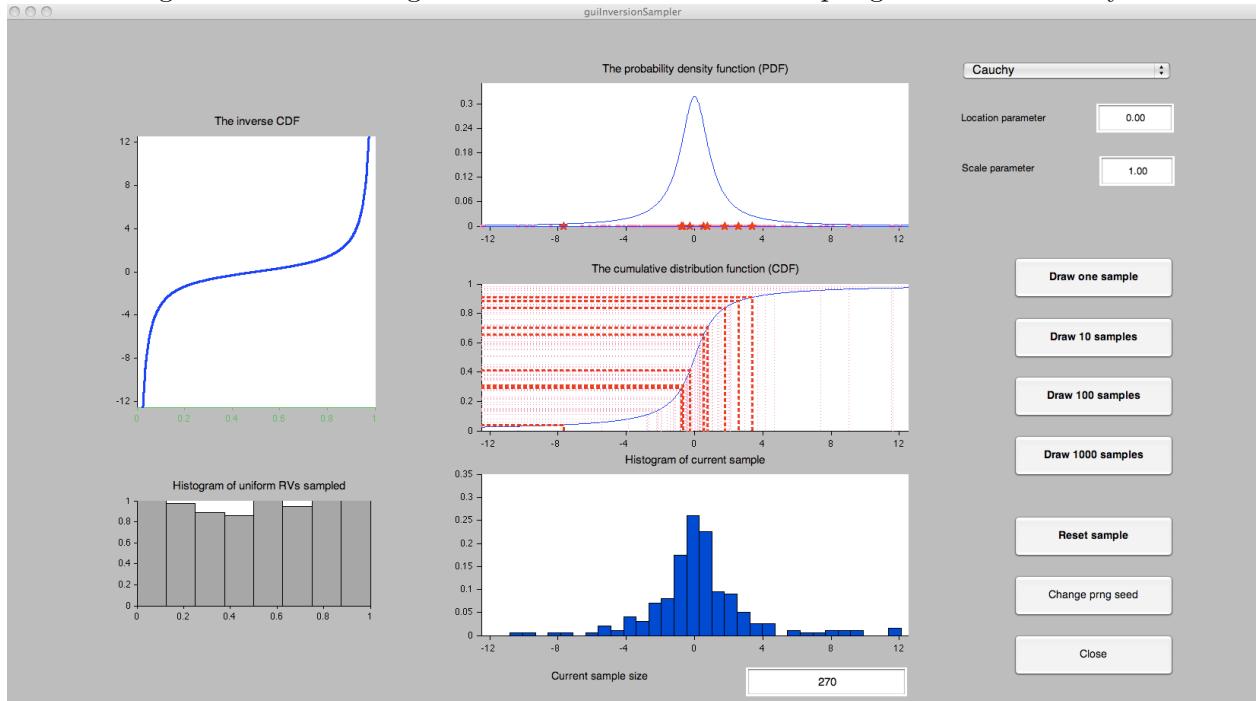
Variance and higher moments cannot be defined when the expectation itself is undefined.

Labwork 68 (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 6.7. Using the drop-down menu change from the default target distribution $\text{Uniform}(-5, 5)$ to Cauchy. 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 (6.8) with an additional location parameter μ and scale parameter σ .

Simulation 69 (Cauchy) We can draw n IID samples from the Cauchy RV X by transforming n IID samples from $\text{Uniform}(0, 1)$ RV U using the inverse DF as follows:

Figure 6.7: Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Cauchy}$.

```

>> 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

```

Recall that the mean of the Cauchy RV X does not exist since $\int |x| dF(x) = \infty$ (6.10). We will investigate this in Labwork 70.

Labwork 70 (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:

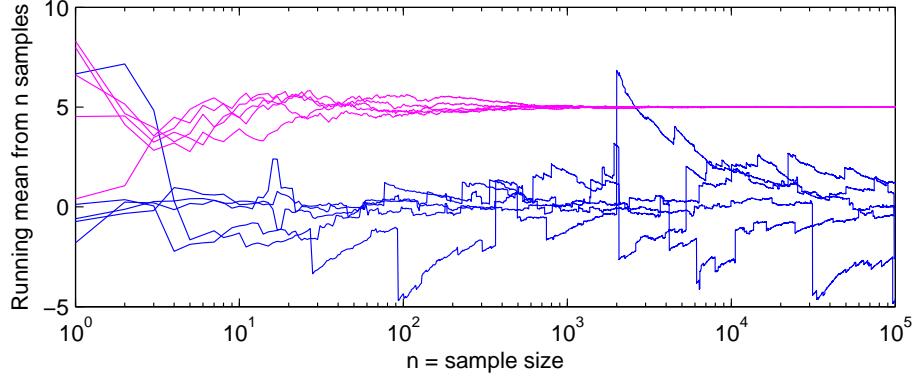
```

% 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
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 6.8: 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 6.8. 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.

For a continuous RV X with a closed-form expression for the inverse DF $F^{[-1]}$, we can employ Algorithm 3 to draw samples from X . Table 6.1 summarises some random variables that are amenable to Algorithm 3.

Table 6.1: Some continuous RVs that can be simulated from using Algorithm 3.

Random Variable X	$F(x)$	$X = F^{[-1]}(U)$, $U \sim \text{Uniform}(0, 1)$	Simplified form
Uniform(a, b)	(6.2)	$a + (b - a)U$	–
Exponential(λ)	(6.3)	$\frac{-1}{\lambda} \log(1 - U)$	$\frac{-1}{\lambda} \log(U)$
Laplace(λ)	(6.7)	$-\frac{1}{\lambda} \text{sign}(U - \frac{1}{2}) \log(1 - 2 U - \frac{1}{2})$	–
Cauchy	(6.9)	$\tan(\pi(U - \frac{1}{2}))$	$\tan(\pi U)$

Next, we familiarise ourselves with the Gaussian or Normal RV.

Model 8 (Normal(μ, σ^2)) X has a Normal(μ, σ^2) or Gaussian(μ, σ^2) distribution with the location parameter $\mu \in \mathbb{R}$ and the scale or variance parameter $\sigma^2 > 0$, if:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right), \quad x \in \mathbb{R} \quad (6.11)$$

Normal(0, 1) distributed RV, which plays a fundamental role in asymptotic statistics, is conventionally denoted by Z . Z is said to have the **Standard Normal** distribution with PDF $f(z; 0, 1)$ and DF $F(z; 0, 1)$ conventionally denoted by $\varphi(z)$ and $\Phi(z)$, respectively.

There is no closed form expression for $\Phi(z)$ or $F(x; \mu, \sigma)$. The latter is simply defined as:

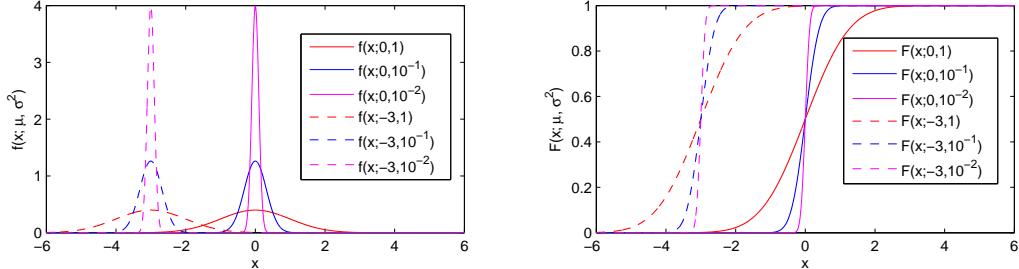
$$F(x; \mu, \sigma^2) = \int_{-\infty}^x f(y; \mu, \sigma) dy$$

We can express $F(x; \mu, \sigma^2)$ in terms of the error function (erf) as follows:

$$F(x; \mu, \sigma^2) = \frac{1}{2} \operatorname{erf}\left(\frac{x - \mu}{\sqrt{2\sigma^2}}\right) + \frac{1}{2} \quad (6.12)$$

We implement the PDF (6.11) and DF (6.12) for a $\text{Normal}(\mu, \sigma^2)$ RV X as MATLAB functions `NormalPdf` and `NormalCdf`, respectively, in Labwork ??, and then produce plots for various $\text{Normal}(\mu, \sigma^2)$ RVs, shown in Figure 6.9. Observe the concentration of probability mass, in terms of the PDF and DF plots, about the location parameter μ as the variance parameter σ^2 decreases.

Figure 6.9: Density and distribution function of several $\text{Normal}(\mu, \sigma^2)$ RVs.



Mean and Variance of $\text{Normal}(\mu, \sigma^2)$: The mean of a $\text{Normal}(\mu, \sigma^2)$ RV X is:

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} xf(x; \mu, \sigma^2) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) dx = \mu ,$$

and the variance is:

$$\mathbf{V}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x; \mu, \sigma^2) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x - \mu)^2 \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) dx = \sigma^2 .$$

Labwork 71 (Compute the the $\mathbf{P}(X \in (a, b))$ for the $\text{Normal}(0, 1)$ RV X) Write a function to evaluate the $\mathbf{P}(X \in (a, b))$ for the $\text{Normal}(0, 1)$ RV X for user-specified values of a and b . [Hint: one option is by making two calls to `NormalCdf` and doing one arithmetic operation.]

Simulations 62 and 64, 67 and 69 produce samples from a continuous RV X with a closed-form expression for the inverse DF $F^{[-1]}$ via Algorithm 3 (Table 6.1). But only a few RVs have an explicit $F^{[-1]}$. For example, $\text{Normal}(0, 1)$ RV does not have an explicit $F^{[-1]}$. Algorithm 4 is a more general but inexact method that relies on an approximate numerical solution of x , for a given u , that satisfies the equation $F(x) = u$.

Simulation 72 (Normal(μ, σ^2)) We may employ Algorithm 4 to sample from the $\text{Normal}(\mu, \sigma^2)$ RV X using the following function.

```
function x = Sample1NormalByNewRap(u, Mu, SigmaSq)
% Returns a sample from Normal(Mu, SigmaSq)
% Newton-Raphson numerical solution of F(x)=u
```

Algorithm 4 Inversion Sampler by Numerical Solution of $F(X) = U$ via Newton-Raphson Method

-
- 1: *input:* $F(x)$, the DF of the target RV X
 - 2: *input:* $f(x)$, the density of X
 - 3: *input:* A reasonable Stopping Rule,
e.g. a specified tolerance $\epsilon > 0$ and a maximum number of iterations MAX
 - 4: *input:* a careful mechanism to specify x_0
 - 5: *output:* a sample from X distributed according to F
 - 6: *draw:* $u \sim \text{Uniform}(0, 1)$
 - 7: *initialise:* $i \leftarrow 0$, $x_i \leftarrow x_0$, $x_{i+1} \leftarrow x_0 - \frac{F(x_0) - u}{f(x_0)}$
 - 8: **while** Stopping Rule is not satisfied,
e.g. $|F(x_i) - F(x_{i-1})| > \epsilon$ AND $i < \text{MAX}$ **do**
 - 9: $x_i \leftarrow x_{i+1}$
 - 10: $x_{i+1} \leftarrow \left(x_i - \frac{F(x_i) - u}{f(x_i)} \right)$
 - 11: $i \leftarrow i + 1$
 - 12: **end while**
 - 13: *return:* $x \leftarrow x_i$
-

```
% Input: u = one random Uniform(0,1) sample
%         Mu = Mean of Normal(Mu, SigmaSq)
%         SigmaSq = Variance of Normal(Mu, SigmaSq)
% Usage: x = Sample1NormalByNewRap(u,Mu,SigmaSq)
% To transform an array Us of uniform samples to array Xs of Normal samples via arrayfun
%     Xs = arrayfun(@(u)(Sample1NormalByNewRap(u,-100.23,0.01)),Us);
Epsilon=1e-5; % Tolerance in stopping rule
MaxIter=10000; % Maximum allowed iterations in stopping rule
x=0; % initialize the output x as 0
% initialize i, xi, and xii
i=0; % Mu is an ideal initial condition since F(x; Mu, SigmaSq)
xi = Mu; % is convex when x < Mu and concave when x > Mu and the
% Newton-Raphson method started at Mu converges
xii = xi - (NormalCdf(xi,Mu,SigmaSq)-u)/NormalPdf(xi,Mu,SigmaSq);
% Newton-Raphson Iterations
while (abs(NormalCdf(xii,Mu,SigmaSq)-NormalCdf(xi,Mu,SigmaSq))...
    > Epsilon & i < MaxIter),
    xi = xii;
    xii = xii - (NormalCdf(xii,Mu,SigmaSq)-u)/NormalPdf(xii,Mu,SigmaSq);
    i=i+1;
end
x=xii; % record the simulated x from the j-th element of u
```

We draw five samples from the $\text{Normal}(0, 1)$ RV Z and store them in z as follows. The vector z can be obtained by a Newton-Raphson-based numerical transformation of the vector u of 5 IID samples from the $\text{Uniform}(0, 1)$ RV. We simply need to apply the function `Sample1NormalByNewRap` to each element of an array of $\text{Uniform}(0, 1)$ samples. MATLAB's `arrayfun` command can be used to apply `@(u)(Sample1NormalByNewRap(u,0,1))` (i.e., `Sample1NormalByNewRap` as a function of u) to every element of our array of $\text{Uniform}(0, 1)$ samples, say `Us`. Note that $F(z)$ is the same as the drawn u from U at least up to four significant digits.

```
>> rand('twister',563987);
>> Us=rand(1,5); % store 5 samples from Uniform(0,1) RV in array Us
>> disp(Us); % display Us
  0.8872    0.2569    0.5275    0.8650    0.8517
>> z=Sample1NormalByNewRap(Us(1),0,1); %transform Us(1) to a Normal(0,1) sample z
```

```
>> disp(z); % display z
    1.2119
>> z = arrayfun(@(u)(Sample1NormalByNewRap(u,0,1)),Us); %transform array Us via arrayfun
>> % display array z obtained from applying Sample1NormalByNewRap to each element of Us
>> disp(z);
    1.2119   -0.6530    0.0691    1.1031    1.0439
>> % check that numerical inversion of F worked, i.e., is F(z)=u ?
>> disp(NormalCdf(z,0,1));
    0.8872    0.2569    0.5275    0.8650    0.8517
```

Next we draw five samples from the $\text{Normal}(-100.23, 0.01)$ RV X , store it in an array x and observe that the numerical method is reasonably accurate by the equality of u and $F(x)$.

```
>> rand('twister',563987);
>> disp(Us); % display Us
    0.8872    0.2569    0.5275    0.8650    0.8517
>> % transform array Us via arrayfun
>> x = arrayfun(@(u)(Sample1NormalByNewRap(u,-100.23,0.01)),Us);
>> disp(x);
    -100.1088 -100.2953 -100.2231 -100.1197 -100.1256
>> disp(NormalCdf(x,-100.23,0.01));
    0.8872    0.2569    0.5275    0.8650    0.8517
```

One has to be extremely careful with this approximate simulation algorithm implemented in floating-point arithmetic. More robust samplers for the $\text{Normal}(\mu, \sigma^2)$ RV exist. However, Algorithm 4 is often the only choice when simulating from an arbitrary RV with an unknown closed-form expression for its $F^{[-1]}$.

Next, we use our simulation capability to gain an informal and intuitive understanding of one of the most elementary theorems in probability and statistics, namely, the Central Limit Theorem (CLT). We will see a formal treatment of CLT later.

Informally, the CLT can be stated as follows:

“The sample mean of a large number of IID samples, none of which is dominant, tends to the Normal distribution as the number of samples increases.”

Labwork 73 (Investigating the Central Limit Theorem with IID Exponential($\lambda = 0.1$) RVs)

Let us investigate the histograms from 10000 simulations of the sample mean of $n = 10, 100, 1000$ IID Exponential($\lambda = 0.1$) RVs as follows:

```
>> rand('twister',1973); % initialise the fundamental sampler
>> % a demonstration of Central Limit Theorem (CLT) -- Details of CLT are in the sequel
>> % the sample mean should be a Normal(1/lambd,a/lambd/n) RV
>> lambda=0.1; Reps=10000; n=10; hist(sum(-1/lambda * log(rand(n,Reps)))/n)
>> lambda=0.1; Reps=10000; n=100; hist(sum(-1/lambda * log(rand(n,Reps)))/n,20)
>> lambda=0.1; Reps=10000; n=1000; hist(sum(-1/lambda * log(rand(n,Reps)))/n,20)
```

Do you see a pattern in the histograms?

See the histograms generated from the following code that produces sample means from the Cauchy RV:

```
>> Reps=10000; n=1000; hist(sum(tan(pi * rand(n,Reps)))/n,20)
>> Reps=10000; n=1000; hist(sum(tan(pi * rand(n,Reps)))/n,20)
>> Reps=10000; n=1000; hist(sum(tan(pi * rand(n,Reps)))/n,20)
```

Classwork 74 (Why doesn't the sample mean of the Cauchy RV ever settle down?) Explain in words why the mean of n IID samples from the Cauchy RV “is **not** obeying” the Central Limit Theorem. Also relate it to Figure 6.8 of Labwork 70.

Model 9 (Gamma(λ, k) RV) Given a shape parameter $\alpha > 0$ and a rate parameter $\beta > 0$, the RV X is said to be Gamma(α, β) distributed if its PDF is:

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

where, the gamma function which interpolates the factorial function is:

$$\Gamma(\alpha) := \int_0^\infty \exp(-y) y^{\alpha-1} dy .$$

When $k \in \mathbb{N}$, then $\Gamma(k) = (k - 1)!$. The DF of X is:

$$F(x; \lambda, k) = \mathbf{1}_{\mathbb{R}_{>0}}(x) \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^x y^{\alpha-1} \exp(-\beta y) dy = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)} & \text{if } x > 0 \end{cases}$$

where $\gamma(\alpha, \beta x)$ is called the lower incomplete Gamma function.

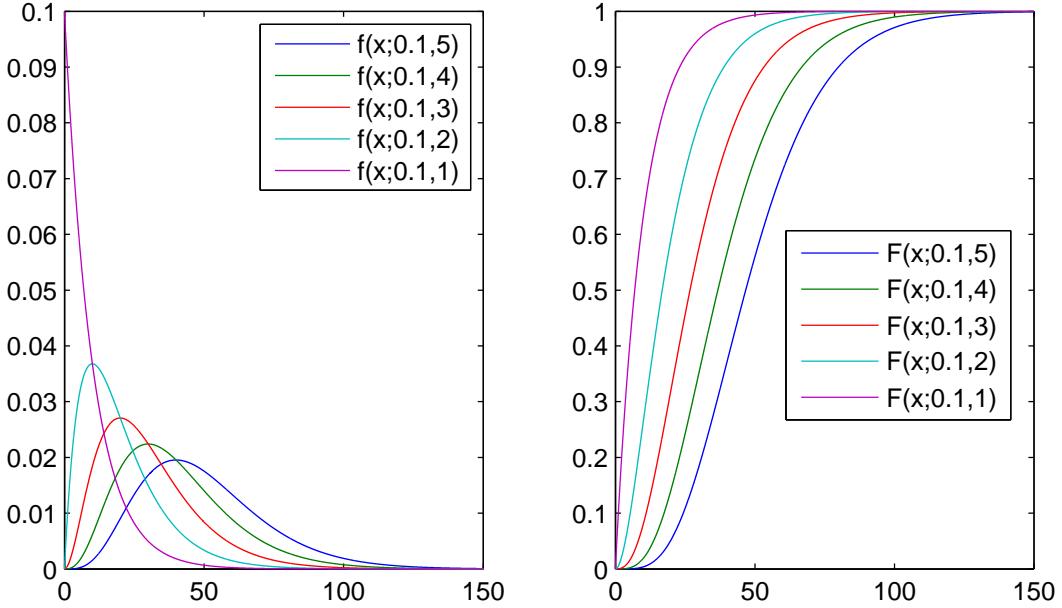
The expectation and variance of a Gamma(α, β) RV are α/β and α/β^2 , respectively. The Gamma function and the incomplete Gamma function are available as MATLAB functions `gamma` and `gammainc`, respectively. Thus, `gamma(k)` returns $\Gamma(k)$ and `gammainc(lambda*x, k)` returns $F(x; \lambda, k)$. Using these functions, it is straightforward to evaluate the PDF and CDF of $X \sim \text{Gamma}(\lambda, k)$. We use the following script to get a sense for the impact upon the PDF and CDF of the shape parameter k as it ranges in $\{1, 2, 3, 4, 5\}$ for a given scale parameter $\lambda = 0.1$.

```
PlotPdfCdfGamma.m
lambda=0.1; % choose some scale parameter
Xs=0:0.01:150; % choose some x values
% Plot PDFs for k=5,4,3,2,1
k=5; fXsk5=(1/gamma(k))*(lambda*exp(-lambda*Xs).*(lambda*Xs).^(k-1));% PDF for k=5
k=4; fXsk4=(1/gamma(k))*(lambda*exp(-lambda*Xs).*(lambda*Xs).^(k-1));% PDF for k=4
k=3; fXsk3=(1/gamma(k))*(lambda*exp(-lambda*Xs).*(lambda*Xs).^(k-1));% PDF for k=3
k=2; fXsk2=(1/gamma(k))*(lambda*exp(-lambda*Xs).*(lambda*Xs).^(k-1));% PDF for k=2
k=1; fXsk1=(1/gamma(k))*(lambda*exp(-lambda*Xs).*(lambda*Xs).^(k-1));% PDF for k=1
clf; % clear any previous figures
subplot(1,2,1); % make first PDF plot
plot(Xs,fXsk5, Xs, fXsk4, Xs, fXsk3, Xs, fXsk2, Xs, fXsk1)
legend('f(x;0.1,5)', 'f(x;0.1,4)', 'f(x;0.1,3)', 'f(x;0.1,2)', 'f(x;0.1,1)')
subplot(1,2,2) % make second CDF plots using MATLAB's gammaintc (incomplete gamma function)
plot(Xs,gammaintc(lambda*Xs,5), Xs,gammaintc(lambda*Xs,4), Xs,gammaintc(lambda*Xs,3),...
      Xs,gammaintc(lambda*Xs,2), Xs,gammaintc(lambda*Xs,1))
legend('F(x;0.1,5)', 'F(x;0.1,4)', 'F(x;0.1,3)', 'F(x;0.1,2)', 'F(x;0.1,1)')
```

Note that if $X \sim \text{Gamma}(1, \beta)$ then $X \sim \text{Exponential}(\beta)$, since:

$$f(x; 1, \beta) = \frac{1}{(1-1)!} \beta \exp(-\beta x) = \beta \exp(-\beta x) .$$

More generally, if $X \sim \text{Gamma}(\alpha, \beta)$ and $\alpha \in \mathbb{N}$, then $X \sim \sum_{i=1}^{\alpha} Y_i$, where $Y_i \stackrel{IID}{\sim} \text{Exponential}(\beta)$ RVS, i.e. the sum of α IID Exponential(β) RVs forms the model for the Gamma(α, β) RV. If you model the inter-arrival time of buses at a bus-stop by IID Exponential(β) RV, then you can think of the arrival time of the k^{th} bus as a Gamma(α, β) RV.

Figure 6.10: PDF and CDF of $X \sim \text{Gamma}(\beta = 0.1, \alpha)$ with $\alpha \in \{1, 2, 3, 4, 5\}$.

6.4 Discrete Random Variables

6.5 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 38 (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 $U \sim \text{Uniform}(0, 1)$ 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:

$$\mathbf{P}(F^{[-1]}(U) \leq x) = \mathbf{P}(U \leq F(x)) = F(x) := \mathbf{P}(X \leq x)$$

6.6 Some Simulations of Discrete Random Variables

Simulation 75 (Bernoulli(θ)) Consider the problem of simulating from a $\text{Bernoulli}(\theta)$ RV based on an input from a $\text{Uniform}(0, 1)$ 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 $\text{Bernoulli}(\theta)$ 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 $\text{Bernoulli}(\theta)$ 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)

```

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.

Model 10 (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:

$$F(x; \theta) = \begin{cases} 0 & \text{if } x < \theta \\ 1 & \text{if } x \geq \theta \end{cases} \quad (6.13)$$

and the PMF is:

$$f(x; \theta) = \begin{cases} 0 & \text{if } x \neq \theta \\ 1 & \text{if } x = \theta \end{cases} \quad (6.14)$$

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.

Mean and variance of Point Mass(θ) RV: Let $X \sim \text{Point Mass}(\theta)$. Then:

$$\mathbf{E}(X) = \sum_x x f(x) = \theta \times 1 = \theta, \quad \mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \theta^2 - \theta^2 = 0.$$

Simulation 76 (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:

```

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 $\text{Uniform}(0, 1)$ 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 let us consider a natural generalization of the $\text{Bernoulli}(\theta)$ RV with more than two outcomes.

Model 11 (de Moivre($\theta_1, \theta_2, \dots, \theta_k$)) Given a specific point $(\theta_1, \theta_2, \dots, \theta_k)$ in the k -Simplex:

$$\Delta_k := \{ (\theta_1, \theta_2, \dots, \theta_k) : \theta_1 \geq 0, \theta_2 \geq 0, \dots, \theta_k \geq 0, \sum_{i=1}^k \theta_i = 1 \},$$

we say that an RV X is de Moivre($\theta_1, \theta_2, \dots, \theta_k$) distributed if its PMF is:

$$f(x; \theta_1, \theta_2, \dots, \theta_k) = \begin{cases} 0 & \text{if } x \notin [k] := \{1, 2, \dots, k\}, \\ \theta_x & \text{if } x \in [k]. \end{cases}$$

The DF for de Moivre($\theta_1, \theta_2, \dots, \theta_k$) RV X is:

$$F(x; \theta_1, \theta_2, \dots, \theta_k) = \begin{cases} 0 & \text{if } -\infty < x < 1 \\ \theta_1 & \text{if } 1 \leq x < 2 \\ \theta_1 + \theta_2 & \text{if } 2 \leq x < 3 \\ \vdots & \\ \theta_1 + \theta_2 + \dots + \theta_{k-1} & \text{if } k-1 \leq x < k \\ \theta_1 + \theta_2 + \dots + \theta_{k-1} + \theta_k = 1 & \text{if } k \leq x < \infty \end{cases} \quad (6.15)$$

The de Moivre($\theta_1, \theta_2, \dots, \theta_k$) 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 $1, 2, \dots, k$ ”. The parameters $\theta_1, \theta_2, \dots, \theta_k$ 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 $\theta_1 = \theta_2 = \dots = \theta_k = 1/k$, we have a probability model for the outcomes of a fair die.

Mean and variance of de Moivre($\theta_1, \theta_2, \dots, \theta_k$) RV: The not too useful expressions for the first two moments of $X \sim \text{de Moivre}(\theta_1, \theta_2, \dots, \theta_k)$ are,

$$\mathbf{E}(X) = \sum_{x=1}^k x\theta(x) = \theta_1 + 2\theta_2 + \dots + k\theta_k, \text{ and}$$

$$\mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = (\theta_1 + 2^2\theta_2 + \dots + k^2\theta_k) - (\theta_1 + 2\theta_2 + \dots + k\theta_k)^2 .$$

However, if $X \sim \text{de Moivre}(1/k, 1/k, \dots, 1/k)$, 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,

$$\mathbf{E}(X) = \frac{1}{k} (1 + 2 + \dots + k) = \frac{1}{k} \frac{k(k+1)}{2} = \frac{k+1}{2} ,$$

$$\mathbf{E}(X^2) = \frac{1}{k} (1^2 + 2^2 + \dots + k^2) = \frac{1}{k} \frac{k(k+1)(2k+1)}{6} = \frac{2k^2 + 3k + 1}{6} ,$$

$$\begin{aligned} \mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 &= \frac{2k^2 + 3k + 1}{6} - \left(\frac{k+1}{2}\right)^2 = \frac{2k^2 + 3k + 1}{6} - \left(\frac{k^2 + 2k + 1}{4}\right) \\ &= \frac{8k^2 + 12k + 4 - 6k^2 - 12k - 6}{24} = \frac{2k^2 - 2}{24} = \frac{k^2 - 1}{12} . \end{aligned}$$

Next we simulate from $\text{de Moivre}(\theta_1, \theta_2, \dots, \theta_k)$ RV X via its inverse DF

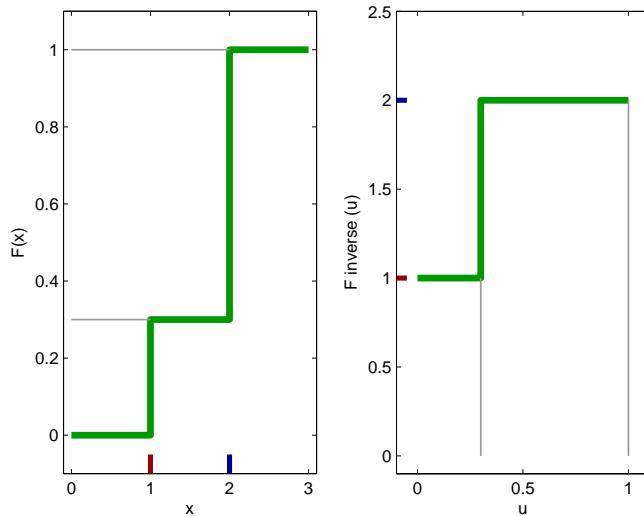
$$F^{[-1]} : [0, 1] \rightarrow [k] := \{1, 2, \dots, k\} ,$$

given by:

$$F^{[-1]}(u; \theta_1, \theta_2, \dots, \theta_k) = \begin{cases} 1 & \text{if } 0 \leq u < \theta_1 \\ 2 & \text{if } \theta_1 \leq u < \theta_1 + \theta_2 \\ 3 & \text{if } \theta_1 + \theta_2 \leq u < \theta_1 + \theta_2 + \theta_3 \\ \vdots & \\ k & \text{if } \theta_1 + \theta_2 + \dots + \theta_{k-1} \leq u < 1 \end{cases} \quad (6.16)$$

When $k = 2$ in the $\text{de Moivre}(\theta_1, \theta_2)$ 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 6.11.

Figure 6.11: The DF $F(x; 0.3, 0.7)$ of the $\text{de Moivre}(0.3, 0.7)$ RV and its inverse $F^{[-1]}(u; 0.3, 0.7)$.



First we simulate from an equi-probable special case of the $\text{de Moivre}(\theta_1, \theta_2, \dots, \theta_k)$ RV, with $\theta_1 = \theta_2 = \dots = \theta_k = 1/k$.

Simulation 77 (de Moivre($1/k, 1/k, \dots, 1/k$)) The equi-probable *de Moivre*($1/k, 1/k, \dots, 1/k$) RV X with a discrete uniform distribution over $[k] = \{1, 2, \dots, k\}$ 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 5 produces samples from the *de Moivre*($1/k, 1/k, \dots, 1/k$) RV X .

Algorithm 5 Inversion Sampler for de Moivre($1/k, 1/k, \dots, 1/k$) RV

1: *input*:

1. k in de Moivre($1/k, 1/k, \dots, 1/k$) RV X
2. $u \sim \text{Uniform}(0, 1)$

2: *output*: a sample from X

3: *return*: $x \leftarrow \lceil ku \rceil$

The M-file implementing Algorithm 5 is:

```
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 $\text{de Moivre}(\theta_1, \theta_2, \dots, \theta_k)$ RV. That is, the values of θ_i need not be equal to $1/k$.

Simulation 78 (de Moivre($\theta_1, \theta_2, \dots, \theta_k$)) We can generate samples from a $\text{de Moivre}(\theta_1, \theta_2, \dots, \theta_k)$ RV X when $(\theta_1, \theta_2, \dots, \theta_k)$ are specifiable as an input vector via the following algorithm.

The M-file implementing Algorithm 6 is:

```
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
```

Algorithm 6 Inversion Sampler for de Moivre($\theta_1, \theta_2, \dots, \theta_k$) RV X

1: *input:*

1. parameter vector $(\theta_1, \theta_2, \dots, \theta_k)$ of de Moivre($\theta_1, \theta_2, \dots, \theta_k$) RV X .
2. $u \sim \text{Uniform}(0, 1)$

2: *output:* a sample from X

3: *initialise:* $F \leftarrow \theta_1$, $i \leftarrow 1$

4: **while** $u > F$ **do**

5: $i \leftarrow i + 1$

6: $F \leftarrow F + \theta_i$

7: **end while**

8: *return:* $x \leftarrow i$

```
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
>> % 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 6 may be used to sample from any de Moivre($\theta_1, \dots, \theta_k$) RV X . We demonstrate this by producing five samples from a randomly generated PMF `f2`.

```
>> 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:

$$1\theta_1 + 2\theta_2 + 3\theta_3 + \dots + k\theta_k = \sum_{i=1}^k i\theta_i$$

For the de Moivre($1/k, 1/k, \dots, 1/k$) 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($1/k, 1/k, \dots, 1/k$) RV using Algorithm 6. Using the simpler Algorithm 5, which exploits the fact that all values of θ_i are equal, we generated samples in constant time, which is denoted by $O(1)$. Let us consider a RV that arises from an IID stochastic process of Bernoulli(θ) RVs $\{X_i\}_{i \in \mathbb{N}}$, ie.

$$\{X_i\}_{i \in \mathbb{N}} := \{X_1, X_2, \dots\} \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta).$$

When we consider the number of IID Bernoulli(θ) trials before the first ‘Head’ occurs we get the following discrete RV.

Model 12 (Geometric(θ) RV) Given a parameter $\theta \in (0, 1)$, the PMF of the Geometric(θ) RV X is

$$f(x; \theta) = \begin{cases} \theta(1 - \theta)^x & \text{if } x \in \mathbb{Z}_+ := \{0, 1, 2, \dots\} \\ 0 & \text{otherwise} \end{cases} \quad (6.17)$$

It is straightforward to verify that $f(x; \theta)$ is indeed a PDF :

$$\sum_{x=0}^{\infty} f(x; \theta) = \sum_{x=0}^{\infty} \theta(1 - \theta)^x = \theta \left(\frac{1}{1 - (1 - \theta)} \right) = \theta \left(\frac{1}{\theta} \right) = 1$$

The above equality is a consequence of the geometric series identity (6.18) with $a = \theta$ and $\vartheta := 1 - \theta$:

$$\sum_{x=0}^{\infty} a\vartheta^x = a \left(\frac{1}{1 - \vartheta} \right), \text{ provided, } 0 < \vartheta < 1. \quad (6.18)$$

Proof:

$$a + a\vartheta + a\vartheta^2 + \dots + a\vartheta^n = \sum_{0 \leq x \leq n} a\vartheta^x = a + \sum_{1 \leq x \leq n} a\vartheta^x = a + \vartheta \sum_{1 \leq x \leq n} a\vartheta^{x-1} = a + \vartheta \sum_{0 \leq x \leq n-1} a\vartheta^x = a + \vartheta \sum_{0 \leq x \leq n} a\vartheta^x - a\vartheta^{n+1}$$

Therefore,

$$\begin{aligned} \sum_{0 \leq x \leq n} a\vartheta^x &= a + \vartheta \sum_{0 \leq x \leq n} a\vartheta^x - a\vartheta^{n+1} \\ \left(\sum_{0 \leq x \leq n} a\vartheta^x \right) - \left(\vartheta \sum_{0 \leq x \leq n} a\vartheta^x \right) &= a - a\vartheta^{n+1} \\ \left(\sum_{0 \leq x \leq n} a\vartheta^x \right) (1 - \vartheta) &= a(1 - \vartheta^{n+1}) \\ \sum_{0 \leq x \leq n} a\vartheta^x &= a \left(\frac{1 - \vartheta^{n+1}}{1 - \vartheta} \right) \\ \sum_{x=0}^{\infty} a\vartheta^x := \lim_{n \rightarrow \infty} \sum_{0 \leq x \leq n} a\vartheta^x &= a \left(\frac{1}{1 - \vartheta} \right), \text{ provided, } 0 < \vartheta < 1 \end{aligned}$$

The outcome of a $\text{Geometric}(\theta)$ 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.”

Mean and variance of $\text{Geometric}(\theta)$ RV: Let $X \sim \text{Geometric}(\theta)$ 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 :

$$\frac{-1}{\theta^2} = \frac{d}{d\theta} \left(\frac{1}{\theta} \right) = \frac{d}{d\theta} \sum_{x=0}^{\infty} (1-\theta)^x = \sum_{x=0}^{\infty} -x(1-\theta)^{x-1}$$

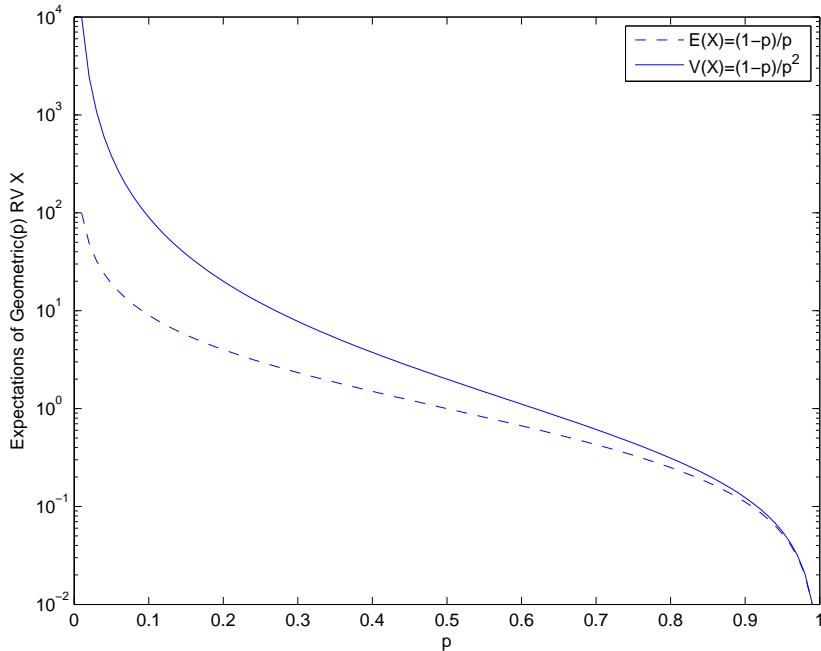
Multiplying the LHS and RHS above by $-(1-\theta)$ and substituting in $\mathbf{E}(X) = \theta \sum_{x=0}^{\infty} x(1-\theta)^x$, we get a much simpler expression for $\mathbf{E}(X)$:

$$\frac{1-\theta}{\theta^2} = \sum_{x=0}^{\infty} x(1-\theta)^x \implies \mathbf{E}(X) = \theta \left(\frac{1-\theta}{\theta^2} \right) = \frac{1-\theta}{\theta} .$$

Similarly, it can be shown that

$$\mathbf{V}(X) = \frac{1-\theta}{\theta^2} .$$

Figure 6.12: Mean and variance of a $\text{Geometric}(\theta)$ RV X as a function of the parameter θ .



Simulation 79 (Geometric(θ)) We can simulate a sample x from a $\text{Geometric}(\theta)$ RV X using the following simple algorithm:

$$x \leftarrow \lfloor \log(u) / \log(1-\theta) \rfloor, \quad \text{where, } u \sim \text{Uniform}(0, 1) .$$

To verify that the above procedure is valid, note that:

$$\begin{aligned} \lfloor \log(U)/\log(1-\theta) \rfloor = x &\iff x \leq \log(U)/\log(1-\theta) < x+1 \\ &\iff x \leq \log_{1-\theta}(U) < x+1 \\ &\iff (1-\theta)^x \geq U > (1-\theta)^{x+1} \end{aligned}$$

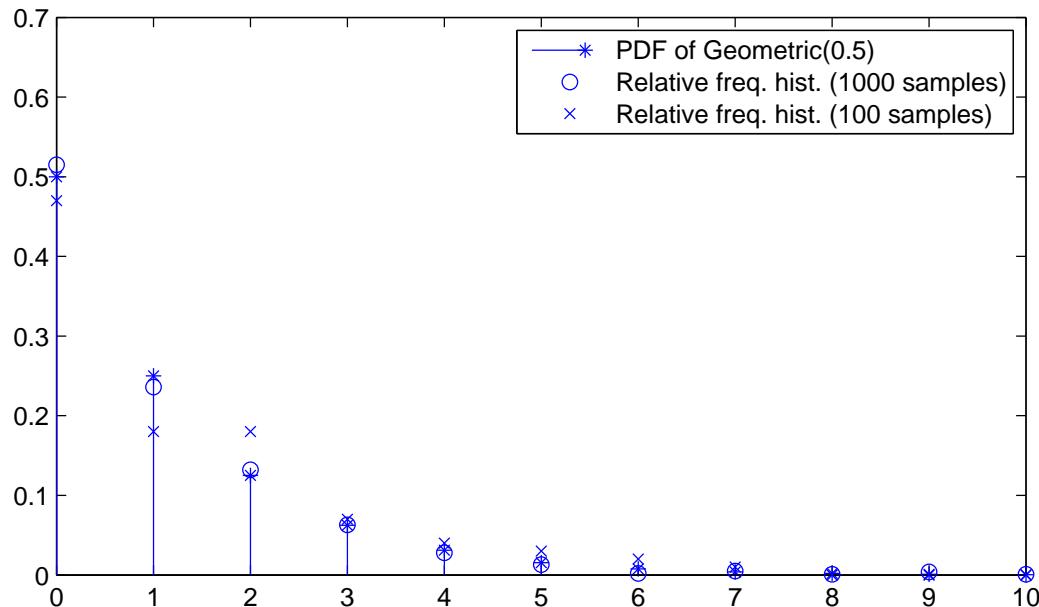
The inequalities are reversed since the base being exponentiated is $1-\theta \leq 1$. The uniform event $(1-\theta)^x \geq U > (1-\theta)^{x+1}$ happens with the desired probability:

$$(1-\theta)^x - (1-\theta)^{x+1} = (1-\theta)^x(1-(1-\theta)) = \theta(1-\theta)^x =: f(x; \theta), \quad X \sim \text{Geometric}(\theta).$$

We implement the sampler to generate samples from $\text{Geometric}(\theta)$ RV with $\theta = 0.5$, for instance:

```
>> theta=0.5; u=rand(); % choose some theta and uniform(0,1) variate
>> % Simulate from a Geometric(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
```

Figure 6.13: PDF of $X \sim \text{Geometric}(\theta = 0.5)$ and the relative frequency histogram based on 100 and 1000 samples from X .



Labwork 80 (Compare PDF to the 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 6.13:

```
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 frequency 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)')
```

The RV Y in Table ?? may be generalized to an experiment \mathcal{E}_θ^n with n coin tosses. Let X_i be the Indicator function of the event ‘Heads on the i -th toss’ as before. Then Y defined by,

$$Y := \sum_{i=1}^n X_i := X_1 + X_2 + \cdots + X_n ,$$

is the number of ‘Heads’ in n tosses. Akin to the second row of Table ??, for the ‘Toss n times’ experiment \mathcal{E}_θ^n the RV Y as defined above will take values in $\{0, 1, 2, \dots, n\}$ and is therefore a discrete RV. This is called the Binomial RV as defined next. But, first we remind ourselves of some elementary definitions involving arrangements of objects from a collection (recall Section 1.5).

Model 13 (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.:

$$X = \sum_{i=1}^n X_i, \quad X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta) .$$

Given two parameters n and θ , the PMF of the Binomial(n, θ) RV X is:

$$f(x; n, \theta) = \begin{cases} \binom{n}{x} \theta^x (1 - \theta)^{n-x} & \text{if } x \in \{0, 1, 2, 3, \dots, n\} , \\ 0 & \text{otherwise} \end{cases} \quad (6.19)$$

where, $\binom{n}{x}$ is:

$$\binom{n}{x} = \frac{n(n-1)(n-2)\dots(n-x+1)}{x(x-1)(x-2)\dots(2)(1)} = \frac{n!}{x!(n-x)!} .$$

$\binom{n}{x}$ is read as “ n choose x .”

Proof: 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 $x \in \{0, 1, \dots, n\}$.

Mean and variance of $\text{Binomial}(n, \theta)$ 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 (3.10) and (3.13) by noting that $X = \sum_{i=1}^n X_i$, where the $\{X_1, X_2, \dots, X_n\} \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta)$, $\mathbf{E}(X_i) = \theta$ and $\mathbf{V}(X_i) = \theta(1 - \theta)$:

$$\begin{aligned}\mathbf{E}(X) &= \mathbf{E}(X_1 + X_2 + \dots + X_n) = \mathbf{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbf{E}(X_i) = n\theta , \\ \mathbf{V}(X) &= \mathbf{V}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbf{V}(X_i) = \sum_{i=1}^n \theta(1 - \theta) = n\theta(1 - \theta) .\end{aligned}$$

Labwork 81 (Binomial coefficient) We may implement the MATLAB function `BinomialCoefficient` to compute:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} = \frac{n(n-1)(n-2)\dots(n-x+1)}{x(x-1)(x-2)\dots(2)(1)} = \frac{\prod_{i=(n-x+1)}^n i}{\prod_{i=2}^x i} ,$$

with the following M-file:

```
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 $\text{Binomial}(n, \theta)$ RV X as follows:

```
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 $\text{Binomial}(8, 0.5)$ RV X , as follows:

```
>> 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 82 ($\text{Binomial}(n, \theta)$ as $\sum_{i=1}^n \text{Bernoulli}(\theta)$) Since the $\text{Binomial}(n, \theta)$ RV X is the sum of n IID $\text{Bernoulli}(\theta)$ RVs we can also simulate from X by first simulating n IID $\text{Bernoulli}(\theta)$ RVs and then adding them up as follows:

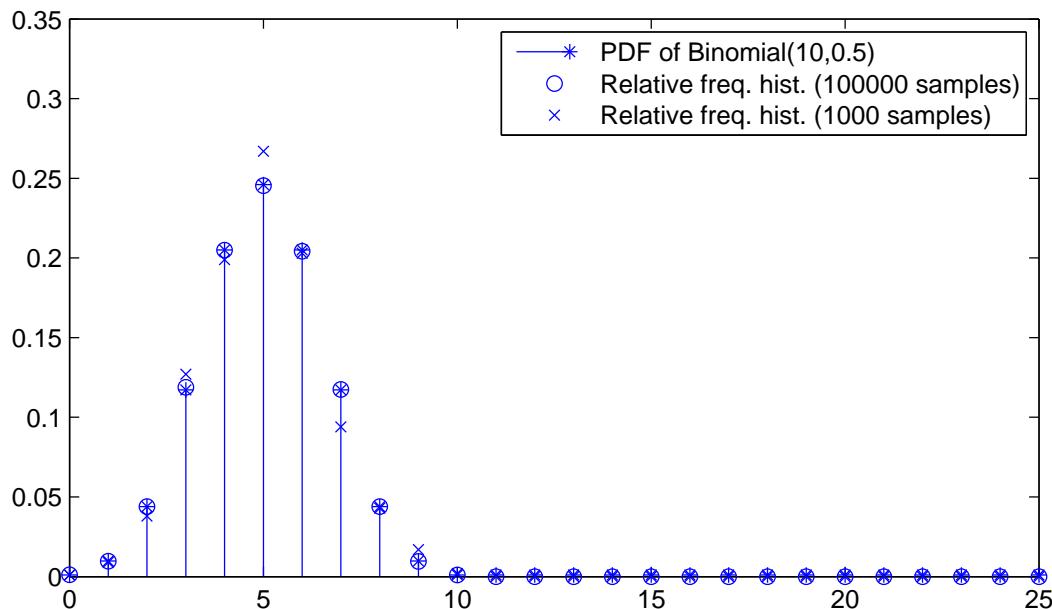
```
>> rand('twister',17678);
>> theta=0.5; % give some desired theta value, say 0.5
>> n=5; % give the parameter n for Binomial(n,theta) RV X, say n=5
>> xis=floor(rand(1,n)+theta) % produce n IID samples from Bernoulli(theta=0.5) RVs X1,X2,...Xn
xis =
    1     1     0     0     0
>> x=sum(xis) % sum 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 82, the number of IID $\text{Bernoulli}(\theta)$ RVs needed to simulate one sample from the $\text{Binomial}(n, \theta)$ RV is exactly n . Thus, as n increases, the amount of time needed to simulate from $\text{Binomial}(n, \theta)$ is $O(n)$, i.e. linear in n . We can simulate more efficiently by exploiting a simple relationship between the $\text{Geometric}(\theta)$ RV and the $\text{Binomial}(n, \theta)$ RV.

Figure 6.14: PDF of $X \sim \text{Binomial}(n = 10, \theta = 0.5)$ and the relative frequency histogram based on 100,000 samples from X .



The Binomial(n, θ) RV X is related to the IID Geometric(θ) RV Y_1, Y_2, \dots : 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 83 (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, \dots , 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:

```
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
```

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 6.14 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.

In several situations it becomes cumbersome to model the events using the Binomial(n, θ) RV, especially when the parameter $\theta \propto 1/n$ and the events become rare. 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 as defined later). 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 $X \sim \text{Binomial}(n, \theta = \lambda/n)$, then for any $x \in \{0, 1, 2, 3, \dots, n\}$,

$$\begin{aligned}
 \mathbf{P}(X = x) &= \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
 &= \frac{n(n-1)(n-2)\cdots(n-x+1)}{x(x-1)(x-2)\cdots(2)(1)} \left(\frac{\lambda^x}{n^x}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \\
 &= \underbrace{\left(\frac{n}{n}\right) \left(\frac{n-1}{n}\right) \left(\frac{n-2}{n}\right) \cdots \left(\frac{n-x+1}{n}\right)}_{\text{overbrace}} \underbrace{\left(\frac{\lambda^x}{x!}\right)}_{\text{overbrace}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\text{overbrace}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-x}}_{\text{overbrace}}
 \end{aligned} \tag{6.20}$$

As $n \rightarrow \infty$, the expression below the first overbrace $\rightarrow 1$, while that below the second overbrace, being independent of n remains the same. By the elementary examples of limits 14 and 15, as $n \rightarrow$

∞ , the expression over the first underbrace approaches $e^{-\lambda}$ while that over the second underbrace approaches 1. Finally, we get the desired limit:

$$\lim_{n \rightarrow \infty} \mathbf{P}(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} .$$

Model 14 (Poisson(λ) RV) Given a real parameter $\lambda > 0$, the discrete RV X is said to be Poisson(λ) distributed if X has PDF:

$$f(x; \lambda) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!} & \text{if } x \in \mathbb{Z}_+ := \{0, 1, 2, \dots\} , \\ 0 & \text{otherwise .} \end{cases} \quad (6.21)$$

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.

Mean and variance of Poisson(λ) RV: Let $X \sim \text{Poisson}(\lambda)$. Then:

$$\mathbf{E}(X) = \sum_{x=0}^{\infty} x f(x; \lambda) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} = e^{-\lambda} \lambda e^{\lambda} = \lambda .$$

Similarly,

$$\mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \lambda + \lambda^2 - \lambda^2 = \lambda .$$

since

$$\begin{aligned} \mathbf{E}(X^2) &= \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda} \lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{x \lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda} \left(1 + \frac{2\lambda}{1} + \frac{3\lambda^2}{2!} + \frac{4\lambda^3}{3!} + \dots \right) \\ &= \lambda e^{-\lambda} \left(\left(1 + \frac{\lambda}{1} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots \right) + \left[\frac{\lambda}{1} + \frac{2\lambda^2}{2!} + \frac{3\lambda^3}{3!} + \dots \right] \right) \\ &= \lambda e^{-\lambda} \left((e^{\lambda}) + \lambda \left(1 + \frac{2\lambda}{2!} + \frac{3\lambda^2}{3!} + \dots \right) \right) = \lambda e^{-\lambda} \left(e^{\lambda} + \lambda \left(1 + \lambda + \frac{\lambda^2}{2!} + \dots \right) \right) \\ &= \lambda e^{-\lambda} (e^{\lambda} + \lambda (e^{\lambda})) = \lambda e^{-\lambda} (e^{\lambda} + \lambda e^{\lambda}) = \lambda(1 + \lambda) = \lambda + \lambda^2 \end{aligned}$$

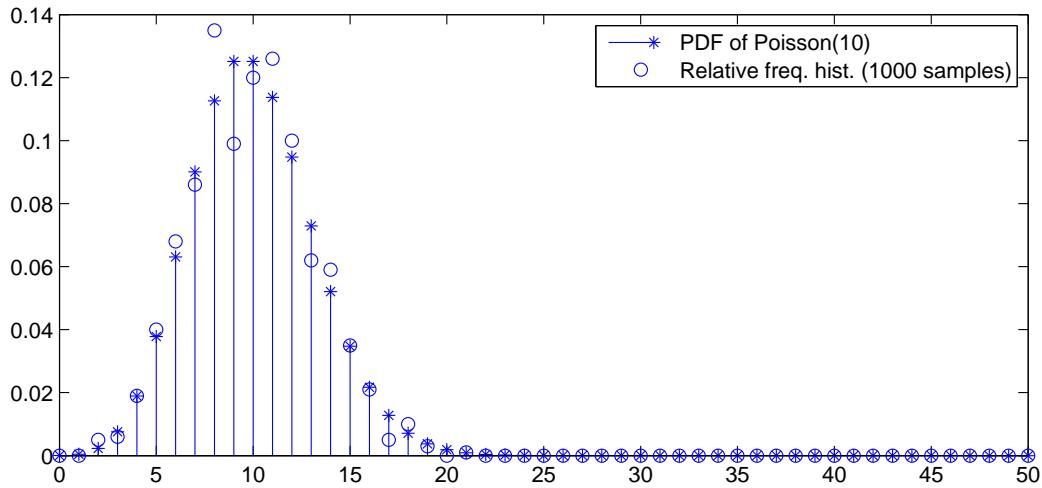
Note that Poisson(λ) distribution is one whose mean and variance are the same, namely λ .

The Poisson(λ) RV X is also related to the IID Exponential(λ) RV Y_1, Y_2, \dots : 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.

Simulation 84 (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, \dots , 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:

Figure 6.15: PDF of $X \sim \text{Poisson}(\lambda = 10)$ and the relative frequency histogram based on 1000 samples from X .



```
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 $\text{Poisson}(\lambda = 10.0)$ and $\text{Poisson}(\lambda = 0.1)$ 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 6.15 depicts a comparison of the PDF of $\text{Poisson}(\lambda = 10)$ RV and a relative frequency histogram based on 1000 simulations from it.

Simulating from a $\text{Poisson}(\lambda)$ RV is also a special case of simulating from the following more general RV.

Model 15 ($GD(\theta_0, \theta_1, \dots)$) We say X is a General Discrete($\theta_0, \theta_1, \dots$) or $GD(\theta_0, \theta_1, \dots)$ RV over the countable discrete state space $\mathbb{Z}_+ := \{0, 1, 2, \dots\}$ with parameters $(\theta_0, \theta_1, \dots)$ if the PMF of X is defined as follows:

$$f(X = x; \theta_0, \theta_1, \dots) = \begin{cases} 0, & \text{if } x \notin \{0, 1, 2, \dots\} \\ \theta_0, & \text{if } x = 0 \\ \theta_1, & \text{if } x = 1 \\ \vdots & \end{cases}$$

Algorithm 7 allows us to simulate from any member of the class of non-negative discrete RVs as specified by the probabilities $(\theta_0, \theta_1, \dots)$. 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 from \mathbb{Z}_+ to \mathbb{X} that allows us to think of $\{0, 1, 2, \dots\}$ as indices of an array D .

Algorithm 7 Inversion Sampler for $GD(\theta_0, \theta_1, \dots)$ RV X

1: *input:*

1. θ_0 and $\{C(i) = \theta_i / \theta_{i-1}\}$ for any $i \in \{1, 2, 3, \dots\}$.
2. $u \sim \text{Uniform}(0, 1)$

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: $i \leftarrow i + 1$, $p \leftarrow p C(i)$, $q \leftarrow q + p$
- 6: **end while**

7: *return:* $x = i$

Simulation 85 ($\text{Binomial}(n, \theta)$) To simulate from a $\text{Binomial}(n, \theta)$ RV X , we can use Algorithm 7 with:

$$\theta_0 = (1 - \theta)^n, \quad C(x+1) = \frac{\theta(n-x)}{(1-\theta)(x+1)}, \quad \text{Mean Efficiency: } O(1 + n\theta).$$

Similarly, with the appropriate θ_0 and $C(x+1)$, we can also simulate from the $\text{Geometric}(\theta)$ and $\text{Poisson}(\lambda)$ RVs.

Labwork 86 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 7 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 $\text{Binomial}(n, p)$ 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 $\text{Binomial}(n = 20, p = 0.5)$ 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 $\text{Binomial}(n, p)$ RV is given above].
3. Show that your code works for drawing samples from a $\text{Geometric}(p)$ RV by doing the following:
 - (a) Seed the fundamental sampler by your Student ID.

- (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].

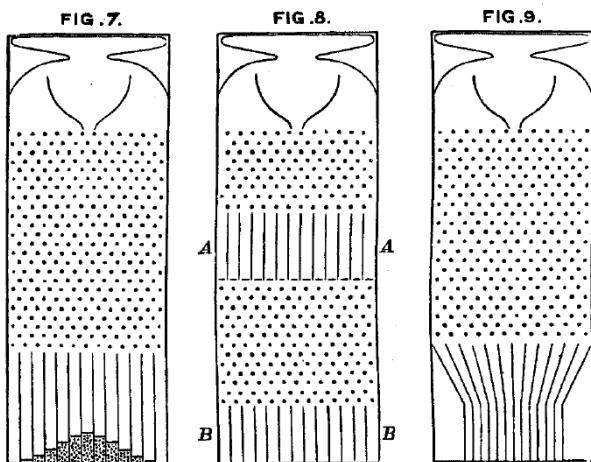
To make concrete sense of the Binomial(n, θ) and other more sophisticated concepts in the sequel, let us take a historical detour into some origins of statistical thinking in 19th century England.

6.7 Sir Francis Galton's Quincunx

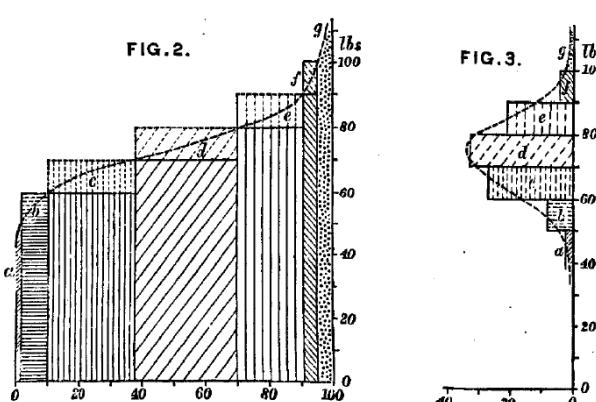
This section is introduced to provide some forms for a kinesthetic (hands-on) and visual understanding of some elementary statistical distributions and laws. The following words are from Sir Francis Galton, F.R.S., *Natural Inheritance*, pp. 62-65, Macmillan, 1889. In here you will already find the kernels 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 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 6.16: 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 mouth 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.”

Summary of Random Variables

Model	PDF	Mean	Variance
Bernoulli(θ)	$\theta^x(1-\theta)^{1-x}\mathbf{1}_{\{0,1\}}(x)$	θ	$\theta(1-\theta)$
Binomial(n, θ)	$\binom{n}{\theta}\theta^x(1-\theta)^{n-x}\mathbf{1}_{\{0,1,\dots,n\}}(x)$	$n\theta$	$n\theta(1-\theta)$
Geometric(θ)	$\theta(1-\theta)^x\mathbf{1}_{\mathbb{Z}_+}(x)$	$\frac{1}{\theta} - 1$	$\frac{1-\theta}{\theta^2}$
Poisson(λ)	$\frac{\lambda^x e^{-\lambda}}{x!}\mathbf{1}_{\mathbb{Z}_+}(x)$	λ	λ
Uniform(θ_1, θ_2)	$\mathbf{1}_{[\theta_1, \theta_2]}(x)/(\theta_2 - \theta_1)$	$\frac{\theta_1 + \theta_2}{2}$	$\frac{(\theta_2 - \theta_1)^2}{12}$
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
Gamma(α, β)	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$	α/β	α/β^2

Table 6.2: Random Variables with PDF, Mean and Variance

Exercises

Ex. 6.1 — 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. 6.2 — 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. 6.3 — 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. 6.4 — 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. 6.5 — 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. 6.6 — 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. 6.7 — Four fair coins are tossed simultaneously. If we count the number of heads that appear then we have a binomial random variable, $X = \text{the number of heads}$.

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. 6.8 — 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 than five of the donors are *not* type A.

Ex. 6.9 — 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. 6.10 — 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. 6.11 — 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. 6.12 — 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).

Ex. 6.13 — 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. 6.14 — 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?

Ex. 6.15 — The number of lacunae (surface pits) on specimens of steel, polished and examined in a metallurgical laboratory, is known to have 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?

Ex. 6.16 — Find the probability that none of the three bulbs in a traffic signal 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

$$f(x) = \begin{cases} 6(0.25 - (x - 1.5)^2) & 1 < x < 2 \\ 0 & \text{otherwise} \end{cases} .$$

Note: X is measured in multiples of 1000 hours.

Ex. 6.17 — 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.

Ex. 6.18 — **Starting from the definition of the variance of a random variable (Definition 22) show that

$$\mathbf{V}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 .$$

Ex. 6.19 — **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 .

6.8 Random Vectors

Let us try to relate some discrete probability models to the Quincunx. First, we need to introduce simple random vectors (\vec{RV}), i.e. ordered pairs, ordered triples, or more generally ordered m -tuples of random variables (X_1, X_2, \dots, X_m) . We focus on elementary definitions needed to define bivariate \vec{RV} obtained from a pair of RVs. Here is a simple example of a discrete bivariate \vec{RV} that illustrates the notions of joint and marginal probabilities.

Ex. 6.20 — Example 87 (Pair of Bernoulli(1/2) RVs) Let X_1 and X_2 be a pair of IID Bernoulli(1/2) RVs each taking values in the set $\{0, 1\}$ with the following joint probabilities:

	$X_2 = 0$	$X_2 = 1$	
$X_1 = 0$	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 $\mathbf{P}((X_1, X_2) = (0, 0)) = 1/4$ and that the marginal probability $\mathbf{P}(X_1 = 0) = 1/2$.

Definition 39 (Joint PDF, PMF, CDF) A function $f(x_1, x_2)$ is called a **joint PDF (or PMF)** for the ordered pair of random variables (X_1, X_2) if:

1. $f(x_1, x_2) \geq 0$ for all $(x_1, x_2) \in \mathbb{R}^2$

- 2.

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dF(x_1, x_2) = \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 dx_2 & \text{if } (X_1, X_2) \text{ are continuous} \\ \sum_{x_1} \sum_{x_2} f(x_1, x_2) & \text{if } (X_1, X_2) \text{ are discrete} \end{cases}$$

3. for any event $A \subset \mathbb{R}^2$,

$$\mathbf{P}(A) = \int \int_A dF(x_1, x_2) = \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbf{1}_A((x_1, x_2)) f(x_1, x_2) dx_1 dx_2 & \text{if } (X_1, X_2) \text{ are continuous} \\ \sum_{x_1} \sum_{x_2} \mathbf{1}_A((x_1, x_2)) f(x_1, x_2) & \text{if } (X_1, X_2) \text{ are discrete} \end{cases}$$

The **joint CDF or joint DF** for discrete or continuous \vec{RV} (X_1, X_2) is:

$$F(x_1, x_2) := \mathbf{P}(X_1 \leq x_1, X_2 \leq x_2).$$

Definition 40 (Marginal PDF or PMF) If the \vec{RV} (X_1, X_2) has $f(x_1, x_2)$ as its joint density, i.e. joint PDF or joint PMF, then the **marginal PDF or PMF** of X_1 is defined by:

$$f(x_1) = \mathbf{P}(X_1 = x_1) = \begin{cases} \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 & \text{if } (X_1, X_2) \text{ are continuous} \\ \sum_{x_2} f(x_1, x_2) & \text{if } (X_1, X_2) \text{ are discrete} \end{cases}$$

and the **marginal PDF or PMF** of X_2 is defined by:

$$f(x_2) = \mathbf{P}(X_2 = x_2) = \begin{cases} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 & \text{if } (X_1, X_2) \text{ are continuous} \\ \sum_{x_1} f(x_1, x_2) & \text{if } (X_1, X_2) \text{ are discrete} \end{cases}$$

Example 88 (Bivariate Uniform) Let (X_1, X_2) be uniformly distributed on the square $[0, 1]^2 := [0, 1] \times [0, 1]$. Then,

$$f(x_1, x_2) = \mathbf{1}_{[0,1]^2}(x_1, x_2) .$$

Let the rectangular event $A = \{X_1 < 1/3, Y < 1/2\} \subset [0, 1]^2$. By integrating the joint PDF over A , which amounts here to finding the area of A , we compute $\mathbf{P}(A) = (1/3)(1/2) = 1/6$. Note that the marginal PDF of X_1 or X_2 is the PDF of the Uniform(0, 1) RV.

Definition 41 (Conditional PDF or PMF) Let (X_1, X_2) be a discrete bivariate RV. The conditional PMF of $X_1|X_2 = x_2$, where $f(X_2 = x_2) := \mathbf{P}(X_2 = x_2) > 0$ is:

$$f(x_1|x_2) := \mathbf{P}(X_1 = x_1|X_2 = x_2) = \frac{\mathbf{P}(X_1 = x_1, X_2 = x_2)}{\mathbf{P}(X_2 = x_2)} = \frac{f(x_1, x_2)}{f(x_2)} .$$

Similarly, if $f(X_1 = x_1) > 0$, then the conditional PMF of $X_2|X_1 = x_1$ is:

$$f(x_2|x_1) := \mathbf{P}(X_2 = x_2|X_1 = x_1) = \frac{\mathbf{P}(X_1 = x_1, X_2 = x_2)}{\mathbf{P}(X_1 = x_1)} = \frac{f(x_1, x_2)}{f(x_1)} .$$

If (X_1, X_2) are continuous RVs such that $f(x_2) > 0$, then the conditional PDF of $X_1|X_2 = x_2$ is:

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f(x_2)}, \quad \mathbf{P}(X_1 \in A|X_2 = x_2) = \int_A f(x_1|x_2) dx_1 .$$

Similarly, if $f(x_1) > 0$, then the conditional PDF of $X_2|X_1 = x_1$ is:

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f(x_1)}, \quad \mathbf{P}(X_2 \in A|X_1 = x_1) = \int_A f(x_2|x_1) dx_2 .$$

We need a new notion for the variance of two RVs.

Definition 42 (Covariance) Suppose X_1 and X_2 are random variables, such that $\mathbf{E}(X_1^2) < \infty$ and $\mathbf{E}(X_2)^2 < \infty$. Then, $\mathbf{E}(|X_1 X_2|) < \infty$ and $\mathbf{E}(|(X_1 - \mathbf{E}(X_1))(X_2 - \mathbf{E}(X_2))|) < \infty$. We therefore define the covariance $\mathbf{Cov}(X_1, X_2)$ of X_1 and X_2 as:

$$\mathbf{Cov}(X_1, X_2) := \mathbf{E}((X_1 - \mathbf{E}(X_1))(X_2 - \mathbf{E}(X_2))) = \mathbf{E}(X_1 X_2) - \mathbf{E}(X_1)\mathbf{E}(X_2)$$

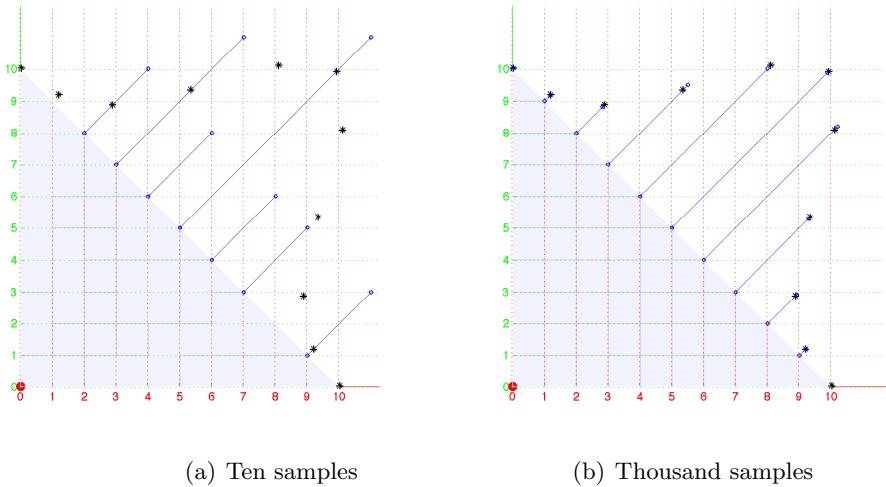
Let us consider the natural two-dimensional analogue of the Bernoulli(θ) RV in the real plane $\mathbb{R}^2 := (-\infty, \infty)^2 := (-\infty, \infty) \times (-\infty, \infty)$. A natural possibility is to use the **ortho-normal basis vectors** in \mathbb{R}^2 :

$$\boxed{e_1 := (1, 0), \quad e_2 := (0, 1)} .$$

Recall that vector addition and subtraction are done component-wise, i.e. $(x_1, x_2) \pm (y_1, y_2) = (x_1 \pm y_1, x_2 \pm y_2)$.

Classwork 89 (Geometry of Vector Addition) Recall elementary vector addition in the plane. What is $(1, 0) + (1, 0)$, $(1, 0) + (0, 1)$, $(0, 1) + (0, 1)$? What is the relationship between $(1, 0)$, $(0, 1)$ and $(1, 1)$ geometrically? How does the diagonal of the parallelogram relate to its two sides in the geometry of addition in the plane? What is $(1, 0) + (0, 1) + (1, 0)$?

Figure 6.17: Quincunx on the Cartesian plane. Simulations of $\text{Binomial}(n = 10, \theta = 0.5)$ 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 $\text{Bernoulli}(\theta = 0.5)$ 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 PDF of our $\text{Binomial}(n = 10, \theta = 0.5)$ RV, respectively.



Model 16 ($\text{Bernoulli}(\theta)$ $\vec{\text{RV}}$) Given a parameter $\theta \in [0, 1]$, we say that $X := (X_1, X_2)$ is a $\text{Bernoulli}(\theta)$ random vector ($\vec{\text{RV}}$) 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 $\vec{\text{RV}}$ $X := (X_1, X_2)$ with realisation $x := (x_1, x_2)$ is:

$$f(x; \theta) := \mathbf{P}(X = x) = \theta \mathbf{1}_{\{e_1\}}(x) + (1 - \theta) \mathbf{1}_{\{e_2\}}(x) = \begin{cases} \theta & \text{if } x = e_1 := (1, 0) \\ 1 - \theta & \text{if } x = e_2 := (0, 1) \\ 0 & \text{otherwise} \end{cases}$$

Classwork 90 (Expectation and Variance of $\text{Bernoulli}(\theta)$ $\vec{\text{RV}}$) What is the Expectation of $\text{Bernoulli}(\theta)$ $\vec{\text{RV}}$?

$$\mathbf{E}_\theta(X) = \mathbf{E}_\theta((X_1, X_2)) = \sum_{(x_1, x_2) \in \{e_1, e_2\}} (x_1, x_2) f((x_1, x_2); \theta) = (1, 0)\theta + (0, 1)(1 - \theta) = (\theta, 1 - \theta).$$

How about the variance? [Hint: Use the definitions of $\mathbf{E}(X)$ and $\mathbf{V}(X)$ for the $\vec{\text{RV}}$ X . $\mathbf{E}(X^2)$ is not a single number and you may need new words such as covariance to deal with terms like $\mathbf{E}(X_1 X_2)$.]

We can write the $\text{Binomial}(n, \theta)$ RV Y as a $\text{Binomial}(n, \theta)$ $\vec{\text{RV}}$ $X := (Y, n - Y)$. In fact, this is the underlying model and the **bi** in the $\text{Binomial}(n, \theta)$ 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 $\text{Binomial}(n, \theta)$ $\vec{\text{RV}}$ $(Y, n - Y)$ is the sum of n IID $\text{Bernoulli}(\theta)$

\vec{RV} s $X_1 := (X_{1,1}, X_{1,2}), X_2 := (X_{2,1}, X_{2,2}), \dots, X_n := (X_{n,1}, X_{n,2})$:

$$(Y, n - Y) = X_1 + X_2 + \dots + X_n = (X_{1,1}, X_{1,2}) + (X_{2,1}, X_{2,2}) + \dots + (X_{n,1}, X_{n,2})$$

Go the Biomathematics Research Centre on the 6th floor of Erskine to play with the Quincunx built by Ryan Lawrence in 2007 (See the project by Ashman and Lawrence at <http://www.math.canterbury.ac.nz/~r.sainudiin/courses/STAT218/projects/Stat218StudentProjects2007.pdf> for details). It is important to gain a physical intimacy with the Quincunx to appreciate the following model of it. We can make a statistical model of Galton's observations earlier regarding the dynamics of lead shots through the Quincunx as the sum of n IID Bernoulli(0.5) \vec{RV} s, where n is number of pegs that each ball bounces on before making a left or right turn with equal probability.

Exercise 91 (Number of paths and the binomial coefficient) How does the number of paths that lead to a bucket (x_1, x_2) with $x_1 + x_2 = n$ relate to the binomial coefficient $\binom{n}{x_1}$?

Labwork 92 (Quincunx Sampler Demo – Sum of n IID Bernoulli(1/2) \vec{RV} s) Let us understand the Quincunx construction of the Binomial($n, 1/2$) $\vec{RV} X$ as the sum of n independent and identical Bernoulli(1/2) \vec{RV} s by calling the interactive visual cognitive tool as follows:

```
>> guiMultinomial
```

The M-file `guiMultinomial.m` will bring a graphical user interface (GUI) as shown in Figure 6.18. Using the drop-down menu at “How many levels?” change the number of levels to 2 ($n = 2$). Now click the “Do one” button as many times as you like and comprehend the simulation process – the path taken by the ball as it falls through two levels. Next, from the drop-down menu at “How many Replication?” change it from 10 to 100. You can press “Do all” to watch all 100 balls drop into their possible values at level 2. Change the number of levels or n in Binomial($n, 1/2$) \vec{RV} to 3 or 5 or 10 and do more simulations until you are comfortable with the construction that the sum of n IID Bernoulli(1/2) \vec{RV} s is the Binomial($n, 1/2$) \vec{RV} .

When we drop 1000 balls into the simulated Quincunx the density histogram is much closer to the PDF of Binomial($n = 10, \theta = 0.5$) RV than when we only drop 10 balls. See Figure 6.17 for a description of the simulations. Try to replicate such a simulation on your own.

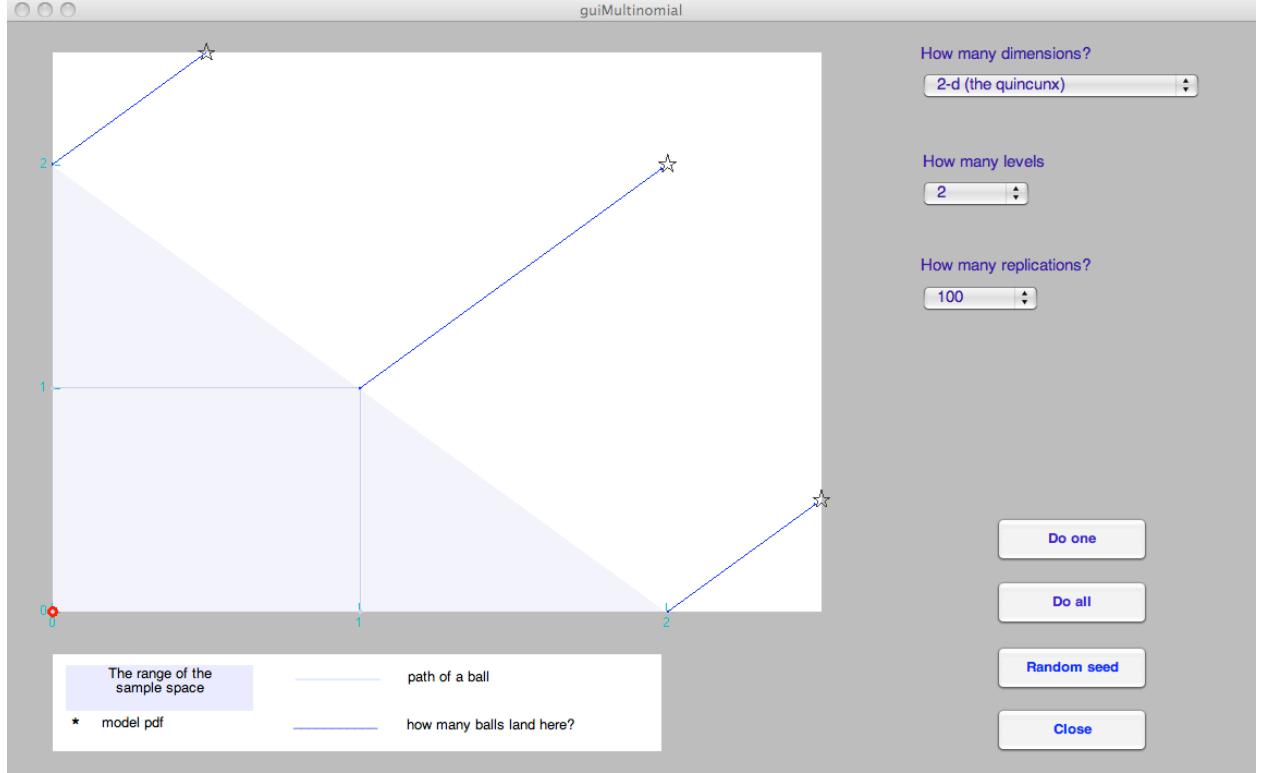
We are now ready to extend the Binomial(n, θ) RV or \vec{RV} to its multivariate version called the Multinomial($n, \theta_1, \theta_2, \dots, \theta_k$) \vec{RV} . We develop this \vec{RV} as the sum of n IID de Moivre($\theta_1, \theta_2, \dots, \theta_k$) \vec{RV} that is defined next.

Model 17 (de Moivre($\theta_1, \theta_2, \dots, \theta_k$) \vec{RV}) The PMF of the de Moivre($\theta_1, \theta_2, \dots, \theta_k$) $\vec{RV} X := (X_1, X_2, \dots, X_k)$ taking value $x := (x_1, x_2, \dots, x_k) \in \{e_1, e_2, \dots, e_k\}$, where the e_i 's are orthonormal basis vectors in \mathbb{R}^k is:

$$f(x; \theta_1, \theta_2, \dots, \theta_k) := \mathbf{P}(X = x) = \sum_{i=1}^k \theta_i \mathbf{1}_{\{e_i\}}(x) = \begin{cases} \theta_1 & \text{if } x = e_1 := (1, 0, \dots, 0) \in \mathbb{R}^k \\ \theta_1 & \text{if } x = e_1 := (0, 1, \dots, 0) \in \mathbb{R}^k \\ \vdots & \\ \theta_k & \text{if } x = e_k := (0, 0, \dots, 1) \in \mathbb{R}^k \\ 0 & \text{otherwise} \end{cases}$$

Of course, $\sum_{i=1}^k \theta_i = 1$.

Figure 6.18: Visual Cognitive Tool GUI: Quincunx.



When we add n IID $\text{de Moivre}(\theta_1, \theta_2, \dots, \theta_k)$ R \vec{V} s together, we get the $\text{Multinomial}(n, \theta_1, \theta_2, \dots, \theta_k)$ R \vec{V} s as defined below.

Model 18 ($\text{Multinomial}(n, \theta_1, \theta_2, \dots, \theta_k)$ R \vec{V}) We say that a R \vec{V} $Y := (Y_1, Y_2, \dots, Y_k)$ obtained from the sum of n IID $\text{de Moivre}(\theta_1, \theta_2, \dots, \theta_k)$ R \vec{V} s with realisations

$$y := (y_1, y_2, \dots, y_k) \in \mathbb{Y} := \{(y_1, y_2, \dots, y_k) \in \mathbb{Z}_+^k : \sum_{i=1}^k y_i = n\}$$

has the PMF given by:

$$f(y; n, \theta) := f(y; n, \theta_1, \theta_2, \dots, \theta_k) := \mathbf{P}(Y = y; n, \theta_1, \theta_2, \dots, \theta_k) = \binom{n}{y_1, y_2, \dots, y_k} \prod_{i=1}^k \theta_i^{y_i},$$

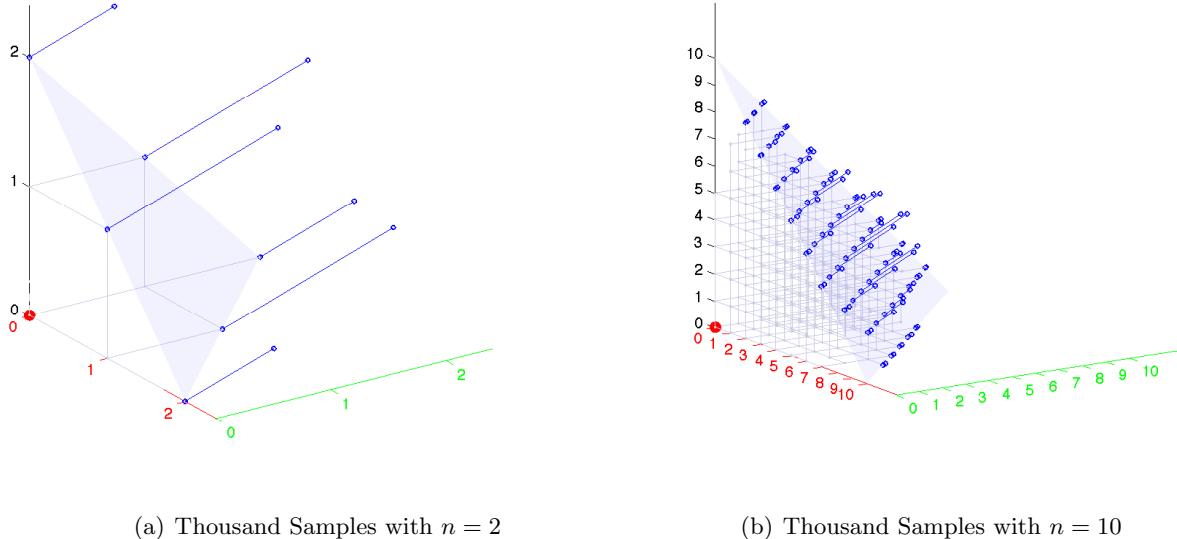
where, the multinomial coefficient:

$$\binom{n}{y_1, y_2, \dots, y_k} := \frac{n!}{y_1! y_2! \cdots y_k!}.$$

Note that the marginal PMF of Y_j is $\text{Binomial}(n, \theta_j)$ for any $j = 1, 2, \dots, k$.

We can visualise the $\text{Multinomial}(n, \theta_1, \theta_2, \theta_3)$ process as a sum of n IID $\text{de Moivre}(\theta_1, \theta_2, \theta_3)$ R \vec{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, $\mathbb{Y} = \{(y_1, y_2, y_3) \in \mathbb{Z}_+^3 : \sum_{i=1}^3 y_i = n\}$

Figure 6.19: Septcunx on the Cartesian co-ordinates. Simulations of Multinomial($n = 2, \theta_1 = 1/3, \theta_2 = 1/3, \theta_3 = 1/3$) \vec{RV} as the sum of n IID de Moivre($\theta_1 = 1/3, \theta_2 = 1/3, \theta_3 = 1/3$) \vec{RV} s over $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ with probabilities $\{\theta_1, \theta_2, \theta_3\}$, respectively. The blue lines perpendicular to the sample space of the Multinomial($3, \theta_1, \theta_2, \theta_3$) \vec{RV} , i.e. the plane in \mathbb{R}^3 connecting $(n, 0, 0)$, $(0, n, 0)$ and $(0, 0, n)$, are the density histogram of the samples.



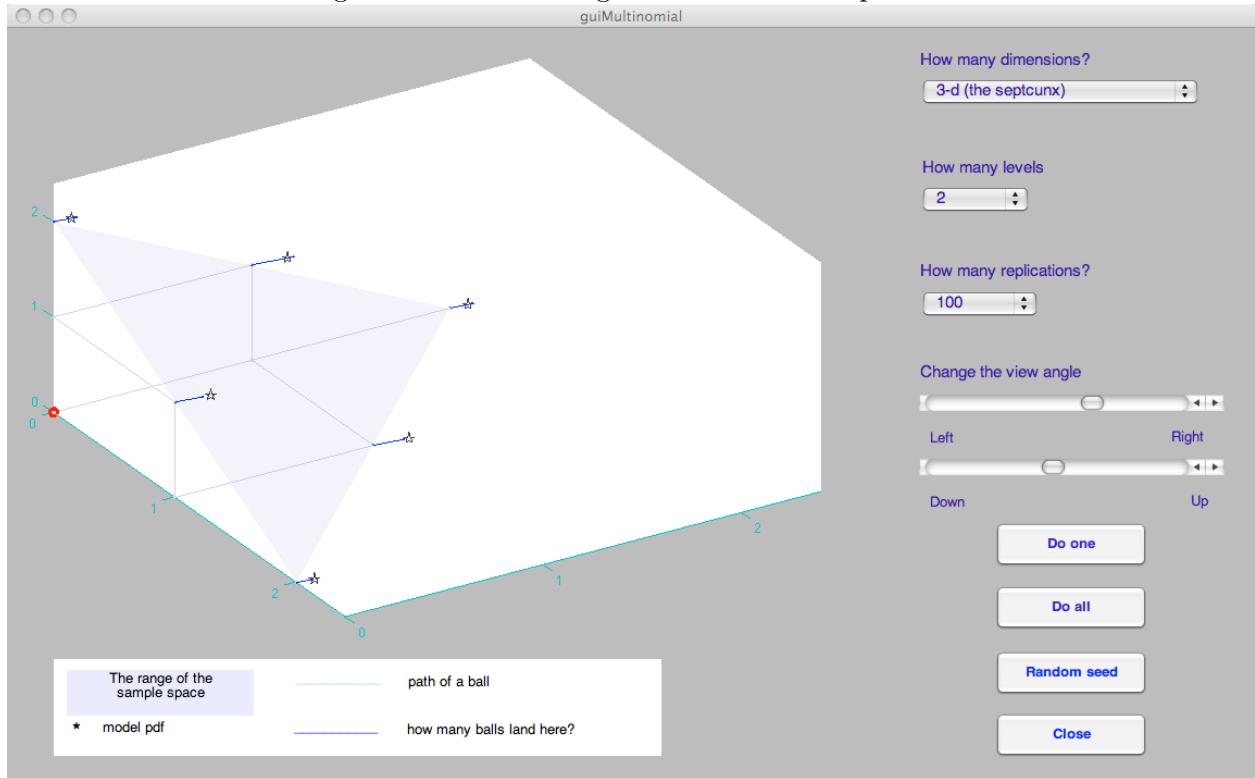
$n\}$. Once again, we can visualise that the sum of n IID de Moivre($\theta_1, \theta_2, \theta_3$) \vec{RV} s constitute the Multinomial($n, \theta_1, \theta_2, \theta_3$) \vec{RV} as depicted in Figure 6.19.

Labwork 93 (Septcunx Sampler Demo – Sum of n IID de Moivre(1/3, 1/3, 13/) \vec{RV} s) Let us understand the Septcunx construction of the Multinomial($n, 1/3, 1/3, 1/3$) \vec{RV} X as the sum of n independent and identical de Moivre($1/3, 1/3, 13/$) \vec{RV} s by calling the interactive visual cognitive tool as follows:

```
>> guiMultinomial
```

The M-file `guiMultinomial.m` will bring a GUI as shown in Figure 6.18. Using the drop-down menu at “How many dimensions?” change to “3-d (the septcunx)” and you will see a septcunx as shown in Figure 6.20. Next, using the drop-down menu at “How many levels?” change the number of levels to 2 ($n = 2$). Now click the “Do one” button as many times as you like and comprehend the simulation process – the path taken by the ball as it falls through two levels in three dimensional space. Feel free to change the up-down and left-right sliders for the view angles. Next, from the drop-down menu at “How many Replication?” change it from 10 to 100. You can press “Do all” to watch all 100 balls drop into their possible values at level 2. Change the number of levels or n in Multinomial($n, 1/3, 1/3, 1/3$) \vec{RV} to 5 or 10 and do more simulations until you are comfortable with the construction that the sum of n IID de Moivre($1/3, 1/3, 1/3$) \vec{RV} s is the Multinomial($n, 1/3, 1/3, 1/3$) \vec{RV} .

Figure 6.20: Visual Cognitive Tool GUI: Septcunx.



Labwork 94 (PDF of Multinomial(n, θ) \vec{R}) We can implement the following MATLAB function `MultinomialPdf` to compute the PDF of the Multinomial(n, θ) \vec{R} where $\theta := (\theta_1, \theta_2, \dots, \theta_k)$ is a point in the k -simplex Δ_k as follows:

```
function MP = MultinomialPdf(x,n,theta)
% returns the multinomial Pdf of x(1),x(2),...,x(k) given
% theta(1),...,theta(k). x and theta are vectors and sum to
% the scalars n and 1, respectively and 0 <= x(i) <= n
% Since double precision numbers only have about 15 digits, the answer is
% only accurate for n <= 21 in factorial function.
NonZeroXs = find(x>0);
MP=exp(log(factorial(n))+sum((log(theta(NonZeroXs)) .* x(NonZeroXs)) ...
 - log(factorial(x(NonZeroXs)))));
```

We can call this function to evaluate the PDF at a specific sample $x = (x_1, x_2, \dots, x_k)$ as follows:

```
>> MultinomialPdf([2 0 0],2,[1/3 1/3 1/3])
ans =    0.1111
>> MultinomialPdf([0 2 0],2,[1/3 1/3 1/3])
ans =    0.1111
>> MultinomialPdf([0 0 2],2,[1/3 1/3 1/3])
ans =    0.1111
>> MultinomialPdf([1 1 0],2,[1/3 1/3 1/3])
ans =    0.2222
>> MultinomialPdf([1 0 1],2,[1/3 1/3 1/3])
ans =    0.2222
>> MultinomialPdf([0 1 1],2,[1/3 1/3 1/3])
ans =    0.2222
```

Simulation 95 (A simple multinomial simulation) Using the identity matrix I in \mathbb{R}^3 that can be created in MATLAB using the `eye(3)` command, and the `de Moivre(1/3, 1/3, 1/3)` RV sampler, simulate vector-valued samples from `de Moivre(1/3, 1/3, 1/3)` \vec{RV} . Finally add up $n = 10$ samples from `de Moivre(1/3, 1/3, 1/3)` \vec{RV} to produce samples from `Multinomial(10, 1/3, 1/3, 1/3)` \vec{RV} .

6.9 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:

$$f(x) \leq ag(x). \quad (6.22)$$

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

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 43 (Fundamental Theorem of Simulation) The von Neumann rejection sampler of Algorithm 8 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 :

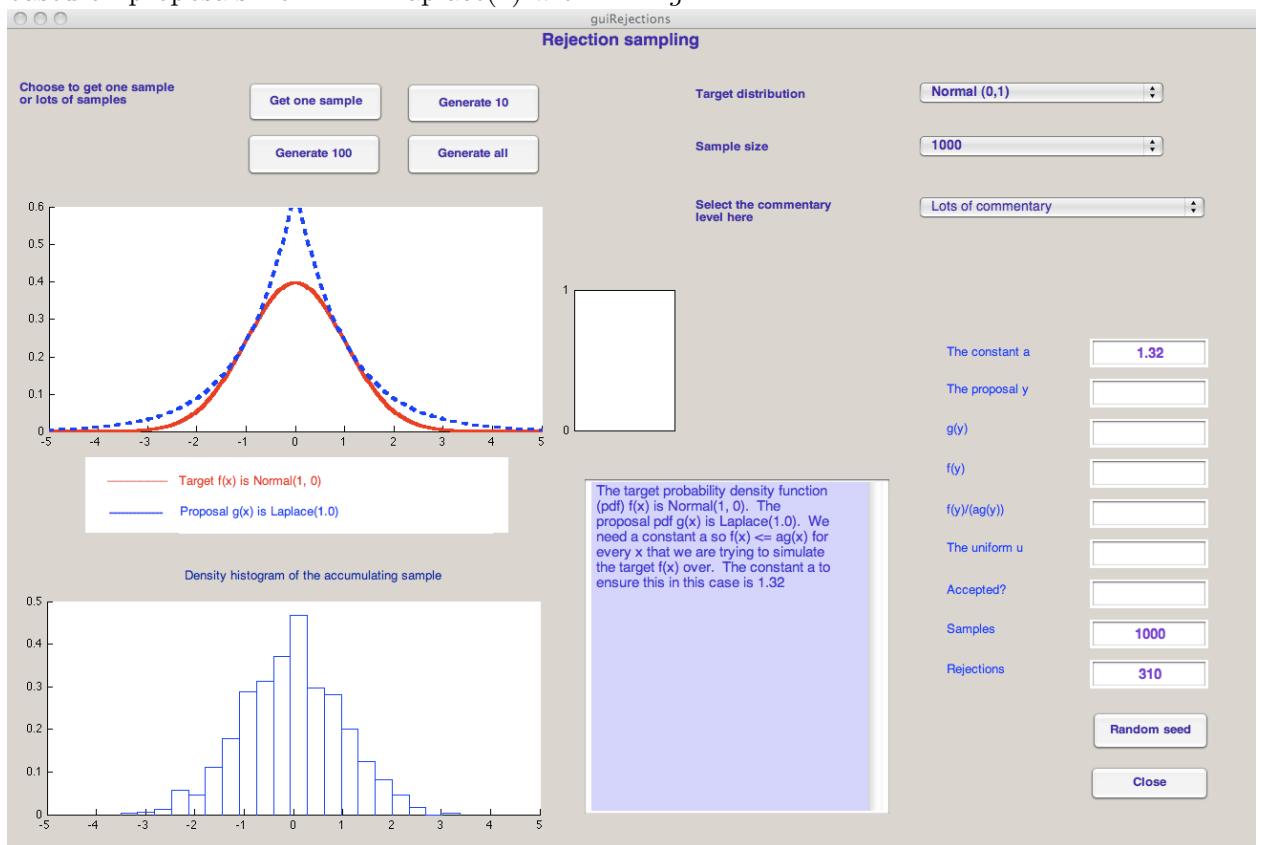
$$\begin{aligned} F(t) &= \mathbf{P}(X \leq t) = \mathbf{P}\left(Y \leq t \mid U \leq \frac{f(Y)}{ag(Y)}\right) = \frac{\mathbf{P}\left(Y \leq t, U \leq \frac{f(Y)}{ag(Y)}\right)}{\mathbf{P}\left(U \leq \frac{f(Y)}{ag(Y)}\right)} \\ &= \frac{\int_{-\infty}^t \left(\int_0^{f(y)/ag(y)} 1 du\right) g(y) dy}{\int_{-\infty}^{\infty} \left(\int_0^{f(y)/ag(y)} 1 du\right) g(y) dy} = \frac{\int_{-\infty}^t \left(\frac{f(y)}{ag(y)}\right) g(y) dy}{\int_{-\infty}^{\infty} \left(\frac{f(y)}{ag(y)}\right) g(y) dy} \\ &= \int_{-\infty}^t f(y) dy \end{aligned}$$

Labwork 96 (Rejection Sampler Demo) Let us understand the rejection sampler by calling the interactive visual cognitive tool:

```
>> guiRejections
```

The M-file `guiRejections.m` will bring a graphical user interface (GUI) as shown in Figure 6.21. 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 6.21: 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 97 (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 67). 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:

$$\frac{f(y)}{ag(y)} = \frac{f(y)}{g(y)a} = \sqrt{\frac{2}{\pi}} \exp\left(|y| - \frac{y^2}{2}\right) \frac{1}{\sqrt{\frac{2}{\pi}} \exp\left(\frac{1}{2}\right)} = \exp\left(|y| - \frac{y^2}{2} - \frac{1}{2}\right)$$

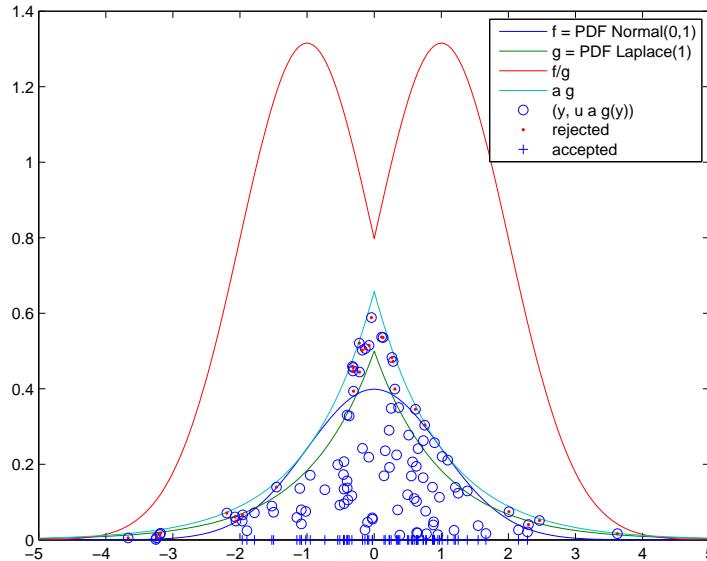
Let us implement a rejection sampler as a function in the M-file `RejectionNormalLaplace.m` by reusing the function in `LaplaceInvCDF.m`.

```
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 and 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 6.22: 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 98 (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 44 (Acceptance Probability of RS) The expected number of iterations of the rejection algorithm to get a sample x is the constant a .

Proof: For the continuous case:

$$\mathbf{P}(\text{'accept } y') = \mathbf{P}\left(u \leq \frac{f(y)}{ag(y)}\right) = \int_{-\infty}^{\infty} \left(\int_0^{f(y)/ag(y)} du \right) g(y) dy = \int_{-\infty}^{\infty} \frac{f(y)}{ag(y)} g(y) dy = \frac{1}{a}.$$

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:

$$\tilde{f}(x) \leq \tilde{a}\tilde{g}(x), \quad (6.23)$$

for any $x \in \mathbb{X}$, the support of X . Then x can be generated from Algorithm 9.

Algorithm 9 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 $u \sim \text{Uniform}(0, 1)$

5: **until** $u \leq \frac{\tilde{f}(y)}{\tilde{a}\tilde{g}(y)}$

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 $Z \sim \text{Normal}(0, 1)$ RV. The MATLAB function `randn` uses this method to produce samples from Z . See http://www.mathworks.com/company/newsletters/news_notes/clevescorner/spring01_cleve.html or http://en.wikipedia.org/wiki/Ziggurat_algorithm for more details.

Labwork 99 (Gaussian Sampling with randn) We can use MATLAB function `randn` that implements the Ziggurat method to draw samples from an RV $Z \sim \text{Normal}(0, 1)$ as follows:

```
>> 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 $Z \sim \text{Normal}(0, 1)$:

$$X \leftarrow \mu + \sigma Z, \quad Z \sim \text{Normal}(0, 1).$$

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 100 (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 $\text{Normal}(\mu, \sigma^2)$ 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.
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 .

6.10 Importance Resampler

The rejection method cannot be used when the constant a or \tilde{a} that guarantees the envelope condition cannot be found. The importance resampler, also known as the method of sampling/importance resampling, does not require the constant, but it produces a random variable that is only approximately distributed according to f . As for the rejection method, we need a density/mass function g that we can generate from and that has support at least as large as the support of f .

Algorithm 10 Importance Resampler

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 only (a) and (b) above.
- (3) a large enough integer m .

2: *output:* a sample x' from RV X' with density f' that is close to f

3: Generate $y_1, \dots, y_m \sim g$

4: Compute

$$w_i = \frac{f(y_i)/g(y_i)}{\sum_{j=1}^m f(y_j)/g(y_j)}, i = 1, \dots, m .$$

5: Resample x' from $\{y_1, \dots, y_m\}$ with weights $\{w_1, \dots, w_m\}$

Proposition 45 The Importance Resampler of Algorithm 10 produces samples from a variable X' that is approximately distributed according to f , in the sense that:

$$\lim_{m \rightarrow \infty} \mathbf{P}(X' \leq t) = \int_{-\infty}^t f(x) dx \quad (6.24)$$

for any real number t .

Proof:

$$\begin{aligned} \mathbf{P}(X' \leq t) &= \sum_{i=1}^m w_i I_{(-\infty, t]}(y_i) = \frac{\frac{1}{m} \sum_{i=1}^m \frac{f(y_i)}{g(y_i)} I_{(-\infty, t]}(y_i)}{\frac{1}{m} \sum_{i=1}^m \frac{f(y_i)}{g(y_i)}} \\ &\xrightarrow{m \rightarrow \infty} \frac{E[\frac{f(y)}{g(y)} I_{(-\infty, t]}(y)]}{E[\frac{f(y)}{g(y)}]} = \frac{\int_{-\infty}^t f(y) dy}{\int_{-\infty}^{\infty} f(y) dy} = \int_{-\infty}^t f(y) dy \end{aligned}$$

Let us visualise the Importance Resampler in action from Labwork ??.

Labwork 101 (Cauchy RV via Importance Resampler) Use the sampling/importance resampling method to generate 1000 approximate Cauchy samples by using the $\text{Normal}(0, 1)$ samples:

$$f(x) = \frac{1}{\pi(1+x^2)} \text{ and } g(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

```
n = 1000;
m = 10000;
y = randn(1,m); % randn is the N(0,1) generator in Matlab
y2 = y .* y;
w = exp(0.5 * y2) ./ (1 + y2);
w = w / sum(w);
x = randsample(y,n,true,w); % resample n values from y weighted by w
```

Note that to get n sample points from f using sampling/importance resampling, we must start with a sample from g of size m larger than n .

As for the rejection method, the sampling/importance resampling method can still be used if only the un-normalised form of f or g (or both) is known, simply by using the un-normalised densities/mass functions to compute the weights.

6.11 Other Continuous Random Variables

Here, we see other common continuous RVs that can be simulated from transforming RVs we have already encountered.

Simulation 102 (Gamma(λ, k) for integer k) Using this relationship we can simulate from $X \sim \text{Gamma}(\lambda, k)$, for an integer-valued k , by simply summing k IID samples from $\text{Exponential}(\lambda)$ RV as follows:

```
>> lambda=0.1; %declare some lambda parameter
>> k=5; % declare some k parameter (has to be integer)
>> rand('twister',7267); % initialise the fundamental sampler
>> % sum k IID Exponential(lambda) samples for one desired sample from Gamma(lambda,k)
>> x= sum(-1/lambda*log(rand(k,1)))
x =
    28.1401
>> % sum the 10 columns of k X 10 IID Exponential(lambda) samples for 10 desired samples from Gamma(lambda,k)
>> x= sum(-1/lambda*log(rand(k,10)))
x =
    83.8150    61.2674    80.3683   103.5748    48.4454    20.2269    93.8310    56.1909    77.0656    29.0851
```

Model 19 (Lognormal(λ, ζ)) X has a Lognormal(λ, ζ) distribution if $\log(X)$ has a $\text{Normal}(\lambda, \zeta^2)$ distribution. The location parameter $\lambda = \mathbf{E}(\log(X)) > 0$ and the scale parameter $\zeta > 0$. The PDF is:

$$f(x; \lambda, \zeta) = \frac{1}{\sqrt{2\pi}\zeta x} \exp\left(-\frac{1}{2\zeta^2}(\log(x) - \lambda)^2\right), \quad x > 0 \quad (6.25)$$

No closed form expression for $F(x; \lambda, \zeta)$ exists and it is simply defined as:

$$F(x; \lambda, \zeta) = \int_0^x f(y; \lambda, \zeta) dy$$

We can express $F(x; \lambda, \zeta)$ in terms of Φ (and, in turn, via the associated error function erf) as follows:

$$F(x; \lambda, \zeta) = \Phi\left(\frac{\log(x) - \lambda}{\zeta}\right) = \frac{1}{2} \text{erf}\left(\frac{\log(x) - \lambda}{\sqrt{2}\zeta}\right) + \frac{1}{2} \quad (6.26)$$

Labwork 103 (Simulations with the Lognormal(λ_C, ζ_C) RV) Transform a sequence of samples obtained from the fundamental sampler to those from the Lognormal(λ_C, ζ_C) RV C by using only Algorithm 4 or MATLAB's `randn` as an intermediate step. [Hint: If Y is a Normal(λ, ζ^2) RV, then $Z = e^Y$ is said to be a Lognormal(λ, ζ) RV.]

1. Seed the fundamental sampler by your Student ID,
2. generate 1000 samples from an RV $C \sim \text{Lognormal}(\lambda = 10.36, \zeta = 0.26)$ by exponentiating the samples from the Normal(10.36, 0.26²) RV and
3. and report:
 - (a) how many of the samples are larger than 35000,
 - (b) the sample mean, and
 - (c) the sample standard deviation.

Beta RV

Chi-Square

F distribution

t-distribution

Weibul

Heavy-tail family

6.12 Other Random Vectors

Multivariate Normal

Uniform Distribution on Sphere

Dirichlet Distribution

Ex. 6.21 — **The covariance of two random variables X and Y is defined as

$$\mathbf{Cov}(X, Y) := \mathbf{E}((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) .$$

(a) Show, starting from the definition, that $\mathbf{Cov}(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$.

(b) When $\mathbf{Cov}(X, Y) = 0$, X and Y are said to be “uncorrelated”. Show that if X and Y are independent, then they are also uncorrelated.

Ex. 6.22 — ** Let X_1, X_2, \dots, X_n be random variables. Their joint CDF is defined as

$$F(x_1, x_2, \dots, x_n) := \mathbf{P}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) .$$

By repeated application of the definition of conditional probability, show that the joint CDF admits the following “telescopic” representation:

$$\begin{aligned} F(x_1, x_2, \dots, x_n) &= F(x_n | x_1, \dots, x_{n-1})F(x_{n-1} | x_1, \dots, x_{n-2}) \cdots F(x_2 | x_1)F(x_1) \\ &= F(x_1) \prod_{i=2}^n F(x_i | x_1, \dots, x_{i-1}), \end{aligned}$$

where, $F(x_i | x_1, \dots, x_{i-1})$ denotes the conditional probability, $\mathbf{P}(X_i \leq x_i | X_1 \leq x_1, \dots, X_{i-1} \leq x_{i-1})$.

6.13 Problems

Exercise 104 If $u \sim U[0, 1]$, show that the distribution of $1 - u$ is also $U[0, 1]$.

Exercise 105 Write a Matlab function to generate n random variables from the distribution with the following mass function:

x	1.7	3.4	5.9	7.2	9.6
$f(x)$	0.15	0.4	0.05	0.1	0.3

Use your Matlab function to generate 1000 sample values from the distribution, and compare the relative frequencies obtained with the mass function probabilities.

Exercise 106 The Laplacian distribution is also called the double exponential distribution because it can be regarded as the extension of the exponential distribution for both positive and negative values. An easy way to generate a Laplacian(0, 1) random variable is to generate an exponential(1) random variable and then change its sign to negative with probability 0.5. Write a Matlab function to generate n Laplacian(0, 1) random variables using the `exprnd` function from Exercise 2.6.5. Call your function `laprnd`. It should take n as input and produce a row vector containing the n Laplacian(0, 1) random variables as output.

Exercise 107 (a) Referring to Example 2.2.3, write a MATLAB function to generate n $N(0, 1)$ random variables using the rejection method with the Laplacian(0, 1) distribution. Include a counter for the number of iterations in your function.

(b) Use your Matlab function to generate 1000 $N(0, 1)$ random variables. Plot the density histogram for your generated values and superimpose the $N(0, 1)$ density onto it. Compare the average number of iterations to get a single $N(0, 1)$ random variable with the constant a .

(c) Now suppose that we know only the un-normalised $N(0, 1)$ and Laplacian(0, 1) densities, i.e.:

$$\tilde{f}(x) = \exp\left(-\frac{x^2}{2}\right) \text{ and } \tilde{g}(x) = \exp(-|x|)$$

What is the constant \tilde{a} for the rejection method in this case? Implement the rejection method in Matlab, including a counter for the number of iterations, and use it to generate 1000 $N(0, 1)$ random variables. Compare the average number of iterations to get a single $N(0, 1)$ random variable with a and \tilde{a} .

Exercise 108 Consider (Ross, p.64.) the use of the rejection method to generate from the density:

$$f(x) = 20x(1-x)^3.$$

for $0 \leq x \leq 1$, using the $U(0, 1)$ distribution as proposal distribution.

- (a) Show that the constant for using the rejection method is $a = 2.1094$.
- (b) Write a MATLAB function to generate n random variables from f using the rejection method. Include a counter for the number of iterations in your function.
- (c) Use your MATLAB function to generate 1000 random variables from f . Plot the density histogram for your generated values and superimpose the density curve onto it. Compare the average number of iterations to get a single random variable with the constant a .

Exercise 109 Consider (Ross, .p65.) the use of the rejection method to generate from the density:

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

for $x \geq 0$, and using the exponential distribution with mean m as proposal distribution.

- (a) Show that the constant for using the rejection method is:

$$a = \sqrt{\frac{2}{\pi e}} \frac{m^{3/2}}{(m-1)^{1/2}}$$

- (b) Show that the best exponential distribution to use is the one with a mean of $3/2$.
- (c) Write a MATLAB function to generate n random variables from f using the rejection method. Include a counter for the number of iterations in your function.
- (d) Use your MATLAB function to generate 1000 random variables from f . Plot the density histogram for your generated values and superimpose the density curve onto it. Compare the average number of iterations to get a single random variable with the constant a .

Exercise 110 (a) Referring to Example 2.3.3, implement the Matlab function to generate 1000 approximate Cauchy($0, 1$) random variables using sampling/importance resampling, starting with $m = 10,000 N(0, 1)$ sample values. Plot the density histogram for your generated values and superimpose the Cauchy($0, 1$) density onto it.

(b) Explore what happens if you start with (i) $m = 1000N(0, 1)$ sample values, (ii) $m = 100000N(0, 1)$ sample values.

Exercise 111 Write a MATLAB function to generate 1000 approximate Laplacian($0, 1$) random variables using sampling/importance resampling with the $N(0, 1)$ distribution. Plot the density histogram for your generated values and superimpose the Laplacian($0, 1$) density onto it.

Exercise 112 Implement the RWMH sampler in Example 2.4.8. Perform 10,000 iterations and plot the outputs sequentially. Comment on the appearance of the plot with regard to convergence to the target density. Plot the density histogram for the last 5000 iterations and superimpose the target density onto it. Investigate what happens when $g(\cdot|x) = U(x - c, x + c)$ is used as the proposal density with different values of c that are smaller or larger than 1. (Note: In MATLAB , the modified Bessel function of the first kind is available as `besseli`.)

Chapter 7

Statistical Experiments

7.1 Introduction

We formalize the notion of a staistical experiment. Let us first motivate the need for a statistical experiment. Recall that statistical inference or learning is the process of using observations or data to infer the distribution that generated it. A generic question is:

Given realizations from $X_1, X_2, \dots, X_n \sim$ some unknown DF F , how do we infer F ?

However, to make this question tractable or even sensible it is best to restrict ourselves to a particular class or family of DFs that may be assumed to contain the unknown DF F .

Definition 46 (Experiment) A statistical experiment \mathcal{E} is a set of probability distributions (DFs, PDFs or PMFs) $\mathbb{P} := \{P_\theta : \theta \in \Theta\}$ associated with a RV X and indexed by the set Θ . We refer to Θ as the parameter space or the index set and $d : \Theta \rightarrow \mathbb{P}$ that associates to each $\theta \in \Theta$ a probability $P_\theta \in \mathbb{P}$ as the index map.

7.2 Some Common Experiments

Next, let's formally consider some experiments we have already encountered.

Experiment 20 (The Fundamental Experiment) The ‘uniformly pick a number in the interval $[0, 1]$ ’ experiment is the following singleton family of DFs :

$$\mathbb{P} = \{ F(x) = x\mathbf{1}_{[0,1]}(x) \}$$

where, the only distribution $F(x)$ in the family \mathbb{P} is a re-expression of (3.7) using the indicator function $\mathbf{1}_{[0,1]}(x)$. The parameter space of the fundamental experiment is a singleton whose DF is its own inverse, ie. $F(x) = F^{[-1]}(x)$.

Experiment 21 (Bernoulli) The ‘toss 1 times’ experiment is the following family of densities (PMFs) :

$$\mathbb{P} = \{ f(x; p) : p \in [0, 1] \}$$

where, $f(x; p)$ is given in (3.4). The one dimensional parameter space or index set for this experiment is $\Theta = [0, 1] \subset \mathbb{R}$.

Figure 7.1: Geometry of the Θ 's for de Moivre[k] Experiments with $k \in \{1, 2, 3, 4\}$.

Experiment 22 (Point Mass) The ‘deterministically choose a specific real number’ experiment is the following family of DFs :

$$\mathbb{P} = \{ F(x; a) : a \in \mathbb{R} \}$$

where, $F(x; a)$ is given in (6.13). The one dimensional parameter space or index set for this experiment is $\Theta = \mathbb{R}$, the entire real line.

Note that we can use the PDF's or the DF's to specify the family \mathbb{P} of an experiment. When an experiment can be parametrized by finitely many parameters it is said to be a **parametric** experiment. Experiment 21 involving discrete RVs as well as Experiment 22 are **parametric** since they both have only one parameter (the parameter space is one dimensional for Experiments 21 and 22). The Fundamental Experiment 20 involving the continuous RV of Model 3 is also parametric since its parameter space, being a point, is zero-dimensional. The next example is also parametric and involves $(k - 1)$ -dimensional families of discrete RVs.

Experiment 23 (de Moivre[k]) The ‘pick a number from the set $[k] := \{1, 2, \dots, k\}$ somehow’ experiment is the following family of densities (PMFs) :

$$\mathbb{P} = \{ f(x; \theta_1, \theta_2, \dots, \theta_k) : (\theta_1, \theta_2, \dots, \theta_k) \in \Delta_k \}$$

where, $f(x; \theta_1, \theta_2, \dots, \theta_k)$ is any PMF such that

$$f(x; \theta_1, \theta_2, \dots, \theta_k) = \theta_x, \quad x \in \{1, 2, \dots, k\} .$$

The $k - 1$ dimensional parameter space Θ is the k -Simplex Δ_k . This as an ‘exhaustive’ experiment since all possible densities over the finite set $[k] := \{1, 2, \dots, k\}$ are being considered that can be thought of as “the outcome of rolling a convex polyhedral die with k faces and an arbitrary center of mass specified by the θ_i 's.”

An experiment with infinite dimensional parameter space Θ is said to be **nonparametric**. Next we consider two nonparametric experiments.

Experiment 24 (All DFs) The ‘pick a number from the Real line in an arbitrary way’ experiment is the following family of distribution functions (DFs) :

$$\mathbb{P} = \{ F(x; F) : F \text{ is a DF} \} = \Theta$$

where, the DF $F(x; F)$ is indexed or parameterized by itself. Thus, the parameter space

$$\Theta = \mathbb{P} = \{\text{all DFs}\}$$

is the infinite dimensional space of **All DFs**”.

Next we consider a **nonparametric** experiment involving continuous RVs.

Experiment 25 (Sobolev Densities) The ‘pick a number from the Real line in some reasonable way’ experiment is the following family of densities (pdfs) :

$$\mathbb{P} = \left\{ f(x; f) : \int (f''(x))^2 < \infty \right\} = \Theta$$

where, the density $f(x; f)$ is indexed by itself. Thus, the parameter space $\Theta = \mathbb{P}$ is the infinite dimensional **Sobolev space** of “not too wiggly functions”.

7.3 Typical Decision Problems with Experiments

Some of the concrete problems involving experiments include:

- **Simulation:** Often it is necessary to simulate a RV with some specific distribution to gain insight into its features or simulate whole systems such as the air-traffic queues at ‘London Heathrow’ to make better management decisions.
- **Estimation:**
 1. **Parametric Estimation:** Using samples from some unknown DF F parameterized by some unknown θ , we can estimate θ from a statistic T_n called the estimator of θ using one of several methods (maximum likelihood, moment estimation, or parametric bootstrap).
 2. **Nonparametric Estimation of the DF:** Based on n IID observations from an unknown DF F , we can estimate it under the general assumption that $F \in \{\text{all DFs}\}$.
 3. **Confidence Sets:** We can obtain a $1 - \alpha$ confidence set for the point estimates, of the unknown parameter $\theta \in \Theta$ or the unknown DF $F \in \{\text{all DFs}\}$
- **Hypothesis Testing:** Based on observations from some DF F that is hypothesized to belong to a subset Θ_0 of Θ called the space of null hypotheses, we will learn to test (attempt to reject) the falsifiable null hypothesis that $F \in \Theta_0 \subset \Theta$.
- ...

Chapter 8

Limits of Random Variables

8.1 Convergence of Random Variables

This important topic is concerned with the limiting behavior of sequences of RVs

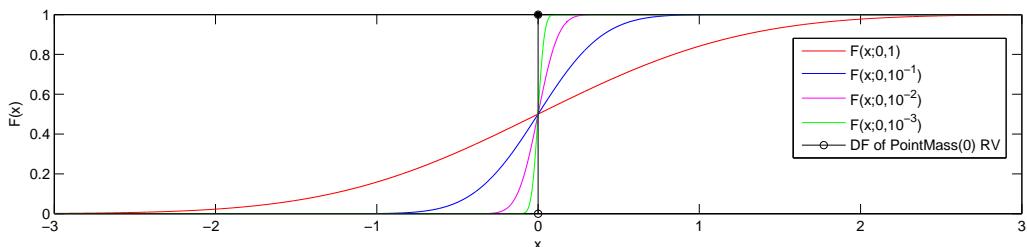
$$\{X_i\}_{i=1}^n := X_1, X_2, X_3, \dots, X_{n-1}, X_n \quad \text{as } n \rightarrow \infty.$$

From a statistical viewpoint $n \rightarrow \infty$ is associated with the amount of data or information $\rightarrow \infty$. Refresh yourself with notions of convergence, limits and continuity in the real line (**S 1.6**) before proceeding further.

Classwork 113 (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 8.1 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 8.1. Based on this observation, can we expect $\lim_{n \rightarrow \infty} X_n = X$, where the limiting RV $X \sim \text{Point Mass}(0)$?

The answer is **no**. This is because $\mathbf{P}(X_n = X) = 0$ for any n , since $X \sim \text{Point Mass}(0)$ 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.

Figure 8.1: Distribution functions of several $\text{Normal}(\mu, \sigma^2)$ 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 three basic propositions in Statistics :

1. Weak Law of Large Numbers,
2. Central Limit Theorem,
3. Gilvenko-Cantelli Theorem.

Definition 47 (Convergence in Distribution) Let X_1, X_2, \dots , 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 . Then 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,

$$\lim_{n \rightarrow \infty} F_n(t) = F(t) \quad [\text{in the sense of Definition 4}].$$

The above limit, by (3.2) in our Definition 17 of a DF, can be equivalently expressed as follows:

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbf{P}(\{\omega : X_n(\omega) \leq t\}) &= \mathbf{P}(\{\omega : X(\omega) \leq t\}), \\ \text{i.e. } \mathbf{P}(\{\omega : X_n(\omega) \leq t\}) &\rightarrow \mathbf{P}(\{\omega : X(\omega) \leq t\}), \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Definition 48 (Convergence in Probability) Let X_1, X_2, \dots , 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 . Then we say that X_n converges to X in probability, and write:

$$X_n \xrightarrow{P} X$$

if for every real number $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbf{P}(|X_n - X| > \epsilon) = 0 \quad [\text{in the sense of Definition 4}].$$

Once again, the above limit, by (3.1) in our Definition 16 of a RV, can be equivalently expressed as follows:

$$\lim_{n \rightarrow \infty} \mathbf{P}(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}) = 0, \quad \text{ie, } \mathbf{P}(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

Let us revisit the problem of convergence in Classwork 113 armed with our new notions of convergence.

Example 114 (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 $X \sim \text{Point Mass}(0)$ with DF F . We can formalize our observation in Classwork 113 that X_n is concentrating about 0 as $n \rightarrow \infty$ by the statement:

$$X_n \text{ is converging in distribution to } X, \text{ ie, } X_n \rightsquigarrow X.$$

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, \dots and the limiting RV X . Thus, we need to verify that for any continuity point t of the Point Mass(0) DF F , $\lim_{n \rightarrow \infty} F_n(t) = F(t)$. First note that

$$X_n \sim \text{Normal}(0, 1/n) \implies Z := \sqrt{n}X_n \sim \text{Normal}(0, 1),$$

and thus

$$F_n(t) = \mathbf{P}(X_n < t) = \mathbf{P}(\sqrt{n}X_n < \sqrt{nt}) = \mathbf{P}(Z < \sqrt{nt}).$$

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 $\sqrt{nt} \rightarrow -\infty$, as $n \rightarrow \infty$,

$$\lim_{n \rightarrow \infty} F_n(t) = \lim_{n \rightarrow \infty} \mathbf{P}(Z < \sqrt{nt}) = 0 = F(t) .$$

And, when $t > 0$, $F(t)$, being the constant 1 function over the interval $(0, \infty)$, is again continuous at t . Since $\sqrt{nt} \rightarrow \infty$, as $n \rightarrow \infty$,

$$\lim_{n \rightarrow \infty} F_n(t) = \lim_{n \rightarrow \infty} \mathbf{P}(Z < \sqrt{nt}) = 1 = F(t) .$$

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: $\lim_{n \rightarrow \infty} F_n(t) = F(t)$.

However, note that

$$F_n(0) = \frac{1}{2} \neq F(0) = 1 ,$$

and so convergence fails at 0, i.e. $\lim_{n \rightarrow \infty} F_n(t) \neq F(t)$ 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 .

For the same sequence of RVs in Classwork 113 and Example 114 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 $X_n \xrightarrow{P} X$. We need some elementary inequalities in Probability to help us answer this question. We visit these inequalities next.

Proposition 49 (Markov's Inequality) Let (Ω, \mathcal{F}, P) be a probability triple and let $X = X(\omega)$ be a non-negative RV. Then,

$$\mathbf{P}(X \geq \epsilon) \leq \frac{\mathbf{E}(X)}{\epsilon}, \quad \text{for any } \epsilon > 0 . \quad (8.1)$$

Proof:

$$\begin{aligned} X &= X\mathbf{1}_{\{y:y \geq \epsilon\}}(x) + X\mathbf{1}_{\{y:y < \epsilon\}}(x) \\ &\geq X\mathbf{1}_{\{y:y \geq \epsilon\}}(x) \\ &\geq \epsilon\mathbf{1}_{\{y:y \geq \epsilon\}}(x) \end{aligned} \quad (8.2)$$

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 (3.14), we get the desired result:

$$\mathbf{E}(X) \geq \epsilon\mathbf{E}(\mathbf{1}_{\{y:y \geq \epsilon\}}(x)) = \epsilon\mathbf{P}(X \geq \epsilon) .$$

Let us look at some immediate consequences of Markov's inequality.

Proposition 50 (Chebychev's Inequality) For any RV X and any $\epsilon > 0$,

$$\mathbf{P}(|X| > \epsilon) \leq \frac{\mathbf{E}(|X|)}{\epsilon} \quad (8.3)$$

$$\mathbf{P}(|X| > \epsilon) = \mathbf{P}(X^2 \geq \epsilon^2) \leq \frac{\mathbf{E}(X^2)}{\epsilon^2} \quad (8.4)$$

$$\mathbf{P}(|X - \mathbf{E}(X)| \geq \epsilon) = \mathbf{P}((X - \mathbf{E}(X))^2 \geq \epsilon^2) \leq \frac{\mathbf{E}(X - \mathbf{E}(X))^2}{\epsilon^2} = \frac{\mathbf{V}(X)}{\epsilon^2} \quad (8.5)$$

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 113 and Example 114.

Example 115 (Convergence in probability) Does 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 $X_n \xrightarrow{P} X$?

To find out if $X_n \xrightarrow{P} X$, we need to show that for any $\epsilon > 0$, $\lim_{n \rightarrow \infty} \mathbf{P}(|X_n - X| > \epsilon) = 0$.

Let ϵ be any real number greater than 0, then

$$\begin{aligned}\mathbf{P}(|X_n| > \epsilon) &= \mathbf{P}(|X_n|^2 > \epsilon^2) \\ &= \frac{\mathbf{E}(X_n^2)}{\epsilon^2} \quad [\text{by Markov's Inequality (8.1)}] \\ &= \frac{\frac{1}{n}}{\epsilon^2} \rightarrow 0, \quad \text{as } n \rightarrow \infty \quad [\text{in the sense of Definition 4].}\end{aligned}$$

Hence, we have shown that for any $\epsilon > 0$, $\lim_{n \rightarrow \infty} \mathbf{P}(|X_n - X| > \epsilon) = 0$ and therefore by Definition 48, $X_n \xrightarrow{P} X$ or $X_n \xrightarrow{P} 0$.

Convention: When X has a Point Mass(θ) distribution and $X_n \xrightarrow{P} X$, we simply write $X_n \xrightarrow{P} \theta$.

Now that we have been introduced to two notions of convergence for sequences of RVs we can begin to appreciate the statements of the basic limit theorems of Statistics.

8.2 Some Basic Limit Laws of Statistics

Proposition 51 (Weak Law of Large Numbers (WLLN)) If we are given a sequence of independent and identically distributed RVs, $X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1$ and if $\mathbf{E}(X_1)$ exists, as per (3.9), then the sample mean \bar{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:

$$\text{If } X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1 \text{ and if } \mathbf{E}(X_1) \text{ exists, then } \bar{X}_n \xrightarrow{P} \mathbf{E}(X_1).$$

Proof: For simplicity, we will prove a slightly weaker result by assuming finite variance of X_1 . Suppose $\mathbf{V}(X_1) < \infty$, then:

$$\begin{aligned}\mathbf{P}(|\bar{X}_n - \mathbf{E}(\bar{X}_n)| \geq \epsilon) &= \frac{\mathbf{V}(\bar{X}_n)}{\epsilon^2} \quad [\text{by applying Chebychev's inequality (8.5) to the RV } \bar{X}_n] \\ &= \frac{\frac{1}{n}\mathbf{V}(X_1)}{\epsilon^2} \quad [\text{by the IID assumption of } X_1, X_2, \dots \text{ we can apply (5.3)}]\end{aligned}$$

Therefore, for any given $\epsilon > 0$,

$$\begin{aligned}\mathbf{P}(|\bar{X}_n - \mathbf{E}(\bar{X}_n)| \geq \epsilon) &= \mathbf{P}(|\bar{X}_n - \mathbf{E}(\bar{X}_n)| \geq \epsilon) \quad [\text{by the IID assumption of } X_1, X_2, \dots, \mathbf{E}(\bar{X}_n) = \mathbf{E}(X_1), \text{ as per (5.2)}] \\ &= \frac{\frac{1}{n}\mathbf{V}(X_1)}{\epsilon^2} \rightarrow 0, \quad \text{as } n \rightarrow \infty,\end{aligned}$$

or equivalently, $\lim_{n \rightarrow \infty} \mathbf{P}(|\bar{X}_n - \mathbf{E}(\bar{X}_n)| \geq \epsilon) = 0$. And the last statement is the definition of the claim made by the weak law of large numbers (WLLN), namely that $\bar{X}_n \xrightarrow{P} \mathbf{E}(X_1)$.

Heuristic Interpretation of WLLN: The distribution of the sample mean RV \bar{X}_n obtained from an independent and identically distributed sequence of RVs X_1, X_2, \dots [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.

Example 116 (Bernoulli WLLN and Galton's Quincunx) We can appreciate the WLLN for $\bar{X}_n = n^{-1}S_n = \sum_{i=1}^n X_i$, where $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} \text{Bernoulli}(p)$ using the paths of balls dropped into a device built by Galton called the Quincunx.

Proposition 52 (Central Limit Theorem (CLT)) Let $X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1$ and suppose $\mathbf{E}(X_1)$ and $\mathbf{V}(X_1)$ exists, then

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \rightsquigarrow X \sim \text{Normal} \left(\mathbf{E}(X_1), \frac{\mathbf{V}(X_1)}{n} \right) , \quad (8.6)$$

$$\bar{X}_n - \mathbf{E}(X_1) \rightsquigarrow X - \mathbf{E}(X_1) \sim \text{Normal} \left(0, \frac{\mathbf{V}(X_1)}{n} \right) , \quad (8.7)$$

$$\sqrt{n} (\bar{X}_n - \mathbf{E}(X_1)) \rightsquigarrow \sqrt{n} (X - \mathbf{E}(X_1)) \sim \text{Normal}(0, \mathbf{V}(X_1)) , \quad (8.8)$$

$$Z_n := \frac{\bar{X}_n - \mathbf{E}(\bar{X}_n)}{\sqrt{\mathbf{V}(\bar{X}_n)}} = \frac{\sqrt{n} (\bar{X}_n - \mathbf{E}(X_1))}{\sqrt{\mathbf{V}(X_1)}} \rightsquigarrow Z \sim \text{Normal}(0, 1) , \quad (8.9)$$

$$\lim_{n \rightarrow \infty} P \left(\frac{\bar{X}_n - \mathbf{E}(\bar{X}_n)}{\sqrt{\mathbf{V}(\bar{X}_n)}} \leq z \right) = \lim_{n \rightarrow \infty} \mathbf{P}(Z_n \leq z) = \Phi(z) := \int_{-\infty}^z \left(\frac{1}{\sqrt{2\pi}} \exp \left(\frac{-x^2}{2} \right) \right) dx . \quad (8.10)$$

Thus, for sufficiently large n (say $n > 30$) we can make the following approximation:

$$P \left(\frac{\bar{X}_n - \mathbf{E}(\bar{X}_n)}{\sqrt{\mathbf{V}(\bar{X}_n)}} \leq z \right) \approx \mathbf{P}(Z \leq z) = \Phi(z) := \int_{-\infty}^z \left(\frac{1}{\sqrt{2\pi}} \exp \left(\frac{-x^2}{2} \right) \right) dx . \quad (8.11)$$

Proof: See any intermediate to advanced undergraduate text in Probability. Start from the index looking for “Central Limit Theorem” to find the page number for the proof

Heuristic Interpretation of CLT: Probability statements about the sample mean RV \bar{X}_n can be approximated using a Normal distribution.

Here is a simulation showing CLT in action.

```
>> % a demonstration of Central Limit Theorem --
>> % the sample mean of a sequence of n IID Exponential(lambda) RVs
>> % itself a Gaussian(1/lambda, lambda/n) RV
>> lambda=0.1; Reps=10000; n=10; hist(sum(-1/lambda * log(rand(n,Reps)))/n)
>> lambda=0.1; Reps=10000; n=100; hist(sum(-1/lambda * log(rand(n,Reps)))/n,20)
>> lambda=0.1; Reps=10000; n=1000; hist(sum(-1/lambda * log(rand(n,Reps)))/n,20)
```

Let us look at an example that makes use of the CLT next.

Example 117 (Errors in computer code (Wasserman03, p. 78)) Suppose the collection of RVs X_1, X_2, \dots, X_n model the number of errors in n computer programs named $1, 2, \dots, n$, respectively. Suppose that the RV X_i modeling the number of errors in the i -th program is the $\text{Poisson}(\lambda = 5)$ for any $i = 1, 2, \dots, n$. Further suppose that they are independently distributed. Succinctly, we suppose that

$$X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} \text{Poisson}(\lambda = 5) .$$

Suppose we have $n = 125$ programs and want to make a probability statement about \bar{X}_n which is the average error per program out of these 125 programs. Since $\mathbf{E}(X_i) = \lambda = 5$ and $\mathbf{V}(X_i) = \lambda = 5$, we may want to know how often our sample mean \bar{X}_{125} differs from the expectation of 5 errors per

program. Using the CLT we can approximate $\mathbf{P}(\bar{X}_n < 5.5)$, for instance, as follows:

$$\begin{aligned}
 \mathbf{P}(\bar{X}_n < 5.5) &= P\left(\frac{\sqrt{n}(\bar{X}_n - \mathbf{E}(X_1))}{\sqrt{\mathbf{V}(X_1)}} < \frac{\sqrt{n}(5.5 - \mathbf{E}(X_1))}{\sqrt{\mathbf{V}(X_1)}}\right) \\
 &\approx P\left(Z < \frac{\sqrt{n}(5.5 - \lambda)}{\sqrt{\lambda}}\right) \quad [\text{by (8.11), and } \mathbf{E}(X_1) = \mathbf{V}(X_1) = \lambda] \\
 &= P\left(Z < \frac{\sqrt{125}(5.5 - 5)}{\sqrt{5}}\right) \quad [\text{Since, } \lambda = 5 \text{ and } n = 125 \text{ in this Example}] \\
 &= \mathbf{P}(Z \leq 2.5) = \Phi(2.5) = \int_{-\infty}^{2.5} \left(\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) \right) dx \approx 0.993790334674224 .
 \end{aligned}$$

The last number above needed the following:

Labwork 118 (Numerical approximation of $\Phi(2.5)$) The numerical approximation of $\Phi(2.5)$ was obtained via the following call to our erf-based `NormalCdf` function from ??.

```

>> format long
>> disp(NormalCdf(2.5,0,1))
0.993790334674224

```

The CLT says that if $X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1$, then $Z_n := \sqrt{n}(\bar{X}_n - \mathbf{E}(X_1))/\sqrt{\mathbf{V}(X_1)}$ is approximately distributed as $\text{Normal}(0, 1)$. In Example 117, we knew $\sqrt{\mathbf{V}(X_1)}$. However, in general, we may not know $\sqrt{\mathbf{V}(X_1)}$. The next proposition says that we may estimate $\sqrt{\mathbf{V}(X_1)}$ using the sample standard deviation S_n of X_1, X_2, \dots, X_n , according to (5.5), and still make probability statements about the sample mean \bar{X}_n using a Normal distribution.

Proposition 53 (CLT based on Sample Variance) Let $X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1$ and suppose $\mathbf{E}(X_1)$ and $\mathbf{V}(X_1)$ exists, then

$$\frac{\sqrt{n}(\bar{X}_n - \mathbf{E}(X_1))}{S_n} \rightsquigarrow \text{Normal}(0, 1) . \quad (8.12)$$

We will use (8.12) for statistical estimation in the sequel.

Chapter 9

Finite Markov Chains

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. Finite Markov chains are among the simplest stochastic processes with a ‘first-order’ dependence called Markov dependence.

9.1 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 54 (Finite Markov Chain) A stochastic sequence,

$$(X_n)_{n \in \mathbb{Z}_+} := (X_0, X_1, \dots),$$

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 $\mathbf{P}(H_{t-1} \cap \{X_t = x\}) > 0$, the following **Markov property** is satisfied:

$$\mathbf{P}(X_{t+1} = y | H_{t-1} \cap \{X_t = x\}) = \mathbf{P}(X_{t+1} = y | X_t = x) =: P(x, y). \quad (9.1)$$

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 $(x_0, x_1, \dots, x_{t-1})$ preceded the current state x . Thus, the $|\mathbb{X}| \times |\mathbb{X}|$ 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.,

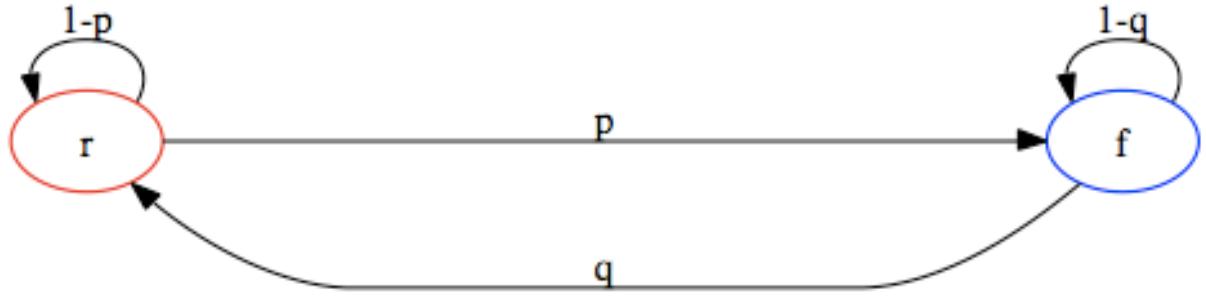
$$P(x, y) \geq 0 \quad \text{for all } (x, y) \in \mathbb{X}^2 \quad \text{and} \quad \sum_{y \in \mathbb{X}} P(x, y) = 1 \quad \text{for all } x \in \mathbb{X}. \quad (9.2)$$

Thus, for a Markov chain $(X_n)_{n \in \mathbb{Z}_+}$, the distribution of X_{t+1} given X_0, \dots, X_t depends on X_t alone. Because of this dependence on the previous state, the stochastic sequence, (X_0, X_1, \dots) , are *not* independent. We introduce the most important concepts using a simple example.

Example 119 (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, \dots) 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 9.1: Transition Diagram of Flippant Freddy's Jumps.



Then Freddy's sequence of jumps (X_0, X_1, \dots) is a Markov chain on \mathbb{X} with transition matrix:

$$P = \begin{matrix} r & f \\ \begin{matrix} r \\ f \end{matrix} & \begin{pmatrix} P(r,r) & P(r,f) \\ P(f,r) & P(f,f) \end{pmatrix} \end{matrix} = \begin{matrix} r & f \\ \begin{matrix} r \\ f \end{matrix} & \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \end{matrix}. \quad (9.3)$$

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 flipopia with probability p and stay with probability $1 - p$, i.e.,

$$\mathbf{P}(X_1 = f | X_0 = r) = p, \quad \mathbf{P}(X_1 = r | X_0 = r) = 1 - p. \quad (9.4)$$

What happens when he is restless for the second time? By considering the two possibilities for X_1 ,

Definition of conditional probability and the Markov property, we see that,

$$\begin{aligned}
 \mathbf{P}(X_2 = f | X_0 = r) &= \mathbf{P}(X_2 = f, X_1 = f | X_0 = r) + \mathbf{P}(X_2 = f, X_1 = r | X_0 = r) \\
 &= \frac{\mathbf{P}(X_2 = f, X_1 = f, X_0 = r)}{\mathbf{P}(X_0 = r)} + \frac{\mathbf{P}(X_2 = f, X_1 = r, X_0 = r)}{\mathbf{P}(X_0 = r)} \\
 &= \mathbf{P}(X_2 = f | X_1 = f, X_0 = r) \frac{\mathbf{P}(X_1 = f, X_0 = r)}{\mathbf{P}(X_0 = r)} \\
 &\quad + \mathbf{P}(X_2 = f | X_1 = r, X_0 = r) \frac{\mathbf{P}(X_1 = r, X_0 = r)}{\mathbf{P}(X_0 = r)} \\
 &= \mathbf{P}(X_2 = f | X_1 = f, X_0 = r) \mathbf{P}(X_1 = f | X_0 = r) \\
 &\quad + \mathbf{P}(X_2 = f | X_1 = r, X_0 = r) \mathbf{P}(X_1 = r | X_0 = r) \\
 &= \mathbf{P}(X_2 = f | X_1 = f) \mathbf{P}(X_1 = f | X_0 = r) \\
 &\quad + \mathbf{P}(X_2 = f | X_1 = r) \mathbf{P}(X_1 = r | X_0 = r) \\
 &= P(f, f)P(r, f) + P(r, f)P(r, r) \\
 &= (1 - q)p + p(1 - p)
 \end{aligned} \tag{9.5}$$

Similarly,

$$\mathbf{P}(X_2 = r | X_0 = r) = P(f, r)P(r, f) + P(r, r)P(r, r) = qp + (1 - p)(1 - p) \tag{9.6}$$

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 (9.4) from $\mu_1 = \mu_0 P$ and the 2-step state probabilities in (9.5) and (9.6) by $\mu_2 = \mu_1 P = \mu_0 P P = \mu_0 P^2$. 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:

$$\mu_t = \mu_{t-1} P \quad \text{for all } t \geq 1 .$$

And for any initial distribution μ_0 ,

$$\mu_t = \mu_0 P^t \quad \text{for all } t \geq 0 .$$

This can be easily implemented in MATLAB as follows:

```

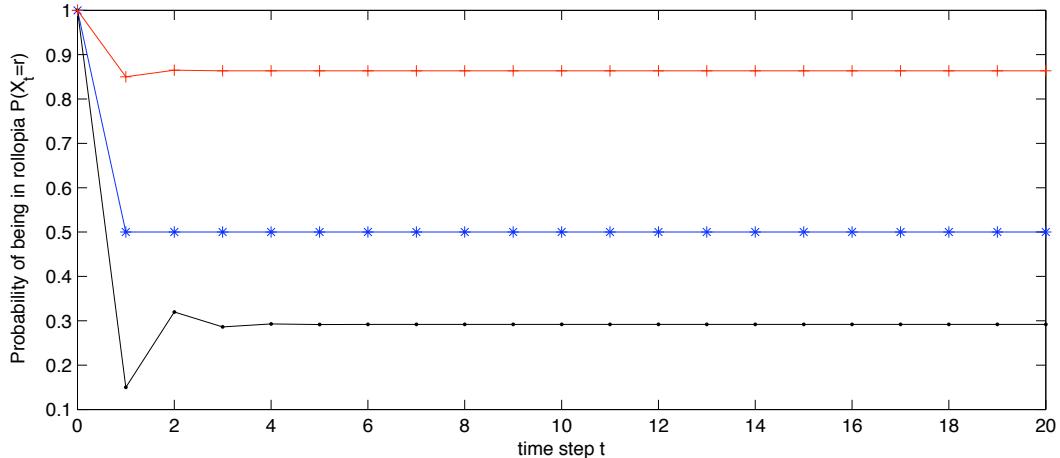
>> p=0.85; q=0.35; P = [1-p p; q 1-q] % assume an unfair coin and an unfair die
P =
    0.1500    0.8500
    0.3500    0.6500
>> mu0 = [1, 0] % initial state vector since Freddy started in rollopia
mu0 =
    1         0
>> mu0*P^0    % initial state distribution at t=0 is just mu0
ans =
    1         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

```

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]; % initial 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 9.2: 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 9.2 that as $t \rightarrow \infty$, 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:

$$\pi(r) = \frac{q}{p+q}, \quad \pi(f) = \frac{p}{p+q} .$$

In Figure 9.2 we see that $\mathbf{P}(X_t = r)$ approaches $\pi(r) = \frac{q}{p+q}$ for the three cases of p and q :

$$\begin{aligned} \text{(i)} \quad & p = 0.50, q = 0.50, & \mathbf{P}(X_t = r) \rightarrow \pi(r) = \frac{q}{p+q} = \frac{0.50}{0.50+0.50} = 0.5000, \\ \text{(ii)} \quad & p = 0.85, q = 0.35, & \mathbf{P}(X_t = r) \rightarrow \pi(r) = \frac{q}{p+q} = \frac{0.35}{0.85+0.35} = 0.2917, \\ \text{(iii)} \quad & p = 0.15, q = 0.95, & \mathbf{P}(X_t = r) \rightarrow \pi(r) = \frac{q}{p+q} = \frac{0.95}{0.15+0.95} = 0.8636. \end{aligned}$$

Now let us generalise the lessons learned from Example 119.

Proposition 55 For a finite Markov chain $(X_t)_{t \in \mathbb{Z}_+}$ with state space $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$, initial distribution

$$\mu_0 := (\mu_0(s_1), \mu_0(s_2), \dots, \mu_0(s_k)),$$

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:

$$\mu_t := (\mu_t(s_1), \mu_t(s_2), \dots, \mu_t(s_k)),$$

where $\mu_t(s_i) = \mathbf{P}(X_t = s_i)$, satisfies:

$$\mu_t = \mu_0 P^t. \quad (9.7)$$

Proof: We will prove this by induction on $\mathbb{Z}_+ := \{0, 1, 2, \dots\}$. 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 $j \in \{1, 2, \dots, k\}$, 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 (9.7) holds for $t = m$ and prove that (9.7) also holds for $t = m + 1$. For each $j \in \{1, 2, \dots, k\}$, 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.

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 $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$ with the same transition matrix P but different initial distributions, we introduce \mathbf{P}_μ and \mathbf{E}_μ 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:

$$\mathbf{1}_{\{x\}}(y) := \begin{cases} 1 & \text{if } y = x \\ 0 & \text{if } y \neq x \end{cases}$$

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 $\mathbf{P}_{\mathbf{1}_{\{x\}}}$ or \mathbf{P}_{e_x} and \mathbf{E}_x for $\mathbf{E}_{\mathbf{1}_{\{x\}}}$ or \mathbf{E}_{e_x} . Thus, Proposition 55 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.

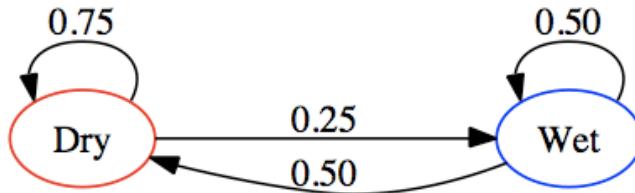
Let the function $f(x) : \mathbb{X} \rightarrow \mathbb{R}$ be represented by the column vector $f := (f(s_1), f(s_2), \dots, f(s_k)) \in \mathbb{R}^{k \times 1}$. Then the x -th entry of $P^t f$ is:

$$P^t f(x) = \sum_y P^t(x, y) f(y) = \sum_y f(y) \mathbf{P}_x(X_t = y) = \mathbf{E}_x(f(X_t)) .$$

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 120 (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 9.3 give the transition matrix P for our dry-wet Markov chain. Using (9.7) we can find

Figure 9.3: 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
  
```

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 121 (Freddy discovers a gold coin) Flippant Freddy of Example 119 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, rollophia 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 rollophia he will jump to flipopia but if he rolls an even number then he will stay in rollophia 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 rollophia.

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 122 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

$$P = \begin{matrix} & \begin{matrix} a & b & c \end{matrix} \\ \begin{matrix} a \\ b \\ c \end{matrix} & \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \end{matrix}.$$

For each t , define $Y_t = \mathbf{1}_{\{b,c\}}(X_t)$. Show that $(Y_t)_{t \in \mathbb{Z}_+}$ is not a Markov chain.

Exercise 123 Let $(X_t)_{t \in \mathbb{Z}_+}$ be a (homogeneous) Markov chain on $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$ 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 54, 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.

Definition 56 (Inhomogeneous finite Markov chain) Let P_1, P_2, \dots be a sequence of $k \times k$ stochastic matrices satisfying the conditions in Equation 9.2. Then, the stochastic sequence $(X_t)_{t \in \mathbb{Z}_+} := (X_0, X_1, \dots)$ with finite state space $\mathbb{X} := \{s_1, s_2, \dots, s_k\}$ is called an inhomogeneous Markov chain with transition matrices P_1, P_2, \dots , 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 $\mathbf{P}(H_{t-1} \cap \{X_t = x\}) > 0$, the following **Markov property** is satisfied:

$$\mathbf{P}(X_{t+1} = y | H_{t-1} \cap \{X_t = x\}) = \mathbf{P}(X_{t+1} = y | X_t = x) =: P_{t+1}(x, y) . \quad (9.8)$$

Proposition 57 For a finite inhomogeneous Markov chain $(X_t)_{t \in \mathbb{Z}_+}$ with state space $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$, initial distribution

$$\mu_0 := (\mu_0(s_1), \mu_0(s_2), \dots, \mu_0(s_k)) ,$$

where $\mu_0(s_i) = \mathbf{P}(X_0 = s_i)$, and transition matrices

$$(P_1, P_2, \dots) , \quad P_t := (P_t(s_i, s_j))_{(s_i, s_j) \in \mathbb{X} \times \mathbb{X}} , \quad t \in \{1, 2, \dots\}$$

we have for any $t \in \mathbb{Z}_+$ that the distribution at time t given by:

$$\mu_t := (\mu_t(s_1), \mu_t(s_2), \dots, \mu_t(s_k)) ,$$

where $\mu_t(s_i) = \mathbf{P}(X_t = s_i)$, satisfies:

$$\mu_t = \mu_0 P_1 P_2 \cdots P_t . \quad (9.9)$$

Proof: Left as Exercise 124.

Exercise 124 Prove Proposition 57 using induction as done for Proposition 55.

Example 125 (a more sophisticated dry-wet chain) Let us make a more sophisticated version of the dry-wet chain of Example 120 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:

$$P_{\text{hot}} = \begin{matrix} d & w \\ \begin{pmatrix} 0.95 & 0.05 \\ 0.75 & 0.25 \end{pmatrix} \end{matrix}, \quad P_{\text{cold}} = \begin{matrix} d & w \\ \begin{pmatrix} 0.65 & 0.35 \\ 0.45 & 0.55 \end{pmatrix} \end{matrix} .$$

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 (9.9) 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
```

```

>> 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 126 For the Markov chain in Example 125 compute the probability that the day after tomorrow is wet if today is dry and hot but tomorrow is supposed to be cold.

9.2 Random Mapping Representation and Simulation

In order to simulate (x_0, x_1, \dots, x_n) , a sequential realisation or sequence of states visited by a Markov chain, say the sequence of lily pads that Flippant Freddy visits on his jumps, we need a random mapping representation of a Markov chain and its computer implementation.

Definition 58 (Random mapping representation (RMR)) A **random mapping representation (RMR)** of a transition matrix $P := (P(x, y))_{(x,y) \in \mathbb{X}^2}$ is a function

$$\rho(x, w) : \mathbb{X} \times \mathbb{W} \rightarrow \mathbb{X} , \quad (9.10)$$

along with the auxiliary \mathbb{W} -valued random variable W , satisfying

$$\mathbf{P}(\{\rho(x, W) = y\}) = P(x, y), \quad \text{for each } (x, y) \in \mathbb{X}^2 . \quad (9.11)$$

Proposition 59 (Markov chain from RMR) If $W_1, W_2, \dots \stackrel{IID}{\sim} W$, the auxiliary RV in a RMR of a transition matrix $P := (P(x, y))_{(x,y) \in \mathbb{X}^2}$, and $X_0 \sim \mu_0$, then $(X_t)_{t \in \mathbb{Z}_+}$ defined by

$$X_t = \rho(X_{t-1}, W_t) , \quad \text{for all } t \geq 1$$

is a Markov chain with transition matrix P and initial distribution μ_0 on state space \mathbb{X} .

Proof: Left as Exercise 127.

Exercise 127 Do the proof of Proposition 59 by using the necessary Definitions.

Example 128 (An RMR for Flippant Freddy) Reconsider the Markov chain of Flippant Freddy with fair dice and fair coin on state space $\mathbb{X} = \{r, f\}$ with transition matrix

$$P = \begin{matrix} & r & f \\ r & \left(\begin{matrix} 1/2 & 1/2 \end{matrix} \right) \\ f & \left(\begin{matrix} 1/2 & 1/2 \end{matrix} \right) \end{matrix} .$$

Let the auxiliary RV W have sample space $\mathbb{W} = \{0, 1\}$. Then an RMR $\rho : \mathbb{X} \times \mathbb{W} \rightarrow \mathbb{X}$ for this P is given by

$$\rho(x, w) : \{r, f\} \times \{0, 1\} \rightarrow \{r, f\}, \quad \rho(r, 0) = r, \quad \rho(r, 1) = f, \quad \rho(f, 0) = f, \quad \rho(f, 1) = r,$$

with $\mathbf{P}(W = 0) = \mathbf{P}(W = 1) = 1/2$. Now let us check that our ρ and W satisfy Equation 9.11:

$$\begin{aligned} r & \quad f \\ r \begin{pmatrix} \mathbf{P}(\{\rho(r, W) = r\}) & \mathbf{P}(\{\rho(r, W) = f\}) \\ \mathbf{P}(\{\rho(f, W) = r\}) & \mathbf{P}(\{\rho(f, W) = f\}) \end{pmatrix} &= r \begin{pmatrix} \mathbf{P}(W = 0) & \mathbf{P}(W = 1) \\ \mathbf{P}(W = 1) & \mathbf{P}(W = 0) \end{pmatrix} \\ &= r \begin{pmatrix} r & f \\ 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} = P . \end{aligned}$$

Thus, by Proposition 59 we can obtain Freddy's Markov chain $(X_t)_{t \in \mathbb{Z}_+}$ by initialising $X_0 \sim \mu_0 = (1, 0)$, i.e., setting $X_0 = r$ since Freddy starts at r , and defining

$$X_t = \rho(X_{t-1}, W_t), \quad \text{for all } t \geq 1, \text{ where, } W_1, W_2, \dots \stackrel{IID}{\sim} \text{Bernoulli}(1/2) \text{ RV} .$$

In other words, we can simulate a sequence of states or lily pads visited by Freddy by merely doing independent Bernoulli(1/2) trials and use the mapping ρ . A MATLAB implementation of this RMR ρ as a MATLAB function is:

```
RMR10fFairFreddy.m
function y = RMR10fFairFreddy(x,w)
% Random Mapping Representation Number 1 of P=[1/2 1/2; 1/2 12/]
% input: character x as 'r' or 'f' and w as 0 or 1
% output: character y as 'r' or 'f'
if (x =='r')
    if (w==0)
        y = 'r';
    elseif (w==1)
        y = 'f';
    else
        y = Nan;
        print "when x = 'r' w is neither 0 nor 1!";
    end
elseif (x =='f')
    if (w==0)
        y = 'f';
    elseif (w==1)
        y = 'r';
    else
        y = Nan;
        print "when x='f' w is neither 0 nor 1!";
    end
else
    y = Nan;
    print "x is neither 'r' nor 'f'";
end
```

We can simulate one realisation of the first two states (x_0, x_1) visited by (X_0, X_1) as follows:

```
>> % set PRNG to be twister with seed 19731511
>> RandStream.setDefaultStream(RandStream('mt19937ar','seed',19731511));
>> x0 = 'r' % set x_0 = 'r'
x0 = r
>> w1 = floor( rand + 0.5 ) % a Bernoulli(0.5) trial
w1 =
0
>> x1 = RMR10fFairFreddy(x0,w1) % x_1 = rho(x_0,w1) is the state at time t=1
x1 = r
```

We can simulate one realisation of the first 10 states (x_0, x_1, \dots, x_9) visited by (X_0, X_1, \dots, X_9) using a for loop as follows:

```
>> % set PRNG to be twister with seed 19731511
>> RandStream.setDefaultStream(RandStream('mt19937ar','seed',19731511));
>> x0 = 'r' % set x_0 = 'r'
x0 =
>> xt = x0; % current state x_t is x_0
>> Visited = x0; % initialise the variable Visited to hold the visited states
>> for t = 1:9 % start a for loop for t = 1,2,...,9
xt = RMR10fFairFreddy(xt, floor(rand+0.5) ); % update the current state at t
Visited = strcat(Visited,',',xt); % store the visited state in string Visited
end
>> Visited % disclose the string of visited state separated by commas
Visited = r,r,f,f,r,r,f,f,f,r
```

If we change the seed to some other number and repeat the code above, we will get another realisation of visits (x_0, x_1, \dots, x_9) of (X_0, X_1, \dots, X_9) . However, there are many distinct RMRs of the same transition matrix P . For example, we can define a new RMR ρ' from our first RMR ρ for P by $\rho'(x, w) = \rho(x, 1 - w)$. The reader should check that ρ' also satisfies Equation 9.11 with $W \sim \text{Bernoulli}(1/2)$. But note that even for the same seed and the same PRNG the sequence of states (x_0, x_1, \dots, x_9) visited by (X_0, X_1, \dots, X_9) under the new RMR ρ' is different from that of the original RMR ρ :

```
>> % set PRNG to be twister with seed 19731511
>> RandStream.setDefaultStream(RandStream('mt19937ar','seed',19731511));
>> x0 = 'r' % set x_0 = 'r'
x0 =
>> xt = x0; % current state x_t is x_0
>> Visited = x0; % initialise the variable Visited to hold the visited states
>> for t = 1:9 % start a for loop for t = 1,2,...,9
xt = RMR10fFairFreddy(xt, 1-floor(rand+0.5) ); % update the current state at t with new RMR rho'
Visited = strcat(Visited,',',xt); % store the visited state in string Visited
end
>> Visited % disclose the string of visited state separated by commas under new RMR rho'
Visited = r,f,f,r,r,f,f,r,f,f
```

Proposition 60 (Existence and non-uniqueness of RMR) Every transition matrix P on a finite state space \mathbb{X} has a random mapping representation (RMR) that is not necessarily unique.

Proof: Let $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$ be sequentially accessible by $\psi(i) = s_i : \{1, 2, \dots, k\} \rightarrow \mathbb{X}$. We will prove the proposition constructively via the inversion sampler for \mathbb{X} -valued family of ψ -transformed de Moivre RVs. Let the auxiliary RV W be Uniform(0, 1) with $\mathbb{W} = [0, 1]$ and let $\rho(x, w) : \mathbb{X} \times \mathbb{W} \rightarrow \mathbb{X}$ be given by $F^{[-1]}(u; \theta_1, \theta_2, \dots, \theta_k)$ of Equation 6.16, the inverse DF of the de Moivre($\theta_1, \theta_2, \dots, \theta_k$) RV, as follows:

$$\rho(x, w) = \psi \left(F^{[-1]}(w; P(x, s_1), P(x, s_2), \dots, P(x, s_k)) \right), \quad \text{for each } x \in \mathbb{X} .$$

Then, by construction, this ρ is indeed an RMR of P since

$$\mathbf{P}(\{\rho(x, W) = y\}) = P(x, y) \quad \text{for each } (x, y) \in \mathbb{X}^2 .$$

Non-uniqueness is established by constructing another RMR for P as $\rho'(x, w) = \rho(x, 1 - w)$.

Labwork 129 (Markov chain from $\{\text{de Moivre}(P(x,.))\}_{x \in \mathbb{X}}$ RVs) Let us implement a function that will take a transition matrix P as input and produce a sequence of n states $(x_0, x_1, \dots, x_{n-1})$ visited by the corresponding Markov chain (X_0, X_1, \dots, X_n) using the function in the following M-file.

```
MCSimBydeMoivre.m
function VisitedStateIdxs = MCSimBydeMoivre(idx0, P, n)
% input: idx0 = index of initial state x_0, psi(idx0) = x_0
%         P = transition probability matrix (has to be stochastic matrix)
%         n = number of time steps to simulate, n >= 0
% output: VisitedStateIdxs = idx0, idx1, ..., idxn
VisitedStateIdxs = zeros(1,n);
VisitedStateIdxs(1, 0+1) = idx0; % initial state index is the input idx0
for i=1:n-1
    CurrentState = VisitedStateIdxs(1, i); % current state
    Thetas = P(CurrentState,:);
    VisitedStateIdxs(1, i+1) = SimdeMoivreOnce(rand,Thetas); % next state
end
end
```

Simulation 130 (Another simulation of Freddy's jumps) Let us simulate a sequence of 10 jumps of Flippant Freddy with fair dice and coin by using the function `MCSimBydeMoivre` defined in Labwork 129 as follows:

```
>> % set PRNG to be twister with seed 19731511
>> RandStream.setDefaultStream(RandStream('mt19937ar','seed',19731511));
>> MCSimBydeMoivre(1,[0.5 0.5; 0.5 0.5], 10)
ans =
     1     1     2     1     2     1     2     1     1     2
```

Here we need to further transform the output by $\psi : \{1, 2\} \rightarrow \{r, f\}$ with $\psi(1) = r$ and $\psi(2) = f$.

Labwork 131 (Markov chain from $\{\text{de Moivre}(P(x,.))\}_{x \in \mathbb{X}}$ RVs by Recursion) Let us implement a recursive function that will take a transition matrix P as input and produce a sequence of n states $(x_0, x_1, \dots, x_{n-1})$ visited by the corresponding Markov chain (X_0, X_1, \dots, X_n) using the function in the following M-file.

```
MCSimBydeMoivreRecurse.m
function VisitedStateIdxs = MCSimBydeMoivreRecurse(VisitedStateIdxs, P, n)
% input: VisitedStateIdxs = array of indexes of states visited so far
%         P = transition probability matrix (has to be stochastic matrix)
%         n = number of time steps to simulate, n >= 0
% output: VisitedStateIdxs = idx0, idx1, ..., idxn
i = length(VisitedStateIdxs);
if i < n
    CurrentState = VisitedStateIdxs(1, i); % current state
    Thetas = P(CurrentState,:);
    % recursion
    VisitedStateIdxs= MCSimBydeMoivreRecurse([VisitedStateIdxs SimdeMoivreOnce(rand,Thetas)],P,n); % next state
end
end
```

Now, let us compare this recursive function to the function `MCSimBydeMoivre` defined in Labwork 129 as follows:

```
CompareMCSimBydeMoivreMethods.m
format compact
P=[1/3 2/3;1/4 4/5]
```

```

initial = 2
visited = [initial];
n = 12;

s = RandStream('mt19937ar','Seed', 5489);
RandStream.setDefaultStream(s) % reset the PRNG to default state Mersenne Twister with seed=5489

VisitByMethod1 = MCSimBydeMoivre(initial, P, n)

s = RandStream('mt19937ar','Seed', 5489);
RandStream.setDefaultStream(s) % reset the PRNG to default state Mersenne Twister with seed=5489

VisitByMethod2 = MCSimBydeMoivreRecurse(visited, P, n)

```

```

>> CompareMCSimBydeMoivreMethods
P =
    0.3333    0.6667
    0.2500    0.8000
initial =
    2
VisitByMethod1 =
    2    2    2    1    2    2    1    1    2    2    2    1
VisitByMethod2 =
    2    2    2    1    2    2    1    1    2    2    2    1

```

Therefore, both methods produce the same output. The recursive version of the function is more versatile and useful in the sequel.

Simulation 132 Using the function `MCSimBydeMoivre` of Labwork 129 simulate twenty states visited by the Markov chain in Exercise 121.

Simulation 133 (Drunkard's walk around the block) Consider the Markov chain $(X_t)_{t \in \mathbb{Z}_+}$ on $\mathbb{X} = \{0, 1, 2, 3\}$ with initial distribution $\mathbf{1}_{\{3\}}(x)$ and transition matrix

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

Draw the transition diagram for this Markov chain. Do you see why this chain can be called the “drunkard's walk around the block”? Using the function `MCSimBydeMoivre` of Labwork 129 simulate a sequence of ten states visited by the drunkard (don't forget to subtract 1 from the output of `MCSimBydeMoivre` since $\psi(i) = i - 1$ here).

There are many distinct and interesting RMRs of any given transition matrix P beyond that constructed in the proof above. Good RMRs will typically simplify the simulation of a Markov chain. Let us consider examples of Markov chains that can be simulated by simpler methods.

Example 134 (Jukes & Cantor Model of DNA mutation) The “blueprint” of organisms on earth are typically given by a long sequence of deoxyribonucleic acid or DNA. A DNA sequence of length n can be thought of as a string made up of n alphabets from the set of four nucleotides $\{a, c, g, t\}$. For example a DNA sequence of length 3 is *agg* and another is *act*. When an organism

goes through time to “stay alive” it has to copy its DNA. This copying process is not perfect and mistakes or mutations are made. We can look at a particular position of a DNA sequence and keep track of its mutations using a simple Markov chain due to Jukes and Cantor [Jukes TH and Cantor CR (1969) Evolution of protein molecules. In Munro HN, editor, Mammalian Protein Metabolism, pp. 21-132, Academic Press, New York.] with the following transition probability matrix:

$$P = \begin{pmatrix} & a & c & g & t \\ a & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ c & \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ g & \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ t & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix}.$$

Suppose you initially observe the particular position of a DNA sequence at state c and want to simulate a sequence of states visited due to mutation under this Markov chain model. We can achieve this by improvising the inversion sampler for the equi-probable de Moivre($1/3, 1/3, 1/3$) RV (Algorithm 5) in the following RMR:

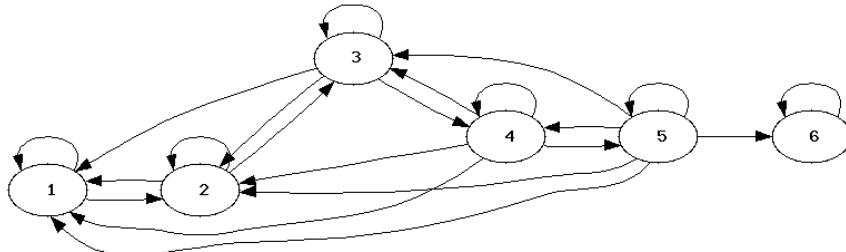
$$\rho(x, U) : \{a, c, g, t\} \times [0, 1] \rightarrow \{a, c, g, t\}, \quad \rho(x, U) = \psi_x(\lceil 3U \rceil), \quad U \sim \text{Uniform}(0, 1),$$

with any fixed bijection $\psi_x(i) : \{1, 2, 3\} \rightarrow \{a, c, g, t\} \setminus \{x\}$ for each $x \in \{a, c, g, t\}$. Then we can produce a sequence of visited states as follows:

$$X_0 \leftarrow c, \quad X_i \leftarrow \rho(X_{i-1}, U_i), \quad i = 1, 2, \dots.$$

Example 135 (Six Lounges) Suppose there are six lounges with doors that allow you to go only in one direction. These lounges are labelled by 1, 2, 3, 4, 5 and 6 and form our state space \mathbb{X} with one-way-doors as shown in Figure 9.4. Every hour an alarm rings and it can be heard in all six

Figure 9.4: Transition diagram over six lounges (without edge probabilities).



lounges. In each lounge $i \in \{1, 2, 3, 4, 5\}$ there is a fair i -sided polyhedral cylinder whose i faces are marked with lounge numbers $1, 2, \dots, i$ but in lounge 6 there is a hexagonal cylinder with all six faces marked by 6. Suppose you start from lounge number 1. When the hourly alarm rings you toss the polyhedral cylinder in the current lounge over the floor. When the cylinder comes to rest, you note the number on the face that touches the floor and go to the lounge labelled by this number. This scheme of lounge hopping can be formalised as a Markov chain starting at lounge

number 1 and evolving according to the transition matrix P :

$$P = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 2 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 3 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 4 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 \\ 5 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 6 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The inversion samplers for the family of equi-probable $\{\text{de Moivre}(1/i, 1/i, \dots, 1/i)\}_{i \in \{1, 2, \dots, 5\}}$ RVs (Algorithm 5) and the Point Mass(6) RV (Simulation 76) can be combined in the random mapping representation:

$$\rho(i, U) : \mathbb{X} \times [0, 1] \rightarrow \mathbb{X}, \quad \rho(i, U) = \lceil iU \rceil \mathbf{1}_{\{1, 2, 3, 4, 5\}}(i) + 6 \mathbf{1}_{\{6\}}(i), \quad U \sim \text{Uniform}(0, 1),$$

in order to simulate a sequence of states from this markov chain as follows:

$$X_0 \leftarrow 1, \quad X_i \leftarrow \rho(X_{i-1}, U_i), \quad i = 1, 2, \dots. \quad (9.12)$$

Simulation 136 (Trapped in lounge 6) Implement the Algorithm described in Equation 9.12 in a MATLAB program to simulate the first ten states visited by the Markov chain in Example 135. Recall the “Hotel California” character of lounge 6 – *you can check out anytime you like, but you can never leave!* Repeat this simulation 1000 times and find the fraction of times your are not trapped in lounge 6 by the tenth time step.

Exercise 137 (Drunkard’s walk around a polygonal block with k corners) Can you think of another way to simulate the “drunkard’s walk around a polygonal block with k corners” labelled by $0, 1, \dots, k - 1$ that is more efficient than using the `MCSimBydeMoivre` function which relies on the `SimdeMoivreOnce` function that implements Algorithm 6 with an average-case efficiency that is linear in k ?

Hint: think of the drunkard tossing a fair coin to make his decision of where to go next from each corner and arithmetic mod k .

9.3 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 61 (Communication between states) Let $(X_t)_{t \in \mathbb{Z}_+}$ be a homogeneous Markov chain with transition matrix P on state space $\mathbb{X} := \{s_1, s_2, \dots, s_k\}$. We say that a state s_i **communicates** with a state s_j and write $s_i \rightarrow s_j$ 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 $s_i \not\rightarrow s_j$ or $s_j \not\leftarrow s_i$.

We say that two states s_i and s_j **intercommunicate** and write $s_i \leftrightarrow s_j$ if $s_i \rightarrow s_j$ and $s_j \rightarrow s_i$. 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 $s_i \not\leftrightarrow s_j$.

Definition 62 (Irreducible) A homogeneous Markov chain $(X_t)_{t \in \mathbb{Z}_+}$ with transition matrix P on state space $\mathbb{X} := \{s_1, s_2, \dots, s_k\}$ 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 9.3. If $(p, q) \in (0, 1)^2$, then the corresponding Markov chain is irreducible because we can go from rollovia 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 rollovia 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 $\{P_{(p,q)} : p \in \{0\} \text{ or } q \in \{0\}\}$ 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. For instance, from the transition diagram in Figure 9.4 of the lounge-hopping Markov chain of Example 135, it is clear that if you start at state 6 you cannot find any arrow going to any other state. Therefore, the chain is reducible since $6 \not\rightarrow i$ for any $i \in \{1, 2, 3, 4, 5\}$.

Exercise 138 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 63 (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 64 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 65 (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 $\gcd(\{1, 2, 3, \dots\}) = 1$. However, the Markov chain model for a drunkard's walk around a block over the state space $\{0, 1, 2, 3\}$ (Simulation 133) is periodic because you can only return to the starting state in an even number of time steps and

$$\gcd(\mathbb{T}(0)) = \gcd(\mathbb{T}(1)) = \gcd(\mathbb{T}(2)) = \gcd(\mathbb{T}(3)) = \gcd(\{2, 4, 6, \dots\}) = 2 \neq 1 .$$

Exercise 139 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 66 Let $A = \{a_1, a_2, \dots\} \subset \mathbb{N}$ that satisfies the following two conditions:

1. A is a **nonlattice**, meaning that $\gcd(A) = 1$ and
2. A is closed under 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 67 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 68 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 $(x, y) \in \mathbb{X}^2$.

Proof: TBD

Exercise 140 (King's random walk on a chessboard) Consider the squares in the chessboard as the state space $\mathbb{X} = \{0, 1, 2, \dots, 7\}^2$ 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 irreducible and/or aperiodic?

Exercise 141 (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 $\mathbb{X} = \{0, 1, 2, \dots, 7\}^2$ 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 irreducible 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.

9.4 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, \dots) the value of X_t will keep fluctuating in the state space \mathbb{X} as $t \rightarrow \infty$ and we cannot hope for convergence to a fixed point state $x^* \in \mathbb{X}$ or to a k -cycle of states $\{x_1, x_2, \dots, x_k\} \subset \mathbb{X}$. 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 $t \rightarrow \infty$. The Markov chain convergence theorem indeed states that the distribution of X_t over \mathbb{X} settles down as $t \rightarrow \infty$, provided the Markov chain is irreducible and aperiodic.

Definition 69 (Stationary distribution) Let $(X_t)_{t \in \mathbb{Z}_+}$ be a Markov chain with state space $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$ and transition matrix $P = (P(x, y))_{(x,y) \in \mathbb{X}^2}$. A row vector

$$\pi = (\pi(s_1), \pi(s_2), \dots, \pi(s_k)) \in \mathbb{R}^{1 \times k}$$

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 70 (Hitting times) If a Markov chain $(X_t)_{t \in \mathbb{Z}_+}$ with state space $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$ and transition matrix $P = (P(x, y))_{(x,y) \in \mathbb{X}^2}$ starts at state x , then we can define the **hitting time**

$$T(x, y) = \min\{t \geq 1 : X_t = y\} .$$

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 71 (Hitting times of irreducible aperiodic Markov chains) If $(X_t)_{t \in \mathbb{Z}_+}$ is an irreducible aperiodic Markov chain with state space $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$, transition matrix $P = (P(x, y))_{(x,y) \in \mathbb{X}^2}$ then for any pair of states $(x, y) \in \mathbb{X}^2$,

$$\mathbf{P}(T(x, y) < \infty) = 1 ,$$

and the mean hitting time is finite, i.e.,

$$\tau(x, y) < \infty .$$

Proposition 72 (Existence of Stationary distribution) For any irreducible and aperiodic Markov chain there exists at least one stationary distribution.

Proof: TBD

Definition 73 (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 $\mathbb{X} := \{s_1, s_2, \dots, s_k\}$, then we define the **total variation distance** between ν_1 and ν_2 as

$$d_{TV}(\nu_1, \nu_2) := \frac{1}{2} \sum_{x \in \mathbb{X}} \text{abs}(\nu_1(x) - \nu_2(x)), \quad d_{TV} : \mathcal{P}(\mathbb{X}) \times \mathcal{P}(\mathbb{X}) \rightarrow [0, 1] . \quad (9.13)$$

If ν_1, ν_2, \dots and ν are probability distributions on \mathbb{X} , then we say that ν_t **converges in total variation** to ν as $n \rightarrow \infty$ and write $\nu_t \xrightarrow{TV} \nu$, if

$$\lim_{t \rightarrow \infty} d_{TV}(\nu_t, \nu) = 0 .$$

Observe that if $d_{TV}(\nu_1, \nu_2) = 0$ then $\nu_1 = \nu_2$. The constant $1/2$ in Equation 9.13 ensures that the range of d_{TV} is in $[0, 1]$. If $d_{TV}(\nu_1, \nu_2) = 1$ 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:

$$d_{TV}(\nu_1, \nu_2) = \max_{A \subset \mathbb{X}} \text{abs}(\nu_1(A) - \nu_2(A)) .$$

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 74 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 $t \rightarrow \infty$.

Proposition 74 (Markov chain convergence theorem) Let $(X_t)_{t \in \mathbb{Z}_+}$ be an irreducible aperiodic Markov chain with state space $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$, 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

$$\mu_t \xrightarrow{TV} \pi . \quad (9.14)$$

Proof: TBD

Proposition 75 (Uniqueness of stationary distribution) Any irreducible aperiodic Markov chain has a unique stationary distribution.

Proof: TBD

Exercise 142 Consider the Markov chain on $\{1, 2, 3, 4, 5, 6\}$ with the following transition matrix:

$$P = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & \left(\begin{array}{cccccc} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \end{array} \right) \\ 2 & \left(\begin{array}{cccccc} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \end{array} \right) \\ 3 & \left(\begin{array}{cccccc} 0 & 0 & \frac{1}{4} & \frac{3}{4} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} & 0 & 0 \end{array} \right) \\ 4 & \left(\begin{array}{cccccc} 0 & 0 & \frac{1}{4} & \frac{3}{4} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} & 0 & 0 \end{array} \right) \\ 5 & \left(\begin{array}{cccccc} 0 & 0 & 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{array} \right) \\ 6 & \left(\begin{array}{cccccc} 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{3}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{array} \right) \end{pmatrix} .$$

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 143 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 144 Show that for a drunkard's walk chain started at state 0 around a polygonal block with k corners labelled $\{0, 1, 2, \dots, k - 1\}$, the state probability vector at time step t

$$\mu_t \xrightarrow{\text{TV}} \pi$$

if and only if k is odd. Explain what happens to μ_t when k is even.

9.5 Reversibility

We introduce another specific property called reversibility. This property will assist in conjuring Markov chains with a desired stationary distribution.

Definition 76 (Reversible) A probability distribution π on $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$ 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 $(x, y) \in \mathbb{X}^2$:

$$\pi(x)P(x, y) = \pi(y)P(y, x) . \quad (9.15)$$

A Markov chain that has a reversible distribution is said to be a reversible Markov chain.

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 77 (A reversible π is a stationary π) Let $(X_t)_{t \in \mathbb{Z}_+}$ be a Markov chain on $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$ 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 $y \in \mathbb{X}$,

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

Definition 78 (Graph) A **Graph** $\mathbb{G} := (\mathbb{V}, \mathbb{E})$ consists of a **vertex set** $\mathbb{V} := \{v_1, v_2, \dots, v_k\}$ together with an **edge set** $\mathbb{E} := \{e_1, e_2, \dots, e_l\}$. 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 **neighbourhood** 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 26 (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 $\mathbb{V} := \{v_1, v_2, \dots, v_k\}$ 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

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

Proposition 79 The random walk on an undirected graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$, with vertex set $\mathbb{V} := \{v_1, v_2, \dots, v_k\}$ and degree sum $d = \sum_{i=1}^k \deg(v_i)$ is a reversible Markov chain with the reversible distribution π given by:

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

Proof: First note that π is a probability distribution provided that $d > 0$. To show that π is reversible we need to verify Equation 9.15 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 77 π is also the stationary distribution.

Exercise 145 Prove Proposition 79 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 146 (Random Walk on a regular graph) A graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ is called regular if every vertex in $\mathbb{V} = \{v_1, v_2, \dots, v_k\}$ 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 79, the stationary distribution of the random walk on \mathbb{G} is the uniform distribution on \mathbb{V} given by

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

Example 147 (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 :

$$A = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \end{matrix}, \quad P = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \end{matrix}, \quad \text{.}$$

By Proposition 79, the stationary distribution of the random walk on \mathbb{G} is

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

Exercise 148 Show that the Drunkard's walk around the block from Simulation 133 is a random walk on the undirected graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ with $\mathbb{V} = \{0, 1, 2, 3\}$ and $\mathbb{E} = \{\langle 0, 1 \rangle, \langle 1, 2 \rangle, \langle 2, 3 \rangle, \langle 0, 3 \rangle\}$. What is its reversible distribution?

Example 149 (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 $\mathbf{1}_{\{3\}}(x)$ and transition matrix

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

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. Since $(X_t)_{t \in \mathbb{Z}_+}$ is irreducible (aperiodicity is not necessary for uniqueness of π) π is the unique stationary distribution. Due to Proposition 77, 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 150 Find the stationary distribution of the Markov chain in Exercise 141.

Model 27 (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 $\mathbb{V} := \{v_1, v_2, \dots, v_k\}$ 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 151 (Directed Triangulated Quadrangle) The random walk on the directed 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 :

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 1 & 1 \\ 3 & 1 & 0 & 0 & 0 \\ 4 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad P = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 3 & 1 & 0 & 0 & 0 \\ 4 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad \text{Diagram: } \begin{array}{c} \text{1} \xrightarrow{\quad} \text{2} \xrightarrow{\quad} \text{4} \xrightarrow{\quad} \text{3} \\ \text{1} \xleftarrow{\quad} \text{2} \end{array} .$$

Exercise 152 Show that there is no reversible distribution for the Markov chain in Example 151.

Example 153 (Random surf on the word wide web) Consider the huge graph with vertices as webpages and hyper-links as undirected edges. Then Model 26 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.

Model 28 (Lazy Random Walk) You can convert a random walk on an undirected graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ into a **lazy random walk** on \mathbb{G} by the following steps:

- Add loops to each vertex in $\mathbb{V} = \{v_1, v_2, \dots, v_k\}$ to obtain a new set of edges $\mathbb{E}' = \mathbb{E} \cup \{\langle v_1, v_1 \rangle, \langle v_2, v_2 \rangle, \dots, \langle v_k, v_k \rangle\}$.
- Construct the lazy graph $\mathbb{G}' = (\mathbb{V}, \mathbb{E}')$.
- Do a random walk on the undirected graph \mathbb{G}' .

The lazy random walk allows us to introduce aperiodicity quite easily.

Exercise 154 (Lazy Random Walk on the Triangulated Quadrangle) Consider the random walk of Example 147 on the undirected graph

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

Construct the lazy random walk on \mathbb{G} , obtain its transition probability matrix and state transition diagram. Show that the stationary distribution of this lazy random walk on \mathbb{G} is

$$\pi = \left(\frac{3}{14}, \frac{4}{14}, \frac{4}{14}, \frac{3}{14} \right) .$$

Model 29 (Random Walks on Groups) Under 

Model 30 (Birth-Death chains) Under 

9.6 Metropolis-Hastings Markov chain

Definition 80 (Metropolis-Hastings Markov chain) If we are given an irreducible Markov chain $(Y_t)_{t \in \mathbb{Z}_+}$ called the **base chain** or the **proposal chain** on a finite state space $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$ with transition probability matrix $Q = (Q(x, y))_{(x,y) \in \mathbb{X}^2}$ and some probability distribution π on \mathbb{X} of interest that may only be known up to a normalizing constant as $\tilde{\pi}$, i.e., $\pi(x) = (\sum_{z \in \mathbb{X}} \tilde{\pi}(z))^{-1} \tilde{\pi}(x)$ for each $x \in \mathbb{X}$, then we can construct a new Markov chain $(X_t)_{t \in \mathbb{Z}_+}$ called the **Metropolis-Hastings** chain on \mathbb{X} with the following transition probabilities:

$$P(x, y) = \begin{cases} Q(x, y)a(x, y) & \text{if } x \neq y \\ 1 - \sum_{z \in \{z \in \mathbb{X}: z \neq x\}} Q(x, z)a(x, z) & \text{if } x = y \end{cases}, \quad (9.16)$$

where the acceptance probability is

$$a(x, y) := \min \left\{ 1, \frac{\pi(y)}{\pi(x)} \frac{Q(y, x)}{Q(x, y)} \right\}. \quad (9.17)$$

Note that we only need to know π up to ratios. Thus, $\pi(y)/\pi(x)$ in $a(x, y)$ can be replaced by $\tilde{\pi}(y)/\tilde{\pi}(x)$ since

$$\frac{\pi(y)}{\pi(x)} = \frac{(\sum_{z \in \mathbb{X}} \tilde{\pi}(z))^{-1} \tilde{\pi}(y)}{(\sum_{z \in \mathbb{X}} \tilde{\pi}(z))^{-1} \tilde{\pi}(x)} = \frac{\tilde{\pi}(y)}{\tilde{\pi}(x)}.$$

Algorithm 11 describes how to simulate samples from a Metropolis-Hastings Markov chain.

Proposition 81 (Stationarity of the Metropolis-Hastings chain) The Metropolis-Hastings chain constructed according to Definition 80 has π as its stationary distribution.

Proof: It suffices to show that π is the reversible distribution for $(X_t)_{t \in \mathbb{Z}_+}$, i.e., for each $(x, y) \in \mathbb{X}^2$, $\pi(x)P(x, y) = \pi(y)P(y, x)$. Fix a pair $(x, y) \in \mathbb{X}^2$ and suppose $x \neq y$. Then,

$$\begin{aligned} \pi(x)P(x, y) &= \pi(x)Q(x, y)a(x, y) \\ &= \pi(x)Q(x, y) \min \left\{ 1, \frac{\pi(y)}{\pi(x)} \frac{Q(y, x)}{Q(x, y)} \right\} \\ &= \min \left\{ \pi(x)Q(x, y), \pi(x)Q(x, y) \frac{\pi(y)}{\pi(x)} \frac{Q(y, x)}{Q(x, y)} \right\} \\ &= \min \{ \pi(x)Q(x, y), \pi(y)Q(y, x) \} \\ &= \min \{ \pi(y)Q(y, x), \pi(x)Q(x, y) \} \\ &= \min \left\{ \pi(y)Q(y, x), \pi(y)Q(y, x) \frac{\pi(x)}{\pi(y)} \frac{Q(x, y)}{Q(y, x)} \right\} \\ &= \pi(y)Q(y, x) \min \left\{ 1, \frac{\pi(x)}{\pi(y)} \frac{Q(x, y)}{Q(y, x)} \right\} \\ &= \pi(y)P(y, x). \end{aligned}$$

When $x = y$, reversibility is trivially satisfied since $\pi(x)P(x, y) = \pi(y)P(y, x) = \pi(x)P(x, x)$.

Definition 82 If the base chain $(Y_t)_{t \in \mathbb{Z}_+}$ in the Metropolis-Hastings Markov chain of Definition 80 has a symmetric transition matrix Q with $Q(x, y) = Q(y, x)$ for each $(x, y) \in \mathbb{X}^2$ then the acceptance probability in Equation 9.17 simplifies to

$$a(x, y) = \min \left\{ 1, \frac{\pi(y)}{\pi(x)} \right\},$$

and the corresponding Metropolis-Hastings chain $(X_t)_{t \in \mathbb{Z}_+}$ is called the **Metropolis chain**.

Algorithm 11 Metropolis-Hastings Markov chain1: *input:*(1) shape of a target density $\tilde{\pi}(x) = (\sum_{x \in \mathbb{X}} \tilde{\pi}(x)) \pi(x)$,(2) sampler for the base chain that can produce samples $y \sim Q(x, \cdot)$.2: *output:* a sequence of samples x_0, x_1, \dots, x_n from the Metropolis-Hastings Markov chain $(X_t)_{t \in \mathbb{Z}_+}$ with stationary distribution π 3: Choose initial state $x_0 \in \mathbb{X}$ according to μ_0 4: **repeat**5: At iteration t ,6: Generate $y \sim Q(x_{t-1}, \cdot)$ and $u \sim \text{Uniform}(0, 1)$,7: Compute *acceptance probability*

$$a(x_{t-1}, y) = \min \left\{ 1, \frac{\tilde{\pi}(y)}{\tilde{\pi}(x_{t-1})} \frac{Q(y, x_{t-1})}{Q(x_{t-1}, y)} \right\},$$

8: **If** $u \leq a(x_{t-1}, y)$ **then** $x_t \leftarrow y$, **else** $x_t \leftarrow x_{t-1}$ 9: **until** desired number of samples n are obtained from $(X_t)_{t \in \mathbb{Z}_+}$

Suppose you know neither the vertex set \mathbb{V} nor the edge set \mathbb{E} entirely for an undirected graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ but you are capable of walking locally on \mathbb{G} . In other words, if you are currently at vertex x you are able to make a move to one of the neighbouring vertices of x . However, you do not know every single vertex in \mathbb{V} or the entire set of edges \mathbb{E} as an adjacency matrix for instance. Several real-world problems fall in this class. Some examples include the random surfer on www to rank web pages (Example 153), social network analyses in facebook or twitter, exact tests for contingency tables, etc.

Model 31 (Metropolis-Hastings Random Walk on Graph) Let $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ be an undirected graph and let $(Y_t)_{t \in \mathbb{Z}_+}$ with transition matrix Q be an irreducible random walk on \mathbb{G} and let π be a probability distribution on $\mathbb{V} = \{v_1, v_2, \dots, v_k\}$ that is known upto a normalizing constant as $\tilde{\pi}$. The **Metropolis-Hastings random walk** on \mathbb{G} is the Metropolis-Hastings Markov chain $(X_t)_{t \in \mathbb{Z}_+}$ on \mathbb{V} with base chain $(Y_t)_{t \in \mathbb{Z}_+}$ and the following transition probabilities:

$$P(x, y) = \begin{cases} \frac{1}{\deg(v_i)} \min \left\{ 1, \frac{\tilde{\pi}(v_j)}{\tilde{\pi}(v_i)} \frac{\deg(v_i)}{\deg(v_j)} \right\} & \text{if } \langle v_i, v_j \rangle \in \mathbb{E} \\ 1 - \sum_{v_l \in \text{nbhd}(v_i)} \left(\frac{1}{\deg(v_i)} \min \left\{ 1, \frac{\tilde{\pi}(v_l)}{\tilde{\pi}(v_i)} \frac{\deg(v_i)}{\deg(v_l)} \right\} \right) & \text{if } v_i = v_j \\ 0 & \text{otherwise} \end{cases}.$$

By Proposition 81, $(X_t)_{t \in \mathbb{Z}_+}$ has π as its stationary distribution. This Markov chain can be simulated as follows:

- Suppose $x_t = v_i$ at time t
- Propose v_j uniformly at random from $\text{nbhd}(v_i)$
- Sample u from $\text{Uniform}(0, 1)$
- If $u < \min\{1, \pi(v_j) \deg(v_i) / \pi(v_i) \deg(v_j)\}$ then $x_{t+1} = v_j$ else $x_{t+1} = x_t$

Model 32 (Metropolis chain on a regular graph) Consider the random walk $(Y_t)_{t \in \mathbb{Z}_+}$ on a regular graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ with $\deg(v_i) = \delta$ for every vertex $v_i \in \mathbb{V} = \{v_1, v_2, \dots, v_k\}$ and the symmetric transition matrix

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

You can sample from a given distribution π on \mathbb{V} by constructing the Metropolis chain with stationary distribution π from the base chain given by $(Y_t)_{t \in \mathbb{Z}_+}$.

Model 33 (sampling from a uniform distribution over an irregular graph) A graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ that is not regular is said to be irregular. Clearly, the stationary distribution of a random walk on \mathbb{G} is not uniform. Suppose you want to sample uniformly from \mathbb{V} according to $\pi(v_i) = (\#\mathbb{V})^{-1}$ for each $v_i \in \mathbb{V}$. We can accomplish this by constructing a Metropolis-Hastings Markov chain with the random walk on \mathbb{G} as the base chain and the following transition probabilities:

$$P(v_i, v_j) = \begin{cases} \frac{1}{\deg(v_i)} \min \left\{ 1, \frac{\deg(v_i)}{\deg(v_j)} \right\} & \text{if } \langle v_i, v_j \rangle \in \mathbb{E} \\ 1 - \sum_{v_l \in \text{nbhd}(v_i)} \left(\frac{1}{\deg(v_i)} \min \left\{ 1, \frac{\deg(v_i)}{\deg(v_l)} \right\} \right) & \text{if } v_i = v_j \\ 0 & \text{otherwise} \end{cases}.$$

Thus the Metropolis-Hastings walk on \mathbb{G} is biased against visiting higher degree vertices and thereby samples uniformly from \mathbb{V} at stationarity.

Example 155 (Stochastic Optimization) Let $f : \mathbb{V} \rightarrow \mathbb{R}$ and $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ be an undirected graph. Let the global maximum be

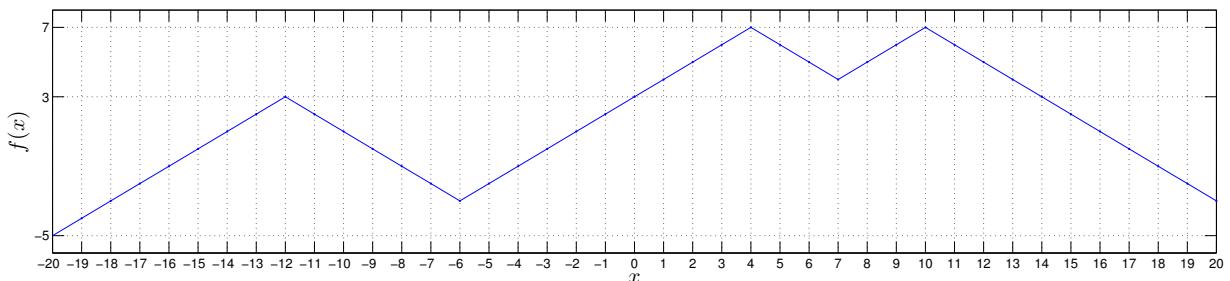
$$f^* := \max_{y \in \mathbb{V}} f(y),$$

and the set of maximizers of f be

$$\mathbb{V}^* := \underset{x \in \mathbb{V}}{\operatorname{argmax}} f(x) = \{x \in \mathbb{V} : f(x) = f^*\}.$$

In many problems such as maximum likelihood estimation, minimizing a cost function by maximizing its negative, etc, one is interested in $\mathbb{V}^* \subset \mathbb{V}$. This global maximization problem is difficult when $\#\mathbb{V}$ is huge. A deterministic hill-climbing or gradient ascent algorithm that iteratively moves from the current state v_i to a neighbouring state v_j if $f(v_j) > f(v_i)$ can easily get trapped in a local peak of f and thereby miss the global peak attained by elements in \mathbb{V}^* .

Figure 9.5: Stochastic Optimization with Metropolis chain.



For example consider the global maximization problem shown in Figure 9.5 with

$$f^* = 7 \text{ and } \mathbb{V}^* = \{4, 10\} \subset \mathbb{V} = \{-20, -19, \dots, 19, 20\} .$$

The deterministic hill-climbing algorithm will clearly miss \mathbb{V}^* and terminate at the local maximum of 3 at -12 if initialised at any element in $\{-20, -19, \dots, -8, -7\}$. Also, this algorithm will not find both elements in \mathbb{V}^* even when initialised more appropriately.

We will construct a Markov chain to solve this global maximization problem. For a fixed parameter $\lambda \in \mathbb{R}_{>0}$, let

$$\pi_\lambda(x) = \frac{\lambda^{f(x)}}{\sum_{z \in \mathbb{V}} \lambda^{f(z)}} .$$

Since $\pi_\lambda(x)$ is increasing in $f(x)$, $\pi_\lambda(x)$ favours vertices with large $f(x)$. First form a graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ by adding edges between the vertices in \mathbb{V} so that you can get from any vertex to any other vertex in \mathbb{V} by following a sequence of edges in \mathbb{E} . Now using the random walk on \mathbb{G} as the base chain let us construct a Metropolis-Hastings chain $(X_t)_{t \in \mathbb{Z}_+}$ on \mathbb{G} with π_λ on \mathbb{V} as its stationary distribution.

For simplicity, let us suppose that \mathbb{G} is a regular graph with a symmetric transition matrix Q for the base chain and thereby making $(X_t)_{t \in \mathbb{Z}_+}$ a Metropolis chain. For instance, in the Example from Figure 9.5 with $\mathbb{V} = \{-20, -19, \dots, 19, 20\}$, we can obtain a Metropolis chain on \mathbb{V} with stationary distribution π_λ by taking \mathbb{E} in $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ to be

$$\mathbb{E} = \{\langle -20, -19 \rangle, \langle -19, -18 \rangle, \langle -18, -17 \rangle, \dots, \langle 17, 18 \rangle, \langle 18, 19 \rangle, \langle 19, 20 \rangle\} .$$

Then, if $f(y) < f(x)$, the Metropolis chain accepts a transition from x to y with probability

$$\frac{\pi_\lambda(y)}{\pi_\lambda(x)} = \frac{\lambda^{f(y)}}{\lambda^{f(x)}} = \lambda^{f(y)-f(x)} = \lambda^{-(f(x)-f(y))} .$$

As $\lambda \rightarrow \infty$, the Metropolis chain approaches the deterministic hill-climbing algorithm and yields a uniform distribution over \mathbb{V}^* as follows:

$$\lim_{\lambda \rightarrow \infty} \pi_\lambda(x) = \lim_{\lambda \rightarrow \infty} \frac{\lambda^{f(x)}/\lambda^{f^*}}{\#\mathbb{V}^* + \sum_{z \in \mathbb{V} \setminus \mathbb{V}^*} \lambda^{f(z)}/\lambda^{f^*}} = \frac{\mathbb{1}_{\mathbb{V}^*}(x)}{\#\mathbb{V}^*} .$$

9.7 Glauber Dynamics

Let \mathbb{S} be a finite set of states. Let \mathbb{V} be a set of vertices. Typically, \mathbb{S} contains characters or colours that can be taken by each site or vertex in \mathbb{V} . Let $x \in \mathbb{S}^\mathbb{V}$ be a configuration, i.e., a function from \mathbb{V} to \mathbb{S} . A configuration can be thought of as a labelling of vertices in \mathbb{V} with elements in \mathbb{S} .

Definition 83 (Glauber dynamics for π) Let \mathbb{V} and \mathbb{S} be finite sets and let $\mathbb{X} \subset \mathbb{S}^\mathbb{V}$ which forms the support of the probability distribution π on $\mathbb{S}^\mathbb{V}$, i.e.,

$$\mathbb{X} = \{x \in \mathbb{S}^\mathbb{V} : \pi(x) > 0\} .$$

The **Glauber dynamics** or **Gibbs sampler** for π is a reversible Markov chain on \mathbb{X} with stationary distribution π under the following transition mechanism. Let the current state at time t be x . To obtain the state at time $t+1$ first choose a vertex v uniformly at random from \mathbb{V} and then choose

a new state according to π conditioned on the set of states equal to x at all vertices other than v . We give the details of this transition mechanism next.

For $x \in \mathbb{X}$ and $v \in \mathbb{V}$, define the set of states identical to x everywhere except possibly at v as

$$\mathbb{X}(x, v) := \{y \in \mathbb{X} : y(w) = x(w) \text{ for all } w \neq v\} .$$

Now let

$$\pi^{x,v}(y) := \pi(y|\mathbb{X}(x, v)) = \begin{cases} \left(\sum_{z \in \mathbb{X}(x, v)} \pi(z)\right)^{-1} \pi(y) & \text{if } y \in \mathbb{X}(x, v) \\ 0 & \text{if } y \notin \mathbb{X}(x, v) \end{cases}$$

be the distribution π conditioned on $\mathbb{X}(x, v)$. Therefore the rule for updating the current state x is:

- pick a vertex v uniformly at random from \mathbb{V} ,
- choose a new configuration by sampling from $\pi^{x,v}$.

Proposition 84 (Stationarity of Glauber dynamics) The Glauber dynamics for π on $\mathbb{X} \subset \mathbb{S}^{\mathbb{V}}$ has π as its reversible and stationary distribution.

Proof: Exercise.

Model 34 (Hard-core model) Let $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ be an undirected graph. An assignment of elements of $\mathbb{S} = \{0, 1\}$ to vertices in \mathbb{V} is called a configuration. Thus, the configuration x is a function $x : \mathbb{V} \rightarrow \mathbb{S}$ and $x \in \mathbb{S}^{\mathbb{V}}$. The vertices v of a configuration x with $x(v) = 1$ are said to be occupied and those with $x(v) = 0$ are said to be vacant. Thus a configuration models a placement of particles on the vertices of \mathbb{V} . A hard-core configuration is a configuration in which no two neighbouring vertices are occupied. More formally, a configuration x is called hard-core if $\sum_{(v_i, v_j) \in \mathbb{E}} x(v_i)x(v_j) = 0$. Let the set of hard-core configurations be \mathbb{X} and let π be the uniform distribution on \mathbb{X} , given by

$$\pi(x) = \begin{cases} \frac{1}{\#\mathbb{X}} & \text{if } x \in \mathbb{X} \\ 0 & \text{otherwise} \end{cases} .$$

The Glauber dynamics $(X_t)_{t \in \mathbb{Z}_+}$ for the uniform distribution π on hard-core configurations can be simulated as follows:

- initialize with vacant vertices, i.e., $X_0(w) = 0$ for each $w \in \mathbb{V}$,
- let the current hard-core configuration be $x_t : \mathbb{V} \rightarrow \{0, 1\}$ at time t ,
- choose a vertex v uniformly at random from \mathbb{V} ,
- if any neighbour of v is occupied then v is left vacant, i.e., $x_{t+1}(v) = 0$
- if every neighbour of v is vacant then v is occupied with probability $1/2$, i.e., $x_{t+1}(v) = 1$,
- leave the values at all other vertices unchanged, i.e., $x_{t+1}(w) = x_t(w)$ for each $w \neq v$,
- the possibly modified configuration x_{t+1} is the updated hard-core configuration at time $t + 1$.

Proposition 85 The Glauber dynamics of Model 34 does indeed have π as its stationary distribution.

Proof: First we need to verify that $(X_t)_{t \in \mathbb{Z}_+}$, the Markov chain given by the Glauber dynamics for π in Model 34, is irreducible and aperiodic. Clearly $(X_t)_{t \in \mathbb{Z}_+}$ is aperiodic since we can get from any hard-core configuration $x \in \mathbb{X}$ to itself in one time step by choosing a vertex with at least one occupied neighbour and leaving the chosen vertex unchanged or by choosing a vertex with no occupied neighbours and leaving the chosen vertex unchanged with probability $1/2$. Next we need to establish irreducibility, i.e., we need to show that we can get from any hardcore configuration x to any other hardcore configuration x' in finitely many steps. Let the vacant configuration be \tilde{x} , i.e., $\tilde{x}(v) = 0$ for every vertex $v \in \mathbb{V}$. In finitely many steps, we can go from any x to \tilde{x} and from \tilde{x} to x' . If x has $s(x) := \sum_{v \in \mathbb{V}} x(v)$ occupied sites or vertices then we can go to the vacant configuration \tilde{x} with $s(\tilde{x}) = 0$ in $s(x)$ time steps by picking one of the currently occupied sites and making it vacant as follows:

$$\mathbf{P}(X_{t+s(x)} = \tilde{x} | X_t = x) = \prod_{i=0}^{s(x)-1} \frac{(s(x) - i)}{\#\mathbb{V}} \frac{1}{2} > 0 .$$

Similarly, we can go from \tilde{x} to any other configuration x' with $s(x')$ many occupied sites in $s(x')$ time steps with the following positive probability:

$$\mathbf{P}(X_{t+s(x')} = x' | X_t = \tilde{x}) = \prod_{i=0}^{s(x')-1} \frac{(s(x') - i)}{\#\mathbb{V}} \frac{1}{2} > 0 .$$

Note that this is not the shortest possible number of steps to go from x to x' but just a finite number of steps. Thus we have established that $x \leftrightarrow x'$ for every $(x, x') \in \mathbb{X}$ and thereby established irreducibility of the chain $(X_t)_{t \in \mathbb{Z}_+}$.

If we now show that π is reversible for $(X_t)_{t \in \mathbb{Z}_+}$ then by Proposition 77 π is also stationary for $(X_t)_{t \in \mathbb{Z}_+}$ and finally π is the unique stationary distribution due to irreducibility and aperiodicity. Let $P(x, y)$ be the probability of going from x to y in one time step of $(X_t)_{t \in \mathbb{Z}_+}$. We need to show that for any pair of hardcore configurations $(x, y) \in \mathbb{X}^2$ the following equality holds:

$$\pi(x)P(x, y) = \pi(y)P(y, x), \quad \pi(x) = \frac{1}{\#\mathbb{X}} .$$

Let the number of vertices at which x and y differ be $d(x, y) := \sum_{v \in \mathbb{V}} \text{abs}(x(v) - y(v))$. Let us consider three cases of $(x, y) \in \mathbb{X}^2$.

Case i: When $d(x, y) = 0$ the two configurations are identical, i.e., $x = y$, and therefore we have the trivial equality:

$$\pi(x)P(x, y) = \pi(x)P(x, x) = \pi(y)P(y, x) .$$

Case ii: When $d(x, y) > 1$ the two configurations differ at more than one vertex and therefore $P(x, y) = 0$ and we have the trivial equality:

$$\pi(x)P(x, y) = \pi(x)0 = 0 = \pi(y)P(y, x) .$$

Case iii: When $d(x, y) = 1$ the two configurations differ at exactly one vertex v and therefore all neighbouring vertices of v must be vacant, i.e., take the value 0, in both x and y with $P(x, y) = P(y, x) = \frac{1}{\#\mathbb{V}} \frac{1}{2}$. Thus,

$$\pi(x)P(x, y) = \frac{1}{\#\mathbb{X}} \left(\frac{1}{\#\mathbb{V}} \frac{1}{2} \right) = \pi(y)P(y, x) .$$

We have established that $\pi(x) = 1/\#\mathbb{X}$ for each $x \in \mathbb{X}$ is the reversible distribution and thereby also the unique stationarity distribution for $(X_t)_{t \in \mathbb{Z}_+}$, the Markov chain given by the Glauber dynamics for π in Model 34.

Exercise 156 (1-D hardcore model) Let \mathbb{X}_n be the set of hardcore configurations on a path graph with n vertices. Recall that a path graph $\mathbb{G}_n = (\mathbb{V}_n, \mathbb{E}_n)$ has n vertices and $n - 1$ edges, as follows:

$$\mathbb{V}_n = \{v_1, v_2, \dots, v_n\}, \quad \mathbb{E}_n = \{\langle v_1, v_2 \rangle, \langle v_2, v_3 \rangle, \dots, \langle v_{n-1}, v_n \rangle\} .$$

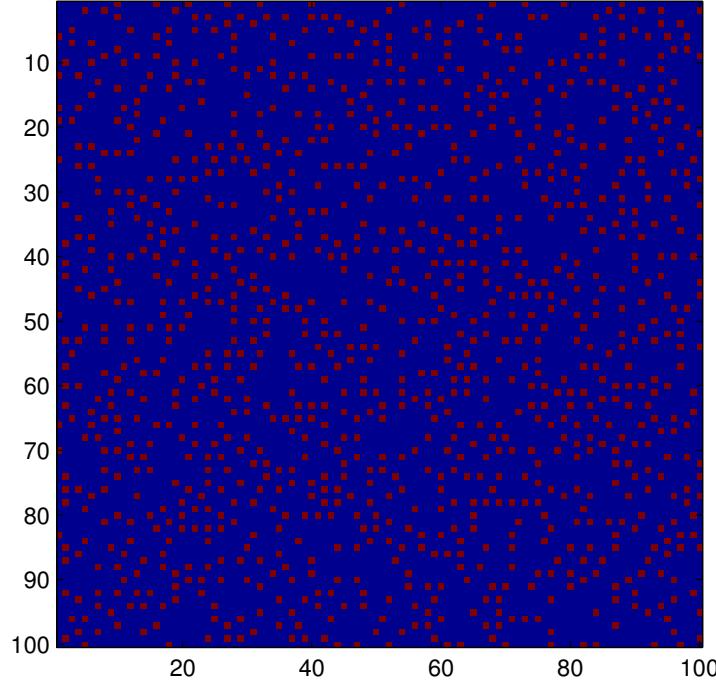
Draw all five hardcore configurations in \mathbb{X}_3 . Show that for any positive integer n ,

$$\#\mathbb{X}_n = \text{fib}(n + 1) ,$$

the $(n + 1)$ -th Fibonacci number, that is defined recursively as follows:

$$\text{fib}(0) := \text{fib}(1) := 1, \quad \text{fib}(n) = \text{fib}(n - 1) + \text{fib}(n - 2), \quad n \geq 1 .$$

Figure 9.6: The sample at time step 10^6 from the Glauber dynamics for the hardcore model on 100×100 regular torus grid. A red site is occupied while a blue site is vacant.



Simulation 157 (Glauber dynamics for the hardcore model on a 2D regular torus) Let us implement a program in MATLAB that will simulate Glauber dynamics to sample uniformly from the hardcore configurations on the undirected regular torus graph. We can report the sample mean of the fraction of occupied sites on this graph from the simulated sequence and make a movie of the simulations (last frame is shown in Figure 9.6).

```
>> Hardcore2D
Avg1s = 0.1128
```

The simulation was implemented in the following M-file:

```
HardCore2D.m
```

```
% simulation of Glauber dynamics for the hardcore model on
% 2D regular torus grid
clf; %clear; clc; % clear current settings
Seed=347632321; rand('twister',Seed); % set seed for PRNG
MaxSteps=1000000; % number of time steps to simulate
DisplayStepSize=10000; % display interval
Steps=0; % initialize time-step to 0
StepsM=1; % index for movie frame
Rows=100; % number of rows
Cols=100; % number of columns
CC = zeros(Rows,Cols,'int8'); %initialize all sites to be vacant
Delta=[-1,0,+1]; % neighbourhood of indices along one coordinate
Avg1s=0.0;%initialise the Average Fraction of occupied sites
while(Steps <= MaxSteps)
    % find a random site with 0 for possible swap
    I=ceil(Rows*rand); J=ceil(Cols*rand);
    % Get the Nbhd of CC(I,J)
    RowNbhd = mod((I-1)+Delta,Rows)+1;
    ColNbhd = mod((J-1)+Delta,Cols)+1;
    Nbhd=CC(RowNbhd, ColNbhd);
    To1Is=find(Nbhd); % find the 1s in Nbhd of CC(I,J)
    Num1s=length(To1Is); % total number of 1s in Nbhd
    if(Num1s > 0)
        CC(I,J)=0; % set site to be vacant
    elseif(rand < 0.5)
        CC(I,J)=1; % set site to be occupied
    else
        CC(I,J)=0; % set site to be vacant
    end
    Steps=Steps+1; % increment time step
    Frac1s=sum(sum(CC))/(Rows*Cols); % fraction of occupied sites
    Avg1s = Avg1s + (Frac1s - Avg1s)/Steps; % online sample mean
    if(mod(Steps,DisplayStepSize)==0)
        A(StepsM)=getframe; % get the frame into A
        imagesc(CC)
        axis square
        StepsM=StepsM+1;
    end
end
Avg1s % print the sample mean of fraction of occupied sites
movie(A,5) % make a movie
```

Model 35 (Ising model) Let $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ be an undirected graph. The Ising model is a probability distribution on $\mathbb{X} = \{-1, +1\}^{\mathbb{V}}$, i.e., a way of randomly assigning elements from the set $\{-1, +1\}$ to vertices of \mathbb{G} . The physical interpretation of the model is that each vertex is the position of an atom in a ferromagnetic material and $+1$'s or -1 's denote the two possible spin orientations of the atoms. There is a parameter β in the model called inverse temperature and $\beta \in [0, \infty)$. Associated with each spin configuration $x \in \mathbb{X}$ is its energy

$$H(x) = - \sum_{\langle u,v \rangle \in \mathbb{E}} x(u)x(v)$$

where $x(u)$ and $x(v)$ give the spin orientations of the atoms at vertices u and v , respectively. So, each edge $\langle u, v \rangle$ adds 1 to the energy $H(x)$ if its neighbouring vertices have opposite spins and subtracts 1 from $H(x)$ otherwise. Thus, lower energy is equivalent to a higher agreement in spins between neighbouring vertices.

The Ising model on \mathbb{G} at inverse temperature β means a random spin configuration X with

$$\mathbf{P}(X = x) = \pi_{\mathbb{G}, \beta}(x) = \frac{1}{Z_{\mathbb{G}, \beta}} \exp(-\beta H(x)) = \frac{1}{Z_{\mathbb{G}, \beta}} \exp\left(\beta \sum_{\langle u, v \rangle \in \mathbb{E}} x(u)x(v)\right),$$

where $Z_{\mathbb{G}, \beta} = \sum_{x \in \mathbb{X}} \exp(-\beta H(x))$ is the normalising constant.

Labwork 158 (Glauber dynamics for the Ising model on a 2D regular torus) Implement a program in MATLAB to simulate from the Ising model on the undirected regular torus graph.

Let us explore the physical interpretation of the Ising model further. If the inverse temperature $\beta = 0$ then we are at infinite temperature and therefore every configuration in \mathbb{X} is equally likely, i.e., $\pi_{\mathbb{G}, 0} = 1/\#\mathbb{X}$. At the other extreme, if $\beta \rightarrow \infty$ then we are approaching zero temperature and the probability over \mathbb{X} under $\pi_{\mathbb{G}, \infty}$ is equally split between “all +1” configuration and “all -1” configuration. However, if $\beta > 0$, then we are at some temperature $1/\beta$ that is neither absolutely hot or absolutely cold and therefore the model will favour configurations with lower energy as opposed to higher energy. Such favourable low energy configurations tend to have neighbouring clumps of identical spins. We say that there is a phase transition in β since the Ising model’s qualitative behaviour depends on whether β is above or below a critical threshold β_c .

Model 36 (Proper q -colourings) A proper q -colouring of an undirected graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ is an assignment of q colours labelled $\{1, 2, \dots, q\}$ to vertices in \mathbb{V} , subject to the constraint that neighbouring vertices do not receive the same colour. Let \mathbb{X} denote the set of all proper q -colourings of \mathbb{G} . If \mathbb{V} is large then \mathbb{X} can be a large and complicated subset of $\{1, 2, \dots, q\}^{\mathbb{V}}$. Note that proper q colourings are a natural generalisation of the hardcore model.

9.7.1 Random Walks on \mathbb{Z} and the reflection principle



9.8 Coupling from the past

MCMC algorithms make it easy to implement a Markov chain that has a given distribution as its stationary distribution. When used on their own, however, MCMC algorithms can only provide sample values that approximate a desired distribution. To obtain sample values that have a desired distribution *exactly* or *perfectly*, MCMC algorithms must be used in conjunction with ideas that make clever use of coupling.

MCMC convergence diagnostics based on *multiple* independent or *coupled* Markov chains running *forward* in time have been suggested, but are not completely reliable. The chains are coupled if the same sequence of random numbers is used to propagate all of them. By adopting a different perspective - running multiple coupled chains from the past or *backward coupling* - Propp & Wilson (1996) developed the *coupling from the past (CFTP)* algorithm, which allowed exact sample values to be obtained from the stationary distribution of an ergodic Markov chain with *finite* state space.

Let us first appreciate the trouble with MCMC algorithms such as Metropolis-Hastings chain, Metropolis chain and Glauber dynamics. Firstly, no matter how large we make time t to be we

cannot avoid the discrepancy between the t -step distribution μ_t and the stationary distribution π . Consider the following transition probability matrix:

$$P = \begin{matrix} & s_1 & s_2 \\ s_1 & \left(\begin{array}{cc} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{array} \right) \\ s_2 & & \end{matrix}$$

We can prove by induction that

$$\mu_t = \left(\frac{1}{2} (1 + 2^{-t}), \frac{1}{2} (1 - 2^{-t}) \right)$$

for every $t \in \mathbb{Z}_+$. The stationary distribution is $\pi = (1/2, 1/2)$. So, as t approaches infinity $\mu_t \xrightarrow{\text{TV}} \pi$, however for any t the total variation distance between $d_{\text{TV}}(\mu_t, \pi) = 2^{-t}$ is strictly positive. Even in this simple example μ_t may never equal π for any finite t , however large. Thus, we have to settle for an approximation to π with some acceptable error ϵ . Secondly, to make the approximation error measured by $d_{\text{TV}}(\mu_t, \pi)$ smaller than ϵ we have to find the ϵ -burnin time τ_ϵ by which $d_{\text{TV}}(\mu_{\tau_\epsilon}, \pi) < \epsilon$. Determining τ_ϵ is nontrivial except in special cases and constitutes an active field of research.

The following material is under .

Demonstration 159 (Applet – Perfect sampling.) The CFTP algorithm starts multiple Markov chains, one for each possible state, at some time $t_0 < 0$ in the past, and uses coupled transitions to propagate them to time 0. If all the chains *coalesce*, (i.e. end up having the same state, at or before time 0), then they will have “forgotten” their starting values and will evolve as a single chain from that point onwards. The common state at time zero ($X^{(0)}$) is an exact sample value from the stationary distribution. Intuitively, if coalescence occurs at some finite time, $t^* < 0$, then if the chains had been started in the infinite past, coupling with the same sequence of random numbers will ensure that they coalesce at t^* , and the common chain at time 0 must be stationary because it had been running for an infinitely long time. Thus, the existence of a finite coalescence time can give a stationary sample value in finite time. The use of coupling is essential to induce coalescence in a finite length of time.

Consider a Markov chain with finite state space, $S = 1, 2, \dots, K$. The CFTP algorithm starts K Markov chains, one from each state in S , at some time $t_0 < 0$ in the past. A sequence of t_0 random vectors, $R^{t+1}, R^{t+2}, \dots, R^0$, is generated and used to propagate all K Markov chains to time 0. Let $X^{t,k(t_0)}$ represent the state of the Markov chain at time t , starting from state $k \in S$ at time $t_0 < t$, and let φ be the update function of the Markov chain, such that:

$$X^{(t+1,k(t_0))} = \varphi(X^{(t,k(t_0))}, R^{(t+1)}) \quad (9.18)$$

9.8.1 Algorithm – Coupling from the past.

Set $t_0 = 0$.

Repeat

 Set $t_0 = t_0 - 1$, (take 1 time-step back)

 Generate $R^{(t_0+1)}$,

 For $k = 1, 2, \dots, K$, (for each state)

Set $X^{(t_0, k(t_0))} = k$, (start chain in that state)
 For $t = t_0, t_0 + 1, \dots, -1$, (propagate chain to time 0)
 Set $X^{(t+1, k(t_0))} = \varphi(X^{(t, k(t_0))}, R^{(t+1)})$.
 Until $X^{(0, 1(t_0))} = X^{(0, 2(t_0))} = \Lambda = X^{(0, K(t_0))}$. (check for coalescence at time 0)
 Return $X^{(0)}$.

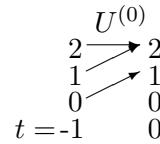
Example 160 Suppose that the Markov chain has the state space, $S = 0, 1, 2$, and a transition matrix:

$$Q = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.4 & 0.4 & 0.2 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$$

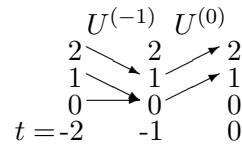
where the (i, j) -element is the conditional probability, $P(X^{(t+1)} = j | X^{(t)} = i)$. The matrix of conditional cumulative probabilities is

$$C = \begin{pmatrix} 0.6 & 0.9 & 1 \\ 0.4 & 0.8 & 1 \\ 0.3 & 0.7 & 1 \end{pmatrix}$$

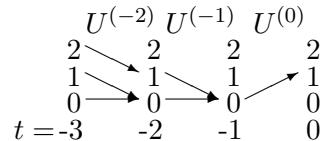
where the (i, j) -element is the probability, $P(X^{(t+1)} = j | X^{(t)} = i)$. Beginning at $t_0 = -1$, three chains are started at 0, 1 and 2. A uniform $(0, 1)$ random number, $U^{(0)}$, is generated (in this example, $R^{(0)} = U^{(0)}$) and used to propagate all three chains to time 0. Suppose that $U^{(0)} \in (0.8, 0.9)$. Then the three chains are updated as shown:



The chains have not coalesced at $t = 0$, so we need to move one time-step back to $t_0 = -2$, start three chains at 0, 1 and 2, generate a second uniform $(0, 1)$ random number, $U^{(-1)}$ and use it along with the previous $U^{(0)}$ to propagate the chains to time 0. Suppose that $U^{(-1)} \in (0.3, 0.4)$. The three chains then evolve as shown:



The chains have still not coalesced at $t = 0$, so we must move another time-step back to $t_0 = -3$ and start again, generating a third uniform $(0, 1)$ random number, $U^{(-2)}$. Suppose that $U^{(-2)} \in (0.3, 0.4)$. This is used with $U^{(-1)}$ and $U^{(0)}$ from before, giving the following transitions:



All three chains have now coalesced at $t = 0$ and so $X^{(0)} = 1$ is accepted as a sample value from the stationary distribution. The whole process is repeated to get another independent sample value. It is important to note that even though the chains have coalesced at $t = 1$, with the common value $X^{(-1)} = 0$; this value at the time of coalescence is not accepted as being from the stationary distribution. This is because the time of coalescence is a random time that depends only on the sequence of random numbers, $U^{(0)}, U^{(-1)}, \dots$; while the time at which a coalesced state has the required stationary distribution must be a fixed time. In the CFTP algorithm, this *fixed* time has been arbitrarily specified to be $t = 0$.

Example 161 To see that the state at the time of coalescence does not have the stationary distribution, suppose that the state space is $S = 1, 2$ and the transition matrix is:

$$Q = \begin{pmatrix} 0.5 & 0.5 \\ 1 & 0 \end{pmatrix}.$$

Since $Q(2, 1) = 1$, the two coupled chains must be in state 1 at the time of coalescence. However, the stationary distribution of this Markov chain is $f(1) = 2/3$ and $f(2) = 1/3$, and so the state at the time of coalescence cannot be from the stationary distribution.

Instead of taking a single step back when the two bounding chains fail to coalesce, any decreasing sequence of time-steps may be used. The “double-until-overshoot” choice of $t_0 = -2^0, -2^1, -2^2, \dots$ is optimal in the sense that it minimises the worst-case number of steps and almost minimises the expected number of steps for coalescence.

Exercise 162 Implement the CFTP algorithm for the Markov chain in Example 2.5.3 and use it to generate 1000 sample points from the stationary distribution of the chain. The stationary distribution can be shown to be:

x	0	1	2
$f(x)$	0.4789	0.3521	0.1690

Compare the relative frequencies of the generated sample with the true stationary probabilities.

Exercise 163 2.6.19 Consider a Markov chain with a state space $S = 0, 1, 2, 3$ and the transition matrix:

$$Q = \begin{pmatrix} 0.6 & 0.4 & 0 & 0 \\ 0.4 & 0.2 & 0.4 & 0 \\ 0.2 & 0.4 & 0 & 0.4 \\ 0 & 0.2 & 0.4 & 0.4 \end{pmatrix}.$$

Let $f = (f_0, f_1, f_2, f_3)$ be a row vector containing the stationary probabilities of the chain.

- (a) By solving $fQ = f$ and $f_0 + f_1 + f_2 + f_3 = 1$ simultaneously, show that the stationary distribution of the chain is $f = (14/35, 11/35, 6/35, 4/35)$.
- (b) Implement the “double-until-overshoot” version of the CFTP algorithm to generate from the stationary distribution, and use it to obtain 1000 sample points. Compare the relative frequencies of the generated sample with the true stationary probabilities.

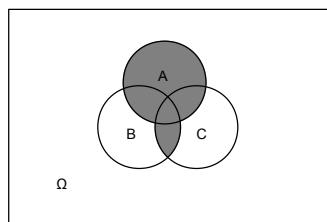
Chapter 10

Answers to Selected Exercises

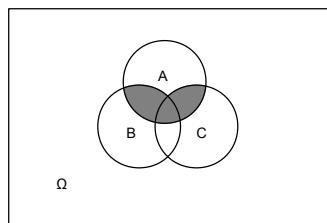
Answer (Ex. 1.1) — By operating with Ω , T , L and S we can obtain the answers as follows:

- | | |
|---|--|
| (a) $T \cap L = \{L_3\}$ | (f) $S \cap L = \emptyset$ |
| (b) $T \cap S = \emptyset$ | (g) $S^c \cap L = \{L_1, L_2, L_3\} = L$ |
| (c) $T \cup L = \{T_1, T_2, T_3, L_3, L_1, L_2\}$ | (h) $T^c = \{L_1, L_2, S_1, S_2, S_3, \dots, S_{50}\}$ |
| (d) $T \cup L \cup S = \Omega$ | (i) $T^c \cap L = \{L_1, L_2\}$ |
| (e) $S^c = \{T_1, T_2, T_3, L_3, L_1, L_2\}$ | (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:

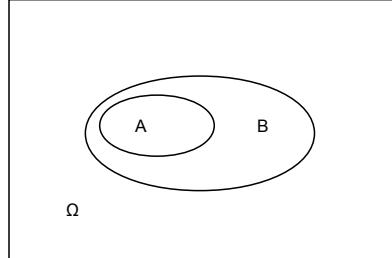


Answer (Ex. 1.4) — To illustrate the idea that $A \subseteq B$ if and only if $A \cup B = B$, we need to illustrate two implications:

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

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

The following Venn diagram illustrates the two implications clearly.



Answer (Ex. 2.1) — (a) $\mathbf{P}(\{Z\}) = 0.1\% = \frac{0.1}{100} = 0.001$

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

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

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

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

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

Answer (Ex. 2.2) — 1. {BB, BW, WB, WW}

2. {RRRR, RRRL, RRRL, RLRR, LRRR, RLRL, RRLL, LLRR, LRLR, LRRL, RLLL, LLLL, LLLR, LLRL, LRLL, RLLL}

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

Answer (Ex. 2.3) — 1. The sample space $\Omega = \{W, A, I, M, K, R\}$.

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

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

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. 2.4) — 1. First, the sample space is: $\Omega = \{B, I, N, G, O\}$.

2. The probabilities of simple events are:

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

3. Using the addition rule for mutually exclusive events,

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

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

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

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

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

Answer (Ex. 2.5) — 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 $\frac{2}{3} \cdot \frac{2}{3} = \frac{4}{9}$, 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 (Ex. 2.6) — 1. The sample space is

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

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

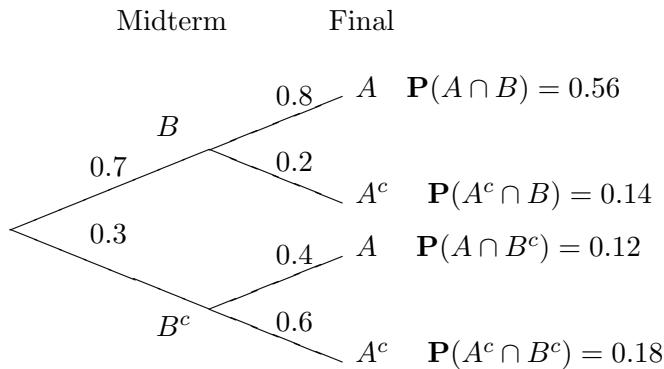
Let A be the event *the sum is 5* and B be the event *the sum is 6*, then A and B are mutually exclusive events with probabilities

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

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. 2.7) — 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:

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

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. 2.8) — 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 given 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,

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

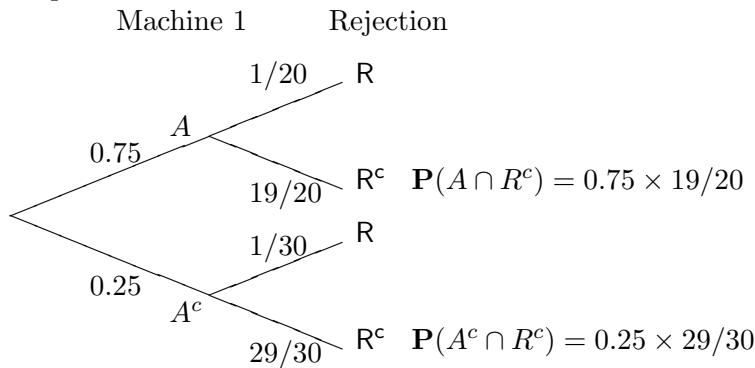
and,

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

Therefore,

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

The tree diagram for this problem is:



So the required probability is

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

Answer (Ex. 2.9) — 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

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

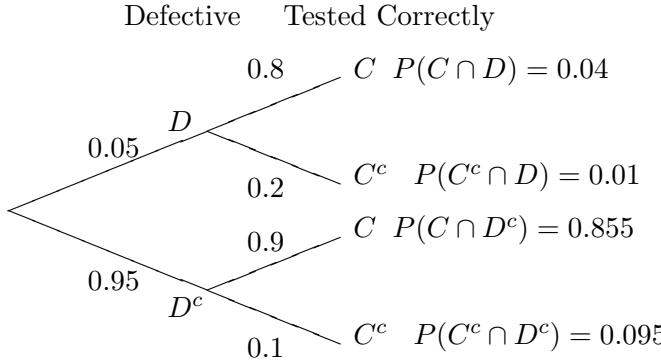
The probability that a micro-chip is defective, and has been declared as effective is

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

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

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

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

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

- (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:

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

Answer (Ex. 2.10) — (a) Let F_1 be the event a gale of force 1 occurs, let F_2 be the event a gale of force 2 occurs and F_3 be the event a gale of force 3 occurs. Now we know that

$$P(F_1) = \frac{2}{3}, \quad P(F_2) = \frac{1}{4}, \quad P(F_3) = \frac{1}{12}.$$

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

$$P(D|F_1) = \frac{1}{4}, \quad P(D|F_2) = \frac{2}{3}, \quad P(D|F_3) = \frac{5}{6}.$$

The probability that a reported gale causes damage is

$$P(D) = P(D \cap F_1) + P(D \cap F_2) + P(D \cap F_3)$$

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 probabilities 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

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

Similarly,

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

and

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

Answer (Ex. 6.1) — $\mathbf{P}(X = 3)$ 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. 6.2) — 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:

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

The graphs of $f(x)$ and $F(x)$ for random variable X are shown below:

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. 6.3) — 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 (Ex. 6.4) — (a)

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

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

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

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

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

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

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

Answer (Ex. 6.5) — Since we are sampling without replacement,

$$\mathbf{P}(X = 0) = \frac{4}{10} \cdot \frac{3}{9} = \frac{2}{15} \quad (\text{one way of drawing two right screws}),$$

$$\mathbf{P}(X = 1) = \frac{6}{10} \cdot \frac{4}{9} + \frac{4}{10} \cdot \frac{6}{9} = \frac{8}{15} \quad (\text{two ways of drawing one left and one right screw}),$$

$$\mathbf{P}(X = 2) = \frac{6}{10} \cdot \frac{5}{9} = \frac{1}{3} \quad (\text{one way of drawing two left screws}).$$

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.

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

2.

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

3.

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

Answer (Ex. 6.6) — 1. Since f is a probability mass function,

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

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

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

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. 6.7) — Note that $\theta = \frac{1}{2}$ here.

1. X has probability mass function

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

2. The required probabilities are:

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

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

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

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

Answer (Ex. 6.8) — 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

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

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

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

Answer (Ex. 6.9) — 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

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

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

$$1 - 0.3487 \approx 0.6513 .$$

Answer (Ex. 6.10) — 1. Since $f(x)$ is a (continuous) probability density function which integrates to one,

$$\int_{-4}^4 kdx = 1 .$$

That is,

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

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

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

If $-4 \leq x \leq 4$, 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 $x \geq 4$, 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

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

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

Answer (Ex. 6.11) — 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.).

Answer (Ex. 6.12) — 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, \dots, 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 $\text{Binomial}(n = 537, \theta = \frac{1}{576})$ 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 $\text{Binomial}(n = 537, \theta = \frac{1}{576})$ random variable by Poisson(λ) random variable with $\lambda = n\theta = \frac{537}{576} \approx 0.933$. 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:

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. 6.13) — 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. 6.14) — 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. 6.15) — 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:

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

Hence

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

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

Answer (Ex. 6.16) — The probability that *one* light bulb doesn't need to be replaced in 1200 hours is:

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

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. 6.17) — 1.

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

2.

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