

# Probability Theory 1 & Inference Theory I

Raazesh Sainudiin\*, Dominic Lee<sup>†</sup> and Michael Nussbaum<sup>•</sup>,

\*Laboratory for Mathematical Statistical Experiments, Uppsala Centre, and

<sup>•</sup>Department of Mathematics, Uppsala University, Uppsala, Sweden

<sup>†</sup>School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

<sup>•</sup>Department of Mathematics, Cornell University, Ithaca, New York, USA

Version Date: January 2, 2020

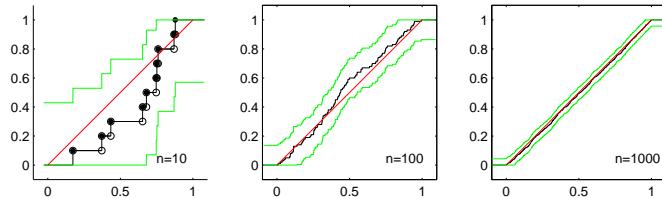
©2007–2019 Raazesh Sainudiin. ©2008–2019 Dominic Lee. ©2010–2019

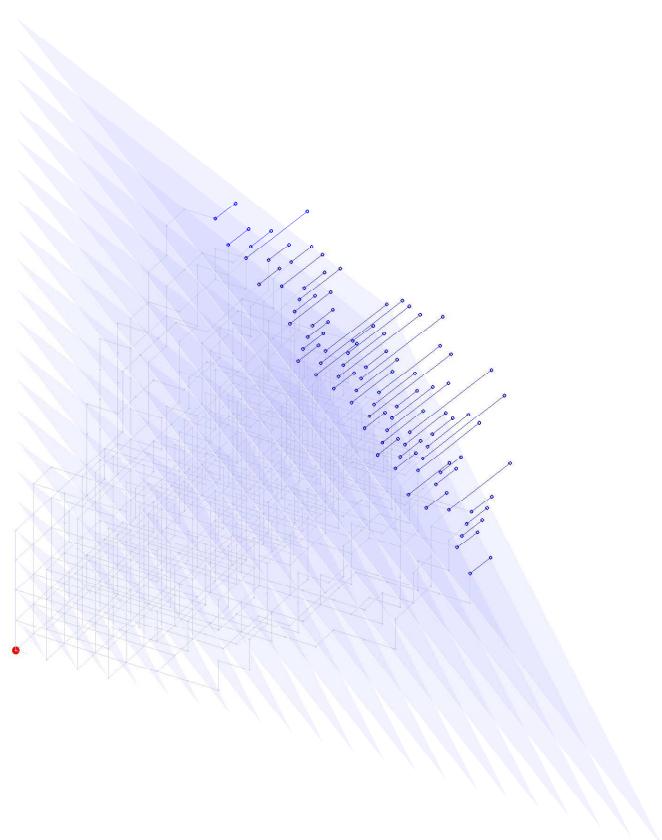
This work is licensed under the Creative Commons Attribution-Noncommercial-Share Alike 4.0

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, NSF/NIGMS grant DMS-02-01037 and Erskine Fellowship at Department of Statistics, University of Oxford..





# 1MS034 Syllabus

Course Code: 1MS034	Semester: Autumn 2019
Report Code: 10504	week: 36 – 44
33%, DAG, NML Uppsala University	$\xrightarrow{\text{work}}$ $12 = \lfloor \frac{33}{100} 37.5 \rfloor$ hours / week 2019-09-02 – 2019-11-03
<b>Probability Theory I, 5.0 c</b>	
Course Syllabus: Friday, 04th of October 2019	
Course Coordinator: Raazesh Sainudiin	

## Official Course Syllabus

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

### LEARNING OUTCOMES

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

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

### CONTENT

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

### ASSESSMENT

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

## Time Table – Prob. Theor. I

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

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

Lec.	Lab.	Week	Topics	Comprehension $\mapsto$ Action $\times$ Content
01	*	36	Preliminaries: Set Theory, Numbers, Functions ,...	RW Sec. 1.1,1.3,1.4 EX 1.2; RW Table. 1.1
			Elementary Combinatorics & Number Theory [optional] Introduction to MATLAB	RD 1.6; UD 6,7; EX 1.5; RD 1.9 PM 5,9,10,11,21
02			Probability	RD 2.1,2.2; EXs 2.3
03			Probability and Conditional Probability	RD 2.2,2.4; UD 31
04		37	Conditional Probability & Bayes Theorem	RD 2.4.1; UD 32,33,34; RD 2.4.2
05			Conditional Probability & Independence	RD 2.4.2; UD 35,36,37; EXs 2.5
06			Random variables	RD 3.1; EX 3.1; UD 1,41i, EX 3.2
07			Discrete Random variables and IID Bernoulli Trials	RD 3.1, 3.2.1, 3.2.3
08		38	Common Discrete Random variables	RD 3.2.3; UD 45; UD 4 ;EX 3.3;
09			Common Discrete Random variables	UD 46, 5, 47, 48, 6, 49, 50, 51
10			Continuous RVs	EX 3.4 & EXs 3.3;
11			Common Continuous RVs	RD 3.4; UD 52, 53, 54, RD 3.4.1
12		39	Common Continuous RVs	RD 3.4.2; UD 55, 56; EX 3.17, 3.18
13			Transformations of RVs – Discrete	EX 57, 58; UD 59; EXs 3.5
14			Expectations	RD 3.6; UD 60, 61; RW 3.6.1
15			Transformations of RVs – Discrete	RD 3.6.2, UD 62, 63, 64
16		40	Transformations of RVs – Continuous (1-to-1 & monotone)	RD 3.6.3; UD 65, 66, 67, 68
17			Transformations of RVs – Continuous (Direct method)	RD 3.6.3; UD 69, 70; EXs 3.7
18			Expectations	RD 3.8, 3.8.1; UD 71, 72, 74, 75, 76; EX 3.29; UD 77, 78, 79, 80; EXs 3.9
19			Multivariate Random Variables	RD 3.10; UD 83, 84, 85, 86,
20		41	Common $\mathbb{R}^m$ -valued RVs	UD 87, 88, 89, 90, 91, 92
21			Characteristic Functions	RD 3.10.2, 3.10.3; UD 95, 96
22			Statistics & Random Number Generation	RD 3.10.4; UD 97, 98;
23			Statistics	EX 3.36, 3.37, EXs 3.11
24		42	Simulation – Inversion & Rejection Samplers	RD 3.12; UD 101, 102, 103, 104, 105; EXs 3.13; [non-examinable] UD 169, 170
25			Convergence of RVs & Limit Laws	RD 3.14, 4.2, 4.3
26			Basic Inequalities & Law of Large Numbers	EXs 3.15
27			Law of Large Numbers & Central Limit Theorem via CFs	EXs 4.4
28		43	Model exam with live-scribed solutions	RD 5.1, UD 157, 158, 159, 160
29				UD 161, 163, 164; EX 5.1
30				UD 165, 166, 167, 168; EXs 5.4

# 1MS035 Syllabus

Course Code: 1MS035	Semester: Autumn 2019
Report Code: 10522	week: 45 – 03
33%, DAG, NML Uppsala University	$\xrightarrow{\text{work}}$ $12 = \lfloor \frac{33}{100} 37.5 \rfloor$ hours / week 2019-11-04 – 2020-01-19
<b>Inference Theory I, 5.0 c</b>	
Course Syllabus: Thursday, 07th of November 2019	
Course Coordinator: Raazesh Sainudiin	

## Official Course Syllabus

See <https://www.uu.se/en/admissions/master/selma/kursplan/?kpid=38972&type=1>.

### LEARNING OUTCOMES

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

- account for the bases of statistical studies and have knowledge of some methods for describing statistics;
- account for basic inference theoretical concepts and definitions;
- illustrate and interpret important concepts in concrete situations;
- design estimations and confidence intervals inclusive of in connection with linear regression;
- translate problems from relevant application fields into a form appropriate for statistical treatment, choose appropriate model and solution method;
- interpret and evaluate received results;
- use statistical software;

### CONTENT

Critical review of how statistics are presented and interpreted. General about statistical studies. Basic theory of point and interval estimations and hypothesis test. Correlation and regression. Parametric methods. Statistical software.

### ASSESSMENT

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

## Time Table – Inf. Theor. I

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

Table 2: Time Table for Inference Theory I

Lec.	Lab.	Week	Topics	Section/Labworks/Simulations
01		45	Overview of Inference & Decisions	RD 7.1–7.4
02			Review of LLN, CLT & Inference for proportions	
03		04	Inference for proportions and Estimators	RD 7.5, 7.5.2 7.5.3; EX 7.1, 7.2, 7.3
			Set up MATLAB scripts directory	download & unzip <code>scripts.zip</code> <sup>↓</sup>
			Introduction to MATLAB	PM 5,9,10,11,21
			Test your familiarity with MATLAB	PM 192, Try Labworks in Prob. Theor. I
05		46	Inference for proportions and Success/Failure Rule	RD 7.5, 7.5.3; EX 7.5; PM 193
06			Confidence Intervals for proportions	RD 7.5, 7.5.3, 7.5.4; EX 7.6
07			Model Student Project (5 bonus points – optional)	RD 9 and brainstorm
			Exact and asymptotic confidence intervals	RD 7.5.4
08		47	Inference for proportions contd. with Exercises	RD 7.5, 7.5.4;
09				RD 7.5.4; EX 7.4, 7.7
10				RD 7.5.4; EX 7.10, 7.11, 7.12, 7.13
11		48	Estimation with the Likelihood Principle	RD 7.8.1, 7.8.2; UD 210 214;
12			Method of Moment Estimator	PM 214; UD 214; EX 7.14, 7.15; UD 7.8.3
13		49	Hypothesis Testing: size, level, power, Wald Test & p-value	UD 7.7.1, 7.7.2, 7.7.4
14			Permutation Test & Chi-square	UD 7.7.5, 7.7.6
15				
16		50	DF Estimation & Plug-in Estimation	UD 7.13,7.13.1, 7.14
17			Bootstrap	UD 7.15.1, 7.15.2; LWS: 226, 227, 230, 231, 233
18			Linear Regression	UD 7.16
19		01	Fisher Information & Delta Method	UD 7.10,7.11,7.12; EX 217,219, 220,223,224
20		01	Review	

<sup>↓</sup>download and unzip `scripts.zip` into your computer with MATLAB :

`scripts.zip` is at <https://github.com/lamastex/computational-statistical-experiments/tree/master/matlab/csebook>

# Contents

<b>1MS034 Syllabus</b>	<b>3</b>
<b>1MS034 Time Table</b>	<b>4</b>
<b>1MS035 Syllabus</b>	<b>5</b>
<b>1MS035 Time Table</b>	<b>6</b>
<b>1 Preliminaries</b>	<b>18</b>
1.1 Elementary Set Theory . . . . .	18
1.2 Exercises . . . . .	20
1.3 Natural Numbers, Integers and Rational Numbers . . . . .	21
1.4 Real Numbers . . . . .	25
1.5 Introduction to MATLAB . . . . .	29
1.6 Elementary Combinatorics . . . . .	32
1.7 Array, Sequence, Limit, . . . . .	36
1.8 Elementary Real Analysis . . . . .	41
1.8.1 Limits of Real Numbers – A Review . . . . .	41
1.9 Elementary Number Theory . . . . .	44
<b>2 Probability Model</b>	<b>45</b>
2.1 Experiments . . . . .	45
2.2 Probability . . . . .	47
2.2.1 Consequences of our Definition of Probability . . . . .	49
2.2.2 Sigma Algebras of Typical Experiments* . . . . .	51
2.3 Exercises in Probability . . . . .	53
2.4 Conditional Probability . . . . .	54
2.4.1 Bayes' Theorem . . . . .	56
2.4.2 Independence and Dependence . . . . .	60
2.5 Exercises in Conditional Probability . . . . .	62

<b>CONTENTS</b>	<b>8</b>
<b>3 Random Variables</b>	<b>65</b>
3.1 Basic Definitions . . . . .	67
3.2 Discrete Random Variables . . . . .	70
3.2.1 An Elementary Family of Bernoulli Random Variables . . . . .	74
3.2.2 Independent Bernoulli Trials . . . . .	75
3.2.3 Some Common Discrete Random Variables . . . . .	76
3.3 Exercises in Discrete Random Variables . . . . .	85
3.4 Continuous Random Variables . . . . .	87
3.4.1 An Elementary Continuous Random Variable . . . . .	90
3.4.2 Some Common Continuous Random Variables . . . . .	91
3.5 Exercises in Continuous Random Variables . . . . .	97
3.6 Transformations of random variables . . . . .	97
3.6.1 A Review of Inverse Images . . . . .	98
3.6.2 Transformations of discrete random variables . . . . .	100
3.6.3 Transformations of continuous random variables . . . . .	101
3.7 Exercises in Transformations of Random Variables . . . . .	108
3.8 Expectations . . . . .	108
3.8.1 Expectations of functions of random variables . . . . .	109
3.8.2 Properties of expectations . . . . .	112
3.8.3 Expectation of Common Random Variables . . . . .	113
3.9 Exercises in Expectations of Random Variables . . . . .	118
3.10 Multivariate Random Variables . . . . .	119
3.10.1 $\mathbb{R}^2$ -valued Random Variables . . . . .	120
3.10.2 Conditional Random Variables . . . . .	130
3.10.3 $\mathbb{R}^m$ -valued Random Variables . . . . .	132
3.10.4 Some Common $\mathbb{R}^m$ -valued RVs . . . . .	136
3.10.5 Dependent Random Variables . . . . .	141
3.11 Exercises in Multivariate Random Variables . . . . .	142
3.12 Characteristic Functions . . . . .	145
3.12.1 Obtaining Moments from Characteristic Function . . . . .	145
3.12.2 Moment Generating Function . . . . .	150
3.13 Exercises in Characteristic Functions . . . . .	150
3.14 Statistics . . . . .	151
3.14.1 Data and Statistics . . . . .	151
3.14.2 Univariate Data . . . . .	158
3.14.3 Bivariate Data . . . . .	160
3.14.4 Trivariate Data . . . . .	161
3.14.5 Multivariate Data . . . . .	162
3.14.6 Loading and Exploring Real-world Data . . . . .	163
3.14.7 Geological Data . . . . .	163
3.14.8 Metereological Data . . . . .	166
3.14.9 Textual Data . . . . .	170
3.14.10 Machine Sensor Data . . . . .	171
3.15 Exercises in Statistics . . . . .	172

<b>CONTENTS</b>	<b>9</b>
<b>4 Simulation</b>	<b>173</b>
4.1 Physical Random Number Generators . . . . .	173
4.2 Pseudo-Random Number Generators . . . . .	173
4.2.1 Linear Congruential Generators . . . . .	174
4.2.2 Generalized Feedback Shift Register and the “Mersenne Twister” PRNG . . .	177
4.3 Simulation of non-Uniform(0, 1) Random Variables . . . . .	179
4.3.1 Inversion Sampler for Continuous Random Variables . . . . .	179
4.3.2 Inversion Sampler for Discrete Random Variables . . . . .	187
4.3.3 von Neumann Rejection Sampler (RS) . . . . .	196
4.4 Exercises in Simulation . . . . .	201
<b>5 Limit Laws of Statistics</b>	<b>202</b>
5.1 Convergence of Random Variables . . . . .	202
5.1.1 Properties of Convergence of RVs** . . . . .	208
5.2 Law of Large Numbers . . . . .	208
5.2.1 Application: Point Estimation of $E(X_1)$ . . . . .	211
5.3 Central Limit Theorem . . . . .	213
5.3.1 Application: Tolerating Errors in our estimate of $E(X_1)$ . . . . .	214
5.3.2 Application: Set Estimation of $E(X_1)$ . . . . .	216
5.4 Exercises in Limit Laws of Statistics . . . . .	217
<b>6 Finite Markov Chains</b>	<b>218</b>
6.1 Stochastic Processes . . . . .	218
6.2 Introduction . . . . .	219
6.3 Irreducibility and Aperiodicity . . . . .	227
6.4 Stationarity . . . . .	229
6.5 Reversibility . . . . .	231
6.6 Standard normal distribution function table . . . . .	235
<b>Summary of Probability Theory I</b>	<b>236</b>
<b>7 Inference for Statistical Experiments</b>	<b>241</b>
7.1 Introduction . . . . .	241
7.2 Some Common Experiments . . . . .	241
7.3 Typical Decision Problems with Experiments . . . . .	243
7.4 Decision Problems and Procedures for Actions . . . . .	244
7.5 Statistics for Bernoulli Trials (Inference for Proportions) . . . . .	245
7.5.1 Testing biasedness of a coin . . . . .	245
7.5.2 Review of underlying probability concepts . . . . .	246

<b>CONTENTS</b>	<b>10</b>
7.5.3 The success / failure rule . . . . .	253
7.5.4 Confidence intervals for a proportion . . . . .	264
7.6 Fundamentals of Estimation . . . . .	279
7.6.1 Introduction . . . . .	279
7.6.2 Point Estimation . . . . .	279
7.6.3 Some Properties of Point Estimators . . . . .	280
7.6.4 Confidence Set Estimation . . . . .	283
7.7 Fundamentals of Hypothesis Testing . . . . .	287
7.7.1 Introduction . . . . .	287
7.7.2 The Wald Test . . . . .	288
7.7.3 A Composite Hypothesis Test . . . . .	290
7.7.4 p-values . . . . .	292
7.7.5 Permutation Test for the equality of any two DFs . . . . .	293
7.7.6 Pearson's Chi-Square Test for Multinomial Trials . . . . .	296
7.8 Parameter Estimation and Likelihood . . . . .	300
7.8.1 Point and Set Estimation – A General Likelihood Approach . . . . .	300
7.8.2 Likelihood . . . . .	300
7.8.3 Moment Estimator (MME) . . . . .	311
7.9 Practical Excursion in One-dimensional Optimisation . . . . .	312
7.10 More Properties of the Maximum Likelihood Estimator . . . . .	316
7.11 Fisher Information . . . . .	316
7.12 Delta Method . . . . .	322
7.13 Non-parametric DF Estimation . . . . .	326
7.13.1 Estimating DF . . . . .	327
7.14 Plug-in Estimators of Statistical Functionals . . . . .	331
7.15 Bootstrap . . . . .	334
7.15.1 Non-parametric Bootstrap for Confidence Sets . . . . .	334
7.15.2 Parametric Bootstrap for Confidence Sets . . . . .	337
7.16 Linear Regression . . . . .	340
7.16.1 Introduction . . . . .	340
7.16.2 Simple Linear Regression . . . . .	340
7.16.3 Least Squares and Maximum Likelihood . . . . .	341
7.16.4 Properties of the Least Squares Estimator (LSE) . . . . .	342
<b>Answers to Selected Exercises</b>	<b>344</b>
<b>8 Appendix</b>	<b>382</b>
8.1 Code . . . . .	382
8.2 Data . . . . .	389

<i>CONTENTS</i>	11
<b>9 Student Projects</b>	<b>391</b>
9.1 Testing the Approximation of $\pi$ by Buffon's Needle Test . . . . .	391
9.1.1 Introduction & Motivation . . . . .	391
9.1.2 Materials & Methods . . . . .	392
9.1.3 Results . . . . .	397
9.1.4 Conclusion . . . . .	397
9.2 Estimating the Binomial probability $p$ for a Galton's Quincunx . . . . .	400
9.2.1 Motivation & Introduction . . . . .	400
9.2.2 Materials and Methods . . . . .	401
9.2.3 Statistical Methodology . . . . .	402
9.2.4 Results & Conclusion . . . . .	403
9.3 Investigation of a Statistical Simulation from the 19th Century . . . . .	404
9.3.1 Introduction and Motivation . . . . .	404
9.3.2 Statistical Methodology . . . . .	405
9.3.3 Results . . . . .	407
9.3.4 Conclusion . . . . .	409
9.4 Testing the average waiting time for the Orbiter Bus Service . . . . .	412
9.4.1 Motivation . . . . .	412
9.4.2 Method . . . . .	413
9.4.3 Results . . . . .	414
9.4.4 Discussion . . . . .	414
9.4.5 Conclusion . . . . .	417
9.5 Diameter of <i>Dosinia</i> Shells . . . . .	422
9.5.1 Introduction and Objective . . . . .	422
9.5.2 Materials and Methods . . . . .	422
9.5.3 Results . . . . .	423
<b>Index</b>	<b>424</b>

# List of Tables

1	Time Table for Virtual Student of Probability Theory I . . . . .	4
2	Time Table for Inference Theory I . . . . .	6
1.1	Symbol Table: Sets and Numbers . . . . .	28
3.1	The 8 $\omega$ 's in the sample space $\Omega$ of the experiment $\mathcal{E}_\theta^3$ are given in the first row above. The RV $Y$ is the number of 'Heads' in the 3 tosses and the RV $Z$ is the number of 'Tails' in the 3 tosses. Finally, the RVs $Y'$ and $Z'$ are the indicator functions of the event that 'all three tosses were Heads' and the event that 'all three tosses were Tails', respectively. . . . .	131
6.1	Symbol Table: Probability and Statistics . . . . .	238
6.2	Random Variables with PDF and PMF (using indicator function), Mean and Variance	239
6.3	Symbol Table: Sets and Numbers . . . . .	239
6.4	Symbol Table: Probability and Statistics . . . . .	240
7.1	Outcomes of an hypothesis test. . . . .	288
7.2	Some terminology in hypothesis testing. . . . .	288
7.3	Evidence scale against the null hypothesis in terms of the range of p-value. . . . .	292

# List of Figures

1.1	Union and intersection of sets shown by Venn diagrams . . . . .	19
1.2	These Venn diagram illustrate De Morgan’s Laws. . . . .	20
1.3	A function $f$ (“father of”) from $\mathbb{X}$ (a set of children) to $\mathbb{Y}$ (their fathers) and its inverse (“children of”). . . . .	23
1.4	A pictorial depiction of addition and its inverse. The domain is plotted in orthogonal <b>Cartesian coordinates</b> . . . . .	24
1.5	A depiction of the real line segment $[-10, 10]$ . . . . .	26
1.6	Point plot and stem plot of the finite sequence $\langle b_{1:10} \rangle$ declared as an array. . . . .	38
1.7	A plot of the sine wave over $[-2\pi, 2\pi]$ . . . . .	41
2.1	A binary tree whose leaves are all possible outcomes. . . . .	47
2.2	First ball number in 1114 NZ Lotto draws from 1987 to 2008. . . . .	50
2.3	Reference to the Venn diagram will help you understand this idea behind the proof of the total probability theorem in Proposition 14 for the four event case. . . . .	58
3.1	The Indicator function of event $A \in \mathcal{F}$ is a RV $\mathbf{1}_A$ with DF $F$ . . . . .	69
3.2	A RV $X$ from a sample space $\Omega$ with 8 elements to $\mathbb{R}$ and its DF $F$ . . . . .	69
3.3	$f(x)$ and $F(x)$ of the <i>fair coin toss</i> random variable $X$ , a discrete uniform RV on $\{0, 1\}$ . . . . .	72
3.4	$f(x)$ and $F(x)$ of the <i>fair die toss</i> random variable $X$ , a discrete uniform RV on $\{1, 2, 3, 4, 5, 6\}$ . . . . .	73
3.5	$f(x)$ and $F(x)$ of surmised <i>astragali toss</i> random variable $X$ , a discrete (non-uniform) RV on $\{1, 2, 3, 4\}$ . . . . .	74
3.6	PMF $f(x; \theta)$ and DF $f(x; \theta)$ with $\theta = 0.33$ . You should see how PMF and DF change as $\theta$ goes from 0 to 1 . . . . .	75
3.7	PMF of $X \sim \text{Geometric}(\theta = 0.5)$ and the relative frequency histogram based on 100 and 1000 samples from $X$ according to Simulation 144 and Labwork 145 you will see in the sequel. . . . .	78
3.8	PDF of $X \sim \text{Binomial}(n = 10, \theta = 0.5)$ and the relative frequency histogram based on 100,000 samples from $X$ obtained according to Simulation 148. . . . .	80
3.9	Figures from Sir Francis Galton, F.R.S., <i>Natural Inheritance</i> , , Macmillan, 1889. . . . .	81
3.10	$f(x)$ and $F(x)$ of the $\text{Uniform}(0, 1)$ random variable $X$ . . . . .	90

3.11 A convenient but mathematically imprecise Matlab plot from a polyline interpolation for the PDF and DF or CDF of the Uniform(0, 1) continuous RV $X$ . . . . .	91
3.12 $f(x; \lambda)$ and $F(x; \lambda)$ of an exponential random variable where $\lambda = 1$ (dotted) and $\lambda = 2$ (dashed). . . . .	92
3.13 Density and distribution functions of $\text{Exponential}(\lambda)$ RVs, for $\lambda = 1, 10, 10^{-1}$ , in four different axes scales. . . . .	93
3.14 $f(x)$ and $F(x)$ of the $\text{Uniform}(\theta_1, \theta_2)$ random variable $X$ . . . . .	94
3.15 PDF and DF of a $\text{Normal}(\mu, \sigma^2)$ RV for different values of $\mu$ and $\sigma^2$ . . . . .	105
3.16 Mean ( $E_\theta(X)$ ), variance ( $V_\theta(X)$ ) and the rate of change of variance ( $\frac{d}{d\theta} V_\theta(X)$ ) of a Bernoulli( $\theta$ ) RV $X$ as a function of the parameter $\theta$ . . . . .	113
3.17 Mean and variance of a Geometric( $\theta$ ) RV $X$ as a function of the parameter $\theta$ . . . . .	115
3.18 PDF of $X \sim \text{Poisson}(\lambda = 10)$ and the relative frequency histogram based on 1000 samples from $X$ according to Simulation 149. . . . .	116
3.19 Diagrams done on the board! . . . . .	130
3.20 Visual Cognitive Tool GUI: Quincunx & Septcunx. . . . .	138
3.21 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 Bernoulli( $\theta = 0.5$ ) random vectors over $\{(1, 0), (0, 1)\}$ with probabilities $\{\theta, 1 - \theta\}$ , respectively. The blue lines and black asterisks perpendicular to and above the diagonal line, i.e. the line connecting $(0, 10)$ and $(10, 0)$ , are the density histogram of the samples and the PMF of our $\text{Binomial}(n = 10, \theta = 0.5)$ RV, respectively. . . . .	139
3.22 JPDF, Marginal PDFs and Frequency Histogram of Bivariate Standard Normal $\vec{V}$ . .	141
3.23 JPDF, Marginal PDFs and Frequency Histogram of a Bivariate Normal $\vec{V}$ for lengths of girths of cylindrical shafts in a manufacturing process (in cm) . . . . .	142
3.24 Sample Space, Random Variable, Realisation, Data, and Data Space. . . . .	152
3.25 Data Spaces $\mathbb{X} = \{0, 1\}^2$ and $\mathbb{X} = \{0, 1\}^3$ for two and three Bernoulli trials, respectively.	152
3.26 Plot of the DF of $\text{Uniform}(0, 1)$ , 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. . . . .	157
3.27 Frequency, Relative Frequency and Density Histograms . . . . .	159
3.28 Frequency, Relative Frequency and Density Histograms . . . . .	160
3.29 2D Scatter Plot . . . . .	161
3.30 3D Scatter Plot . . . . .	162
3.31 Plot Matrix of uniformly generated data in $[0, 1]^5$ . . . . .	162
3.32 Matrix of Scatter Plots of the latitude, longitude, magnitude and depth of the 22-02-2011 earth quakes in Christchurch, New Zealand. . . . .	165
3.33 Google Earth Visualisation of the earth quakes . . . . .	167
3.34 Daily rainfalls in Christchurch since March 27 2010 . . . . .	168
3.35 Daily temperatures in Christchurch for one year since March 27 2010 . . . . .	169
3.36 Wordle of JOE 2010 . . . . .	170
3.37 Double Pendulum . . . . .	171

4.1	The linear congruential sequence of LinConGen(256, 137, 0, 123, 257) with non-maximal period length of 32 as a line plot over $\{0, 1, \dots, 256\}$ , scaled over $[0, 1]$ by a division by 256 and a histogram of the 256 points in $[0, 1]$ with 15 bins. . . . .	175
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. . . . .	177
4.3	Triplet point clouds from the “Mersenne Twister” with two different seeds (see Lab-work 130). . . . .	179
4.4	A plot of the PDF, DF or CDF and inverse DF of the Uniform( $-1, 1$ ) RV $X$ . . . . .	180
4.5	Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Uniform}(-5, 5)$ . . . . .	181
4.6	The PDF $f$ , DF $F$ , and inverse DF $F^{[-1]}$ of the the Exponential( $\lambda = 1.0$ ) RV. . . . .	182
4.7	Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Exponential}(0.5)$ . . . . .	183
4.8	Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Laplace}(5)$ . . . . .	185
4.9	Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Cauchy}$ . . . . .	186
4.10	The DF $F(x; 0.3, 0.7)$ of the de Moivre( $0.3, 0.7$ ) RV and its inverse $F^{[-1]}(u; 0.3, 0.7)$ . .	189
4.11	Visual Cognitive Tool GUI: Rejection Sampling from $X \sim \text{Normal}(0, 1)$ with PDF $f$ based on proposals from $Y \sim \text{Laplace}(1)$ with PDF $g$ . . . . .	197
4.12	Rejection Sampling from $X \sim \text{Normal}(0, 1)$ with PDF $f$ based on 100 proposals from $Y \sim \text{Laplace}(1)$ with PDF $g$ . . . . .	198
5.1	Sequence of $\{\text{Point Mass}(17)\}_{i=1}^{\infty}$ RVs (left panel) and $\{\text{Point Mass}(1/i)\}_{i=1}^{\infty}$ RVs (only the first seven are shown on right panel) and their limiting RVs in red. . . . .	203
5.2	Distribution functions of several $\text{Normal}(\mu, \sigma^2)$ RVs for $\sigma^2 = 1, \frac{1}{10}, \frac{1}{100}, \frac{1}{1000}$ . . . . .	203
5.3	PDF $f_{X_n}(x) := \mathbb{1}_{(0,1)}(x)(1 - \cos(2\pi nx))$ of the RV $X_n$ [the left sub-figure] and its DF $F_n(x) := \int_{-\infty}^x \mathbb{1}_{(0,1)}(v)(1 - \cos(2\pi nv))dv$ [the right sub-figure], for $n = 1$ [red ' - - '], $n = 10$ [blue ' - . '], and $n = 100$ [green ' - ], respectively. One can see clear convergence of the DFs $F_n$ to $\mathbb{1}_{(0,1)}(x)x$ , the DF of the Uniform( $0, 1$ ) RV, while the corresponding PDFs $f_n(x)$ keep oscillating wildly with $n$ across $[0, 2]$ about $\mathbb{1}_{(0,1)}(x)$ , the PDF of the Uniform( $0, 1$ ) RV $X$ . Thus giving a counter-example to the claim that convergence in DFs does not imply convergence in PDFs. . . . .	205
5.4	Sample mean $\bar{X}_n$ as a function of sample size $n$ for 20 replications from independent realizations of a fair die (blue), fair coin (magenta), Uniform( $0, 30$ ) RV (green) and Exponential( $0.1$ ) RV (red) with population means $(1+2+3+4+5+6)/6 = 21/6 = 3.5$ , $(0+1)/2 = 0.5$ , $(30-0)/2 = 15$ and $1/0.1 = 10$ , respectively. . . . .	210
5.5	Unending fluctuations of the running means based on $n$ IID samples from the Standard Cauchy RV $X$ in each of five replicate simulations (blue lines). The running means, based on $n$ IID samples from the Uniform( $0, 10$ ) RV, for each of five replicate simulations (magenta lines). . . . .	211
6.1	Transition Diagram of Flippant Freddy's Jumps. . . . .	220
6.2	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). . . . .	222
6.3	Transition Diagram of Dry and Wet Days in Christchurch. . . . .	224

7.1	Geometry of the $\Theta$ 's for de Moivre[ $k$ ] Experiments with $k \in \{1, 2, 3, 4\}$ . . . . .	242
7.2	Density and Confidence Interval of the Asymptotically Normal Point Estimator . . . . .	285
7.3	Plot of power function $\beta(\lambda)$ for different values of the critical value $c$ and the size $\alpha$ as function of the critical values. . . . .	290
7.4	The smallest $\alpha$ at which a size $\alpha$ test rejects the null hypothesis $H_0$ is the p-value. . . . .	292
7.5	Data Spaces $\mathbb{X}_1 = \{0, 1\}$ , $\mathbb{X}_2 = \{0, 1\}^2$ and $\mathbb{X}_3 = \{0, 1\}^3$ for one, two and three IID Bernoulli trials, respectively and the corresponding likelihood functions. . . . .	302
7.6	Plot of $\log(L(\lambda))$ as a function of the parameter $\lambda$ and the MLE $\hat{\lambda}_{132}$ of 0.1102 for Fenemore-Wang Orbiter Waiting Times Experiment from STAT 218 S2 2007. The density or PDF and the DF at the MLE of 0.1102 are compared with a histogram and the empirical DF. . . . .	306
7.7	Comparing the Exponential( $\hat{\lambda}_{6128} = 28.6694$ ) PDF and DF with a histogram and empirical DF of the times (in units of days) between earth quakes in NZ. The epicenters of 6128 earth quakes are shown in left panel. . . . .	307
7.8	Plots of the log likelihood $\ell_n(\theta) = \log(L(1, 0, 0, 0, 1, 1, 0, 0, 1, 0; \theta))$ as a function of the parameter $\theta$ over the parameter space $\Theta = [0, 1]$ and the MLE $\hat{\theta}_{10}$ of 0.4 for the coin-tossing experiment shown in standard scale (left panel) and log scale for $x$ -axis (right panel). . . . .	310
7.9	100 realizations of 95% confidence intervals based on samples of size $n = 10, 100$ and 1000 simulated from IID Bernoulli( $\theta^* = 0.5$ ) RVs. The MLE $\hat{\theta}_n$ (cyan dot) and the log-likelihood function (magenta curve) for each of the 100 replications of the experiment for each sample size $n$ are depicted. The approximate normal-based 95% confidence intervals with blue boundaries are based on the exact $\text{se}_n = \sqrt{\theta^*(1 - \theta^*)/n} = \sqrt{1/4}$ , while those with red boundaries are based on the estimated $\widehat{\text{se}}_n = \sqrt{\hat{\theta}_n(1 - \hat{\theta}_n)/n} = \sqrt{\bar{x}_n(1 - \bar{x}_n)/n}$ . The fraction of times the true parameter $\theta^* = 0.5$ was contained by the exact and approximate confidence interval (known as <i>empirical coverage</i> ) over the 100 replications of the simulation experiment for each of the three sample sizes are given by the numbers after <i>Cvrg.</i> = and $\sim$ =, above each sub-plot, respectively. . . . .	310
7.10	The ML fitted Rayleigh( $\hat{\alpha}_{10} = 2$ ) PDF and a histogram of the ocean wave heights. . . . .	314
7.11	Plot of $\log(L(\lambda))$ as a function of the parameter $\lambda$ , the MLE $\hat{\lambda}_{132} = 0.1102$ and 95% confidence interval $C_n = [0.0914, 0.1290]$ for Fenemore-Wang Orbiter Waiting Times Experiment from STAT 218 S2 2007. The PDF and the DF at (1) the MLE 0.1102 (black), (2) lower 95% confidence bound 0.0914 (red) and (3) upper 95% confidence bound 0.1290 (blue) are compared with a histogram and the empirical DF. . . . .	321
7.12	Plots of ten distinct ECDFs $\hat{F}_n$ based on 10 sets of $n$ IID samples from Uniform(0, 1) RV $X$ , as $n$ increases from 10 to 100 to 1000. The DF $F(x) = x$ over $[0, 1]$ is shown in red. The script of Labwork 240 was used to generate this plot. . . . .	327
7.13	The empirical DFs $\hat{F}_n^{(1)}$ from sample size $n = 10, 100, 1000$ (black), is the point estimate of the fixed and known DF $F(x) = x, x \in [0, 1]$ of Uniform(0, 1) RV (red). The 95% confidence band for each $\hat{F}_n$ are depicted by green lines. . . . .	329
7.14	The empirical DF $\hat{F}_{6128}$ for the inter earth quake times and the 95% confidence bands for the non-parametric experiment. . . . .	329
7.15	The empirical DF $\hat{F}_{132}$ for the Orbiter waiting times and the 95% confidence bands for the non-parametric experiment. . . . .	330

7.16 The empirical DFs $\hat{F}_{n_1}^{(1)}$ with $n_1 = 56485$ , for the web log times starting October 1, and $\hat{F}_{n_2}^{(2)}$ with $n_2 = 53966$ , for the web log times starting October 2. Their 95% confidence bands are indicated by the green. . . . .	331
7.17 Data from Bradley Efron's LSAT and GPA scores for fifteen individuals (left). The confidence interval of the sample correlation, the plug-in estimate of the population correlation, is obtained from the sample correlation of one thousand bootstrapped data sets (right). . . . .	337
9.1 Example of Needle Tosses . . . . .	393
9.2 Explaining the outcomes of the needle toss . . . . .	393
9.3 Outcome Space of Buffon's Needle Experiment for General Case . . . . .	394
9.4 Plot showing the midpoints mapping back to specific values on the x axis. . . . .	406
9.5 Plot showing both the GN and GDN CDFs. They are very similar. . . . .	407
9.6 Plot showing the empirical DF of our results against GDN. Our values take on a stair case appearance and are very close to GDN. The main deviations occur mostly in the tails. . . . .	408
9.7 Plot showing the Standard Normal Distribution against Galton's Normal Distribution.	409
9.8 From the graphs above, we can see that often, a short wait is followed by a long wait, in both directions. Also, the anticlockwise times are generally much closer to 10 minutes waiting time. It is also seen that around rush hour times (8:30, 15:00, 16:45), a pattern emerged where several buses in quick succession were followed by a long wait for the next bus to arrive. This could be because of the time taken for more passengers than usual to aboard and depart, and areas where traffic volume is greater at these times. . . . .	415
9.9 This graph shows the probability distribution function for the exponential function with the green line indicating a $\lambda$ value of 0.1, the claimed $\lambda$ . The red line indicates the value of $\lambda$ estimated, 0.1105. From this graph, you can see the probability of getting a short waiting time is high - approximately 0.06, while the probability of a long waiting time is much much lower - approximately 0.01. The Matlab code for this graph is shown in Appendix I. . . . .	416
9.10 This plot is the Empirical CDF plot(black), with a 95% confidence interval (red) and the actual CDF based on claimed $\lambda = 0.1$ (blue). The Matlab code for this graph is given in Appendix II. This graph shows the accuracy of the empirical distribution and hence the accuracy of the data we collected. There are some inconsistencies caused by the randomness of inter-arrival times but our empirical CDF is generally good as the actual CDF lies mostly within the interval lines. With more data points, our accuracy would greatly improve. . . . .	416

# 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  $\text{E}\mathbb{G}G$  of twenty two Greek lower-case alphabets that we may encounter later:

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

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:

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

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:

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

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

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

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

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

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

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

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

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:

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

**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:

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

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.:

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

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

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

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

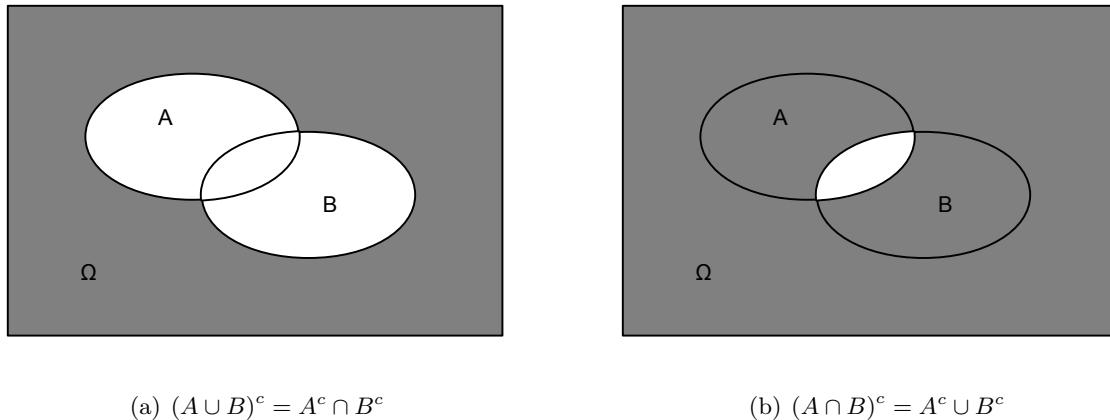


Figure 1.2: These Venn diagram illustrate De Morgan's Laws.

**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,

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

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

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

## 1.2 Exercises

**Ex. 1.1** — Let  $\Omega$  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.
- $\Omega$  — universal set.
- $A^c$  — the complement of  $A$ , meaning the set of all elements in  $\Omega$ , the universal set, which are not in  $A$ .

## 1.3 Natural Numbers, Integers and Rational Numbers

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

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

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  $E \otimes G$ ,  $\#A = 2$  and  $\#E \otimes G = 22$ . 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.

## 1.4 Real Numbers

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

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

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  $\pi$  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  $\infty$ ) 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 $\{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$
$\mathbb{Z}_+$	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

## 1.5 Introduction to MATLAB

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

## 1.6 Elementary Combinatorics

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

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

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

$$n_1 \times n_2 \times n_3 \times \dots$$

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

First letter: 26 possibilities

Fourth letter: 26 possibilities

Second letter: 26 possibilities

First digit: 10 possibilities

Third letter: 26 possibilities

Second digit: 10 possibilities

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

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

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

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

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

“I have 17 probability texts on my bottom shelf”

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

“The combination of my PIN is math99”

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

So in mathematics, we use more precise language:

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

**Permutations:** There are basically two types of permutations:

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

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

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

but since we want four letters only here, we have

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

choices.

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

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

**Combinations:** There are also two types of combinations:

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

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

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

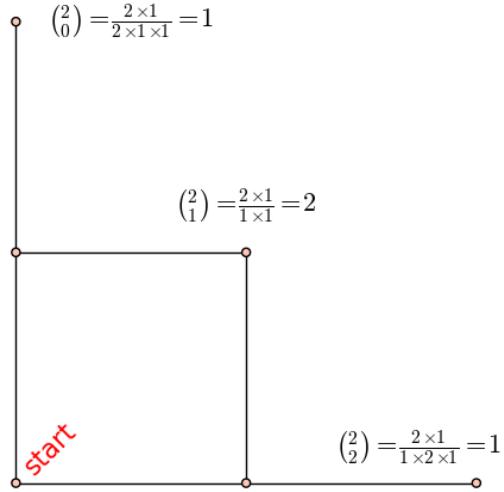
**Example 8** Let us imagine being in the lower Manhattan in New York city with its perpendicular grid of streets and avenues. If you start at a given intersection and are asked to only proceed in a north-easterly direction then how many ways are there to reach another intersection by walking exactly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

Solution:

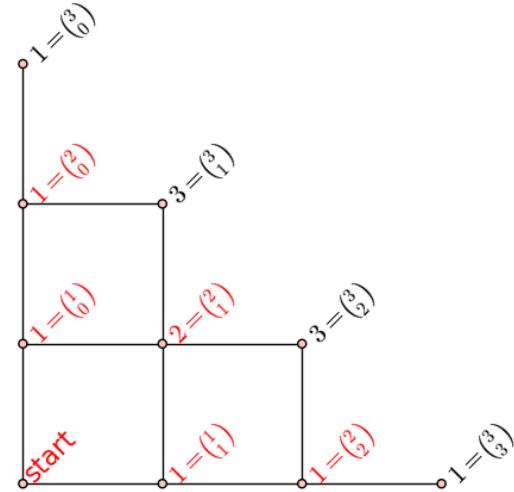
Let us answer this question of combinations by drawing Fig. 1.6. Let us denote the number of easterly turns you take by  $r$  and the total number of blocks you are allowed to walk either easterly or northerly by  $n$ . From Fig. 1.6(a) it is clear that the number of ways to reach each of the three intersections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

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

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



(a) Walking two blocks north-easterly.



(b) Walking three blocks north-easterly.

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

$$12, \quad 21,$$

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) :

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

**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:

$$\binom{n}{k} := \frac{p_{n,k}}{p_{k,k}} = \frac{n(n-1)(n-2)\dots(n-k+1)}{k(k-1)(k-2)\dots2\ 1}.$$

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 Chandadhśāstra*, due to Halāyudha (10-th century AD).

## 1.7 Array, Sequence, Limit, ...

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

For any natural number  $n$ , we write

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

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:

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

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:

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

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.

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

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

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

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

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

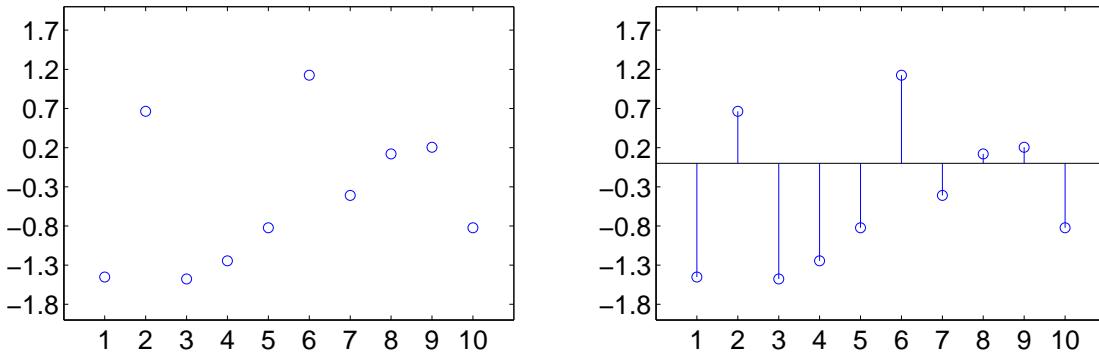
```

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

```

We may visualise (as per Figure 1.6) the finite sequences  $\langle b_{1:n} \rangle$  stored in the array **b** as the set of ordered pairs  $\{(1, b_1), (2, b_2), \dots, (10, b_{10})\}$  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.

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



```

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

```

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

```

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

```

```

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

```

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

```

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

```

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

```

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

```

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

Similarly, operations can be performed with matrices.

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

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

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

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

---

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

---

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

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

## 1.8 Elementary Real Analysis

### 1.8.1 Limits of Real Numbers – A Review

Let us first recall some elementary ideas from real analysis.

**Definition 3 (Convergent sequence of real numbers)** A sequence of real numbers  $\langle x_i \rangle_{i=1}^{\infty} := x_1, x_2, \dots$  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  $j \geq N_m$ ,  $|x_j - a| \leq \frac{1}{m}$ .

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

**Example 12 (Limit of a sequence of 17s)** Let  $\langle x_i \rangle_{i=1}^{\infty} = 17, 17, 17, \dots$ . 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 13 (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 14 (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  $\infty \notin \mathbb{R}$ .

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

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

the Euler's constant  $e$ :

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

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

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 16 (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 17 (Limit of  $(1 - \frac{\lambda}{n})^n$ )** The limit of  $f(n) = (1 - \frac{\lambda}{n})^n$  as  $n$  approaches  $\infty$  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 18 (Limit of  $(1 - \frac{\lambda}{n})^{-\alpha}$ )** The limit of  $f(n) = (1 - \frac{\lambda}{n})^{-\alpha}$ , for some  $\alpha > 0$ , as  $n$  approaches  $\infty$  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 19 (Discontinuity of  $f(x) = (1+x)^{\frac{1}{x}}$  at 0)** Let us reconsider the function  $f(x) = (1+x)^{\frac{1}{x}} : \mathbb{R} \rightarrow \mathbb{R}$ . 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  $\mathbb{R}$ .

## 1.9 Elementary Number Theory

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

For any real number  $x$ :

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

### Example 20 (Floors and ceilings)

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

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

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

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

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

Let us define modular arithmetic next. Suppose  $x$  and  $y$  are any real numbers, i.e.  $x, y \in \mathbb{R}$ , we define the binary operation called “ $x \bmod y$ ” 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  $\Omega$ .

The subsets of  $\Omega$  are called **events**. A single outcome,  $\omega$ , when seen as a subset of  $\Omega$ , 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 23** 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  $\Omega$  are  $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}$ , and  $\{6\}$ .

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

**Example 24** Consider a generic die-tossing experiment by a human experimenter. Here  $\Omega = \{\omega_1, \omega_2, \omega_3, \dots, \omega_6\}$ , 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 25** Some standard examples of a trial are:

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

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

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

**Example 26 (Toss a coin  $n$  times)** Suppose our experiment entails tossing a coin  $n$  times and recording 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  $\Omega$ :

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

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 27 (A thrice-bifurcating tree of outcomes)** Can you think of a graphical way to enumerate the outcomes of the Experiment 26? Draw a diagram of this under the caption of Figure 2.1, using the caption as a hint (in other words, draw your own Figure 2.1).

#### EXPERIMENT SUMMARY

Experiment	–	an activity producing distinct outcomes.
$\Omega$	–	set of all outcomes of the experiment.
$\omega$	–	an individual outcome in $\Omega$ , called a simple event.
$A \subseteq \Omega$	–	a subset $A$ of $\Omega$ 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  $P(A)$ , to intuitively mean the following: if one were to perform a super-experiment  $\mathcal{E}^\infty$  by independently repeating the experiment  $\mathcal{E}$  and recording  $N(A, n)$ , the fraction of times  $A$  occurs in the first  $n$  performances of  $\mathcal{E}$  within the super-experiment  $\mathcal{E}^\infty$ . Then the LTRF idea suggests:

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

Now, we are finally ready to define probability.

**Definition 10 (Probability)** Let  $\mathcal{E}$  be an experiment with sample space  $\Omega$ . Let  $\mathcal{F}$  denote a suitable collection of events in  $\Omega$  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  $\Omega$  and a collection of events  $\mathcal{F}$  that satisfy the three condition.

Given a double, e.g.  $(\Omega, \mathcal{F})$ , **probability** is just a function  $P$  which assigns each event  $A \in \mathcal{F}$  a number  $P(A)$  in the real interval  $[0, 1]$ , i.e.  $\boxed{P : \mathcal{F} \rightarrow [0, 1]}$ , such that:

1. The ‘Something Happens’ axiom holds, i.e.  $\boxed{P(\Omega) = 1}$ .
2. The ‘Addition Rule’ axiom holds, i.e. for events  $A$  and  $B$ :

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

### 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  $A$ ,  $\boxed{P(A^c) = 1 - P(A)}.$

**Proof:** One line proof.

$$\overbrace{P(A) + P(A^c)}^{LHS} \underset{\substack{+ \text{ rule } A \cap A^c = \emptyset}}{=} P(A \cup A^c) \underset{A \cup A^c = \Omega}{=} P(\Omega) \underset{P(\Omega) = 1}{=} \overbrace{1}^{RHS} \Rightarrow P(A^c) = 1 - P(A)$$

- If  $A = \Omega$  then  $A^c = \Omega^c = \emptyset$  and  $\boxed{P(\emptyset) = 1 - P(\Omega) = 1 - 1 = 0}.$

2. For any two events  $A$  and  $B$ , we have the **inclusion-exclusion principle**:

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}.$$

**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:

$$\begin{aligned} P(A) &= P(A \setminus B) + P(A \cap B) \\ P(A \cup B) &= P(A \setminus B) + P(B) \end{aligned}$$

Substituting the first equality above into the second, we get:

$$P(A \cup B) = P(A \setminus B) + P(B) = P(A) - P(A \cap B) + P(B)$$

3. From inclusion-exclusion principle we get **Boole’s inequality**: for any two events  $A, B$

$$P(A \cup B) \leq P(A) + P(B)$$

4. The inclusion-exclusion principle extends similarly to any three events  $A_1, A_2, A_3$  as follows:

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$$

and generalises to any  $n$  events  $A_1, A_2, \dots, A_n$  as follows:

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) + \dots + (-1)^{n-1} \sum_{i < \dots < n} P\left(\bigcap_{i=1}^n A_i\right)$$

**Proof:** See the counting argument in [https://en.wikipedia.org/wiki/Inclusion%20%93exclusion\\_principle](https://en.wikipedia.org/wiki/Inclusion%20%93exclusion_principle) if you are curious.

5. Once again by the inclusion-exclusion principle, the Boole’s inequality generalises to any  $n$  events  $A_1, A_2, \dots, A_n$  as follows:

$$P\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n P(A_i)$$

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

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

**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:

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

The  $n$ -event case follows by mathematical induction.

We have formally defined the **probability model** specified by the **probability triple**  $(\Omega, \mathcal{F}, P)$  that can be used to model an **experiment**  $\mathcal{E}$ .

**Example 28 (First Ball out of NZ Lotto)** Let us observe the number on *the first ball that pops out in a New Zealand Lotto trial*. There are forty balls labelled 1 through 40 for this experiment and so the sample space is

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

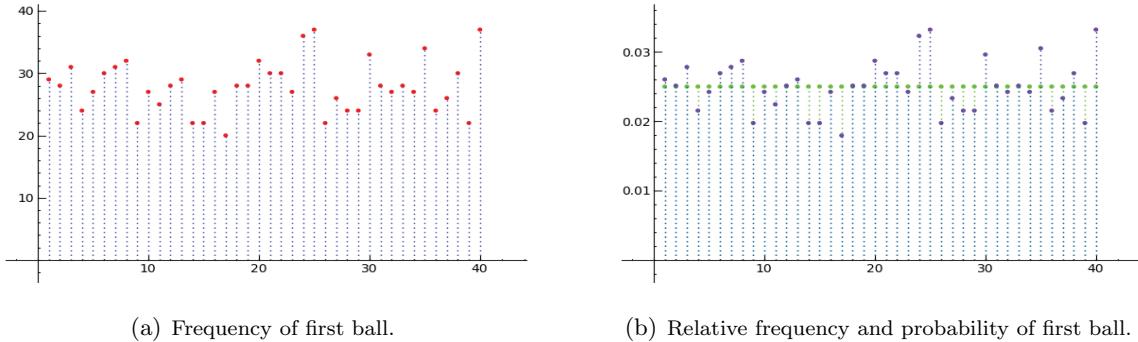
Because the balls are vigorously whirled around inside the Lotto machine, modelled as a well-stirred urn, before the first one pops out, we can model each ball to pop out first with the same probability. So, we assign each outcome  $\omega \in \Omega$  the same probability of  $\frac{1}{40}$ , i.e., our probability model for this experiment is:

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

Note: We sometimes abuse notation and write  $P(\omega)$  instead of the more accurate but cumbersome  $P(\{\omega\})$  when writing down probabilities of simple events.

Crucially, by  $\omega = 17$  for example, we mean all the detailed dynamics inside the Lotto machine that lead to the event that the ball labelled by the number 17 ends up popping out. So,  $\Omega$  here is indeed a more complicated set although it only leads to 40 possible outcomes.

Figure 2.2 (a) shows the frequency of the first ball number in 1114 NZ Lotto draws. Figure 2.2 (b) shows the relative frequency, i.e., the frequency divided by 1114, the number of draws. Figure 2.2 (b) also shows the equal probabilities under our model.



(a) Frequency of first ball.

(b) Relative frequency and probability of first ball.

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

Next, let us take a detour into how one might interpret it in the real world. The following is an adaptation from Williams D, *Weighing the Odds: A Course in Probability and Statistics*, Cambridge University Press, 2001, which henceforth is abbreviated as WD2001.

**Probability Model**

Sample space $\Omega$	Set of all outcomes of an experiment
Sample point $\omega$ (No counterpart)	Possible outcome of an experiment
Event A, a (suitable) subset of $\Omega$	Actual outcome $\omega^*$ of an experiment
$P(A)$ , a number between 0 and 1	The real-world event corresponding to A occurs if and only if $\omega^* \in A$

**Real-world Interpretation**

Set of all outcomes of an experiment
Possible outcome of an experiment
Actual outcome $\omega^*$ 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 $\Omega$	The certain even ‘something happens’
The $\emptyset$ of $\Omega$	The impossible event ‘nothing happens’
The intersection $A \cap B$	‘Both A and B occur’
$A_1 \cap A_2 \cap \dots \cap A_n$	‘All of the events $A_1, A_2, \dots, A_n$ occur simultaneously’
The union $A \cup B$	‘At least one of A and B occurs’
$A_1 \cup A_2 \cup \dots \cup A_n$	‘At least one of the events $A_1, A_2, \dots, A_n$ occurs’
$A^c$ , the complement of A	‘A does not occur’
$A \setminus B$	‘A occurs, but B does not occur’
$A \subset B$	‘If A occurs, then B must occur’

In the probability model of Example 28, show that for any event  $E \subset \Omega$ ,

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

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:

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

### 2.2.2 Sigma Algebras of Typical Experiments\*

**Example 29** (‘Toss a fair coin once’) Consider the ‘Toss a fair coin once’ experiment. What is its sample space  $\Omega$  and a reasonable collection of events  $\mathcal{F}$  that underpin this experiment?

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

A function that will satisfy the definition of probability for this collection of events  $\mathcal{F}$  and assign  $P(H) = \frac{1}{2}$  is summarized below. First check that the above  $\mathcal{F}$  is a sigma-algebra. Draw a picture for  $P$  with arrows that map elements in the domain  $\mathcal{F}$  given above to elements in its range.

Event $A \in \mathcal{F}$	$P : \mathcal{F} \rightarrow [0, 1]$	$P(A) \in [0, 1]$
$\Omega = \{H, T\} \bullet$	$\longrightarrow$	1
$T \bullet$	$\longrightarrow$	$1 - \frac{1}{2}$
$H \bullet$	$\longrightarrow$	$\frac{1}{2}$
$\emptyset \bullet$	$\longrightarrow$	0

**Classwork 30 (The trivial sigma algebra)** Note that  $\mathcal{F}' = \{\Omega, \emptyset\}$  is also a sigma algebra of the sample space  $\Omega = \{H, T\}$ . Can you think of a probability for the collection  $\mathcal{F}'$ ?

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

Thus,  $\mathcal{F}$  and  $\mathcal{F}'$  are two distinct sigma algebras over our  $\Omega = \{\text{H, T}\}$ . 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  $\Omega$  (think of intersecting an arbitrary family of sigma algebras)?

Generally one encounters four types of sigma algebras (you will understand the last two types after taking more advanced courses in mathematics, so it is fine to understand the ideas intuitively for now!) and they are:

1. When the sample space  $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$  is a finite set with  $k$  outcomes and  $P(\omega_i)$ , 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  $\Omega$  called the **power set** and denoted by  $2^\Omega$ . The probability of each event  $A \in 2^\Omega$  can be obtained by adding the probabilities of the outcomes in  $A$ , i.e.,  $P(A) = \sum_{\omega_i \in A} P(\omega_i)$ . Clearly,  $2^\Omega$  is indeed a sigma-algebra and it contains  $2^{\#\Omega}$  events in it.
2. 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  $\Omega$ . Note that this is very similar to the case with finite  $\Omega$  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  $\Omega$  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  $\mathbb{S}^\infty$  is the smallest sigma-algebra containing the cylinder sets.

- **A most primitive sigma-algebra for probability theory:** For example if  $\mathbb{S} = \{0, 1\}$ , then  $\Omega = \{0, 1\}^\infty$  is the set of all infinite sequences made of 0's and 1's. To take advantage of arithmetic and analysis,  $\Omega$  can be seen as the binary representation of all real numbers in the unit interval  $[0, 1]$ . We can take advantage of combinatorics and algebra if we further represent the dyadic partition of  $[0, 1]$  by a binary tree (as drawn in lectures). Then, a cylinder set such as  $1 \times 1 \times 0 \times \{0, 1\} \times \{0, 1\} \times \dots$ , an event here, can be interpreted as the finite binary sequence  $(1, 1, 0)$  — corresponding to the third leaf of a finite binary tree with four leaves obtained by splitting the right-most leaf twice. This cylindrical event  $(1, 1, 0)$  contains all real numbers in the interval  $[\frac{3}{4}, \frac{7}{8}] \subset [0, 1] =: \Omega$ .

**Exercise 2.1 (Intuiting a most primitive sigma-algebra – this is optional)** Try to carefully recollect and understand the most primitive sigma-algebra in the last item above as it was explained in lectures.

## PROBABILITY SUMMARY

Axioms:

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

Rules:

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

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

### 2.3 Exercises in Probability

**Ex. 2.2** — In English language text, the twenty six letters in the alphabet occur with the following frequencies:

E	13%	R	7.7%	A	7.3%	H	3.5%	F	2.8%	M	2.5%	W	1.6%	X	0.5%	J	0.2%
T	9.3%	O	7.4%	S	6.3%	L	3.5%	P	2.7%	Y	1.9%	V	1.3%	K	0.3%	Z	0.1%
N	7.8%	I	7.4%	D	4.4%	C	3%	U	2.7%	G	1.6%	B	0.9%	Q	0.3%		

Suppose you pick one letter at random from a randomly chosen English book from our central library with  $\Omega = \{A, B, C, \dots, Z\}$  (ignoring upper/lower cases), then what is the probability of these events?

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

**Ex. 2.3** — Find the sample spaces for the following experiments:

1. Tossing 2 coins whose faces are sprayed with black paint denoted by  $B$  and white paint denoted by  $W$ .
2. Drawing 4 screws from a bucket of left-handed and right-handed screws denoted by  $L$  and  $R$ , respectively.
3. Rolling a die and recording the number on the upturned face until the first 6 appears.

**Ex. 2.4** — Suppose we pick a letter at random from the word WAIMAKARIRI.

1. What is the sample space  $\Omega$ ?
2. What probabilities should be assigned to the outcomes?
3. What is the probability of *not* choosing the letter R?

**Ex. 2.5** — There are seventy five balls in total inside the Bingo Machine. Each ball is labelled by one of the following five letters: B, I, N, G, and O. There are fifteen balls labelled by each letter. The letter on the first ball that comes out of a BINGO machine after it has been well-mixed is the outcome of our experiment.

- (a) Write down the sample space of this experiment.
- (b) Find the probabilities of each simple event.
- (c) Show that  $P(\Omega)$  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  $P(C \cup D)$ .

## 2.4 Conditional Probability

Conditional probabilities arise when we have partial information about the result of an experiment which restricts the sample space to a range of outcomes. For example, if there has been a lot of recent seismic activity in Christchurch, then the probability that an already damaged building will collapse tomorrow is clearly higher than if there had been no recent seismic activity.

Conditional probabilities are often expressed in English by phrases such as:

“If  $A$  happens, what is the probability that  $B$  happens?”

or

“What is the probability that  $A$  happens if  $B$  happens?”

or

“What is the probability that  $A$  occurs given that  $B$  occurs?”

Next, we define conditional probability and the notion of independence of events. We use the LTRF idea to motivate the definition.

**Idea 11 (LTRF intuition for conditional probability)** Let  $A$  and  $B$  be any two events associated with our experiment  $\mathcal{E}$  with  $P(A) \neq 0$ . The ‘conditional probability that  $B$  occurs given that  $A$  occurs’ denoted by  $P(B|A)$  is again intuitively underpinned by the super-experiment  $\mathcal{E}^\infty$  which is the ‘independent’ repetition of our original experiment  $\mathcal{E}$  ‘infinitely’ often. The LTRF idea is that  $P(B|A)$  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

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

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  $(\Omega, \mathcal{F}, P)$ . Let  $A$  and  $B$  be events, ie.  $A, B \in \mathcal{F}$ , such that  $P(A) \neq 0$ . Then, we define the **conditional probability** of  $B$  given  $A$  by,

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

Note that  $A$  serves as the new reduced sample space so that conditional probabilities given  $A$  are indeed probabilities. Thus, for a **fixed** event  $A \in \mathcal{F}$  with  $P(A) > 0$  and **any** event  $B \in \mathcal{F}$ , the conditional probability  $P(B|A)$  is a probability as in Definition 10, ie. a function:

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

that assigns to each  $B \in \mathcal{F}$  a number in the interval  $[0, 1]$ , such that, the axioms of probability are satisfied:

Axiom (1): For any event  $B$ ,  $0 \leq P(B|A) \leq 1$ .

Axiom (2):  $P(\Omega|A) = 1$       Meaning ‘Something Happens given the event A happens’

Axiom (3): The ‘Addition Rule’ axiom holds, ie. for events  $B_1, B_2 \in \mathcal{F}$ ,

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

Axiom (4): For mutually exclusive events,  $B_1, B_2, \dots$ ,

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

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

**Complementation rule:**  $P(B|A) = 1 - P(B^c|A)$  .

**Addition rule for two arbitrary events  $B_1$  and  $B_2$ :**

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

Solving for  $P(A \cap B)$  with these definitions of conditional probability gives another rule:

**Multiplication rule for two likely events:**

If  $A$  and  $B$  are events, and if  $P(A) \neq 0$  and  $P(B) \neq 0$ , then

$$P(A \cap B) = P(A) P(B|A) = P(B) P(A|B) .$$

**Example 31 (Wasserman03, p. 11)** A medical test for a disease  $D$  has outcomes + and -. the probabilities are:

	Have Disease ( $D$ )	Don't have disease ( $D^c$ )
Test positive (+)	0.009	0.099
Test negative (-)	0.001	0.891

Using the definition of conditional probability, we can compute the conditional probability that you test positive given that you have the disease:

$$P(+|D) = \frac{P(+ \cap D)}{P(D)} = \frac{0.009}{0.009 + 0.001} = 0.9 ,$$

and the conditional probability that you test negative given that you don't have the disease:

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

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 ?

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

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.

### 2.4.1 Bayes' Theorem

Next we look at one of the most elegant applications of the definition of conditional probability along with the addition rule for a partition of  $\Omega$  called *Bayes' Theorem*. We will present a two event case first called *Bayes' Rule* and then present the more general case of the Theorem.

This is useful because many problems involve reversing the order of conditional probabilities. Suppose we want to investigate some phenomenon  $A$  and have an observation  $B$  that is evidence about  $A$ : for example,  $A$  may be breast cancer and  $B$  may be a positive mammogram. Then Bayes' Theorem tells us how we should update our probability of  $A$ , given the new evidence  $B$ .

Or, put more simply, Bayes' Rule is useful when you know  $P(B|A)$  but want  $P(A|B)$ !

**Proposition 13 (Bayes' Rule)**

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} . \quad (2.3)$$

**Proof:** From the definition of conditional probability and the multiplication rule for two likely events  $A$  and  $B$  we get:

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

**Example 32 (Mammogram)** Approximately 1% of women aged 40–50 have breast cancer. A woman with breast cancer has a 90% chance of a positive test from a mammogram, while a woman without breast cancer has a 10% chance of a false positive result from the test. What is the probability that a woman indeed has breast cancer given that she just had a positive test?

Solution:

Let  $A$  = “the woman has breast cancer”, and  $B$  = “a positive test.”

We want  $P(A|B)$  but what we are given is  $P(B|A) = 0.9$ .

By the definition of conditional probability,

$$P(A|B) = P(A \cap B)/P(B)$$

To evaluate the numerator we use the multiplication rule

$$P(A \cap B) = P(A)P(B|A) = 0.01 \times 0.9 = 0.009$$

Similarly,

$$P(A^c \cap B) = P(A^c)P(B|A^c) = 0.99 \times 0.1 = 0.099$$

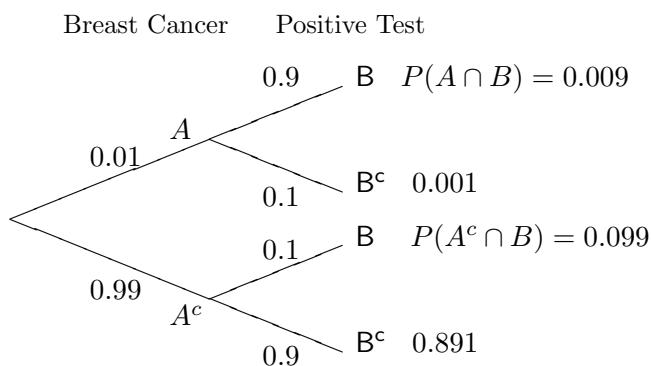
Now  $P(B) = P(A \cap B) + P(A^c \cap B)$  so

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.009}{0.009 + 0.099} = \frac{9}{108}$$

or a little less than 9%. This situation comes about because it is much easier to have a false positive for a healthy woman, which has probability 0.099, than to find a woman with breast cancer having a positive test, which has probability 0.009.

This answer is somewhat surprising. Indeed when ninety-five physicians were asked this question their average answer was 75%. The two statisticians who carried out this survey indicated that physicians were better able to see the answer when the data was presented in frequency format. 10 out of 1000 women have breast cancer. Of these 9 will have a positive mammogram. However of the remaining 990 women without breast cancer 99 will have a positive reaction, and again we arrive at the answer  $9/(9 + 99)$ .

*Alternative solution using a tree diagram:*



So the probability that a woman has breast cancer given that she has just had a positive test is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.009}{0.009 + 0.099} = \frac{9}{108}$$

\*In the exam, there won't be any need for electronic calculators and you may leave the answer in either of the last two numerical forms for full credit, provided you show the steps in your reasoning.

Before we see the more general form of Bayes' Rule, let us make a simple observation called the *total probability theorem*.

**Proposition 14 (Total probability theorem)** Suppose  $A_1 \cup A_2 \dots \cup A_k$  is a sequence of events with positive probability that partition the sample space, that is,  $A_1 \cup A_2 \dots \cup A_k = \Omega$  and  $A_i \cap A_j = \emptyset$  for any  $i \neq j$ , then for some arbitrary event  $B$ .

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

**Proof:** The first equality is due to the addition rule for mutually exclusive events,

$$B \cap A_1, B \cap A_2, \dots, B \cap A_k$$

and the second equality is due to the multiplication rule for two likely events.

Reference to the Venn diagram (you should draw below as done in lectures) will help you understand this idea for the four event case.

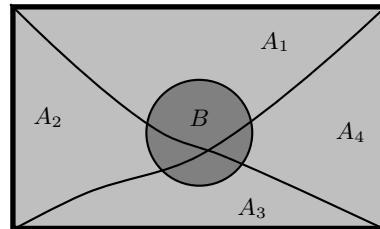
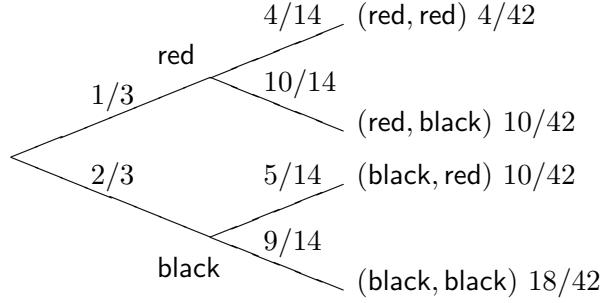


Figure 2.3: Reference to the Venn diagram will help you understand this idea behind the proof of the total probability theorem in Proposition 14 for the four event case.

**Example 33 (Urn with red and black balls)** A well-mixed urn contains five red and ten black balls. We draw two balls from the urn without replacement. What is the probability that the second ball drawn is red?

This is easy to see if we draw a probability tree diagram. The first split in the tree is based on the outcome of the first draw and the second on the outcome of the last draw. The outcome of the first draw dictates the probabilities for the second one since we are sampling without replacement. We multiply the probabilities on the edges to get probabilities of the four endpoints, and then sum the ones that correspond to red in the second draw, that is

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



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  $\Omega$  so we can write:

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

**Proposition 15 (Bayes' Theorem, 1763)** Suppose the events  $A_1, A_2, \dots, A_k \in \mathcal{F}$ , with  $P(A_h) > 0$  for each  $h \in \{1, 2, \dots, k\}$ , partition the sample space  $\Omega$ , ie. they are mutually exclusive (disjoint) and exhaustive events with positive probability:

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

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  $P(B) > 0$ , then

$$P(A_h|B) = \frac{P(B|A_h)P(A_h)}{\sum_{h=1}^k P(B|A_h)P(A_h)} \quad (2.5)$$

**Proof:** We apply elementary set theory, the definition of conditional probability  $k+2$  times and the addition rule once:

$$\begin{aligned}
 P(A_h|B) &= \frac{P(A_h \cap B)}{P(B)} = \frac{P(B \cap A_h)}{P(B)} = \frac{P(B|A_h)P(A_h)}{P(B)} \\
 &= \frac{P(B|A_h)P(A_h)}{P\left(\bigcup_{h=1}^k (B \cap A_h)\right)} = \frac{P(B|A_h)P(A_h)}{\sum_{h=1}^k P(B \cap A_h)} \\
 &= \frac{P(B|A_h)P(A_h)}{\sum_{h=1}^k P(B|A_h)P(A_h)}
 \end{aligned}$$

The operations done to the denominator in the proof above is merely the total probability theorem:

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

We call  $P(A_h)$  the **prior probability** of  $A_h$ , i.e., before observing  $B$  or *a priori*, and  $P(A_h|B)$  the **posterior probability** of  $A_h$ , i.e., after observing  $B$  or *a posteriori*.

This theorem is at the heart of solving Bayesian *Decision Problems* which fall into several sub-problems called *inference*, *learning* and *control* problems. Let's see one of the simplest such *learning problems* called *prediction*, more specifically *classification*, where we need to choose between finitely many possible choices based on past information next.

**Example 34 (Wasserman2003 p.12)** Suppose Larry divides his email into three categories:  $A_1$  = “spam”,  $A_2$  = “low priority”, and  $A_3$  = “high priority”. From previous experience, he finds that  $P(A_1) = 0.7$ ,  $P(A_2) = 0.2$  and  $P(A_3) = 0.1$ . Note that  $P(A_1 \cup A_2 \cup A_3) = P(\Omega) = 0.7 + 0.2 + 0.1 = 1$ . Let  $B$  be the event that the email contains the word “free.” From previous experience,  $P(B|A_1) = 0.9$ ,  $P(B|A_2) = 0.01$  and  $P(B|A_3) = 0.01$ . Note that  $P(B|A_1) + P(B|A_2) + P(B|A_3) = 0.9 + 0.01 + 0.01 \neq 1$ . Now, suppose Larry receives an email with the word “free.” What is the probability that it is “spam,” “low priority,” and “high priority” ? Solution:

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1) P(A_1)}{P(B|A_1) P(A_1) + P(B|A_2) P(A_2) + P(B|A_3) P(A_3)} = \frac{0.9 \times 0.7}{(0.9 \times 0.7) + (0.01 \times 0.2) + (0.01 \times 0.1)} = \frac{0.63}{0.633} \approx 0.995 \\ P(A_2|B) &= \frac{P(B|A_2) P(A_2)}{P(B|A_1) P(A_1) + P(B|A_2) P(A_2) + P(B|A_3) P(A_3)} = \frac{0.01 \times 0.2}{(0.9 \times 0.7) + (0.01 \times 0.2) + (0.01 \times 0.1)} = \frac{0.002}{0.633} \approx 0.003 \\ P(A_3|B) &= \frac{P(B|A_3) P(A_3)}{P(B|A_1) P(A_1) + P(B|A_2) P(A_2) + P(B|A_3) P(A_3)} = \frac{0.01 \times 0.1}{(0.9 \times 0.7) + (0.01 \times 0.2) + (0.01 \times 0.1)} = \frac{0.001}{0.633} \approx 0.002 \end{aligned}$$

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

This is essentially the idea behind *Bayes classifiers*, that are used to solve such *prediction* problems across different problem domains in *statistical machine learning*, where solutions are given from computer programs.

### 2.4.2 Independence and Dependence

In general,  $P(A|B)$  and  $P(A)$  are different, but sometimes the occurrence of  $B$  makes no difference, and gives no new information about the chances of  $A$  occurring. This is the idea behind independence. Events like “having blue eyes” and “having blond hair” are associated due to common genetic ancestry, but events like “my neighbour wins Lotto” and “I win Lotto” are not due to the Lotto machine being chaotically whirled around before ejection (as modelled by a well-stirred urn).

**Definition 16 (Independence of two events)** Any two events  $A$  and  $B$  are said to be **independent** if and only if

$$P(A \cap B) = P(A) P(B) . \quad (2.6)$$

Let us make sense of this definition in terms of our previous definitions. When  $P(A) = 0$  or  $P(B) = 0$ , both sides of the above equality are 0. If  $P(A) \neq 0$ , then rearranging the above equation we get:

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

But, the LHS is  $P(B|A)$  by definition 2.2, and thus for independent events  $A$  and  $B$ , we get:

$$P(B|A) = P(B) .$$

This says that information about the occurrence of  $A$  does not affect the occurrence of  $B$ . If  $P(B) \neq 0$ , then an analogous argument:

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

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

**Definition 17 (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

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

**Example 35 (Some Standard Examples)** A sequence of events in a sequence of independent trials is independent.

- (a) Suppose you toss a fair coin twice such that the first toss is independent of the second. Then,

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

- (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:

$$\begin{aligned} P(E_1 \cap E_2 \cap E_3) &= P(E_1)P(E_2)P(E_3) \\ &= (P(\{2, 4, 6\}))^3 \\ &= (P(\{2\} \cup \{4\} \cup \{6\}))^3 \\ &= (P(\{2\}) + P(\{4\}) + P(\{6\}))^3 \\ &= \left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right)^3 = \left(\frac{1}{2}\right)^3 = \frac{1}{8} . \end{aligned}$$

- (c) Suppose you toss a fair coin independently  $m$  times. Then each of the  $2^m$  possible outcomes in the sample space  $\Omega$  has equal probability of  $\frac{1}{2^m}$  due to independence.

**Example 36 (dependence and independence)** Suppose we toss two fair dice. Let  $A$  denote the event that the sum of the dice is six and  $B$  denote the event that the first die equals four. The sample space encoding the thirty six ordered pairs of outcomes for the two dice is  $\Omega = \{(1, 1), (1, 2), \dots, (1, 6), (2, 1), \dots, (2, 6), \dots, (5, 6), (6, 6)\}$  and due to independence  $P(\omega) = 1/36$  for each  $\omega \in \Omega$ . Then

$$P(A \cap B) = P(\{(4, 2)\}) = \frac{1}{36} ,$$

but

$$\begin{aligned} P(A)P(B) &= P(\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\})P(\{(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6)\}) \\ &= \frac{5}{36} \times \frac{6}{36} = \frac{5}{36} \times \frac{1}{6} = \frac{5}{216} , \end{aligned}$$

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

$$P(C \cap B) = P(\{(4, 3)\}) = \frac{1}{36},$$

while

$$\begin{aligned} P(C \cap B) &= P(\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}) P(\{(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6)\}) \\ &= \frac{6}{36} \times \frac{6}{36} = \frac{1}{36}, \end{aligned}$$

and therefore  $C$  and  $B$  are independent events. Once again this is clear because the chance of getting a total of seven does not depend any more on the outcome of the first die (it is allowed to be any one of the six possible outcomes).

**Example 37 (Pairwise independent events that are not jointly independent)** Let a ball be drawn from an well-stirred urn containing four balls labelled 1,2,3,4. Consider the events  $A = \{1, 2\}$ ,  $B = \{1, 3\}$  and  $C = \{1, 4\}$ . Then,

$$\begin{aligned} P(A \cap B) &= P(A) P(B) = \frac{2}{4} \times \frac{2}{4} = \frac{1}{4}, \\ P(A \cap C) &= P(A) P(C) = \frac{2}{4} \times \frac{2}{4} = \frac{1}{4}, \\ P(B \cap C) &= P(B) P(C) = \frac{2}{4} \times \frac{2}{4} = \frac{1}{4}, \end{aligned}$$

but,

$$\frac{1}{4} = P(\{1\}) = P(A \cap B \cap C) \neq P(A) P(B) P(C) = \frac{2}{4} \times \frac{2}{4} \times \frac{2}{4} = \frac{1}{8}.$$

Therefore, inspite of being pairwise independent, the events  $A$ ,  $B$  and  $C$  are not jointly independent.

#### CONDITIONAL PROBABILITY SUMMARY

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

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

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

Conditional probabilities obey the axioms and rules of probability.

## 2.5 Exercises in Conditional Probability

**Ex. 2.6 —** What gives the greater probability of hitting some target at least once:

- 1.hitting in a shot with probability  $\frac{1}{2}$  and firing 1 shot, or
- 2.hitting in a shot with probability  $\frac{1}{3}$  and firing 2 shots?

First guess. Then calculate.

**Ex. 2.7** — Suppose we independently roll two fair dice each of whose faces are marked by numbers 1,2,3,4, 5 and 6.

- 1.List the sample space for the experiment if we note the numbers on the 2 upturned faces.
- 2.What is the probability of obtaining a sum greater than 4 but less than 7?

**Ex. 2.8** — Based on past experience, 70% of students in a certain course pass the midterm test. The final exam is passed by 80% of those who passed the midterm test, but only by 40% of those who fail the midterm test. What fraction of students pass the final exam?

**Ex. 2.9** — A small brewery has two bottling machines. Machine 1 produces 75% of the bottles and machine 2 produces 25%. One out of every 20 bottles filled by machine 1 is rejected for some reason, while one out of every 30 bottles filled by machine 2 is rejected. What is the probability that a randomly selected bottle comes from machine 1 given that it is accepted?

**Ex. 2.10** — A process producing micro-chips, produces 5% defective, at random. Each micro-chip is tested, and the test will correctly detect a defective one  $\frac{4}{5}$  of the time, and if a good micro-chip is tested the test will declare it is defective with probability  $\frac{1}{10}$ .

- (a)If a micro-chip is chosen at random, and tested to be good, what was the probability that it was defective anyway?
- (b)If a micro-chip is chosen at random, and tested to be defective, what was the probability that it was good anyway?
- (c)If 2 micro-chips are tested and determined to be good, what is the probability that at least one is in fact defective?

**Ex. 2.11** — Suppose that  $\frac{2}{3}$  of all gales are force 1,  $\frac{1}{4}$  are force 2 and  $\frac{1}{12}$  are force 3. Furthermore, the probability that force 1 gales cause damage is  $\frac{1}{4}$ , the probability that force 2 gales cause damage is  $\frac{2}{3}$  and the probability that force 3 gales cause damage is  $\frac{5}{6}$ .

- (a)If a gale is reported, what is the probability of it causing damage?
- (b)If the gale caused damage, find the probabilities that it was of: force 1; force 2; force 3.
- (c)If the gale did NOT cause damage, find the probabilities that it was of: force 1; force 2; force 3.

**Ex. 2.12** — \*\*The sensitivity and specificity of a medical diagnostic test for a disease are defined as follows:

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

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

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

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. 2.14 —** \*\*Let  $A$  and  $B$  be events such that  $P(A) \neq 0$  and  $P(B) \neq 0$ . When  $A$  and  $B$  are disjoint, are they also independent? Explain clearly why or why not.

# Chapter 3

## Random Variables

We are used to classical variables such as  $x$  as an “unknown” in the equation:  $x + 3 = 7$ .

We also use classical variables to represent geometric objects such as a line:

$$y = 3x - 2,$$

where the variable  $y$  for the  $y$ -axis is determined by the value taken by the variable  $x$ , as  $x$  varies over the real line  $\mathbb{R} = (-\infty, \infty)$ .

Yet another example is the use of variables to represent sequences such as:

$$\{a_n\}_{n=1}^{\infty} = a_1, a_2, a_3, \dots .$$

What these *classical variables* have in common is that they *take a fixed or deterministic value* when we can solve for them.

We need a different kind of variable to deal with real-world situations where the same variable may take different values in a non-deterministic manner. **Random variables** do this job for us. Random variables, unlike classical deterministic variables, can take a bunch of different values.

Crucially, it can become inconvenient to work with a set of outcomes  $\Omega$  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	$\mathbb{R}$

Thus, we want a **random variable** to be a function from the sample space  $\Omega$  to the set of real numbers  $\mathbb{R}$ , that is,  $X : \Omega \rightarrow \mathbb{R}$  that should satisfy certain conditions to keep the meaning of the underlying probability space  $(\Omega, \mathcal{F}, P)$ . Let us go through some examples before giving the formal definition of such a real-valued or  $\mathbb{R}$ -valued random variable.

**Example 38 (Rain or Shine)** Suppose our experiment is to observe whether it will rain or not rain tomorrow. The sample space of this experiment is  $\Omega = \{\text{rain, not rain}\}$ . We can associate a random variable  $X$  with this experiment as follows:

$$X(\omega) = \begin{cases} 1, & \text{if } \omega = \text{rain} \\ 0, & \text{if } \omega = \text{not rain} \end{cases}$$

Thus,  $X$  will take the value 1 if it will rain tomorrow and 0 otherwise. Note that another equally valid (though possibly not so useful) random variable, say  $Y$ , for this experiment is:

$$Y(\omega) = \begin{cases} \pi, & \text{if } \omega = \text{rain} \\ \sqrt{2}, & \text{if } \omega = \text{not rain} \end{cases}$$

**Example 39 (Rain Fall on Angstrom)** Suppose our experiment instead is to measure the volume of rain that falls into a large funnel stuck on top of a graduated cylinder that is placed on top of the middle of House 1 of Angstrom Laboratory. Suppose the cylinder is graduated in millimeters then our random variable  $X(\omega)$  can report a non-negative real number given by the lower miniscus of the water column, if any, in the cylinder tomorrow. Thus,  $X(\omega)$  will measure the volume of rain in millilitres that will fall into our funnel tomorrow.

**Example 40 (Counting Seedlings)** Suppose ten seeds are planted. Perhaps fewer than ten will actually germinate. The number which do germinate, say  $X$ , must be one of the integer numbers in  $\mathbb{R}$  given by the set:

$$\mathbb{X} := \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} .$$

But until the seeds are actually planted and allowed to germinate it is impossible to say which number  $X(\omega) : \Omega \rightarrow \mathbb{X}$  will take. The number of seeds which germinate is a variable, but it is not necessarily the same for each group of ten seeds planted, but takes values from the same set  $\mathbb{X}$ . As  $X$  is not known in advance it is called a **random variable**. Its value cannot be known until we actually perform the experiment, i.e., plant the seeds.

Certain things can be said about the value a random variable might take. In the case of these ten seeds we can be sure the number that germinate is less than eleven, and not less than zero! It may also be known that that the probability of seven seeds germinating is greater than the probability of one seed; or perhaps that the number of seeds germinating averages eight. These statements are based on probabilities unlike the sort of statements made about deterministic variables.

### Discrete versus continuous random variables.

A **discrete** random variable is one in which the set of possible values of the random variable is finite or at most countably infinite, whereas a **continuous** random variable may take on any value in some range, and its value may be any real value in that range (Think: uncountably infinite). Examples 38 and 40 are about discrete random variables and Example 39 is about a continuous random variable.

Discrete random variables are usually generated from experiments where things are “counted” rather than “measured” such as the seed planting experiment in Example 40. Continuous random variables appear in experiments in which we measure, such as the amount of rain, in millilitres in Example 39.

### Random variables as functions.

In fact, random variables are actually functions, more formally measurable maps from  $\mathcal{F}$  to certain subsets of  $\mathbb{R}$  that you will learn carefully in more advanced courses. They take you from the “world of random processes and phenomena” to the world of real numbers. In other words, a random variable is a numerical value determined by the outcome of the experiment.

We said that a random variable can take one of many values, but we cannot be certain of which value it will take. However, *we can make probabilistic statements about the value  $x$  the random variable  $X$  will take.* A question like,

“What is the probability of it raining tomorrow?”

in the rain/not experiment of Example 38 becomes

“What is  $P(\{\omega : X(\omega) = 1\})$ ?”

or, more simply,

“What is  $P(X = 1)$ ?”

With this motivation we are ready to formally define such a random variable.

### 3.1 Basic Definitions

To take advantage of our measurements over the real numbers, in terms of its metric structure and arithmetic, we need to formally define this measurement process using the notion of a random variable.

**Definition 18 (Random Variable)** Let  $(\Omega, \mathcal{F}, P)$  be some probability triple. Then, a **Random Variable (RV)**, say  $X$ , is a function from the sample space  $\Omega$  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:

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

**Definition 19 (Distribution Function)** The **Distribution Function (DF)** or **Cumulative Distribution Function (CDF)** of any RV  $X$ , over a probability triple  $(\Omega, \mathcal{F}, P)$ , denoted by  $F$  is:

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

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  $X \sim F$ .

**Remark 20 (Notation)** It is enough to understand the idea of random variables as explained above, and work with random variables using simplified notation like

$$P(2 \leq X \leq 3)$$

rather than

$$P(\{\omega : 2 \leq X(\omega) \leq 3\})$$

but note that when learning or doing more advanced work this sample space notation is usually needed to clarify the true meaning of the simplified notation at least to yourself! But in the exam you can use the simpler notation as done in the solutions to exercises.

From the idea of a distribution function, we get:

**Proposition 21** The probability that the random variable  $X$  takes a value  $x$  in the half-open interval  $(a, b]$ , i.e.,  $a < x \leq b$ , is:

$$P(a < X \leq b) = F(b) - F(a) . \quad (3.3)$$

**Proof:** Since  $(X \leq a)$  and  $(a < X \leq b)$  are disjoint events whose union is the event  $(X \leq b)$ ,

$$F(b) = P(X \leq b) = P(X \leq a) + P(a < X \leq b) = F(a) + P(a < X \leq b) .$$

Subtraction of  $F(a)$  from both sides of the above equation yields Equation 3.3.

A special RV that often plays the role of ‘building-block’ in Probability and Statistics is the indicator function of an event  $A$  that tells us whether the event  $A$  has occurred or not. Recall that an event belongs to the collection of possible events  $\mathcal{F}$  for our experiment.

**Definition 22 (Indicator Function)** Given a probability triple  $(\Omega, \mathcal{F}, P)$ , the **Indicator Function** of an event  $A \in \mathcal{F}$  which is denoted  $\mathbb{1}_A$  is defined as follows:

$$\mathbb{1}_A(\omega) := \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} \quad (3.4)$$

**Model 1 (Indicator of an event as Bernoulli RV)** This is the most primitive RV from which all others are obtained. Let us convince ourselves that  $\mathbb{1}_A$  is really a RV. For  $\mathbb{1}_A$  to be a RV, we need to verify that for any real number  $x \in \mathbb{R}$ , the inverse image  $\mathbb{1}_A^{[-1]}( (-\infty, x] )$  is an event, ie :

$$\mathbb{1}_A^{[-1]}( (-\infty, x] ) := \{\omega : \mathbb{1}_A(\omega) \leq x\} \in \mathcal{F} .$$

All we can assume about the collection of events  $\mathcal{F}$  is that it contains the event  $A$  and that it is a sigma algebra. A careful look at the Figure 3.1 yields:

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

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)  $\Omega$  depending on the value taken by  $x$  relative to the interval  $[0, 1]$ . We have proved that  $\mathbb{1}_A$  is indeed a RV.

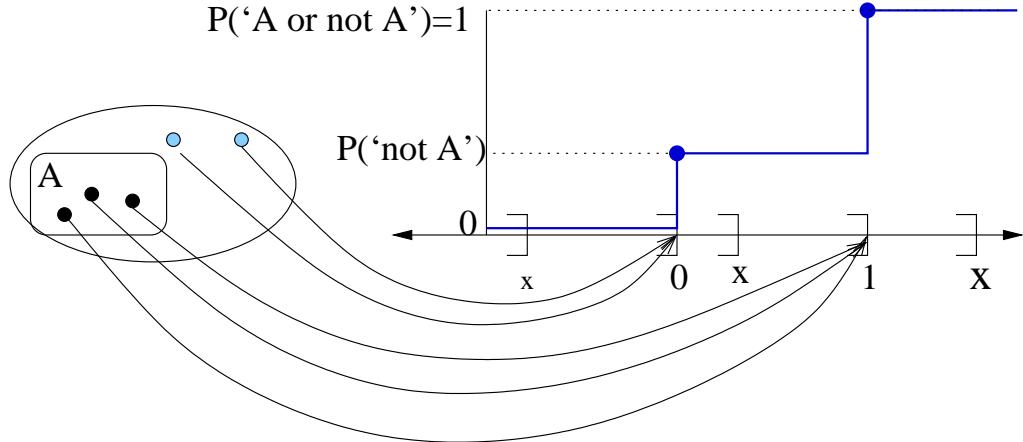
Model 1 is called the Bernoulli RV for event  $A$  with a known probability  $P(A)$ . We will define as our next model the Bernoulli( $\theta$ ) RV by introducing a parameter  $\theta \in [0, 1]$  for the typically unknown probability  $P(A)$ .

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.

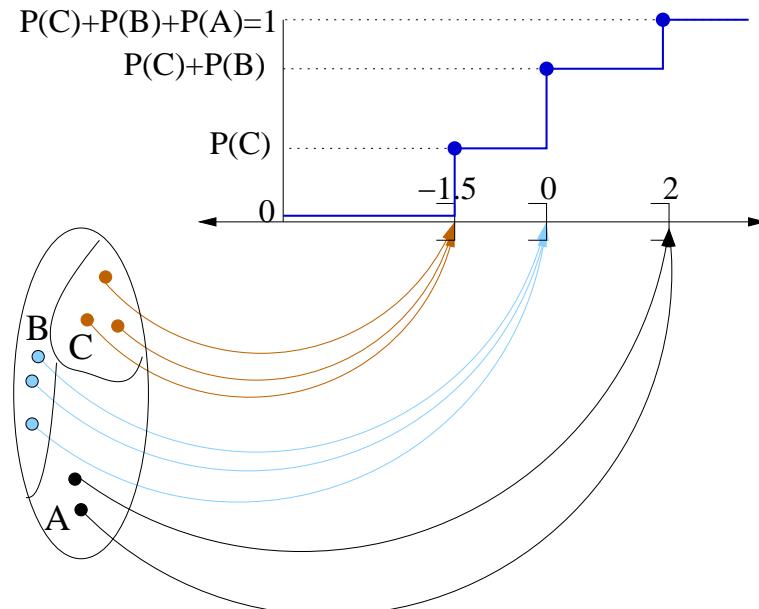
Figure 3.1: The Indicator function of event  $A \in \mathcal{F}$  is a RV  $\mathbb{1}_A$  with DF  $F$   
DF



**Exercise 3.1 (Drawing discontinuous functions)** Identify the mistakes in how the  $\mathbb{1}_A$  is drawn as a discontinuous function in Figure 3.1.

**Classwork 41 (A random variable with three values and eight sample points)** Consider the RV  $X$  of Figure 3.2. First draw this properly as done in Ex. 3.1. Let the events  $A = \{\omega_1, \omega_2\}$ ,  $B = \{\omega_3, \omega_4, \omega_5\}$  and  $C = \{\omega_6, \omega_7, \omega_8\}$ . Define the RV  $X$  formally. What sets should  $\mathcal{F}$  minimally include? What do you need to do to make sure that  $\mathcal{F}$  is a sigma algebra?

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



**Exercise 3.2 (Fair coin toss RV)** Consider the *fair coin toss experiment* with  $\Omega = \{\text{H}, \text{T}\}$  and  $P(\text{H}) = P(\text{T}) = 1/2$ .

We can associate a Bernoulli random variable  $X$  (in Model 1) for the event that the coin lands as H, with this experiment as follows:

$$X(\omega) = \begin{cases} 1, & \text{if } \omega = \text{H} \\ 0, & \text{if } \omega = \text{T} \end{cases}$$

Find the distribution function for  $X$ .

## 3.2 Discrete Random Variables

When a RV takes at most countably many values from a discrete set  $\mathbb{X}$ , we call it a **discrete** RV. Recall that a set  $\mathbb{X}$  is said to be discrete if we can enumerate its elements, i.e., find an enumerating or counting function  $\mathbb{X} \ni x \mapsto i \in \mathbb{N}$  that associates each element  $x \in \mathbb{X}$  to a natural number  $i \in \mathbb{N}$ . So,  $\mathbb{X}$  is either finite with  $k$  elements in  $\mathbb{X} = \{x_1, x_2, \dots, x_k\}$  or countably infinite with the same cardinality as  $\mathbb{N}$  with  $\mathbb{X} = \{x_1, x_2, \dots\}$ . When  $\mathbb{X} \subset \mathbb{R}$ , we have a real-valued or  $\mathbb{R}$ -valued discrete random variable.

**Definition 23 (probability mass function (PMF))** Let  $X$  be a  $\mathbb{R}$ -valued discrete RV over a probability triple  $(\Omega, \mathcal{F}, P)$ . We define the **probability mass function** (PMF)  $f$  of  $X$  to be the function  $f : \mathbb{R} \rightarrow [0, 1]$  defined as follows:

$$f(x) := P(X = x) = P(\{\omega : X(\omega) = x\}) = \begin{cases} \theta_i & \text{if } x = x_i \in \mathbb{X}. \\ 0 & \text{otherwise.} \end{cases} \quad (3.5)$$

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

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

$$P(X \leq x) = F(x) = \sum_{x_i \leq x} f(x_i) = \sum_{x_i \leq x} \theta_i. \quad (3.6)$$

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

$$P(a < X \leq b) = F(b) - F(a) = \sum_{a < x_i \leq b} \theta_i. \quad (3.7)$$

This is just the sum of all probabilities  $\theta_i$  for which  $x_i$  satisfies  $a < x_i \leq b$ .

3. From the fact that  $P(\Omega) = 1$ , we get that the sum of all the probabilities is 1:

$$\sum_i \theta_i = 1. \quad (3.8)$$

4. When  $X$  only has finitely many possibilities, say  $k$  with  $\mathbb{X} = \{x_1, x_2, \dots, x_k\}$ , then we may think of the probability  $P$  specified by  $(\theta_1, \theta_2, \dots, \theta_k)$  as a point in the **unit**  $(k-1)$  **simplex**:

$$\Delta^{k-1} := \{(\theta_1, \theta_2, \dots, \theta_k) \in \mathbb{R}^k : \sum_i \theta_i = 1 \text{ and } \theta_i \geq 0, \text{ for all } i\} \quad (3.9)$$

In particular when  $X$  has only two possible values with  $\mathbb{X} = \{x_1, x_2\}$  then  $\theta_2 = 1 - \theta_1$ , so we can avoid subscripts and take  $\theta := \theta_1$  and realize that the probability  $P$  is now specified by the point  $(\theta, 1 - \theta)$  in the **unit** 1 **simplex**:

$$\Delta^1 := \{(\theta, 1 - \theta) \in \mathbb{R}^2 : 0 \leq \theta \leq 1\} . \quad (3.10)$$

See <https://en.wikipedia.org/wiki/Simplex> for the images scribed on the board.

#### DISCRETE RANDOM VARIABLES - SIMPLIFIED NOTATION

Notice that in equations (3.5), (3.6) and (3.7) the use of the “ $\omega \in \Omega$ ” notation, where random variables are defined as functions, is much reduced. The reason is that in straightforward examples it is convenient to associate the possible values  $x_1, x_2, \dots$  with the outcomes  $\omega_1, \omega_2, \dots$ . Hence, we can describe a discrete random variable by the table:

Possible values: $x_i$	$x_1$	$x_2$	$x_3$	$\dots$
Probability: $P(X = x_i) = \theta_i$	$\theta_1$	$\theta_2$	$\theta_3$	$\dots$

It is customary to use  $p_i$  instead of  $\theta_i$  for the probabilities. But we try to avoid it as it will hurt us when we start doing Inference Theory soon!

Note that this table hides the more complex notation but it is still there, under the surface. In Probability Theory I, you should be able to work with and manipulate discrete random variables using the simplified notation given above. The same comment applies to the continuous random variables discussed later. But you are students of mathematics and should know more about what is “under the hood”.

Out of the class of discrete random variables we will define specific kinds as they arise often in applications. We classify discrete random variables into three types for convenience as follows:

- Discrete uniform random variables with finitely many possibilities
- Discrete non-uniform random variables with finitely many possibilities
- Discrete non-uniform random variables with (countably) infinitely many possibilities

**Model 2 (Discrete Uniform)** We say that a discrete random variable  $X$  is uniformly distributed over  $k$  possible values in  $\mathbb{X} = \{x_1, x_2, \dots, x_k\}$  if its probability mass function is:

$$f(x) = \begin{cases} \theta_i = \frac{1}{k} & \text{if } x = x_i, \text{ where } i = 1, 2, \dots, k , \\ 0 & \text{otherwise .} \end{cases} \quad (3.11)$$

The distribution function for the discrete uniform random variable  $X$  is:

$$F(x) = \sum_{x_i \leq x} f(x_i) = \sum_{x_i \leq x} \theta_i = \begin{cases} 0 & \text{if } -\infty < x < x_1 , \\ \frac{1}{k} & \text{if } x_1 \leq x < x_2 , \\ \frac{2}{k} & \text{if } x_2 \leq x < x_3 , \\ \vdots & \\ \frac{k-1}{k} & \text{if } x_{k-1} \leq x < x_k , \\ 1 & \text{if } x_k \leq x < \infty . \end{cases} \quad (3.12)$$

The discrete uniform RV with values in  $\mathbb{X} = \{1, 2, \dots, k\}$  is called the equi-probable de Moivre( $k$ ) RV as we will see in the sequel.

**Example 42** The *fair coin toss experiment* of Exercise 3.2 is an example of a discrete uniform random variable with finitely many possibilities. Its probability mass function is given by

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

and its distribution function is given by

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

Let us sketch the probability mass function and distribution function for  $X$  below.

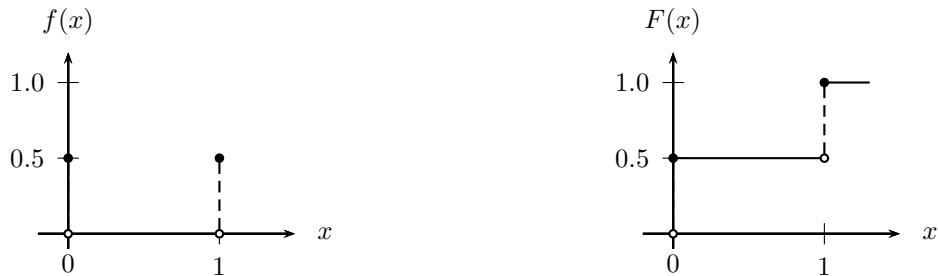


Figure 3.3:  $f(x)$  and  $F(x)$  of the fair coin toss random variable  $X$ , a discrete uniform RV on  $\{0, 1\}$ .

**Example 43 (Fair dice RV)** Now consider the *toss a fair die* experiment and define  $X$  to be the number that shows up on the top face. Note that here  $\Omega$  is the set of numerical symbols  $\{1, 2, 3, 4, 5, 6\}$  that label each face while each of these symbols are associated with the real number  $x \in \{1, 2, 3, 4, 5, 6\}$ . We can describe this random variable by the table

Find the probability mass function and distribution function for this random variable, and sketch their graphs.

Solution:

The probability mass function of this random variable is:

$$f(x) = P(X = x) = \begin{cases} \frac{1}{6} & \text{if } x = 1 \\ \frac{1}{6} & \text{if } x = 2 \\ \frac{1}{6} & \text{if } x = 3 \\ \frac{1}{6} & \text{if } x = 4 \\ \frac{1}{6} & \text{if } x = 5 \\ \frac{1}{6} & \text{if } x = 6 \\ 0 & \text{otherwise} \end{cases}$$

and the distribution function is:

$$F(x) = P(X \leq x) = \begin{cases} 0, & \text{if } -\infty < x < 1 \\ \frac{1}{6}, & \text{if } 1 \leq x < 2 \\ \frac{1}{3}, & \text{if } 2 \leq x < 3 \\ \frac{1}{2}, & \text{if } 3 \leq x < 4 \\ \frac{2}{3}, & \text{if } 4 \leq x < 5 \\ \frac{5}{6}, & \text{if } 5 \leq x < 6 \\ 1, & \text{if } 6 \leq x < \infty \end{cases}$$

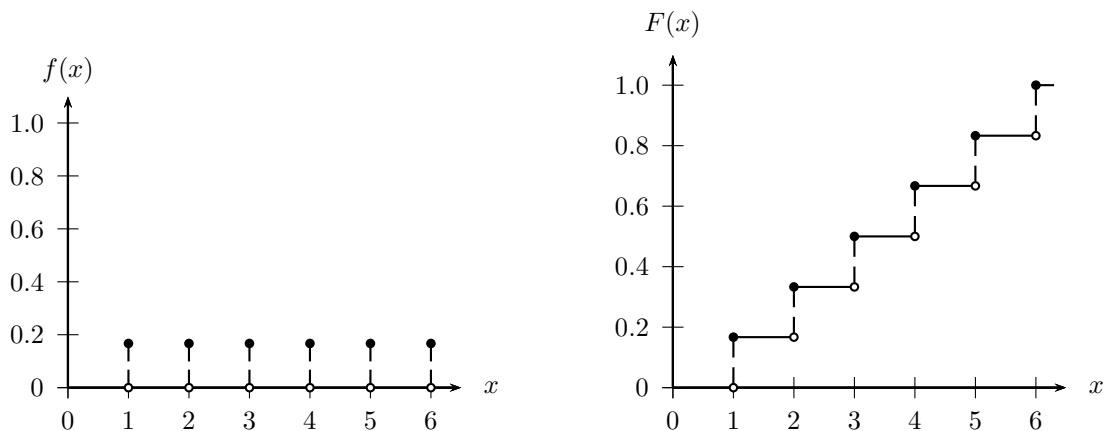


Figure 3.4:  $f(x)$  and  $F(x)$  of the fair die toss random variable  $X$ , a discrete uniform RV on  $\{1, 2, 3, 4, 5, 6\}$ .

**Example 44 (Astragali with a Kiwi sheep ankle bone) Astragali.** Board games involving chance were known in Egypt, 3000 years before Christ. The element of chance needed for these games was at first provided by tossing astragali, the ankle bones of sheep. These bones could come to rest on only four sides, the other two sides being rounded. The upper side of the bone, broad

and slightly convex counted four; the opposite side broad and slightly concave counted three; the lateral side flat and narrow, one, and the opposite narrow lateral side, which is slightly hollow, six. You may examine an astragali of a kiwi sheep.

This is an example of a discrete non-uniform random variable with finitely many possibilities. A surmised probability mass function with  $f(4) = \frac{4}{10}$ ,  $f(3) = \frac{3}{10}$ ,  $f(1) = \frac{2}{10}$ ,  $f(6) = \frac{1}{10}$  and distribution function are shown below.

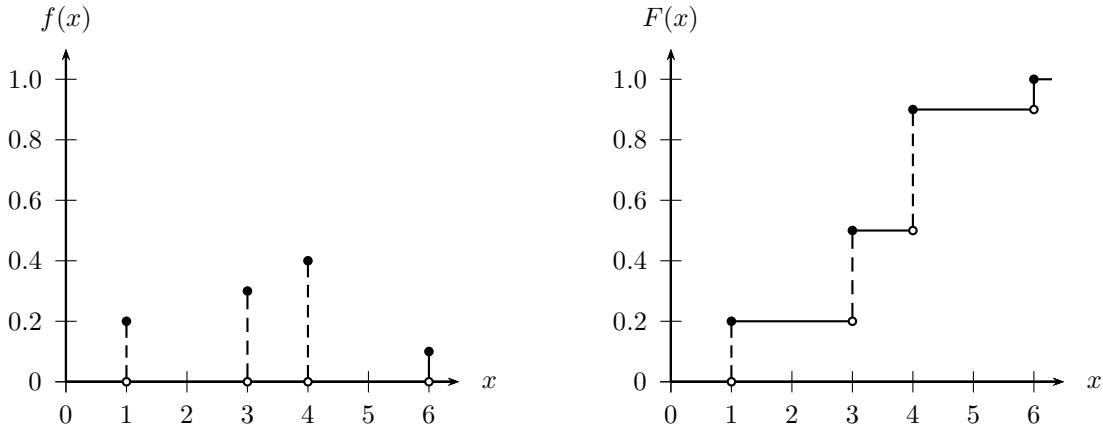


Figure 3.5:  $f(x)$  and  $F(x)$  of surmised *astragali* toss random variable  $X$ , a discrete (non-uniform) RV on  $\{1, 2, 3, 4\}$ .

### 3.2.1 An Elementary Family of Bernoulli Random Variables

In many experiments there are only two outcomes. For instance:

- Flip a coin to see whether it is defective.
- Roll a die and determine whether it is a 6 or not.
- Determine whether it will be below 0 degrees Celsius at 0600 hours in Uppsala tomorrow or not.

Performing such an experiment  $\mathcal{E}$  once to see if an event of interest  $A$  occurs is called a **Bernoulli trial** and its probability model over a triple  $(\Omega, \mathcal{F}, P)$ , with  $A \in \mathcal{F}$ , given by the Indicator Function  $\mathbb{1}_A$  in Model 1 is called the Bernoulli RV.

If we do not know the probability  $\theta$  that ‘ $A$  occurs’, i.e., the Bernoulli RV will equal 1, then we can define a whole family of Bernoulli RVs for each  $\theta \in [0, 1]$  or more precisely for each  $(\theta, 1 - \theta) \in \Delta^1$ , the unit 1-Simplex. Note that this family includes the fair Bernoulli trial of Example 42 when  $\theta = 0.5$ . Let us formalise this as the  $\text{Bernoulli}(\theta)$  RV for each  $\theta \in [0, 1]$  next.

**Model 3 (Bernoulli( $\theta$ ) RV)** Given a parameter  $\theta \in [0, 1]$ , the probability mass function (PMF) for the  $\text{Bernoulli}(\theta)$  RV  $X$  is:

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x} \mathbb{1}_{\{0,1\}}(x) = \begin{cases} \theta & \text{if } x = 1, \\ 1 - \theta & \text{if } x = 0, \\ 0 & \text{otherwise} \end{cases} \quad (3.13)$$

and its DF is:

$$F(x; \theta) = \begin{cases} 1 & \text{if } 1 \leq x, \\ 1 - \theta & \text{if } 0 \leq x < 1, \\ 0 & \text{otherwise} \end{cases} \quad (3.14)$$

We emphasise the dependence of the probabilities on the parameter  $\theta$  by specifying it following the semicolon in the argument for  $f$  and  $F$  and by subscripting the probabilities, i.e.  $P_\theta(X = 1) = \theta$  and  $P_\theta(X = 0) = 1 - \theta$ .

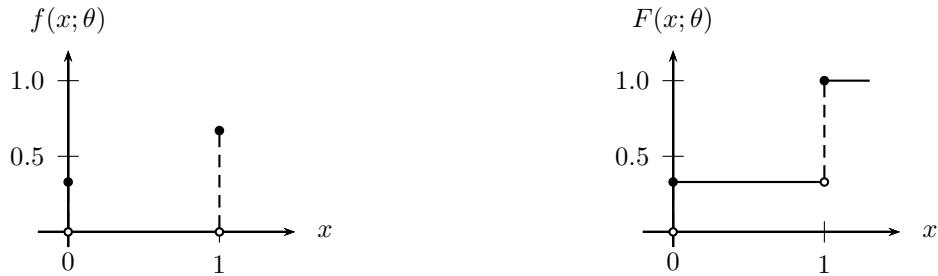


Figure 3.6: PMF  $f(x; \theta)$  and DF  $F(x; \theta)$  with  $\theta = 0.33$ . You should see how PMF and DF change as  $\theta$  goes from 0 to 1

### 3.2.2 Independent Bernoulli Trials

Random variables make sense for a series of trials as well as just a single trial of an experiment. We now look at what happens when we perform a sequence of independent Bernoulli trials. For instance:

- Flip a coin 10 times; count the number of heads
  - by possibly allowing for the coin's  $P(H)$  to change each time because each of them are manufactured in a terrible mint.
- Test 50 randomly selected circuits from an assembly line; count the number of defective circuits.
- Roll a die 100 times; count the number of sixes you throw.
- Provide a property near a particular bridge in our archipelago with flood insurance for 20 years; count the number of years, during the 20-year period, during which the property is flooded. Note: we assume that flooding is independent from year to year, and that the probability of flooding is the same each year.

Since the  $\text{Bernoulli}(\theta)$  RV has only two outcomes, i.e., simple events, we know how to obtain the probability of each of the two outcomes in a given Bernoulli trial with the probability given by the deterministic variable or parameter  $\theta$ . Now consider doing more than one trial so we have sequence of  $\text{Bernoulli}(\theta_i)$  trials, say,

$$X_i \sim \text{Bernoulli}(\theta_i) \text{ with } i \in \mathbb{N},$$

with each  $\theta_i \in [0, 1]$  being possibly unknown but fixed as a parameter. Now, if we assume independence across trials, so one trial's outcome does not affect the outcome of any of the other trials, in the sense of Definition 17 about *independence of a sequence of events*, then we can obtain the

probability of the entire sequence of outcomes for this sequence of **independently distributed** Bernoulli( $\theta_i$ ) **trails** which can be any infinite sequence of 0's and 1's, i.e., any element of  $\{0, 1\}^\infty$ , by simply multiplying the corresponding probabilities given by  $\theta_i$ 's in  $(\theta_1, \theta_2, \dots) \in [0, 1]^\infty$ , an infinite dimensional parameter space, as follows:

$$\begin{aligned} P(x; (\theta_1, \theta_2, \dots)) &= \prod_i f(x_i; \theta_i) = \prod_i \theta_i^{x_i} (1 - \theta_i)^{1-x_i} \mathbb{1}_{\{0,1\}}(x_i), \\ &\text{where } x := (x_1, x_2, \dots) \in \mathbb{X}_\infty = \{0, 1\}^\infty := \{0, 1\} \times \{0, 1\} \times \dots \end{aligned} \quad (3.15)$$

By further assuming that all the  $\theta_i$ 's are identical, say  $\theta = \theta_1 = \theta_2 = \dots$ , with  $\theta \in [0, 1]$ , a one-dimensional parameter space, we get the much simpler expression for the **independent and identically distributed (IID)** Bernoulli( $\theta$ ) **trails** as follows:

$$\begin{aligned} P(x; \theta) &= \prod_i f(x_i; \theta) = \prod_i \theta^{x_i} (1 - \theta)^{1-x_i} \mathbb{1}_{\{0,1\}}(x_i) = \mathbb{1}_{\{0,1\}^\infty}(x) \theta^{\sum_i x_i} (1 - \theta)^{\sum_i (1-x_i)} \\ &= \begin{cases} \theta^{\sum_i x_i} (1 - \theta)^{n - \sum_i x_i} & \text{if } x := (x_1, x_2, \dots) \in \mathbb{X}_\infty = \{0, 1\}^\infty \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (3.16)$$

Remembering that all other RVs can be derived from such IID Bernoulli( $\theta$ ) trials using  $\theta = 1/2$ , as we will see in the sequel, we are ready to take a tour through some common discrete and continuous random variables that are useful in many applications.

### 3.2.3 Some Common Discrete Random Variables

Let us start with the simplest example to fix ideas carefully.

**Example 45 (Waiting For the First Heads)** Suppose our experiment is to toss a fair coin independently and identically (that is, the same coin is tossed in essentially the same manner independent of the other tosses in each trial) as often as necessary until we have a head, H. Let the random variable  $X$  denote the *Number of trials until the first H appears*.

Let's first find the probability mass function of  $X$ .

Now  $X$  can take on the values  $\{1, 2, 3, \dots\}$ , so we have a non-uniform random variable with infinitely many possibilities. Since

$$\begin{aligned} f(1) &= P(X = 1) = P(H) = \frac{1}{2}, \\ f(2) &= P(X = 2) = P(TH) = \frac{1}{2} \cdot \frac{1}{2} = \left(\frac{1}{2}\right)^2, \\ f(3) &= P(X = 3) = P(TTH) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \left(\frac{1}{2}\right)^3, \quad \text{etc.} \end{aligned}$$

the probability mass function of  $X$  is:

$$f(x) = P(X = x) = \left(\frac{1}{2}\right)^x, \quad x = 1, 2, \dots.$$

In the previous Example, noting that we have independent trials, we get:

$$f(x) = P(X = x) = P(\underbrace{TT \dots T}_{n-1} H) = P(T)^{x-1} P(H) = \left(\frac{1}{2}\right)^{x-1} \frac{1}{2}.$$

More generally, let there be two possibilities, success (S) or failure (F), with  $P(S) = \theta$  and  $P(F) = 1 - \theta$  so that:

$$P(X = x) = P(\underbrace{FF \dots FS}_{x-1}) = (1 - \theta)^{x-1} \theta.$$

This is called a **geometric random variable** with “success probability” parameter  $\theta$ . We can spot a geometric distribution because there will be *a sequence of independent trials with a constant probability of success. We are counting the number of trials until the first success appears*. Let us define this random variable formally next.

**Model 4 (Geometric( $\theta$ ) RV)** Given a parameter  $\theta \in (0, 1)$ , the PMF of the Geometric( $\theta$ ) RV  $X$  is

$$f(x; \theta) = \begin{cases} \theta(1 - \theta)^x & \text{if } x \in \mathbb{Z}_+ := \{0, 1, 2, \dots\} \\ 0 & \text{otherwise} \end{cases} \quad (3.17)$$

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

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

The above equality is a consequence of the geometric series identity (3.18) with  $a = \theta$  and  $\vartheta := 1 - \theta$ :

$$\sum_{x=0}^{\infty} a\vartheta^x = a \left( \frac{1}{1 - \vartheta} \right), \text{ provided, } 0 < \vartheta < 1. \quad (3.18)$$

**Proof:**

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

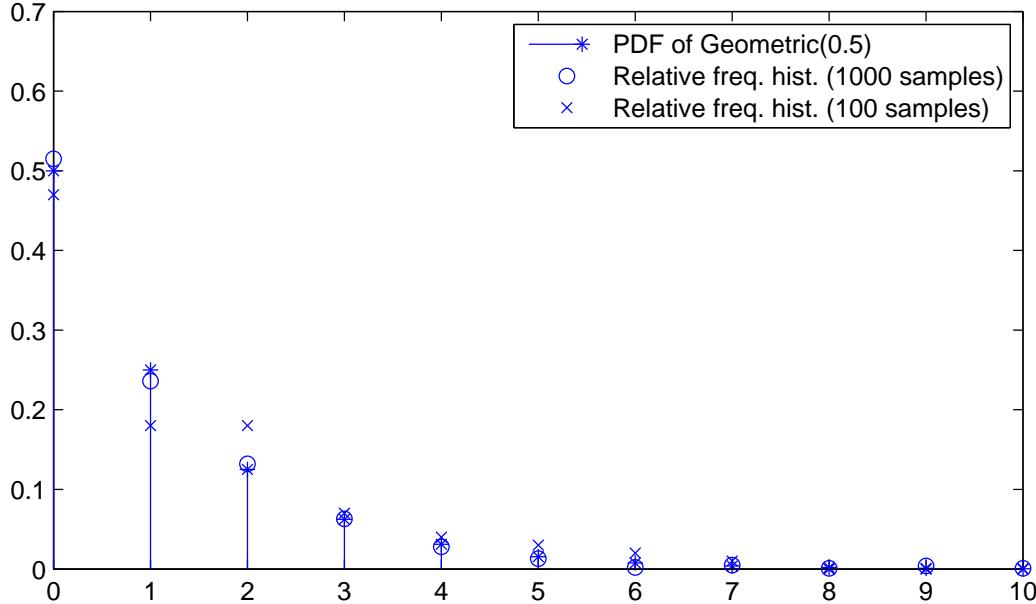
Therefore,

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

The outcome of a 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  $\theta$  in a independent and identical manner.”

**Exercise 3.3 (Coupon Collector’s Problem)** Recall the Coupon Collector’s Problem from lectures.

Figure 3.7: PMF of  $X \sim \text{Geometric}(\theta = 0.5)$  and the relative frequency histogram based on 100 and 1000 samples from  $X$  according to Simulation 144 and Labwork 145 you will see in the sequel.



**Example 46** Suppose we flip a coin 10 times and count the number of heads. Let's consider the probability of getting three heads, say. The probability that the first three flips are heads and the last seven flips are tails, *in order*, is

$$\underbrace{\frac{1}{2} \frac{1}{2} \frac{1}{2}}_{3 \text{ successes}} \quad \underbrace{\frac{1}{2} \frac{1}{2} \cdots \frac{1}{2}}_{7 \text{ failures}}.$$

But there are

$$\binom{10}{3} = \frac{10!}{7! 3!} = 120$$

ways of ordering three heads and seven tails, so the probability of getting three heads and seven tails *in any order*, is

$$P(\text{'3 heads'}) = \binom{10}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^7 \approx 0.117$$

We can describe this sort of situation by considering a random variable  $X$  which counts the number of successes, as follows:

**Model 5 (Binomial( $n, \theta$ ) RV)** Let the RV  $X = \sum_{i=1}^n X_i$  be the sum of  $n$  independent and identically distributed Bernoulli( $\theta$ ) 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  $\theta$ , the PMF of the Binomial( $n, \theta$ ) RV  $X$  is:

$$f(x; n, \theta) = \begin{cases} \binom{n}{x} \theta^x (1 - \theta)^{n-x} & \text{if } x \in \{0, 1, 2, 3, \dots, n\}, \\ 0 & \text{otherwise} \end{cases} \quad (3.19)$$

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  $x$ .”

**A Quick Justification:** The argument from Example 46 generalises as follows. Since the trials are independent and identical, the probability of  $x$  successes followed by  $n - x$  failures, *in order*, is given by

$$\underbrace{\text{SS}\dots\text{S}}_x \underbrace{\text{FF}\dots\text{F}}_{n-x} = \theta^x(1-\theta)^{n-x}.$$

Since the  $n$  symbols SS  $\dots$  SFF  $\dots$  F may be arranged in

$$\binom{n}{x} = \frac{n!}{(n-x)!x!}$$

ways, the probability of  $x$  successes and  $n - x$  failures, *in any order*, is given by

$$\binom{n}{x} \theta^x(1-\theta)^{n-x}.$$

**Proof:** This is only a sketch. A formal proof should start with the mathematical induction for the very formula for the binomial coefficient.

Observe that for the Binomial( $n, \theta$ ) RV  $X$ ,  $P(X = x) = f(x; n, \theta)$  is the probability that  $x$  of the  $n$  Bernoulli( $\theta$ ) 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( $\theta$ ) RVs  $\{X_1, X_2, \dots, X_n\}$ . Now choose  $x$  indices from  $\{1, 2, \dots, n\}$  to mark those trials in a particular realization of  $\{x_1, x_2, \dots, x_n\}$  with the Bernoulli outcome of 1. The probability of each such event is  $\theta^x(1-\theta)^{n-x}$  due to the IID assumption. For each realization  $\{x_1, x_2, \dots, x_n\} \in \{0, 1\}^n := \{\text{all binary } (0-1) \text{ strings of length } n\}$ , specified by a choice of  $x$  trial indices with Bernoulli outcome 1, the binomial RV  $X = \sum_{i=1}^n X_i$  takes the value  $x$ . Since there are exactly  $\binom{n}{x}$  many ways in which we can choose  $x$  trial indices (with outcome 1) from the set of  $n$  trial indices  $\{1, 2, \dots, n\}$ , we get the desired product for  $f(x; n, \theta) = \binom{n}{x} \theta^x(1-\theta)^{n-x}$  when  $x \in \{0, 1, \dots, n\}$ .

**Example 47** Find the probability that seven of ten persons will recover from a tropical disease where the probability is identically 0.80 that any one of them will recover from the disease.

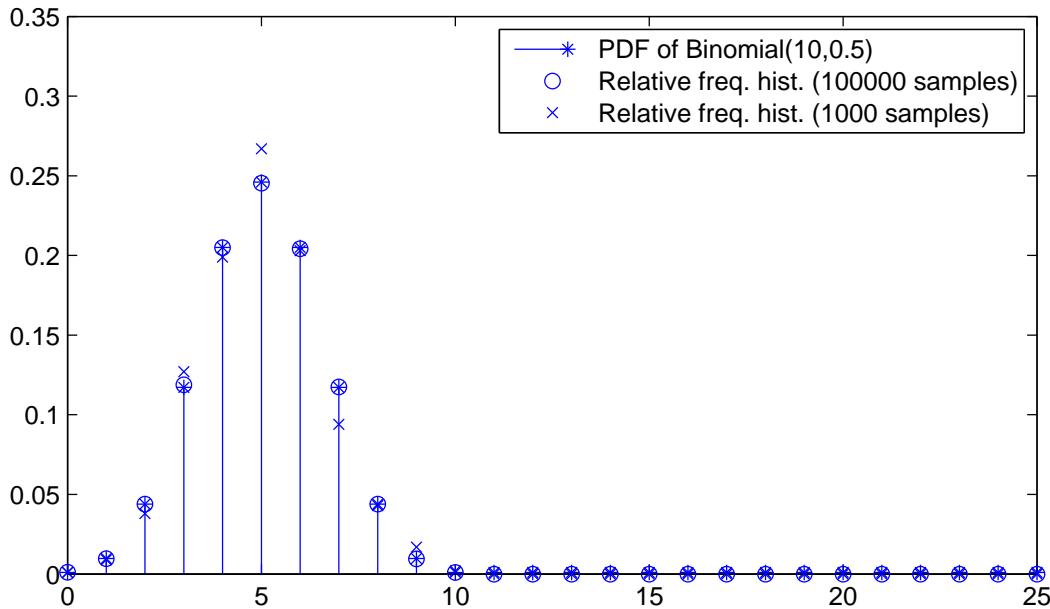
Solution:

We can assume independence here, so we have a binomial situation with  $x = 7$ ,  $n = 10$ , and  $\theta = 0.8$ . Substituting these into the formula for the probability mass function for Binomial(10, 0.8) random variable, we get:

$$\begin{aligned} f(7; 10, 0.8) &= \binom{10}{7} \times (0.8)^7 \times (1-0.8)^{10-7} \\ &= \frac{10!}{(10-7)!7!} \times (0.8)^7 \times (1-0.8)^{10-7} \\ &= 120 \times (0.8)^7 \times (1-0.8)^{10-7} \\ &\approx 0.20 \end{aligned}$$

\*In the exam you can give your answer as such an expression.

Figure 3.8: PDF of  $X \sim \text{Binomial}(n = 10, \theta = 0.5)$  and the relative frequency histogram based on 100,000 samples from  $X$  obtained according to Simulation 148.



**Example 48** Compute the probability of obtaining *at least two 6's* in rolling a fair die independently and identically four times.

Solution:

In any given toss let  $\theta = P(\{6\}) = 1/6$ ,  $1 - \theta = 5/6$ ,  $n = 4$ .

The event *at least two 6's* occurs if we obtain two or three or four 6's. Hence the answer is:

$$\begin{aligned}
 P(\text{at least two 6's}) &= f\left(2; 4, \frac{1}{6}\right) + f\left(3; 4, \frac{1}{6}\right) + f\left(4; 4, \frac{1}{6}\right) \\
 &= \binom{4}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{4-2} + \binom{4}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^{4-3} + \binom{4}{4} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^{4-4} \\
 &= \frac{1}{6^4} (6 \cdot 25 + 4 \cdot 5 + 1) \\
 &\approx 0.132
 \end{aligned}$$

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

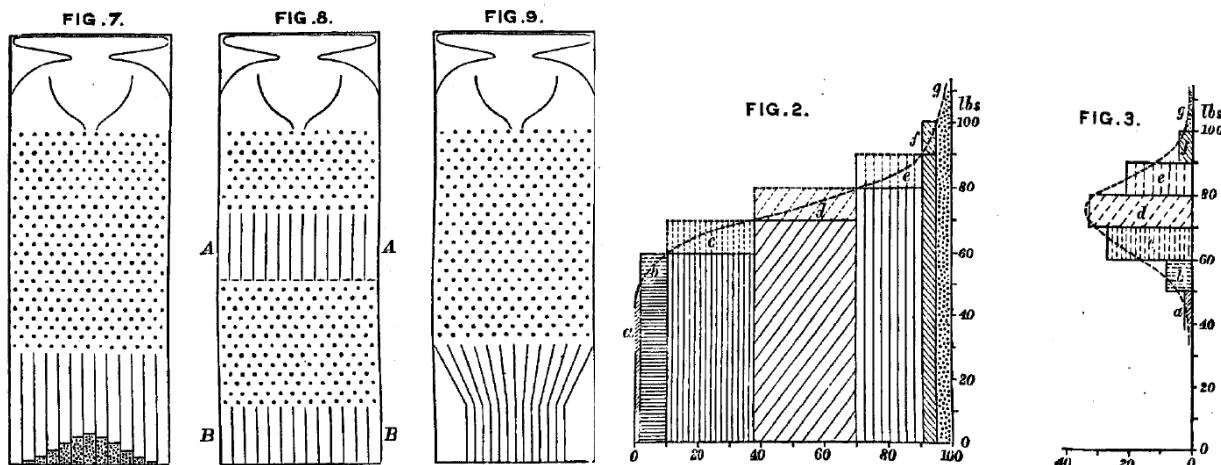
### Sir Francis Galton's Quincunx

This section is introduced to provide some forms for a kinesthetic (hands-on) and visual understanding of some elementary statistical distributions and laws. The following words are from Sir Francis Galton, F.R.S., *Natural Inheritance*, pp. 62-65, Macmillan, 1889. In here you will already find the kernels behind the construction of  $\text{Binomial}(\theta)$  RV as sum of IID  $\text{Bernoulli}(\theta)$  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, if its mountains could be thrown into its lakes, two nuances would be got rid of at once. An Average is but a solitary fact, whereas if a single other fact be added to it, an entire Normal Scheme, which nearly corresponds to the observed one, starts potentially into existence.

Some people hate the very name of statistics, but I find them full of beauty and interest. Whenever they are not brutalised, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomenon is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of man.

Figure 3.9: Figures from Sir Francis Galton, F.R.S., *Natural Inheritance*, Macmillan, 1889.



(a) FIG. 7, FIG. 8, and FIG. 9 (p. 63)

(b) FIG. 2 and FIG. 3 (p. 38)

Mechanical Illustration of the Cause of the Curve of Frequency.—*The Curve of Frequency, and that of Distribution, are convertible: therefore if the genesis of either of them can be made clear, that of the other also becomes intelligible. I shall now illustrate the origin of the Curve of Frequency, by means of an apparatus shown in Fig. 7, that mimics in a very pretty way the conditions on which Deviation depends.* It is a frame glazed in front, leaving a depth of about a quarter of an inch behind the glass. Strips are placed in the upper part to act as a funnel. Below the outlet of the funnel stand a succession of rows of pins stuck squarely into the backboard, and below these again are a series of vertical compartments. A charge of small shot is inclosed. When the frame is held topsy-turvy, all the shot runs to the upper end; then, when it is turned back into its working position, the desired action commences. Lateral strips, shown in the diagram, have the effect of directing all the shot that had collected at the upper end of the frame to run into the wide 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.”*

We now consider the last of our common discrete random variables for now, the **Poisson** case. A Poisson random variable counts the number of times an event occurs.

We might, for example, ask:

- How many customers visit Cafe Angstrom each day?
- How many sixes are scored in a cricket season? Cricket is a game played in the English-speaking worlds.
- How many bombs hit a city block in south London during World War II?

A Poisson experiment has the following characteristics:

- The average rate of an event occurring is known. This rate is constant.
- The probability that an event will occur during a short continuum is proportional to the size of the continuum.
- Events occur independently.

The number of events occurring in a Poisson experiment is referred to as a **Poisson random variable**.

**Model 6 (Poisson( $\lambda$ ) RV)** Given a real parameter  $\lambda > 0$ , the discrete RV  $X$  is said to be Poisson( $\lambda$ ) distributed if  $X$  has PDF:

$$f(x; \lambda) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!} & \text{if } x \in \mathbb{Z}_+ := \{0, 1, 2, \dots\}, \\ 0 & \text{otherwise.} \end{cases} \quad (3.20)$$

Note that the PDF integrates to 1:

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

where we exploit the Taylor series of  $e^\lambda$  to obtain the second-last equality above.

We interpret  $X$  as the number of times an event occurs during a specified continuum given that the average value in the continuum is  $\lambda$ .

**Example 49** If on the average, 2 cars enter a certain parking lot per minute, what is the probability that during any given minute three cars or fewer will enter the lot?

Think: Why are the assumptions for a Poisson random variable likely to be correct here?

Note: Use calculators, or Excel or Maple, etc. In an exam you may be given needed values from Poisson tables.

Let the random variable  $X$  denote the number of cars arriving per minute. Note that the continuum is 1 minute here. Then  $X$  can be considered to have a Poisson distribution with  $\lambda = 2$  because 2 cars enter on average.

The probability that three cars or fewer enter the lot is:

$$\begin{aligned} P(X \leq 3) &= f(0; 2) + f(1; 2) + f(2; 2) + f(3; 2) \\ &= e^{-2} \left( \frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!} \right) && \text{*This is a perfectly fine answer in the exam.} \\ &= 0.857 \quad (3 \text{ sig. fig.}) \end{aligned}$$

**Example 50 (Arrivals at a Service Station)** The proprietor of a service station finds that, on average, 8 cars arrive *per hour* on Saturdays. What is the probability that during a randomly chosen 15 *minute period* on a Saturday:

- (a) No cars arrive?
- (b) At least three cars arrive?

Solution:

Let the random variable  $X$  denote the number of cars arriving in a 15 minute interval. The continuum is 15 minutes here so we need the average number of cars that arrive in a 15 minute period, or  $\frac{1}{4}$  of an hour. We know that 8 cars arrive per hour, so  $X$  has a Poisson distribution with

$$\lambda = \frac{8}{4} = 2.$$

(a)

$$P(X = 0) = f(0; 2) = \frac{e^{-2} 2^0}{0!} = 0.135 \quad (3 \text{ sig. fig})$$

(b)

$$\begin{aligned} P(X \geq 3) &= 1 - P(X < 3) \\ &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - f(0; 2) - f(1; 2) - f(2; 2) \\ &= 1 - 0.1353 - 0.2707 - 0.2707 \\ &= 0.323 \quad (3 \text{ sig. fig.}) \end{aligned}$$

**Remark 24** In the binomial case where  $\theta$  is small and  $n$  is large, it can be shown that the binomial distribution with parameters  $n$  and  $\theta$  is closely approximated by the Poisson distribution having  $\lambda = n\theta$ . The smaller the value of  $\theta$  and larger the value of  $n$ , the better the approximation.

In the sequel we will see more formally, after understanding notions of convergence of RVs, that the sum of a sequence of  $n$  IID Bernoulli( $\theta$ ) RVs with  $\lambda = n\theta$  converges to the Poisson( $\lambda$ ) RV as  $n \rightarrow \infty$  and  $\theta \rightarrow 0$  in a specific sense.

**Example 51 (Still-born Babies)** About 0.01% of babies are stillborn in a certain hospital. We find the probability that of the next 5000 babies born, there will be no more than 1 stillborn baby.

Let the random variable  $X$  denote the number of stillborn babies. Then  $X$  has a binomial distribution with parameters  $n = 5000$  and  $\theta = 0.0001$ . Since  $\theta$  is so small and  $n$  is large, this binomial distribution may be approximated by a Poisson distribution with parameter

$$\lambda = n\theta = 5000 \times 0.0001 = 0.5.$$

Hence

$$P(X \leq 1) = P(X = 0) + P(X = 1) = f(0; 0.5) + f(1; 0.5) = 0.910 \quad (3 \text{ sig. fig.})$$

**Exercise 3.4 (Nazi Bombs on London)** Feller discusses the probability and statistics of flying bomb hits in an area of southern London during II world war. The area in question was partitioned into  $24 \times 24 = 576$  small squares. The total number of hits was 537. There were 229 squares with 0 hits, 211 with 1 hit, 93 with 2 hits, 35 with 3 hits, 7 with 4 hits and 1 with 5 or more hits. Assuming the hits were purely random, use the Poisson approximation to find the probability that a particular square would have exactly  $k$  hits. Compute the expected number of squares that would have 0, 1, 2, 3, 4, and 5 or more hits and compare this with the observed results (Snell 9.2.14).

### THINKING POISSON

The Poisson distribution has been described as a limiting version of the Binomial. In particular, Exercise 49 thinks of a Poisson distribution as a model for the number of events (cars) that occur in a period of time (1 minute) when in each little chunk of time one car arrives with constant probability, independently of the other time intervals. This leads to the general view of the Poisson distribution as a good model when:

*You count the number of events in a continuum when the events occur at constant rate, one at a time and independent of each other.*

## DISCRETE RANDOM VARIABLE SUMMARY

Probability mass function

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

Distribution function

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

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

### 3.3 Exercises in Discrete Random Variables

**Ex. 3.5** — One number in the following table for the probability function of a random variable  $X$  is incorrect. Which is it, and what should the correct value be?

$x$	1	2	3	4	5
$P(X = x)$	0.07	0.10	1.10	0.32	0.40

**Ex. 3.6** — Let  $X$  be the number of years before a particular type of machine will need replacement. Assume that  $X$  has the probability function  $f(1) = 0.1$ ,  $f(2) = 0.2$ ,  $f(3) = 0.2$ ,  $f(4) = 0.2$ ,  $f(5) = 0.3$ .

1. Find the distribution function,  $F$ , for  $X$ , and graph both  $f$  and  $F$ .
2. Find the probability that the machine needs to be replaced during the first 3 years.
3. Find the probability that the machine needs no replacement during the first 3 years.

**Ex. 3.7** — Of 200 adults, 176 own one TV set, 22 own two TV sets, and 2 own three TV sets. A person is chosen at random. What is the probability mass function of  $X$ , the number of TV sets owned by that person?

**Ex. 3.8** — Suppose a discrete random variable  $X$  has probability function give by

$x$	3	4	5	6	7	8	9	10	11	12	13
$P(X = x)$	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.

$$(i) P(X \leq 5)$$

$$(ii) P(X < 12)$$

$$(iii) P(X > 9)$$

$$(iv) P(X \geq 9)$$

$$(v) P(4 < X \leq 9)$$

$$(vi) P(4 < X < 11)$$

**Ex. 3.9** — A box contains 4 right-handed and 6 left-handed screws. Two screws are drawn at random without replacement. Let  $X$  be the number of left-handed screws drawn. Find the probability mass function for  $X$ , and then calculate the following probabilities:

1.  $P(X \leq 1)$
2.  $P(X \geq 1)$
3.  $P(X > 1)$

**Ex. 3.10** — Suppose that a random variable  $X$  has geometric probability mass function,

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

1. Find the value of  $k$ .
2. What is  $P(X \geq 4)$ ?

**Ex. 3.11** — Four fair coins are tossed simultaneously. If we count the number of heads that appear then we have a binomial random variable,  $X = \text{the number of heads}$ .

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. 3.12** — The distribution of blood types in a certain population is as follows:

Blood type	Type O	Type A	Type B	Type AB
Proportion	0.45	0.40	0.10	0.05

A random sample of 15 blood donors is observed from this population. Find the probabilities of the following events.

1. Only one type  $AB$  donor is included.
2. At least three of the donors are type  $B$ .
3. More than ten of the donors are *either* type  $O$  or type  $A$ .
4. Fewer than five of the donors are *not* type  $A$ .

**Ex. 3.13** — If the probability of hitting a target in a single shot is 10% and 10 shots are fired independently, what is the probability that the target will be hit at least once?

**Ex. 3.14** — Suppose that a certain type of magnetic tape contains, on the average, 2 defects per 100 meters. What is the probability that a roll of tape 300 meters long will contain no defects?

**Ex. 3.15** — In 1910, E. Rutherford and H. Geiger showed experimentally that the number of alpha particles emitted per second in a radioactive process is a random variable  $X$  having a Poisson distribution. If the average number of particles emitted per second is 0.5, what is the probability of observing two or more particles during any given second?

**Ex. 3.16 —** The number of lacunae (surface pits) on specimens of steel, polished and examined in a metallurgical laboratory, is thought to have a Poisson distribution.

1. Write down the formula for the probability that a specimen has  $x$  defects, explaining the meanings of the symbols you use.
2. Simplify the formula in the case  $x = 0$ .
3. In a large homogeneous collection of specimens, 10% have one or more lacunae. Find (approximately) the percentage having exactly two.
4. Why might the Poisson distribution not apply in this situation?

[HINT: Recall the *emphasised sentence* in THINKING POISSON and what the continuum on which the number of events occur is for the problem, and what could possibly go wrong in your imagination of the manufacturing process of the steel specimens (normally you need to melt and manipulate iron with other elements and cast them in moulds and this needs energy and raw materials of possibly varying quality and the machines used in the process could break down, etc.) to violate the Poisson assumption about the occurrence of pits on the surface of the specimens.]

### 3.4 Continuous Random Variables

If  $X$  is a measurement of a continuous quantity, such as,

- the maximum diameter in millimeters of a venus shell I picked up at New Brighton beach,
- the distance you transported yourself to lectures today in meters,
- the volume of rain that fell on the roof of this building over the past 365 days in litres,
- the vertical position (in micro meters above sea-level) since the release of a pollen grain at a location in Lake Rogen in Härjedalen, as it traces through Göta älv—Klarälven, the longest river of Sweden before discharging in a delta into Vänern at Karlstad.
- the volume of water (in cubic meters) that fell on the southern Alps of the South Island of New Zealand throughout last year.
- etc.,

then  $X$  is a continuous random variable. Continuous random variables are based on measurements in a continuous scale of a given precision as opposed to discrete random variables that are based on counting.

**Example 52** Suppose that  $X$  is the time, in minutes, before the next student leaves the lecture room. This is an example of a continuous random variable that takes one of (uncountably) infinitely many values. When a student leaves,  $X$  will take on the value  $x$  and this  $x$  could be 2.1 minutes, or 2.1000000001 minutes, or 2.9999999 minutes, etc., depending the measurement precision of the clock being used to measure time.

Finding  $P(X = 2)$ , for example, doesn't make sense because how can it ever be *exactly* 2.00000... minutes? It is more sensible to consider probabilities like  $P(X > x)$  or  $P(X < x)$  or  $P(a < X < b)$  with  $a < b$ , up to measurement precision of the time-measuring clock rather than the discrete approach of trying to compute  $P(X = x)$ .

The characteristics of continuous random variables are:

- The outcomes are measured, not counted.
- Geometrically, *the probability of an outcome is equal to an area under a mathematical curve.*
- Each individual value has zero probability of occurring. So we find the probability that the value is between two endpoints of an interval, or a set of intervals, including half-lines in  $\mathbb{R}$ .

**Definition 25 (probability density function (PDF))** A RV  $X$  with distribution function (DF) given by  $F$  is said to be **continuous** if there exists a piecewise-continuous function  $f$ , called the **probability density function (PDF)** of  $X$ , such that

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(v) dv \quad (3.21)$$

where  $f : \mathbb{R} \rightarrow \mathbb{R}$  is a non-negative function, i.e.,  $f(x) \geq 0$ . We write  $v$  because  $x$  is needed as the upper limit of the integral. Piecewise-continuity of  $f$  means  $f$  is continuous, perhaps possibly at the  $x$ -values where  $f$  is discontinuous between the continuous pieces (see <https://en.wikipedia.org/wiki/Piecewise>).

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

1. For any  $x \in \mathbb{R}$ ,  $P(X = x) = P(X \in [x, x]) = \int_x^x f(v)dv = 0$ .
2. By the fundamental theorem of calculus:

$$f(x) = \frac{d}{dx} F(x) =: F'(x), \quad (3.22)$$

for every  $x$  at which  $f(x)$  is continuous.

3. Consequentially, for any  $a, b \in \mathbb{R}$  with  $a < b$ ,

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b) \quad (3.23)$$

$$= F(b) - F(a) = \int_a^b f(v)dv. \quad (3.24)$$

4. And  $P(\Omega) = 1$  implies that:

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

The next set of examples illustrate notation and typical applications of the formulae above.

**Example 53** Consider the continuous random variable,  $X$ , whose probability density function is:

$$f(x) = \begin{cases} 3x^2 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

- (a) Find the distribution function,  $F(x)$ .
- (b) Find  $P(\frac{1}{3} \leq X \leq \frac{2}{3})$ .

*Solution*

(a) First note that if  $x \leq 0$ , then

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

If  $0 < x < 1$ , then

$$\begin{aligned} F(x) &= \int_{-\infty}^0 0 dv + \int_0^x 3v^2 dv \\ &= 0 + [v^3]_0^x \\ &= x^3 \end{aligned}$$

If  $x \geq 1$ , then

$$\begin{aligned} F(x) &= \int_{-\infty}^0 0 dv + \int_0^1 3v^2 dv + \int_1^x 0 dv \\ &= 0 + [v^3]_0^1 + 0 \\ &= 1 \end{aligned}$$

Hence

$$F(x) = \begin{cases} 0 & x \leq 0 \\ x^3 & 0 < x < 1 \\ 1 & x \geq 1 \end{cases}$$

(b)

$$\begin{aligned} P\left(\frac{1}{3} \leq X \leq \frac{2}{3}\right) &= F\left(\frac{2}{3}\right) - F\left(\frac{1}{3}\right) \\ &= \left(\frac{2}{3}\right)^3 - \left(\frac{1}{3}\right)^3 \\ &= \frac{7}{27} \end{aligned}$$

**Example 54** Consider the continuous random variable,  $X$ , whose distribution function is:

$$F(x) = \begin{cases} 0 & x \leq 0 \\ \sin(x) & 0 < x < \frac{\pi}{2} \\ 1 & x \geq \frac{\pi}{2} \end{cases}$$

(a) Find the probability density function,  $f(x)$ .

(b) Find  $P(X > \frac{\pi}{4})$

*Solution*

(a) The probability density function,  $f(x)$  is given by

$$f(x) = F'(x) = \begin{cases} 0 & x < 0 \\ \cos x & 0 < x < \frac{\pi}{2} \\ 0 & x \geq \frac{\pi}{2} \end{cases}$$

(b)

$$P\left(X > \frac{\pi}{4}\right) = 1 - P\left(X \leq \frac{\pi}{4}\right) = 1 - F\left(\frac{\pi}{4}\right) = 1 - \sin\left(\frac{\pi}{4}\right) = 0.293 \text{ (3 sig. fig.)}$$

\* You may stop at  $1 - \sin\left(\frac{\pi}{4}\right)$  for full credit in the exam.

Note:  $f(x)$  is not defined at  $x = 0$  as  $F(x)$  is not differentiable at  $x = 0$ . There is a “kink” in the distribution function at  $x = 0$  causing this problem. It is standard to define  $f(0) = 0$  in such situations, as  $f(x) = 0$  for  $x < 0$ . This choice is arbitrary but it simplifies things and makes no difference to the calculated probability.

Now that we have warmed-up with two examples of continuous RVs, let us define the most elementary continuous RV next.

### 3.4.1 An Elementary Continuous Random Variable

An elementary and fundamental example of a continuous RV is the Uniform(0, 1) RV of Model 7. It forms the foundation for all non-uniform random variate generation and simulation as we will see in Chapter 4. In fact, it is appropriate to call this the fundamental model since every other probability model can be obtained from this one!

**Model 7 (The Fundamental Model)** The probability density function (PDF) of the fundamental model or the Uniform(0, 1) RV is

$$f(x) = \mathbb{1}_{[0,1]}(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise} \end{cases} \quad (3.25)$$

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

$$F(x) := \int_{-\infty}^x f(y) dy = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x > 1 \end{cases} \quad (3.26)$$

Note that the DF is the identity map in  $[0, 1]$ . The PDF and DF are depicted in Figure 3.11.

Let us draw the PDF and DF for Uniform(0, 1) RV next by hand.

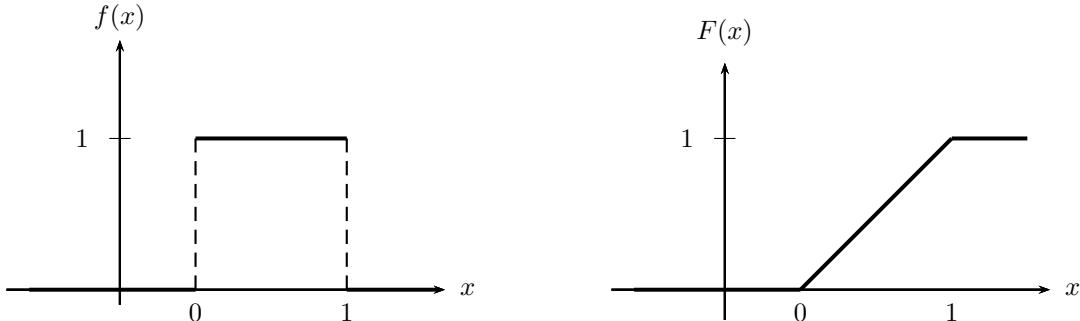
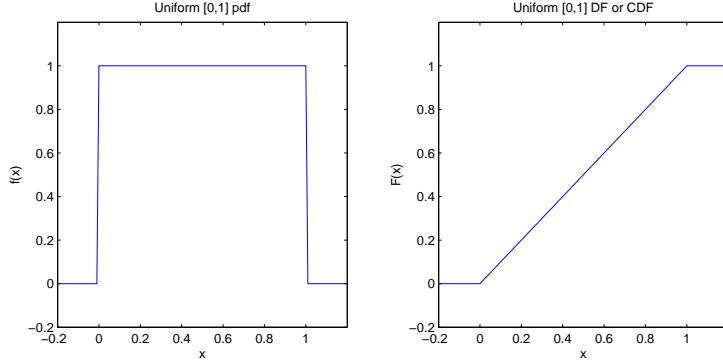


Figure 3.10:  $f(x)$  and  $F(x)$  of the Uniform(0, 1) random variable  $X$ .

Figure 3.11: A convenient but mathematically imprecise Matlab plot from a polyline interpolation for the PDF and DF or CDF of the Uniform(0, 1) continuous RV  $X$ .



**\*\*tossing a fair coin infinitely often, i.e., IID sequence of Bernoulli(1/2) trials, and the fundamental model**

- The fundamental model is equivalent to infinite tosses of a fair coin (see using binary expansion of any  $x \in (0, 1)$  if you want as suggested in optional Exercise 2.1 on intuiting a most primitive sigma-algebra)
- The fundamental model has infinitely many copies of itself within it! You can see this since its DF  $F$  is the identity function on  $[0, 1]$  or equivalently how the dyadic binary tree is identical below a given node in the tree no matter which node in the tree you choose.

**\*\*universality of the fundamental model**

- one can obtain any other random variable from the fundamental model whose unique DF is its own inverse, i.e.,  $F(x) = F^{[-1]}(x)$ , as you will See from von Neumann's Fundamental Theorem of Simulation in Chapter 4.

### 3.4.2 Some Common Continuous Random Variables

Let us warm-up with an example.

**Example 55** Let  $X$  have density function  $f(x) = e^{-x}$ , if  $x \geq 0$ , and zero otherwise.

- (a) Find the distribution function.
- (b) Find the probabilities,  $P(\frac{1}{4} \leq X \leq 2)$  and  $P(-\frac{1}{2} \leq X \leq \frac{1}{2})$ .
- (c) Find  $x$  such that  $P(X \leq x) = 0.95$ .

Solution:

(a)

$$F(x) = \int_0^x e^{-v} dv = -e^{-v}]_0^x = -e^{-x} + 1 = 1 - e^{-x} \quad \text{if } x \geq 0$$

Therefore,

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

(b)

$$P\left(\frac{1}{4} \leq X \leq 2\right) = F(2) - F\left(\frac{1}{4}\right) = 0.634 \text{ (3 sig. fig.)}$$

$$P\left(-\frac{1}{2} \leq X \leq \frac{1}{2}\right) = F\left(\frac{1}{2}\right) - F\left(-\frac{1}{2}\right) = 0.394 \text{ (3 sig. fig.)}$$

(c)

$$P(X \leq x) = F(x) = 1 - e^{-x} = 0.95$$

Therefore,

$$x = -\log(1 - 0.95) = 3.00 \text{ (3 sig. fig.) .}$$

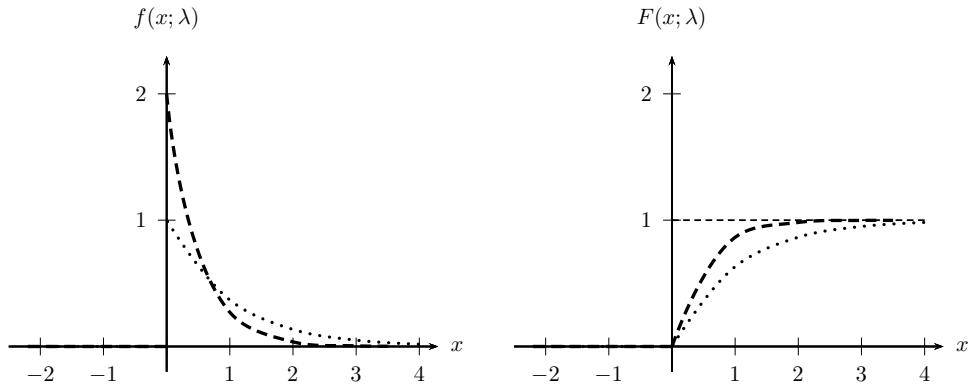


Figure 3.12:  $f(x; \lambda)$  and  $F(x; \lambda)$  of an exponential random variable where  $\lambda = 1$  (dotted) and  $\lambda = 2$  (dashed).

The previous example is a special case of the following parametric family of random variables.

**Model 8** (Exponential( $\lambda$ )) For a given  $\lambda > 0$ , an Exponential( $\lambda$ ) RV has the following PDF  $f$  and DF  $F$  and its complementary distribution function denoted by  $\bar{F}(x; \lambda) := P(X > x) = 1 - F(x; \lambda)$ :

$$f(x; \lambda) = \mathbb{1}_{(0, \infty)} \lambda e^{-\lambda x} = \begin{cases} \lambda \exp(-\lambda x) & x > 0 , \\ 0 & \text{otherwise} , \end{cases} \quad (3.27)$$

$$F(x; \lambda) = 1 - e^{-\lambda x} , \quad (3.28)$$

$$\bar{F}(x; \lambda) = e^{-\lambda x} . \quad (3.29)$$

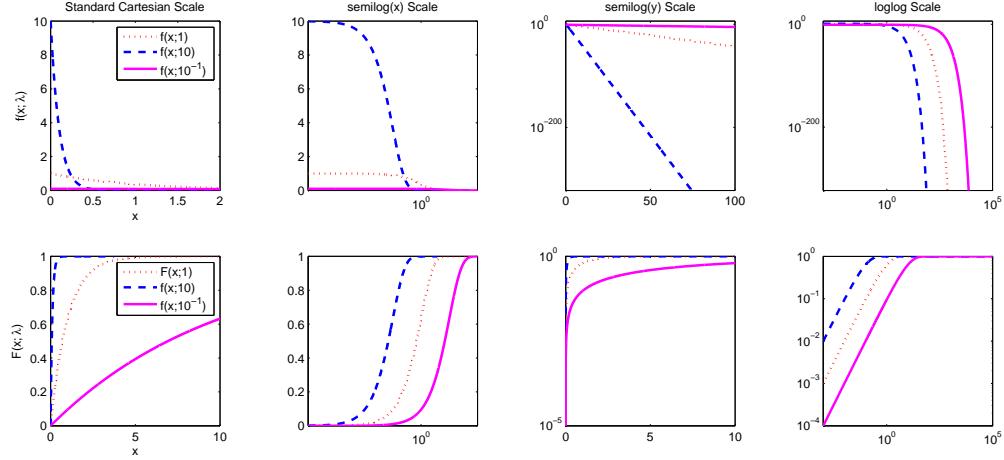
The last two equations are derived from definitions as follows:

$$\begin{aligned} F(x; \lambda) &= \int_{-\infty}^x \mathbb{1}_{(0, \infty)} \lambda e^{-\lambda v} dv = \lambda \int_0^x e^{-\lambda v} dv = \lambda \left( -\frac{1}{\lambda} e^{-\lambda v} \right)_0^x = \left( -e^{-\lambda v} \right)_0^x \\ &= -e^{-\lambda x} - (-e^{-0}) = -e^{-\lambda x} - (-1/e^0) = -e^{-\lambda x} - (-1/1) = -e^{-\lambda x} - (-1) = -e^{-\lambda x} + 1 \end{aligned}$$

$$P(X > x) = 1 - P(X \leq x) = 1 - F(x; \lambda) = 1 - \left( 1 - e^{-\lambda x} \right) = e^{-\lambda x}$$

This distribution is unique because of its property of **memorylessness**, i.e.,  $P(X > x+y | X > y) = e^{-\lambda x}$ , and plays a fundamental role in modeling continuous time processes, such as time between occurrence of events of interest, as we will see in the sequel.

Figure 3.13: Density and distribution functions of  $\text{Exponential}(\lambda)$  RVs, for  $\lambda = 1, 10, 10^{-1}$ , in four different axes scales.



**Example 56 (On a dark desert highway)** At a certain location on a dark desert highway, the time in minutes between arrival of cars that exceed the speed limit is an  $\text{Exponential}(\lambda = 1/60)$  random variable. If you just saw a car that exceeded the speed limit then what is the probability of waiting less than 5 minutes before seeing another car that will exceed the speed limit?

*Solution:*

The waiting time in minutes is simply given by the  $\text{Exponential}(\lambda = 1/60)$  random variable. Thus, the desired probability is

$$P(0 \leq X < 5) = \int_0^5 \frac{1}{60} e^{-\frac{1}{60}x} dx = -e^{-\frac{1}{60}x} \Big|_0^5 = -e^{-\frac{1}{12}} + 1 \approx 0.07996.$$

In exam you can stop at the expression  $-e^{-\frac{1}{12}} + 1$  for full credit. You may need a calculator for the last step (with answer 0.07996).

Note: We could use the distribution function directly:

$$P(0 \leq X < 5) = F\left(5; \frac{1}{60}\right) - F\left(0; \frac{1}{60}\right) = F\left(5; \frac{1}{60}\right) = 1 - e^{-\frac{1}{60}5} = 1 - e^{-\frac{1}{12}} \approx 0.07996$$

**Proposition 26 (Memorylessness of  $\text{Exponential}(\lambda)$  RV)** If  $X \sim \text{Exponential}(\lambda)$ , then  $X$  has the property of **memorylessness**, i.e.,

$$\boxed{P(X > x + y | X > y) = P(X > x)} . \quad (3.30)$$

**Proof:** By the definition of conditional probability,

$$P(X > x + y | X > y) = P(\{X > x + y\} | \{X > y\}) = \frac{P(\{X > x + y\} \cap \{X > y\})}{P(\{X > y\})}$$

Due to redundancy, i.e.,  $\{X > x + y\} \subset \{X > y\} \implies \{X > x + y\} \cap \{X > y\} = \{X > x + y\}$ , so

$$\begin{aligned} P(X > x + y | X > y) &= \frac{P(\{X > x + y\})}{P(\{X > y\})} = \frac{\bar{F}(x + y; \lambda)}{\bar{F}(y; \lambda)} = \frac{e^{-\lambda(x+y)}}{e^{-\lambda y}} = \frac{e^{-\lambda x} e^{-\lambda y}}{e^{-\lambda y}} = e^{-\lambda x} = \bar{F}(x; \lambda) \\ &= P(X > x) \end{aligned}$$

**Exercise 3.17 (Memoryless Server Times)** Suppose customers in a Queue are served one at a time by a server whose service time is an independent and identical Exponential( $\lambda$ ) RV, with  $\lambda = 1/10$ . The server is immediately free to serve the next customer once the current customer being served is done. Suppose you just arrive and are the first in the queue and know that the server is busy serving another customer. You do not know how long the customer has already been in service. What is the probability that the server will be free after 2 units of time?

Let us introduce parameters for the lower and upper bounds of the interval upon which a continuous RV is uniformly distributed using the following probability model.

**Model 9 (Uniform( $\theta_1, \theta_2$ ))** Given two real parameters  $\theta_1, \theta_2 \in \mathbb{R}$ , such that  $\theta_1 < \theta_2$ , the PDF of the  $Uniform(\theta_1, \theta_2)$  RV  $X$  is:

$$f(x; \theta_1, \theta_2) = \begin{cases} \frac{1}{\theta_2 - \theta_1} & \text{if } \theta_1 \leq x \leq \theta_2, \\ 0 & \text{otherwise} \end{cases} \quad (3.31)$$

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

$$F(x; \theta_1, \theta_2) = \begin{cases} 0 & \text{if } x < \theta_1 \\ \frac{x - \theta_1}{\theta_2 - \theta_1} & \text{if } \theta_1 \leq x \leq \theta_2, \\ 1 & \text{if } x > \theta_2 \end{cases} \quad (3.32)$$

Recall that we emphasise the dependence of the probabilities on the two parameters  $\theta_1$  and  $\theta_2$  by specifying them following the semicolon in the argument for  $f$  and  $F$ .

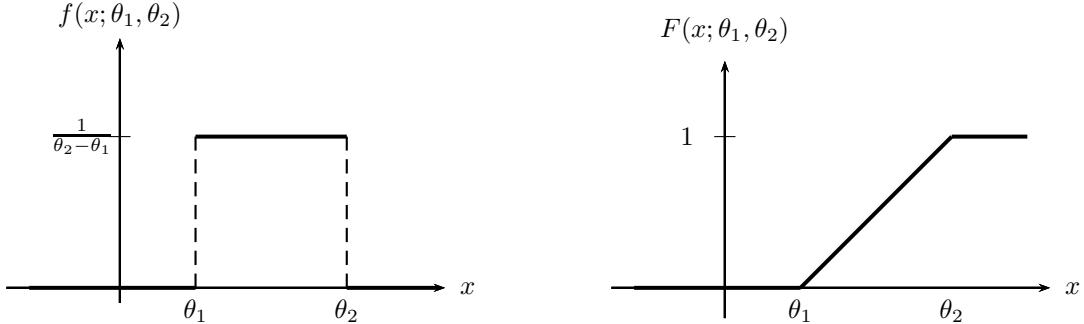


Figure 3.14:  $f(x)$  and  $F(x)$  of the  $Uniform(\theta_1, \theta_2)$  random variable  $X$ .

**Exercise 3.18** Consider a random variable with a probability density function

$$f(x) = \begin{cases} k & \text{if } 2 \leq x \leq 6, \\ 0 & \text{otherwise} \end{cases}$$

- (a) Find the value of  $k$ .
- (b) Sketch the graphs of  $f(x)$  and  $F(x)$ .

The standard normal distribution is the most important continuous probability distribution. It was first described by De Moivre in 1733 and subsequently by C. F. Gauss (1777 - 1885). Many random variables have a normal distribution, or they are approximately normal, or can be transformed into normal random variables in a relatively simple fashion. Furthermore, the normal distribution is a useful approximation of more complicated distributions.

**Model 10 (Normal(0, 1) or standard normal or Gaussian RV)** A continuous random variable  $Z$  is called **standard normal or standard Gaussian** if its probability density function is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right). \quad (3.33)$$

An exercise in calculus yields the first two derivatives of  $\phi$  as follows:

$$\frac{d\phi}{dz} = -\frac{1}{\sqrt{2\pi}} z \exp\left(-\frac{z^2}{2}\right) = -z\phi(z), \quad \frac{d^2\phi}{dz^2} = \frac{1}{\sqrt{2\pi}} (z^2 - 1) \exp\left(-\frac{z^2}{2}\right) = (z^2 - 1)\phi(z).$$

Thus,  $\phi$  has a global maximum at 0, it is concave down if  $z \in (-1, 1)$  and concave up if  $z \in (-\infty, -1) \cup (1, \infty)$ . This shows that the graph of  $\phi$  is shaped like a smooth symmetric bell centred at the origin over the real line.

**Classwork 57** From the above exercise in calculus let us draw the graph of  $\phi$  by hand now!

Do it step by step:  $z^2$ ,  $-z^2$ ,  $-z^2/2$ ,  $\exp(-z^2/2)$ ,  $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$  now!

The distribution function of  $Z$  is given by

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-v^2/2} dv. \quad (3.34)$$

**Remark 27** The integral for  $\Phi(z)$  has no closed form expression and cannot be evaluated exactly by standard methods of calculus, but its values can be obtained numerically and tabulated. Values of  $\Phi(z)$  are tabulated in the “Standard Normal Distribution Function Table” in Sec. 6.6.

We can express  $\Phi(z)$  in terms of the error function (erf) as follows:

$$\Phi(z) = \frac{1}{2} \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right) + \frac{1}{2} \quad (3.35)$$

And use MATLAB’s `erf` function to get  $\Phi(z)$  numerically instead of looking up the Table.

**Classwork 58** Note that the curve of  $\Phi(z)$  is *S*-shaped, increasing in a strictly monotone way from 0 at  $-\infty$  to 1 at  $\infty$ , and intersects the vertical axis at  $1/2$ . Draw this by hand too.

just do it!

**Example 59** Find the probabilities, using normal tables, that a random variable having the standard normal distribution will take on a value:

(a) less than 1.72

(c) between 1.30 and 1.75

(b) less than -0.88

(d) between -0.25 and 0.45

(a)

$$P(Z < 1.72) = \Phi(1.72) = 0.9573$$

(b) First note that  $P(Z < 0.88) = 0.8106$ , so that

$$\begin{aligned} P(Z < -0.88) &= P(Z > 0.88) \\ &= 1 - P(Z < 0.88) \\ &= 1 - \Phi(0.88) \\ &= 1 - 0.8106 = 0.1894 \end{aligned}$$

(c)  $P(1.30 < Z < 1.75) = \Phi(1.75) - \Phi(1.30) = 0.9599 - 0.9032 = 0.0567$

(d)

$$\begin{aligned} P(-0.25 < Z < 0.45) &= P(Z < 0.45) - P(Z < -0.25) \\ &= P(Z < 0.45) - (1 - P(Z < 0.25)) \\ &= \Phi(0.45) - (1 - \Phi(0.25)) \\ &= (0.6736) - (1 - 0.5987) \\ &= 0.2723 \end{aligned}$$

#### CONTINUOUS RANDOM VARIABLES: NOTATION

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

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

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

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

### 3.5 Exercises in Continuous Random Variables

**Ex. 3.19** — Consider the probability density function

$$f(x) = \begin{cases} k & -4 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases} .$$

1. Find the value of  $k$ .
2. Find the distribution function,  $F$ .
3. Graph  $f$  and  $F$ .

**Ex. 3.20** — Assume that a new light bulb will burn out at time  $t$  hours according to the probability density function given by

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

In this context,  $\lambda$  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  $\tau$  hours.  
This  $\tau$ -specific probability is often called the reliability of the bulb.  
Hint: Use the distribution function for an Exponential( $\lambda$ ) random variable (recall,  $F(\tau; \lambda) = \int_{-\infty}^{\tau} f(t)dt$ !)
- (b) For what value of  $\tau$  is the reliability of the bulb exactly  $\frac{1}{2}$ ?

**Ex. 3.21** — Let the random variable  $X$  be the time after which certain ball bearings wear out, with density

$$f(x) = \begin{cases} ke^{-x} & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases} .$$

Note:  $X$  is measured in years.

1. Find  $k$ .
2. Find the probability that a bearing will last at least 1 year.

### 3.6 Transformations of random variables

Suppose we know the distribution of a random variable  $X$ . How do we find the distribution of a transformation of  $X$ , say  $g(X)$ ? Before we answer this question let us ask a motivational question. Why are we interested in functions of random variables?

**Example 60** Consider a simple financial example where an individual sells  $X$  items per day, the profit per item is \$5 and the overhead costs are \$500 per day. The original random variable is  $X$ , but the random variable  $Y$  which gives the daily profit is of more interest, where

$$Y = 5X - 500 .$$

**Example 61** In a cell-phone system a mobile signal may have a signal-to-noise-ratio of  $X$ , but engineers prefer to express such ratios in decibels, i.e.,

$$Y = 10 \log_{10}(X) .$$

### 3.6.1 A Review of Inverse Images

Hence in a great many situations we are more interested in functions of random variables. Let us return to our original question of determining the distribution of a transformation or function of  $X$ . First note that this transformation of  $X$  is itself another random variable, say  $Y = g(X)$ , where  $g$  is a function from a subset  $\mathbb{X}$  of  $\mathbb{R}$  to a subset  $\mathbb{Y}$  of  $\mathbb{R}$ , i.e.,  $g : \mathbb{X} \rightarrow \mathbb{Y}$ ,  $\mathbb{X} \subset \mathbb{R}$  and  $\mathbb{Y} \subset \mathbb{R}$ .

The **inverse image** of a set  $A$  is the set of all real numbers in  $\mathbb{X}$  whose image is in  $A$ , i.e.,

$$g^{[-1]}(A) = \{x \in \mathbb{X} : g(x) \in A\} .$$

In other words,

$$x \in g^{[-1]}(A) \text{ if and only if } g(x) \in A .$$

For example,

- if  $g(x) = 2x$  then  $g^{[-1]}([4, 6]) = [2, 3]$
- if  $g(x) = 2x + 1$  then  $g^{[-1]}([5, 7]) = [2, 3]$
- if  $g(x) = x^3$  then  $g^{[-1]}([1, 8]) = [1, 2]$
- if  $g(x) = x^2$  then  $g^{[-1]}([1, 4]) = [-2, -1] \cup [1, 2]$
- if  $g(x) = \sin(x)$  then  $g^{[-1]}([-1, 1]) = \mathbb{R}$
- if ...

For the singleton set  $A = \{y\}$ , we write  $g^{[-1]}(y)$  instead of  $g^{[-1]}(\{y\})$ . For example,

- if  $g(x) = 2x$  then  $g^{[-1]}(4) = \{2\}$
- if  $g(x) = 2x + 1$  then  $g^{[-1]}(7) = \{3\}$
- if  $g(x) = x^3$  then  $g^{[-1]}(8) = \{2\}$
- if  $g(x) = x^2$  then  $g^{[-1]}(4) = \{-2, 2\}$
- if  $g(x) = \sin(x)$  then  $g^{[-1]}(0) = \{k\pi : k \in \mathbb{Z}\} = \{\dots, -3\pi, -2\pi, -\pi, 0, \pi, 2\pi, 3\pi, \dots\}$
- if ...

If  $g : \mathbb{X} \rightarrow \mathbb{Y}$  is one-to-one (injective) and onto (surjective), then the inverse image of a singleton set is itself a singleton set. Thus, the inverse image of such a function  $g$  becomes itself a function and is called the **inverse function**. One can find the inverse function, if it exists by the following steps:

Step 1; write  $y = g(x)$

Step 2; solve for  $x$  in terms of  $y$

Step 3; set  $g^{-1}(y)$  to be this solution

We write  $g^{-1}$  whenever the inverse image  $g^{[-1]}$  exists as an inverse function of  $g$ . Thus, the inverse function  $g^{-1}$  is a specific type of inverse image  $g^{[-1]}$ . For example,

- if  $g(x) = 2x$  then  $g : \mathbb{R} \rightarrow \mathbb{R}$  is injective and surjective and therefore its inverse function is:  
Step 1;  $y = 2x$ , Step 2;  $x = \frac{y}{2}$ , Step 3;  $g^{-1}(y) = \frac{y}{2}$

- if  $g(x) = 2x + 1$  then  $g : \mathbb{R} \rightarrow \mathbb{R}$  is injective and surjective and therefore its inverse function is:  
Step 1;  $y = 2x + 1$ , Step 2;  $x = \frac{y-1}{2}$ , Step 3;  $g^{-1}(y) = \frac{y-1}{2}$
- if  $g(x) = x^3$  then  $g : \mathbb{R} \rightarrow \mathbb{R}$  is injective and surjective and therefore its inverse function is:  
Step 1;  $y = x^3$ , Step 2;  $x = y^{\frac{1}{3}}$ , Step 3;  $g^{-1}(y) = y^{\frac{1}{3}}$

However, you need to be careful by limiting the domain to obtain the inverse function for the following examples:

- if  $g(x) = x^2$  and domain of  $g$  is  $[0, +\infty)$  then its inverse function is  $g^{-1}(y) = \sqrt{y}$ , i.e., if  $g(x) = x^2 : [0, +\infty) \rightarrow [0, +\infty)$  then the inverse image  $g^{[-1]}(y)$  for  $y \in [0, +\infty)$  is given by the inverse function  $g^{-1}(y) = \sqrt{y} : [0, +\infty) \rightarrow [0, +\infty)$ .
- if  $g(x) = x^2$  and domain of  $g$  is  $(-\infty, 0]$  then its inverse function is  $g^{-1}(y) = -\sqrt{y}$ , i.e., if  $g(x) = x^2 : (-\infty, 0] \rightarrow [0, +\infty)$  then the inverse image  $g^{[-1]}(y)$  for  $y \in [0, +\infty)$  is given by the inverse function  $g^{-1}(y) = -\sqrt{y} : [0, +\infty) \rightarrow (-\infty, 0]$ .
- if  $g(x) = \sin(x)$  and domain of  $g$  is  $[0, \frac{\pi}{2}]$  then its inverse function  $g^{-1}(y) = \arcsin(y)$ , i.e., if  $g(x) = \sin(x) : [0, \frac{\pi}{2}] \rightarrow [0, 1]$  then the inverse image  $g^{[-1]}(y)$  for  $y \in [0, 1]$  is given by the inverse function  $g^{-1}(y) = \arcsin(y) : [0, 1] \rightarrow [0, \frac{\pi}{2}]$ .
- if  $g(x) = \sin(x)$  and domain of  $g$  is  $[-\frac{\pi}{2}, \frac{\pi}{2}]$  then its inverse function  $g^{-1}(y) = \arcsin(y)$ , i.e., if  $g(x) = \sin(x) : [-\frac{\pi}{2}, \frac{\pi}{2}] \rightarrow [-1, 1]$  then the inverse image  $g^{[-1]}(y)$  for  $y \in [-1, 1]$  is given by the inverse function  $g^{-1}(y) = \arcsin(y) : [-1, 1] \rightarrow [-\frac{\pi}{2}, \frac{\pi}{2}]$ .
- if ...

Now, let us return to our question of determining the distribution of the transformation  $g(X)$ . To answer this question we must first observe that the inverse image  $g^{[-1]}$  satisfies the following properties:

- $g^{[-1]}(\mathbb{Y}) = \mathbb{X}$
- For any set  $A$ ,  $g^{[-1]}(A^c) = (g^{[-1]}(A))^c$
- For any collection of sets  $\{A_1, A_2, \dots\}$ ,

$$g^{[-1]}(A_1 \cup A_2 \cup \dots) = g^{[-1]}(A_1) \cup g^{[-1]}(A_2) \cup \dots .$$

Consequently,

$$P(g(X) \in A) = P\left(X \in g^{[-1]}(A)\right) \quad (3.36)$$

satisfies the axioms of probability and gives the desired probability of the event  $A$  from the transformation  $Y = g(X)$  in terms of the probability of the event given by the inverse image of  $A$  underpinned by the random variable  $X$ . It is crucial to understand this from the sample space  $\Omega$  of the underlying experiment in the sense that Equation (3.36) is just short-hand for its actual meaning:

$$P(\{\omega \in \Omega : g(X(\omega)) \in A\}) = P\left(\left\{\omega \in \Omega : X(\omega) \in g^{[-1]}(A)\right\}\right) .$$

Because we have more than one random variable to consider, namely,  $X$  and its transformation  $Y = g(X)$  we will subscript the probability density or mass function and the distribution function by the random variable itself. For example we denote the distribution function of  $X$  by  $F_X(x)$  and that of  $Y$  by  $F_Y(y)$ .

### 3.6.2 Transformations of discrete random variables

For a discrete random variable  $X$  with probability mass function  $f_X$  we can obtain the probability mass function  $f_Y$  of  $Y = g(X)$  using Equation (3.36) as follows:

$$\begin{aligned} f_Y(y) &= P(Y = y) = P(Y \in \{y\}) \\ &= P(g(X) \in \{y\}) = P\left(X \in g^{[-1]}\(\{y\}\)\right) \\ &= P\left(X \in g^{[-1]}(y)\right) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{x: g(x)=y\}} f_X(x) . \end{aligned}$$

This gives the formula:

$f_Y(y) = P(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{x: g(x)=y\}} f_X(x) .$		(3.37)
---	--	--------

**Example 62** Let  $X$  be the discrete random variable with probability mass function  $f_X$  as tabulated below:

x	-1	0	1
$f_X(x) = P(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

If  $Y = 2X$  then the transformation  $g(X) = 2X$  has inverse image  $g^{[-1]}(y) = \{y/2\}$ . Then, by Equation (3.37) the probability mass function of  $Y$  is expressed in terms of the known probabilities of  $X$  as:

$$f_Y(y) = P(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{y/2\}} f_X(x) = f_X(y/2) ,$$

and tabulated below:

y	-2	0	2
$f_Y(y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

**Example 63** If  $X$  is the random variable in the previous Example then what is the probability mass function of  $Y = 2X + 1$ ? Once again,

$$f_Y(y) = P(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{(y-1)/2\}} f_X(x) = f_X((y-1)/2) ,$$

and tabulated below:

y	-1	1	3
$f_Y(y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

In fact, obtaining the probability of a one-to-one transformation of a discrete random variable as in Examples 62 and 63 is merely a matter of looking up the probability at the image of the inverse function. This is because there is only one term in the sum that appears in Equation (3.37). When the transformation is not one-to-one the number of terms in the sum can be more than one as shown in the next Example.

**Example 64** Reconsider the random variable  $X$  of the last two Examples and let  $Y = X^2$ . Recall that  $g(x) = x^2$  does not have an inverse function unless the domain is restricted to the positive or the negative parts of the real line. Since our random variable  $X$  takes values on both sides of the real line, namely  $\{-1, 0, 1\}$ , let us note that the transformation  $g(X) = X^2$  is no longer a one-to-one function. Then, by Equation (3.37) the probability mass function of  $Y$  is expressed in terms of the known probabilities of  $X$  as:

$$f_Y(y) = P(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{\{x: g(x)=y\}} f_X(x) = \sum_{\{x: x^2=y\}} f_X(x) ,$$

computed for each  $y \in \{0, 1\}$  as follows:

$$\begin{aligned} f_Y(0) &= \sum_{\{x: x^2=0\}} f_X(x) = f_X(0) = \frac{1}{2} , \\ f_Y(1) &= \sum_{\{x: x^2=1\}} f_X(x) = f_X(-1) + f_X(1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} , \end{aligned}$$

and finally tabulated below:

$y$	0	1	
$f_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$	

### 3.6.3 Transformations of continuous random variables

Suppose we know  $F_X$  and/or  $f_X$  of a continuous random variable  $X$ . Let  $Y = g(X)$  be a transformation of  $X$ . Our objective is to obtain  $F_Y$  and/or  $f_Y$  of  $Y$  from  $F_X$  and/or  $f_X$ .

#### One-to-one transformations

The easiest case for transformations of continuous random variables is when  $g$  is **one-to-one and monotone**.

- First, let us consider the case when  $g$  is **monotone and increasing** on the range of the random variable  $X$ . In this case  $g^{-1}$  is also an increasing function and we can obtain the distribution function of  $Y = g(X)$  in terms of the distribution function of  $X$  as

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) .$$

Now, let us use a form of chainrule to compute the density of  $Y$  as follows:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \frac{d}{dy}(g^{-1}(y)) .$$

- Second, let us consider the case when  $g$  is **monotone and decreasing** on the range of the random variable  $X$ . In this case  $g^{-1}$  is also a decreasing function and we can obtain the distribution function of  $Y = g(X)$  in terms of the distribution function of  $X$  as

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)) ,$$

and the density of  $Y$  as

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} (1 - F_X(g^{-1}(y))) = -f_X(g^{-1}(y)) \frac{d}{dy} (g^{-1}(y)) .$$

For a monotonic and decreasing  $g$ , its inverse function  $g^{-1}$  is also decreasing and consequently the density  $f_Y$  is indeed positive because  $\frac{d}{dy} (g^{-1}(y))$  is negative.

We can combine the above two cases and obtain the following **change of variable formula** for the probability density of  $Y = g(X)$  when  $g$  is one-to-one and monotone on the range of  $X$ .

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| .$$

(3.38)

The steps involved in finding the density of  $Y = g(X)$  for a one-to-one and monotone  $g$  are:

1. Write  $y = g(x)$  for  $x$  in range of  $x$  and check that  $g(x)$  is monotone over the required range to apply the change of variable formula.
2. Write  $x = g^{-1}(y)$  for  $y$  in range of  $y$ .
3. Obtain  $\left| \frac{d}{dy} g^{-1}(y) \right|$  for  $y$  in range of  $y$ .
4. Finally, from Equation (3.38) get  $f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$  for  $y$  in range of  $y$ .

Let us use these four steps to obtain the density of monotone transformations of continuous random variables.

**Example 65** Let  $X$  be Uniform(0, 1) random variable and let  $Y = g(X) = 1 - X$ . We are interested in the density of the transformed random variable  $Y$ . Let us follow the four steps and use the change of variable formula to obtain  $f_Y$  from  $f_X$  and  $g$ .

1.  $y = g(x) = 1 - x$  is a monotone decreasing function over  $0 \leq x \leq 1$ , the range of  $X$ . So, we can apply the change of variable formula.
  2.  $x = g^{-1}(y) = 1 - y$  is a monotone decreasing function over  $1 - 0 \geq 1 - x \geq 1 - 1$ , i.e.,  $0 \leq y \leq 1$ .
  3. For  $0 \leq y \leq 1$ ,
- $$\left| \frac{d}{dy} g^{-1}(y) \right| = \left| \frac{d}{dy} (1 - y) \right| = |-1| = 1 .$$
4. we can use Equation (3.38) to find the density of  $Y$  as follows:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| = f_X(1 - y) 1 = 1 ,$$

for  $0 \leq y \leq 1$

Thus, we have shown that if  $X$  is a Uniform(0, 1) random variable then  $Y = 1 - X$  is also a Uniform(0, 1) random variable.

**Example 66** Let  $X$  be a Uniform(0, 1) random variable and let  $Y = g(X) = -\log(X)$ . We are interested in the density of the transformed random variable  $Y$ . Once again, since  $g$  is a one-to-one monotone function let us follow the four steps and use the change of variable formula to obtain  $f_Y$  from  $f_X$  and  $g$ .

1.  $y = g(x) = -\log(x)$  is a monotone decreasing function over  $0 < x < 1$ , the range of  $X$ . So, we can apply the change of variable formula.
2.  $x = g^{-1}(y) = \exp(-y)$  is a monotone decreasing function over  $0 < y < \infty$ .
3. For  $0 < y < \infty$ ,

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

4. We can use Equation (3.38) to find the density of  $Y$  as follows:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| = f_X(\exp(-y)) \exp(-y) = 1 \exp(-y) = \exp(-y) .$$

Note that  $0 < \exp(-y) < 1$  for  $0 < y < \infty$ .

Thus, we have shown that if  $X$  is a Uniform(0, 1) random variable then  $Y = -\log(X)$  is a random variable with PDF  $f_Y(y) = \mathbf{1}_{(0,\infty)}(y) \exp(-y)$ . We can similarly show that for a parameter  $\lambda > 0$ , if  $X \sim \text{Uniform}(0, 1)$  then  $Y = -\lambda^{-1} \log(X)$  yields a probability model of RVs that are parameterized by  $\lambda$  and extremely useful in applications. This is noting but our Exponential( $\lambda$ ) RV.

The next example yields the *location-scale* family of normal random variables via a family of linear transformations of the standard normal random variable.

**Example 67** Let  $Z$  be the standard Gaussian or standard normal random variable with probability density function  $\phi(z)$  given by Equation (3.33). For real numbers  $\sigma > 0$  and  $\mu$  consider the linear transformation of  $Z$  given by

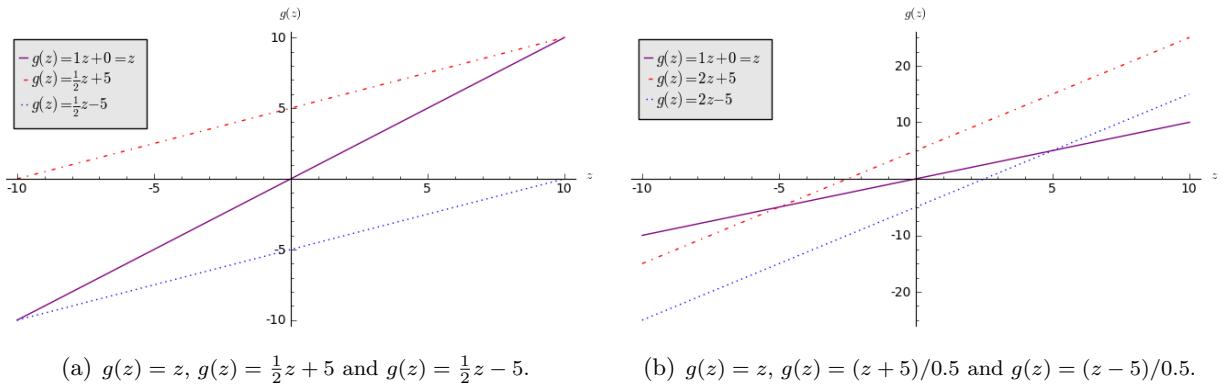
$$Y = g(Z) = \sigma Z + \mu .$$

Some graphs of such linear transformations of  $Z$  are shown in Figures (a) and (b).

We are interested in the density of the transformed random variable  $Y = g(Z) = \sigma Z + \mu$ . Once again, since  $g$  is a one-to-one monotone function let us follow the four steps and use the change of variable formula to obtain  $f_Y$  from  $f_Z = \phi$  and  $g$ .

1.  $y = g(z) = \sigma z + \mu$  is a monotone increasing function over  $-\infty < z < \infty$ , the range of  $Z$ . So, we can apply the change of variable formula.
2.  $z = g^{-1}(y) = (y - \mu)/\sigma$  is a monotone increasing function over the range of  $y$  given by,  $-\infty < y < \infty$ .
3. For  $-\infty < y < \infty$ ,

$$\left| \frac{d}{dy} g^{-1}(y) \right| = \left| \frac{d}{dy} \left( \frac{y - \mu}{\sigma} \right) \right| = \left| \frac{1}{\sigma} \right| = \frac{1}{\sigma} .$$



4. we can use Equation (3.38) and Equation (3.33) which gives

$$f_Z(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) ,$$

to find the density of  $Y$  as follows:

$$f_Y(y) = f_Z(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| = \phi\left(\frac{y - \mu}{\sigma}\right) \frac{1}{\sigma} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{y - \mu}{\sigma}\right)^2\right] ,$$

for  $-\infty < y < \infty$ .

Thus, we have obtained the expression for the probability density function of the linear transformation  $\sigma Z + \mu$  of the standard normal random variable  $Z$ . This analysis leads to the following definition.

**Model 11** (Normal( $\mu, \sigma^2$ ) RV) Given a location parameter  $\mu \in (-\infty, +\infty)$  and a scale parameter  $\sigma^2 > 0$ , the Normal( $\mu, \sigma^2$ ) or Gaussian( $\mu, \sigma^2$ ) random variable  $X$  has probability density function:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2\right] \quad (\sigma > 0) . \quad (3.39)$$

This is simpler than it may at first look.  $f(x; \mu, \sigma^2)$  has the following features.

- $\mu$  is the expected value or mean parameter and  $\sigma^2$  is the variance parameter. These concepts, mean and variance, are described in more detail in the next section on expectations.
- $1/(\sigma\sqrt{2\pi})$  is a constant factor that makes the area under the curve of  $f(x)$  from  $-\infty$  to  $\infty$  equal to 1, as it must be.
- The curve of  $f(x)$  is symmetric with respect to  $x = \mu$  because the exponent is quadratic. Hence for  $\mu = 0$  it is symmetric with respect to the  $y$ -axis  $x = 0$ .
- The exponential function decays to zero very fast — the faster the decay, the smaller the value of  $\sigma$ .

The normal distribution has the **distribution function**

$$F(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left[-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right] dv . \quad (3.40)$$

Here we need  $x$  as the upper limit of integration and so we write  $v$  in the integrand.

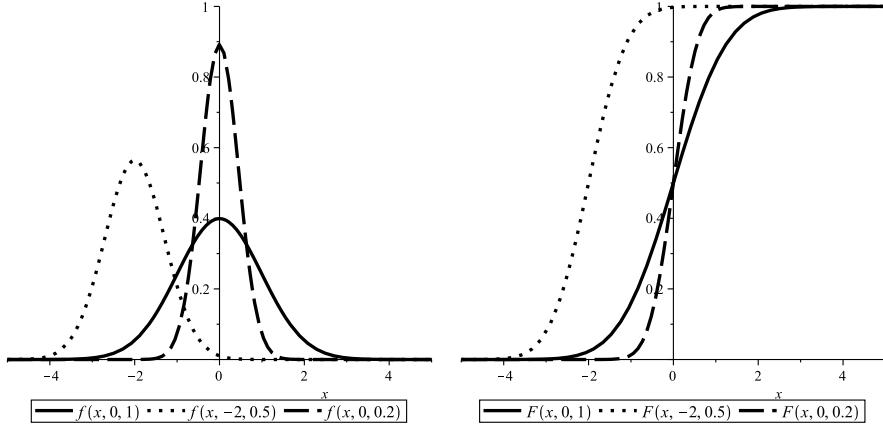


Figure 3.15: PDF and DF of a  $\text{Normal}(\mu, \sigma^2)$  RV for different values of  $\mu$  and  $\sigma^2$

Using the direct method's Equation 3.41, we can obtain the distribution function of the  $\text{Normal}(\mu, \sigma^2)$  random variable from that of the tabulated distribution function of the  $\text{Normal}(0, 1)$  in the Standard normal distribution function table in Sec. 6.6.

**Proposition 28 (One Table to Rule Them All Gaussians)** The distribution function  $F_X(x; \mu, \sigma^2)$  of the  $\text{Normal}(\mu, \sigma^2)$  random variable  $X$  and the distribution function  $F_Z(z) = \Phi(z)$  of the standard normal random variable  $Z$  are related by:

$$F_X(x; \mu, \sigma^2) = F_Z\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right) .$$

**Proof:** Let  $Z$  be a  $\text{Normal}(0, 1)$  random variable with distribution function  $\Phi(z) = P(Z \leq z)$ . We know that if  $X = g(Z) = \sigma Z + \mu$  then  $X$  is the  $\text{Normal}(\mu, \sigma^2)$  random variable. Therefore,

$$\begin{aligned} F_X(x; \mu, \sigma^2) &= P(X \leq x) = P(g(Z) \leq x) = P(\sigma Z + \mu \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right) . \end{aligned}$$

Hence we often transform a general  $\text{Normal}(\mu, \sigma^2)$  random variable,  $X$ , to a standardised  $\text{Normal}(0, 1)$  random variable,  $Z$ , by the substitution:

$$Z = \frac{X - \mu}{\sigma} .$$

**Example 68** Suppose that the amount of cosmic radiation to which a person is exposed when flying by jet across the United States is a random variable,  $X$ , having a normal distribution with a mean of 4.35 mrem and a standard deviation of 0.59 mrem. What is the probability that a person will be exposed to more than 5.20 mrem of cosmic radiation on such a flight?

*Solution:*

$$\begin{aligned}
 P(X > 5.20) &= 1 - P(X \leq 5.20) \\
 &= 1 - F(5.20) \\
 &= 1 - \Phi\left(\frac{5.20 - 4.35}{0.59}\right) \\
 &= 1 - \Phi(1.44) \\
 &= 1 - 0.9251 \\
 &= 0.0749
 \end{aligned}$$

After some more notions you will see that  $\text{Normal}(0, 1)$  RV can actually be obtained from an IID process of  $\text{Bernoulli}(\theta)$  RVs. This is an instance of the central limit theorem. To appreciate this we first need to understand what we mean by statistics and then familiarise ourselves with notions of convergence of random variables.

### Direct method

If the transformation  $g$  in  $Y = g(X)$  is not necessarily one-to-one then special care is needed to obtain the distribution function or density of  $Y$ . For a continuous random variable  $X$  with a known distribution function  $F_X$  we can obtain the distribution function  $F_Y$  of  $Y = g(X)$  using Equation (3.36) as follows:

$$\begin{aligned}
 F_Y(y) &= P(Y \leq y) = P(Y \in (-\infty, y]) \\
 &= P(g(X) \in (-\infty, y]) = P\left(X \in g^{[-1]}((-\infty, y])\right) = P(X \in \{x : g(x) \in (-\infty, y]\}) \quad (3.41)
 \end{aligned}$$

In words, the above equalities just mean that the probability that  $Y \leq y$  is the probability that  $X$  takes a value  $x$  that satisfies  $g(x) \leq y$ . We can use this approach if it is reasonably easy to find the set  $g^{[-1]}((-\infty, y]) = \{x : g(x) = (-\infty, y]\}$ .

**Example 69** Let  $X$  be any random variable with distribution function  $F_X$ . Let  $Y = g(X) = X^2$ . Then we can find  $F_Y$ , the distribution function of  $Y$  from  $F_X$  as follows:

- Since  $Y = X^2 \geq 0$ , if  $y < 0$  then  $F_Y(y) = P(X \in \{x : x^2 < y\}) = P(X \in \emptyset) = 0$ .
- If  $y \geq 0$  then

$$\begin{aligned}
 F_Y(y) = P(Y \leq y) &= P(X^2 \leq y) \\
 &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\
 &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) .
 \end{aligned}$$

By differentiation we get:

- If  $y < 0$  then  $f_Y(y) = \frac{d}{dy}(F_Y(y)) = \frac{d}{dy}0 = 0$ .
- If  $y \geq 0$  then

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

Therefore, the distribution function of  $Y = X^2$  is:

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \text{if } y \geq 0 \end{cases} . \quad (3.42)$$

and the probability density function of  $Y = X^2$  is:

$$f_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{2\sqrt{y}}(f_X(\sqrt{y}) + f_X(-\sqrt{y})) & \text{if } y \geq 0 \end{cases} . \quad (3.43)$$

**Example 70** If  $X$  is the standard normal random variable with density

$$f_X(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

then by Equation (3.43) the density of  $Y = X^2$  is:

$$f_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{2\sqrt{y}}(f_X(\sqrt{y}) + f_X(-\sqrt{y})) = \frac{1}{\sqrt{2\pi y}} \exp(-\frac{y}{2}) & \text{if } y \geq 0 \end{cases} .$$

$Y$  is called the **chi-square** random variable with one degree of freedom. This distribution plays a fundamental role in hypothesis testing as we will see in Inference Theory and was derived at the beginning of last century to settle “supposedly evidence-based disputes” among scientists using mathematics.

### 3.7 Exercises in Transformations of Random Variables

**Ex. 3.22** — Let  $X$  be the outcome of a fair die roll with probability mass function given by

$$f_X(x) = \begin{cases} \frac{1}{6} & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise.} \end{cases}$$

If  $Y = (X - 3)^2$  then find the probability mass function of  $Y$ ,  $f_Y(y)$ .

**Ex. 3.23** — Given a natural number  $n$  as a parameter, i.e., given a parameter  $n \in \{1, 2, 3, \dots\}$ , let  $X$  be a discrete uniform random variable on the finite set

$$\mathbb{X} = \{-n, -n + 1, \dots, -1, 0, 1, \dots, n - 1, n\}$$

i.e. the probability mass function of  $X$  is:

$$f_X(x; n) = \begin{cases} \frac{1}{2n+1} & \text{if } x \in \mathbb{X} \\ 0 & \text{otherwise.} \end{cases}$$

Find the probability mass function  $f_Y(y; n)$  for  $Y = |X|$ , the absolute value of  $X$ .

**Ex. 3.24** — If  $X$  is a Geometric( $\theta$ ) random variable and  $Y = (\frac{1}{2})^X$  then find an expression for  $f_Y(y)$ .

**Ex. 3.25** — If  $X$  is a Poisson( $\lambda$ ) random variable find the probability mass function,  $f_Y(y)$ , of

$$Y = \frac{1}{(X + 1)^2} .$$

**Ex. 3.26** — If  $X$  is a continuous random variable with probability density function

$$f_X(x) = \begin{cases} xe^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases},$$

find the probability density function of  $Y = e^X$ .

**Ex. 3.27** — If  $X$ , the received power at an antenna is an Exponential( $\lambda$ ) random variable then find the probability density function of the amplitude  $Y = \sqrt{X}$ .

**Ex. 3.28** — If  $X$  is a Uniform( $a, b$ ) random variable where  $0 < a < b$ , find the probability density function,  $f_Y(y)$ , of

$$Y = \log_e(X).$$

### 3.8 Expectations

Expectation is perhaps the most fundamental concept in probability theory. In fact, probability is itself an expectation as you will soon see!

Expectation is one of the fundamental concepts in probability. The expected value of a real-valued random variable gives the population mean, a measure of the centre of the distribution of the variable in some sense. Its variance measures its spread and so on.

**Definition 29 (Expectation of a RV)** The **expectation**, or **expected value**, or **mean**, or **first moment**, of a random variable  $X$ , with distribution function  $F$  and density  $f$ , is defined to be

$$E(X) := \int x dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous,} \end{cases} \quad (3.44)$$

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

$$\int |x| dF(x) < \infty. \quad (3.45)$$

Sometimes, we denote  $E(X)$  by  $E X$  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.

$$E_\theta(X) := \int x dF(x; \theta)$$

**Definition 30 (Variance of a RV)** Let  $X$  be a RV with mean or expectation  $E(X)$ . Variance of  $X$  denoted by  $V(X)$  or simply  $V X$  is

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

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

**Definition 31 ( $k$ -th moment of a RV)** We call

$$E(X^k) = \int x^k dF(x)$$

as the  $k$ -th moment of the RV  $X$  and say that the  $k$ -th moment exists when  $E(|X|^k) < \infty$ . We call the following expectation as the  $k$ -th central moment:

$$E((X - E(X))^k).$$

### 3.8.1 Expectations of functions of random variables

More generally, by taking the expected value of various functions of a random variable, we can measure many interesting features of its distribution, including spread and correlation.

**Definition 32 (Expectation of a function of a RV)** The **Expectation** of a function  $g(X)$  of a random variable  $X$  is defined as:

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

provided  $E(g(X))$  exists, i.e.,  $\int |g(x)| dF(x) < \infty$ .

The **mean** which characterises the central location of the random variable  $X$  is merely the expectation of the identity function  $g(x) = x$ :

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

Often, mean is denoted by  $\mu$ .

The **variance** which characterises the spread or the variability of the random variable  $X$  is also the expectation of the function  $g(x) = (x - \mathbb{E}(X))^2$ :

$$\mathbb{V}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \begin{cases} \sum_x (x - \mathbb{E}(X))^2 f(x) & \text{if } X \text{ is a discrete RV} \\ \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f(x) dx & \text{if } X \text{ is a continuous RV} \end{cases}$$

Often, variance is denoted by  $\sigma^2$ .

### INTUITIVELY, WHAT IS EXPECTATION?

Definition 32 gives expectation as a “weighted average” of the possible values. This is true but some intuitive idea of expectation is also helpful.

- Expectation is what you expect.

Consider tossing a fair coin. If it is heads you lose \$10. If it is tails you win \$10. What do you expect to win? Nothing. If  $X$  is the amount you win then

$$\mathbb{E}(X) = -10 \times \frac{1}{2} + 10 \times \frac{1}{2} = 0.$$

So what you expect (nothing) and the weighted average ( $\mathbb{E}(X) = 0$ ) agree.

- Expectation is a long run average.

Suppose you are able to repeat an experiment independently, over and over again. Each experiment produces one value  $x$  of a random variable  $X$ . If you take the average of the  $x$  values for a large number of trials, then this average converges to  $\mathbb{E}(X)$  as the number of trials grows. In fact, this is called the **law of large numbers**.

We can concretize the above two intuitive insights by the following two examples.

**Example 71 (Winnings on Average)** Let  $Y = r(X)$ . Then

$$\mathbb{E}(Y) = \mathbb{E}(r(X)) = \int r(x) dF(x).$$

Think of playing a game where we draw  $x \sim X$  and then I pay you  $y = r(x)$ . Then your average income is  $r(x)$  times the chance that  $X = x$ , summed (or integrated) over all values of  $x$ .

**Example 72 (Probability is an Expectation)** Let  $A$  be an event and let  $r(X) = \mathbf{1}_A(x)$ . Recall  $\mathbf{1}_A(x)$  is 1 if  $x \in A$  and  $\mathbf{1}_A(x) = 0$  if  $x \notin A$ . Then

$$\mathbb{E}(\mathbf{1}_A(X)) = \int \mathbf{1}_A(x) dF(x) = \int_A dF(x) = \mathbb{P}(X \in A) = \mathbb{P}(A) \quad (3.46)$$

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

### Expectations of functions of $\mathbb{R}^2$ -valued random variables

In the case of a single random variable we saw that its expectation gives the population mean, a measure of the center of the distribution of the variable in some sense. Similarly, by taking the expected value of various functions of a  $\mathbb{R}^2$ -valued random variable, we can measure many interesting features of its joint distribution.

**Definition 33** The **Expectation** of a function  $g(X, Y)$  of the  $\mathbb{R}^2$ -valued RV  $(X, Y)$  is defined as:

$$E(g(X, Y)) = \begin{cases} \sum_{(x,y)} g(x, y) f_{X,Y}(x, y) & \text{if } (X, Y) \text{ is a discrete RV} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy & \text{if } (X, Y) \text{ is a continuous RV} \end{cases}$$

Some typical expectations for  $\mathbb{R}^2$ -valued random variables are:

#### 1. Joint Moments

$$E(X^r Y^s)$$

When  $r = s = 1$ , we have  $E(XY)$ , the expectation of the product of two RVs.

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

If  $E(X^2) < \infty$  and  $E(Y^2) < \infty$  then  $E(|XY|) < \infty$  and  $E(|(X - E(X))(Y - E(Y))|) < \infty$ . This allows the definition of **covariance** of  $X$  and  $Y$  as

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

The same ideas naturally extend, via multiple sums and integrals, to define the expectation of functions of  $\mathbb{R}^k$ -valued random variables with  $k > 2$ .

### Viewing a deterministic real variable as a random variable

Consider the class of discrete RVs with distributions that place all probability mass on a single real number. This is the probability model for the deterministic real variable, which is often thought of as an unknown constant  $\theta \in \mathbb{R}$ .

**Model 12** (Point Mass( $\theta$ )) Given a specific point  $\theta \in \mathbb{R}$ , we say an RV  $X$  has point mass at  $\theta$  or is Point Mass( $\theta$ ) distributed if the DF is:

$$F(x; \theta) = \begin{cases} 0 & \text{if } x < \theta \\ 1 & \text{if } x \geq \theta \end{cases} \quad (3.47)$$

and the PMF is:

$$f(x; \theta) = \begin{cases} 0 & \text{if } x \neq \theta \\ 1 & \text{if } x = \theta \end{cases} \quad (3.48)$$

Thus, Point Mass( $\theta$ ) RV  $X$  is deterministic in the sense that every realisation of  $X$  is exactly equal to  $\theta \in \mathbb{R}$ . We will see that this distribution plays a central limiting role in asymptotic statistics.

**Example 73 (Mean and variance of Point Mass( $\theta$ ) RV)** Let  $X \sim \text{Point Mass}(\theta)$ . Then:

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

### 3.8.2 Properties of expectations

The following results, where  $a$  is a constant, may easily be proved using the properties of summations and integrals:

$$\boxed{\mathbb{E}(a) = a}$$

$$\boxed{\mathbb{E}(a g(X)) = a \mathbb{E}(g(X))}$$

$$\boxed{\mathbb{E}(g(X) + h(X)) = \mathbb{E}(g(X)) + \mathbb{E}(h(X))}$$

Note that here  $g(X)$  and  $h(X)$  are functions of the random variable  $X$ : e.g.  $g(X) = X^2$ .

Using these results we can obtain the following useful formula for variance:

$$\begin{aligned} \mathbb{V}(X) &= \mathbb{E}((X - \mathbb{E}(X))^2) \\ &= \mathbb{E}(X^2 - 2X \mathbb{E}(X) + (\mathbb{E}(X))^2) \\ &= \mathbb{E}(X^2) - \mathbb{E}(2X \mathbb{E}(X)) + \mathbb{E}((\mathbb{E}(X))^2) \\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + (\mathbb{E}(X))^2 \\ &= \mathbb{E}(X^2) - 2(\mathbb{E}(X))^2 + (\mathbb{E}(X))^2 \\ &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 . \end{aligned}$$

That is,

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

The above properties of expectations imply that for constants  $a$  and  $b$ ,

$$\boxed{\mathbb{V}(aX + b) = a^2 \mathbb{V}(X)} . \quad (3.49)$$

More generally, for random variables  $X_1, X_2, \dots, X_n$  and constants  $a_1, a_2, \dots, a_n$

- $\mathbb{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mathbb{E}(X_i) . \quad (3.50)$

- $\mathbb{V}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \mathbb{V}(X_i), \text{ provided } X_1, X_2, \dots, X_n \text{ are independent} . \quad (3.51)$

- Let  $X_1, X_2, \dots, X_n$  be independent RVs, then

$$\mathbb{E}\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n \mathbb{E}(X_i), \text{ provided } X_1, X_2, \dots, X_n \text{ are independent} . \quad (3.52)$$

### 3.8.3 Expectation of Common Random Variables

Let us compute the mean and variance of our familiar RVs.

**Example 74 (Mean and variance of Bernoulli( $\theta$ ) RV)** Let  $X \sim \text{Bernoulli}(\theta)$ . Then,

$$\mathbb{E}(X) = \sum_{x=0}^1 xf(x) = (0 \times (1 - \theta)) + (1 \times \theta) = 0 + \theta = \theta ,$$

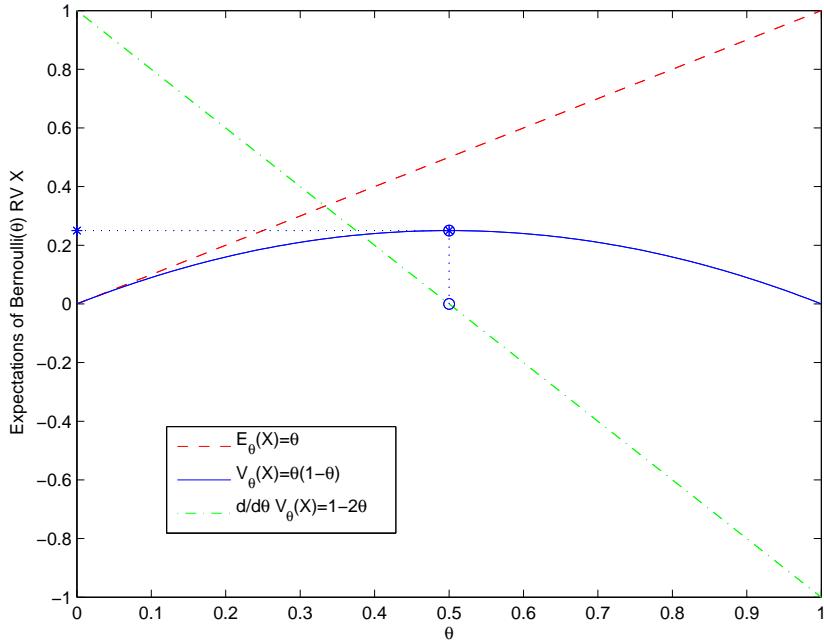
$$\mathbb{E}(X^2) = \sum_{x=0}^1 x^2 f(x) = (0^2 \times (1 - \theta)) + (1^2 \times \theta) = 0 + \theta = \theta ,$$

$$\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \theta - \theta^2 = \theta(1 - \theta) .$$

Parameter specifically,

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

Figure 3.16: Mean ( $\mathbb{E}_\theta(X)$ ), variance ( $\mathbb{V}_\theta(X)$ ) and the rate of change of variance ( $\frac{d}{d\theta} \mathbb{V}_\theta(X)$ ) of a Bernoulli( $\theta$ ) RV  $X$  as a function of the parameter  $\theta$ .



Maximum of the variance  $\mathbb{V}_\theta(X)$  is found by setting the derivative to zero, solving for  $\theta$  and showing the second derivative is locally negative, i.e.  $\mathbb{V}_\theta(X)$  is concave down:

$$\mathbb{V}'_\theta(X) := \frac{d}{d\theta} \mathbb{V}_\theta(X) = 1 - 2\theta = 0 \iff \theta = \frac{1}{2} , \quad \mathbb{V}''_\theta(X) := \frac{d}{d\theta} \left( \frac{d}{d\theta} \mathbb{V}_\theta(X) \right) = -2 < 0 ,$$

$$\max_{\theta \in [0,1]} \mathbb{V}_\theta(X) = \frac{1}{2} \left( 1 - \frac{1}{2} \right) = \frac{1}{4} , \text{ since } \mathbb{V}_\theta(X) \text{ is maximized at } \theta = \frac{1}{2}$$

The plot depicting these expectations as well as the rate of change of the variance are depicted in Figure 3.16. Note from this Figure that  $V_\theta(X)$  attains its maximum value of  $1/4$  at  $\theta = 0.5$  where  $\frac{d}{d\theta} V_\theta(X) = 0$ . Furthermore, we know that we don't have a minimum at  $\theta = 0.5$  since the second derivative  $V''_\theta(X) = -2$  is negative for any  $\theta \in [0, 1]$ . This confirms that  $V_\theta(X)$  is concave down and therefore we have a maximum of  $V_\theta(X)$  at  $\theta = 0.5$ . We will revisit this example when we employ a numerical approach called Newton-Raphson method to solve for the maximum of a differentiable function by setting its derivative equal to zero.

**Example 75 (Mean and variance of Uniform(0, 1) RV)** Let  $X \sim \text{Uniform}(0, 1)$ . Then,

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

**Exercise 3.29 (Mean and variance of Uniform( $\theta_1, \theta_2$ ) RV)** Let  $X \sim \text{Uniform}(\theta_1, \theta_2)$  of Model 9. Derive expressions for  $E(X)$  and  $V(X)$  in terms of the parameters  $\theta_1$  and  $\theta_2$ . Make sure that when  $\theta_2 = 1$  and  $\theta_1 = 0$  you recover the expectation and variance of the Uniform(0, 1) RV in Example 75.

**Example 76 (Expected Exponential of the Uniform(0, 1) RV)** Let  $X \sim \text{Uniform}(0, 1)$  and  $Y = r(X) = e^X$ . Compute  $E(Y)$ .

We can simply apply the definition of  $E(r(X))$ , since  $Y = r(X)$ , is just a function of  $X$ , as follows:

$$E(Y) = \int_0^1 e^x f(x) dx = \int_0^1 e^x \cdot 1 dx = e - 1.$$

**Example 77 (Mean and variance of Exponential( $\lambda$ ) RV)** Show that the mean of an Exponential( $\lambda$ ) RV  $X$  is:

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

and the variance is:

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

**Example 78 (Mean and variance of Geometric( $\theta$ ) RV)** Let  $X \sim \text{Geometric}(\theta)$  RV. Then,

$$E(X) = \sum_{x=0}^{\infty} x \theta (1 - \theta)^x = \theta \sum_{x=0}^{\infty} x (1 - \theta)^x$$

In order to simplify the RHS above, let us employ differentiation with respect to  $\theta$ :

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

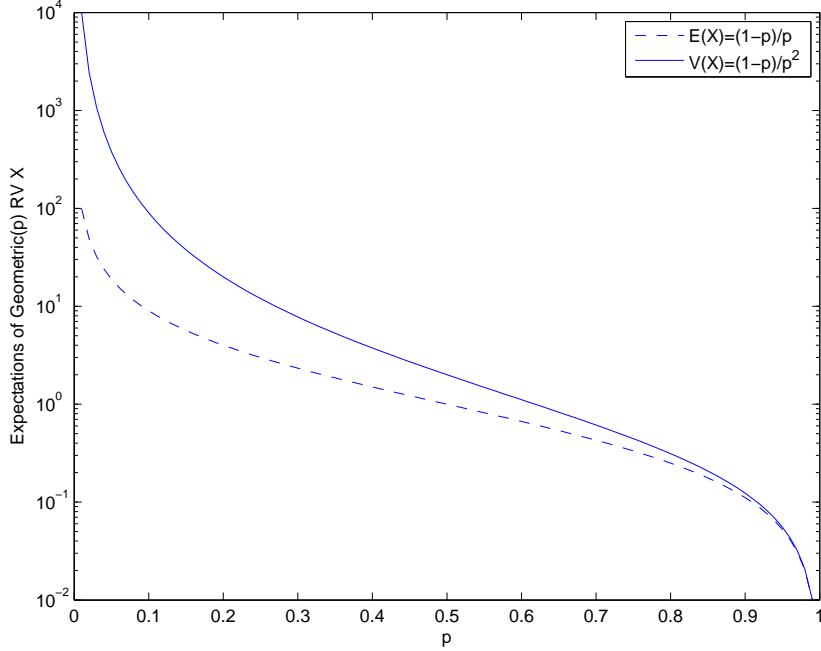
Multiplying the LHS and RHS above by  $-(1 - \theta)$  and substituting in  $E(X) = \theta \sum_{x=0}^{\infty} x (1 - \theta)^x$ , we get a much simpler expression for  $E(X)$ :

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

Similarly, it can be shown that

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

Figure 3.17: Mean and variance of a Geometric( $\theta$ ) RV  $X$  as a function of the parameter  $\theta$ .



**Example 79 (Mean and variance of Binomial( $n, \theta$ ) RV)** Let  $X \sim \text{Binomial}(n, \theta)$ . Based on the definition of expectation:

$$E(X) = \int x dF(x; n, \theta) = \sum_x x f(x; n, \theta) = \sum_{x=0}^n x \binom{n}{x} \theta^x (1-\theta)^{n-x} .$$

However, this is a nontrivial sum to evaluate. Instead, we may use (3.50) and (3.51) by noting that  $X = \sum_{i=1}^n X_i$ , where the  $\{X_1, X_2, \dots, X_n\} \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta)$ ,  $E(X_i) = \theta$  and  $V(X_i) = \theta(1-\theta)$ :

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

**Example 80 (Mean and variance of Poisson( $\lambda$ ) RV)** Let  $X \sim \text{Poisson}(\lambda)$ . Then:

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

Similarly,

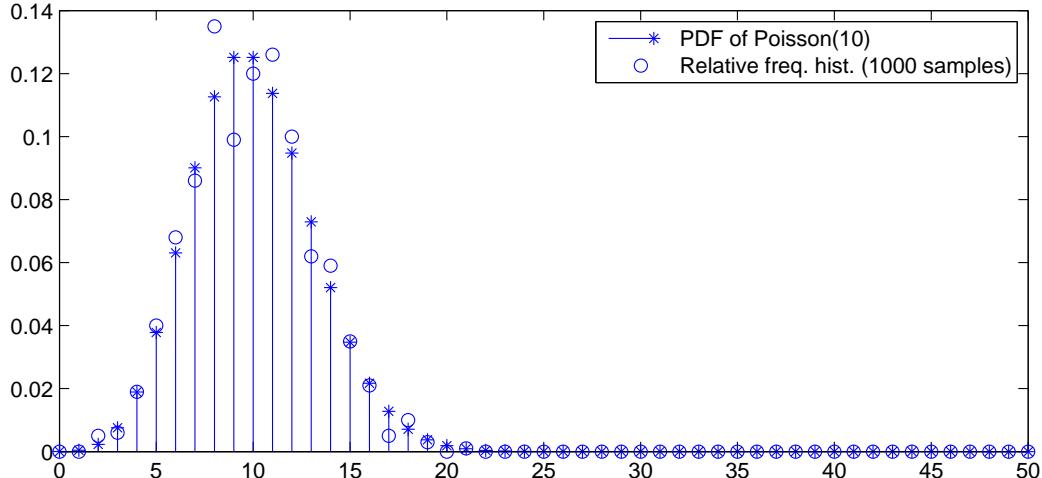
$$V(X) = E(X^2) - (E(X))^2 = \lambda + \lambda^2 - \lambda^2 = \lambda .$$

since

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

Note that  $\text{Poisson}(\lambda)$  distribution is one whose mean and variance are the same, namely  $\lambda$ .

Figure 3.18: PDF of  $X \sim \text{Poisson}(\lambda = 10)$  and the relative frequency histogram based on 1000 samples from  $X$  according to Simulation 149.



The  $\text{Poisson}(\lambda)$  RV  $X$  is also related to the IID Exponential( $\lambda$ ) 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( $\lambda$ ) RV. For example, the number of buses arriving at our bus-stop in the next minute, with exponentially distributed inter-arrival times, has a Poisson distribution.

**Example 81 (Mean and variance of Normal( $\mu, \sigma^2$ ) RV)** The location-scale family of RVs is indeed parameterised by its mean and variance, i.e., if  $X \sim \text{Normal}(\mu, \sigma^2)$  where  $X = g(Z) = \sigma Z + \mu$  and  $Z \sim \text{Normal}(0, 1)$  then  $E(X) = \mu$  and  $V(X) = \sigma^2$  follows directly from the properties of Expectations, provided  $E(Z) = 0$  and  $V(Z) = E(Z^2) - (E(Z))^2 = E(Z^2) = 1$ .

The mean of a  $\text{Normal}(0, 1)$  RV  $Z$  is:

$$E(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \exp\left(-\frac{1}{2}z^2\right) dz = \frac{1}{\sqrt{2\pi}} \left[ -\exp\left(-\frac{1}{2}z^2\right) \right]_{-\infty}^{\infty} = 0,$$

and the variance is:

$$V(Z) = E(Z^2) - (E(Z))^2 = E(Z^2) - 0 = E(Z^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz.$$

Using integration by parts with  $u = z, dv = ze^{-z^2/2} \implies du = 1, v = -e^{-z^2/2}$ ,  $\int u dv = uv - \int v du$

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \left( -ze^{-z^2/2} \right)_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz = 0 + 1 = 1$$

The first term after the first equality above equals 0 because the exponential goes to 0 much faster than  $z$  grows to  $\pm\infty$ . The second term equals 1 because it is exactly the total probability integral of the PDF of the  $\text{Normal}(0, 1)$  RV.

Next, let us become familiar with an RV for which the expectation does not exist.

**Model 13 (Cauchy)** The density of the Cauchy RV  $Y$  is:

$$f(y) = \frac{1}{\pi(1+y^2)}, \quad -\infty < y < \infty , \quad (3.53)$$

and its DF is:

$$F(y) = \frac{1}{\pi} \tan^{-1}(y) + \frac{1}{2} . \quad (3.54)$$

Randomly spinning a LASER emitting improvisation of “Darth Maul’s double edged lightsaber” that is centered at  $(1, 0)$  in the plane  $\mathbb{R}^2$  and recording its intersection with the  $y$ -axis, in terms of the  $y$  coordinates of the point  $(0, y)$ , gives rise to the *Standard Cauchy* RV.

The Cauchy RV  $Y$  can be derived from a RV  $X \sim \text{Uniform}(-\pi/2, \pi/2)$  by the simple transformation  $Y = \tan(X)$  for the above construction. Since  $\tan(x)$  is one-to-one and monotone on the range of  $X$  given by  $(-\pi/2, \pi/2)$ , we can use the change of variable formula in Equation 3.38 to obtain the PDF  $f_Y(y)$  from the PDF  $f_X(x) = \frac{1}{\pi} \mathbf{1}_{(-\pi/2, \pi/2)}(x)$  as follows:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| = f_X(\tan^{-1}(y)) \left| \frac{d}{dy} \tan^{-1}(y) \right| = \frac{1}{\pi} \left| \frac{1}{1+y^2} \right|$$

Note that the construction is valid even if we sample  $X$  uniformly from  $(0, \pi)$  and take its  $\tan(X)$ .

**Example 82 (Mean of Cauchy RV)** The expectation of the Cauchy RV  $X$ , obtained via integration by parts (set  $u = x$  and  $v = \tan^{-1}(x)$ ) does not exist, since:

$$\int |x| dF(x) = \frac{2}{\pi} \int_0^\infty \frac{x}{1+x^2} dx = (x \tan^{-1}(x))_0^\infty - \int_0^\infty \tan^{-1}(x) dx = \infty . \quad (3.55)$$

Note that we consider symmetry of integral about the origin and take twice the integral over  $(0, \infty)$  above. Variance and higher moments cannot be defined when the expectation itself is undefined.

Next let us consider a natural generalization of the  $\text{Bernoulli}(\theta)$  RV with more than two outcomes but in the set  $\{1, 2, \dots, k\}$ .

**Model 14 (de Moivre( $\theta_1, \theta_2, \dots, \theta_k$ ))** Given a specific point  $(\theta_1, \theta_2, \dots, \theta_k)$  in the unit  $k-1$ -Simplex:

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

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:

$$F(x; \theta_1, \theta_2, \dots, \theta_k) = \begin{cases} 0 & \text{if } -\infty < x < 1 \\ \theta_1 & \text{if } 1 \leq x < 2 \\ \theta_1 + \theta_2 & \text{if } 2 \leq x < 3 \\ \vdots & \\ \theta_1 + \theta_2 + \dots + \theta_{k-1} & \text{if } k-1 \leq x < k \\ \theta_1 + \theta_2 + \dots + \theta_{k-1} + \theta_k = 1 & \text{if } k \leq x < \infty \end{cases} \quad (3.56)$$

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,

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

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

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,

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

### 3.9 Exercises in Expectations of Random Variables

**Ex. 3.30** — Let  $X$  be the number of air conditioners a store sells each day, and assume that  $X$  has probability mass function  $f(10) = 0.1, f(11) = 0.3, f(12) = 0.4, f(13) = 0.2$ .

1. Find the expected number of conditioners that the store sells each day.
2. If the profit per conditioner is \$55, what is the expected daily profit?

**Ex. 3.31** — A small petrol station is supplied with fuel every Saturday afternoon. Assume that its volume of sales  $X$ , in ten thousands of litres, has density

$$f(x) = \begin{cases} 6x(1-x) & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} .$$

Determine the mean and variance of  $X$ .

**Ex. 3.32** — Starting from the definition of the variance of a random variable (Definition 30) show that

$$V(X) = E(X^2) - (E(X))^2 .$$

**Ex. 3.33** — Show that  $V(aX + b) = a^2V(X)$  for constants  $a$  and  $b$  and a random variable  $X$ .

**Ex. 3.34** — \*\*Let  $X$  be a discrete random variable with PMF given by

$$f(x) = \begin{cases} \frac{x}{10} & \text{if } x \in \{1, 2, 3, 4\}, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find:

- (i)  $P(X = 0)$
- (ii)  $P(2.5 < X < 5)$
- (iii)  $E(X)$
- (iv)  $V(X)$

(b) Write down the DF (or CDF) of  $X$ .

(c) Plot the PMF and CDF of  $X$ .

**Ex. 3.35** — Find the mean and the variance of the following random variables.

1.  $X$  a discrete uniform random variable on  $\{1, 2, 3, 4, 5, 6\}$ , i.e., *the number a fair die turns up*.

2.  $X$  is a  $\text{Uniform}(0, 8)$  random variable, i.e., *a continuous uniform random variable from the interval  $[0, 8]$* .

3.  $X$  has a density function

$$f(x) = \begin{cases} 2e^{-2x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

## 3.10 Multivariate Random Variables

Often, in experiments we are measuring two or more aspects simultaneously. For example, we may be measuring the diameters and lengths of cylindrical shafts manufactured in a plant or heights, weights and blood-sugar levels of individuals in a clinical trial. Thus, the underlying outcome  $\omega \in \Omega$  needs to be mapped to measurements as realizations of random vectors in the real plane  $\mathbb{R}^2 = (-\infty, \infty) \times (-\infty, \infty)$  or the real space  $\mathbb{R}^3 = (-\infty, \infty) \times (-\infty, \infty) \times (-\infty, \infty)$ :

$$\omega \mapsto (X(\omega), Y(\omega)) : \Omega \rightarrow \mathbb{R}^2 \quad \omega \mapsto (X(\omega), Y(\omega), Z(\omega)) : \Omega \rightarrow \mathbb{R}^3$$

More generally, we may be interested in heights, weights, blood-sugar levels, family medical history, known allergies, etc. of individuals in the clinical trial and thus need to make  $m$  measurements of the outcome in  $\mathbb{R}^m$  using a “measurable mapping” from  $\Omega \rightarrow \mathbb{R}^m$ . To deal with such multivariate measurements we need the notion of **random vectors** ( $\text{R}\vec{\text{Vs}}$ ), i.e. ordered pairs of random variables  $(X, Y)$ , ordered triples of random variables  $(X, Y, Z)$ , or more generally ordered  $m$ -tuples of random variables  $(X_1, X_2, \dots, X_m)$ .

### 3.10.1 $\mathbb{R}^2$ -valued Random Variables

We first focus on understanding  $(X, Y)$ , a bivariate R $\vec{V}$  or  $\mathbb{R}^2$ -valued RV that is obtained from a pair of discrete or continuous RVs. We then generalize to  $\mathbb{R}^m$ -valued RVs with  $m > 2$  in the next section.

**Definition 34 (JDF)** The **joint distribution function (JDF)** or **joint cumulative distribution function (JCDF)**,  $F_{X,Y}(x, y) : \mathbb{R}^2 \rightarrow [0, 1]$ , of the bivariate random vector  $(X, Y)$  is

$$\begin{aligned} F_{X,Y}(x, y) &= P(X \leq x \cap Y \leq y) = P(X \leq x, Y \leq y) \\ &= P(\{\omega : X(\omega) \leq x, Y(\omega) \leq y\}), \text{ for any } (x, y) \in \mathbb{R}^2, \end{aligned} \quad (3.57)$$

where the right-hand side represents the probability that the random vector  $(X, Y)$  takes on a value in  $\{(x', y') : x' \leq x, y' \leq y\}$ , the set of points in the plane that are south-west of the point  $(x, y)$ .

The JDF  $F_{X,Y}(x, y) : \mathbb{R}^2 \rightarrow \mathbb{R}$  satisfies the following conditions to remain a probability:

1.  $0 \leq F_{X,Y}(x, y) \leq 1$
2.  $F_{X,Y}(x, y)$  is a non-decreasing function of both  $x$  and  $y$
3.  $F_{X,Y}(x, y) \rightarrow 1$  as  $x \rightarrow \infty$  and  $y \rightarrow \infty$
4.  $F_{X,Y}(x, y) \rightarrow 0$  as  $x \rightarrow -\infty$  and  $y \rightarrow -\infty$

**Definition 35 (JPMF)** If  $(X, Y)$  is a **discrete random vector** that takes values in a discrete support set  $\mathcal{S}_{X,Y} = \{(x_i, y_j) : i = 1, 2, \dots, j = 1, 2, \dots\} \subset \mathbb{R}^2$  with probabilities  $p_{i,j} = P(X = x_i, Y = y_j) > 0$ , then its **joint probability mass function** (or JPMF) is:

$$f_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j) = \begin{cases} p_{i,j} & \text{if } (x_i, y_j) \in \mathcal{S}_{X,Y} \\ 0 & \text{otherwise} \end{cases}. \quad (3.58)$$

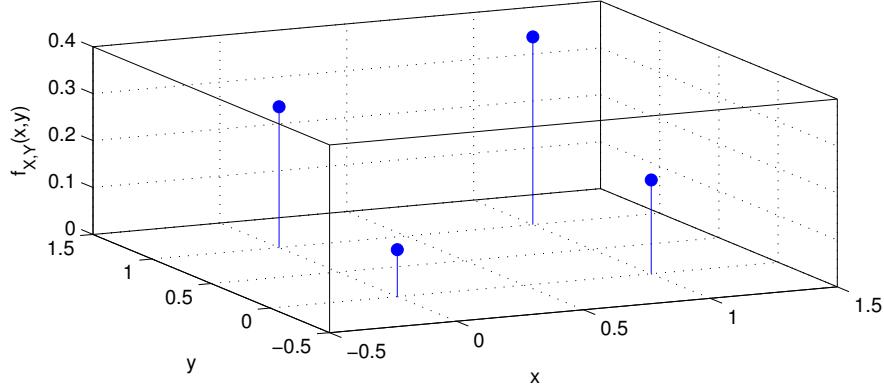
Since  $P(\Omega) = 1$ ,  $\sum_{(x_i, y_j) \in \mathcal{S}_{X,Y}} f_{X,Y}(x_i, y_j) = 1$ .

From JPMF  $f_{X,Y}$  we can get the values of the JDF  $F_{X,Y}(x, y)$  and the probability of any event  $B$  by simply taking sums,

$$\boxed{F_{X,Y}(x, y) = \sum_{x_i \leq x, y_j \leq y} f_{X,Y}(x_i, y_j)}, \quad \boxed{P(B) = \sum_{(x_i, y_j) \in B \cap \mathcal{S}_{X,Y}} f_{X,Y}(x_i, y_j)}, \quad (3.59)$$

**Example 83** Let  $(X, Y)$  be a discrete bivariate R $\vec{V}$  with the following joint probability mass function (JPMF):

$$f_{X,Y}(x, y) := P(X = x, Y = y) = \begin{cases} 0.1 & \text{if } (x, y) = (0, 0) \\ 0.3 & \text{if } (x, y) = (0, 1) \\ 0.2 & \text{if } (x, y) = (1, 0) \\ 0.4 & \text{if } (x, y) = (1, 1) \\ 0.0 & \text{otherwise.} \end{cases}$$



It is helpful to write down the JPMF  $f_{X,Y}(x,y)$  in a tabular form:

	$Y = 0$	$Y = 1$
$X = 0$	0.1	0.3
$X = 1$	0.2	0.4

From the above Table we can read for instance that the joint probability  $f_{X,Y}(0,0) = 0.1$ .

Find  $P(B)$  for the event  $B = \{(0,0), (1,1)\}, F_{X,Y}(1/2, 1/2), F_{X,Y}(3/2, 1/2), F_{X,Y}(4, 5)$  and  $F_{X,Y}(-4, -1)$ .

1.  $P(B) = \sum_{(x,y) \in \{(0,0), (1,1)\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) + f_{X,Y}(1,1) = 0.1 + 0.4$
2.  $F_{X,Y}(1/2, 1/2) = \sum_{\{(x,y): x \leq 1/2, y \leq 1/2\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) = 0.1$
3.  $F_{X,Y}(3/2, 1/2) = \sum_{\{(x,y): x \leq 3/2, y \leq 1/2\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) + f_{X,Y}(1,0) = 0.1 + 0.2 = 0.3$
4.  $F_{X,Y}(4, 5) = \sum_{\{(x,y): x \leq 4, y \leq 5\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) + f_{X,Y}(0,1) + f_{X,Y}(1,0) + f_{X,Y}(1,1) = 1$
5.  $F_{X,Y}(-4, -1) = \sum_{\{(x,y): x \leq -4, y \leq -1\}} f_{X,Y}(x,y) = 0$

**Definition 36 (JPDF)** We say  $(X, Y)$  is a **continuous  $R^2$ -valued random variable** if its JDF  $F_{X,Y}(x,y)$  is differentiable and its **joint probability density function (JPDF)** is given by:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) .$$

For notational convenience, we sometimes suppress the subscripting when the random variables are clear from the context and write  $f(x,y)$  and  $F(x,y)$  instead of  $f_{X,Y}(x,y)$  and  $F_{X,Y}(x,y)$ , respectively.

From JPDF  $f_{X,Y}$  we can compute the JDF  $F_{X,Y}$  at any point  $(x,y) \in \mathbb{R}^2$  and more generally we can compute the probability of any event  $B$ , that can be cast as a region in  $\mathbb{R}^2$ , by simply taking two-dimensional integrals:

$$\boxed{F_{X,Y}(x,y) = \int_{-\infty}^y \int_{-\infty}^x f_{X,Y}(u,v) du dv} , \quad (3.60)$$

and

$$\boxed{P(B) = \int \int_B f_{X,Y}(x,y) dx dy} . \quad (3.61)$$

In particular, if  $\mathbb{B}_\delta(x, y)$  denotes a square of a small area  $\delta > 0$  that is centered at  $(x, y)$ , then the following approximate equality holds and improves as  $\delta \rightarrow 0$ :

$$\mathbb{P}((X, Y) \in \mathbb{B}_\delta(x, y)) \approx \delta f_{X,Y}(x, y) . \quad (3.62)$$

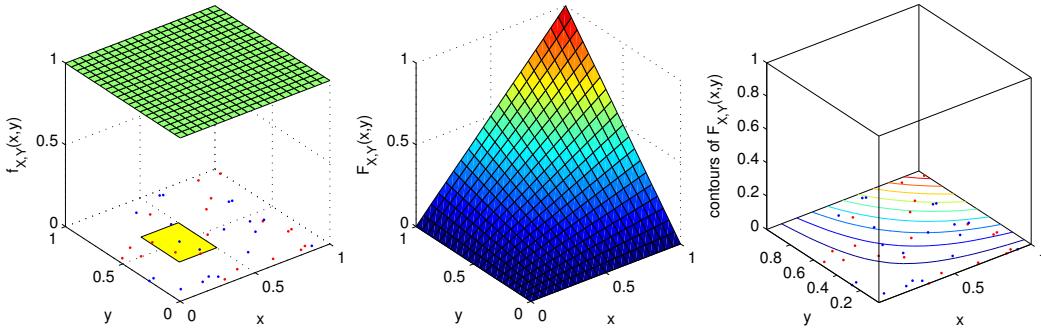
The JPDF satisfies the following two properties:

1. integrates to 1, i.e.,  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1$
2. is a non-negative function, i.e.,  $f_{X,Y}(x, y) \geq 0$  for every  $(x, y) \in \mathbb{R}^2$ .

**Example 84** Let  $(X, Y)$  be a continuous RV that is uniformly distributed on the unit square  $[0, 1]^2 := [0, 1] \times [0, 1]$  with following JPDF:

$$f(x, y) = \mathbb{1}_{[0,1]^2}(x) \begin{cases} 1 & \text{if } (x, y) \in [0, 1]^2 \\ 0 & \text{otherwise.} \end{cases}$$

Find explicit expressions for the following: (1) DF  $F(x, y)$  for any  $(x, y) \in [0, 1]^2$ , (2)  $\mathbb{P}(X \leq 1/3, Y \leq 1/2)$ , (3)  $\mathbb{P}((X, Y) \in [1/4, 1/2] \times [1/3, 2/3])$ .



Let us begin to find the needed expressions.

1. Let  $(x, y) \in [0, 1]^2$  then by Equation (3.60):

$$F_{X,Y}(x, y) = \int_{-\infty}^y \int_{-\infty}^x f_{X,Y}(u, v) du dv = \int_0^y \int_0^x 1 du dv = \int_0^y [u]_{u=0}^x dv = \int_0^y x dv = [xv]_{v=0}^y = xy$$

2. We can obtain  $\mathbb{P}(X \leq 1/3, Y \leq 1/2)$  by evaluating  $F_{X,Y}$  at  $(1/3, 1/2)$ :

$$\mathbb{P}(X \leq 1/3, Y \leq 1/2) = F_{X,Y}(1/3, 1/2) = \frac{1}{3} \frac{1}{2} = \frac{1}{6}$$

We can also find  $\mathbb{P}(X \leq 1/3, Y \leq 1/2)$  by integrating the JPDF over the rectangular event  $A = \{X < 1/3, Y < 1/2\} \subset [0, 1]^2$  according to Equation (3.61). This amounts here to finding the area of  $A$ , we compute  $\mathbb{P}(A) = (1/3)(1/2) = 1/6$ .

3. We can find  $\mathbb{P}((X, Y) \in [1/4, 1/2] \times [1/3, 2/3])$  by integrating the JPDF over the rectangular event  $B = [1/4, 1/2] \times [1/3, 2/3]$  according to Equation (3.61):

$$\begin{aligned} \mathbb{P}((X, Y) \in [1/4, 1/2] \times [1/3, 2/3]) &= \int \int_B f_{X,Y}(x, y) dx dy = \int_{1/3}^{2/3} \int_{1/4}^{1/2} 1 dx dy \\ &= \int_{1/3}^{2/3} [x]_{1/4}^{1/2} dy = \int_{1/3}^{2/3} \left[ \frac{1}{2} - \frac{1}{4} \right] dy = \left( \frac{1}{2} - \frac{1}{4} \right) [y]_{1/3}^{2/3} \\ &= \left( \frac{1}{2} - \frac{1}{4} \right) \left( \frac{2}{3} - \frac{1}{3} \right) = \frac{1}{4} \left( \frac{1}{3} \right) = \frac{1}{12} \end{aligned}$$

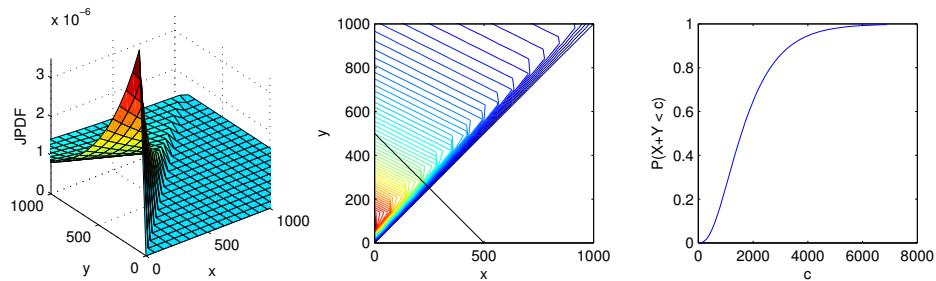
In general, for a bivariate uniform RV  $\vec{V}$  on the unit square the  $P([a, b] \times [c, d]) = (b-a)(d-c)$  for any event given by the rectangular region  $[a, b] \times [c, d]$  inside the unit square  $[0, 1] \times [0, 1]$ . Thus any two events with the same rectangular area have the same probability (imagine sliding a small rectangle inside the unit square... no matter where you slide this rectangle to while remaining in the unit square, the probability of  $\omega \mapsto (X(\omega), Y(\omega)) = (x, y)$  falling inside this “slidable” rectangle is the same...).

**Example 85** Let the RV  $X$  denote the time until a web server connects to your computer, and let the RV  $Y$  denote the time until the server authorizes you as a valid user. Each of these RVs measures the waiting time from a common starting time (in milliseconds) and  $X < Y$ . From past response times of the web server we know that a good approximation for the JPDF of the RV  $(X, Y)$  is

$$f_{X,Y}(x, y) = \begin{cases} \frac{6}{10^6} \exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right) & \text{if } x > 0, y > 0, x < y \\ 0 & \text{otherwise.} \end{cases}$$

Answer the following:

1. identify the support of  $(X, Y)$ , i.e., the region in the plane where  $f_{X,Y}$  takes positive values
2. check that  $f_{X,Y}$  indeed integrates to 1 as it should
3. Find  $P(X \leq 400, Y \leq 800)$
4. It is known that humans prefer a response time of under 1/10 seconds ( $10^2$  milliseconds) from the web server before they get impatient. What is  $P(X + Y < 10^2)$ ?



Let us answer the questions.

1. The support is the intersection of the positive quadrant with the  $y > x$  half-plane.

2.

$$\begin{aligned}
\int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} f_{X,Y}(x,y) dx dy &= \int_{x=0}^{\infty} \int_{y=x}^{\infty} f_{X,Y}(x,y) dy dx \\
&= \int_{x=0}^{\infty} \int_{y=x}^{\infty} \frac{6}{10^6} \exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right) dy dx \\
&= \frac{6}{10^6} \int_{x=0}^{\infty} \left( \int_{y=x}^{\infty} \exp\left(-\frac{2}{1000}y\right) dy \right) \exp\left(-\frac{1}{1000}x\right) dx \\
&= \frac{6}{10^6} \int_{x=0}^{\infty} \left[ -\frac{1000}{2} \exp\left(-\frac{2}{1000}y\right) \right]_{y=x}^{\infty} \exp\left(-\frac{1}{1000}x\right) dx \\
&= \frac{6}{10^6} \int_{x=0}^{\infty} \left[ 0 + \frac{1000}{2} \exp\left(-\frac{2}{1000}x\right) \right] \exp\left(-\frac{1}{1000}x\right) dx \\
&= \frac{6}{10^6} \int_{x=0}^{\infty} \frac{1000}{2} \exp\left(-\frac{2}{1000}x - \frac{1}{1000}x\right) dx \\
&= \frac{6}{10^6} \frac{1000}{2} \left[ -\frac{1000}{3} \exp\left(-\frac{3}{1000}x\right) \right]_{x=0}^{\infty} \\
&= \frac{6}{10^6} \frac{1000}{2} \left[ 0 + \frac{1000}{3} \right] \\
&= 1
\end{aligned}$$

3. First, identify the region with positive JPDF for the event ( $X \leq 400, Y \leq 800$ )

$$\begin{aligned}
P(X \leq 400, Y \leq 800) &= \int_{x=0}^{400} \int_{y=x}^{800} f_{X,Y}(x,y) dy dx \\
&= \int_{x=0}^{400} \int_{y=x}^{800} \frac{6}{10^6} \exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right) dy dx \\
&= \frac{6}{10^6} \int_{x=0}^{400} \left[ -\frac{1000}{2} \exp\left(-\frac{2}{1000}y\right) \right]_{y=x}^{800} \exp\left(-\frac{1}{1000}x\right) dx \\
&= \frac{6}{10^6} \frac{1000}{2} \int_{x=0}^{400} \left( -\exp\left(-\frac{1600}{1000}\right) + \exp\left(-\frac{2}{1000}x\right) \right) \exp\left(-\frac{1}{1000}x\right) dx \\
&= \frac{6}{10^6} \frac{1000}{2} \int_{x=0}^{400} \left( \exp\left(-\frac{3}{1000}x\right) - e^{-8/5} \exp\left(-\frac{1}{1000}x\right) \right) dx \\
&= \frac{6}{10^6} \frac{1000}{2} \left( \left( -\frac{1000}{3} \exp\left(-\frac{3}{1000}x\right) \right)_{x=0}^{400} - e^{-8/5} \left( -1000 \exp\left(-\frac{1}{1000}x\right) \right)_{x=0}^{400} \right) \\
&= \frac{6}{10^6} \frac{1000}{2} 1000 \left( \frac{1}{3} \left( 1 - e^{-6/5} \right) - e^{-8/5} \left( 1 - e^{-2/5} \right) \right) \\
&= 3 \left( \frac{1}{3} \left( 1 - e^{-6/5} \right) - e^{-8/5} \left( 1 - e^{-2/5} \right) \right) \\
&\approx 0.499 .
\end{aligned}$$

4. First, identify the region with positive JPDF for the event ( $X + Y \leq c$ ), say  $c = 500$  (but generally  $c$  can be any positive number). This is the triangular region at the intersection of the four half-planes:  $x > 0$ ,  $x < c$ ,  $y > x$  and  $y < c - x$ . (Draw picture here) Let's integrate

the JPDF over our triangular event as follows:

$$\begin{aligned}
 P(X + Y \leq c) &= \int_{x=0}^{c/2} \int_{y=x}^{c-x} f_{X,Y}(x,y) dy dx \\
 &= \int_{x=0}^{c/2} \int_{y=x}^{c-x} \frac{6}{10^6} \exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right) dy dx \\
 &= \frac{6}{10^6} \int_{x=0}^{c/2} \int_{y=x}^{c-x} \exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right) dy dx \\
 &= \frac{6}{10^6} \frac{1000}{2} \int_{x=0}^{c/2} \left[-\exp\left(-\frac{2}{1000}y\right)\right]_{y=x}^{c-x} \exp\left(-\frac{1}{1000}x\right) dx \\
 &= \frac{3}{10^3} \int_{x=0}^{c/2} \left[-\exp\left(-\frac{2c-2x}{1000}\right) + \exp\left(-\frac{2x}{1000}\right)\right] \exp\left(-\frac{x}{1000}\right) dx \\
 &= \frac{3}{10^3} \int_{x=0}^{c/2} \left(\exp\left(-\frac{3x}{1000}\right) - \exp\left(\frac{x-2c}{1000}\right)\right) dx \\
 &= 3 \left( \left[ -\frac{1}{3} \exp\left(-\frac{3x}{1000}\right) \right]_{x=0}^{c/2} - \left[ e^{-2c/1000} \exp\left(\frac{x}{1000}\right) \right]_{x=0}^{c/2} \right) \\
 &= 3 \left( \frac{1}{3} (1 - e^{-3c/2000}) - e^{-2c/1000} (e^{c/2000} - 1) \right) \\
 &= 1 - e^{-3c/2000} + 3e^{-2c/1000} - 3e^{-3c/2000} \\
 &= 1 - 4e^{-3c/2000} + 3e^{-c/500}
 \end{aligned}$$

5.  $P(X + Y < 100) = 1 - 4e^{-300/2000} + 3e^{-100/500} \approx 0.134$ . This means only about one in one hundred requests to this server will be processed within 100 milliseconds.

We can obtain  $P(X + Y < c)$  for several values of  $c$  using MATLAB and note that about 96% of requests are processed in less than 3000 milliseconds or 3 seconds.

```

>> c = [100 1000 2000 3000 4000]
c = 100      1000      2000      3000      4000

>> p = 1 - 4 * exp(-3*c/2000) + 3 * exp(-c/500)

p = 0.0134    0.5135    0.8558    0.9630    0.9911

```

**Definition 37 (Marginal PDF or PMF)** If the R $\vec{V}$   $(X, Y)$  has  $f_{X,Y}(x, y)$  as its joint PDF or joint PMF, then the **marginal PDF or PMF** of a random vector  $(X, Y)$  is defined by :

$$f_X(x) = \begin{cases} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy & \text{if } (X, Y) \text{ is a continuous R}\vec{V} \\ \sum_y f_{X,Y}(x,y) & \text{if } (X, Y) \text{ is a discrete R}\vec{V} \end{cases}$$

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

$$f_Y(y) = \begin{cases} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx & \text{if } (X, Y) \text{ is a continuous R}\vec{V} \\ \sum_x f_{X,Y}(x,y) & \text{if } (X, Y) \text{ is a discrete R}\vec{V} \end{cases}$$

**Example 86** Obtain the marginal PMFs  $f_Y(y)$  and  $f_X(x)$  from the joint PMF  $f_{X,Y}(x,y)$  of the discrete R $\vec{V}$  in Example 83. Just sum  $f_{X,Y}(x,y)$  over  $x$ 's and  $y$ 's (reported in a tabular form):

	$Y = 0$	$Y = 1$
$X = 0$	0.1	0.3
$X = 1$	0.2	0.4

From the above Table we can find:

$$f_X(x) = P(X = x) = \sum_y f_{X,Y}(x, y)$$

$$= f_{X,Y}(x, 0) + f_{X,Y}(x, 1) = \begin{cases} f_{X,Y}(0, 0) + f_{X,Y}(0, 1) = 0.1 + 0.3 = 0.4 & \text{if } x = 0 \\ f_{X,Y}(1, 0) + f_{X,Y}(1, 1) = 0.2 + 0.4 = 0.6 & \text{if } x = 1 \end{cases}$$

Similarly,

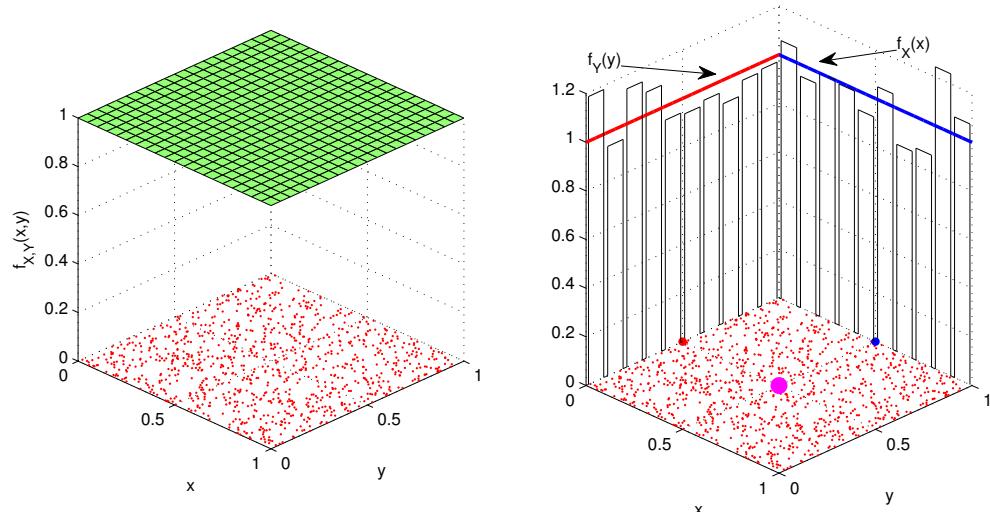
$$f_Y(y) = P(Y = y) = \sum_x f_{X,Y}(x, y)$$

$$= f_{X,Y}(0, y) + f_{X,Y}(1, y) = \begin{cases} f_{X,Y}(0, 0) + f_{X,Y}(1, 0) = 0.1 + 0.2 = 0.3 & \text{if } y = 0 \\ f_{X,Y}(0, 1) + f_{X,Y}(1, 1) = 0.3 + 0.4 = 0.7 & \text{if } y = 1 \end{cases}$$

Just report the marginal probabilities as row and column sums of the JPDF table.

Thus marginal PMF gives us the probability of a specific RV, within a R $\vec{V}$ , taking a value irrespective of the value taken by the other RV in this R $\vec{V}$ .

**Example 87** Obtain the marginal PDFs  $f_Y(y)$  and  $f_X(x)$  from the joint PDF  $f_{X,Y}(x, y)$  of the continuous R $\vec{V}$  in Example 84 (the bivariate uniform R $\vec{V}$  on  $[0, 1]^2$ ).



Let us suppose  $(x, y) \in [0, 1]^2$  and note that  $f_{X,Y} = 0$  if  $(x, y) \notin [0, 1]^2$ . We can obtain marginal PMFs  $f_X(x)$  and  $f_Y(y)$  by integrating the JPDF  $f_{X,Y} = 1$  along  $y$  and  $x$ , respectively.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy = \int_0^1 f_{X,Y}(x, y) dy = \int_0^1 1 dy = [y]_0^1 = 1 - 0 = 1$$

Similarly,

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

We are seeing a histogram of the **marginal samples** and their marginal PDFs in the Figure.

Thus marginal PDF gives us the probability density of a specific RV in a R $\vec{V}$ , irrespective of the value taken by the other RV in this R $\vec{V}$ .

**Example 88** Obtain the marginal PDF  $f_Y(y)$  from the joint PDF  $f_{X,Y}(x,y)$  of the continuous R $\vec{V}$  in Example 85 that gave the response times of a web server.

$$f_{X,Y}(x,y) = \begin{cases} \frac{6}{10^6} \exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right) & \text{if } x > 0, y > 0, x < y \\ 0 & \text{otherwise.} \end{cases}$$

Use  $f_Y(y)$  to compute the probability that  $Y$  exceeds 2000 milliseconds.

First we need to obtain an expression for  $f_Y(y)$ . For  $y > 0$ ,

$$\begin{aligned} f_Y(y) &= \int_{x=-\infty}^{\infty} f_{X,Y}(x,y) dx \\ &= \int_{x=-\infty}^{\infty} 6 \times 10^{-6} e^{-0.001x-0.002y} dx \\ &= 6 \times 10^{-6} \int_{x=0}^y e^{-0.001x-0.002y} dx \\ &= 6 \times 10^{-6} e^{-0.002y} \int_{x=0}^y e^{-0.001x} dx \\ &= 6 \times 10^{-6} e^{-0.002y} \left[ \frac{e^{-0.001x}}{-0.001} \right]_{x=0}^{x=y} \\ &= 6 \times 10^{-6} e^{-0.002y} \left( \frac{e^{-0.001y}}{-0.001} - \frac{e^{-0.001 \times 0}}{-0.001} \right) \\ &= 6 \times 10^{-6} e^{-0.002y} \left( \frac{1 - e^{-0.001y}}{0.001} \right) \\ &= 6 \times 10^{-3} e^{-0.002y} (1 - e^{-0.001y}) \end{aligned}$$

We have the marginal PDF of  $Y$  and from this we can obtain

$$\begin{aligned} P(Y > 2000) &= \int_{2000}^{\infty} f_Y(y) dy \\ &= \int_{2000}^{\infty} 6 \times 10^{-3} e^{-0.002y} (1 - e^{-0.001y}) dy \\ &= 6 \times 10^{-3} \int_{2000}^{\infty} e^{-0.002y} dy - \int_{2000}^{\infty} e^{-0.003y} dy \\ &= 6 \times 10^{-3} \left( \left[ \frac{e^{-0.002y}}{-0.002} \right]_{2000}^{\infty} - \left( \left[ \frac{e^{-0.003y}}{-0.003} \right]_{2000}^{\infty} \right) \right) \\ &= 6 \times 10^{-3} \left( \frac{e^{-4}}{0.002} - \frac{e^{-6}}{0.003} \right) \\ &= 0.05 \end{aligned}$$

Alternatively, you can obtain  $P(Y > 2000)$  by directly integrating the joint PDF  $f_{X,Y}(x,y)$  over the appropriate region (but you may now have to integrate two pieces: rectangular infinite strip  $(x,y) : 0 < x < 2000, y > 2000$  and a triangular infinite piece  $\{(x,y) : y > x, y > 2000, x > 2000\}$ )... more involved but we get the same answer.

$$\begin{aligned} P(Y > 2000) &= \int_{x=0}^{2000} \left( \int_{y=2000}^{\infty} 6 \times 10^{-6} e^{-0.001x-0.002y} dy \right) dx + \\ &\quad \int_{x=2000}^{\infty} \left( \int_{y=x}^{\infty} 6 \times 10^{-6} e^{-0.001x-0.002y} dy \right) dx \\ &\quad \vdots \text{(try as a tutorial problem)} \end{aligned}$$

$$P(Y > 2000) = 0.0475 + 0.0025 = 0.05$$

We have seen the notion of independence of two events in Definition 16 or of a sequence of events in Definition 17. Recall that independence amounts to having the probability of the joint occurrence of the events to be given by the product of the probabilities of each of the events.

We can use the definition of independence of two events to define the independence of two random variables using their distribution functions.

**Definition 38 (Independence of Two RVs)** Consider an  $\mathbb{R}^2$ -valued RV  $X := (X_1, X_2)$ . Then the  $\mathbb{R}$ -valued RVs  $X_1$  and  $X_2$  are said to be independent or independently distributed if and only if

$$P(X_1 \leq x_1, X_2 \leq x_2) = P(X_1 \leq x_1) P(X_2 \leq x_2)$$

or equivalently,

$$F_{X_1, X_2}(x_1, x_2) = F_{X_1}(x_1) F_{X_2}(x_2) ,$$

for any pair of real numbers  $(x_1, x_2) \in \mathbb{R}^2$ .

By the above definition, for **discrete** RVs  $X_1, X_2$  that are independent, the following equality is satisfied between the joint and marginal PMFs:

$$f_{X_1, X_2}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1) P(X_2 = x_2) = f_{X_1}(x_1) f_{X_2}(x_2) \text{ for any } (x_1, x_2) \in \mathbb{R}^2 ,$$

and for **continuous** RVs  $X_1, X_2$  that are independent, the following equality is satisfied between the joint and marginal PDFs:

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) \text{ for any } (x_1, x_2) \in \mathbb{R}^2 .$$

In summary, two RVs  $X$  and  $Y$  are said to be **independent** if and only if for every  $(x, y)$

$F_{X,Y}(x, y) = F_X(x) \times F_Y(y)$	or	$f_{X,Y}(x, y) = f_X(x) \times f_Y(y)$
--	----	--

Let us confirm that our familiar experiment of tossing a fair coin twice independently when encoded by a pair of independent Bernoulli(1/2) RVs satisfies the above definition.

**Example 89 (Pair of independent Bernoulli(1/2) RVs)** Let  $X_1$  and  $X_2$  be a pair of independent Bernoulli(1/2) RVs each taking values in the set  $\{0, 1\}$  with the following tabulated probabilities. Verify that the JPMF  $f_{X_1, X_2}(x_1, x_2) = 1/4$  for each  $(x_1, x_2) \in \{0, 1\}^2$  is indeed given by the marginal PMF  $f_{X_i}(x_i) = 1/2$  for each  $i \in \{1, 2\}$  and each  $x_i \in \{0, 1\}$ .

	$X_2 = 0$	$X_2 = 1$	
$X_1 = 0$	1/4	1/4	1/2
$X_1 = 1$	1/4	1/4	1/2
	1/2	1/2	1

From the above Table we can read for instance that the *joint probability* that  $\mathbb{R}^2$ -valued RV  $(X_1, X_2)$  takes the value or realization  $(0, 0)$  is 1/4 from the first entry of the inner-most tabulated rectangle, i.e.,  $P((X_1, X_2) = (0, 0)) = 1/4$ , and that the *marginal probability* that the RV  $X_1$  takes the value or realization 0 is 1/2, i.e.,  $P(X_1 = 0) = 1/2$ . Clearly,  $1/4 = 1/2 \times 1/2$ , and so our familiar experiment when seen as an  $\mathbb{R}^2$ -valued RV is indeed composed of two independent  $\mathbb{R}$ -valued Bernoulli(1/2) RVs.

**Example 90** Recall the  $\mathbb{R}^2$ -valued continuous RV  $(X, Y)$  of Example 84 that is uniformly distributed on the unit square  $[0, 1]^2$ . First show that  $X$  and  $Y$  independent. Then show that both  $X$  and  $Y$  are identically distributed according to the  $\text{Uniform}(0, 1)$  RV.

*Solution:*

This can be shown by checking that the joint PDF is indeed equal to the product of the marginal PDFs of  $\text{Uniform}(0, 1)$  RVs as follows:

$$\begin{cases} 1 = f_{X,Y}(x, y) = f_X(x) \times f_Y(y) = 1 \times 1 = 1 & \text{if } (x, y) \in [0, 1]^2 \\ 0 = f_{X,Y}(x, y) = f_X(x) \times f_Y(y) = 0 \times 0 = 0 & \text{if } (x, y) \notin [0, 1]^2 \end{cases}$$

Are  $X$  and  $Y$  independent in the server times  $R\vec{V}$  from Example 85?

We can compute  $f_X(x)$  and use the already computed  $f_Y(y)$  to mechanically check if the JPDF is the product of the marginal PDFs. But intuitively, we know that these RVs (connection time and authentication time) are dependent – one is strictly greater than the other. Also the JPDF has zero density when  $x > y$ , but the product of the marginal densities won't.

Now, let us take advantage of independent random variables and solve some problems.

**Example 91 (distance between random faults in a manufactured line)** Suppose two points are tossed independently and uniformly at random onto a line segment of unit length. What is the probability that the distance between the two points does not exceed a given length  $l$ ?

done in lectures...

**Example 92 (Buffon's Needle Experiment to Physically Estimate  $\pi$ )** Suppose a needle is tossed at random onto a plane ruled with parallel lines a distance  $L$  apart. By a “needle” we mean a line segment of length  $l \leq L$ .

What is the probability that the needle intersects one of the parallel lines? Can you use repeated trials of this experiment to find an approximation to  $\pi$ ?

*Solution:*

Let  $X_1$  be the angle between the needle and the direction of the rulings, and let  $X_2$  be the distance between the bottom point of the needle and the nearest line above this point (see left sub-figure of Figure 3.19). Then the conditions of the “needle tossing at random” experiment are such that the RV  $X_1$  is uniformly distributed in the interval  $[0, \pi]$ , while the RV  $X_2$  is uniformly distributed in the interval  $[0, L]$ . Hence *assuming that the RVs  $X_1$  and  $X_2$  are independent*, we find that their joint probability density function (JPDF) is:

$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{\pi} \mathbf{1}_{[0, \pi]}(x_1) \times \frac{1}{L} \mathbf{1}_{[0, L]}(x_2) = \frac{1}{\pi L} \mathbf{1}_{[0, \pi]}(x_1) \mathbf{1}_{[0, L]}(x_2) = \frac{1}{\pi L} \mathbf{1}_{[0, \pi] \times [0, L]}(x_1, x_2).$$

Figure 3.19: Diagrams done on the board!

The event  $A$  that the needle intersects one of the parallel ruled lines occurs if and only if

$$X_2 \leq l \sin(X_1) ,$$

i.e., if and only if the corresponding point  $X := (X_1, X_2)$  falls in the region  $B$ , where  $B$  is part of the rectangle  $[0, \pi] \times [0, L]$  lying between the  $x_1$ -axis and the curve  $x_2 = \sin(x_1)$  (area under the curve in right-subfigure of Figure 3.19). Hence, we can integrate the JPDF to get the probability of the event  $A$  of interest:

$$P(A) = P((X_1, X_2) \in B) = \int_B \int \frac{dx_1 dx_2}{\pi L} = \frac{2l}{\pi L}$$

where,

$$l \int_0^\pi \pi \sin(x_1) dx_1 = l (-\cos(x_1)]_0^\pi = l(1 - (-1)) = l(1 + 1) = 2l ,$$

is the area of  $B$ .

Thus, if the needle is repeatedly tossed onto the ruled plane and  $n(A)$  is the number of times  $A$  occurs out of  $n$  trials, then the relative frequency of the event  $A$  should approach  $P(A)$  as  $n \rightarrow \infty$  (we will see this as the Law of Large Numbers in the sequel, but recall that this is also how we motivated the LTRF or long-term relative frequency idea of probability):

$$\frac{n(A)}{n} \rightarrow \frac{2l}{\pi L}$$

Hence, for large  $n$ ,

$$\frac{2l}{L} \frac{n}{n(A)}$$

should be a good approximation to  $\pi = 3.14 \dots$ . This is indeed the case.

### 3.10.2 Conditional Random Variables

Often we will have a condition where one of the two random variables that make up a random vector  $(X_1, X_2)$  already occurs and takes a value. And we might want to compute the probability of the occurrence of the other random variable given this conditional information. For this all we need to do is extend the idea of conditional probabilities to  $\mathbb{R}^2$ -valued random variables as defined below.

**Definition 39 (Conditional PDF or PMF)** Let  $(X_1, X_2)$  be a discrete bivariate RV. The conditional PMF of  $X_1|X_2 = x_2$ , where  $f_{X_2}(x_2) := P(X_2 = x_2) > 0$  is:

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

Similarly, if  $f_{X_1}(x_1) := \text{P}(X_1 = x_1) > 0$ , then the conditional PMF of  $X_2|X_1 = x_1$  is:

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

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

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

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

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

Let us consider a few discrete RVs for the simple coin tossing experiment  $\mathcal{E}_\theta^3$  that build on the Bernoulli( $\theta$ ) RV  $X_i$  for the  $i$ -th toss in an **independent and identically distributed (IID.)** manner.

Table 3.1: The 8  $\omega$ 's in the sample space  $\Omega$  of the experiment  $\mathcal{E}_\theta^3$  are given in the first row above. The RV  $Y$  is the number of ‘Heads’ in the 3 tosses and the RV  $Z$  is the number of ‘Tails’ in the 3 tosses. Finally, the RVs  $Y'$  and  $Z'$  are the indicator functions of the event that ‘all three tosses were Heads’ and the event that ‘all three tosses were Tails’, respectively.

$\omega$ :	H	H	H	T	H	T	H	T	T	RV Definitions / Model
$P(\omega)$ :	$\frac{1}{8}$	$X_i \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\frac{1}{2})$								
$Y(\omega)$ :	3	2	2	1	2	1	1	0	0	$Y := X_1 + X_2 + X_3$
$Z(\omega)$ :	0	1	1	2	1	2	2	3	3	$Z := (1 - X_1) + (1 - X_2) + (1 - X_3)$
$Y'(\omega)$ :	1	0	0	0	0	0	0	0	0	$Y' := X_1 X_2 X_3$
$Z'(\omega)$ :	0	0	0	0	0	0	0	1	1	$Z' := (1 - X_1)(1 - X_2)(1 - X_3)$

**Classwork 93 (Two random variables of ‘toss a coin thrice’ experiment)** Describe the probability of the RV  $Y$  and  $Y'$  of Table 3.1 in terms of its PMF. Repeat the process for the RV  $Z$  in your spare time.

$$P(Y = y) = \left\{ \begin{array}{l} \dots \\ \dots \end{array} \right. \quad P(Y' = y') = \left\{ \begin{array}{l} \dots \\ \dots \end{array} \right.$$

**Classwork 94 (The number of ‘Heads’ given there is at least one ‘Tails’)** Consider the following two questions.

1. What is conditional probability  $P(Y|Y' = 0)$  ?

$P(Y = y Y' = 0)$	$= \frac{P(Y=y, Y'=0)}{P(Y'=0)}$	$= \frac{P(\{\omega: Y(\omega)=y \cap Y'(\omega)=0\})}{P(\{\omega: Y'(\omega)=0\})}$	$= ?$
$P(Y = 0 Y' = 0)$	$\frac{P(Y=0, Y'=0)}{P(Y'=0)}$	$\frac{\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}}{\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}}$	$\frac{1}{7}$
$P(Y = 1 Y' = 0)$	$\frac{P(Y=1, Y'=0)}{P(Y'=0)}$	$\frac{\frac{1}{8} + \frac{1}{8} + \frac{1}{8}}{\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}}$	$\frac{3}{7}$
$P(Y = 2 Y' = 0)$	$\frac{P(Y=2, Y'=0)}{P(Y'=0)}$	$\frac{\frac{1}{8} + \frac{1}{8} + \frac{1}{8}}{\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}}$	$\frac{3}{7}$
$P(Y = 3 Y' = 0)$	$\frac{P(Y=3, Y'=0)}{P(Y'=0)}$	$\frac{P(\emptyset)}{\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}}$	$0$
$P(Y \in \{0, 1, 2, 3\} Y' = 0)$	$\frac{\sum_{y=0}^3 P(Y=y, Y'=0)}{P(Y'=0)}$	$\frac{\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}}{\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}}$	$1$

2. What is  $P(Y|Y' = 1)$  ?

$$P(Y = y|Y' = 1) = \begin{cases} 1 & \text{if } y = 3 \\ 0 & \text{otherwise} \end{cases}$$

### 3.10.3 $\mathbb{R}^m$ -valued Random Variables

Consider the R $\vec{V}$   $X$  whose components are the RVs  $X_1, X_2, \dots, X_m$ , i.e.,  $X := (X_1, X_2, \dots, X_m)$ , where  $m \geq 2$ . A particular realization of this RV is a point  $(x_1, x_2, \dots, x_m)$  in  $\mathbb{R}^m$ . Now, let us extend the notions of JCDF, JPMF and JPDF to  $\mathbb{R}^m$ .

**Definition 40 (multivariate JDF)** The **joint distribution function (JDF)** or **joint cumulative distribution function (JCDF)**,  $F_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) : \mathbb{R}^m \rightarrow [0, 1]$ , of the multivariate random vector  $(X_1, X_2, \dots, X_m)$  is

$$\begin{aligned} F_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) &= P(X \leq x_1 \cap X_2 \leq x_2 \cap \dots \cap X_m \leq x_m) \\ &= P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_m \leq x_m) \\ &= P(\{\omega : X_1(\omega) \leq x_1, X_2(\omega) \leq x_2, \dots, X_m(\omega) \leq x_m\}), \end{aligned} \quad (3.63)$$

for any  $(x_1, x_2, \dots, x_m) \in \mathbb{R}^m$ , where the right-hand side represents the probability that the random vector  $(X_1, X_2, \dots, X_m)$  takes on a value in  $\{(x'_1, x'_2, \dots, x'_m) : x'_1 \leq x_1, x'_2 \leq x_2, \dots, x'_m \leq x_m\}$ , the set of points in  $\mathbb{R}^m$  that are less than the point  $(x_1, x_2, \dots, x_m)$  in each coordinate  $1, 2, \dots, m$ .

The JDF  $F_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) : \mathbb{R}^m \rightarrow \mathbb{R}$  satisfies the following conditions to remain a probability:

1.  $0 \leq F_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) \leq 1$
2.  $F_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m)$  is an increasing function of  $x_1, x_2, \dots$  and  $x_m$
3.  $F_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) \rightarrow 1$  as  $x_1 \rightarrow \infty, x_2 \rightarrow \infty, \dots$  and  $x_m \rightarrow \infty$

4.  $F_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) \rightarrow 0$  as  $x_1 \rightarrow -\infty, x_2 \rightarrow -\infty, \dots$  and  $x_m \rightarrow -\infty$

**Definition 41 (Multivariate JPMF)** If  $(X_1, X_2, \dots, X_m)$  is a **discrete random vector** that takes values in a discrete support set  $\mathcal{S}_{X_1, X_2, \dots, X_m}$ , then its **joint probability mass function** (or JPMF) is:

$$f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) = P(X_1 = x_1, X_2 = x_2, \dots, X_m = x_m) . \quad (3.64)$$

Since  $P(\Omega) = 1$ ,  $\sum_{(x_1, x_2, \dots, x_m) \in \mathcal{S}_{X_1, X_2, \dots, X_m}} f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) = 1$ .

From JPMF  $f_{X_1, X_2, \dots, X_m}$  we can get the JCDF  $F_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m)$  and the probability of any event  $B$  by simply taking sums as in Equation (3.59) but now over all  $m$  coordinates.

**Definition 42 (Multivariate JPFD)**  $(X_1, X_2, \dots, X_m)$  is a **continuous random vector** if its JDF  $F_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m)$  is differentiable and the **joint probability density function (JPDF)** is given by:

$$f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) = \frac{\partial^m}{\partial x_1 \partial x_2 \dots \partial x_m} F_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) ,$$

From JPFD  $f_{X_1, X_2, \dots, X_m}$  we can compute the JDF  $F_{X_1, X_2, \dots, X_m}$  at any point  $(x_1, x_2, \dots, x_m) \in \mathbb{R}^m$  and more generally we can compute the probability of any event  $B$ , that can be cast as a region in  $\mathbb{R}^m$ , by “simply” taking  $m$ -dimensional integrals (you have done such iterated integrals when  $m = 3$ ):

$$\boxed{F_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) = \int_{-\infty}^{x_m} \dots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) dx_1 dx_2 \dots dx_m} , \quad (3.65)$$

and

$$\boxed{P(B) = \int \dots \int \int_B f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) dx_1 dx_2 \dots dx_m} . \quad (3.66)$$

The JPFD satisfies the following two properties:

1. integrates to 1, i.e.,  $\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) dx_1 dx_2 \dots dx_m = 1$
2. is a non-negative function, i.e.,  $f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) \geq 0$ .

The marginal PDF (marginal PMF) is obtained by integrating (summing) the JPFD (JPMF) over all other random variables. For example, the marginal PDF of  $X_1$  is

$$f_{X_1}(x_1) = \int_{x_2=-\infty}^{\infty} \dots \int_{x_m=-\infty}^{\infty} f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) dx_2 \dots dx_m$$

**Definition 43 (Independence of Sequence of RVs)** A finite or infinite sequence of RVs  $X_1, X_2, \dots$  is said to be independent or independently distributed if and only if

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

or equivalently,

$$F_{X_{i_1}, X_{i_2}, \dots, X_{i_m}}(x_{i_1}, x_{i_2}, \dots, x_{i_m}) = F_{X_{i_1}}(x_{i_1}) F_{X_{i_2}}(x_{i_2}) \dots F_{X_{i_m}}(x_{i_m}) ,$$

for any distinct subset of indices  $\{i_1, i_2, \dots, i_m\}$  of  $\{1, 2, \dots\}$ , the index set of the sequence of RVs and any sequence of real numbers  $x_{i_1}, x_{i_2}, \dots, x_{i_m}$ .

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:

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

or equivalently,

$$f_{X_{i_1}, X_{i_2}, \dots, X_{i_k}}(x_{i_1}, x_{i_2}, \dots, x_{i_k}) = f_{X_{i_1}}(x_{i_1}) f_{X_{i_2}}(x_{i_2}) \cdots f_{X_{i_k}}(x_{i_k}).$$

From Definition 43, we say  $m$  random variables  $X_1, X_2, \dots, X_m$  are jointly independent or mutually independent if and only if for every  $(x_1, x_2, \dots, x_m) \in \mathbb{R}^m$

$$F_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_m}(x_m), \quad (3.67)$$

$$f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_m}(x_m). \quad (3.68)$$

**Proposition 44 (Conditional probability of independent sequence of RVs)** For an independent sequence of RVs  $\{X_1, X_2, \dots\}$ , we have

$$P(X_{i+1} \leq x_{i+1} | X_i \leq x_i, X_{i-1} \leq x_{i-1}, \dots, X_1 \leq x_1) = P(X_{i+1} \leq x_{i+1}) \quad (3.69)$$

**Proof:**

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

Equation (3.69) simply says that the conditional distribution of the RV  $X_{i+1}$  given all previous RVs  $X_i, X_{i-1}, \dots, X_1$  is simply determined by the distribution of  $X_{i+1}$ .

**Example 95** If  $X_1$  and  $X_2$  are independent random variables then what is their covariance  $\text{Cov}(X_1, X_2)$ ?

*Solution:*

We know for independent RVs from the properties of expectations that

$$E(X_1 X_2) = E(X_1) E(X_2)$$

From the formula for covariance

$$\begin{aligned} \text{Cov}(X_1, X_2) &= E(X_1 X_2) - E(X_1) E(X_2) \\ &= E(X_1) E(X_2) - E(X_1) E(X_2) \quad \text{due to independence} \\ &= 0 \end{aligned}$$

**Remark 45** The converse is not true: two random variables that have zero covariance are not necessarily independent.

### Linear Combination of Independent Normal RVs is a Normal RV

We can get the following special property of normal RVs using Eqn. (3.74). If  $X_1, X_2, \dots, X_m$  be jointly independent RVs, where  $X_i$  is  $\text{Normal}(\mu_i, \sigma_i^2)$ , for  $i = 1, 2, \dots, m$  then  $Y = c + \sum_{i=1}^m a_i X_i$  for some constants  $c, a_1, a_2, \dots, a_m$  is the  $\text{Normal}(c + \sum_{i=1}^m a_i \mu_i, \sum_{i=1}^m a_i^2 \sigma_i^2)$  RV.

**Example 96** Let  $X$  be  $\text{Normal}(2, 4)$ ,  $Y$  be  $\text{Normal}(-1, 2)$  and  $Z$  be  $\text{Normal}(0, 1)$  RVs that are jointly independent. Obtain the following:

1.  $E(3X - 2Y + 4Z)$
2.  $V(2Y - 3Z)$
3. the distribution of  $6 - 2Z + X - Y$
4. the probability that  $6 - 2Z + X - Y > 0$
5.  $\text{Cov}(X, W)$ , where  $W = X - Y$ .

*Solution*

1.

$$E(3X - 2Y + 4Z) = 3E(X) - 2E(Y) + 4E(Z) = (3 \times 2) + (-2 \times (-1)) + 4 \times 0 = 6 + 2 + 0 = 8$$

2.

$$V(2Y - 3Z) = 2^2 V(Y) + (-3)^2 V(Z) = (4 \times 2) + (9 \times 1) = 8 + 9 = 17$$

3. From the special property of normal RVs, the distribution of  $6 - 2Z + X - Y$  is

$$\begin{aligned} & \text{Normal}(6 + (-2 \times 0) + (1 \times 2) + (-1 \times -1), ((-2)^2 \times 1) + (1^2 \times 4) + ((-1)^2 \times 2)) \\ &= \text{Normal}(6 + 0 + 2 + 1, 4 + 4 + 2) \\ &= \text{Normal}(9, 10) \end{aligned}$$

4. Let  $U = 6 - 2Z + X - Y$  and we know  $U$  is  $\text{Normal}(9, 10)$  RV.

$$\begin{aligned} P(6 - 2Z + X - Y > 0) &= P(U > 0) = P(U - 9 > 0 - 9) = P\left(\frac{U - 9}{\sqrt{10}} > \frac{-9}{\sqrt{10}}\right) \\ &= P\left(Z > \frac{-9}{\sqrt{10}}\right) \\ &= P\left(Z < \frac{9}{\sqrt{10}}\right) \\ &\approx P(Z < 2.85) = 0.9978 \end{aligned}$$

5.

$$\begin{aligned} \text{Cov}(X, W) &= E(XW) - E(X)E(W) = E(X(X - Y)) - E(X)E(X - Y) \\ &= E(X^2 - XY) - E(X)(E(X) - E(Y)) = E(X^2) - E(XY) - 2 \times (2 - (-1)) \\ &= E(X^2) - E(X)E(Y) - 6 = E(X^2) - (2 \times (-1)) - 6 \\ &= (V(X) + (E(X))^2) + 2 - 6 = (4 + 2^2) - 4 = 4 \end{aligned}$$

### 3.10.4 Some Common $\mathbb{R}^m$ -valued RVs

So far, we have treated our random vectors as random points in  $\mathbb{R}^m$  and not been explicit about whether they are row or column vectors. We need to be more explicit now in order to perform arithmetic operations and transformations with them.

Let  $X = (X_1, X_2, \dots, X_{m_X})$  be a R $\vec{V}$  in  $\mathbb{R}^{1 \times m_X}$ , i.e.,  $X$  is a random row vector with 1 row and  $m_X$  columns, with JCDF  $F_{X_1, X_2, \dots, X_{m_X}}$  and JPDF  $f_{X_1, X_2, \dots, X_{m_X}}$ . Similarly, let  $Y = (Y_1, Y_2, \dots, Y_{m_Y})$  be a R $\vec{V}$  in  $\mathbb{R}^{1 \times m_Y}$ , i.e.,  $Y$  is a random row vector with 1 row and  $m_Y$  columns, with JCDF  $F_{Y_1, Y_2, \dots, Y_{m_Y}}$  and JPDF  $f_{Y_1, Y_2, \dots, Y_{m_Y}}$ . Let the JCDF of the random vectors  $X$  and  $Y$  together be  $F_{X_1, X_2, \dots, X_{m_X}, Y_1, Y_2, \dots, Y_{m_Y}}$  and JPDF be  $f_{X_1, X_2, \dots, X_{m_X}, Y_1, Y_2, \dots, Y_{m_Y}}$ .

#### Independent Random Vectors and their sums

The notion of mutual independence or joint independence of  $n$  random vectors is obtained similarly from ensuring the independence of any subset of the  $n$  vectors in terms of their JCDFs (JPMFs or JPDFs) being equal to the product of their marginal CDFs (PMFs or PDFs).

Thus, for a given  $m_X < \infty$  and  $m_Y < \infty$ , two **random vectors are independent** if and only if for any  $(x_1, x_2, \dots, x_{m_X}) \in \mathbb{R}^{1 \times m_X}$  and any  $(y_1, y_2, \dots, y_{m_Y}) \in \mathbb{R}^{1 \times m_Y}$

$$\begin{aligned} F_{X_1, X_2, \dots, X_{m_X}, Y_1, Y_2, \dots, Y_{m_Y}}(x_1, x_2, \dots, x_{m_X}, y_1, y_2, \dots, y_{m_Y}) \\ = F_{X_1, X_2, \dots, X_{m_X}}(x_1, x_2, \dots, x_{m_X}) \times F_{Y_1, Y_2, \dots, Y_{m_Y}}(y_1, y_2, \dots, y_{m_Y}) \end{aligned}$$

or, equivalently

$$\begin{aligned} f_{X_1, X_2, \dots, X_{m_X}, Y_1, Y_2, \dots, Y_{m_Y}}(x_1, x_2, \dots, x_{m_X}, y_1, y_2, \dots, y_{m_Y}) \\ = f_{X_1, X_2, \dots, X_{m_X}}(x_1, x_2, \dots, x_{m_X}) \times f_{Y_1, Y_2, \dots, Y_{m_Y}}(y_1, y_2, \dots, y_{m_Y}) \end{aligned}$$

Let us consider the natural two-dimensional analogue of the Bernoulli( $\theta$ ) 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  $\mathbb{R}^2$ :

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

Recall that vector addition and subtraction are done component-wise, i.e.  $(x_1, x_2) \pm (y_1, y_2) = (x_1 \pm y_1, x_2 \pm y_2)$ . We introduce a useful function called the indicator function of a set, say  $A$ .

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbb{1}_A(x)$  returns 1 if  $x$  belongs to  $A$  and 0 otherwise.

**Example 97** Let us recall the geometry and arithmetic of vector addition in the plane.

1. What is  $(1, 0) + (1, 0)$ ,  $(1, 0) + (0, 1)$ ,  $(0, 1) + (0, 1)$ ?
2. What is the relationship between  $(1, 0)$ ,  $(0, 1)$  and  $(1, 1)$  geometrically?
3. How does the diagonal of the parallelogram relate the its two sides in the geometry of addition in the plane?
4. What is  $(1, 0) + (0, 1) + (1, 0)$ ?

*Solution:*

1. addition is component-wise

$$(1, 0) + (1, 0) = (1 + 1, 0 + 0) = (2, 0)$$

$$(1, 0) + (0, 1) = (1 + 0, 0 + 1) = (1, 1)$$

$$(0, 1) + (0, 1) = (0 + 0, 1 + 1) = (0, 2)$$

2.  $(1, 0)$  and  $(0, 1)$  are vectors for the two sides of unit square and  $(1, 1)$  is its diagonal.

3. Generally, the diagonal of the parallelogram is the resultant or sum of the vectors representing its two sides

- 4.

$$(1, 0) + (0, 1) + (1, 0) = (1 + 0 + 1, 0 + 1 + 0) = (2, 1)$$

**Model 15** (Bernoulli( $\theta$ )  $\vec{RV}$ ) Given a parameter  $(\theta, 1 - \theta) \in \Delta^1$ , the unit 1-Simplex, we say that  $X := (X_1, X_2)$  is a Bernoulli( $\theta$ ) random vector ( $\vec{RV}$ ) 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  $\vec{RV}$   $X := (X_1, X_2)$  with realization  $x := (x_1, x_2)$  is:

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

**Example 98** Let us find the Expectation of Bernoulli( $\theta$ )  $\vec{RV}$  in Model 15.

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

**Remark 46** We can write the Binomial( $n, \theta$ ) RV  $Y$  as a Binomial( $n, \theta$ )  $\vec{RV}$   $X := (Y, n - Y)$ . In fact, this is the underlying model and the **bi** in the 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  $\theta$  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  $\theta$ . In other words, the  $\text{Binomial}(n, \theta)$  RV  $(Y, n - Y)$  is the sum of  $n$  IID Bernoulli( $\theta$ ) RVs  $X_1 := (X_{1,1}, X_{1,2}), X_2 := (X_{2,1}, X_{2,2}), \dots, X_n := (X_{n,1}, X_{n,2})$ :

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

**Exercise 3.36 (Random walk in the first Quadrant)** Consider an independent and identical random walk starting from  $(0, 0)$  in the first quadrant where you go east, i.e., add  $(1, 0)$  to your current position with probability  $\theta$ , and go north, i.e., add  $(0, 1)$  to your current position with probability  $1 - \theta$ . Suppose you take  $n$  such IID steps according to the Bernoulli( $\theta$ ) RV. Answer the following questions:

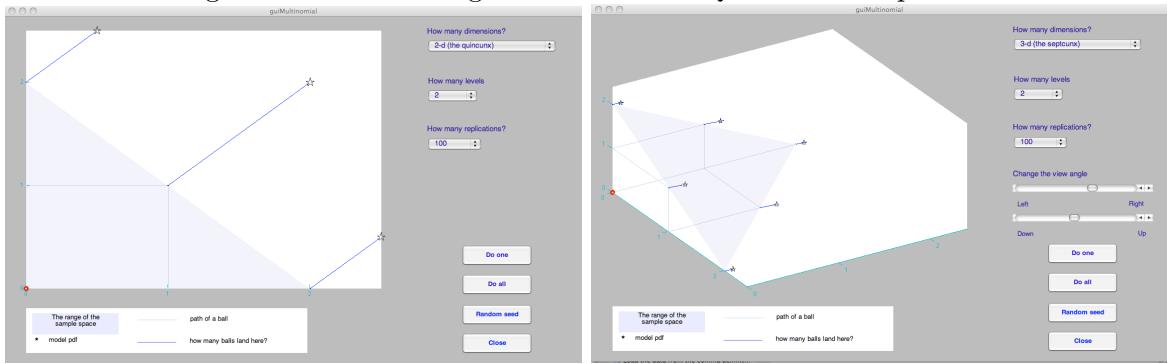
1. How does the number of paths that lead to a  $(x_1, x_2)$  with  $x_1 + x_2 = n$  relate to the binomial coefficient  $\binom{n}{x_1}$ ?
2. What is the probability of taking  $x_1$  steps east and  $x_2$  steps north?

**Exercise 3.37 (Random walks in the first Quadrant and Galton's Quincunx)** Compare the probability models for the Random walk in the first quadrant and Galton's Quincunx and explain how they are related.

**Labwork 99 (Quincunx Sampler Demo – Sum of  $n$  IID Bernoulli(1/2) RVs)** Let us understand the Quincunx construction of the  $\text{Binomial}(n, 1/2)$  RV  $X$  as the sum of  $n$  independent and identical Bernoulli(1/2) RVs by calling the interactive visual cognitive tool as follows:

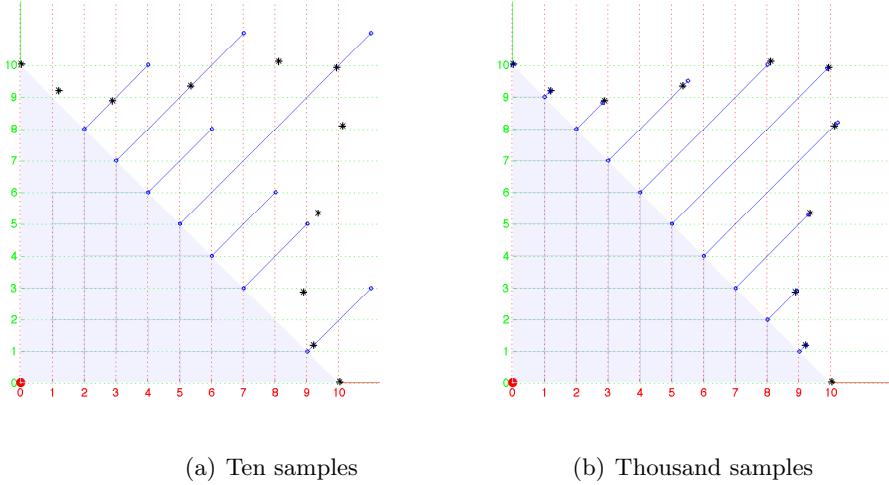
```
>> guiMultinomial
```

Figure 3.20: Visual Cognitive Tool GUI: Quincunx & Septcunx.



We are now ready to extend the  $\text{Binomial}(n, \theta)$  RV or  $\vec{Y}$  to its multivariate version called the  $\text{Multinomial}(n, \theta_1, \theta_2, \dots, \theta_k)$  RV  $\vec{Y}$ . We develop this RV as the sum of  $n$  IID de Moivre( $\theta_1, \theta_2, \dots, \theta_k$ ) RV that is defined next by extending de Moivre( $\theta_1, \theta_2, \dots, \theta_k$ ) RV taking values in  $\{1, 2, \dots, k\}$  of Model 14 to its vector-valued cousin taking values in  $\{e_1, e_2, \dots, e_k\}$ , the ortho-normal basis vectors in  $\mathbb{R}^k$ .

Figure 3.21: 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 PMF of our  $\text{Binomial}(n = 10, \theta = 0.5)$  RV, respectively.



**Model 16** (de Moivre( $\theta_1, \theta_2, \dots, \theta_k$ )  $\vec{\text{RV}}$ ) The PMF of the de Moivre( $\theta_1, \theta_2, \dots, \theta_k$ )  $\vec{\text{RV}}$   $X := (X_1, X_2, \dots, X_k)$  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:

$$f(x; \theta_1, \theta_2, \dots, \theta_k) := P(X = x) = \sum_{i=1}^k \theta_i \mathbf{1}_{\{e_i\}}(x) = \begin{cases} \theta_1 & \text{if } x = e_1 := (1, 0, \dots, 0) \in \mathbb{R}^k \\ \theta_2 & \text{if } x = e_2 := (0, 1, \dots, 0) \in \mathbb{R}^k \\ \vdots & \\ \theta_k & \text{if } x = e_k := (0, 0, \dots, 1) \in \mathbb{R}^k \\ 0 & \text{otherwise} \end{cases}$$

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

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

**Model 17** (Multinomial( $n, \theta_1, \theta_2, \dots, \theta_k$ )  $\vec{\text{RV}}$ ) We say that a  $\vec{\text{RV}}$   $Y := (Y_1, Y_2, \dots, Y_k)$  obtained from the sum of  $n$  IID de Moivre( $\theta_1, \theta_2, \dots, \theta_k$ )  $\vec{\text{RV}}$ s with realizations

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

has the PMF given by:

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

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 Binomial( $n, \theta_j$ ) for any  $j = 1, 2, \dots, k$ .

We can visualize the Multinomial( $n, \theta_1, \theta_2, \theta_3$ ) process as a sum of  $n$  IID de Moivre( $\theta_1, \theta_2, \theta_3$ ) R $\vec{V}$ 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  $\theta_1, \theta_2$  and  $\theta_3$ , respectively, in an IID manner at each of the  $n$  levels, before they collect at buckets placed at the integral points in the 3-simplex,  $\mathbb{Y} = \{(y_1, y_2, y_3) \in \mathbb{Z}_+^3 : \sum_{i=1}^3 y_i = n\}$ . Once again, we can visualize that the sum of  $n$  IID de Moivre( $\theta_1, \theta_2, \theta_3$ ) R $\vec{V}$ s constitute the Multinomial( $n, \theta_1, \theta_2, \theta_3$ ) R $\vec{V}$ .

**Labwork 100 (Septcunx Sampler Demo – Sum of n IID de Moivre(1/3, 1/3, 1/3) R $\vec{V}$ s)** Let us understand the Septcunx construction of the Multinomial( $n, 1/3, 1/3, 1/3$ ) R $\vec{V}X$  as the sum of  $n$  independent and identical de Moivre( $1/3, 1/3, 1/3$ ) R $\vec{V}$ s by calling the interactive visual cognitive tool as follows:

```
>> guiMultinomial
```

Multinomial distributions are at the very foundations of various machine learning algorithms, including, filtering junk email, learning from large knowledge-based resources like www, Wikipedia, word-net, etc.

**Model 18** (Normal( $\mu, \Sigma$ ) R $\vec{V}$ ) The univariate Normal( $\mu, \sigma^2$ ) RV has two parameters,  $\mu \in \mathbb{R}$  and  $\sigma^2 \in (0, \infty)$ . In the multivariate version  $\mu \in \mathbb{R}^{m \times 1}$  is a column vector and  $\sigma^2$  is replaced by a matrix  $\Sigma$ . To begin, let

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_m \end{pmatrix}$$

where,  $Z_1, Z_2, \dots, Z_m$  are jointly independent Normal( $0, 1$ ) RVs. Then the JPDF of  $Z$  is

$$f_Z(z) = f_{Z_1, Z_2, \dots, Z_m}(z_1, z_2, \dots, z_m) = \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2} \sum_{j=1}^m z_j^2\right) = \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2} z^T z\right)$$

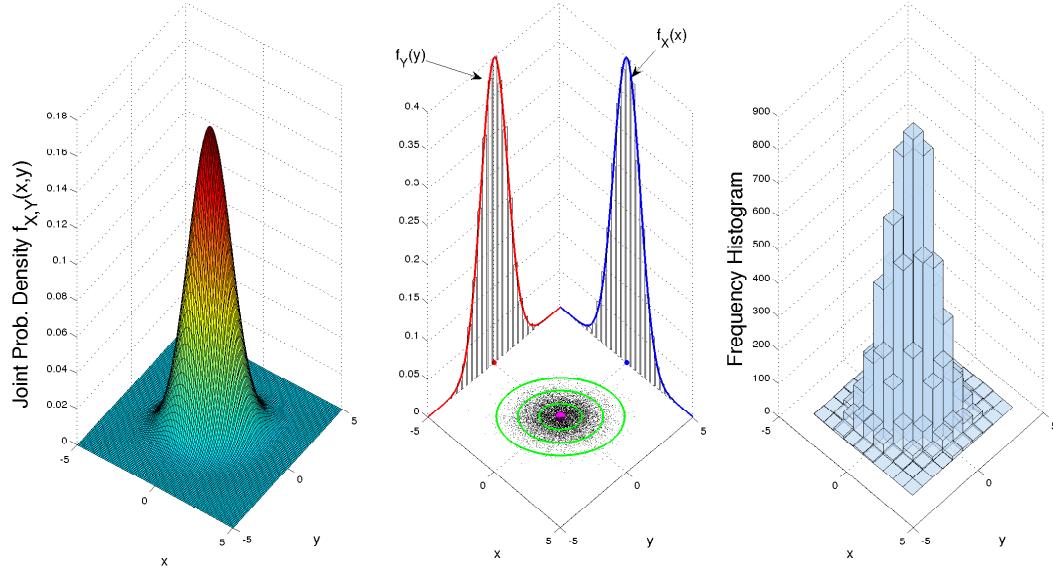
We say that  $Z$  has a standard multivariate normal distribution and write  $Z \sim \text{Normal}(0, I)$ , where it is understood that 0 represents the vector of  $m$  zeros and  $I$  is the  $m \times m$  identity matrix (with 1 along the diagonal entries and 0 on all off-diagonal entries).

More generally, a vector  $X$  has a multivariate normal distribution denoted by  $X \sim \text{Normal}(\mu, \Sigma)$ , if it has joint probability density function

$$f_X(x; \mu, \Sigma) = f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m; \mu, \Sigma) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

where  $|\Sigma|$  denotes the determinant of  $\Sigma$ ,  $\mu$  is a vector of length  $m$  and  $\Sigma$  is a  $m \times m$  symmetric, positive definite matrix. Setting  $\mu = 0$  and  $\Sigma = I$  gives back the standard multivariate normal R $\vec{V}$ .

Figure 3.22: JPDF, Marginal PDFs and Frequency Histogram of Bivariate Standard Normal RV.



When we have a non-zero mean vector

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} = \begin{pmatrix} 6.49 \\ 5.07 \end{pmatrix}$$

for the mean lengths and girths of cylindrical shafts from a manufacturing process with variance-covariance matrix

$$\Sigma = \begin{pmatrix} \text{Cov}(X, X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Cov}(Y, Y) \end{pmatrix} = \begin{pmatrix} V(X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & V(Y) \end{pmatrix} = \begin{pmatrix} 0.59 & 0.24 \\ 0.24 & 0.26 \end{pmatrix}$$

then the  $\text{Normal}(\mu, \Sigma)$  RV has JPDF, marginal PDFs and samples with frequency histograms as shown in Figure 3.23.

We can use MATLAB to compute for instance the probability that a cylinder has length and girth below 6.0 cms as follows:

```
>> mvncdf([6.0 6.0],[6.49 5.07],[0.59 0.24; 0.24 0.26])
ans =      0.2615
```

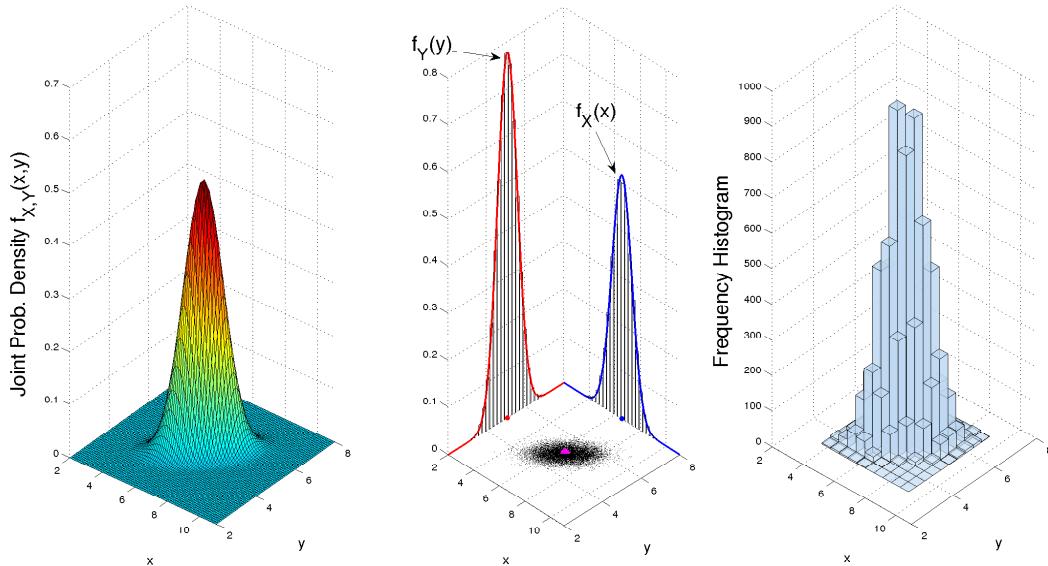
Or find the probability (with numerical error tolerance) that the cylinders are within the rectangular specifications of  $6 \pm 1.0$  along  $x$  and  $y$  as follows:

```
>> [F err] = mvncdf([5.0 5.0], [7.0 7.0], [6.49 5.07],[0.59 0.24; 0.24 0.26])
F =      0.3352
err =    1.0000e-08
```

### 3.10.5 Dependent Random Variables

When a sequence of RVs are not independent they are said to be **dependent**. The simplest form of dependence is *Markov dependence* that we will briefly see via a couple examples in Chapter 6.

Figure 3.23: JPDF, Marginal PDFs and Frequency Histogram of a Bivariate Normal R $\vec{V}$  for lengths of girths of cylindrical shafts in a manufacturing process (in cm).



### 3.11 Exercises in Multivariate Random Variables

**Ex. 3.38** — Find the probability that none of the three bulbs in a traffic signal, that are assumed to have independent life-times (i.e., the time during which they are operational), need to be replaced during the first 1200 hours of operation if the length of time before a single bulb needs to be replaced is a continuous random variable  $X$  with density

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

Note:  $X$  is measured in multiples of 1000 hours.

**Ex. 3.39** — Let  $(X, Y)$  be a continuous R $\vec{V}$  with joint probability density function (JPDF)

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

Find the following:

1. the normalizing constant  $a$  which will ensure  $P(\Omega) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$
2.  $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy$  called the marginal probability density function (MPDF) of  $X$
3.  $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx$  called the marginal probability density function (MPDF) of  $Y$
4. Check if  $f_X(x)f_Y(y) = f_{X,Y}(x,y)$  for every  $(x,y)$  and decide whether  $X$  and  $Y$  are independent random variables. Hint:  $X$  and  $Y$  are said to be independent if  $f_X(x)f_Y(y) = f_{X,Y}(x,y)$  for every  $(x,y)$ .
5.  $F_{X,Y}(x,y)$ , the joint cumulative distribution function (JCDF) of  $(X, Y)$  for any  $(x,y) \in (0,1) \times (0,1)$
6. the probability that  $X > 0.5$  and  $Y < 0.6$ , i.e.,  $P(X > 0.5, Y < 0.6)$

7.  $E(X)$ , the expectation of  $X$  or the first moment of  $X$
8.  $E(Y)$ , the expectation of  $Y$  or the first moment of  $Y$
9.  $E(XY)$ , the expectation of  $XY$
10.  $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$ , the covariance of  $X$  and  $Y$ .

**Ex. 3.40** — Logs are milled to have a width of  $\mu$ . The actual width of a randomly selected item is  $X$ . If  $X$  is a  $\text{Normal}(\mu, \sigma^2)$  random variable then find the probability density function of the *squared-error* of the milling process,

$$Y = (X - \mu)^2.$$

**Ex. 3.41** — Let  $(X, Y)$  be a discrete random vector ( $\vec{RV}$ ) with support:

$$\mathcal{S}_{X,Y} = \{(0,0), (0,1), (1,0), (1,1)\} .$$

Let its joint probability mass function (JPMF) be:

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{4} & \text{if } (x,y) = (0,0) \\ \frac{1}{4} & \text{if } (x,y) = (0,1) \\ \frac{1}{4} & \text{if } (x,y) = (1,0) \\ \frac{1}{4} & \text{if } (x,y) = (1,1) \\ 0 & \text{otherwise} . \end{cases}$$

Are  $X$  and  $Y$  independent?

**Ex. 3.42** — A semiconductor product consists of three layers that are fabricated independently. If the variances in thickness of the first, second and third layers are 25, 40 and 30 nanometers squared, what is the variance of the thickness of the final product?

**Ex. 3.43** — Find the covariance for the discrete  $\vec{RV}$   $(X, Y)$  with joint probability mass function

$$f_{X,Y}(x,y) = \begin{cases} 0.2 & \text{if } (x,y) = (0,0) \\ 0.1 & \text{if } (x,y) = (1,1) \\ 0.1 & \text{if } (x,y) = (1,2) \\ 0.1 & \text{if } (x,y) = (2,1) \\ 0.1 & \text{if } (x,y) = (2,2) \\ 0.4 & \text{if } (x,y) = (3,3) \\ 0 & \text{otherwise} . \end{cases}$$

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

**Ex. 3.44** — Consider two random variables (RVs)  $X$  and  $Y$  having marginal distribution functions

$$F_X(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{1}{2} & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1 - \frac{1}{2}e^{-y} - \frac{1}{2}e^{-2y} & \text{if } y \geq 0 \end{cases}$$

If  $X$  and  $Y$  are independent, what is their joint distribution function  $F_{X,Y}(x,y)$ ? [Hint: you need to express  $F_{X,Y}(x,y)$  for any  $(x,y) \in \mathbb{R}^2$ .]

**Ex. 3.45** — Let  $(X, Y)$  be a continuous RV with joint probability density function (JPDF):

$$f_{X,Y}(x, y) = \begin{cases} e^{-x} & \text{if } x \in [0, \infty) \text{ and } y \in [2, 3] \\ 0 & \text{otherwise .} \end{cases}$$

Are  $X$  and  $Y$  independent?

**Ex. 3.46** — In an electronic assembly, let the RVs  $X_1, X_2, X_3, X_4$  denote the lifetimes of four components in hours. Suppose that the JPDF of these variables is

$$\begin{aligned} f_{X_1, X_2, X_3, X_4}(x_1, x_2, x_3, x_4) \\ = \begin{cases} 9 \times 10^{-12} e^{-0.001x_1 - 0.002x_2 - 0.0015x_3 - 0.003x_4} & \text{if } x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0 \\ 0 & \text{otherwise .} \end{cases} \end{aligned}$$

What is the probability that the device operates for more than 1000 hours without any failures? [Hint: The requested probability is  $P(X_1 > 1000, X_2 > 1000, X_3 > 1000, X_4 > 1000)$  since each one of the four components of the device must not fail before 1000 hours.]

**Ex. 3.47** — Suppose the RVs  $Y_1$ ,  $Y_2$  and  $Y_3$  represent the thickness in micrometers of a substrate, an active layer, and a coating layer of a chemical product. Assume  $Y_1$ ,  $Y_2$  and  $Y_3$  are  $\text{Normal}(10000, 250^2)$ ,  $\text{Normal}(1000, 20^2)$  and  $\text{Normal}(80, 4^2)$  RVs, respectively. Further suppose that they are independent. The required specifications for the thickness of the substrate, active layer and coating layer are  $[9500, 10500]$ ,  $[950, 1050]$  and  $[75, 85]$ , respectively. What proportion of chemical products meets all thickness specifications? [Hint: this is just  $P(9500 < Y_1 < 10500, 950 < Y_2 < 1050, 75 < Y_3 < 85)$ ] Which one of the three thicknesses has the least probability of meeting specifications?

**Ex. 3.48** — Soft drink cans are filled by an automated filling machine. Assume the fill volumes of the cans are independent  $\text{Normal}(12.1, 0.01)$  RVs. What is the probability that the average volume of ten cans selected from this process is less than 12.01 fluid ounces?

**Ex. 3.49** — Let  $X_1, X_2, X_3, X_4$  be RVs that denote the number of bits received in a digital channel that are classified as *excellent*, *good*, *fair* and *poor*, respectively. In a transmission of 10 bits, what is the probability that 6 of the bits received are *excellent*, 2 are *good*, 2 are *fair* and none are *poor* under the assumption that the classification of bits are independent events and that the probabilities of each bit being *excellent*, *good*, *fair* and *poor* are 0.6, 0.3, 0.08 and 0.02, respectively. [Hint: Think of  $\text{Multinomial}(n = 10, \theta_1 = 0.6, \theta_2 = 0.3, \theta_3 = 0.08, \theta_4 = 0.02)$  as a model for bit classification in this digital channel.]

## 3.12 Characteristic Functions

The characteristic function (CF) of a random variable gives another way to specify its distribution. Thus CF is a powerful tool for analytical results involving random variables (more).

**Definition 47 (Characteristic Function (CF))** Let  $X$  be a RV and  $\imath = \sqrt{-1}$ . The function  $\varphi_X(t) : \mathbb{R} \rightarrow \mathbb{C}$  defined by

$$\varphi_X(t) := E(\exp(\imath tX)) = \begin{cases} \sum_x \exp(\imath tx) f_X(x) & \text{if } X \text{ is discrete RV} \\ \int_{-\infty}^{\infty} \exp(\imath tx) f_X(x) dx & \text{if } X \text{ is continuous RV} \end{cases} \quad (3.70)$$

is called the **characteristic function** of  $X$ .

NOTE:  $\varphi_X(t)$  exists for any  $t \in \mathbb{R}$ , because

$$\begin{aligned} \varphi_X(t) &= E(\exp(\imath tX)) \\ &= E(\cos(tX) + \imath \sin(tX)) \\ &= E(\cos(tX)) + \imath E(\sin(tX)) \end{aligned}$$

and the last two expected values are well-defined, because the sine and cosine functions are bounded by  $[-1, 1]$ .

For a continuous RV,  $\int_{-\infty}^{\infty} \exp(-\imath tx) f_X(x) dx$  is called the *Fourier transform* of  $f_X$ . This is the CF but with  $t$  replaced by  $-t$ . You will also encounter Fourier transforms when solving differential equations.

### 3.12.1 Obtaining Moments from Characteristic Function

Recall that the  $k$ -th moment of  $X$  is  $E(X^k)$  for any  $k \in \mathbb{N} := \{1, 2, 3, \dots\}$  is

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

The characteristic function can be used to derive the moments of  $X$  due to the following nice relationship between the the  $k$ -th moment of  $X$  and the  $k$ -th derivative of the CF of  $X$ .

**Proposition 48 (Moment & CF.)** Let  $X$  be a random variable and  $\varphi_X(t)$  be its CF. If  $E(X^k)$  exists and is finite, then  $\varphi_X(t)$  is  $k$  times continuously differentiable and

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

where  $\left[ \frac{d^k \varphi_X(t)}{dt^k} \right]_{t=0}$  is the  $k$ -th derivative of  $\varphi_X(t)$  with respect to  $t$ , evaluated at the point  $t = 0$ .

**Proof:** The proper proof is very messy so we just give a sketch of the ideas in the proof. Due to the linearity of the expectation (integral) and the derivative operators, we can change the order of operations:

$$\frac{d^k \varphi_X(t)}{dt^k} = \frac{d^k}{dt^k} E(\exp(itX)) = E\left(\frac{d^k}{dt^k} \exp(itX)\right) = E\left((it)^k \exp(itX)\right) = it^k E\left(X^k \exp(itX)\right)$$

The RHS evaluated at  $t = 0$  is

$$\left[ \frac{d^k \varphi_X(t)}{dt^k} \right]_{t=0} = \left[ it^k E\left(X^k \exp(itX)\right) \right]_{t=0} = it^k E\left(X^k\right)$$

This completes the sketch of the proof.

The above Theorem gives us the relationship between the moments and the derivatives of the CF if we already know that the moment exists. When one wants to compute a moment of a random variable, what we need is the following Theorem.

**Proposition 49 (Moments from CF.)** Let  $X$  be a random variable and  $\varphi_X(t)$  be its CF. If  $\varphi_X(t)$  is  $k$  times differentiable at the point  $t = 0$ , then

1. if  $k$  is even, the  $n$ -th moment of  $X$  exists and is finite for any  $0 \leq n \leq k$ ;
2. if  $k$  is odd, the  $n$ -th moment of  $X$  exists and is finite for any  $0 \leq n \leq k - 1$ .

In both cases,

$$E(X^k) = \frac{1}{i^k} \left[ \frac{d^k \varphi_X(t)}{dt^k} \right]_{t=0}. \quad (3.71)$$

where  $\left[ \frac{d^k \varphi_X(t)}{dt^k} \right]_{t=0}$  is the  $k$ -th derivative of  $\varphi_X(t)$  with respect to  $t$ , evaluated at the point  $t = 0$ .

**Proof:** For proof see e.g., Ushakov, N. G. (1999) Selected topics in characteristic functions, VSP (p. 39).

**Example 101** Let  $X$  be the Bernoulli( $\theta$ ) RV. Find the CF of  $X$ . Then use CF to find  $E(X)$ ,  $E(X^2)$  and from this obtain the variance  $V(X) = E(X^2) - (E(X))^2$ .

Solution:

### Part 1

Recall the PMF for this discrete RV with parameter  $\theta \in (0, 1)$  is

$$f_X(x; \theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Let's first find the CF of  $X$

$$\begin{aligned} \varphi_X(t) &= E(\exp(itX)) = \sum_x \exp(itx) f_X(x; \theta) \quad \text{By Defn. in Equation (3.70)} \\ &= \exp(it \times 0)(1 - \theta) + \exp(it \times 1)\theta = \exp(0)(1 - \theta) + \exp(it)\theta = 1 - \theta + \theta \exp(it) \end{aligned}$$

**Part 2:**

Let's differentiate CF

$$\frac{d}{dt}\varphi_X(t) = \frac{d}{dt}(1 - \theta + \theta e^{it}) = \theta i \exp(it)$$

We get  $E(X)$  by evaluating  $\frac{d}{dt}\varphi_X(t)$  at  $t = 0$  and dividing by  $i$  according to Equation (3.71) as follows:

$$E(X) = \frac{1}{i} \left[ \frac{d}{dt}\varphi_X(t) \right]_{t=0} = \frac{1}{i} [\theta i \exp(it)]_{t=0} = \frac{1}{i} (\theta i \exp(i0)) = \theta .$$

Similarly from Equation (3.71) we can get  $E(X^2)$  as follows:

$$\begin{aligned} E(X^2) &= \frac{1}{i^2} \left[ \frac{d^2}{dt^2}\varphi_X(t) \right]_{t=0} = \frac{1}{i^2} \left[ \frac{d}{dt} \frac{d}{dt}\varphi_X(t) \right]_{t=0} = \frac{1}{i^2} \left[ \frac{d}{dt} \theta i \exp(it) \right]_{t=0} \\ &= \frac{1}{i^2} [\theta i^2 \exp(it)]_{t=0} = \frac{1}{i^2} (\theta i^2 \exp(i0)) = \theta . \end{aligned}$$

Finally, from the first and second moments we can get the variance as follows:

$$V(X) = E(X^2) - (E(X))^2 = \theta - \theta^2 = \theta(1 - \theta) .$$

Let's check that this is what we have as variance for the Bernoulli( $\theta$ ) RV if we directly computed it using weighted sums in the definition of expectations:  $E(X) = 1 \times \theta + 0 \times (1 - \theta) = \theta$ ,  $E(X^2) = 1^2 \times \theta + 0^2 \times (1 - \theta) = \theta$  and thus giving the same  $V(X) = E(X^2) - (E(X))^2 = \theta - \theta^2 = \theta(1 - \theta)$ .

**Example 102** Let  $X$  be an Exponential( $\lambda$ ) RV. First show that its CF is  $\lambda/(\lambda - it)$ . Then use CF to find  $E(X)$ ,  $E(X^2)$  and from this obtain the variance  $V(X) = E(X^2) - (E(X))^2$ .

Solution:

Recall that the PDF of an Exponential( $\lambda$ ) RV for a given parameter  $\lambda \in (0, \infty)$  is  $\lambda e^{-\lambda x}$  if  $x \in [0, \infty)$  and 0 if  $x \notin [0, \infty)$ .

**Part 1:** Find the CF.

We will use the fact that

$$\int_0^\infty \alpha e^{-\alpha x} dx = [-e^{-\alpha x}]_0^\infty = 1$$

$$\begin{aligned} \varphi_X(t) &= E(\exp(itX)) = E(e^{itX}) = \int_{-\infty}^\infty e^{itx} \lambda e^{-\lambda x} dx = \lambda \int_0^\infty e^{-(\lambda - it)x} dx \\ &= \frac{\lambda}{\lambda - it} \int_0^\infty (\lambda - it)e^{-(\lambda - it)x} dx = \frac{\lambda}{\lambda - it} \int_0^\infty \alpha e^{-\alpha x} dx = \frac{\lambda}{\lambda - it} , \end{aligned}$$

where  $\alpha = \lambda - it$  with  $\lambda > 0$ .

Alternatively, you can use  $e^{itx} = \cos(tx) + i \sin(tx)$  and do integration by parts to arrive at the same answer starting from:

$$\varphi_X(t) = \int_{-\infty}^\infty e^{itx} \lambda e^{-\lambda x} dx = \int_{-\infty}^\infty \cos(tx) e^{-\lambda x} dx + i \int_{-\infty}^\infty \sin(tx) e^{-\lambda x} dx = \frac{\lambda}{\lambda - it} .$$

**Part 2:**

Let us differentiate the CF to get moments using Equation (3.71) (CF has to be once and twice differentiable at  $t = 0$  to get the first and second moments).

$$\begin{aligned}\frac{d}{dt}\varphi_X(t) &= \frac{d}{dt}\left(\frac{\lambda}{\lambda - it}\right) = \lambda\left(-1 \times (\lambda - it)^{-2} \times \frac{d}{dt}(\lambda - it)\right) \\ &= \lambda\left(\frac{-1}{(\lambda - it)^2} \times (-i)\right) = \frac{\lambda i}{(\lambda - it)^2}\end{aligned}$$

We get  $E(X)$  by evaluating  $\frac{d}{dt}\varphi_X(t)$  at  $t = 0$  and dividing by  $i$  according to Equation (3.71) as follows:

$$E(X) = \frac{1}{i} \left[ \frac{d}{dt}\varphi_X(t) \right]_{t=0} = \frac{1}{i} \left[ \frac{\lambda i}{(\lambda - it)^2} \right]_{t=0} = \frac{1}{i} \left( \frac{\lambda i}{\lambda^2} \right) = \frac{1}{i} \left( \frac{i}{\lambda} \right) = \frac{1}{\lambda}$$

Let's pause and see if this makes sense.... Yes, because the expected value of Exponential( $\lambda$ ) RV is indeed  $1/\lambda$  (recall from when we introduced this RV).

Similarly from Equation (3.71) we can get  $E(X^2)$  as follows:

$$\begin{aligned}E(X^2) &= \frac{1}{i^2} \left[ \frac{d^2}{dt^2}\varphi_X(t) \right]_{t=0} = \frac{1}{i^2} \left[ \frac{d}{dt} \frac{d}{dt}\varphi_X(t) \right]_{t=0} = \frac{1}{i^2} \left[ \frac{d}{dt} \frac{\lambda i}{(\lambda - it)^2} \right]_{t=0} \\ &= \frac{1}{i^2} \left[ \lambda i \times \frac{d}{dt}(\lambda - it)^{-2} \right]_{t=0} = \frac{1}{i^2} \left[ \lambda i \left( -2(\lambda - it)^{-3} \frac{d}{dt}(\lambda - it) \right) \right]_{t=0} \\ &= \frac{1}{i^2} \left[ \lambda i (-2(\lambda - it)^{-3} \times (-i)) \right]_{t=0} = \frac{1}{i^2} \left[ \frac{2\lambda i^2}{(\lambda - it)^3} \right]_{t=0} = \frac{1}{i^2} \left( \frac{2\lambda i^2}{\lambda^3} \right) = \frac{2}{\lambda^2}.\end{aligned}$$

Finally, from the first and second moments we can get the variance as follows:

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

Let's check that this is what we had as variance for the Exponential( $\lambda$ ) RV when we first introduced it and directly computed using integrals for definition of expectation.

Characteristic functions can be used to characterize the distribution of a random variable.

Two RVs  $X$  and  $Y$  have the same DFs , i.e.,  $F_X(x) = F_Y(x)$  for all  $x \in \mathbb{R}$ , if and only if they have the same characteristic functions, i.e.  $\varphi_X(t) = \varphi_Y(t)$  for all  $t \in \mathbb{R}$  (for proof see Resnick, S. I. (1999) A Probability Path, Birkhauser).

Thus, if we can show that two RVs have the same CF then we know they are the same. This can be much more challenging or impossible to do directly with their DFs.

Let  $Z$  be  $\text{Normal}(0, 1)$ , the standard normal RV. We can find the CF for  $Z$  using couple of tricks as follows

$$\begin{aligned}\varphi_Z(t) &= E(e^{itZ}) \\ &= \int_{-\infty}^{\infty} e^{itz} f_Z(z) dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itz} e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itz - z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(t^2 + (z-it)^2)/2} dz = e^{-t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z-it)^2/2} dz \\ &= e^{-t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy \quad \text{substituting } y = z - it, dy = dz \\ &= e^{-t^2/2} \frac{1}{\sqrt{2\pi}} \sqrt{2\pi} \quad \text{using the normalizing constant in PDF of Normal}(0, 1) \text{ RV} \\ &= e^{-t^2/2}\end{aligned}$$

Thus the CF of the standard normal RV  $Z$  is

$$\boxed{\varphi_Z(t) = e^{-t^2/2}} \quad (3.72)$$

Let  $X$  be a RV with CF  $\varphi_X(t)$ . Let  $Y$  be a linear transformation of  $X$

$$Y = a + bX$$

where  $a$  and  $b$  are two constant real numbers and  $b \neq 0$ . Then the CF of  $Y$  is

$$\boxed{\varphi_Y(t) = \exp(iat)\varphi_X(bt)} \quad (3.73)$$

**Proof:** This is easy to prove using the definition of CF as follows:

$$\begin{aligned} \varphi_Y(t) &= E(\exp(itY)) = E(\exp(it(a+bX))) = E(\exp(ita+itbX)) \\ &= E(\exp(ita)\exp(itbX)) = \exp(ita)E(\exp(itbX)) = \exp(ita)\varphi_X(bt) \end{aligned}$$

**Example 103** Let  $Y$  be a  $\text{Normal}(\mu, \sigma^2)$  RV. Recall that  $Y$  is a linear transformation of  $Z$ , i.e.,  $Y = \mu + \sigma Z$  where  $Z$  is a  $\text{Normal}(0, 1)$  RV. Using Equations (3.72) and (3.73) find the CF of  $Y$ .

Solution:

$$\begin{aligned} \varphi_Y(t) &= \exp(i\mu t)\varphi_Z(\sigma t), \quad \text{since } Y = \mu + \sigma Z \\ &= e^{i\mu t}e^{(-\sigma^2 t^2)/2}, \quad \text{since } \varphi_Z(t) = e^{-t^2/2} \\ &= e^{i\mu t - (\sigma^2 t^2)/2} \end{aligned}$$

A generalization of (3.73) is the following. If  $X_1, X_2, \dots, X_n$  are independent RVs and  $a_1, a_2, \dots, a_n$  are some constants, then the CF of the linear combination  $Y = \sum_{i=1}^n a_i X_i$  is

$$\boxed{\varphi_Y(t) = \varphi_{X_1}(a_1 t) \times \varphi_{X_2}(a_2 t) \times \cdots \times \varphi_{X_n}(a_n t) = \prod_{i=1}^n \varphi_{X_i}(a_i t)} \quad (3.74)$$

**Example 104** Using the following three facts:

- Eqn. (3.74)
- the Binomial( $n, \theta$ ) RV  $Y$  is the sum of  $n$  independent Bernoulli( $\theta$ ) RVs (from Probability Course)
- the CF of Bernoulli( $\theta$ ) RV (from lecture notes for Inference Course)

find the CF of the Binomial( $n, \theta$ ) RV  $Y$ .

Solution:

Let  $X_1, X_2, \dots, X_n$  be independent Bernoulli( $\theta$ ) RVs with CF  $(1 - \theta + \theta e^{it})$  then  $Y = \sum_{i=1}^n X_i$  is the Binomial( $n, \theta$ ) RV and by Eqn. (3.74) with  $a_1 = a_2 = \cdots = 1$ , we get

$$\varphi_Y(t) = \varphi_{X_1}(t) \times \varphi_{X_2}(t) \cdots \varphi_{X_n}(t) = \prod_{i=1}^n \varphi_{X_i}(t) = \prod_{i=1}^n (1 - \theta + \theta e^{it}) = (1 - \theta + \theta e^{it})^n .$$

**Example 105** Let  $Z_1$  and  $Z_2$  be independent  $\text{Normal}(0, 1)$  RVs.

1. Use Eqn. (3.74) to find the CF of  $Z_1 + Z_2$ .
2. From the CF of  $Z_1 + Z_2$  identify what RV it is.
3. Use Eqn. (3.74) to find the CF of  $2Z_1$ .
4. From the CF of  $2Z_1$  identify what RV it is.
5. Try to understand the difference between the distributions of  $Z_1 + Z_2$  and  $2Z_1$  inspite of  $Z_1$  and  $Z_2$  having the same distribution.

Hint: from lectures we know that  $\varphi_X(t) = e^{i\mu t - (\sigma^2 t^2)/2}$  for a  $\text{Normal}(\mu, \sigma^2)$  RV  $X$ .

Solution:

1. By Eqn. (3.74) we just multiply the characteristic functions of  $Z_1$  and  $Z_2$ , both of which are  $e^{-t^2/2}$ ,
$$\varphi_{Z_1+Z_2}(t) = \varphi_{Z_1}(t) \times \varphi_{Z_2}(t) = e^{-t^2/2} \times e^{-t^2/2} = e^{-2t^2/2} = e^{-t^2} .$$
2. The CF of  $Z_1 + Z_2$  is that of the  $\text{Normal}(\mu, \sigma^2)$  RV with  $\mu = 0$  and  $\sigma^2 = 2$ . Thus  $Z_1 + Z_2$  is the  $\text{Normal}(0, 2)$  RV with mean parameter  $\mu = 0$  and variance parameter  $\sigma^2 = 2$ .
3. We can again use Eqn. (3.74) to find the CF of  $2Z_1$  as follows

$$\varphi_{2Z_1} = \varphi_{Z_1}(2t) = e^{-2^2 t^2/2} .$$

4. The CF of  $2Z_1$  is that of the  $\text{Normal}(\mu, \sigma^2)$  RV with  $\mu = 0$  and  $\sigma^2 = 2^2 = 4$ . Thus  $2Z_1$  is the  $\text{Normal}(0, 4)$  RV with mean parameter  $\mu = 0$  and variance parameter  $\sigma^2 = 4$ .
5.  $2Z_1$  has a bigger variance from multiplying the standard normal RV by 2 while  $Z_1 + Z_2$  has a smaller variance from adding two independent standard normal RVs. Thus, the result of adding the same RV twice does not have the same distribution as that of multiplying it by 2. In other words  $2 \times Z$  is not equal to  $Z + Z$  in terms of its probability distribution!

### 3.12.2 Moment Generating Function

Moment generating functions are special cases of characteristic functions and we won't be explicitly using them here as it is more convenient to work in the complex plane.

## 3.13 Exercises in Characteristic Functions

**Ex. 3.50 —** Let  $X$  be a discrete random variable (RV) with probability mass function (PMF)

$$f_X(x) = \begin{cases} \frac{1}{3} & \text{if } x = 0 \\ \frac{1}{3} & \text{if } x = 1 \\ \frac{1}{3} & \text{if } x = 2 \\ 0 & \text{otherwise} . \end{cases}$$

1. Find the characteristic function (CF) of  $X$
2. Using the CF find  $V(X)$ , the variance of  $X$ . Hint:  $V(X) = E(X^2) - (E(X))^2$

**Ex. 3.51** — Recall that the Geometric( $\theta$ ) RV  $X$  has the following PMF

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

1. Find the CF of  $X$ . (Hint: the sum of the infinite geometric series  $\sum_{x=0}^{\infty} ar^x = \frac{a}{1-r}$ .)
2. Using the CF find  $E(X)$ .

**Ex. 3.52** — Let  $X$  be the Uniform( $a, b$ ) RV with the following probability density function (PDF)

$$f_X(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

Find the CF of  $X$ .

**Ex. 3.53** — Recall that the Poisson( $\lambda$ ) RV has the following PMF

$$f_X(x; \lambda) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{if } x \in \{0, 1, 2, 3, \dots\} \\ 0 & \text{otherwise} \end{cases}$$

Hint: the power series of  $e^\alpha = \sum_{x=0}^{\infty} \frac{\alpha^x}{x!}$ .

1. Find the CF of  $X$ .
2. Find the variance of  $X$  using its CF.

**Ex. 3.54** — Let  $X$  be a Poisson( $\lambda$ ) RV and  $Y$  be another Poisson( $\mu$ ) RV. Suppose  $X$  and  $Y$  are independent. Use Eqn. (3.74) to first find the CF of the RV  $W = X + Y$ . From the CF of  $W$  try to identify what RV it is.

**Ex. 3.55** — Recall from lecture that if  $Y = a + bX$  for some constants  $a$  and  $b$  with  $b \neq 0$  then  $\varphi_Y(t) = e^{iat} \varphi_X(bt)$  and that  $\varphi_Z(t) = e^{-t^2/2}$  if  $Z$  is the Normal( $0, 1$ ) RV. Using these facts find the CF of  $-Z$ , the RV obtained from  $Z$  by simply switching its sign. From the CF of  $-Z$  identify what RV it is.

## 3.14 Statistics

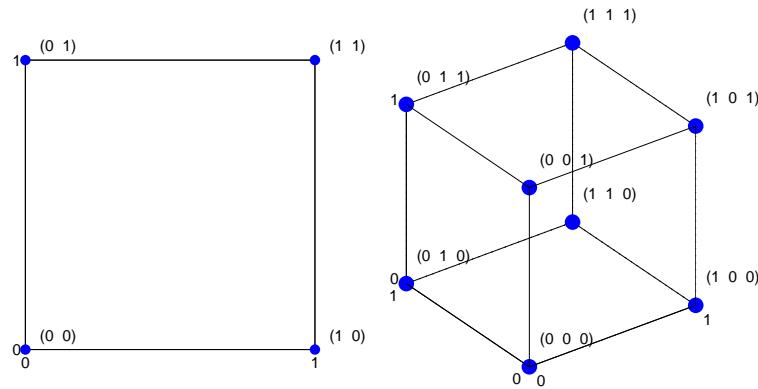
### 3.14.1 Data and Statistics

**Definition 50 (Data)** The function  $X$  measures the outcome  $\omega$  of an experiment with sample space  $\Omega$  [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 3.24: Sample Space, Random Variable, Realisation, Data, and Data Space.

Figure 3.25: Data Spaces  $\mathbb{X} = \{0, 1\}^2$  and  $\mathbb{X} = \{0, 1\}^3$  for two and three Bernoulli trials, respectively.

**Example 106 (Tossing a coin  $n$  times)** For some given parameter  $\theta \in \Theta := [0, 1]$ , consider  $n$  IID  $\text{Bernoulli}(\theta)$  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  $(x_1, x_2, \dots, x_{n-1}, x_n) = (0, 0, \dots, 0, 1)$ .

**Definition 51 (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 107 (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  $E(X_1)$  and  $V(X_1)$ , respectively.

**Definition 52 (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:

$$T_n( (X_1, X_2, \dots, X_n) ) = \bar{X}_n( (X_1, X_2, \dots, X_n) ) := \frac{1}{n} \sum_{i=1}^n X_i . \quad (3.75)$$

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:

$$\begin{aligned} E(\bar{X}_n) &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) && [\text{by definition (3.75)}] \\ &= \frac{1}{n} \sum_{i=1}^n E(X_i) && [\text{by property (3.50)}] \end{aligned}$$

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  $E(X_1)$ , then:

$$E(\bar{X}_n) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n E(X_1) = \frac{1}{n} n E(X_1) = E(X_1) . \quad (3.76)$$

Similarly, we can show that:

$$\begin{aligned} V(\bar{X}_n) &= V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) && [\text{by definition (3.75)}] \\ &= \left(\frac{1}{n}\right)^2 V\left(\sum_{i=1}^n X_i\right) && [\text{by property (3.49)}] \end{aligned}$$

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

$$V(\bar{X}_n) = \left(\frac{1}{n}\right)^2 V\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(X_i) \quad [\text{by property (3.51)}]$$

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

$$V(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^n V(X_i) = \frac{1}{n^2} \sum_{i=1}^n V(X_1) = \frac{1}{n^2} n V(X_1) = \frac{1}{n} V(X_1) . \quad (3.77)$$

**Labwork 108 (Sample mean)** After initializing the fundamental sampler, we draw five samples and then obtain the sample mean using the MATLAB function `mean`. In the following, we will reuse the samples stored in the array `XsFromUni01Twstr101`.

```
>> rand('twister',101); % initialise the fundamental Uniform(0,1) sampler
>> XsFromUni01Twstr101=rand(1,5); % simulate n=5 IID samples from Uniform(0,1) RV
>> SampleMean=mean(XsFromUni01Twstr101);% find sample mean
>> disp(XsFromUni01Twstr101); % The data-points x_1,x_2,x_3,x_4,x_5 are:
    0.5164    0.5707    0.0285    0.1715    0.6853
>> disp(SampleMean); % The Sample mean is :
    0.3945
```

We can thus use `mean` to obtain the sample mean  $\bar{x}_n$  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 53 (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 :

$$T_n( (X_1, X_2, \dots, X_n) ) = S_n^2( (X_1, X_2, \dots, X_n) ) := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 . \quad (3.78)$$

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:

$$S_n( (X_1, X_2, \dots, X_n) ) = \sqrt{S_n^2( (X_1, X_2, \dots, X_n) )} \quad (3.79)$$

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:

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

**Labwork 109 (Sample variance and sample standard deviation)** We can compute the sample variance and sample standard deviation for the five samples stored in the array `XsFromUni01Twstr101` from Labwork 108 using MATLAB's functions `var` and `std`, respectively.

```
>> disp(XsFromUni01Twstr101); % The data-points x_1,x_2,x_3,x_4,x_5 are :
    0.5164    0.5707    0.0285    0.1715    0.6853
>> SampleVar=var(XsFromUni01Twstr101);% find sample variance
>> SampleStd=std(XsFromUni01Twstr101);% find sample standard deviation
>> disp(SampleVar) % The sample variance is:
    0.0785
>> disp(SampleStd) % The sample standard deviation is:
    0.2802
```

It is important to bear in mind that the statistics such as sample mean and sample variance are random variables and have an underlying distribution.

**Definition 54 (Order Statistics)** Suppose  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} F$ , where  $F$  is the DF from the set of all DFs over the real line. Then, the  $n$ -sample **order statistics**  $X_{([n])}$  is:

$$X_{([n])}( (X_1, X_2, \dots, X_n) ) := (X_{(1)}, X_{(2)}, \dots, X_{(n)}), \text{ such that, } X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}. \quad (3.80)$$

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 110 (Order statistics and sorting)** The order statistics for the five samples stored in `XsFromUni01Twstr101` from Labwork 108 can be computed using `sort` as follows:

```
>> disp(XsFromUni01Twstr101); % display the sample points
    0.5164    0.5707    0.0285    0.1715    0.6853
>> SortedXsFromUni01Twstr101=sort(XsFromUni01Twstr101); % sort data
>> disp(SortedXsFromUni01Twstr101); % display the order statistics
    0.0285    0.1715    0.5164    0.5707    0.6853
```

Therefore, we can use `sort` to obtain our order statistics  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  from  $n$  sample points  $x_1, x_2, \dots, x_n$ .

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

**Definition 55 (Inverse DF or Inverse CDF or Quantile Function)** Let  $X$  be an RV with DF  $F$ . The **inverse DF** or **inverse CDF** or **quantile function** is:

$$F^{[-1]}(q) := \inf \{x : F(x) > q\}, \quad \text{for some } q \in [0, 1]. \quad (3.81)$$

If  $F$  is strictly increasing and continuous then  $F^{[-1]}(q)$  is the unique  $x \in \mathbb{R}$  such that  $F(x) = q$ .

A **functional** is merely a function of another function. Thus,  $T(F) : \{\text{All DFs}\} \rightarrow \mathbb{T}$ , 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  $F$ :

$$T(F) = E(X) = \int x dF(x) .$$

2. The **variance** of RV  $X \sim F$  is a function of the DF  $F$ :

$$T(F) = E(X - E(X))^2 = \int (x - E(X))^2 dF(x) .$$

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  $q^{\text{th}}$  **quantile** of RV  $X \sim F$ :

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

2. The **first quartile** or the  $0.25^{\text{th}}$  **quantile** of the RV  $X \sim F$ :

$$T(F) = F^{[-1]}(0.25) .$$

3. The **median** or the **second quartile** or the  $0.50^{\text{th}}$  **quantile** of the RV  $X \sim F$ :

$$T(F) = F^{[-1]}(0.50) .$$

4. The **third quartile** or the  $0.75^{\text{th}}$  **quantile** of the RV  $X \sim F$ :

$$T(F) = F^{[-1]}(0.75) .$$

**Definition 56 (Empirical Distribution Function (EDF or ECDF))** Suppose we have  $n$  IID RVs,  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} F$ , 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  $x_i$ :

$$\hat{F}_n(x) = \frac{\sum_{i=1}^n \mathbf{1}(X_i \leq x)}{n} , \quad \text{where} \quad \mathbf{1}(X_i \leq x) := \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > x \end{cases} \quad (3.82)$$

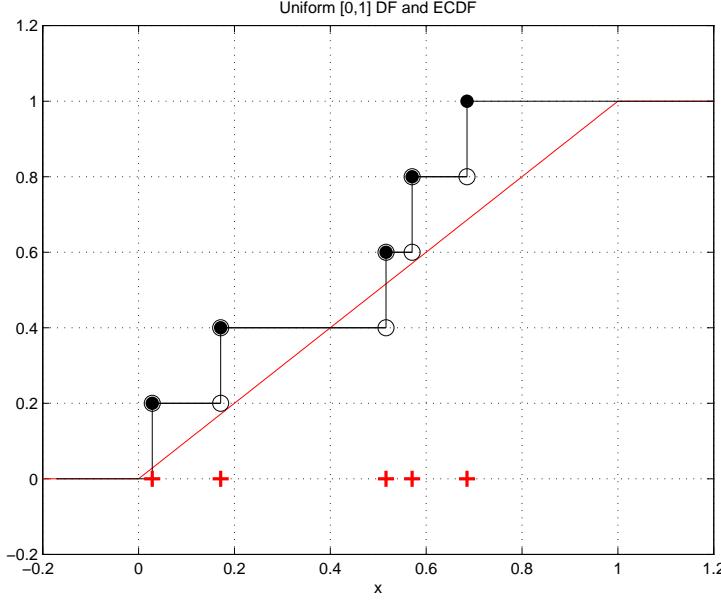
**Labwork 111 (Plot of empirical CDF)** Let us plot the ECDF for the five samples drawn from the  $Uniform(0, 1)$  RV in Labwork 108 using the MATLAB function ECDF. Let us super-impose the samples and the true DF as depicted in Figure 3.26 with the following script:

---

```
plotunifecdf.m
xs = -1:0.01:2; % vector xs from -1 to 2 with increment .05 for x values
% get the [0,1] uniform DF or cdf of xs in vector cdf
cdf=zeros(size(xs));% initialise cdf as zero
indices = find(xs>=1); cdf(indices) = 1; % set cdf as 1 when xs >= 1
indices = find(xs>=0 & xs<=1); cdf(indices)=xs(indices); % cdf=x 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.
```

---

Figure 3.26: 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.



**Definition 57 ( $q^{\text{th}}$  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 (3.82). The  $q^{\text{th}}$  sample quantile is defined as the statistic (statistical functional):

$$T(\hat{F}_n) = \hat{F}_n^{[-1]}(q) := \inf \{x : \hat{F}_n^{[-1]}(x) \geq q\}. \quad (3.83)$$

By replacing  $q$  in this definition of the  $q^{\text{th}}$  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  $q^{\text{th}}$  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  $q^{\text{th}}$  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:

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

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.

**Labwork 112 (The  $q^{\text{th}}$  sample quantile)** Use the implementation of Algorithm 1 as the MATLAB function `qthSampleQuantile` to find the  $q^{\text{th}}$  sample quantile of two simulated data arrays:

1. `SortedXsFromUni01Twstr101`, the order statistics that was constructed in Labwork 110 and
2. Another sorted array of 7 samples called `SortedXs`

**Algorithm 1**  $q^{\text{th}}$  Sample Quantile of Order Statistics1: *input:*

1.  $q$  in the  $q^{\text{th}}$  sample quantile, i.e. the argument  $q$  of  $\widehat{F}_n^{[-1]}(q)$ ,
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:*  $\widehat{F}_n^{[-1]}(q)$ , the  $q^{\text{th}}$  sample quantile3:  $i \leftarrow \lfloor (n - 1)q \rfloor$ 4:  $\delta \leftarrow (n - 1)q - i$ 5: **if**  $i = n - 1$  **then**6:    $\widehat{F}_n^{[-1]}(q) \leftarrow x_{(i+1)}$ 7: **else**8:    $\widehat{F}_n^{[-1]}(q) \leftarrow (1 - \delta)x_{(i+1)} + \delta x_{(i+2)}$ 9: **end if**10: *return:*  $\widehat{F}_n^{[-1]}(q)$ 

```

>> disp(SortedXsFromUni01Twstr101)
    0.0285    0.1715    0.5164    0.5707    0.6853
>> rand('twister',420);
>> SortedXs=sort(rand(1,7));
>> disp(SortedXs)
    0.1089    0.2670    0.3156    0.3525    0.4530    0.6297    0.8682
>> for q=[0, 0.25, 0.5, 0.75, 1.0]
    disp([q, qthSampleQuantile(q,SortedXsFromUni01Twstr101) ...
        qthSampleQuantile(q,SortedXs)])
end
    0    0.0285    0.1089
    0.2500    0.1715    0.2913
    0.5000    0.5164    0.3525
    0.7500    0.5707    0.5414
    1.0000    0.6853    0.8682

```

### 3.14.2 Univariate Data

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

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

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**:

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

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  $w = \bar{b}_i - b_i$  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:

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

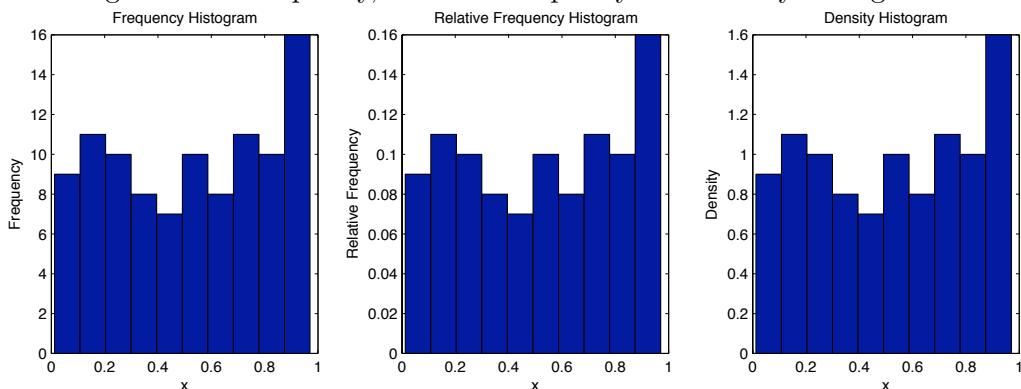
a **relative frequency histogram** is the histogram:

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

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

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

Figure 3.27: Frequency, Relative Frequency and Density Histograms



**Labwork 113 (Histograms with specified number of bins for univariate data)** Let us use samples from the `rand('twister',5489)` as our data set  $x$  and plot various histograms. Let us use `hist` function (read `help hist`) to make a default histogram with ten bins. Then we can make three types of histograms as shown in Figure 3.27 as follows:

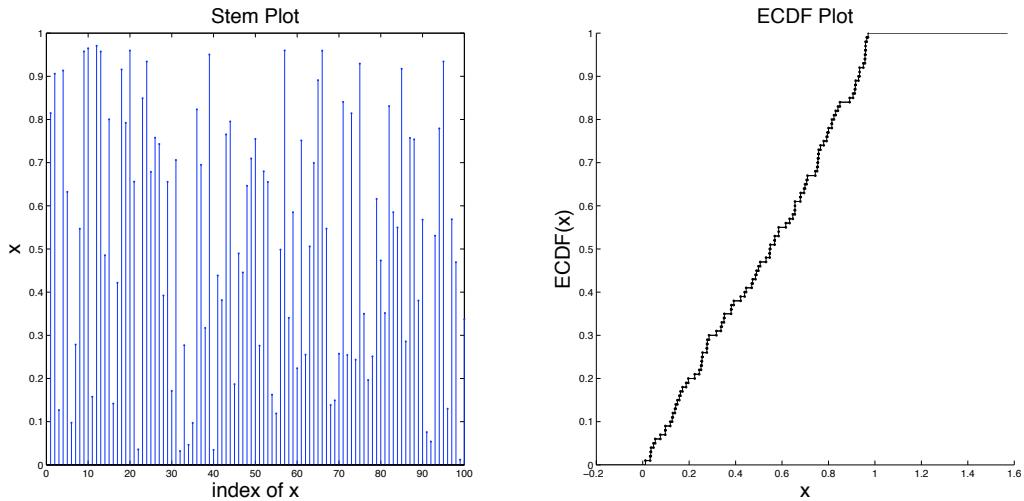
```
>> rand('twister',5489);
>> x=rand(1,100); % generate 100 PRNs
>> hist(x) % see what default hist does in Figure Window
>> % Now let us look deeper into the last hist call
>> [Fs, Cs] = hist(x) % Cs is the bin centers and Fs is the frequencies of data set x
Fs =
    9     11     10      8      7     10      8     11     10     16
Cs =
    0.0598    0.1557    0.2516    0.3474    0.4433    0.5392    0.6351    0.7309    0.8268    0.9227
>> % produce a histogram plot the last argument 1 is the width value for immediately adjacent bars -- help bar
>> bar(Cs,Fs,1) % create a frequency histogram
>> bar(Cs,Fs/100,1) % create a relative frequency histogram
>> bar(Cs,Fs/(0.1*100),1) % create a density histogram (area of bars sum to 1)
>> sum(Fs/(0.1*100) .* ones(1,10)*0.1) % checking if area does sum to 1
>> ans = 1
```

Try making a density histogram with 1000 samples from `rand` with 15 bins. You can specify the number of bins by adding an extra argument to `hist`, for e.g. `[Fs, Cs] = hist(x,15)` will produce 15 bins of equal width over the data range  $\mathcal{R}(x)$ .

**Labwork 114 (Stem plots and ECDF plots for univariate data)** We can also visualise the 100 data points in the array  $x$  using stem plot and ECDF plot as shown in Figure 3.28 as follows:

```
>> rand('twister',5489);
>> x=rand(1,100); % produce 100 samples with rand
>> stem(x,'.') % make a stem plot of the 100 data points in x (the option '.' gives solid circles for x)
>>% ECDF (type help ECDF) plot is extended to left and right by .2 and .6, respectively
>>% (second parameter 6 makes the dots in the plot smaller).
>> ECDF(x,.2,.6);
```

Figure 3.28: Frequency, Relative Frequency and Density Histograms



We can also visually summarise univariate data using the **box plot** or **box-whisker plot** available in the Stats Toolbox of MATLAB. These family of plots display a set of sample quantiles, typically they are include, the median, the first and third quartiles and the minimum and maximum values of our data array  $x$ .

### 3.14.3 Bivariate Data

By bivariate data array  $x$  we mean a  $2 \times n$  matrix of real numbers or equivalently  $n$  ordered pairs of points  $(x_{1,i}, x_{2,i})$  as  $i = 1, 2, \dots, n$ . 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 115 (Visualising bivariate data)** Let us generate a  $2 \times 5$  array representing samples of 5 ordered pairs sampled uniformly at random over the unit square  $[0, 1] \times [0, 1]$ . We can make 2D scatter plot as shown in Figure 3.29 as follows:

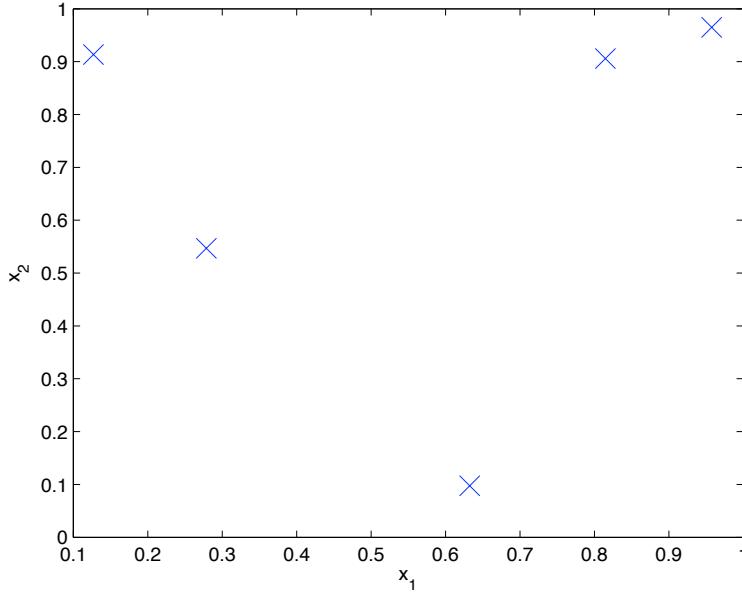
```
>> rand('twister',5489);
>> x=rand(2,5)% create a sequence of 5 ordered pairs uniformly from unit square [0,1]X[0,1]
x =
    0.8147    0.1270    0.6324    0.2785    0.9575
```

```

0.9058    0.9134    0.0975    0.5469    0.9649
>> plot(x(1,:),x(2,:),'x') % a 2D scatter plot with marker cross or 'x'
>> plot(x(1,:),x(2,:),'x', 'MarkerSize',15) % a 2D scatter plot with marker cross or 'x' and larger Marker size
>> xlabel('x_1'); ylabel('x_2'); % label the axes

```

Figure 3.29: 2D Scatter Plot



There are several other techniques for visualising bivariate data, including, 2D histograms, surface plots, heat plots, and we will encounter some of them in the sequel.

### 3.14.4 Trivariate Data

Trivariate data is more difficult to visualise on paper but playing around with the rotate 3D feature in MATLAB's Figure window can help bring a lot more perspective.

**Labwork 116 (Visualising trivariate data)** We can make **3D scatter plots** as shown in Figure 3.30 as follows:

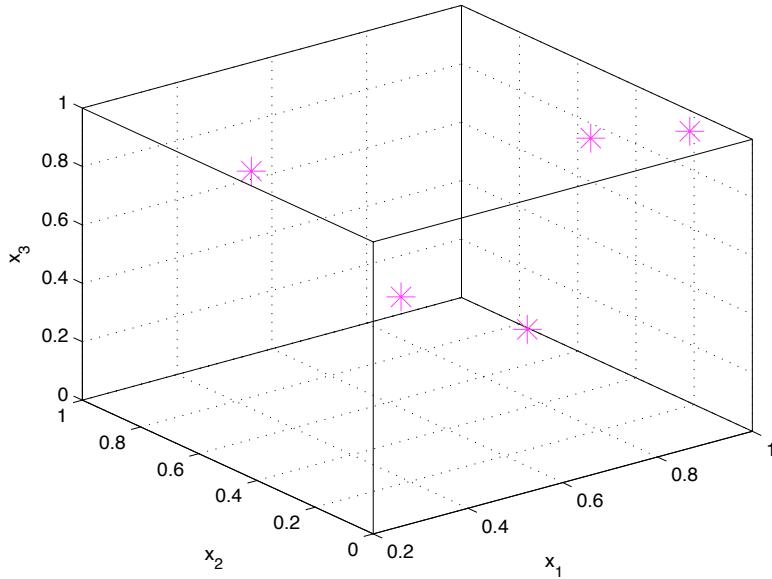
```

>> rand('twister',5489);
>> x=rand(3,5)% create a sequence of 5 ordered triples uniformly from unit cube [0,1]X[0,1]X[0,1]
x =
    0.8147    0.9134    0.2785    0.9649    0.9572
    0.9058    0.6324    0.5469    0.1576    0.4854
    0.1270    0.0975    0.9575    0.9706    0.8003
>> plot3(x(1,:),x(2,:),x(3,:),'x') % a simple 3D scatter plot with marker 'x'
>>% a more interesting one with options that control marker type, line-style,
>>% colour in [Red Green Blue] values and marker size - read help plot3 for more options
>> plot3(x(1,:),x(2,:),x(3,:),'Marker','*','LineStyle','none','Color',[1 0 1],'MarkerSize',15)
>> plot3(x(1,:),x(2,:),x(3,:),'m*','MarkerSize',15) % makes same figure as before but shorter to write
>> box on % turn on the box and see the effect on the Figure
>> grid on % turn on the grid and see the effect on the Figure
>> xlabel('x_1'); ylabel('x_2'); zlabel('x_3'); % assign labels to x,y and z axes

```

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

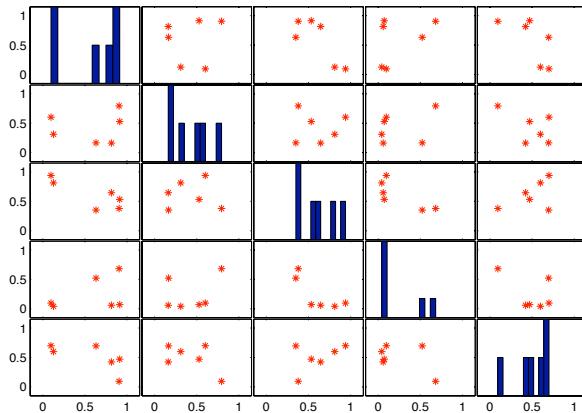
Figure 3.30: 3D Scatter Plot



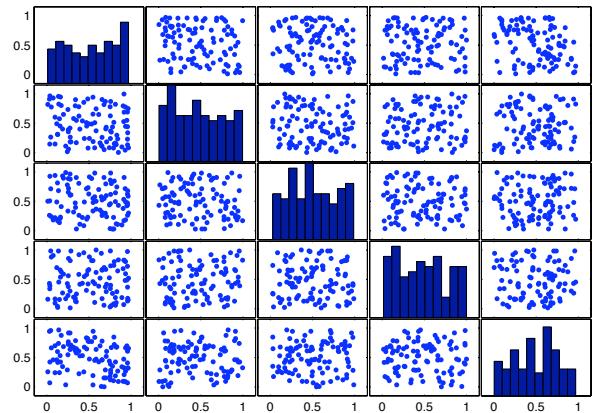
There are several other techniques for visualising trivariate data, including, iso-surface plots, moving surface or heat plots, and you will encounter some of them in the future.

### 3.14.5 Multivariate Data

For high-dimensional data in  $d$ -dimensional space  $\mathbb{R}^d$  with  $d \geq 3$  you have to look at several lower dimensional projections of the data. We can simultaneously look at 2D scatter plots for every pair of co-ordinates  $\{(i, j) \in \{1, 2, \dots, d\}^2 : i \neq j\}$  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 3.31: Plot Matrix of uniformly generated data in  $[0, 1]^5$ 

(a) First six samples



(b) All thousand samples

**Labwork 117** Let us make matrix plots from a uniformly generated sequence of 100 points in 5D unit cube  $[0, 1]^5$  as shown in Figure 3.31.

```
>> rand('twister',5489);
>> % generate a sequence of 1000 points uniformly distributed in 5D unit cube [0,1]X[0,1]X[0,1]X[0,1]X[0,1]
>> x=rand(1000,5);
>> x(1:6,:) % first six points in our 5D unit cube, i.e., the first six rows of x
ans =
    0.8147    0.6312    0.7449    0.3796    0.4271
    0.9058    0.3551    0.8923    0.3191    0.9554
    0.1270    0.9970    0.2426    0.9861    0.7242
    0.9134    0.2242    0.1296    0.7182    0.5809
    0.6324    0.6525    0.2251    0.4132    0.5403
    0.0975    0.6050    0.3500    0.0986    0.7054
>> plotmatrix(x(1:5,:),'r*') % make a plot matrix
>> plotmatrix(x) % make a plot matrix of all 1000 points
```

### 3.14.6 Loading and Exploring Real-world Data

All of the data we have played with so far were computer-generated. It is time to get our hands dirty with real-world data. The first step is to obtain the data. Often, publicly-funded institutions allow the public to access their databases. Such data can be fetched from appropriate URLs in one of the two following ways:

Method A: Manually download by filling the appropriate fields in an online request form.

Method B: Automagically download directly from your MATLAB session.

Then we want to inspect it for inconsistencies, missing values and replace them with `NaN` values in MATLAB that stand for not-any-number. Finally, we can visually explore, transform and interact with the data to discover interesting patterns that are hidden in the data. This process is called *exploratory data analysis* and is the foundational first step towards subsequent computational statistical experiments [*John W. Tukey, Exploratory Data Analysis, Addison-Wesley, New York, 1977*].

### 3.14.7 Geological Data

Let us focus on the data of earth quakes that heavily damaged Christchurch on February 22 2011. This data can be fetched from the URL <http://magma.geonet.org.nz/resources/quakesearch/> by Method A and loaded into MATLAB for exploratory data analysis as done in Labwork 118.

**Labwork 118** Let us go through the process one step at a time using Method A.

1. Download the data as a CSV or *comma separated variable* file in plain ASCII text (this has been done for this data already for you and saved as `NZ20110222earthquakes.csv` in the `CSEMatlabScripts` directory).
2. Open the file in a simple text editor such as `Note Pad` in Windows or one of the following editors in OS X, Unix, Solaris, Linux/GNU variants such as Ubuntu, SUSE, etc: `vi`, `vim`, `emacs`, `geany`, etc. The first three and last two lines of this file look as follows:

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

```
>> EQall(any(isnan(EQall),2),:) = []; %Remove any rows containing NaNs from the matrix EQall
>> % report the size of EQall and see if it is different from before we removed and NaN containing rows
>> size(EQall)
ans =    145      14
>> % remove locations outside Chch and assign it to a new variable called EQ
>> EQ = EQall(-43.75<EQall(:,2) & EQall(:,2)<-43.45 ...
& 172.45<EQall(:,3) & EQall(:,3)<172.9 & EQall(:,12)>3, :);
>> % now report the size of the earthquakes in Christchurch in variable EQ
>> size(EQ)
ans =    124      14
>> % assign the four variables of interest
```

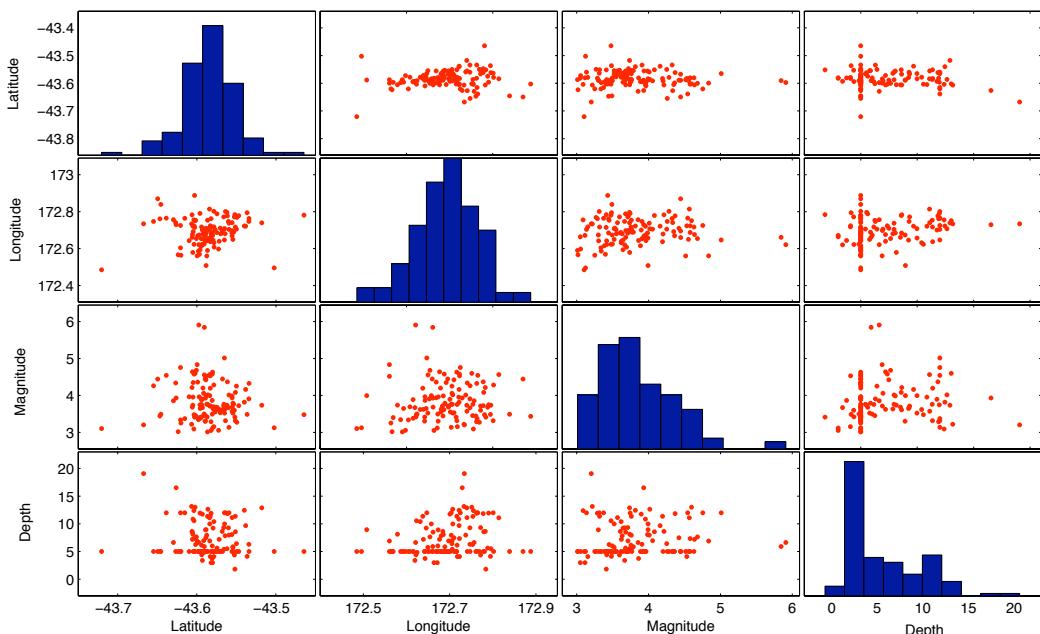
```
>> LatData=EQ(:,2); LonData=EQ(:,3); MagData=EQ(:,12); DepData=EQ(:,13);
>> % finally make a plot matrix of these 124 4-tuples as red points
>> plotmatrix([LatData,LonData,MagData,DepData], 'r.');
```

All of these commands have been put in a script M-file `NZEQChCch20110222.m` and you can simply call it from the command window to automatically load the data and assign it to the variables `EQAll`, `EQ`, `LatData`, `LonData`, `MagData` and `DepData`, instead of retyping each command above every time you need these matrices in MATLAB, as follows:

```
>> NZEQChCch20110222
ans =    145    14
ans =    145    14
ans =    124    14
```

In fact, we will do exactly this to conduct more exploratory data analysis with these earth quake measurements in Labwork 119.

Figure 3.32: Matrix of Scatter Plots of the latitude, longitude, magnitude and depth of the 22-02-2011 earth quakes in Christchurch, New Zealand.



**Labwork 119** Try to understand how to manipulate time stamps of events in MATLAB and the Figures being output by following the comments in the script file `NZEQChCch20110222EDA.m`.

```
>> NZEQChCch20110222
ans =    145    14
ans =    145    14
ans =    124    14
ans =    145    14
ans =    145    14
ans =    124    14
ans = 22-Feb-2011 00:00:31
ans = 22-Feb-2011 23:50:01
```

---

```
NZEQChCch20110222EDA.m
%% Load the data from the comma delimited text file 'NZ20110222earthquakes.csv'
% using the script M-file NZEQChCch20110222.m
NZEQChCch20110222
%% working with time stamps is tricky
%% time is encoded by columns 6 through 11
%% as origin of earthquake in year, month, day, hour, minute, sec:
%% ORI_YEAR,ORI_MONTH,ORI_DAY,ORI_HOUR,ORI_MINUTE,ORI_SECOND
%% datenum is Matlab's date encoding function see help datenum
TimeData=datenum(EQ(:,6:11)); % assign origin times of earth quakes in datenum coordinates
MaxD=max(TimeData); % get the latest time of observation in the data
MinD=min(TimeData); % get the earliest time of observation in the data
datestr(MinD) % a nice way to conver to calendar time!
datestr(MaxD) % ditto

% recall that there four variables were assigned in NZEQChCch20110222.m
% LatData=EQ(:,2); LonData=EQ(:,3); MagData=EQ(:,12); DepData=EQ(:,13);

%clear any existing Figure windows
clf
plot(TimeData,MagData,'o-') % plot origin time against magnitude of each earth quake

figure % tell matlab you are about to make another figure
plotmatrix([LatData,LonData,MagData,DepData],'r.');

figure % tell matlab you are about to make another figure
scatter(LonData,LatData,'.') % plot the LONGitude Vs. LATtitude

figure % tell matlab you are about to make another figure
% relative frequency histogram of magnitudes from 0 to 12 on Richter Scale with 15 bins
hist(MagData,15)

%max(MagData)

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

%%%%%
% more advanced topic - uncomment and read help if bored
%tri = delaunay(LatData,LonData);
%triplot(tri,LatData,LonData,DepData);
```

---

### Geostatistical exploratory data analysis with Google Earth

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

Date Range: 2011 2 22 to 2011 2 22

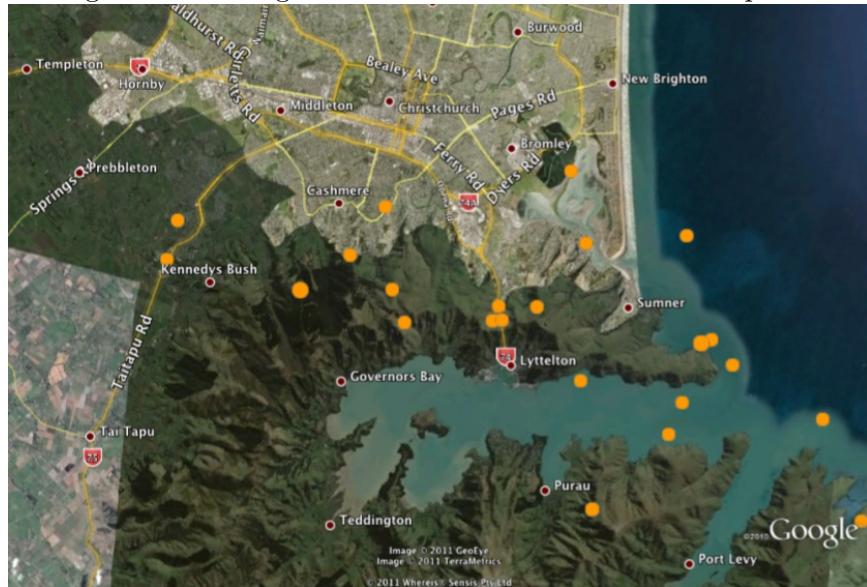
Catalog: USGS/NEIC (PDE-Q)

produced 43 earth quakes world-wide, including those in Christchurch as shown in Figure 3.33. One can do a lot more than a mere visualisation with the USGS/NEIC database of earth-quakes world-wide, the freely available Google earth software bundle <http://www.google.com/earth/index.html> and the freely available MATLAB package googleearth from [http://www.mathworks.com/matlabcentral/fx\\_files/12954/4/content/googleearth/html/html\\_product\\_page.html](http://www.mathworks.com/matlabcentral/fx_files/12954/4/content/googleearth/html/html_product_page.html).

#### 3.14.8 Metereological Data

New Zealand's meteorological service NIWA provides weather data under its TERMS AND CONDITIONS FOR ACCESS TO DATA (See [http://cliflo.niwa.co.nz/doc/terms\\_print.html](http://cliflo.niwa.co.nz/doc/terms_print.html)).

Figure 3.33: Google Earth Visualisation of the earth quakes



We will explore some data of rainfall and temperatures from NIWA.

### Daily Rainfalls in Christchurch

Automagic downloading of the data by Method B can be done if the data provider allows automated queries. It can be accomplished by `urlread` for instance.

Paul Brouwers has a basic CliFlo datafeed on <http://www.math.canterbury.ac.nz/php/lib/cliflo/rainfall.php>. This returns the date and rainfall in milli meters as measured from the CHCH aeroclub station. It is assumed that days without readings would not be listed. The data doesn't go back much before 1944.

**Labwork 120** Understand how Figure 3.34 is obtained by the script file `RainFallsInChch.m` by typing and following the comments:

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

---

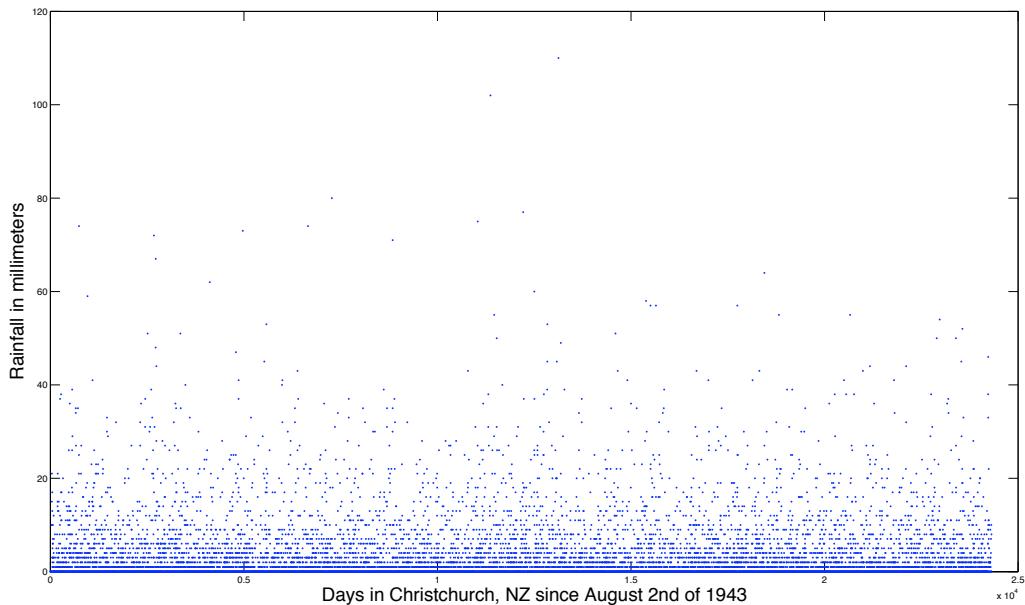
```
_____  
RainFallsInChch.m  
_____
%% How to download data from an URL directly without having to manually
%% fill out forms
% first make a string of the data using urlread (read help urlread if you want details)
StringData = urlread('http://www.math.canterbury.ac.nz/php/lib/cliflo/rainfall.php');
RainFallsChch = textscan(StringData, '%d %f', 'delimiter', ',')
RC = [RainFallsChch{1} RainFallsChch{2}]; % assign Matlab cells as a matrix
size(RC) % find the size of the matrix

FirstDayOfData = min(RC(:,1))
LastDayOfData = max(RC(:,1))

plot(RC(:,2),'.')
xlabel('Days in Christchurch, NZ since August 2nd of 1943','FontSize',20);
ylabel('Rainfall in millimeters','FontSize',20)
```

---

Figure 3.34: Daily rainfalls in Christchurch since March 27 2010



### Daily Temperatures in Christchurch

**Labwork 121** Understand how Figure 3.35 is being generated by following the comments in the script file ChchTempsLoad.m by typing:

```
>> ChchTempsLoad
```

---

```
ChchTempsLoad.m
%% Load the data from the comma delimited text file 'NIWACliFloChchAeroClubStationTemps.txt'
%% with the following column IDs
%% Max_min: Daily Temperature in Christchurch New Zealand
%% Stationate(NZST),Tmax(C),Period(Hrs),Tmin(C),Period(Hrs),Tgmin(C),Period(Hrs),Tmean(C),RHmean(%),Period(Hrs)

% the matrix T is about to be assigned the data as a matrix; the option [27,1,20904,5] to
% specify the upper-left and lower-right corners of an imaginary rectangle
% over the text file 'NIWACliFloChchAeroClubStationTemps.txt'.
% here we start from line number 27 and end at the last line number 20904
% and we read only columns NZST,Tmax(C),Period(Hrs),Tmin(C),Period(Hrs)

T = dlmread('NIWACliFloChchAeroClubStationTemps.txt','','',[27,1,20904,5]);
% just keep column 1,2 and 4 named NZST,Tmax(C),Period(Hrs),Tmin(C),
% i.e. date in YYYYMMDD foramt, maximum temperature, minimum temperature
T = T(:,[1,2,4]); % just pull the time
% print size before removing missig data rows are removed
size(T) % report the dimensions or size of the matrix T

% This file has a lot of missing data points and they were replaced with
% NaN values - see the file for various manipulations that were done to the
% raw text file from NIWA (Copyright NIWA 2011 Subject to NIWA's Terms and
% Conditions. See: http://cliflo.niwa.co.nz/pls/niwp/doc/terms.html)
T(any(isnan(T),2),:) = [];% Remove any rows containing NaNs from a matrix

size(T) % if the matrix has a different size now then the data-less days now!
```

```

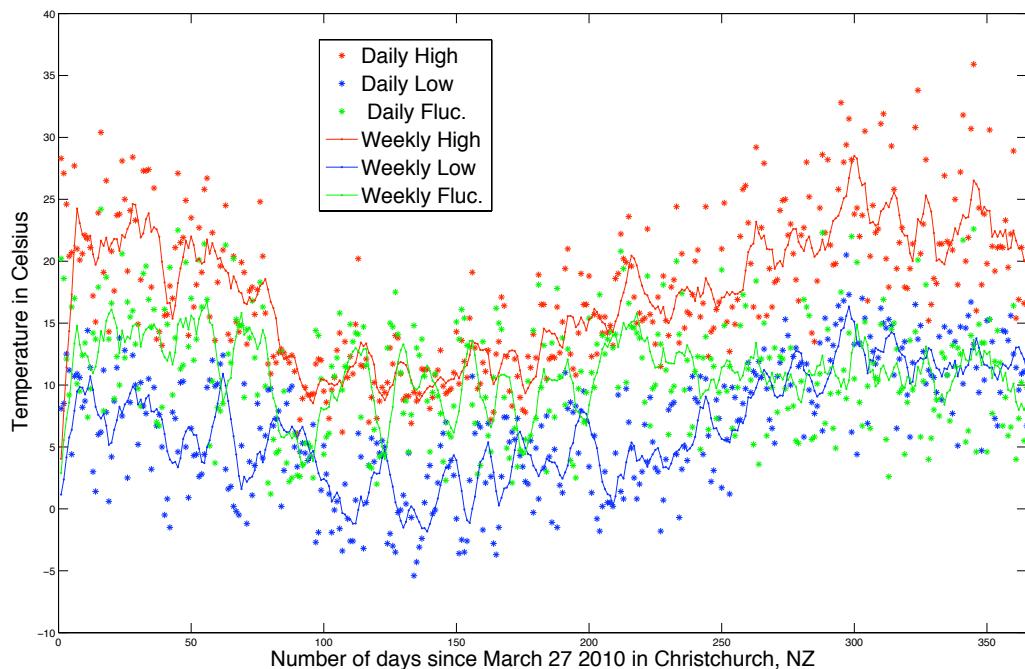
clf % clears all current figures

% Daily max and min temperature in the 100 days with good data
% before last date in this data, i.e., March 27 2011 in Christchurch NZ
H365Days = T(end-365:end,2);
L365Days = T(end-365:end,3);
F365Days = H365Days-L365Days; % assign the maximal fluctuation, i.e. max-min
plot(H365Days,'r*') % plot daily high or maximum temperature = Tmax
hold on; % hold the Figure so that we can overlay more plots on it
plot(L365Days,'b*') % plot daily low or minimum temperature = Tmin
plot(F365Days, 'g*') % plot daily Fluctuation = Tmax - Tmin
% filter for running means
windowSize = 7;
WeeklyHighs = filter(ones(1,windowSize)/windowSize,1,H365Days);
plot(WeeklyHighs,'r.-')
WeeklyLows = filter(ones(1,windowSize)/windowSize,1,L365Days);
plot(WeeklyLows,'b.-')
WeeklyFlucs = filter(ones(1,windowSize)/windowSize,1,F365Days);
plot(WeeklyFlucs,'g.-')
xlabel('Number of days since March 27 2010 in Christchurch, NZ','FontSize',20);
ylabel('Temperature in Celsius','FontSize',20)
MyLeg = legend('Daily High','Daily Low','Daily Fluc.', 'Weekly High','Weekly Low',...
    'Weekly Fluc.', 'Location','NorthEast')
% Create legend
% legend1 = legend(axes1,'show');
set(MyLeg,'FontSize',20);
xlim([0 365]); % set the limits or boundary on the x-axis of the plots
hold off % turn off holding so we stop overlaying new plots on this Figure

```

---

Figure 3.35: Daily temperatures in Christchurch for one year since March 27 2010



### 3.14.9 Textual Data

Processing and analysing textual data to make a decision is another important computational statistical experiment. An obvious example is machine translation and a less obvious one is exploratory data analysis of the textual content of

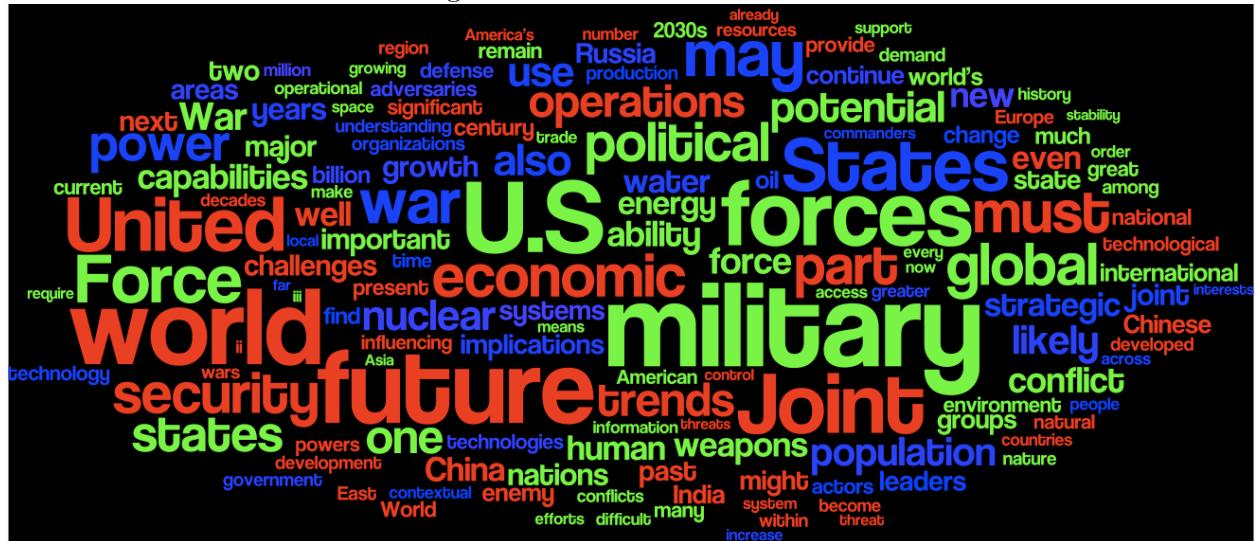
- a large document
  - twitter messages within an online social network of interest
  - etc.

An interesting document with a current affairs projection is the Joint Operating Environment 2010 Report by the US Department of Defense. This document was downloaded from [http://www.jfcom.mil/newslink/storyarchive/2010/JOE\\_2010\\_o.pdf](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 3.36: Wordle of JOE 2010



We can try to produce a statistic of this document by recording the frequency of words in its textual content. Then we can produce a “word histogram” or “word cloud” to explore the document visually at one of the coarsest possible resolutions of the textual content in the JOE 2010 Report. The “word cloud” shown in Figure 3.36 was produced by Phillip Wilson using *wordle* from <http://www.wordle.net/>. A description from the wordle URL says:

Wordle is a toy for generating “word clouds” from text that you provide. The clouds give greater prominence to words that appear more frequently in the source text. You can tweak your clouds with different fonts, layouts, and color schemes. The images you create with Wordle are yours to use however you like. You can print them out, or save them to the Wordle gallery to share with your friends.

**Labwork 122 (favourite word cloud)** This is just for fun. Produce a “word cloud” of your honours thesis or summer project or any other document that fancies your interest by using *wordle* from <http://www.wordle.net/>. Play with the aesthetic features to change colour, shapes, etc.

### 3.14.10 Machine Sensor Data

Instrumentation of modern machines, such as planes, rockets and cars allow the sensors in the machines to collect live data and dynamically take *decisions* and subsequent *actions* by executing algorithms to drive their devices in response to the data that is streaming into their sensors. For example, a rocket may have to adjust its boosters to compensate for the prevailing directional changes in wind in order to keep going up and launch a satellite. These types of decisions and actions, theorised by *controlled Markov processes*, typically arise in various fields of engineering such as, aerospace, civil, electrical, mechanical, robotics, etc.

In an observational setting, without an associated control problem, one can use machine sensor data to get information about some state of the system or phenomenon, i.e., what is it doing? or where is it?, etc. Sometimes sensors are attached to a sample of individuals from a wild population, say Emperor Penguins in Antarctica where the phenomenon of interest may be the diving habits of this species after the eggs hatch. As an other example we can attach sensors to a double pendulum and find what it is doing when we give it a spin.

Based on such observational data the experimenter typically tries to learn about the behaviour of the system from the sensor data to estimate parameters, test hypotheses, etc. Such types of experiments are typically performed by scientists in various fields of science, such as, astronomy, biology, chemistry, geology, physics, etc.

#### Chaotic Time Series of a Double Pendulum

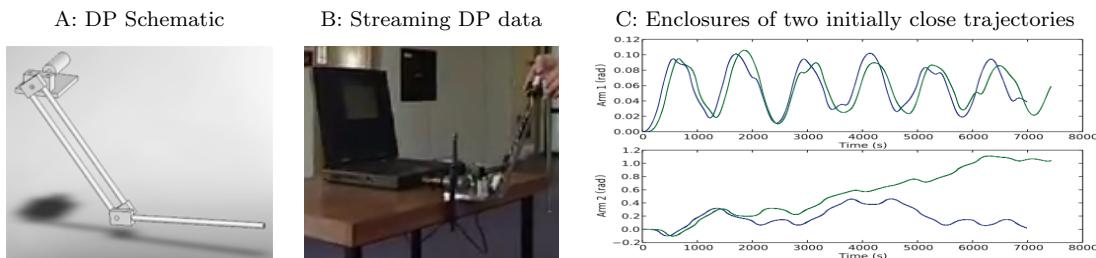


Figure 3.37: Double Pendulum

Sensors called *optical encoders* have been attached to the top end of each arm of a chaotic double pendulum in order to obtain the angular position of each arm through time as shown in Figure 3.37. Time series of the angular position of each arm for two trajectories that were initialized very similarly, say the angles of each arm of the double pendulum are almost the same at the initial time of release. Note how quickly the two trajectories diverge! System with such a sensitivity to initial conditions are said to be *chaotic*.

**Labwork 123 (A Challenging Task)** Try this if you are interested. Read any of the needed details about the design and fabrication of the double pendulum at <http://www.math.canterbury.ac.nz/~r.sainudiin/lmse/double-pendulum/>. Then use MATLAB to generate a plot similar to Figure 3.37(C) using time series data of trajectory 1 and trajectory 2 linked from the bottom of the above URL.

### 3.15 Exercises in Statistics

**Ex. 3.56** — What is the sample mean and sample variance of the following dataset:

$$1, 3, 2, 1, 2, 3, 3$$

# Chapter 4

## Simulation

“Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin.” — John von Neumann (1951)

### 4.1 Physical Random Number Generators

Physical devices such as the BINGO machine demonstrated in class can be used to produce an integer uniformly at random from a finite set of possibilities. Such “ball bouncing machines” used in the British national lottery as well as the New Zealand LOTTO are complex nonlinear systems that are extremely sensitive to initial conditions (“chaotic” systems) and are physical approximations of the probability model called a “well-stirred urn” or an equi-probable de Moivre( $1/k, \dots, 1/k$ ) 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 2** Linear Congruential Generator (LCG)

---

1: *input:* five suitable integers:

1.  $m$ , the modulus;  $0 < m$
2.  $a$ , the multiplier;  $0 \leq a < m$
3.  $c$ , the increment;  $0 \leq c < m$
4.  $x_0$ , the seed;  $0 \leq x_0 < m$
5.  $n$ , the number of desired pseudo-random numbers

2: *output:*  $(x_0, x_1, \dots, x_{n-1})$ , 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 124 (Generic Linear Congruential Sequence)** Let us implement Algorithm 2 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  $(m, a, c, x_0, n)$ .

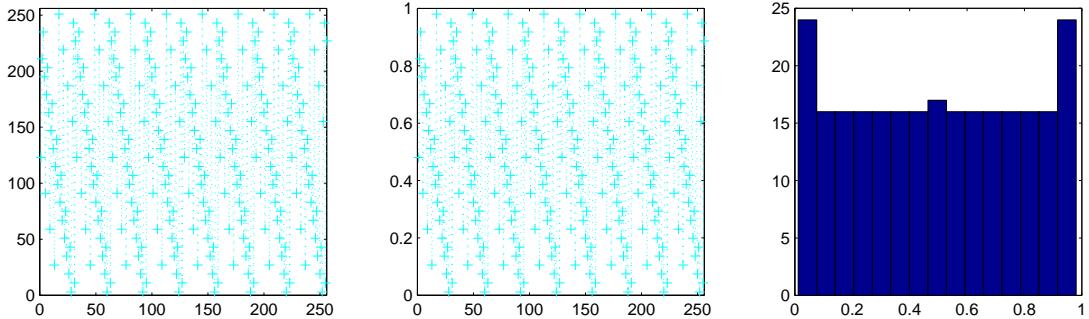
**Labwork 125 (LCG with period length of 32)** Consider the linear congruential sequence with  $(m, a, c, x_0, n) = (256, 137, 0, 123, 257)$  with period length of only  $32 < m = 256$ . We can visualise the sequence as plots in Figure 4.1 after calling the following M-file.

---

```
LinConGenKnuth334T1L5Plots.m
LCGSeq=LinConGen(256,137,0,123,257)
subplot(1,3,1)
plot(LCGSeq,'+')
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 124, the generator `LinConGen(13,12,11,10,12)` has period  $(10, 1)$  of length 2, the generator `LinConGen(13,10,9,8,12)` has period  $(8, 11, 2, 3, 0, 9)$  of length 6 and the generator `LinConGen(256,137,0,123,257)` has a period of length 32. All these generators have a non-maximal period length less than their modulus  $m$ . A good generator should have a maximal period of  $m$ . Let us try to implement a generator with a maximal period of  $m = 256$ .

The period of a general LCG is at most  $m$ , and for some choices of  $a$  the period can be much less than  $m$  as shown in the examples considered earlier. The LCG will have a full period if and only if:

1.  $c$  and  $m$  are relatively prime,
2.  $a - 1$  is divisible by all prime factors of  $m$ ,
3.  $a - 1$  is a multiple of 4 if  $m$  is a multiple of 4

**Labwork 126 (LCG with maximal period length of 256)** Consider the linear congruential sequence with  $(m, a, c, x_0, n) = (256, 137, 123, 13, 256)$ . First check that these parameters do indeed satisfy the three condition above and therefore can produce the maximal period length of only  $m = 256$ . Modify the input parameter to `LinConGen` and repeat Labwork 125 in order to first produce a sequence of length 257. Do you see that the period is of maximal length of 256 as opposed to the generator of Labwork 125? Next produce a Figure to visualise the sequence as done in Figure 4.1.

A useful sequence should clearly have a relatively long period, say at least  $2^{30}$ . Therefore, the **modulus  $m$  has to be rather large** because the **period** cannot have more than  $m$  elements. Moreover, the quality of pseudo-random numbers of a LCG is extremely sensitive to the choice of  $m$ ,  $a$  and  $c$  even if the maximal period is attained. The next example illustrates this point.

**Labwork 127 (The infamous RANDU)** RANDU is an infamous LCG, which has been used since the 1960s. It is widely considered to be one of the most ill-conceived random number generators designed. Notably, it fails the **spectral test** badly for dimensions greater than 2. The following commands help visualise the sequence of first 5001/3 triplets  $(x_i, x_{i+1}, x_{i+2})$  seeded from  $x_0 = 1$  (Figure 4.2). Read `help reshape` and `help plot3`.

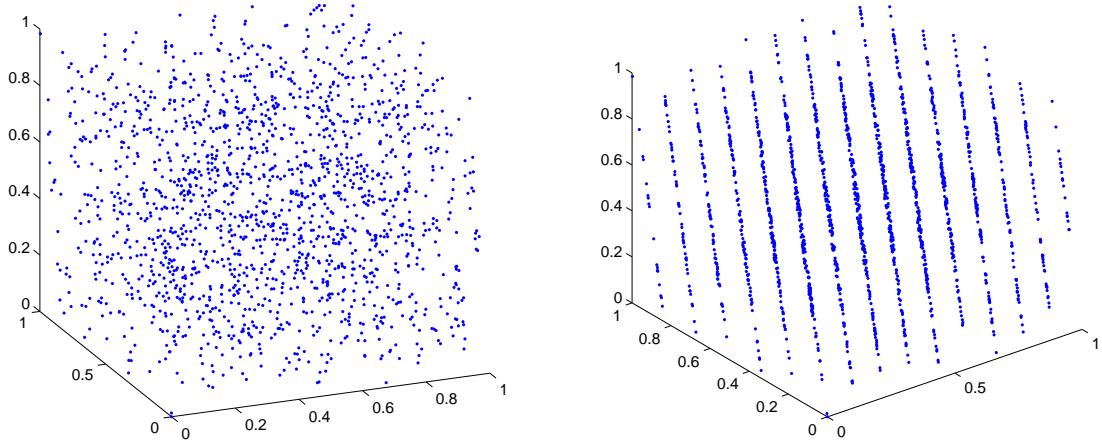
```
>> x=reshape( (LinConGen(2147483648,65539,0,1,5001) ./ 2147483648) ,3,[]);
>> plot3(x(1,:),x(2,:),x(3,:),'.'
```

**Labwork 128 (Fishman20 and Lecuyer21 LCGs)** The following two LCGs are recommended in Knuth's Art of Computer Programming, vol. 2, for generating pseudo-random numbers for simple simulation tasks.

```
>> LinConGen(2147483647,48271,0,08787458,10) ./ 2147483647
ans =    0.0041    0.5239    0.0755    0.7624    0.6496    0.0769    0.9030    0.4259    0.9948    0.8868

>> LinConGen(2147483399,40692,0,01234567,10) ./ 2147483399
ans =    0.0006    0.3934    0.4117    0.7893    0.3913    0.6942    0.6790    0.3337    0.2192    0.1883
```

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](http://en.wikipedia.org/wiki/Mersenne_twister) for details.

Let us learn to implement the MATLAB function that generates PRNs. In MATLAB the function `rand` produces a deterministic PRN sequence. First, read `help rand`. We can generate PRNs as follows.

**Labwork 129 (Calling PRNG in MATLAB)** In MATLAB `rand` is basic PRNG command.

```
>> rand(1,10) % generate a 1 X 10 array of PRNs
ans =
    0.8147    0.9058    0.1270    0.9134    0.6324    0.0975    0.2785    0.5469    0.9575    0.9649
>> rand(1,10) % generate another 1 X 10 array of PRNs
ans =
    0.1576    0.9706    0.9572    0.4854    0.8003    0.1419    0.4218    0.9157    0.7922    0.9595
>> rand('twister',5489) % reset the PRNG to default state Mersenne Twister with seed=5489
>> rand(1,10) % reproduce the first array
ans =
    0.8147    0.9058    0.1270    0.9134    0.6324    0.0975    0.2785    0.5469    0.9575    0.9649
>> rand(1,10) % reproduce the second array
ans =
    0.1576    0.9706    0.9572    0.4854    0.8003    0.1419    0.4218    0.9157    0.7922    0.9595
```

In general, you can use any seed value to initiate your PRNG. You may use the `clock` command to set the seed:

```
>> SeedFromClock=sum(100*clock); % save the seed from clock
>> rand('twister',SeedFromClock) % initialize the PRNG
>> rand(1,10)
ans =
    0.3696    0.3974    0.6428    0.6651    0.6961    0.7311    0.8982    0.6656    0.6991    0.8606
>> rand(2,10)
ans =
    0.3432    0.9511    0.3477    0.1007    0.8880    0.0853    0.6067    0.6976    0.4756    0.1523
    0.5827    0.5685    0.0125    0.1555    0.5551    0.8994    0.2502    0.5955    0.5960    0.5700
>> rand('twister',SeedFromClock) % initialize the PRNG to same SeedFromClock
>> rand(1,10)
ans =
    0.3696    0.3974    0.6428    0.6651    0.6961    0.7311    0.8982    0.6656    0.6991    0.8606
```

**Labwork 130 (3D plots of triplets generated by the “Mersenne Twister”)** Try to find any correlation between triplets generated by the “Mersenne Twister” by rotating the 3D plot generated by the following code:

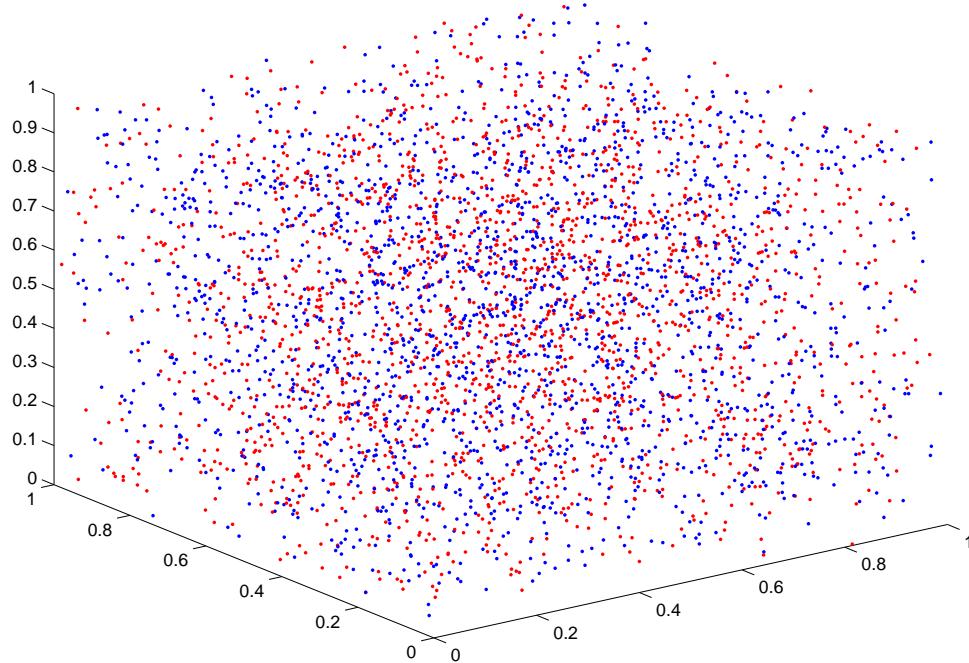
```
>> rand('twister',1234)
>> x=rand(3,2000); % store PRNs in a 3X2000 matrix named x
>> plot3(x(1,:),x(2,:),x(3,:),'.'
```

Compare this with the 3D plot of triplets from RANDU of Labwork 127. Which of these two PRNGs do you think is “more random” looking? and why?

Change the seed value to the recommended default by the authors and look at the point cloud (in red) relative to the previous point cloud (in blue). Rotate the plots to visualise from multiple angles. Are they still random looking?

```
>> rand('twister',1234)% same seed as before
>> x=rand(3,2000); % store PRNs in a 3X2000 matrix named x
>> rand('twister',5489)% the recommended default seed
>> y=rand(3,2000);% store PRNs seeded by 5489 in a 3X2000 matrix named y
>> plot3(x(1,:),x(2,:),x(3,:),'b.') % plot triplets as blue dots
>> hold on;
>> plot3(y(1,:),y(2,:),y(3,:),'r.') % plot triplets as red dots
```

Figure 4.3: Triplet point clouds from the “Mersenne Twister” with two different seeds (see Lab-work 130). .



### 4.3 Simulation of non-Uniform(0, 1) Random Variables

The Uniform(0, 1) RV of Model 7 forms the foundation for random variate generation and simulation. This is appropriately called the fundamental model or experiment, since every other experiment can be obtained from this one.

Next, we simulate or generate samples from other RVs by making the following two assumptions:

1. independent samples from the Uniform(0, 1) RV can be generated, and
2. real arithmetic can be performed exactly in a computer.

Both these assumptions are, in fact, not true and require a more careful treatment of the subject. We may return to these careful treatments later on.

#### 4.3.1 Inversion Sampler for Continuous Random Variables

**Proposition 58 (Inversion sampler)** Let  $F(x) := \int_{-\infty}^x f(y) dy : \mathbb{R} \rightarrow [0, 1]$  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:

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

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

---

**Algorithm 3** Inversion Sampler or Inverse (C)DF Sampler

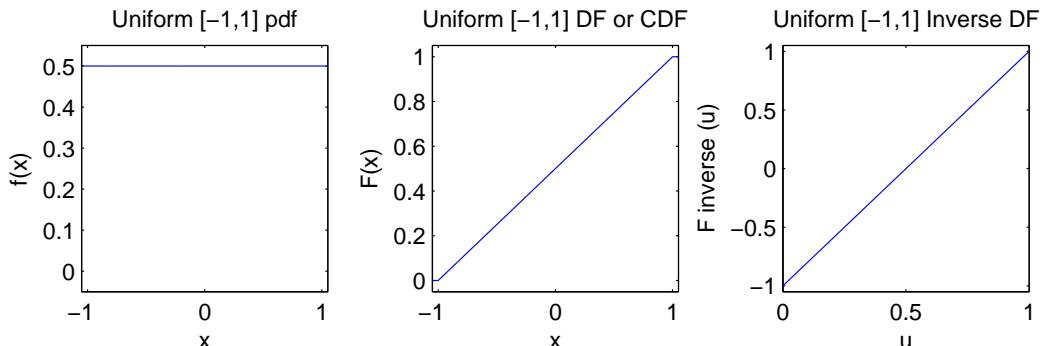
---

- 1: *input*: (1)  $F^{[-1]}(x)$ , inverse of the DF of the target RV  $X$ , (2) the fundamental sampler
  - 2: *initialise*: set the seed, if any, for the fundamental sampler
  - 3: *output*: a sample from  $X$  distributed according to  $F$
  - 4: *draw*  $u \sim \mathrm{Uniform}(0, 1)$
  - 5: *return*:  $x = F^{[-1]}(u)$
- 

This algorithm emphasises the fundamental sampler’s availability in an *input* step, and its set-up needs in an *initialise* step. In the following sections, we will not mention these universal steps; they will be taken for granted. The direct applicability of Algorithm 3 is limited to univariate densities for which the inverse of the cumulative distribution function is explicitly known. The next section will consider some examples.

Recall the  $\mathrm{Uniform}(\theta_1, \theta_2)$  RV of Model 9 with the following PDF, DF and inverse DF. Let us simulate from it using the inversion sampler.

Figure 4.4: A plot of the PDF, DF or CDF and inverse DF of the  $\mathrm{Uniform}(-1, 1)$  RV  $X$ .



**Simulation 131** ( $\mathrm{Uniform}(\theta_1, \theta_2)$ ) To simulate from  $\mathrm{Uniform}(\theta_1, \theta_2)$  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  $\mathrm{Uniform}(\theta_1, \theta_2)$  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:

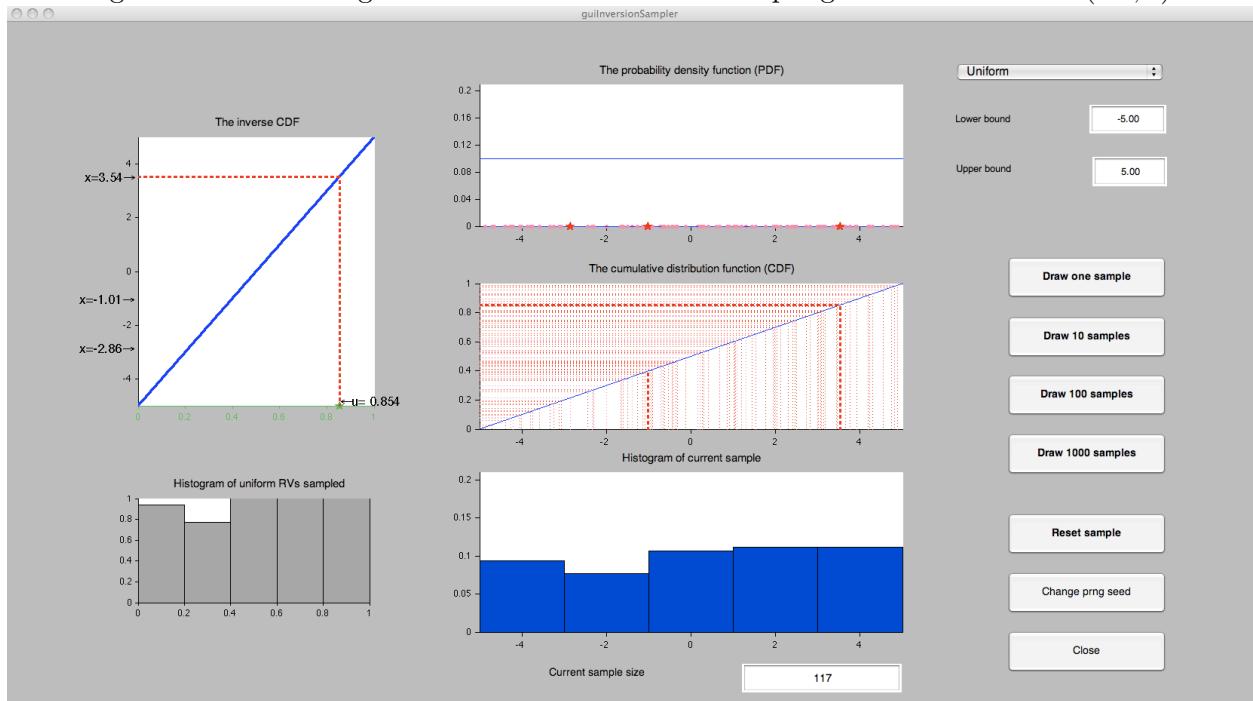
```
>> rand('twister',786543); % initialise the fundamental sampler
>> theta1=-83; theta2=1004; % declare values for parameters a and b
>> u=rand(1,5); % now u is an array of 5 samples from Uniform(0,1)
>> x=theta1+(theta2 - theta1)*u; % x is an array of 5 samples from Uniform(-83,1004]) RV
>> disp(x); % display the 5 samples just drawn from Uniform(-83,1004) RV
465.3065 111.4994 14.3535 724.8881 254.0168
```

**Labwork 132 (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
```

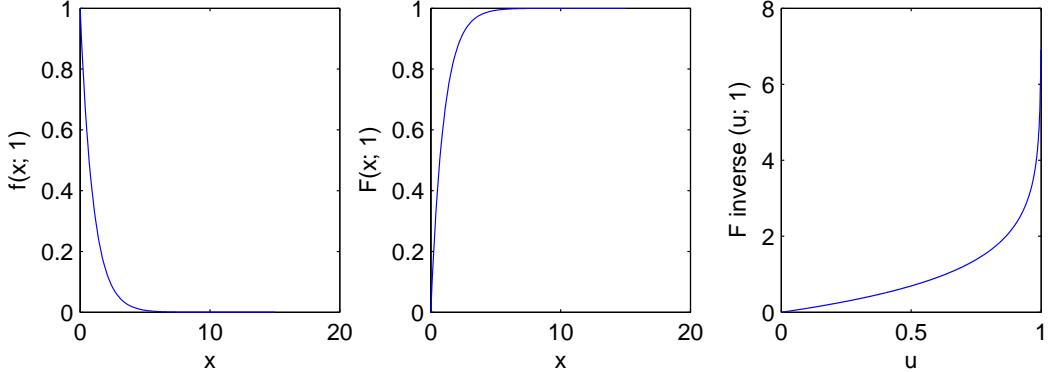
The M-file `guiInversionSampler.m` will bring a graphical user interface (GUI) as shown in Figure 4.5. The default target distribution is  $\text{Uniform}(-5, 5)$ . Now repeatedly push the “Draw one sample” button several times and comprehend the simulation process. You can press “Draw 100 samples” to 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  $\theta_1$  and  $\theta_2$  of  $\text{Uniform}(\theta_1, \theta_2)$  RV.

Figure 4.5: Visual Cognitive Tool GUI: Inversion Sampling from  $X \sim \text{Uniform}(-5, 5)$ .



Recall the  $\text{Exponential}(\lambda)$  RV of Model 8. Let us simulate from it using the inversion sampler.

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.

Figure 4.6: The PDF  $f$ , DF  $F$ , and inverse DF  $F^{[-1]}$  of the Exponential( $\lambda = 1.0$ ) RV.

**Simulation 133 (Exponential( $\lambda$ ))** For a given  $\lambda > 0$ , an Exponential( $\lambda$ ) RV has the following PDF  $f$ , DF  $F$  and inverse DF  $F^{[-1]}$ :

$$f(x; \lambda) = \lambda e^{-\lambda x} \quad F(x; \lambda) = 1 - e^{-\lambda x} \quad F^{[-1]}(u; \lambda) = \frac{-1}{\lambda} \log_e(1-u) \quad (4.1)$$

We write the natural logarithm  $\log_e$  as  $\log$  for notational simplicity. An implementation of the Inversion Sampler for 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 Exponential( $\lambda = 1.0$ ) RV by:

```
lambda=1.0; % some value for lambda
u=rand; % rand is the Fundamental Sampler
ExpInvCDF(u,lambda) % sample from Exponential(1) RV via function in ExpInvCDF.m
```

Because of the following (recall Example 65):

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

we could save a subtraction operation in the above algorithm by replacing  $-(1/\lambda) * \log(1-u)$  by  $-(1/\lambda) * \log(u)$ . Recall that the transformation of  $U \sim \text{Uniform}(0, 1)$  by  $X = -(1/\lambda) \log(U)$  is exactly how we defined  $X$  as the Exponential( $\lambda$ ) RV in Model 8. 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);
```

---

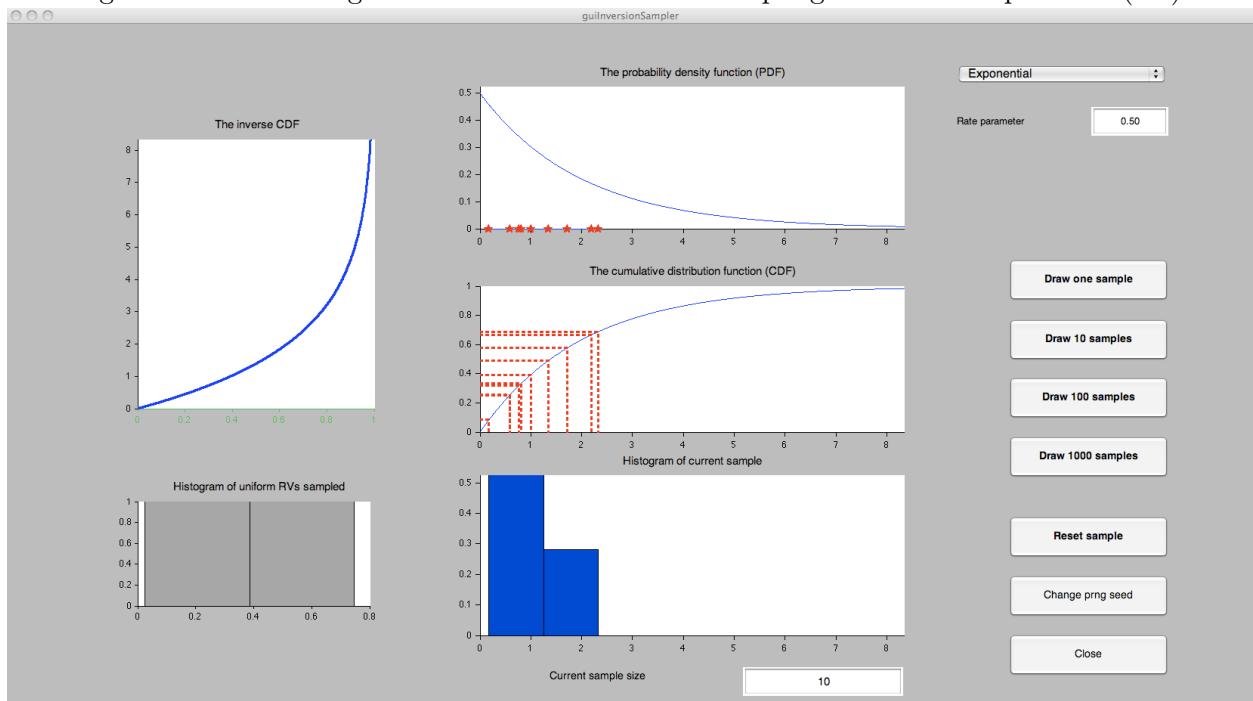
```
>> rand('twister',46678); % initialise the fundamental sampler
>> Lambda=1.0; % declare Lambda=1.0
>> x=ExpInvSam(rand(1,5),Lambda); % pass an array of 5 Uniform(0,1) samples from rand
>> disp(x); % display the Exponential(1.0) distributed samples
0.5945    2.5956    0.9441    1.9015    1.3973
```

**Labwork 134 (Inversion Sampler Demo – Exponential(0.5))** Let us understand the inversion sampler by calling the interactive visual cognitive tool:

```
>> guiInversionSampler
```

The M-file `guiInversionSampler.m` will bring a graphical user interface (GUI) as shown in Figure 4.7. First change the target distribution from the default Uniform( $-5, 5$ ) to Exponential( $0.5$ ) from the drop-down menu. Now push the “Draw 10 samples” button and comprehend the simulation process. Next try changing the “Rate Parameter” from  $0.5$  to  $10.0$  for example and generate several inversion samples and see the density histogram of the accumulating samples. You can press “Draw one sample” to really comprehend the inversion sampler in action one step at a time.

Figure 4.7: Visual Cognitive Tool GUI: Inversion Sampling from  $X \sim \text{Exponential}(0.5)$ .



It is straightforward to do replicate experiments. Consider the experiment of drawing five independent samples from the  $\text{Exponential}(\lambda = 1.0)$  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 135 (Next seven buses at your bus-stop)** Consider the problem of modelling the arrival of buses at a bus stop. Suppose that the time between arrivals is an  $\text{Exponential}(\lambda = 0.1)$  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 19 (Laplace( $\lambda$ ) or Double Exponential( $\lambda$ ) RV)** If a RV  $X$  is equally likely to be either positive or negative with an exponential density, then the Laplace( $\lambda$ ) or Double Exponential( $\lambda$ ) RV, with the rate parameter  $\lambda > 0, \lambda \in \mathbb{R}$ , may be used to model it. The density function for the Laplace( $\lambda$ ) RV given by  $f(x; \lambda)$  is

$$f(x; \lambda) = \frac{\lambda}{2} e^{-\lambda|x|} = \begin{cases} \frac{\lambda}{2} e^{\lambda x} & \text{if } x < 0 \\ \frac{\lambda}{2} e^{-\lambda x} & \text{if } x \geq 0 \end{cases}. \quad (4.2)$$

Let us define the sign of a real number  $x$  by

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}.$$

Then, the DF of the Laplace( $\lambda$ ) RV  $X$  is

$$F(x; \lambda) = \int_{-\infty}^x f(y; \lambda) dy = \frac{1}{2} \left( 1 + \text{sign}(x) \left( 1 - e^{-\lambda|x|} \right) \right), \quad (4.3)$$

and its inverse DF is

$$F^{[-1]}(u; \lambda) = -\frac{1}{\lambda} \text{sign} \left( u - \frac{1}{2} \right) \log \left( 1 - 2 \left| u - \frac{1}{2} \right| \right), \quad u \in [0, 1] \quad (4.4)$$

**Mean and Variance of Laplace( $\lambda$ ) RV  $X$ :** Show that the mean of a Laplace( $\lambda$ ) RV  $X$  is

$$\mathbb{E}(X) = \int_0^\infty x f(x; \lambda) dx = \int_0^\infty x \frac{\lambda}{2} e^{-\lambda|x|} dx = 0,$$

and the variance is

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

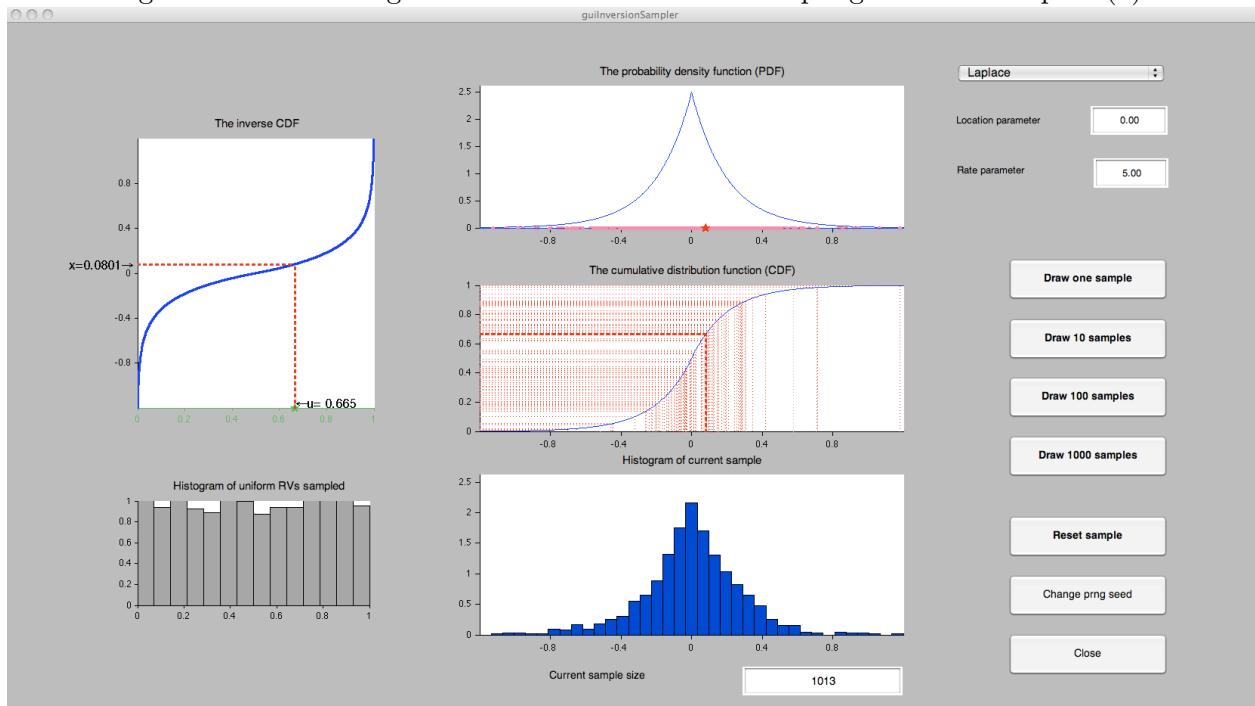
Note that the mean is 0 due to the symmetry of the density about 0 and the variance is twice that of the  $\text{Exponential}(\lambda)$  RV.

**Labwork 136 (Rejection Sampler Demo – Laplace(5))** Let us comprehend the rejection sampler by calling the interactive visual cognitive tool:

```
>> guiInversionSampler
```

The M-file `guiInversionSampler.m` will bring a graphical user interface (GUI) as shown in Figure 4.8. Using the drop-down menu change from the default target distribution Uniform( $-5, 5$ ) to Laplace(5). Now repeatedly push the “Draw one sample” button several times and comprehend the simulation process. You can also press “Draw 1000 samples” and see the density histogram of the generated samples. Next try changing the numbers in the “Rate parameter” box from 5.00 to 1.00 in order to alter the parameter  $\lambda$  of 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 Laplace( $\lambda$ ) implicitly had a location parameter of 0.00, we can easily introduce a location parameter  $\mu$  into the PDF. With a pencil and paper try to rewrite the PDF in (4.2) with an additional location parameter  $\mu$ .

Figure 4.8: Visual Cognitive Tool GUI: Inversion Sampling from  $X \sim \text{Laplace}(5)$ .



**Simulation 137 (Laplace( $\lambda$ ))** Here is an implementation of an inversion sampler to draw IID samples from a Laplace( $\lambda$ ) RV  $X$  by transforming IID samples from the Uniform(0, 1) RV  $U$ :

---

```
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 Laplace( $\lambda = 1.0$ ) RV by

```
>> lambda=1.0; % some value for lambda
>> rand('twister',6567); % initialize the fundamental sampler
>> u=rand(1,5); % draw 5 IID samples from Uniform(0,1) RV
>> disp(u); % display the samples in u
0.6487 0.9003 0.3481 0.6524 0.8152

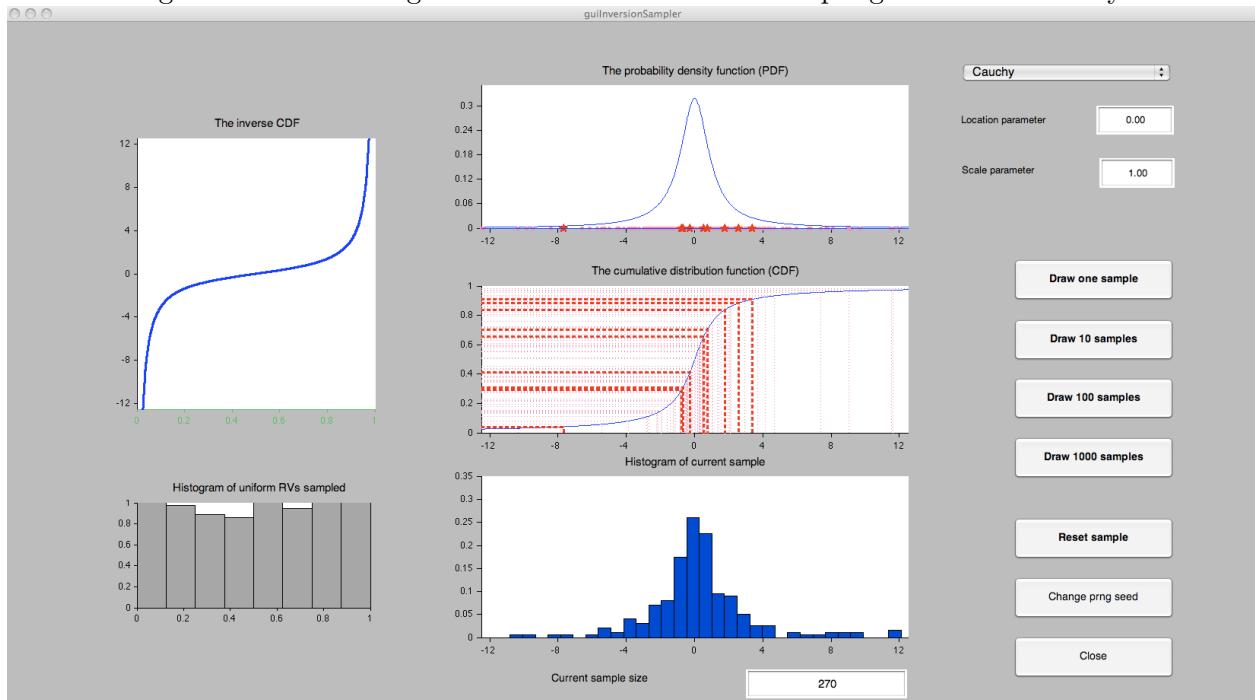
>> x=LaplaceInvCDF(u,lambda); % draw 5 samples from Laplace(1) RV using inverse CDF
>> disp(x); % display the samples
0.3530 1.6127 -0.3621 0.3637 0.9953
```

**Labwork 138 (Inversion Sampler Demo – Cauchy)** Let us comprehend the inversion sampler by calling the interactive visual cognitive tool:

```
>> guiInversionSampler
```

The M-file `guiInversionSampler.m` will bring a graphical user interface (GUI) as shown in Figure 4.9. Using the drop-down menu change from the default target distribution Uniform( $-5, 5$ ) to Cauchy RV of Model 13. Now repeatedly push the “Draw one sample” button several times and comprehend the simulation process. You can also press “Draw 10 samples” several times and see the density histogram of the generated samples. Next try changing the numbers in the “Scale parameter” and “Location Parameter” boxes from the default values of 1.00 and 0.00, respectively. Although our formulation of Cauchy RV is also called *Standard Cauchy* as it implicitly had a location parameter of 0.00 and scale parameter of 1. With a pencil and paper (in conjunction with a wikipedia search if you have to) try to rewrite the PDF in (3.53) with an additional location parameter  $\mu$  and scale parameter  $\sigma$ .

Figure 4.9: Visual Cognitive Tool GUI: Inversion Sampling from  $X \sim \text{Cauchy}$ .



**Simulation 139 (Cauchy)** We can draw  $n$  IID samples from the Cauchy RV  $X$  by transforming  $n$  IID samples from Uniform(0, 1) RV  $U$  using the inverse DF as follows:

```
>> rand('twister',2435567);      % initialise the fundamental sampler
>> u=rand(1,5);                % draw 5 IID samples from Uniform(0,1) RV
>> disp(u);
    0.7176    0.6655    0.9405    0.9198    0.2598
>> x=tan(pi * u);            % draw 5 samples from Standard cauchy RV using inverse CDF
>> disp(x); % display the samples in x
   -1.2272   -1.7470   -0.1892   -0.2575    1.0634
```

### 4.3.2 Inversion Sampler for Discrete Random Variables

Next, consider the problem of **sampling from a random variable  $X$  with a discontinuous or discrete DF** using the inversion sampler. We need to define the inverse more carefully here.

**Proposition 59 (Inversion sampler with compact support)** Let the support of the RV  $X$  be over some real interval  $[a, b]$  and let its inverse DF be defined as follows:

$$F^{[-1]}(u) := \inf\{x \in [a, b] : F(x) \geq u, 0 \leq u \leq 1\}.$$

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

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

**Simulation 140 (Bernoulli( $\theta$ ))** Consider the problem of simulating from a Bernoulli( $\theta$ ) RV based on an input from a Uniform(0, 1) RV. Recall that  $\lfloor x \rfloor$  (called the ‘floor of  $x$ ’) is the largest integer that is smaller than or equal to  $x$ , e.g.  $\lfloor 3.8 \rfloor = 3$ . Using the floor function, we can simulate a Bernoulli( $\theta$ ) 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,n)+theta); % vectorize the operation
>> disp(x) % display the outcomes of the simulation
    0     0     1     0     0     0     0     0     1     1
```

Again, it is straightforward to do replicate experiments, e.g. to demonstrate the Central Limit Theorem for a sequence of  $n$  IID Bernoulli( $\theta$ ) 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)
```

Recall the Point Mass( $\theta$ ) RV. Formally, we can simulate from it trivially as follows.

**Simulation 141** (Point Mass( $\theta$ )) Let us simulate a sample from the Point Mass( $\theta$ ) RV  $X$ . Since this RV produces the same realisation  $\theta$  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 = Sim1PointMass(u,theta);
% Input      : u = one uniform random number eg. rand()
%               theta = a real number (scalar)
% Output     : x = sample from X
x=theta;
```

---

Here is call to the function.

```
>> Sim1PointMass(rand(),2)
ans =
    2
>> % % we can use arrayfun to apply Sim1Pointmass to any array of Uniform(0,1) samples
>> arrayfun(@(u)(Sim1PointMass(u,17)),rand(2,10))
ans =
    17    17    17    17    17    17    17    17    17    17
    17    17    17    17    17    17    17    17    17    17
```

Note that it is not necessary to have input IID samples from Uniform(0, 1) RV via `rand` in order to draw samples from the Point Mass( $\theta$ ) RV. For instance, an input matrix of zeros can do the job:

```
>> arrayfun(@(u)(Sim1PointMass(u,17)),zeros(2,8))
ans =
    17    17    17    17    17    17    17    17
    17    17    17    17    17    17    17    17
```

Next we simulate from de Moivre( $\theta_1, \theta_2, \dots, \theta_k$ ) RV  $X$  of Model 14 via its inverse DF

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

given by:

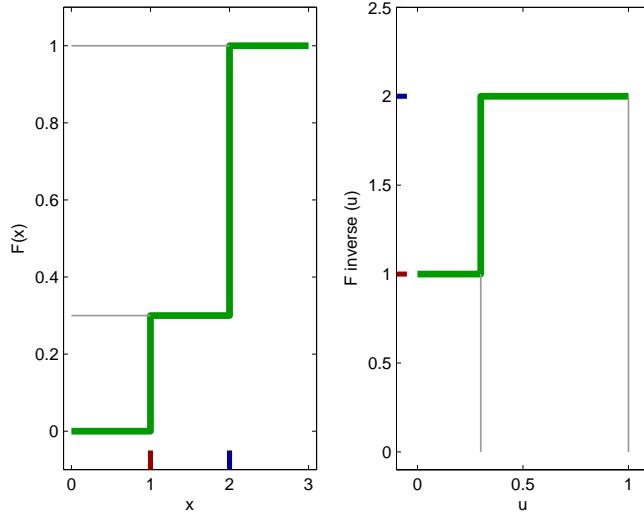
$$F^{[-1]}(u; \theta_1, \theta_2, \dots, \theta_k) = \begin{cases} 1 & \text{if } 0 \leq u < \theta_1 \\ 2 & \text{if } \theta_1 \leq u < \theta_1 + \theta_2 \\ 3 & \text{if } \theta_1 + \theta_2 \leq u < \theta_1 + \theta_2 + \theta_3 \\ \vdots & \\ k & \text{if } \theta_1 + \theta_2 + \dots + \theta_{k-1} \leq u < 1 \end{cases} \quad (4.5)$$

When  $k = 2$  in the 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 4.10.

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

**Simulation 142** (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 4 produces samples from the *de Moivre*( $1/k, 1/k, \dots, 1/k$ ) RV  $X$ .

The M-file implementing Algorithm 4 is:

Figure 4.10: The DF  $F(x; 0.3, 0.7)$  of the de Moivre(0.3, 0.7) RV and its inverse  $F^{-1}(u; 0.3, 0.7)$ .**Algorithm 4** Inversion Sampler for de Moivre( $1/k, 1/k, \dots, 1/k$ ) RV1: *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:*  $x \leftarrow \lceil ku \rceil$ 

```
function x = SimdeMoivreEqui(u,k);
% return samples from de Moivre(1/k,1/k,...,1/k) RV X
% Call Syntax: x = SimdeMoivreEqui(u,k);
% Input      : u = array of uniform random numbers eg. rand
%               k = number of equi-probabble outcomes of X
% Output     : x = samples from X
x = ceil(k * u); % ceil(y) is the smallest integer larger than y
%x = floor(k * u); if outcomes are in {0,1,...,k-1}
```

Let us use the function `SimdeMoivreEqui` to draw five samples from a fair seven-faced cylindrical dice.

```
>> k=7; % number of faces of the fair dice
>> n=5; % number of trials
>> rand('twister',78657); % initialise the fundamental sampler
>> u=rand(1,n); % draw n samples from Uniform(0,1)
>> % inverse transform samples from Uniform(0,1) to samples
>> % from de Moivre(1/7,1/7,1/7,1/7,1/7,1/7,1/7)
>> outcomes=SimdeMoivreEqui(u,k); % save the outcomes in an array
>> disp(outcomes);
6      5      5      5      2
```

Now, let us consider the more general problem of implementing a sampler for an arbitrary but specified de Moivre( $\theta_1, \theta_2, \dots, \theta_k$ ) RV. That is, the values of  $\theta_i$  need not be equal to  $1/k$ .

**Simulation 143** (de Moivre( $\theta_1, \theta_2, \dots, \theta_k$ )) We can generate samples from a 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.

---

**Algorithm 5** 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*:  $x \leftarrow i$

---

The M-file implementing Algorithm 5 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
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 5 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 5. Using the simpler Algorithm 4, which exploits the fact that all values of  $\theta_i$  are equal, we generated samples in constant time, which is denoted by  $O(1)$ .

**Simulation 144** (Geometric( $\theta$ )) We can simulate a sample  $x$  from a 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 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

```

**Labwork 145 (PMF versus relative frequency histogram of simulated Geometric( $\theta$ ) RV)**  
It is a good idea to make a relative frequency histogram of a simulation algorithm and compare that to the PDF of the discrete RV we are simulating from. We use the following script to create Figure 3.7:

---

```
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)')
```

---

Let us simulate from the Binomial( $n, \theta$ ) RV of Model 5.

**Labwork 146 (Binomial coefficient)** The MATLAB function `BinomialCoefficient` can be used to compute:

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

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 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 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 147** ( $\text{Binomial}(n, \theta)$  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 147, 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.

The  $\text{Binomial}(n, \theta)$  RV  $X$  is related to the IID  $\text{Geometric}(\theta)$  RV  $Y_1, Y_2, \dots$ :  $X$  is the number of successful  $\text{Bernoulli}(\theta)$  outcomes (outcome is 1) that occur in a total of  $n$   $\text{Bernoulli}(\theta)$  trials, with the number of trials between consecutive successes distributed according to IID  $\text{Geometric}(\theta)$  RV.

**Simulation 148** ( $\text{Binomial}(\theta)$  from IID  $\text{Geometric}(\theta)$  RVs) By this principle, we can simulate from the  $\text{Binomial}(\theta)$   $X$  by Step 1: generating IID  $\text{Geometric}(\theta)$  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  $\text{Binomial}(n = 10, \theta = 0.5)$  RV:

```
>> theta=0.5; % declare theta
>> n=10; % say n=10
>> SampleSize=12;% say you want to simulate 12 samples
>> rand('twister',10001) % seed the fundamental sampler
>> Samples=arrayfun(@(T)Sim1BinomByGeoms(n,T),theta*ones(1,SampleSize))
Samples =    7     5     8     8     4     1     4     8     2     4     6     5
```

Figure 3.8 depicts a comparison of the PDF of  $\text{Binomial}(n = 10, \theta = 0.5)$  RV and a relative frequency histogram based on 100,000 simulations from it.

Let us simulate from the  $\text{Poisson}(\lambda)$  RV of Model 6 as shown in Figure 3.18.

**Simulation 149** ( $\text{Poisson}(\lambda)$  from IID  $\text{Exponential}(\lambda)$  RVs) By this principle, we can simulate from the  $\text{Poisson}(\lambda)$   $X$  by Step 1: generating IID  $\text{Exponential}(\lambda)$  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:

---

```
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 3.18 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 20** ( $GD(\theta_0, \theta_1, \dots)$ ) 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 6 allows us to simulate from any member of the class of non-negative discrete RVs as specified by the probabilities  $(\theta_0, \theta_1, \dots)$ . 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  $D(i) = x : \mathbb{Z}_+ \rightarrow \mathbb{X}$  that allows us to think of  $(0, 1, 2, \dots)$  as indices of an array  $D$  giving  $\mathbb{X} = (D(0), D(1), \dots)$ .

---

**Algorithm 6** Inversion Sampler for  $GD(\theta_0, \theta_1, \dots)$  RV  $X$ 


---

1: *input:*

1.  $\theta_0$  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 150** ( $\text{Binomial}(n, \theta)$ ) To simulate from a  $\text{Binomial}(n, \theta)$  RV  $X$ , we can use Algorithm 6 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  $\theta_0$  and  $C(x + 1)$ , we can also simulate from the  $\text{Geometric}(\theta)$  and  $\text{Poisson}(\lambda)$  RVs.

**Labwork 151** This is a challenging exercise for the student who is finding the other Labworks too easy. So those who are novice to MATLAB may skip this Labwork.

1. Implement Algorithm 6 via a function named `MyGenDiscInvSampler` in MATLAB. Hand in the M-file named `MyGenDiscInvSampler.m` giving detailed comments explaining your understanding of each step of the code. [Hint:  $C(i)$  should be implemented as a function (use function handles via @) that can be passed as a parameter to the function `MyGenDiscInvSampler`].
2. Show that your code works for drawing samples from a  $\text{Binomial}(n, p)$  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  $2 \times 2$  table with column headings `x` and No. of observations. [Hint: the inputs  $\theta_0$  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  $\theta_0$  and  $C(i)$  for the Geometric( $\theta$ ) RV should be derived and the workings shown].

### 4.3.3 von Neumann Rejection Sampler (RS)

Rejection sampling [John von Neumann, 1947, in *Stanislaw Ulam 1909-1984*, a special issue of Los Alamos Science, Los Alamos National Lab., 1987, p. 135-136] is a Monte Carlo method to draw independent samples from a target RV  $X$  with probability density  $f(x)$ , where  $x \in \mathbb{X} \subset \mathbb{R}^k$ . Typically, the target density  $f$  is only known up to a constant and therefore the (normalised) density  $f$  itself may be unknown and it is difficult to generate samples directly from  $X$ .

Suppose we have another density or mass function  $g$  for which the following are true:

- (a) we can generate random variables from  $g$ ;
- (b) the support of  $g$  contains the support of  $f$ , i.e.  $\mathbb{Y} \supset \mathbb{X}$ ;
- (c) a constant  $a > 1$  exists, such that:

$$f(x) \leq ag(x). \quad (4.6)$$

for any  $x \in \mathbb{X}$ , the support of  $X$ . Then  $x$  can be generated from Algorithm 7.

---

#### Algorithm 7 Rejection Sampler (RS) of von Neumann

1: *input*:

- (1) a target density  $f(x)$ ,
- (2) a proposal density  $g(x)$  satisfying (a), (b) and (c) above.

2: *output*: a sample  $x$  from RV  $X$  with density  $f$

3: **repeat**

4:   Generate  $y \sim g$  and  $u \sim \text{Uniform}(0, 1)$

5: **until**  $u \leq \frac{f(y)}{ag(y)}$

6: *return*:  $x \leftarrow y$

---

**Proposition 60 (Fundamental Theorem of Simulation)** The von Neumann rejection sampler of Algorithm 7 produces a sample  $x$  from the random variable  $X$  with density  $f(x)$ .

**Proof:** We shall prove the result for the continuous case. For any real number  $t$ :

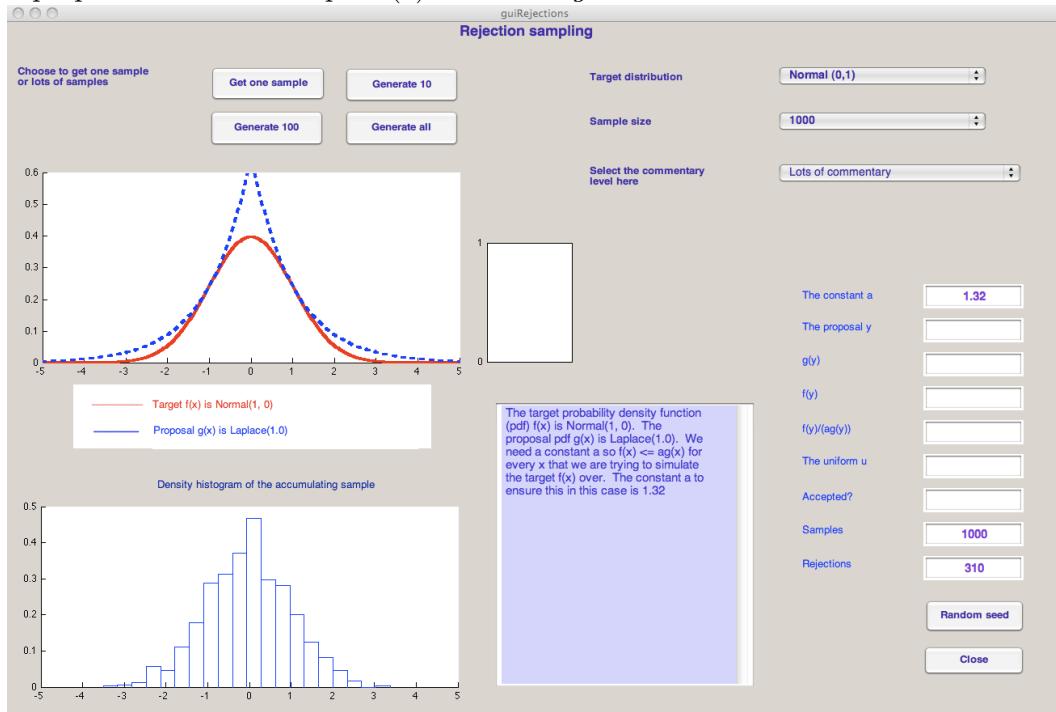
$$\begin{aligned} F(t) &= P(X \leq t) = P\left(Y \leq t \mid U \leq \frac{f(Y)}{ag(Y)}\right) = \frac{P\left(Y \leq t, U \leq \frac{f(Y)}{ag(Y)}\right)}{P\left(U \leq \frac{f(Y)}{ag(Y)}\right)} \\ &= \frac{\int_{-\infty}^t \left( \int_0^{f(y)/ag(y)} 1 du \right) g(y) dy}{\int_{-\infty}^{\infty} \left(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

**Labwork 152 (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 4.11. Try various buttons and see how the output changes with explanations. Try switching the “Target distribution” to “Mywavy4” and generate several rejection samples and see the density histogram of the accumulating samples.

Figure 4.11: Visual Cognitive Tool GUI: Rejection Sampling from  $X \sim \text{Normal}(0, 1)$  with PDF  $f$  based on proposals from  $Y \sim \text{Laplace}(1)$  with PDF  $g$ .



**Simulation 153 (Rejection Sampling Normal(0, 1) with Laplace(1) proposals)** Suppose we wish to generate from  $X \sim \text{Normal}(0, 1)$ . Consider using the rejection sampler with proposals from  $Y \sim \text{Laplace}(1)$  (using inversion sampler of Simulation 137). The support of both RVs is  $(-\infty, \infty)$ . Next:

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

and therefore:

$$\frac{f(x)}{g(x)} = \sqrt{\frac{2}{\pi}} \exp\left(|x| - \frac{x^2}{2}\right) \leq \sqrt{\frac{2}{\pi}} \exp\left(\frac{1}{2}\right) = a \approx 1.3155 .$$

Hence, we can use the rejection method with:

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

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

---

```
RejectionNormalLaplace.m
```

```

function x = RejectionNormalLaplace()
Accept = 0; % a binary variable to indicate whether a proposed point is accepted
while ~Accept % ~ is the logical NOT operation
    y = LaplaceInvCDF(rand(),1); % sample Laplace(1) RV
    Bound = exp( abs(y) - (y*y+1)/2 );
    u = rand();
    if u <= Bound
        x = y;
        Accept = 1;
    end % if
end % while

```

---

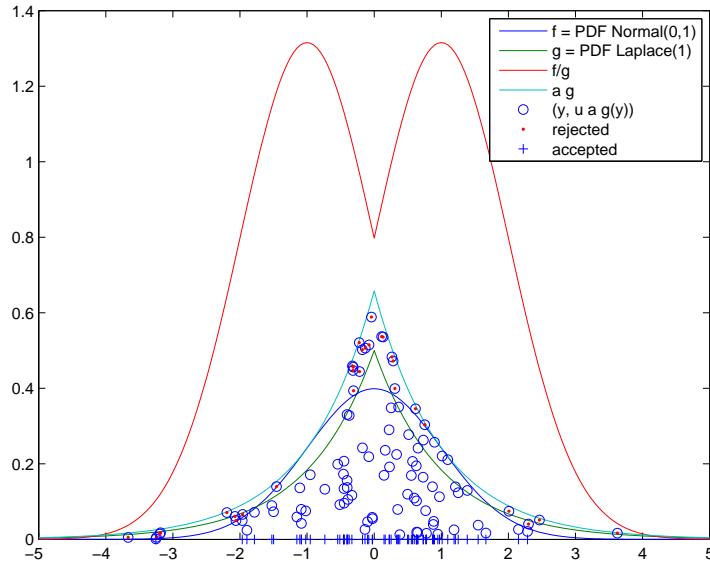
We may obtain a large number of samples and plot them as a histogram using the following commands:

```

>> % 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 4.12: Rejection Sampling from  $X \sim \text{Normal}(0, 1)$  with PDF  $f$  based on 100 proposals from  $Y \sim \text{Laplace}(1)$  with PDF  $g$ .



**Classwork 154 (A note on the proposal's tail in rejection sampling)** The condition  $f(x) \leq ag(x)$  is equivalent to  $f(x)/g(x) \leq a$ , which says that  $f(x)/g(x)$  must be bounded; therefore,  $g$  must have higher tails than  $f$ . The rejection method cannot be used to generate from a Cauchy distribution using a normal distribution, because the latter has lower tails than the former.

The next result tells us how many iterations of the algorithm are needed, on average, to get a sample value from a RV with PDF  $f$ .

**Proposition 61 (Acceptance Probability of RS)** The expected number of iterations of the rejection algorithm to get a sample  $x$  is the constant  $a$ .

**Proof:** For the continuous case:

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

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:

$$\tilde{f}(x) \leq \tilde{a}\tilde{g}(x), \quad (4.7)$$

for any  $x \in \mathbb{X}$ , the support of  $X$ . Then  $x$  can be generated from Algorithm 8.

---

**Algorithm 8** Rejection Sampler (RS) of von Neumann – target shape

---

1: *input:*

- (1) shape of a target density  $\tilde{f}(x) = \left( \int \tilde{f}(x) dx \right) f(x)$ ,
- (2) a proposal density  $g(x)$  satisfying (a), (b) and (c) above.

2: *output:* a sample  $x$  from RV  $X$  with density  $f$

3: **repeat**

4:   Generate  $y \sim g$  and  $u \sim \text{Uniform}(0, 1)$

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  $Z$ .<sup>1</sup>

**Labwork 155 (Gaussian Sampling with `randn`)** We can use MATLAB function `randn` that implements the Ziggurat method to draw samples from an RV  $Z \sim \text{Normal}(0, 1)$  as follows:

---

<sup>1</sup>See [http://en.wikipedia.org/wiki/Ziggurat\\_algorithm](http://en.wikipedia.org/wiki/Ziggurat_algorithm) for more details.

```
>> 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  $\mu$  and  $\sigma$ , 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 156 (Sampling from truncated normal distributions)** [Christian P. Robert, Simulation of truncated normal variables, Statistics and Computing (1995) 5, 121-125] Let  $N_+(\mu, \tau, \sigma^2)$  denote the left-truncated normal distribution with truncation point  $\tau$  and density given by

$$f(x|\mu, \tau, \sigma^2) = \frac{\exp(-(x-\mu)^2/2\sigma^2)}{\sqrt{2\pi}\sigma[1 - \Phi((\tau-\mu)/\sigma)]} \mathbf{1}_{x \geq \tau}.$$

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  $\tau$  is obtained. When  $\tau > \mu$ , however, this can be inefficient and increasingly so as  $\tau$  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:

$$g(y|\lambda, \tau) = \lambda \exp(-\lambda(y-\tau)) \mathbf{1}_{y \geq \tau}.$$

1. Show that for simulating from  $N_+(\mu = 0, \tau, \sigma^2 = 1)$  when  $\tau \geq 0$ , the best choice of  $\lambda$  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  $\tau$  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  $\tau$ . 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,  $\mu$ ,  $\sigma^2$  and  $\tau$ .

## 4.4 Exercises in Simulation

**Ex. 4.1** — Suppose the continuous RV  $X$  has PDF:

$$f_X(x) = (\pi(1 + x^2))^{-1}$$

Devise an algorithm to transform samples from Uniform(0, 1) RV to those from  $X$ . Present your answer as pseudo-code.

# Chapter 5

## Limit Laws of Statistics

### 5.1 Convergence of Random Variables

This important topic is concerned with the limiting behavior of sequences of RVs. We want to understand what it means for a sequence of random variables  $\{X_n\}_{n=1}^{\infty} := X_1, X_2, \dots$  to converge to another random variable  $X$ , when all RVs are defined on the same probability space  $(\Omega, \mathcal{F}, P)$ .

$$\{X_i\}_{i=1}^n := X_1, X_2, X_3, \dots, X_{n-1}, X_n \quad \text{as } n \rightarrow \infty .$$

From a statistical or decision-making viewpoint, as you will see in Inference Theory I course,  $n \rightarrow \infty$  is associated with the amount of data or information  $\rightarrow \infty$ . More abstractly, we are interested in what happens to the limiting RV  $X := \lim_{n \rightarrow \infty} X_n$  when given the DFs  $F_n(x)$  for each  $X_n$ .

We need different notions of convergence to characterize such a behavior: two simplest behaviors are that the sequence eventually takes a constant value  $\theta$ , i.e.  $X_n$  approaches  $X \sim \text{Point Mass}(\theta)$  RV, or that values in the sequence continue to change but can be described by an unchanging probability distribution, i.e.,  $X_n$  approaches  $X \sim F(x)$ . See [https://en.wikipedia.org/wiki/Convergence\\_of\\_random\\_variables](https://en.wikipedia.org/wiki/Convergence_of_random_variables).

Let us first refresh ourselves with notions of convergence, limits and continuity in the real line (Sec. 1.8.1) before proceeding further.

Can the sequences of  $\{\text{Point Mass}(\theta_i = 17)\}_{i=1}^{\infty}$  and  $\{\text{Point Mass}(\theta_i = 1/i)\}_{i=1}^{\infty}$  RVs be the same as the two sequences of real numbers  $\{x_i\}_{i=1}^{\infty} = 17, 17, 17, \dots$  and  $\{x_i\}_{i=1}^{\infty} = \frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \dots$  we saw in Examples 12 and 13?

Yes why not – just move to space of distributions over the reals! See Figure 5.1.

**Classwork 157 (Convergence of  $X_i \sim \text{Normal}(0, 1/i)$ )** Suppose you are given an independent sequence of RVs  $\{X_i\}_{i=1}^n$ , where  $X_i \sim \text{Normal}(0, 1/i)$ . How would you talk about the convergence of  $X_n \sim \text{Normal}(0, 1/n)$  as  $n$  approaches  $\infty$ ? Take a look at Figure 5.2 for insight. The probability mass of  $X_n$  increasingly concentrates about 0 as  $n$  approaches  $\infty$  and the variance  $1/n$  approaches 0, as depicted in Figure 5.2. 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  $P(X_n = X) = 0$  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 5.1: Sequence of  $\{\text{Point Mass}(17)\}_{i=1}^{\infty}$  RVs (left panel) and  $\{\text{Point Mass}(1/i)\}_{i=1}^{\infty}$  RVs (only the first seven are shown on right panel) and their limiting RVs in red.

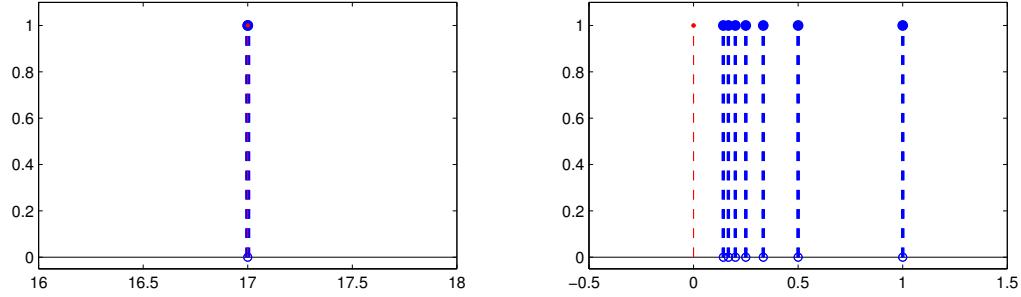
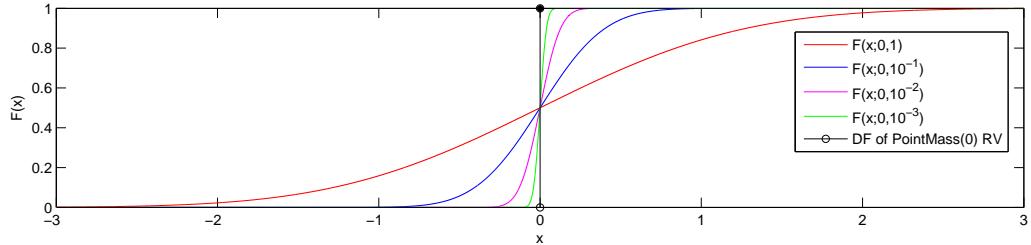


Figure 5.2: Distribution functions of several  $\text{Normal}(\mu, \sigma^2)$  RVs for  $\sigma^2 = 1, \frac{1}{10}, \frac{1}{100}, \frac{1}{1000}$ .



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 two basic propositions in Statistics :

1. Law of Large Numbers,
2. Central Limit Theorem,

**Definition 62 (Convergence in Distribution (or Weakly, or in Law))** Let  $X_1, X_2, \dots$ , 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 19 of a DF, can be equivalently expressed as follows:

$$\begin{aligned} \lim_{n \rightarrow \infty} P(\{\omega : X_n(\omega) \leq t\}) &= P(\{\omega : X(\omega) \leq t\}), \\ \text{i.e. } P(\{\omega : X_n(\omega) \leq t\}) &\rightarrow P(\{\omega : X(\omega) \leq t\}), \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Let us revisit the problem of convergence in Classwork 157 armed with our new notions of convergence.

**Example 158 (Convergence in distribution)** Suppose you are given an independent sequence of RVs  $\{X_i\}_{i=1}^n$ , where  $X_i \sim \text{Normal}(0, 1/i)$  with DF  $F_n$  and let  $X \sim \text{Point Mass}(0)$  with DF  $F$ .

We can formalize our observation in Classwork 157 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

$$F_n(t) = P(X_n < t) = P(\sqrt{n}X_n < \sqrt{nt}) = P(Z < \sqrt{nt}).$$

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  $n \rightarrow \infty$ ,

$$\lim_{n \rightarrow \infty} F_n(t) = \lim_{n \rightarrow \infty} P(Z < \sqrt{nt}) = 0 = F(t).$$

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  $n \rightarrow \infty$ ,

$$\lim_{n \rightarrow \infty} F_n(t) = \lim_{n \rightarrow \infty} P(Z < \sqrt{nt}) = 1 = F(t).$$

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  $F$ .

Convergence in distribution does not in general imply that the sequence of corresponding probability density functions will also converge. Consider for example RV  $X_n$  with density  $\mathbb{1}_{(0,1)}(x)(1 - \cos(2\pi nx))$ . These RVs converge in distribution to  $X \sim \text{Uniform}(0, 1)$ , but their densities (PDFs) do not converge at all as evident in Figure 5.3.

**Proposition 63 (Scheffé's Theorem)** According to **Scheffé's Theorem** convergence of the probability density function (for a continuous RV) or probability mass function (for a discrete RV) implies convergence in distribution.

**Proof:** We will state this without a Proof here as Proof of the Theorem requires measure theory in generality<sup>1</sup>. However, you should be able to see why convergence of PMFs  $f_n(x)$  for discrete RVs  $X_n$ , to  $f(x)$ , the PMF of another discrete RV  $X$ , implies convergence in their corresponding DFs, i.e.,  $F_n(x) \rightarrow F(x)$  for each  $x$  as  $n \rightarrow \infty$ .

---

<sup>1</sup>See [https://en.wikipedia.org/wiki/Scheff%C3%A9%27s\\_lemma](https://en.wikipedia.org/wiki/Scheff%C3%A9%27s_lemma).

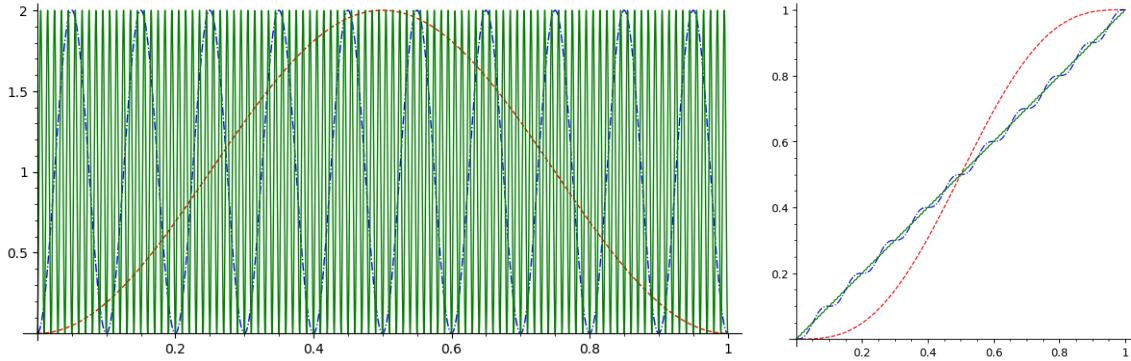


Figure 5.3: PDF  $f_{X_n}(x) := \mathbb{1}_{(0,1)}(x)(1 - \cos(2\pi nx))$  of the RV  $X_n$  [the left sub-figure] and its DF  $F_{X_n}(x) := \int_{-\infty}^x \mathbb{1}_{(0,1)}(v)(1 - \cos(2\pi nv))dv$  [the right sub-figure], for  $n = 1$  [red '---'],  $n = 10$  [blue '-.-'], and  $n = 100$  [green '-'], respectively. One can see clear convergence of the DFs  $F_n$  to  $\mathbb{1}_{(0,1)}(x)x$ , the DF of the Uniform( $0, 1$ ) RV, while the corresponding PDFs  $f_n(x)$  keep oscillating wildly with  $n$  across  $[0, 2]$  about  $\mathbb{1}_{(0,1)}(x)$ , the PDF of the Uniform( $0, 1$ ) RV  $X$ . Thus giving a counter-example to the claim that convergence in DFs does not imply convergence in PDFs.

Since  $F(x) = P(X \leq x)$ , convergence in distribution means that the probability for  $X_n$  to be in a given range is approximately equal to the probability that the value of the limiting RV  $X$  is in that range, provided  $n$  is sufficiently large.

Thus, for a discrete sequence of RVs  $X_n$  'n to converge in distribution to another discrete RV  $X$  taking values in  $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$ , it is sufficient to show that  $\lim_{n \rightarrow \infty} P(X_n = x) = P(X = x)$  for each  $x \in \mathbb{Z}_+$ . We will use this fact to prove why we can approximate Binomial RVs by a Poisson under some limiting conditions.

**Example 159** ( $\text{Binomial}(n, \lambda/n) \rightsquigarrow \text{Poisson}(\lambda)$ ) In several situations, as we saw already, it becomes cumbersome to model the events using the  $\text{Binomial}(n, \theta)$  RV, especially when the parameter  $\theta \propto 1/n$  and the events become rare.

$\text{Binomial}(n, \lambda/n)$  converges in distribution to  $\text{Poisson}(\lambda)$  as  $n \rightarrow \infty$ ,  $\theta = \lambda/n \rightarrow 0$

However, for some real parameter  $\lambda > 0$ , the  $\text{Binomial}(n, \lambda/n)$  RV with probability of the number of successes in  $n$  trials, with per-trial success probability  $\lambda/n$ , approaches the Poisson distribution with expectation  $\lambda$ , as  $n$  approaches  $\infty$  (actually, it converges in distribution). The  $\text{Poisson}(\lambda)$  RV is much simpler to work with than the combinatorially laden  $\text{Binomial}(n, \theta = \lambda/n)$  RV. We sketch the details of this next.

Let  $X_n \sim \text{Binomial}(n, \theta = \lambda/n)$  and  $Y \sim \text{Poisson}(\lambda)$  and let  $\lambda = n\theta$  remain constant as  $n \rightarrow \infty$ ,  $\theta \rightarrow 0$ . We need to show that  $\lim_{n \rightarrow \infty} P(X_n = x) = P(Y = x) = e^{-\lambda} \lambda^x / x!$  for any  $x \in \{0, 1, 2, 3, \dots, n\}$ .

$$\begin{aligned}
 P(X = x) &= \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
 &= \frac{n(n-1)(n-2)\cdots(n-x+1)}{x(x-1)(x-2)\cdots(2)(1)} \left(\frac{\lambda^x}{n^x}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \\
 &= \underbrace{\left(\frac{n}{n}\right) \left(\frac{n-1}{n}\right) \left(\frac{n-2}{n}\right) \cdots \left(\frac{n-x+1}{n}\right)}_{\left(\frac{\lambda^x}{x!}\right)} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\left(1 - \frac{\lambda}{n}\right)^{-x}}
 \end{aligned} \tag{5.1}$$

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 17 and 18, as  $n \rightarrow \infty$ , the expression over the first underbrace approaches  $e^{-\lambda}$  while that over the second underbrace approaches 1. Finally, we get the desired limit:

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

The second notion of convergence of RVs is convergence in probability.

**Definition 64 (Convergence in Probability)** Let  $X_1, X_2, \dots$  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  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \epsilon) = 0 \quad [\text{in the sense of Definition 4}].$$

Once again, the above limit, by (3.1) in our Definition 18 of a RV, can be equivalently expressed as follows:

$$\lim_{n \rightarrow \infty} P(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}) = 0, \quad \text{ie,} \quad P(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}) \rightarrow 0, \quad \text{as } n \rightarrow \infty .$$

For the same sequence of RVs in Classwork 157 and Example 158 we are tempted to ask whether  $X_n \sim \text{Normal}(0, 1/n)$  converges in probability to  $X \sim \text{Point Mass}(0)$ , i.e. whether  $X_n \xrightarrow{P} X$ . We need some elementary inequalities in Probability to help us answer this question. We visit these inequalities next.

**Proposition 65 (Markov's Inequality)** Let  $(\Omega, \mathcal{F}, P)$  be a probability triple and let  $X = X(\omega)$  be a non-negative RV. Then,

$$P(X \geq \epsilon) \leq \frac{E(X)}{\epsilon}, \quad \text{for any } \epsilon > 0 . \quad (5.2)$$

**Proof:**

$$\begin{aligned} X &= X \mathbf{1}_{\{y:y \geq \epsilon\}}(x) + X \mathbf{1}_{\{y:y < \epsilon\}}(x) \\ &\geq X \mathbf{1}_{\{y:y \geq \epsilon\}}(x) \\ &\geq \epsilon \mathbf{1}_{\{y:y \geq \epsilon\}}(x) \end{aligned} \quad (5.3)$$

Finally, taking expectations on both sides of the above inequality and then using the fact that the expectation of an indicator function of an event is simply the probability of that event (3.46), we get the desired result:

$$E(X) \geq \epsilon E(\mathbf{1}_{\{y:y \geq \epsilon\}}(x)) = \epsilon P(X \geq \epsilon) .$$

Let us look at some immediate consequences of Markov's inequality.

**Proposition 66 (Chebychev's Inequality)** For any RV  $X$  and any  $\epsilon > 0$ ,

$$P(|X| > \epsilon) \leq \frac{E(|X|)}{\epsilon} \quad (5.4)$$

$$P(|X| > \epsilon) = P(X^2 \geq \epsilon^2) \leq \frac{E(X^2)}{\epsilon^2} \quad (5.5)$$

$$P(|X - E(X)| \geq \epsilon) = P((X - E(X))^2 \geq \epsilon^2) \leq \frac{E(X - E(X))^2}{\epsilon^2} = \frac{V(X)}{\epsilon^2} \quad (5.6)$$

**Proof:** All three forms of Chebychev's inequality are mere corollaries (careful reapplications) of Markov's inequality.

Armed with Markov's inequality we next enquire the convergence in probability for the sequence of RVs in Classwork 157 and Example 158.

**Example 160 (Convergence in probability)** Does the the sequence of RVs  $\{X_n\}_{n=1}^{\infty}$ , where  $X_n \sim \text{Normal}(0, 1/n)$ , converge in probability to  $X \sim \text{Point Mass}(0)$ , i.e. does  $X_n \xrightarrow{P} X$ ?

To find out if  $X_n \xrightarrow{P} X$ , we need to show that for any  $\epsilon > 0$ ,  $\lim_{n \rightarrow \infty} P(|X_n - X| > \epsilon) = 0$ .

Let  $\epsilon$  be any real number greater than 0, then

$$\begin{aligned} P(|X_n| > \epsilon) &= P(|X_n|^2 > \epsilon^2) \\ &\leq \frac{E(X_n^2)}{\epsilon^2} \quad [\text{by Markov's Inequality (5.2)}] \\ &= \frac{\frac{1}{n}}{\epsilon^2} \rightarrow 0, \quad \text{as } n \rightarrow \infty \quad [\text{in the sense of Definition 4}]. \end{aligned}$$

Hence, we have shown that for any  $\epsilon > 0$ ,  $\lim_{n \rightarrow \infty} P(|X_n - X| > \epsilon) = 0$  and therefore by Definition 64,  $X_n \xrightarrow{P} X$  or  $X_n \xrightarrow{P} 0$ .

**Convention:** When  $X$  has a Point Mass( $\theta$ ) distribution and  $X_n \xrightarrow{P} X$ , we simply write  $X_n \xrightarrow{P} \theta$ .

**Definition 67 (Convergence Almost Surely (or with Probability 1))** To say that the sequence of RVs  $\{X_n\}_{n=1}^{\infty}$  converges almost surely (or with probability 1 or strongly) towards another RV  $X$  on the same probability space  $(\Omega, \mathcal{F}, P)$ , as denoted by

$$X_n \xrightarrow{a.s.} X$$

means that

$$P\left(\left\{\lim_{n \rightarrow \infty} X_n = X\right\}\right) = 1 \iff P\left(\left\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

This means that the values of  $X_n$  approach the value of  $X$ , in the sense that events for which  $X_n$  does not converge to  $X$  have probability 0.

Other notions of convergence are termed sure convergence or pointwise convergence, such as convergence in mean. But the above three types of convergence are elementary.

### 5.1.1 Properties of Convergence of RVs\*\*

We will merely state some properties (without proofs that are hyper-linked for the curious student as they are advanced for this course) and relations between the three notions of convergence with some examples to better appreciate the subtleties among them. You will study the proofs of these statements in Probability Theory II. Just remember that subtle implication relations exist between the three notions.

- Convergence almost surely implies convergence in probability<sup>2</sup>

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X .$$

- By the Borel-Cantelli Lemma<sup>3</sup>, convergence in probability does not imply almost sure convergence in the discrete case<sup>4</sup>
- Convergence in probability implies convergence in distribution<sup>5</sup>

$$X_n \xrightarrow{P} X \implies X_n \rightsquigarrow X .$$

- Convergence in distribution to a constant  $\theta$  implies convergence in probability to  $\theta$ :<sup>6</sup>

$$X_n \rightsquigarrow \text{Point Mass}(\theta) \implies X_n \xrightarrow{P} \text{Point Mass}(\theta) .$$

- In general, convergence in distribution does not imply convergence in probability.

## 5.2 Law of Large Numbers

**Proposition 68 (Law of Large Numbers (LLN):  $\bar{X}_n \xrightarrow{P} E(X_1)$ )** 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  $E(X_1)$  exists, as per (3.45), i.e.,  $E(\text{abs}(X_1)) < \infty$ , and the variance is finite, i.e.,  $V(X_1) < \infty$ , then the sample mean  $\bar{X}_n$  converges in probability to the expectation of any one of the IID RVs, say  $E(X_1)$  by convention. More formally, we write:

$$\text{If } X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1 \text{ and if } E(X_1) \text{ exists, then } \bar{X}_n \xrightarrow{P} E(X_1) .$$

**Proof:** Because  $V(X_1) < \infty$ , we have:

$$\begin{aligned} P(|\bar{X}_n - E(\bar{X}_n)| \geq \epsilon) &= \frac{V(\bar{X}_n)}{\epsilon^2} && [\text{by applying Chebychev's inequality (5.6) to the RV } \bar{X}_n] \\ &= \frac{\frac{1}{n} V(X_1)}{\epsilon^2} && [\text{by the IID assumption of } X_1, X_2, \dots \text{ we can apply (3.77)}] \end{aligned}$$

---

<sup>2</sup> [https://en.wikipedia.org/wiki/Proofs\\_of\\_convergence\\_of\\_random\\_variables#Convergence\\_almost\\_surely\\_implies\\_convergence\\_in\\_probability](https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_almost_surely_implies_convergence_in_probability)

<sup>3</sup> [https://en.wikipedia.org/wiki/Borel%20%93Cantelli\\_lemma](https://en.wikipedia.org/wiki/Borel%20%93Cantelli_lemma)

<sup>4</sup> [https://en.wikipedia.org/wiki/Proofs\\_of\\_convergence\\_of\\_random\\_variables#Convergence\\_in\\_probability\\_does\\_not\\_imply\\_almost\\_sure\\_convergence\\_in\\_the\\_discrete\\_case](https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_in_probability_does_not_imply_almost_sure_convergence_in_the_discrete_case)

<sup>5</sup> [https://en.wikipedia.org/wiki/Proofs\\_of\\_convergence\\_of\\_random\\_variables#Convergence\\_in\\_probability\\_implies\\_convergence\\_in\\_distribution](https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_in_probability_implies_convergence_in_distribution)

<sup>6</sup> [https://en.wikipedia.org/wiki/Proofs\\_of\\_convergence\\_of\\_random\\_variables#Convergence\\_in\\_distribution\\_to\\_a\\_constant\\_implies\\_convergence\\_in\\_probability](https://en.wikipedia.org/wiki/Proofs_of_convergence_of_random_variables#Convergence_in_distribution_to_a_constant_implies_convergence_in_probability)

Therefore, for any given  $\epsilon > 0$ ,

$$\begin{aligned} P(|\bar{X}_n - E(X_1)| \geq \epsilon) &= P(|\bar{X}_n - E(\bar{X}_n)| \geq \epsilon) \quad [\text{by the IID assumption of } X_1, X_2, \dots, E(\bar{X}_n) = E(X_1), \text{ as per (3.76)}] \\ &= \frac{\frac{1}{n} V(X_1)}{\epsilon^2} \rightarrow 0, \quad \text{as } n \rightarrow \infty, \end{aligned}$$

or equivalently,  $\lim_{n \rightarrow \infty} P(|\bar{X}_n - E(X_1)| \geq \epsilon) = 0$ . And the last statement is the definition of the claim made by the law of large numbers (LLN), namely that  $\bar{X}_n \xrightarrow{P} E(X_1)$ .

**Proposition 69 (Weak Law of Large Numbers (WLLN):**  $\bar{X}_n \rightsquigarrow \text{Point Mass}(E(X_1))$ ) If we are given a sequence of independently and identically distributed (IID) RVs,  $X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1$  and if  $E(X_1)$  exists, i.e.  $E(\text{abs}(X)) < \infty$ , then the sample mean  $\bar{X}_n$  converges in distribution to the expectation of any one of the IID RVs, say  $\text{Point Mass}(E(X_1))$  by convention. More formally, we write:

If  $X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1$  and if  $E(X_1)$  exists, then  $\bar{X}_n \rightsquigarrow \text{Point Mass}(E(X_1))$  as  $n \rightarrow \infty$ .

**Proof:** Our proof now is based on the convergence of characteristic functions (CFs) pointwise to the CF of the limiting RV, as this implies, by Lévy's Continuity Theorem on CFs<sup>7</sup>, the convergence of the corresponding distribution functions (DFs).

First, the CF of  $\text{Point Mass}(E(X_1))$  is

$$E(e^{itE(X_1)}) = e^{itE(X_1)},$$

since  $E(X_1)$  is just a constant, i.e., a Point Mass RV that puts all of its probability mass at  $E(X_1)$ .

Second, the CF of  $\bar{X}_n$  is

$$\begin{aligned} E(e^{it\bar{X}_n}) &= E\left(e^{it\frac{1}{n}\sum_{k=1}^n X_k}\right) = E\left(\prod_{k=1}^n e^{itX_k/n}\right) = \prod_{k=1}^n E\left(e^{itX_k/n}\right) = \prod_{k=1}^n \varphi_{X_k}(t/n) \\ &= \prod_{k=1}^n \varphi_{X_1}(t/n) = (\varphi_{X_1}(t/n))^n. \end{aligned}$$

Let us recall Landau's "small o" notation for the relation between two functions. We say,  $f(x)$  is **small o** of  $g(x)$  if  $f$  is dominated by  $g$  as  $x \rightarrow \infty$ , i.e.,  $\frac{|f(x)|}{|g(x)|} \rightarrow 0$  as  $x \rightarrow \infty$ . More formally, for every  $\epsilon > 0$ , there exists an  $x_\epsilon$  such that for all  $x > x_\epsilon$   $|f(x)| < \epsilon|g(x)|$ . For example,  $\log(x)$  is  $o(x)$ ,  $x^2$  is  $o(x^3)$  and  $x^m$  is  $o(x^{m+1})$  for  $m \geq 1$ .

Third, we can expand any CF whose expectation exists as a Taylor series with a remainder term that is  $o(t)$  as follows:

$$\varphi_X(t) = 1 + itE(X) + o(t).$$

Hence,

$$\varphi_{X_1}(t/n) = 1 + i\frac{t}{n}E(X_1) + o\left(\frac{t}{n}\right)$$

and

$$E\left(e^{it\bar{X}_n}\right) = \left(1 + i\frac{t}{n}E(X_1) + o\left(\frac{t}{n}\right)\right)^n \rightarrow e^{itE(X_1)} \text{ as } n \rightarrow \infty.$$

---

<sup>7</sup>[https://en.wikipedia.org/wiki/L%C3%A9vy%27s\\_continuity\\_theorem](https://en.wikipedia.org/wiki/L%C3%A9vy%27s_continuity_theorem)

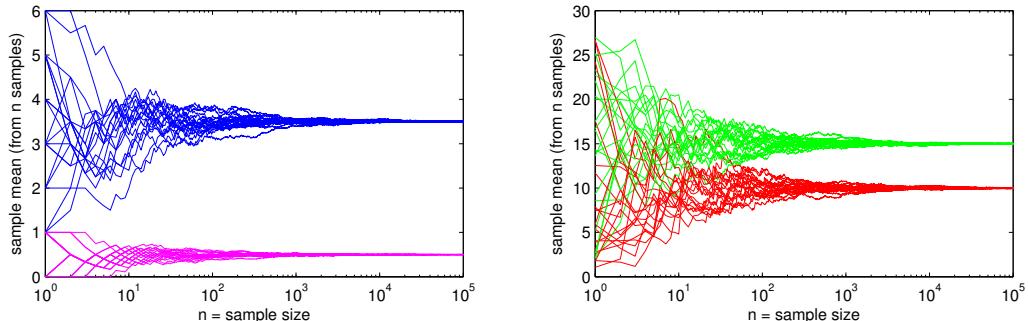
For the last limit we have used  $\left(1 + \frac{x}{n}\right)^n \rightarrow e^x$  as  $n \rightarrow \infty$ .

Finally, we have shown that  $E(e^{it\bar{X}_n})$ , the CF of the  $n$ -sample mean RV  $\bar{X}_n$ , converges to  $E(e^{itE(X_1)}) = e^{itE(X_1)}$ , the CF of the Point Mass( $E(X_1)$ ) RV, as the sample size  $n$  tends to infinity.

### Heuristic Interpretation of LLN

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  $E(X_1)$  [without loss of generality], as  $n$  approaches infinity. See Figure 5.4 for examples of 20 replicates of the sample mean of IID sequences from four RVs. All the sample mean trajectories converge to the corresponding population mean.

Figure 5.4: Sample mean  $\bar{X}_n$  as a function of sample size  $n$  for 20 replications from independent realizations of a fair die (blue), fair coin (magenta), Uniform(0, 30) RV (green) and Exponential(0.1) RV (red) with population means  $(1+2+3+4+5+6)/6 = 21/6 = 3.5$ ,  $(0+1)/2 = 0.5$ ,  $(30-0)/2 = 15$  and  $1/0.1 = 10$ , respectively.



**Example 161 (Bernoulli WLLN and Galton's Quincunx)** We can appreciate the WLLN for  $\bar{X}_n = n^{-1}S_n = \sum_{i=1}^n X_i$ , where  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} \text{Bernoulli}(p)$  using the paths of balls dropped into Galton's Quincunx of Sec. 3.2.3.

### Cauchy whose expectations does not exist has no Law of Large Numbers

Recall that the mean of the Cauchy RV  $X$  does not exist since  $\int |x| dF(x) = \infty$  (3.55). We will investigate this in Labwork 162.

**Labwork 162 (Running mean of the Standard Cauchy RV)** Let us see what happens when we plot the running sample mean for an increasing sequence of IID samples from the Standard Cauchy RV  $X$  by implementing the following script file:

---

```
PlotStandardCauchyRunningMean.m
% script to plot the oscillating running mean of Std Cauchy samples
% relative to those for the Uniform(0,10) samples
rand('twister',25567); % initialize the fundamental sampler
for i=1:5
    N = 10^5; % maximum sample size
```

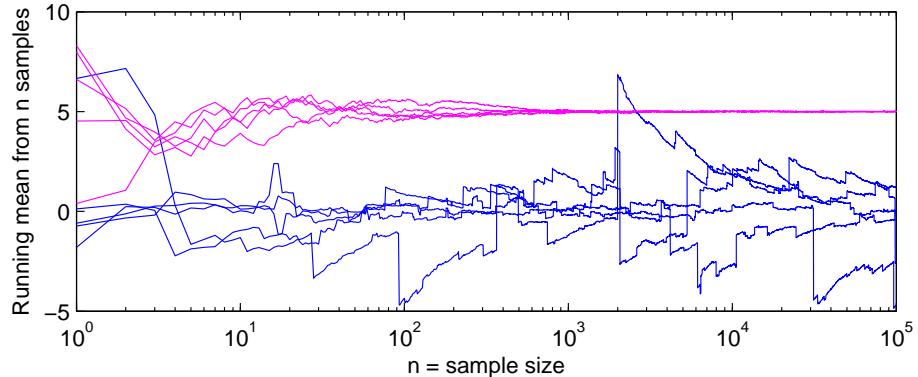
```

u=rand(1,N); % draw N IID samples from Uniform(0,1)
x=tan(pi * u); % draw N IID samples from Standard cauchy RV using inverse CDF
n=1:N; % make a vector n of current sample size [1 2 3 ... N-1 N]
CSx=cumsum(x); % CSx is the cumulative sum of the array x (type 'help cumsum')
% Runnign Means <- vector division of cumulative sum of samples by n
RunningMeanStdCauchy = CSx ./ n; % Running Mean for Standard Cauchy samples
RunningMeanUnif010 = cumsum(u*10.0) ./ n; % Running Mean for Uniform(0,10) samples
semilogx(n, RunningMeanStdCauchy) %
hold on;
semilogx(n, RunningMeanUnif010, 'm')
end
xlabel('n = sample size');
ylabel('Running mean from n samples')

```

---

Figure 5.5: Unending fluctuations of the running means based on  $n$  IID samples from the Standard Cauchy RV  $X$  in each of five replicate simulations (blue lines). The running means, based on  $n$  IID samples from the Uniform(0, 10) RV, for each of five replicate simulations (magenta lines).



The resulting plot is shown in Figure 5.5. Notice that the running means or the sample mean of  $n$  samples as a function of  $n$ , for each of the five replicate simulations, never settles down to a particular value. This is because of the “thick tails” of the density function for this RV which produces extreme observations. Compare them with the running means, based on  $n$  IID samples from the Uniform(0, 10) RV, for each of five replicate simulations (magenta lines). The latter sample means have settled down stably to the mean value of 5 after about 700 samples.

### 5.2.1 Application: Point Estimation of $E(X_1)$

LLN gives us a method to obtain a **point estimator** that gives “the single best guess” for the possibly unknown population mean  $E(X_1)$  based on  $\bar{X}_n$ , the sample mean, of a simple random sequence (SRS) or independent and identically distributed (IID) sequence of  $n$  RVs  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} X_1$ .

**Example 163** Let  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} X_1$ , where  $X_1$  is an  $\text{Exponential}(\lambda^*)$  RV, i.e., let

$$X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} \text{Exponential}(\lambda^*) .$$

Typically, we do not know the “true” parameter  $\lambda^* \in \mathbf{A} = (0, \infty)$  or the population mean  $E(X_1) = 1/\lambda^*$ . But by LLN, we know that

$$\bar{X}_n \rightsquigarrow \text{Point Mass}(E(X_1)) ,$$

and therefore, we can use the sample mean  $\bar{X}_n$  as a point estimator of  $E(X_1) = 1/\lambda^*$ .

Now, suppose you model seven waiting times in nearest minutes between Orbiter buses at Balgay street as follows:

$$X_1, X_2, \dots, X_7 \stackrel{\text{IID}}{\sim} \text{Exponential}(\lambda^*) ,$$

and have the following realization as your observed data:

$$(x_1, x_2, \dots, x_7) = (2, 12, 8, 9, 14, 15, 11) .$$

Then you can use the observed sample mean  $\bar{x}_7 = (2 + 12 + 8 + 9 + 14 + 15 + 11)/7 = 71/7 \cong 10.14$  as a **point estimate** of the population mean  $E(X_1) = 1/\lambda^*$ . By the rearrangement  $\lambda^* = 1/E(X_1)$ , we can also obtain a point estimate of the “true” parameter  $\lambda^*$  from  $1/\bar{x}_7 = 7/71 \cong 0.0986$ .

**Remark 70 (Point estimates are realizations of the Point Estimator)** We say the statistic  $\bar{X}_n$ , which is a random variable that depends on the data  $\vec{X}(X_1, X_2, \dots, X_n)$ , is a **point estimator** of  $E(X_1)$ . But once we have a realization of the data  $\vec{X}$ , i.e., our observed data vector  $(x_1, x_2, \dots, x_n)$  and its corresponding realization as observed sample mean  $\bar{x}_n$ , we say  $\bar{x}_n$  is a **point estimate** of  $E(X_1)$ . In other words, the point estimate  $\bar{x}_n$  is a realization of the the random variable  $\bar{X}_n$  called the point estimator of  $E(X_1)$ . Therefore, when we observe a new data vector  $(x'_1, x'_2, \dots, x'_n)$  that is different from our first data vector  $(x_1, x_2, \dots, x_n)$ , our point estimator of  $E(X_1)$  is still  $\bar{X}_n$  but the point estimate  $n^{-1} \sum_{i=1}^n x'_i$  may be different from the first point estimate  $n^{-1} \sum_{i=1}^n x_i$ . The sample means from  $n$  samples for 20 replications (repeats of the experiment) are typically distinct especially for small  $n$  as shown in Figure 5.4.

**Example 164** Let  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} X_1$ , where  $X_1$  is an  $\text{Bernoulli}(\theta^*)$  RV, i.e., let

$$X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta^*) .$$

Typically, we do not know the “true” parameter  $\theta^* \in \Theta = [0, 1]$ , which is the same as the population mean  $E(X_1) = \theta^*$ . But by LLN, we know that

$$\bar{X}_n \rightsquigarrow \text{Point Mass}(E(X_1)) ,$$

and therefore, we can use the sample mean  $\bar{X}_n$  as a point estimator of  $E(X_1) = \theta^*$ .

Now, suppose you model seven coin tosses (encoding **Heads** as 1 with probability  $\theta^*$  and **Tails** as 0 with probability  $1 - \theta^*$ ) as follows:

$$X_1, X_2, \dots, X_7 \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta^*) ,$$

and have the following realization as your observed data:

$$(x_1, x_2, \dots, x_7) = (0, 1, 1, 0, 0, 1, 0) .$$

Then you can use the observed sample mean  $\bar{x}_7 = (0 + 1 + 1 + 0 + 0 + 1 + 0)/7 = 3/7 \cong 0.4286$  as a **point estimate** of the population mean  $E(X_1) = \theta^*$ . Thus, our “single best guess” for  $E(X_1)$  which is the same as the probability of **Heads** is  $\bar{x}_7 = 3/7$ .

Of course, if we tossed the same coin in the same IID manner another seven times or if we observed another seven waiting times of orbiter buses at a different bus-stop or on a different day we may get a different point estimate for  $E(X_1)$ . See the intersection of the twenty magenta sample mean

trajectories for simulated tosses of a fair coin from IID Bernoulli( $\theta^* = 1/2$ ) RVs and the twenty red sample mean trajectories for simulated waiting times from IID Exponential( $\lambda^* = 1/10$ ) RVs in Figure 5.4 with  $n = 7$ . Clearly, the point estimates for such a small sample size are fluctuating wildly! However, the fluctuations in the point estimates settles down for larger sample sizes.

The next natural question is how large should the sample size be in order to have a small interval of width, say  $2\epsilon$ , “contain”  $E(X_1)$ , the quantity of interest, with a high probability, say  $1 - \alpha$ ? If we can answer this then we can make probability statements like the following:

$$P(\text{error} < \text{tolerance}) = P(|\bar{X}_n - E(X_1)| < \epsilon) = P(-\epsilon < \bar{X}_n - E(X_1) < \epsilon) = 1 - \alpha .$$

In order to ensure the  $\text{error} = |\bar{X}_n - E(X_1)|$  in our estimate of  $E(X_1)$  is within a required  $\text{tolerance} = \epsilon$  we need to know the full distribution of  $\bar{X}_n - E(X_1)$  itself. The Central Limit Theorem (CLT) helps us here.

### 5.3 Central Limit Theorem

What if we scale the sum of  $X_i$ ’s by  $\sqrt{n}$  instead of  $n$ ?

**Exercise 5.1 (What if we scale by  $\sqrt{n}$ )** After reading Sec. 5.1 up to now, think carefully about what you need to be able to show that  $Z_n := 1/\sqrt{n} \sum_{i=1}^n X_i$  converges in distribution to the Normal(0, 1/3) RV, where  $X_i \stackrel{\text{IID}}{\sim} \text{Uniform}(-1, 1)$ . Hint: Characteristic functions

**Proposition 71 (Central Limit Theorem (CLT))** If we are given a sequence of independently and identically distributed (IID) RVs,  $X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1$  and if  $E(X) < \infty$  and  $V(X_1) < \infty$ , then the sample mean  $\bar{X}_n$  converges in distribution to the Normal RV with mean given by any one of the IID RVs, say  $E(X_1)$  by convention, and variance given by  $\frac{1}{n}$  times the variance of any one of the IID RVs, say  $V(X_1)$  by convention. More formally, we write:

$$\begin{aligned} \text{If } X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1 \text{ and if } E(X_1) < \infty, V(X_1) < \infty \\ \text{then } \bar{X}_n \rightsquigarrow \text{Normal}\left(E(X_1), \frac{V(X_1)}{n}\right) \text{ as } n \rightarrow \infty , \end{aligned} \quad (5.7)$$

or equivalently after standardization:

$$\begin{aligned} \text{If } X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1 \text{ and if } E(X_1) < \infty, V(X_1) < \infty \\ \text{then } \frac{\bar{X}_n - E(X_1)}{\sqrt{V(X_1)/n}} \rightsquigarrow Z \sim \text{Normal}(0, 1) \text{ as } n \rightarrow \infty . \end{aligned} \quad (5.8)$$

**Proof:** Our proof is based on the convergence of characteristic functions (CFs). We will prove the standardized form of the CLT in Equation (5.8) by showing that the CF of

$$U_n := \frac{\bar{X}_n - E(X_1)}{\sqrt{V(X_1)/n}}$$

converges to the CF of  $Z$ , the Normal(0, 1) RV. First, note from Equation (3.72) that the CF of  $Z \sim \text{Normal}(0, 1)$  is:

$$\varphi_Z(t) = E(e^{itZ}) = e^{-t^2/2} .$$

Second,

$$U_n := \frac{\bar{X}_n - E(X_1)}{\sqrt{V(X_1)/n}} = \frac{\sum_{k=1}^n X_k - nE(X_1)}{\sqrt{nV(X_1)}} = \frac{1}{\sqrt{n}} \sum_{k=1}^n \left( \frac{X_k - E(X_1)}{\sqrt{V(X_1)}} \right) .$$

Therefore, the CF of  $U_n$  is

$$\begin{aligned} \varphi_{U_n}(t) &= E(\exp(itU_n)) = E\left(\exp\left(i\frac{t}{\sqrt{n}} \sum_{k=1}^n \frac{X_k - E(X_1)}{\sqrt{V(X_1)}}\right)\right) = \prod_{k=1}^n E\left(\exp\left(i\frac{t}{\sqrt{n}} \frac{X_k - E(X_1)}{\sqrt{V(X_1)}}\right)\right) \\ &= \left(E\left(\exp\left(i\frac{t}{\sqrt{n}} \frac{X_1 - E(X_1)}{\sqrt{V(X_1)}}\right)\right)\right)^n . \end{aligned}$$

Now, if we let

$$Y = \frac{X_1 - E(X_1)}{\sqrt{V(X_1)}}$$

then

$$E(Y) = 0, \quad E(Y^2) = 1, \text{ and } V(Y) = 1 .$$

So, the CF of  $U_n$  is

$$\varphi_{U_n}(t) = \left(\varphi_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n ,$$

and since we can Taylor expand  $\varphi_Y(t)$  as follows:

$$\varphi_Y(t) = 1 + itE(Y) + i^2 \frac{t^2}{2} E(Y^2) + o(t^2) ,$$

which implies

$$\varphi_Y\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{it}{\sqrt{n}} E(Y) + \frac{i^2 t^2}{2n} E(Y^2) + o\left(\frac{t^2}{n}\right) ,$$

we finally get

$$\varphi_{U_n}(t) = \left(\varphi_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n = \left(1 + \frac{it}{\sqrt{n}} \times 0 + \frac{i^2 t^2}{2n} \times 1 + o\left(\frac{t^2}{n}\right)\right)^n = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \rightarrow e^{-t^2/2} = \varphi_Z(t) .$$

For the last limit we have used  $(1 + \frac{x}{n})^n \rightarrow e^x$  as  $n \rightarrow \infty$ . Thus, we have proved Equation (5.8) which is equivalent to Equation (5.7) by a standardization argument that if  $W \sim \text{Normal}(\mu, \sigma^2)$  then  $Z = \frac{W-\mu}{\sigma} \sim \text{Normal}(0, 1)$  through the linear transformation  $W = \sigma Z + \mu$  of Example 67.

### 5.3.1 Application: Tolerating Errors in our estimate of $E(X_1)$

Recall that we wanted to ensure the  $\text{error} = |\bar{X}_n - E(X_1)|$  in our estimate of  $E(X_1)$  is within a required  $\text{tolerance} = \epsilon$  and make the following probability statement:

$$P(\text{error} < \text{tolerance}) = P(|\bar{X}_n - E(X_1)| < \epsilon) = P(-\epsilon < \bar{X}_n - E(X_1) < \epsilon) = 1 - \alpha .$$

To be able to do this we needed to know the full distribution of  $\bar{X}_n - E(X_1)$  itself.

Due to the Central Limit Theorem (CLT) we now know that (assuming  $n$  is large)

$$\begin{aligned} P(-\epsilon < \bar{X}_n - E(X_1) < \epsilon) &\approx P\left(-\frac{\epsilon}{\sqrt{V(X_1)/n}} < \frac{\bar{X}_n - E(X_1)}{\sqrt{V(X_1)/n}} < \frac{\epsilon}{\sqrt{V(X_1)/n}}\right) \\ &= P\left(-\frac{\epsilon}{\sqrt{V(X_1)/n}} < Z < \frac{\epsilon}{\sqrt{V(X_1)/n}}\right) , \end{aligned}$$

where  $Z \sim \text{Normal}(0, 1)$ .

**Example 165** Suppose an IID sequence of observations  $(x_1, x_2, \dots, x_{80})$  was drawn from a distribution with variance  $V(X_1) = 4$ . What is the probability that the error in  $\bar{x}_n$  used to estimate  $E(X_1)$  is less than 0.1?

By CLT,

$$P(\text{error} < 0.1) \approx P\left(-\frac{0.1}{\sqrt{4/80}} < Z < \frac{0.1}{\sqrt{4/80}}\right) = P(-0.447 < Z < 0.447) = 0.345 .$$

Suppose you want the **error** to be less than **tolerance** =  $\epsilon$  with a certain probability  $1 - \alpha$ . Then we can use CLT to do such **sample size calculations**. Recall the DF  $\Phi(z) = P(Z < z)$  is tabulated in the standard normal table and now we want

$$P\left(-\frac{\epsilon}{\sqrt{V(X_1)/n}} < Z < \frac{\epsilon}{\sqrt{V(X_1)/n}}\right) = 1 - \alpha .$$

We know,

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha ,$$

make the picture here of  $f_Z(z) = \Phi'(z)$  to recall what  $z_{\alpha/2}$ ,  $z_{-\alpha/2}$ , and the various areas below  $f_Z(\cdot)$  in terms of  $\Phi(\cdot)$  from the table really mean... (See Example 59).

where,  $\Phi(z_{\alpha/2}) = 1 - \alpha/2$  and  $\Phi(z_{-\alpha/2}) = 1 - \Phi(z_{\alpha/2}) = \alpha/2$ . So, we set

$$\frac{\epsilon}{\sqrt{V(X_1)/n}} = z_{\alpha/2}$$

and rearrange to get

$$n = \left( \frac{\sqrt{V(X_1)} z_{\alpha/2}}{\epsilon} \right)^2 \quad (5.9)$$

for the needed sample size that will ensure that our **error** is less than our **tolerance** =  $\epsilon$  with probability  $1 - \alpha$ . Of course, if  $n$  given by Equation (5.9) is not a natural number then we naturally round up to make it one!

A useful  $z_{\alpha/2}$  value to remember: If  $\alpha = 0.05$  when the probability of interest  $1 - \alpha = 0.95$  then  $z_{\alpha/2} = z_{0.025} = 1.96$ .

**Example 166** How large a sample size is needed to make the **error** in our estimate of the population mean  $E(X_1)$  to be less than 0.1 with probability  $1 - \alpha = 0.95$  if we are observing IID samples from a distribution with a population variance  $V(X_1)$  of 4?

Using Equation (5.9) we see that the needed sample size is

$$n = \left( \frac{\sqrt{4} \times 1.96}{0.1} \right)^2 \approx 1537$$

Thus, it pays to check the sample size needed in advance of experimentation, provided you already know the population variance of the distribution whose population mean you are interested in estimating within a given tolerance and with a high probability.

### 5.3.2 Application: Set Estimation of $E(X_1)$

A useful byproduct of the CLT is the  **$(1 - \alpha)$  confidence interval**, a random interval (or bivariate  $\vec{R}\vec{V}$ ) that contains  $E(X_1)$ , the quantity of interest, with probability  $1 - \alpha$ :

$$(\bar{X}_n \pm z_{\alpha/2} \sqrt{V(X_1)/n}) := (\bar{X}_n - z_{\alpha/2} \sqrt{V(X_1)/n}, \bar{X}_n + z_{\alpha/2} \sqrt{V(X_1)/n}) . \quad (5.10)$$

We can easily see how Equation (5.10) is derived from CLT as follows:

$$\begin{aligned} P(-z_{\alpha/2} < Z < z_{\alpha/2}) &= 1 - \alpha \\ P\left(-z_{\alpha/2} < \frac{\bar{X}_n - E(X_1)}{\sqrt{V(X_1)/n}} < z_{\alpha/2}\right) &= 1 - \alpha \\ P\left(-\bar{X}_n - z_{\alpha/2} \sqrt{V(X_1)/n} < -E(X_1) < -\bar{X}_n + z_{\alpha/2} \sqrt{V(X_1)/n}\right) &= 1 - \alpha \\ P\left(\bar{X}_n + z_{\alpha/2} \sqrt{V(X_1)/n} > E(X_1) > \bar{X}_n - z_{\alpha/2} \sqrt{V(X_1)/n}\right) &= 1 - \alpha \\ P\left(\bar{X}_n - z_{\alpha/2} \sqrt{V(X_1)/n} < E(X_1) < \bar{X}_n + z_{\alpha/2} \sqrt{V(X_1)/n}\right) &= 1 - \alpha \\ P(E(X_1) \in (\bar{X}_n - z_{\alpha/2} \sqrt{V(X_1)/n}, \bar{X}_n + z_{\alpha/2} \sqrt{V(X_1)/n})) &= 1 - \alpha . \end{aligned}$$

**Remark 72 (Heuristic interpretation of the  $(1 - \alpha)$  confidence interval)** If we repeatedly produced samples of size  $n$  to contain  $E(X_1)$  within a  $(\bar{X}_n \pm z_{\alpha/2} \sqrt{V(X_1)/n})$ , say 100 times, then on average,  $(1 - \alpha) \times 100$  repetitions will actually contain  $E(X_1)$  within the random interval and  $\alpha \times 100$  repetitions will fail to contain  $E(X_1)$ .

So far, we have assumed we know the population variance  $V(X_1)$  in an IID experiment with  $n$  samples and tried to estimate the population mean  $E(X_1)$ . But in general, we will not know  $V(X_1)$ . We can still get a point estimate of  $E(X_1)$  from the sample mean due to LLN but we won't be able to get a confidence interval for  $E(X_1)$ . Fortunately, a more elaborate form of the CLT tells us that even when we substitute the sample variance  $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$  for the population variance  $V(X_1)$  the following  $1 - \alpha$  confidence interval for  $E(X_1)$  works!

$$(\bar{X}_n \pm z_{\alpha/2} S_n / \sqrt{n}) := (\bar{X}_n - z_{\alpha/2} S_n / \sqrt{n}, \bar{X}_n + z_{\alpha/2} S_n / \sqrt{n}) , \quad (5.11)$$

where,  $S_n = \sqrt{S_n^2}$  is the sample standard deviation.

Let's return to our two examples again.

**Example 167** We model the waiting times between Orbiter buses with unknown  $E(X_1) = 1/\lambda^*$  as

$$X_1, X_2, \dots, X_n \stackrel{IID}{\sim} \text{Exponential}(\lambda^*)$$

and observed the following data, sample mean, sample variance and sample standard deviation:

$$(x_1, x_2, \dots, x_7) = (2, 12, 8, 9, 14, 15, 11), \bar{x}_7 = 10.143, s_7^2 = 19.143, s_7 = 4.375 ,$$

respectively. Our point estimate and  $1 - \alpha = 95\%$  confidence interval for  $E(X_1)$  are:

$$\bar{x}_7 = 10.143 \quad \text{and} \quad (\bar{x}_7 \pm z_{\alpha/2} s_7 / \sqrt{7}) = (10.143 \pm 1.96 \times 4.375 / \sqrt{7}) = (6.9016, 13.3841) ,$$

respectively. So with 95% probability the true population mean  $E(X_1) = 1/\lambda^*$  is contained in  $(6.9016, 13.3841)$  and since the mean waiting time of 10 minutes promised by the Orbiter bus company is also within  $(6.9016, 13.3841)$  we can be fairly certain that the company sticks to its promise.

**Example 168** We model the tosses of a coin with unknown  $E(X_1) = \theta^*$  as

$$X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta^*)$$

and observed the following data, sample mean, sample variance and sample standard deviation:

$$(x_1, x_2, \dots, x_7) = (0, 1, 1, 0, 0, 1, 0), \bar{x}_7 = 0.4286, s_7^2 = 0.2857, s_7 = 0.5345 ,$$

respectively. Our point estimate and  $1 - \alpha = 95\%$  confidence interval for  $E(X_1)$  are:

$$\bar{x}_7 = 0.4286 \quad \text{and} \quad (\bar{x}_7 \pm z_{\alpha/2}s_7/\sqrt{7}) = (0.4286 \pm 1.96 \times 0.5345/\sqrt{7}) = (0.0326, 0.8246) ,$$

respectively. So with 95% probability the true population mean  $E(X_1) = \theta^*$  is contained in  $(0.0326, 0.8246)$  and since  $1/2$  is contained in this interval of width 0.792 we cannot rule out that the flipped coin is not fair with  $\theta^* = 1/2$ .

**Remark 73** The normal-based confidence interval for  $\theta^*$  (as well as  $\lambda^*$  in the previous example) may not be a valid approximation here with just  $n = 7$  samples. After all, the CLT only tells us that the point estimator  $\hat{\Theta}_n$  can be approximated by a normal distribution for large sample sizes. When the sample size  $n$  was increased from 7 to 100 by tossing the same coin another 93 times, a total of 57 trials landed as Heads. Thus the point estimate and confidence interval for  $E(X_1) = \theta^*$  based on the sample mean and sample standard deviations are:

$$\hat{\theta}_{100} = \frac{57}{100} = 0.57 \quad \text{and} \quad (0.57 \pm 1.96 \times 0.4975/\sqrt{100}) = (0.4725, 0.6675) .$$

Thus our confidence interval shrank considerably from a width of 0.792 to 0.195 after an additional 93 Bernoulli trials. Thus, we can make the width of the confidence interval as small as we want by making the number of observations or sample size  $n$  as large as we can.

## 5.4 Exercises in Limit Laws of Statistics

**Ex. 5.2** — Suppose you plan to obtain a simple random sequence (SRS) — also known as independent and identically distributed (IID) sequence — of  $n$  measurements from an instrument. This instrument has been calibrated so that the distribution of measurements made with it have population variance of  $1/4$ . Your boss wants you to make a point estimate of the unknown population mean from a SRS of sample size  $n$ . He also insists that the tolerance for error has to be  $1/10$  and the probability of meeting this tolerance should be just above 95%. Use CLT to find how large should  $n$  be to meet the specifications of your boss.

**Ex. 5.3** — Suppose the collection of RVs  $X_1, X_2, \dots, X_n$  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  $\bar{X}_{125}$  which is the average error per program out of these 125 programs. Since  $E(X_i) = \lambda = 5$  and  $V(X_i) = \lambda = 5$ , we want to know how often our sample mean  $\bar{X}_{125}$  differs from the expectation of 5 errors per program. Using the CLT find the  $P(\bar{X}_{125} < 5.5)$ .

**Ex. 5.4** — What is the distribution of  $\sum_{i=1}^n X_i/n$  as  $n \rightarrow \infty$  when  $X_i \stackrel{\text{IID}}{\sim} \text{Uniform}(-10, 10)$ ?

**Ex. 5.5** — What is the distribution of  $\sum_{i=1}^n X_i / \sqrt{V(X_i)n}$  as  $n \rightarrow \infty$  when  $X_i \stackrel{\text{IID}}{\sim} \text{Uniform}(-10, 10)$ ?

# Chapter 6

## Finite Markov Chains

**NOTE:** No materials from this Chapter will be in Probability Theory I exam!

*This topic is only introduced briefly in Probability Theory I to give concrete instances of dependent sequence of random variables. We will revisit these ideas in the sequel. You only need to understand the ideas behind:*

- Example 169 and
- Example 170.

as done in lectures.

### 6.1 Stochastic Processes

**Definition 74 (Stochastic Process)** A collection of RVs

$$(X_\alpha)_{\alpha \in N} := (X_\alpha : \alpha \in \mathbb{A})$$

is called a **stochastic process**. Thus, for every  $\alpha \in \mathbb{A}$ , the index set of the stochastic process,  $X_\alpha$  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.

Of course the above process is quite general and can allow for arbitrary dependence among the RVs. Generally, we cannot produce useful models without making simplifying assumptions. The absolutely simplest but extremely useful assumption is that of the **Independent and Identically Distributed or IID Process** or merely **IID Sequence of RVs** when the index set is a subset of  $\mathbb{N}$ .

This is exactly the sequence of RVs associated with our product experiment  $\mathcal{E}^{\otimes \infty} := (\Omega, \mathcal{F}_\mathcal{X}, P_\theta)^{\otimes \infty}$ .

**Definition 75 (Independent and Identically Distributed (IID) Process)** 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 43, 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 76 (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 43.

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

As an example of such a class consider the sequence of RVs that are independent but non-identically distributed with each  $X_i \sim \text{Bernoulli}(\theta_i)$ .

When a stochastic process  $(X_\alpha)_{\alpha \in \mathbb{A}}$  is not independent it is said to be dependent. So far we have mostly concerned ourselves with independent processes. In this chapter we introduce finite Markov chains and their simulation methods. Finite Markov chains are among the simplest stochastic processes with a 'first-order' dependence called Markov dependence.

## 6.2 Introduction

A finite Markov chain is a stochastic process that moves among elements in a finite set  $\mathbb{X}$  as follows: when at  $x \in \mathbb{X}$  the next position is chosen at random according to a fixed probability distribution  $P(\cdot|x)$ . We define such a process more formally below.

**Definition 77 (Finite Markov Chain)** A stochastic sequence,

$$(X_n)_{n \in \mathbb{Z}_+} := (X_0, X_1, \dots),$$

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  $P(H_{t-1} \cap \{X_t = x\}) > 0$ , the following **Markov property** is satisfied:

$$P(X_{t+1} = y | H_{t-1} \cap \{X_t = x\}) = P(X_{t+1} = y | X_t = x) =: P(x, y). \quad (6.1)$$

The Markov property means that the conditional probability of going to state  $y$  at time  $t+1$  from state  $x$  at current time  $t$  is always given by the  $(x, y)$ -th entry  $P(x, y)$  of the transition matrix  $P$ , no matter what sequence of states  $(x_0, x_1, \dots, x_{t-1})$  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.,

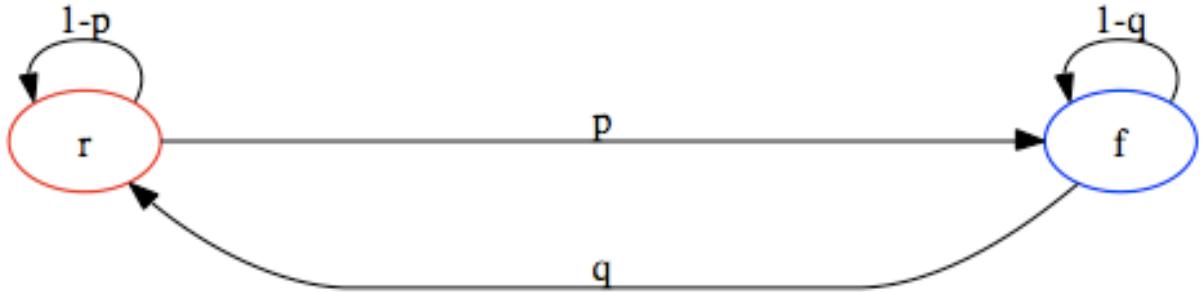
$$P(x, y) \geq 0 \quad \text{for all } (x, y) \in \mathbb{X}^2 \quad \text{and} \quad \sum_{y \in \mathbb{X}} P(x, y) = 1 \quad \text{for all } x \in \mathbb{X}. \quad (6.2)$$

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 169 (Flippant Freddy)** Freddy the flippant frog lives in an enchanted pond with only two lily pads, *rollopia* and *flipopia*. A wizard gave a die and a silver coin to help flippant Freddy decide where to jump next. Freddy left the die on rollopia and the coin on flipopia. When Freddy got restless in rollopia he would roll the die and if the die landed odd he would leave the die behind and jump to flipopia, otherwise he would stay put. When Freddy got restless in flipopia he would flip the coin and if it landed Heads he would leave the coin behind and jump to rollopia, otherwise he would stay put.

Let the state space  $\mathbb{X} = \{r, f\}$ , and let  $(X_0, X_1, \dots)$  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 6.1: 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:

$$P = \begin{matrix} r & f \\ \begin{matrix} r \\ f \end{matrix} & \begin{pmatrix} P(r, r) & P(r, f) \\ P(f, r) & P(f, f) \end{pmatrix} \end{matrix} = \begin{matrix} r & f \\ \begin{matrix} r \\ f \end{matrix} & \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \end{matrix}. \quad (6.3)$$

Suppose we first see Freddy in rollopia, i.e.,  $X_0 = r$ . When he gets restless for the first time we know from the first row of  $P$  that he will leave to flipopia with probability  $p$  and stay with probability  $1 - p$ , i.e.,

$$P(X_1 = f | X_0 = r) = p, \quad P(X_1 = r | X_0 = r) = 1 - p. \quad (6.4)$$

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,

$$\begin{aligned}
 P(X_2 = f | X_0 = r) &= P(X_2 = f, X_1 = f | X_0 = r) + P(X_2 = f, X_1 = r | X_0 = r) \\
 &= \frac{P(X_2 = f, X_1 = f, X_0 = r)}{P(X_0 = r)} + \frac{P(X_2 = f, X_1 = r, X_0 = r)}{P(X_0 = r)} \\
 &= P(X_2 = f | X_1 = f, X_0 = r) \frac{P(X_1 = f, X_0 = r)}{P(X_0 = r)} \\
 &\quad + P(X_2 = f | X_1 = r, X_0 = r) \frac{P(X_1 = r, X_0 = r)}{P(X_0 = r)} \\
 &= P(X_2 = f | X_1 = f, X_0 = r) P(X_1 = f | X_0 = r) \\
 &\quad + P(X_2 = f | X_1 = r, X_0 = r) P(X_1 = r | X_0 = r) \\
 &= P(X_2 = f | X_1 = f) P(X_1 = f | X_0 = r) \\
 &\quad + P(X_2 = f | X_1 = r) P(X_1 = r | X_0 = r) \\
 &= P(f, f) P(r, f) + P(r, f) P(r, r) \\
 &= (1 - q)p + p(1 - p)
 \end{aligned} \tag{6.5}$$

Similarly,

$$P(X_2 = r | X_0 = r) = P(f, r) P(r, f) + P(r, r) P(r, r) = qp + (1 - p)(1 - p) \tag{6.6}$$

Instead of elaborate computations of the probabilities of being in a given state after Freddy's  $t$ -th restless moment, we can store the state probabilities at time  $t$  in a row vector:

$$\mu_t := (P(X_t = r | X_0 = r), P(X_t = f | X_0 = r)) ,$$

Now, we can conveniently represent Freddy starting in rollovia by the **initial distribution**  $\mu_0 = (1, 0)$  and obtain the 1-step **state probability vector** in (6.4) from  $\mu_1 = \mu_0 P$  and the 2-step state probabilities in (6.5) and (6.6) by  $\mu_2 = \mu_1 P = \mu_0 P P = \mu_0 P^2$ . In general, multiplying  $\mu_t$ , 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  $\mu_0$ ,

$$\mu_t = \mu_0 P^t \quad \text{for all } t \geq 0 .$$

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 rollovia
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

```

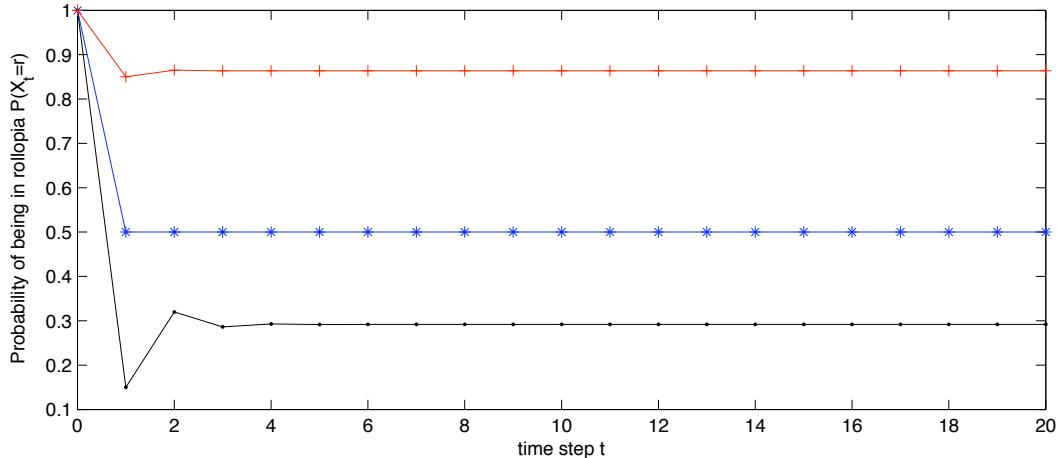
Now, let us compute and look at the probability of being in rollopia after having started there for three values of  $p$  and  $q$  according to the following script:

---

```
----- 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 6.2: The probability of being back in rollopia in  $t$  time steps after having started there under transition matrix  $P$  with (i)  $p = q = 0.5$  (blue line with asterisks), (ii)  $p = 0.85, q = 0.35$  (black line with dots) and (iii)  $p = 0.15, q = 0.95$  (red line with pluses).



It is evident from Figure 6.2 that as  $t \rightarrow \infty$ ,  $\mu_t$  approaches a distribution, say  $\pi$ , 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 6.2 we see that  $P(X_t = r)$  approaches  $\pi(r) = \frac{q}{p+q}$  for the three cases of  $p$  and  $q$ :

$$\begin{aligned} \text{(i)} \quad & p = 0.50, q = 0.50, & P(X_t = r) \rightarrow \pi(r) = \frac{q}{p+q} = \frac{0.50}{0.50 + 0.50} = 0.5000, \\ \text{(ii)} \quad & p = 0.85, q = 0.35, & P(X_t = r) \rightarrow \pi(r) = \frac{q}{p+q} = \frac{0.35}{0.85 + 0.35} = 0.2917, \\ \text{(iii)} \quad & p = 0.15, q = 0.95, & P(X_t = r) \rightarrow \pi(r) = \frac{q}{p+q} = \frac{0.95}{0.15 + 0.95} = 0.8636. \end{aligned}$$

Now let us generalise the lessons learned from Example 169.

**Proposition 78** 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  $\mu_0(s_i) = P(X_0 = s_i)$ , 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  $\mu_t(s_i) = P(X_t = s_i)$ , satisfies:

$$\mu_t = \mu_0 P^t. \quad (6.7)$$

**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

$$\begin{aligned} \mu_1(s_j) &= P(X_1 = s_j) = \sum_{i=1}^k P(X_1 = s_j, X_0 = s_i) \\ &= \sum_{i=1}^k P(X_1 = s_j | X_0 = s_i) P(X_0 = s_i) \\ &= \sum_{i=1}^k P(s_i, s_j) \mu_0(s_i) \\ &= (\mu_0 P)(s_j), \quad \text{the } j\text{-th entry of the row vector } (\mu_0 P). \end{aligned}$$

Hence,  $\mu_1 = \mu_0 P$ . Now, we will fix  $m$  and suppose that (6.7) holds for  $t = m$  and prove that (6.7) also holds for  $t = m + 1$ . For each  $j \in \{1, 2, \dots, k\}$ , we get

$$\begin{aligned} \mu_{m+1}(s_j) &= P(X_{m+1} = s_j) = \sum_{i=1}^k P(X_{m+1} = s_j, X_m = s_i) \\ &= \sum_{i=1}^k P(X_{m+1} = s_j | X_m = s_i) P(X_m = s_i) \\ &= \sum_{i=1}^k P(s_i, s_j) \mu_m(s_i) \\ &= (\mu_m P)(s_j), \quad \text{the } j\text{-th entry of the row vector } (\mu_m P). \end{aligned}$$

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  $\mu_0$  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  $P_\mu$  and  $E_\mu$  for probabilities and expectations given that the initial distribution is  $\mu$ , 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  $P_x$  for  $P_{\mathbf{1}_{\{x\}}}$  or  $P_{e_x}$  and  $E_x$  for  $E_{\mathbf{1}_{\{x\}}}$  or  $E_{e_x}$ . Thus, Proposition 78 along with our new notations means that:

$$P_x(X_t = y) = (e_x P^t)(y) = P^t(x, y) .$$

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.

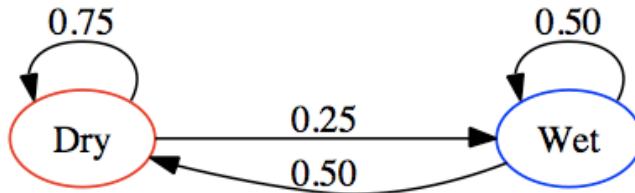
Let the function  $f(x) : \mathbb{X} \rightarrow \mathbb{R}$  be represented by the column vector  $f := (f(s_1), f(s_2), \dots, f(s_k)) \in \mathbb{R}^{k \times 1}$ . Then the  $x$ -th entry of  $P^t f$  is:

$$P^t f(x) = \sum_y P^t(x, y) f(y) = \sum_y f(y) P_x(X_t = y) = E_x(f(X_t)) .$$

This is the expected value of  $f$  under the distribution of states in  $t$  steps given that we start at state  $x$ . Thus multiplying a column vector  $f$  by  $P^t$  from the left takes you from a function on the state space to its expected value in  $t$  steps of the chain.

**Example 170 (Dry-Wet Christchurch Weather)** Consider a toy weather model for dry or wet days in Christchurch using a Markov chain with state space  $\{d, w\}$ . Let the transition diagram in Figure 6.3 give the transition matrix  $P$  for our dry-wet Markov chain. Using (6.7) we can find

Figure 6.3: Transition Diagram of Dry and Wet Days in Christchurch.



that the probability of being dry on the day after tomorrow is 0.625 given that it is wet today as follows:

```

>> P=[0.75 0.25; 0.5 0.5] % Transition Probability Matrix
P =
    0.7500    0.2500
    0.5000    0.5000
>> mu0=[0 1] % it is wet today gives the initial distribution
mu0 =
    0    1
>> mu0 * P^2 % the distribution in 2 days from today
ans =
    0.6250    0.3750
  
```

Suppose you sell \$100 of lemonade at a road-side stand on a hot day but only \$50 on a cold day. Then we can compute your expected sales tomorrow if today is dry as follows:

```
>> P=[0.75 0.25; 0.5 0.5] % Transition Probability Matrix
P =
    0.7500    0.2500
    0.5000    0.5000
>> f = [100; 50] % sales of lemonade in dollars on a dry and wet day
f =
    100
    50
>> P*f % expected sales tomorrow
ans =
    87.5000
    75.0000
>> mu0 = [1 0] % today is dry
mu0 =
    1     0
>> mu0*P*f % expected sales tomorrow if today is dry
ans =    87.5000
```

**Exercise 171 (Freddy discovers a gold coin)** Flippant Freddy of Example 169 found a gold coin at the bottom of the pond. Since this discovery he jumps around differently in the enchanted pond. He can be found now in one of three states: flipopia, 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 172** Let  $(X_t)_{t \in \mathbb{Z}_+}$  be a Markov chain with state space  $\{a, b, c\}$ , initial distribution  $\mu_0 = (1/3, 1/3, 1/3)$  and transition matrix

$$P = \begin{matrix} & \begin{matrix} a & b & c \end{matrix} \\ \begin{matrix} a \\ b \\ c \end{matrix} & \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \end{matrix}.$$

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 173** 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  $\mu_0$ . For a given  $m \in \mathbb{N}$ , let  $(Y_t)_{t \in \mathbb{Z}_+}$  be a stochastic sequence with  $Y_t = X_{mt}$ . Show that  $(Y_t)_{t \in \mathbb{Z}_+}$  is a Markov chain with transition matrix  $P^m$ . This establishes that Markov chains that are sampled at regular time steps are also Markov chains.

Until now our Markov chains have been **homogeneous** in time according to Definition 77, i.e., the transition matrix  $P$  does not change with time. We define inhomogeneous Markov chains that allow their transition matrices to possibly change with time. Such Markov chains are more realistic as models in some situations and more flexible as algorithms in the sequel.

**Definition 79 (Inhomogeneous finite Markov chain)** Let  $P_1, P_2, \dots$  be a sequence of  $k \times k$  stochastic matrices satisfying the conditions in Equation 6.2. 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  $P(H_{t-1} \cap \{X_t = x\}) > 0$ , the following **Markov property** is satisfied:

$$P(X_{t+1} = y | H_{t-1} \cap \{X_t = x\}) = P(X_{t+1} = y | X_t = x) =: P_{t+1}(x, y) . \quad (6.8)$$

**Proposition 80** For a finite inhomogeneous Markov chain  $(X_t)_{t \in \mathbb{Z}_+}$  with state space  $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$ , initial distribution

$$\mu_0 := (\mu_0(s_1), \mu_0(s_2), \dots, \mu_0(s_k)) ,$$

where  $\mu_0(s_i) = P(X_0 = s_i)$ , 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  $\mu_t(s_i) = P(X_t = s_i)$ , satisfies:

$$\mu_t = \mu_0 P_1 P_2 \cdots P_t . \quad (6.9)$$

**Proof:** Left as Exercise 174.

**Exercise 174** Prove Proposition 80 using induction as done for Proposition 78.

**Example 175 (a more sophisticated dry-wet chain)** Let us make a more sophisticated version of the dry-wet chain of Example 170 with state space  $\{d, w\}$ . In order to take some seasonality into account in our weather model for dry and wet days in Christchurch, let us have two transition matrices for hot and cold days:

$$P_{\text{hot}} = \begin{pmatrix} d & w \\ \begin{matrix} 0.95 & 0.05 \\ 0.75 & 0.25 \end{matrix} \end{pmatrix}, \quad P_{\text{cold}} = \begin{pmatrix} d & w \\ \begin{matrix} 0.65 & 0.35 \\ 0.45 & 0.55 \end{matrix} \end{pmatrix} .$$

We say that a day is hot if its maximum temperature is more than 20° Celsius, otherwise it is cold. We use the transition matrix for today to obtain the state probabilities for tomorrow. If today is dry and hot and tomorrow is supposed to be cold then what is the probability that the day after tomorrow will be wet? We can use (6.9) to obtain the answer as 0.36:

```
>> Phot = [0.95 0.05; 0.75 0.25] % Transition Probability Matrix for hot day
Phot =
    0.9500    0.0500
    0.7500    0.2500
>> Pcold = [0.65 0.35; 0.45 0.55] % Transition Probability Matrix for cold day
Pcold =
    0.6500    0.3500
    0.4500    0.5500
>> mu0 = [1 0] % today is dry
mu0 =      1      0
```

```

>> mu1 = mu0 * Phot % distribution for tomorrow since today is hot
mu1 =
    0.9500    0.0500
>> mu2 = mu1 * Pcold % distribution for day after tomorrow since tomorrow is supposed to be cold
mu2 =
    0.6400    0.3600
>> mu2 = mu0 * Phot * Pcold % we can also get the distribution for day after tomorrow directly
mu2 =
    0.6400    0.3600

```

**Exercise 176** For the Markov chain in Example 175 compute the probability that the day after tomorrow is wet if today is dry and hot but tomorrow is supposed to be cold.

### 6.3 Irreducibility and Aperiodicity

The utility of our mathematical constructions with Markov chains depends on a delicate balance between generality and specificity. We introduce two specific conditions called irreducibility and aperiodicity that make Markov chains more useful to model real-word phenomena.

**Definition 81 (Communication between states)** Let  $(X_t)_{t \in \mathbb{Z}_+}$  be a homogeneous Markov chain with transition matrix  $P$  on state space  $\mathbb{X} := \{s_1, s_2, \dots, s_k\}$ . 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:

$$P(X_{t+\eta(s_i, s_j)} = s_j | X_t = s_i) = P^{\eta(s_i, s_j)}(s_i, s_j) > 0 .$$

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 82 (Irreducible)** A homogeneous Markov chain  $(X_t)_{t \in \mathbb{Z}_+}$  with transition matrix  $P$  on state space  $\mathbb{X} := \{s_1, s_2, \dots, s_k\}$  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 6.3. 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.

**Exercise 177** Revisit all the Markov chains we have considered up to now and determine whether they are reducible or irreducible by checking that from each state there is a sequence of arrows leading to any other state in their transition graphs.

**Definition 83 (Return times and period)** Let  $\mathbb{T}(x) := \{t \in \mathbb{N} : P^t(x, x) > 0\}$  be the set of **possible return times** to the starting state  $x$ . The **period** of state  $x$  is defined to be  $\text{gcd}(\mathbb{T}(x))$ , the greatest common divisor of  $\mathbb{T}(x)$ . When the period of a state  $x$  is 1, i.e.,  $\text{gcd}(\mathbb{T}(x)) = 1$ , then  $x$  is said to be an **aperiodic state**.

**Proposition 84** If the Markov chain  $(X_t)_{t \in \mathbb{Z}_+}$  with transition matrix  $P$  on state space  $\mathbb{X}$  is irreducible then  $\text{gcd}(\mathbb{T}(x)) = \text{gcd}(\mathbb{T}(y))$  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  $\text{gcd}(\mathbb{T}(y))$  divides all elements in  $\mathbb{T}(x)$ . Thus,  $\text{gcd}(\mathbb{T}(y)) \leq \text{gcd}(\mathbb{T}(x))$ . By a similar argument we can also conclude that  $\text{gcd}(\mathbb{T}(x)) \leq \text{gcd}(\mathbb{T}(y))$ . Therefore  $\text{gcd}(\mathbb{T}(x)) = \text{gcd}(\mathbb{T}(y))$ .

**Definition 85 (Aperiodic)** A Markov chain  $(X_t)_{t \in \mathbb{Z}_+}$  with transition matrix  $P$  on state space  $\mathbb{X}$  is said to be aperiodic if all of its states are aperiodic, i.e.,  $\text{gcd}(\mathbb{T}(x)) = 1$  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  $\text{gcd}(\{1, 2, 3, \dots\}) = 1$ . However, the Markov chain model for a drunkard's walk around a block over the state space  $\{0, 1, 2, 3\}$  is periodic because you can only return to the starting state in an even number of time steps and

$$\text{gcd}(\mathbb{T}(0)) = \text{gcd}(\mathbb{T}(1)) = \text{gcd}(\mathbb{T}(2)) = \text{gcd}(\mathbb{T}(3)) = \text{gcd}(\{2, 4, 6, \dots\}) = 2 \neq 1 .$$

**Exercise 178** Show that the Markov chain corresponding to a drunkard's walk around a polygonal block with  $k$  corners is irreducible for any integer  $k > 1$ . Show that it is aperiodic only when  $k$  is odd and has period 2 when  $k$  is even.

**Proposition 86** Let  $A = \{a_1, a_2, \dots\} \subset \mathbb{N}$  that satisfies the following two conditions:

1.  $A$  is a **nonlattice**, meaning that  $\text{gcd}(A) = 1$  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 87** If the Markov chain  $(X_t)_{t \in \mathbb{Z}_+}$  with transition matrix  $P$  on state space  $\mathbb{X}$  is irreducible and aperiodic then there is an integer  $\tau$  such that  $P^t(x, x) > 0$  for all  $t \geq \tau$  and all  $x \in \mathbb{X}$ .

**Proof:** TBD

**Proposition 88** If the Markov chain  $(X_t)_{t \in \mathbb{Z}_+}$  with transition matrix  $P$  on state space  $\mathbb{X}$  is irreducible and aperiodic then there is an integer  $\tau$  such that  $P^t(x, y) > 0$  for all  $t \geq \tau$  and all  $(x, y) \in \mathbb{X}^2$ .

**Proof:** TBD

**Exercise 179 (King's random walk on a chessboard)** Consider the squares in the chessboard as the state space  $\mathbb{X} = \{0, 1, 2, \dots, 7\}^2$  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 180 (King's random walk on a chesstorus)** We can obtain a chesstorus from a pliable chessboard by identifying the eastern edge with the western edge (roll the chessboard into a cylinder) and then identifying the northern edge with the southern edge (gluing the top and bottom end of the cylinder together by turning into a doughnut or torus). Consider the squares in the chesstorus as the state space  $\mathbb{X} = \{0, 1, 2, \dots, 7\}^2$  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:

$$P(W_t = w) = \begin{cases} \frac{1}{8} & \text{if } w \in \{(1, 1), (1, 0), (1, -1), (0, -1), (-1, -1), (-1, 0), (-1, 1), (0, 1)\}, \\ 0 & \text{otherwise.} \end{cases}$$

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.

## 6.4 Stationarity

We are interested in statements about a Markov chain that has been running for a long time. For any nontrivial Markov chain  $(X_0, X_1, \dots)$  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 sattes that the distribution of  $X_t$  over  $\mathbb{X}$  settles down as  $t \rightarrow \infty$ , provided the Markov chain is irreducible and aperiodic.

**Definition 89 (Stationary distribution)** Let  $(X_t)_{t \in \mathbb{Z}_+}$  be a Markov chain with state space  $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$  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 90 (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**

$$\tau(x, y) := E(T(x, y)),$$

be the expected time taken to reach  $y$  after having started at  $x$ . Note that  $\tau(x, x)$  is the **mean return time** to state  $x$ .

**Proposition 91 (Hitting times of irreducible aperiodic Markov chains)** If  $(X_t)_{t \in \mathbb{Z}_+}$  is an irreducible aperiodic Markov chain with state space  $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$ , transition matrix  $P = (P(x, y))_{(x, y) \in \mathbb{X}^2}$  then for any pair of states  $(x, y) \in \mathbb{X}^2$ ,

$$P(T(x, y) < \infty) = 1 ,$$

and the mean hitting time is finite, i.e.,

$$\tau(x, y) < \infty .$$

**Proposition 92 (Existence of Stationary distribution)** For any irreducible and aperiodic Markov chain there exists at least one stationary distribution.

**Proof:** TBD

**Definition 93 (Total variation distance)** If  $\nu_1 := (\nu_1(x))_{x \in \mathbb{X}}$  and  $\nu_2 := (\nu_2(x))_{x \in \mathbb{X}}$  are elements of  $\mathcal{P}(\mathbb{X})$ , the set of all probability distributions on  $\mathbb{X} := \{s_1, s_2, \dots, s_k\}$ , then we define the **total variation distance** between  $\nu_1$  and  $\nu_2$  as

$$d_{TV}(\nu_1, \nu_2) := \frac{1}{2} \sum_{x \in \mathbb{X}} \text{abs}(\nu_1(x) - \nu_2(x)) , \quad d_{TV} : \mathcal{P}(\mathbb{X}) \times \mathcal{P}(\mathbb{X}) \rightarrow [0, 1] . \quad (6.10)$$

If  $\nu_1, \nu_2, \dots$  and  $\nu$  are probability distributions on  $\mathbb{X}$ , then we say that  $\nu_t$  **converges in total variation** to  $\nu$  as  $t \rightarrow \infty$  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 6.10 ensures that the range of  $d_{TV}$  is in  $[0, 1]$ . If  $d_{TV}(\nu_1, \nu_2) = 1$  then  $\nu_1$  and  $\nu_2$  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  $\nu_1$  and  $\nu_2$  is the maximal difference in probabilities that the two distributions assign to any one event  $A \in \sigma(\mathbb{X}) = 2^\mathbb{X}$ .

In words, Proposition 94 says that if you run the chain for a sufficiently long enough time  $t$ , then, regardless of the initial distribution  $\mu_0$ , the distribution at time  $t$  will be close to the stationary distribution  $\pi$ . This is referred to as the Markov chain **approaching equilibrium or stationarity** as  $t \rightarrow \infty$ .

**Proposition 94 (Markov chain convergence theorem)** Let  $(X_t)_{t \in \mathbb{Z}_+}$  be an irreducible aperiodic Markov chain with state space  $\mathbb{X} = \{s_1, s_2, \dots, s_k\}$ , transition matrix  $P = (P(x, y))_{(x,y) \in \mathbb{X}^2}$  and initial distribution  $\mu_0$ . Then for any distribution  $\pi$  which is stationary for the transition matrix  $P$ , we have

$$\mu_t \xrightarrow{\text{TV}} \pi . \quad (6.11)$$

**Proof:** TBD

**Proposition 95 (Uniqueness of stationary distribution)** Any irreducible aperiodic Markov chain has a unique stationary distribution.

**Proof:** TBD

**Exercise 181** Consider the Markov chain on  $\{1, 2, 3, 4, 5, 6\}$  with the following transition matrix:

$$P = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & \left( \begin{array}{cccccc} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{array} \right) \end{pmatrix} .$$

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 182** If there are two stationary distributions  $\pi$  and  $\pi'$  then show that there is a infinite family of stationary distributions  $\{\pi_p : p \in [0, 1]\}$ , called the convex combinations of  $\pi$  and  $\pi'$ .

**Exercise 183** Show that for a drunkard's walk chain started at state 0 around a polygonal block with  $k$  corners labelled  $\{0, 1, 2, \dots, k - 1\}$ , 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  $\mu_t$  when  $k$  is even.

## 6.5 Reversibility

We introduce another specific property called reversibility. This property will assist in conjuring Markov chains with a desired stationary distribution.

**Definition 96 (Reversible)** A probability distribution  $\pi$  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  $(x, y) \in \mathbb{X}^2$ :

$$\pi(x)P(x, y) = \pi(y)P(y, x) . \quad (6.12)$$

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  $\pi$ , i.e.,  $\mu_0 = \pi$ , then the probability of going from  $x$  to  $y$  is the same as that of going from  $y$  to  $x$ .

**Proposition 97 (A reversible  $\pi$  is a stationary  $\pi$ )** 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  $\pi$  is a reversible distribution for  $(X_t)_{t \in \mathbb{Z}_+}$  then  $\pi$  is a stationary distribution for  $(X_t)_{t \in \mathbb{Z}_+}$ .

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

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

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

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

**Definition 98 (Graph)** A Graph  $\mathbb{G} := (\mathbb{V}, \mathbb{E})$  consists of a **vertex set**  $\mathbb{V} := \{v_1, v_2, \dots, v_k\}$  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 21 (Random Walk on an Undirected Graph)** A random walk on an undirected graph  $\mathbb{G} = (\mathbb{V}, \mathbb{E})$  is a Markov chain with state space  $\mathbb{V} := \{v_1, v_2, \dots, v_k\}$  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 99** 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  $\pi$  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  $\pi$  is a probability distribution provided that  $d > 0$ . To show that  $\pi$  is reversible we need to verify Equation 6.12 for each  $(v_i, v_j) \in \mathbb{V}^2$ . Fix a pair of states  $(v_i, v_j) \in \mathbb{V}^2$ , then

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

By Proposition 97  $\pi$  is also the stationary distribution.

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

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

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

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

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

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

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

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

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

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

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

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

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

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  $\pi$ )  $\pi$  is the unique stationary distribution. Due to Proposition 97,  $\pi$  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  $\pi$  since,

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

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

**Model 22 (Random Walk on a Directed Graph)** A random walk on a directed graph  $\mathbb{G} = (\mathbb{V}, \mathbb{E})$  is a Markov chain with state space  $\mathbb{V} := \{v_1, v_2, \dots, v_k\}$  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 189 (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  $P$ :

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 1 & 1 \\ 3 & 1 & 0 & 0 & 0 \\ 4 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad P = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 3 & 1 & 0 & 0 & 0 \\ 4 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad \text{Diagram: } \begin{array}{c} \text{1} \rightarrow \text{2} \rightarrow \text{4} \rightarrow \text{3} \\ \text{1} \leftarrow \text{2} \leftarrow \text{4} \leftarrow \text{3} \end{array}.$$

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

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

## 6.6 Standard normal distribution function table

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

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

# Summary of Probability Theory I

## 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.
$\Omega$	— universal set.
$A^c$	— the complement of $A$ , meaning the set of all elements in $\Omega$ , the universal set, which are not in $A$ .

## EXPERIMENT SUMMARY

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

## PROBABILITY SUMMARY

Axioms:

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

Rules:

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

## CONDITIONAL PROBABILITY SUMMARY

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

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

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

Conditional probabilities obey the 4 axioms of probability.

## DISCRETE RANDOM VARIABLE SUMMARY

Probability mass function

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

Distribution function

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

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

## CONTINUOUS RANDOM VARIABLES: NOTATION

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

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

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

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

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

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

## Some Common Expectations

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

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

Table 6.1: Symbol Table: Probability and Statistics

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

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

Symbol	Meaning
$A = \{\star, \circ, \bullet\}$	$A$ is a set containing the elements $\star, \circ$ and $\bullet$
$\circ \in A$	$\circ$ belongs to $A$ or $\circ$ is an element of $A$
$A \ni \circ$	$\circ$ belongs to $A$ or $\circ$ is an element of $A$
$\odot \notin A$	$\odot$ does not belong to $A$
$\#A$	Size of the set $A$ , for e.g. $\#\{\star, \circ, \bullet, \odot\} = 4$
$\mathbb{N}$	The set of natural numbers $\{1, 2, 3, \dots\}$
$\mathbb{Z}$	The set of integers $\{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$
$\mathbb{Z}_+$	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 6.3: Symbol Table: Sets and Numbers

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

Table 6.4: Symbol Table: Probability and Statistics

# Chapter 7

## Inference for 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 100 (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  $\Theta$ . We refer to  $\Theta$  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:

$$\Theta \ni \theta \mapsto P_\theta \in \mathbb{P} .$$

### 7.2 Some Common Experiments

Next, let's formally consider some experiments we have already encountered.

**Experiment 23 (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 (3.26) 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.  $F(x) = F^{[-1]}(x)$ . Recall from Exercise 2.1 that this is equivalent to infinitely many independent and identical Bernoulli( $1/2$ ) trials, i.e., independently tossing a fair coin infinitely many times.

Figure 7.1: Geometry of the  $\Theta$ 's for de Moivre[ $k$ ] Experiments with  $k \in \{1, 2, 3, 4\}$ .

**Experiment 24 (Bernoulli)** The ‘toss 1 times’ experiment is the following family of densities (PMFs) :

$$\mathbb{P} = \{ f(x; \theta) : \theta \in [0, 1] \}$$

where,  $f(x; \theta)$  is given in (3.13). The one dimensional parameter space or index set for this experiment is  $\Theta = [0, 1] \subset \mathbb{R}$ .

**Experiment 25 (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 (3.47). 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 a **parametric** experiment. Experiment 24 involving discrete RVs as well as Experiment 25 are **parametric** since they both have only one parameter (the parameter space is one dimensional for Experiments 24 and 25). The Fundamental Experiment 23 involving the continuous RV of Model 7 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 26 (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  $\Theta$  is the  $k$ -Simplex  $\Delta_k$ . 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  $\theta_i$ ’s.”

An experiment with infinite dimensional parameter space  $\Theta$  is said to be **nonparametric**. Next we consider two nonparametric experiments.

**Experiment 27 (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 28 (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  $\theta$ , we can estimate  $\theta$  from a statistic  $T_n$  called the estimator of  $\theta$  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  $\Theta_0$  of  $\Theta$  called the space of null hypotheses, we will learn to test (attempt to reject) the falsifiable null hypothesis that  $F \in \Theta_0 \subset \Theta$ .
- ...

## 7.4 Decision Problems and Procedures for Actions

Write down the Table from lectures 1 & 2 giving examples of decision problems, procedures and action spaces for typical estimation, hypothesis testing and prediction problems with associated principles (Maximum Likelihood, Empirical Risk Minimisation where risk is expectation of specific loss functions, etc.) and algorithms (including optimisation (Stochastic)Newton/gradient-descent, etc.).

## 7.5 Statistics for Bernoulli Trials (Inference for Proportions)

### 7.5.1 Testing biasedness of a coin

Suppose we have a coin where  $\theta$  is the probability of coming up with Heads and  $1 - \theta$  is the probability of coming up with Tails. Here  $\theta$  is a number between 0 and 1. The coin is fair (unbiased) if  $\theta = 1/2$ ; otherwise it is a biased coin. Throwing a coin constitutes a **Bernoulli trial** if we identify Heads with the number 1 and Tails with 0. The random variable  $X$  symbolizing the outcome of this experiment is a **Bernoulli random variable**. Thus we have

$$P(X = 0) = 1 - \theta, \quad P(X = 1) = \theta, \quad 0 < \theta < 1.$$

Generally we abbreviate “random variable” by RV.

Recall that the distribution of a discrete RV is the entirety of its possible values  $x_1, x_2, \dots$  along with the associated probabilities  $\theta_i = P(X = x_i)$ . The **Bernoulli distribution**  $\text{Bernoulli}(\theta)$  is by definition the distribution of  $X$ , i.e. the possible values are 0 and 1 and the associated probabilities are  $1 - \theta$  and  $\theta$ .

How can we test whether a given coin is biased or not? Obviously we should throw the coin repeatedly and compare the number of Heads with the number of Tails. If we throw the coin  $n$  times, this provides us with  $n$  independent Bernoulli trials, symbolized by independent Bernoulli random variables  $X_1, \dots, X_n$ .

**First example of a hypothesis test.** To phrase the biasedness question in statistical terms, introduce two hypotheses:

$$\begin{aligned} H_0 &: \theta = 1/2 \text{ null hypothesis} \\ H_1 &: \theta \neq 1/2 \text{ alternative hypothesis} \end{aligned}$$

To test if the null hypothesis is true, we throw the coin  $n$  times and let  $X_i = 1$  if Heads comes up on the  $i$ th trial and 0 otherwise, so that  $\bar{X}_n$  is the fraction of times Heads comes up in the first  $n$  trials. The test is specified by giving a critical region  $\mathcal{C}_n$  so that we reject  $H_0$  (that is, decide  $H_0$  is incorrect) when  $\bar{X}_n \in \mathcal{C}_n$ . One possible choice in this case is

$$\mathcal{C}_n = \left\{ x : \left| x - \frac{1}{2} \right| > 1/\sqrt{n} \right\}.$$

This choice is motivated by the fact that if  $H_0$  is true then using the central limit theorem ( $Z$  is a standard normal variable), with  $\theta = 1/2$

$$P(\bar{X}_n \in \mathcal{C}_n) = P\left(\left| \frac{\bar{X}_n - \theta}{\sqrt{\theta(1-\theta)/n}} \right| \geq 2\right) \approx P(|Z| \geq 2) = 0.05. \quad (7.1)$$

and  $2\sqrt{\frac{1}{2} \cdot \frac{1}{2}} = 1$ . Rejecting  $H_0$  when it is true is called a **type I error**. In this test we have set the type I error to be 5%.

### 7.5.2 Review of underlying probability concepts

Although Probability Theory I is a pre-requisite for Inference Theory I, we will review the concepts again as everyone may not have met the pre-requisites at the level needed for the sequel. In the process we will use slightly different notational conventions as these are more common in mathematical statistics.

Recall the notion of independence of RV's: suppose  $X_1, X_2$  are RV's which can be jointly observed, i.e. they have a joint distribution. Suppose also that both  $X_1, X_2$  have the same finite range of possible values:  $\mathbb{X} = \{\xi_1, \dots, \xi_m\}$ . Then independence means

$$P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2), \quad x_1, x_2 \in \mathbb{X}.$$

Independence of  $n$  RV's  $X_1, \dots, X_n$  is defined analogously, where for any set of values  $x_1, \dots, x_n$  (all from  $\mathbb{X}$ ) we require

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i).$$

Throwing a coin  $n$  times, our Bernoulli RV's  $X_1, \dots, X_n$  are not only independent but also *identically distributed*. This means that all  $X_i$  considered "alone" (i.e. in their marginal distribution) have the same distribution  $\text{Bernoulli}(\theta)$ . Such an array of RV's is often called **independent and identically distributed**, abbreviated IID. The case of observed IID RV's is the most frequently assumed and encountered one in statistics. Specialized to Bernoulli RV's  $X_1, \dots, X_n$  the IID assumption means: if  $x_1, \dots, x_n$  is any sequence of 0's and 1's then

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i) = \theta^k (1 - \theta)^{n-k}$$

where  $k$  is the number of 1's in the  $n$ -tuple  $x_1, \dots, x_n$ . Note that we may write

$$k = \sum_{i=1}^n x_i$$

hence

$$P(X_1 = x_1, \dots, X_n = x_n) = \theta^{(\sum_{i=1}^n x_i)} (1 - \theta)^{n - (\sum_{i=1}^n x_i)}.$$

**Exercise 7.1 (Subsets of independent RVs)** Suppose that  $X_1, \dots, X_n$  is a set of RV's all having the same finite range of possible values:  $\mathbb{X} = \{\xi_1, \dots, \xi_m\}$ , i.e. we have  $X_i \in \mathbb{X}$ ,  $i = 1, \dots, n$ . Define independence of  $X_1, \dots, X_n$  by the property: for all  $n$ -tuples  $(x_1, \dots, x_n) \in \mathbb{X}^n$  we have

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i).$$

Show that independence of  $X_1, \dots, X_n$  implies independence of any subset  $X_{i_1}, \dots, X_{i_k}$  where  $\{i_1, \dots, i_k\}$  is an arbitrary subset of size  $k$  of the indices  $\{1, \dots, n\}$ , and  $2 \leq k < n$ .

**Exercise 7.2 (Convergence in distribution and probability to Point Mass RV)** Let  $X_1, X_2, \dots$  be a sequence of RV's and  $\mu$  be a real number. Let  $Y$  be a RV taking value  $\mu$  with probability one

$(P(Y = \mu) = 1)$ . The law, or probability distribution, of  $Y$  is called the *degenerate law concentrated at  $\mu$*  or the Point Mass( $\mu$ ) RV; it has distribution function

$$G_\mu(x) = P(Y \leq x) = \begin{cases} 0, & x < \mu \\ 1, & x \geq \mu. \end{cases}$$

Show that as  $n \rightarrow \infty$

$$X_n \rightsquigarrow Y \text{ if and only if } X_n \xrightarrow{P} \mu,$$

i.e. convergence in distribution to the degenerate law means convergence in probability to  $\mu$ . For these convergence notions, see Definition (103) ( regarding  $\rightsquigarrow$ ), and the well known convergence in probability:  $X_n \xrightarrow{P} \mu$  if  $P(|X_n - \mu| \geq \varepsilon) \rightarrow 0$  for every  $\varepsilon > 0$ .

**Moments of the Bernoulli distribution.** Suppose that  $X$  has the Bernoulli law  $\text{Bernoulli}(\theta)$ . A notation we will frequently use is

$$\mathcal{L}(X) = \text{Bernoulli}(\theta).$$

Here “ $\mathcal{L}(X)$ ” means “the law of the RV  $X$ ”, where the law is a short word for the distribution (derived from “probability law”, an older term for “distribution”). Now it is easy to see, with the shorter notation for expectations without the  $(\cdot)$  for convenience, that

$$\begin{aligned} E(X) &= EX = 0 \cdot (1 - \theta) + 1 \cdot \theta = \theta, \\ E(X^2) &= EX^2 = 0^2 \cdot (1 - \theta) + 1^2 \cdot \theta = \theta \end{aligned}$$

and hence

$$V(X) = EX^2 - (EX)^2 = \theta - \theta^2 = \theta(1 - \theta).$$

**The law of large numbers.** Specialized to our case of IID Bernoulli's it gives

$$\bar{X}_n = n^{-1} \sum_{i=1}^n X_i \xrightarrow{\theta} EX = \theta$$

where  $\xrightarrow{P}$  (or equivalently  $\xrightarrow{P}$ ) denotes convergence in probability. Here  $\bar{X}_n$  is the arithmetic mean of  $X_1, \dots, X_n$ , also called the **sample mean**. Recall that convergence in probability  $\bar{X}_n \xrightarrow{P} \theta$  means: for every  $\varepsilon > 0$

$$P(|\bar{X}_n - \theta| > \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Here  $\theta = EX$  may be also be called “**population mean**”. This derives from the fact that in many examples other than coin throw, a Bernoulli RV  $X$  is obtained from randomly selecting an individual from a large population. For instance, we might select a random individual from the U.S. population and observe whether it is a smoker or nonsmoker. Provided our selection is “truly random”, we obtain a Bernoulli distribution  $\text{Bernoulli}(\theta)$  where  $\theta$  is the proportion of smokers in the population at large. By this reasoning, we often identify a random variable  $X$  with a “population” and its expectation and variance  $EX, V(X)$  with the population mean and variance. Thus there is a correspondence between sample mean  $\bar{X}_n$  and population mean  $EX$  etc. Of course there are many random variables occurring in practice which do not arise from “selecting from a population”, for example a coin throw, hitting a target when shooting, getting rain on a hike etc.,

**Population proportion and sample proportion.** For Bernoulli RV's  $X_1, \dots, X_n$ , the sample mean  $\bar{X}_n$  may be identified with the *sample proportion of 1's*. Indeed

$$\bar{X}_n = n^{-1} \sum_{i=1}^n X_i = \frac{\# \text{ of 1's in the sample } X_1, \dots, X_n}{n}$$

thus  $\bar{X}_n$  is the number of 1's in the sample relative to sample size  $n$ , briefly called **sample proportion**  $\hat{\theta}_n$ . Thus for Bernoulli RV's  $X_1, \dots, X_n$  we have

$$\bar{X}_n = \hat{\theta}_n.$$

The same correspondence exists on the population level:  $E X = \theta$  where  $\theta$  may be called the **population proportion**. To repeat it,  $X$  may not actually be the result of selecting from some population;  $X$  may be the outcome of a random experiment like a coin throw. Still the terminology “sample proportion/ population proportion” is widely used in statistics.

### Chebyshev's inequality and the Law of Large Numbers (LLN)

The law of large numbers (LLN) holds under a general assumption that  $E |X_i| < \infty$ , see [D]<sup>1</sup> p. 223. or Proof via CFs in Probability Theory I earlier. But we recall here from scratch for reinforcement of your learning.

**Proposition 101 (Weak Law of Large Numbers)** Suppose  $X_1, X_2, \dots$  are IID RV's and have  $E |X_i| < \infty$ . Let  $\mu = E X_i$ . Then as  $n \rightarrow \infty$

$$\bar{X}_n \xrightarrow{P} \mu$$

in other words: for every  $\varepsilon > 0$

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \rightarrow 0.$$

The general proof is not given in [D], but is argued under an additional assumption that the IID  $X_i$  have a finite variance. In this case the LLN follows from *Chebyshev's inequality* ([D] p. 222): if  $Y$  is an RV with finite variance  $\sigma^2$  then for any  $t > 0$

$$P(|Y - E Y| \geq t) \leq \frac{V(Y)}{t^2}. \quad (7.2)$$

This easily yields a proof of the weak LLN under an additional assumption  $V(X_i) = \sigma^2 < \infty$ : note that  $V(\bar{X}_n) = \sigma^2/n$  so that

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{V(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0.$$

**Proof:** [Proof of Chebyshev's inequality (7.2)] Let  $X$  be a nonnegative RV ( $X \geq 0$ ) with finite expectation:  $E X < \infty$ . Let  $\mathbf{1}_A$  be the indicator function of an event  $A$ . Then for  $u > 0$

$$E X = E X \mathbf{1}_{\{X < u\}} + E X \mathbf{1}_{\{X \geq u\}} \geq E X \mathbf{1}_{\{X \geq u\}} \geq u E \mathbf{1}_{\{X \geq u\}} = u P(X \geq u).$$

Thus we obtain *Markov's inequality*

$$P(X \geq u) \leq \frac{E X}{u}.$$

Setting  $X = |Y - E Y|^2$  we obtain

$$P(|Y - E Y| \geq t) = P(|Y - E Y|^2 \geq t^2) \leq \frac{E |Y - E Y|^2}{t^2} = \frac{V(Y)}{t^2}.$$

To understand what the assumption of a finite variance means, it is instructive to find an example of a RV having  $E |X| < \infty$  but with infinite variance.

---

<sup>1</sup>Throughout we will use [D] for the reference: Durrett, R., *The Essentials of Probability*, Duxbury Press, 1994.

### Normal approximation and the Central Limit Theorem (CLT)

Let us state the Central Limit Theorem, following [D], p. 228.

**Proposition 102** Suppose  $X_1, X_2, \dots$  are IID RV's and have  $E X_i = \mu$  and  $V(X_i) = \sigma^2$  with  $0 < \sigma^2 < \infty$ . Then as  $n \rightarrow \infty$

$$P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right) \longrightarrow P(Z \leq x) \text{ for all } x$$

where  $Z$  denotes a random variable with the standard normal distribution.

Specialized to IID Bernoulli RV's  $X_1, \dots, X_n$  with law  $\text{Bernoulli}(\theta)$ , this says

$$\sqrt{n} \frac{(\hat{\theta}_n - \theta)}{\sqrt{\theta(1-\theta)}} \rightsquigarrow \text{Normal}(0, 1) \text{ as } n \rightarrow \infty. \quad (7.3)$$

Here  $\text{Normal}(0, 1)$  is the *standard normal distribution* (or standard Gaussian distribution) and  $\rightsquigarrow$  denotes *convergence in distribution* (or in law) of a RV. A random variable  $Z$  has the standard normal distribution if

$$P(Z \leq z) = \Phi(z) = \int_{-\infty}^z \varphi(t) dt$$

where  $\Phi$  is the *standard normal distribution function*, defined in terms of the *standard normal density*

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2).$$

Thus we have a number of symbols associated with the standard normal distribution:  $\text{Normal}(0, 1)$  is the distribution itself,  $Z$  is a common symbol for a RV having that law, i.e.  $\mathcal{L}(Z) = \text{Normal}(0, 1)$ ;  $\Phi$  is the distribution function  $\Phi(t) = P(Z \leq t)$  and  $\varphi$  is the density. The *general normal* (or Gaussian) distribution with mean  $\mu$  and variance  $\sigma^2$  is denoted by  $\text{Normal}(\mu, \sigma^2)$ ; it is defined as

$$\text{Normal}(\mu, \sigma^2) := \mathcal{L}(\mu + \sigma Z).$$

The convergence stated in the CLT is a special case of *convergence in distribution*.

**Definition 103** A sequence of RV's  $Y_n$  converges in distribution (or in law) to a RV  $Y$ , written

$$Y_n \rightsquigarrow Y \text{ as } n \rightarrow \infty$$

if

$$P(Y_n \leq z) \longrightarrow P(Y \leq z) \text{ as } n \rightarrow \infty \quad (7.4)$$

for every point of continuity  $z$  of the distribution function of  $Y$ .

Since for a standard normal  $Z$  the distribution function is  $\Phi$  which is continuous everywhere, in the case of the CLT we simply have (7.4) for every  $z$ , and the CLT as stated above indeed gives a convergence in distribution. Other ways of writing a convergence in law (in distribution) are

$$\mathcal{L}(Y_n) \rightsquigarrow \mathcal{L}(Y) \text{ or } Y_n \rightsquigarrow \mathcal{L}(Y)$$

which is justified since convergence in distribution is a statement about the laws (or distribution functions) of  $Y_n$  and  $Y$ . Thus in the case of the CLT we may write

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \rightsquigarrow \text{Normal}(0, 1) \text{ as } n \rightarrow \infty$$

since  $\text{Normal}(0, 1) = \mathcal{L}(Z)$ , and for Bernoulli's this specializes to (7.3).

In [D] the CLT is proved under the assumption that  $E\exp(tX) < \infty$  for  $t \in (-t_0, t_0)$  and some  $t_0 > 0$ . For a Bernoulli  $X$  this is trivially fulfilled since  $X \leq 1$  and hence  $E\exp(tX) \leq \exp(t)$ .

**Exercise 7.3 (CLT implies LLN)** Show that the CLT implies the LLN. More precisely, assume that a sequence of RV's  $Y_n$  satisfies

$$\frac{Y_n - \mu}{\sigma/\sqrt{n}} \rightsquigarrow \text{Normal}(0, 1) \text{ as } n \rightarrow \infty$$

for certain  $\mu, \sigma$  where  $\sigma > 0$ . Show that  $Y_n \xrightarrow{\text{P}} \mu$ .

### Normal approximation for the binomial distribution

Recall the definition of the binomial distribution: if  $X_1, \dots, X_n$  are IID Bernoulli  $\text{Bernoulli}(\theta)$  then the distribution of the sum  $S_n = \sum_{i=1}^n X_i$  is the *binomial distribution*  $\text{Binomial}(n, \theta)$ . The probability function of  $\text{Binomial}(n, \theta)$  is

$$P(S_n = k) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}, \quad k = 0, \dots, n. \quad (7.5)$$

The binomial law has two parameters-  $n$ , the number of trials, and  $\theta$ , the probability of “success”, i.e. of 1. Thus the binomial law is the distribution of the number of successes in  $n$  independent Bernoulli trials (with the same probability of success).

**Proof:** [Proof of (7.5)] Let  $x_1, \dots, x_n$  be an arbitrary collection of 0's and 1's. We have seen above that

$$P(X_1 = x_1, \dots, X_n = x_n) = \theta^{(\sum_{i=1}^n x_i)} (1 - \theta)^{n - (\sum_{i=1}^n x_i)}$$

where  $\sum_{i=1}^n x_i$  is the number of 1's among  $x_1, \dots, x_n$ , i.e. the number of successes. Thus

$$\begin{aligned} P\left(\sum_{i=1}^n X_i = k\right) &= \sum_{(x_1, \dots, x_n) : \sum_{i=1}^n x_i = k} \theta^k (1 - \theta)^{n-k} \\ &= \theta^k (1 - \theta)^{n-k} \cdot \#\left\{(x_1, \dots, x_n) : \sum_{i=1}^n x_i = k\right\} \\ &= \theta^k (1 - \theta)^{n-k} \binom{n}{k} \end{aligned}$$

(indeed the number of  $n$ -tuples of  $(x_1, \dots, x_n)$  of 0's and 1's having exactly  $k$  1's is  $\binom{n}{k}$ - choose the  $k$  positions among positions  $1, \dots, n$  where you place 1's).

Thus the CLT, specialized to IID Bernoullis in (7.3), is a statement about the normal approximation of the binomial law. Indeed let  $S_n = \sum_{i=1}^n X_i$ ; we may write

$$\begin{aligned} \hat{\Theta}_n &= \bar{X}_n = n^{-1} \sum_{i=1}^n X_i = n^{-1} S_n, \\ \sqrt{n} \frac{(\hat{\Theta}_n - \theta)}{\sqrt{\theta(1 - \theta)}} &= \\ &= \frac{S_n - n\theta}{\sqrt{n\theta(1 - \theta)}} \rightsquigarrow \text{Normal}(0, 1) \text{ as } n \rightarrow \infty. \end{aligned} \quad (7.6)$$

Here  $S_n$  has the binomial law  $\text{Binomial}(n, \theta)$ . The form (7.6) of the CLT is also called the **De Moivre- Laplace theorem**; historically it was the first version of the CLT (De Moivre (1733) for  $\theta = 1/2$ , Laplace (1812) for  $0 < \theta < 1$ ). We see that the left side of (7.6) is just the standardized sum  $S_n$ , by calculating its first two moments:

$$\begin{aligned}\mathbb{E} S_n &= \sum_{i=1}^n \mathbb{E} X_i = n\theta, \\ \mathbb{V}(S_n) &= \sum_{i=1}^n \mathbb{V}(X_i) = n\theta(1 - \theta).\end{aligned}$$

Recall that standardizing a RV  $Y$  means subtracting its expectation  $\mathbb{E} Y$  and then dividing by the standard deviation  $\text{SD}(Y) = \sqrt{\text{Var}(Y)}$  so that the standardized expression

$$\frac{Y - \mathbb{E} Y}{\text{SD}(Y)}$$

has mean 0 and variance 1. Hence the verbal summary of the CLT: “the standardized sum of IID RV’s is approximately standard normal”. Of course standardizing the sum  $S_n = \sum_{i=1}^n X_i$  yields the same result as standardizing the sample mean  $\bar{X}_n$ :

$$\begin{aligned}\mathbb{E} S_n &= n \mathbb{E} \bar{X}_n \text{ and } \mathbb{V}(S_n) = n^2 \mathbb{V}(\bar{X}_n), \text{ hence} \\ \frac{S_n - \mathbb{E} S_n}{\text{SD}(S_n)} &= \frac{\bar{X}_n - \mathbb{E} \bar{X}_n}{\text{SD}(\bar{X}_n)}\end{aligned}$$

so it is also true that “the standardized (sample) mean of IID RV’s is . . .”.

### A visualization of the De Moivre-Laplace central limit theorem

Let  $X_1, \dots, X_n$  be independent identically distributed random variables having the Bernoulli law  $\text{Binomial}(1, \theta)$  with probability of success  $\theta$ . Recall that  $S_n = \sum_{i=1}^n X_i$  then has the binomial law  $\text{Binomial}(n, \theta)$  with probabilities

$$P(Y = k) = \frac{n!}{k! \cdot (n-k)!} \cdot \theta^k (1 - \theta)^{n-k}.$$

To plot these probabilities we use the Gamma-function  $\Gamma(x)$  which is defined for all  $x > 0$ . and which has the property that for integers  $k$

$$\Gamma(k+1) = k!$$

Thus we are able to plot a continuous function for all  $x$  in a range, and we obtain a visualization of the factorial and derived expressions. We let  $x$  be the continuous variable taking the place of  $k = 0, 1, \dots, n$ .

Define a function

$$b_n(x) = \frac{\Gamma(n+1)}{\Gamma(x+1) \cdot \Gamma(n-x+1)} \cdot \theta^x \cdot (1 - \theta)^{n-x}$$

Here  $b_n(x)$  represents the binomial law  $\text{Binomial}(n, \theta)$  in the following sense:

$$b_n(k) = \frac{n!}{k!(n-k)!} \theta^k (1 - \theta)^{n-k}, \quad k = 0, 1, \dots, n.$$

Set  $\theta = 1/3$  and  $n = 600$ . The plot of the function  $b_n(x)$ , which interpolates the binomial probabilities, is:

[done in Lecture 3 - get/read notes to draw by hand here.]

Let us look at this picture around the expected value  $n\theta = 600/3 = 200$  in the range of three standard deviations. The standard deviation is  $\sigma = \sqrt{n} \cdot \sqrt{\theta \cdot (1 - \theta)} = 11.55$ , thus  $3\sigma = 34.65 \approx 35$  and we take a range values of  $x$  from 165 to 235.

[done in Lecture 3 - get/read notes to draw by hand here.]

The curve is visually indistinguishable from a normal density. For a comparison we plot the normal density with expectation  $\mu = 200$  and standard deviation  $\sigma = 11.55$ , i. e. the function

$$f(x) = \frac{1}{\sigma} \varphi \left( \frac{x - \mu}{\sigma} \right)$$

where

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$$

is the standard normal density  $\varphi(t)$ . The density  $f(x)$  is plotted in (red) dots over the previous function  $b_n(x)$ .

[done in Lecture 3 - get/read notes to draw by hand here.]

In this reasoning, we plotted the interpolated probability function of the sum  $S_n$  against the density of the normal distribution  $\text{Normal}(\mu, \sigma^2) = \text{Normal}(n\theta, n\theta(1 - \theta))$ . In fact since the CLT says that the standardized sum  $\frac{S_n - n\theta}{\sqrt{n\theta(1 - \theta)}}$  is approximately standard normal:

$$\frac{S_n - n\theta}{\sqrt{n\theta(1 - \theta)}} \rightsquigarrow Z \quad (7.7)$$

by multiplying by  $\sigma = \sqrt{n\theta(1-\theta)}$  and then adding  $\mu = n\theta$  we would infer  $S_n$  that should be approximately

$$S_n \approx \sqrt{n\theta(1-\theta)}Z + n\theta \quad (7.8)$$

where  $\mathcal{L}(\sigma Z + \mu) = \text{Normal}(\mu, \sigma^2)$ . But here we should be cautious as to the meaning of the sign “ $\approx$ ”: in (7.8) both the left side and the right side depend on  $n$ , so we do not have a limit relation. Nevertheless it is common in statistics to state the normal approximation to the binomial as

$$S_n \approx \text{Normal}(n\theta, n\theta(1-\theta)).$$

This is correct if we understand it to mean the CLT (7.7). When we plot the distributions  $\text{Binomial}(n, \theta)$  and  $\text{Normal}(n\theta, n\theta(1-\theta))$ , we make certain scale transforms anyway, e.g. we look at the distributions around their expectation, on the scale of the standard deviation. This amounts to standardization, and what we see is in fact the CLT (7.7).

Moreover (7.8) can be made rigorous by introducing a certain distance for distributions and then claiming that the distance between  $\mathcal{L}(S_n)$  and  $\text{Normal}(n\theta, n\theta(1-\theta))$  tends to zero; we will not elaborate this here.

### 7.5.3 The success / failure rule

In applied statistics one finds a rule which limits the applicability of the normal approximation to the binomial  $\text{Binomial}(n, \theta)$ : it is required that both  $n\theta \geq 10$  and  $n(1-\theta) \geq 10$ . This is called the “success / failure rule” since  $n\theta$  is the expected number of successes:  $E S_n = n\theta$  and  $n(1-\theta)$  is the expected number of “failures”:  $E(n - S_n) = n(1-\theta)$ . This rule is based on the fact that the CLT “breaks down” for small values of  $\theta$ . More precisely, when  $\theta$  is small, a larger  $n$  is needed to make the normal approximation good; for large enough  $n$  a small  $\theta$  can always be compensated. For an illustration, select  $\theta = 1/200$  and again  $n = 600$ ; then  $n\theta = 3$  and the success / failure rule is violated. The plot of  $b_n(k)$  is as follows:

[done in Lecture 3 - get/read notes to draw by hand here.]

Next we will look at this distribution around  $\mu = n\theta$ , i.e.  $\mu = 3$  in the range of three standard deviations, i. e.  $3\sigma$  where  $\sigma = \sqrt{n\theta(1-\theta)}$ , thus  $\sigma = 1.73$ . Thus  $3 \cdot \sigma = 5.2 \approx 5$  and we select a range of  $x$  from 0 to  $3 + 5 = 10$ .

[done in Lecture 3 - get/read notes to draw by hand here.]

Clearly the normal approximation is not convincing; moreover on that scale, we should take into account that the binomial probabilities  $b_n(k)$  are only defined for integer values  $k$ , whereas we plotted the interpolating continuous function  $b_n(x)$  defined for all  $x \geq 0$ .

**Labwork 192** Write a MATLAB script to visualise the above figures drawn in lectures and algorithmically as well as visually understand *the success / failure rule*.

You need to use `gammaln` in MATLAB for  $\log_e(\Gamma)$  function and use laws of exponents and logarithms for the PDF of  $\text{Binomial}(n, \theta)$  RVs with large  $n$ .

**Exercise 7.4** Suppose the probability of having blue eyes is 0.15 for any given person in the U.S.. The town of Springfield, USA has 800 people. Suppose the residents of Springfield are all unrelated as they are immigrants arriving from all over the U.S., and therefore have eye colors that are independent of each other.

- a) Find the expected number of people with blue eyes in Springfield, USA.
- b) Find the variance and standard deviation of the number of blue-eyed people in Springfield, USA.
- c) Use the Normal approximation to the Binomial to calculate the probability that there are between 110 and 125 blue-eyed residents of Springfield, USA. Be sure to verify the success/failure condition for validity of the normal approximation.

### The Poisson approximation to the binomial

The breakdown of the normal approximation for small  $\theta$  is related to the **Poisson approximation** for  $\text{Binomial}(n, \theta)$ , according to which  $\text{Binomial}(n, \theta) \approx \text{Poisson}(n\theta)$  if  $\theta$  is small and  $n$  is large such that  $n\theta \rightarrow \lambda$ . Here  $\text{Poisson}(\lambda)$  is the Poisson distribution given by probabilities

$$P(Y = k) = \frac{\lambda^k}{k!} \exp(-\lambda), \quad k = 0, 1, \dots$$

**Proposition 104** Suppose  $\lambda > 0$  and consider the binomial law  $\text{Binomial}(n, \theta)$  for  $\theta = \lambda/n$ . Let  $p_{n,k}$  be the probability function of  $\text{Binomial}(n, \lambda/n)$  and let  $q_{n,k}$  be the probability function of the Poisson law  $\text{Poisson}(\lambda)$ . Then for  $n \rightarrow \infty$  and fixed  $\lambda$ ,

$$p_{n,k} \rightarrow q_{n,k} \text{ as } n \rightarrow \infty, \text{ for every } k = 0, 1, \dots$$

**Proof:** We have

$$\begin{aligned} p_{n,k} &= \frac{n!}{k!(n-k)!} \theta^k (1-\theta)^{n-k} \\ &= \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{1}{k!} \cdot \frac{n!}{(n-k)!n^k} \cdot \lambda^k \cdot \frac{1}{(1-\lambda/n)^k} \cdot \left(1 - \frac{\lambda}{n}\right)^n. \end{aligned} \tag{7.9}$$

For the second factor we have

$$\frac{n!}{(n-k)!n^k} = \frac{(n-k+1)}{n} \cdot \frac{(n-k+2)}{n} \cdot \dots \cdot \frac{n}{n}$$

i.e. it is a product of  $k$  factors each of which tends to 1, hence  $\frac{n!}{(n-k)!n^k} \rightarrow 1$ . For the 4th factor in (7.9) we obviously have

$$\frac{1}{(1 - \lambda/n)^k} \rightarrow 1.$$

For the 5th factor in (7.9) we have

$$\left(1 - \frac{\lambda}{n}\right)^n \rightarrow \exp(-\lambda)$$

which proves that

$$p_{n,k} \rightarrow \frac{1}{k!} \cdot \lambda^k \cdot \exp(-\lambda)$$

as  $n \rightarrow \infty$ , for a fixed  $k$ .

We have established the Poisson approximation to Binomial( $n, \theta$ ) if  $n\theta = \lambda$  exactly; a slight modification gives the result when  $\theta = \theta_n$  depends on  $n$  in such a way that  $n\theta_n \rightarrow \lambda$  as  $n \rightarrow \infty$ . This result is sometimes called the **law of small numbers** because it is the probability distribution of the number of occurrences of an event that happens rarely but has very many opportunities to happen. Another name is **law of rare events**.

Analogously to what we did for the binomial distribution, for visualization purposes we will interpolate the probability function by a smooth function

$$h(x) = \frac{1}{\Gamma(x+1)} \cdot \lambda^x \cdot \exp(-\lambda).$$

Then  $h(x)$  represents the Poisson law Poisson( $\lambda$ ) in the following sense:

$$h(k) = \frac{1}{k!} \cdot \lambda^k \cdot \exp(-\lambda), k = 0, 1, \dots$$

Set as before  $n = 600$ , and consider the binomial Binomial( $n, \theta$ ) for  $\theta = 1/200$ ; then our appropriate  $\lambda$  is  $\lambda = 3$ . The picture below is analogous to the last figure, where the dotted line now represents the Poisson law Poisson(3) instead of the normal law Normal( $\mu, \sigma^2$ ) with mean  $\mu = n\theta$  and variance  $n\theta(1 - \theta)$ .

[done in Lecuture 5 - get/read notes to draw by hand here.]

We see a visually perfect approximation of the binomial law by the Poisson law. Here both laws are discrete (concentrated on the integers  $0, 1, \dots$ ) and the picture represents a continuous interpolation.

The success/failure rule thus can be explained by the Poisson approximation, in the sense that the *Poisson approximation contradicts the normal for small  $\lambda = n\theta$* . In the requirement  $\lambda \geq 10$ , the 10 is a limit chosen by convention; in the figures we saw that at least for  $\lambda = 3$ , the Poisson and the normal curves are visually different. Can we argue that for  $\lambda \geq 10$ , the Poisson and the normal approximation are not in contradiction? For that we have to observe that Poisson( $\lambda$ ) approximates a normal distribution as  $\lambda \rightarrow \infty$ .

**Exercise 7.5 (Sum of independent Poisson RVs is a Poisson RV)** Suppose  $X, Y$  are independent RV's with  $\mathcal{L}(X) = \text{Poisson}(t), \mathcal{L}(Y) = \text{Poisson}(u)$  where  $t, u > 0$ . Then  $\mathcal{L}(X + Y) = \text{Poisson}(u + t)$ .

As a consequence, we can represent  $\text{Poisson}(n)$  as the law of a sum of IID RV's, each with law  $\text{Poisson}(1)$ :

$$\text{Poisson}(n) = \mathcal{L}(S_n), S_n = \sum_{i=1}^n Y_i, Y_i \text{ indep.}, \mathcal{L}(Y_i) = \text{Poisson}(1).$$

Therefore a CLT holds for the normalized  $S_n$ ; recall that expectation and variance of  $\text{Poisson}(\lambda)$  are both  $\lambda$ : if  $\mathcal{L}(Y) = \text{Poisson}(\lambda)$  then  $EY = \lambda, V(Y) = \lambda$ ; hence

$$E(S_n) = E S_n = n, V(S_n) = V S_n = n \quad (7.10)$$

$$\frac{S_n - n}{\sqrt{n}} \rightsquigarrow \text{Normal}(0, 1) \text{ by the CLT.} \quad (7.11)$$

We can express the latter relation also as

$$\text{Poisson}(n) \approx \text{Normal}(n, n) \text{ as } n \rightarrow \infty$$

where  $\approx$  means “closeness in distribution”, with a rigorous meaning (7.11).

**Remark 105** It can be verified that these limiting relations are also true for general  $\text{Poisson}(\lambda)$ :

$$\text{Poisson}(\lambda) \approx \text{Normal}(\lambda, \lambda) \text{ as } \lambda \rightarrow \infty$$

which is plausible (the limits holds for all sequences of  $\lambda_n \rightarrow \infty$ , not only for a limit along the integers  $1, 2, \dots$ ). This can be verified using the characteristic function (or moment generating function) of  $\text{Poisson}(\lambda)$ . An alternative argument is: represent  $\text{Poisson}(\lambda)$  as a sum  $X + Y$  where  $\lfloor \lambda \rfloor$  is the floor of  $\lambda$ , i.e., largest integer  $n \leq \lambda$ ,  $\mathcal{L}(X) = \text{Poisson}(\lfloor \lambda \rfloor)$  and  $\mathcal{L}(Y) = \text{Poisson}(\lambda - \lfloor \lambda \rfloor)$ , then show that  $X$  can be approximated by a normal and  $Y$  has negligible influence. We will acquire the tools for a rigorous argument later.

Let us try to illustrate the CLT for the Poisson law, using the tools we have. First we choose a small  $\lambda$ , e.g.  $\lambda = 1.5$  and plot the Poisson  $\text{Poisson}(\lambda)$  and the normal  $\text{Normal}(\lambda, \lambda)$ :

[done in Lecture 5 - get/read notes to draw by hand here.]

Consider the value  $\lambda = 10$  which is “borderline” according to the convention of the success / failure rule:

[done in Lecture 5 - get/read notes to draw by hand here.]

A value  $\lambda = 50$  gives

[done in Lecture 5 - get/read notes to draw by hand here.]

showing a nearly perfect fit again, as an illustration of the CLT in action for sums of Poisson variables.

**Labwork 193** Write a MATLAB script to visualise the above figures drawn in lectures and algorithmically as well as visually understand *the success / failure rule* with **law of rare events** or Poisson approximation.

You need to use `gammaln` in MATLAB for  $\log_e(\Gamma)$  function and use laws of exponents and logarithms for the factorial term in the PDF of  $\text{Binomial}(n, \theta)$  and  $\text{Poisson}(\lambda)$  RVs with large  $n$ .

### An example for use of normal approximation of the binomial

**Exercise 7.6** A multiple-choice examination has 100 questions, each with five possible answers of which only one is correct. Suppose a student just guesses at all the answers.

- (a) What is the probability that he or she gets exactly 2 out of the first 5 questions correct?
- (b) What is approximately the probability that he or she gets between 20 and 30 questions correct on the entire test?

**Exercise 7.7** Suppose the probability of having blue eyes is 0.15 for any given person in the U.S.. The town of Springfield, USA recently experienced economic downturn and most people lost their jobs. There are only 60 people now in Springfield. Suppose the residents of Springfield are all unrelated as they are immigrants arriving from all over the U.S., and therefore have eye colors that are independent of each other.

- a) Find the expected number of people with blue eyes in Springfield, USA.
- b) Find the variance and standard deviation of the number of blue-eyed people in Springfield, USA.
- c) Use either the Normal or Poisson approximation to the Binomial, whichever is most appropriate via the success / failure condition, to calculate the probability that there are between 20 and 25 blue-eyed residents of Springfield, USA. You may leave your answer as a simplified expression instead of a numerical value.

### The correction for continuity

Suppose that in the last example, we substitute question (b) “between 20 and 30 questions correct” by “between 21 and 30 questions correct”, that is, we are asking for the approximate probability

$$P(a+1 \leq X \leq b)$$

with  $a = 20$ ,  $b = 30$  or equivalently, for an approximation to

$$P(a < X \leq b)$$

(since  $X$  is a binomial RV, taking only integer values). By the reasoning above, we find the Z-score of  $a+1$ :

$$\frac{a+1-\mu}{\sigma} = \frac{21-20}{\sigma} = \frac{1}{4} = 0.25.$$

The approximate probability is then

$$P(a+1 \leq X \leq b) \approx P\left(\frac{a+1-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right)$$

This will yield a different value from our previous approximation since the lower Z-score is now  $\frac{a+1-\mu}{\sigma}$  instead of  $\frac{a-\mu}{\sigma}$ ; the difference can be described by

$$\begin{aligned} & P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) - P\left(\frac{a+1-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{a-\mu}{\sigma} + \frac{1}{\sigma}\right). \end{aligned}$$

Recall that  $\sigma = \sqrt{np(1-p)}$  here, so that  $1/\sigma$  is small. At the same time, this difference approximates the difference

$$P(a < X \leq b) - P(a+1 \leq X \leq b) = P(X = a)$$

for  $a = 20$ , so that it turns out that we approximate an individual probability  $P(X = a)$  for a binomial  $X$  by (setting  $z(a) = \frac{a-\mu}{\sigma}$ )

$$P(X = a) \approx \int_{z(a)}^{z(a)+1/\sigma} \varphi(t) dt. \quad (7.12)$$

where  $1/\sigma$  is small. The right hand side is an integral of the standard normal density over a small piece of size  $1/\sigma$ . Extending this reasoning to all integers between  $a = 20$  and  $b = 30$ , we obtain

$$P(a \leq X \leq b) = \sum_{k=0}^{10} P(X = a+k)$$

so we have 11 individual probabilities to approximate. On the other hand, our original approximation was

$$P(a \leq X \leq b) \approx \int_{z(a)}^{z(b)} \varphi(t) dt. \quad (7.13)$$

and noting  $z(a) + 10/\sigma = z(b)$ , we note that the interval  $(z(a), z(b))$  contains only 10 of the small intervals of length  $1/\sigma$ . So there is a slight inconsistency in our method.

This can be corrected by replacing the approximation (7.12), that is, integration over the interval  $(z(a), z(a) + 1/\sigma)$ , by integration over a symmetric interval around  $z(a)$  having the same length:  $(z(a) - 1/2\sigma, z(a) + 1/2\sigma)$ , so that we now use

$$P(X = a) \approx \int_{z(a)-1/2\sigma}^{z(a)+1/2\sigma} \varphi(t) dt. \quad (7.14)$$

Since

$$\begin{aligned} z(a) - 1/2\sigma &= \frac{a - \mu}{\sigma} - \frac{1}{2\sigma} = \frac{a - 1/2 - \mu}{\sigma} = z(a - 1/2) \\ z(a) + 1/2\sigma &= z(a + 1/2), \end{aligned}$$

we may write (7.14) as

$$P(X = a) \approx \int_{z(a-1/2)}^{z(a+1/2)} \varphi(t) dt. \quad (7.15)$$

The principle of using (7.15) for individual binomial probabilities is known as the *correction for continuity*. It implies that, for  $P(a \leq X \leq b)$ , we now use an approximation

$$P(a \leq X \leq b) \approx \int_{z(a-1/2)}^{z(b+1/2)} \varphi(t) dt \quad (7.16)$$

rather than (7.13). This corrected method is now consistent in itself, that is, the left side splits up into 11 individual probabilities  $P(X = k)$ :

$$P(a \leq X \leq b) = \sum_{k=0}^{10} P(X = a + k),$$

and similarly the right side in (7.16) splits into 11 integrals over small pieces of length  $1/\sigma$ .

Whether we use the continuity correction or not, the individual binomial probabilities will be small for large  $n$ : for some  $t^*$

$$\begin{aligned} P(X = a) &\approx \int_{z(a)-1/2\sigma}^{z(a)+1/2\sigma} \varphi(t) dt = \frac{1}{\sigma} \varphi(t^*) = \frac{1}{\sqrt{np(1-p)}} \varphi(t^*) \\ &\leq \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{2\pi p(1-p)}}, \end{aligned}$$

for some  $t^*$  with  $z(a) - 1/2\sigma < t^* < z(a) + 1/2\sigma$ , and in view of

$$\varphi(t^*) = \frac{1}{\sqrt{2\pi}} \exp(-t^*/2) \leq \frac{1}{\sqrt{2\pi}}.$$

Thus the individual binomial probabilities will decrease like  $1/\sqrt{n}$  as  $n \rightarrow \infty$  at most (or will be even smaller).

### The normal table and quantiles

The so-called  $Z$ -table gives the areas under the standard normal curve

$$P(Z \leq z) = \int_{-\infty}^z \varphi(t) dt.$$

It is well known that the normal integral

$$\int_{-\infty}^z \varphi(t) dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp(-t^2/2) dt$$

does not have an explicit analytic solution. Hence the necessity to use a numerically computed table; statistical software has these values stored (or sometimes computes them).

As an example, suppose  $z = -1.26$ . First, look up the value of  $z$  without the second decimal place, i.e.  $-1.2$ , in one of the two columns headed “ $z$ ”. To the left of  $-1.2$  you find all the probabilities  $P(Z \leq -1.2 - s)$  for  $s = 0.00, \dots, 0.09$ , i.e. for values of the second decimal place. In our example we have  $s = 0.06$ , so we go to the column headed “ $0.06$ ” and find  $P(Z \leq -1.26) = 0.1038$ .

Consider a positive value of  $z$ , e.g.  $z = 2.73$ . First, look up the value of  $z$  without the second decimal place, i.e.  $2.7$ , in one of the two columns headed “ $z$ ”. To the right of  $2.73$  you find all the probabilities  $P(Z \leq 2.7 + s)$  for  $s = 0.00, \dots, 0.09$ , i.e. for values of the second decimal place. In our example we have  $s = 0.03$ , so we go to the column headed “ $0.03$ ” and find  $P(Z \leq 2.73) = 0.9968$ .

Useful rules to find other normal probabilities are

$$P(Z < z) = P(Z \leq z)$$

(since for a continuous RV having a density,  $P(Z = z) = 0$ ),

$$\begin{aligned} P(Z \geq z) &= 1 - P(Z \leq z) \\ &= P(Z \leq -z) \end{aligned}$$

The last two equalities give two ways of looking up  $P(Z \geq z)$  in the table. The second equality derives from the symmetry of the normal density: since  $\varphi(t) = \varphi(-t)$ , we have

$$\begin{aligned} P(Z \geq z) &= \int_z^\infty \varphi(t) dt = \lim_{x \rightarrow \infty} \int_z^x \varphi(t) dt = \lim_{x \rightarrow \infty} \int_{-x}^{-z} \varphi(t) dt \\ &= \int_{-\infty}^{-z} \varphi(t) dt = P(Z \leq -z). \end{aligned}$$

It also follows that  $P(Z \leq 0) = P(Z \geq 0) = 1/2$ .

The table begins at  $z = -3.9$  and ends at  $z = 3.9$ , for which the probabilities are given as 0.0000 and 1.0000 respectively, i.e. they are given up to 4 decimal places. There is a famous rule for the practitioner giving the probability content of certain intervals around 0.

**The 68-95-99.7 rule** (De Moivre, 1733). *In a normal model  $\text{Normal}(\mu, \sigma^2)$ , about 68% of the values fall within one standard deviation of the mean, about 95% of the values fall within two standard deviations of the mean, and about 99.7% of the values fall within three standard deviations of the mean.*

The last part (99.7 part) is also called the  **$3\sigma$ -rule**. Let us verify these claims, using the  $Z$ -table. Suppose  $\mathcal{L}(X) = \text{Normal}(\mu, \sigma^2)$ ; then for  $k = 1, 2, 3$

$$\begin{aligned} P(\mu - k\sigma \leq X \leq \mu + k\sigma) &= P\left(-k \leq \frac{X - \mu}{\sigma} \leq k\right) = P(-k \leq Z \leq k) \\ &= P(Z \leq k) - P(Z \leq -k). \end{aligned}$$

For  $k = 1, 2, 3$  we find

$$\begin{aligned} P(Z \leq 1) - P(Z \leq -1) &= 0.8413 - 0.1568 = 0.6845, \\ P(Z \leq 2) - P(Z \leq -2) &= 0.9772 - 0.0228 = 0.9544, \\ P(Z \leq 3) - P(Z \leq -3) &= 0.9987 - 0.0013 = 0.9974. \end{aligned}$$

The rule (as an approximation statement) is confirmed.

**Reverse lookup and quantiles.** It is often of interest to find a value of  $z$  which matches a certain probability, say  $\alpha$  ( $0 < \alpha < 1$ ), such that, if  $X$  is a RV

$$P(X \geq z) = \alpha.$$

In this case we write  $z = z_\alpha$  and call this the **upper  $\alpha$ -quantile** of the distribution of  $X$ . Note that if  $X$  is  $\text{Normal}(\mu, \sigma^2)$  then  $z_\alpha$  is uniquely defined: the equation

$$\int_z^\infty \varphi_{\mu, \sigma^2}(t) dt = \alpha$$

has a unique solution in  $z$ , since the left side is continuous, strictly monotone decreasing in  $z$  and ranges between 0 and 1. Here  $\varphi_{\mu, \sigma^2}(t) = \varphi((t - \mu)/\sigma)/\sigma$  is the density of  $\text{Normal}(\mu, \sigma^2)$ ; since this density is strictly positive everywhere, we find

$$\frac{d}{dz} \int_z^\infty \varphi_{\mu, \sigma^2}(t) dt = -\varphi_{\mu, \sigma^2}(z) < 0$$

and indeed  $P(X \geq z)$  is strictly decreasing in  $z$ . For  $\alpha = 0.25$  the  $z_\alpha$  called **the upper quartile** of the distribution; for  $\alpha = 0.75$  the  $z_\alpha$  called **the lower quartile**. For  $\alpha = 0.5$  we obtain the **median** of distribution of  $X$ ; for  $\text{Normal}(\mu, \sigma^2)$  it coincides with the mean  $\mu$ . We can also define **lower  $\alpha$ -quantiles** by solving  $P(X \leq z) = \alpha$  for  $z$ .

**Example 194** Suppose that Verbal SAT test scores  $X$  are described by a normal curve, for which the mean is 500 and the standard deviation is 100. A student's score is better than 75% of all the scores. What is the student's score?

**Solution.** We are asked the upper 25%-quantile (or 0.25-quantile, or the upper quartile) of  $\text{Normal}(\mu, \sigma^2)$  with  $\mu = 500$  and  $\sigma = 100$ . Call this  $s$  now, i.e.  $s$  is the student's score. We must have

$$P(X \geq s) = 0.25$$

hence

$$P(X \leq s) = 0.75 = P\left(\frac{X - \mu}{\sigma} \leq \frac{s - \mu}{\sigma}\right) = P\left(Z \leq \frac{s - \mu}{\sigma}\right).$$

Let  $s^* = (s - \mu)/\sigma$ , then

$$P(Z \leq s^*) = 0.75$$

i.e.  $s^* = z_{1/4}^*$  is the upper quartile of  $\text{Normal}(0, 1)$ . By “reverse lookup” in the  $Z$ -table, we find the two closest to 0.75 entries (probabilities) to be 0.7486 and 0.7517, corresponding to  $z$ -values 0.67 and 0.68 respectively (i.e.  $P(Z \leq 0.67) = 0.7486$ ). A common method now is to interpolate between these two values of  $z$ , which would give us  $z_{1/4}^* = s^* = 0.675$ . This gives a value for  $s$

$$s = \sigma s^* + \mu = 67.5 + 500 = 567.5.$$

The method can be summarized: if  $z_\alpha$  denotes the  $\alpha$ -quantile of  $\text{Normal}(0, 1)$  and  $x_\alpha^*$  denotes the  $\alpha$ -quantile of  $\text{Normal}(\mu, \sigma^2)$  then  $x_\alpha^* = \sigma z_\alpha + \mu$

### Drawing the normal curve

Below we are plotting the normal density with mean  $\mu = 5$  and standard deviation  $\sigma = 5$ .

We see that at  $x = 0$  the curve changes curvature, i. e. left of 0 it is convex (downward bent) and right of 0 it is concave (upward bent). Such a point is called an **inflection point**. We see that at  $x = 10$  there is another inflection point, and both inflection points are one standard deviation away from the mean. We will show that this a general feature of any normal distribution  $\text{Normal}(\mu, \sigma^2)$ .

To see this, note that a smooth function  $f$  (which has at least 2 derivatives) is convex at  $x$  if  $f'$  is increasing at  $x$  (strictly increasing, say), which means  $f''(x) > 0$ . Similarly,  $f$  is concave at  $x$  if  $f''(x) < 0$ . The density of  $\text{Normal}(\mu, \sigma^2)$  is

$$f(x) = \frac{1}{\sigma} \varphi\left(\frac{x-\mu}{\sigma}\right)$$

where  $\varphi(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$  is the density of  $\text{Normal}(0, 1)$ . Now

$$\begin{aligned} f'(x) &= \frac{1}{\sigma^2} \varphi'\left(\frac{x-\mu}{\sigma}\right), \\ f''(x) &= \frac{1}{\sigma^3} \varphi''\left(\frac{x-\mu}{\sigma}\right) \end{aligned} \tag{7.17}$$

and

$$\begin{aligned} \varphi'(t) &= \frac{d}{dt} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) = -\frac{1}{\sqrt{2\pi}} \exp(-t^2/2) \cdot t, \\ \varphi''(t) &= \frac{1}{\sqrt{2\pi}} (\exp(-t^2/2) \cdot t^2 - \exp(-t^2/2)) \\ &= \varphi(t) \cdot (t^2 - 1). \end{aligned}$$

It follows that  $\varphi''(t) < 0$  for  $|t| < 1$  and  $\varphi''(t) > 0$  for  $|t| > 1$ , hence the inflection points of  $\varphi$  are  $-1$  and  $1$ . From (7.17) it follows that  $f''(x) < 0$  if  $\left|\frac{x-\mu}{\sigma}\right| < 1$  etc, so that the inflection points of  $f$  are at  $x = \mu \pm \sigma$ .

### Chebyshev's inequality and the normal tail

A *tail estimate* for a RV  $X$  is an estimate for the probability  $P(|X| > t)$ . We may compare the tail estimates obtained from the standard normal  $Z$  with those from the Chebyshev inequality. The latter tells us that for any RV  $X$  with  $E X = 0$  and  $V(X) = 1$  and any  $t > 0$

$$P(|X| > t) \leq t^{-2}$$

which for  $t = 1$  gives 1 (i.e. it is trivial), for  $t = 2$  it gives 0.25 and for  $t = 3$  it gives  $1/9 = 0.11$ . For the standard normal we obtain the corresponding  $P(|Z| > t)$  from the table (or approximately from the  $3\sigma$ -rule) as 0.3155 for  $t = 1$ , 0.0456 for  $t = 2$  and 0.0026 for  $t = 3$ . Thus the normal tail decreases much faster than  $1/t^2$ , the upper bound from the Chebyshev inequality. This is not surprising in view of the form of the normal density: we have

$$P(|Z| > t) = 2 \int_t^\infty \varphi(u) du = \sqrt{\frac{2}{\pi}} \int_t^\infty \exp(-u^2/2) du$$

and we would expect that the integral decreases with a similarly fast rate as the density  $\varphi$  itself, as the lower bound  $t$  tends to infinity. This is made precise by the following result.

**Proposition 106 (Mill's inequality)** Let  $\mathcal{L}(Z) = \text{Normal}(0, 1)$ . Then

$$P(|Z| > t) \leq \sqrt{\frac{2}{\pi}} \frac{1}{t} \exp(-t^2/2). \quad (7.18)$$

**Proof:** Observe that  $u/t \geq 1$  for  $u \geq t$ , hence

$$\begin{aligned} P(|Z| > t) &\leq \sqrt{\frac{2}{\pi}} \frac{1}{t} \int_t^\infty u \exp(-u^2/2) du \\ &= \sqrt{\frac{2}{\pi}} \frac{1}{t} \lim_{x \rightarrow \infty} \int_t^x u \exp(-u^2/2) du \\ &= \sqrt{\frac{2}{\pi}} \frac{1}{t} \lim_{x \rightarrow \infty} [-\exp(-u^2/2)]_t^x \\ &= \sqrt{\frac{2}{\pi}} \frac{1}{t} \exp(-t^2/2). \end{aligned}$$

Suppose again that  $S_n = \sum_{i=1}^n X_i$  where  $X_i$  are IID Bernoulli  $\text{Bernoulli}(p)$ . The Chebyshev inequality gives for a tail probability

$$P\left(\left|\frac{S_n - np}{\sqrt{np(1-p)}}\right| > t\right) \leq 1/t^2$$

and it is *exact* (holds for every  $n$ ). In contrast, the normal tail estimate holds only as a limiting result for large  $n$ , i.e.  $P(|\cdot| > t) \rightarrow P(|Z| > t)$ . But then the upper bound on  $P(|\cdot| > t)$  suggested is much smaller than the one of Chebyshev's inequality, as shown by Mill's inequality. Obviously, for applications a choice is to be made. Probability estimates which hold only as limits for  $n \rightarrow \infty$  are called *asymptotic*. Much of the basic statistical methods to be discussed (confidence intervals, tests) are asymptotic in this sense, based on an assumption that  $n$  is large enough.

**Exercise 7.8** Show that Mill's inequality is sharp in the following sense: as  $t \rightarrow \infty$ , the ratio of the left and right sides of (7.18) tends to one. A common notation for this is:

$$P(|Z| > t) \sim \sqrt{\frac{2}{\pi}} \frac{1}{t} \exp(-t^2/2) \text{ as } t \rightarrow \infty$$

where the symbol " $\sim$ " applied to two functions  $g(t), h(t)$  means that  $g(t)/h(t) \rightarrow 1$  as  $t$  tends to a limit ( $t \rightarrow \infty$  in our case).

### 7.5.4 Confidence intervals for a proportion

#### Basic reasoning for a normal mean

Suppose that a random variable  $X$  has a distribution  $\text{Normal}(\mu, 1)$ , i.e. it can be written  $X = \mu + Z$ . Suppose further that we do not know  $\mu$ , but we observe  $X$  (*one observation only*, i.e. we obtain one realization of  $X$ ). What statements can be made about the unknown  $\mu$ ?

From the  $3\sigma$ -rule we know that with probability 99.7%,  $Z$  falls within a distance 3 from 0. Consequently,  $X$  falls within a distance 3 from  $\mu$ , with the same probability 99.7%. Now we have observed  $X$ , and we know that

$$P(|X - \mu| \leq 3) = 0.997$$

which can equivalently be expressed as: “the interval  $[X - 3, X + 3]$  covers  $\mu$  with probability 99.7%” or formally

$$P([X - 3, X + 3] \ni \mu) = 0.997. \quad (7.19)$$

Here “ $\ni$ ” is the inverted “element of” sign  $\in$  which should be read “the interval contains” or “the interval covers”. Of course we could have written  $\mu \in [X - 3, X + 3]$ , but to stress the fact that *the interval is random, not  $\mu$* , we write  $[X - 3, X + 3] \ni \mu$ .

Suppose that  $X$  has been observed and takes the value  $x$ . Then, based on (7.19), the interval  $[x - 3, x + 3]$  is called a *confidence interval* and the probability 0.997 is called the *confidence level*, usually denoted by  $C$ . As an example, assume  $x = 2$  was observed. Then  $[-1, 5]$  is a confidence interval of level  $C = 99.7$  percent. Based on the probability estimate (7.19), the statement usually associated to the interval is: “we are 99.7 % confident that  $[-1, 5]$  covers the unknown mean  $\mu$ ”.

A confidence statement like this is not the same as a probability statement: it is not claimed that, after  $x$  is already observed, that “the probability that the interval  $[-1, 5]$  covers the true mean  $\mu$  is 99.7%”. Indeed after  $X$  took the value  $x = 2$ , there is no randomness left, when we assume that  $\mu$  is merely unknown, but not random. What can be said about the interval  $[-1, 5]$  is a *confidence statement*, not a probability statement. This is based on the probability statement: in 99.7% of all cases, the interval obtained by this method covers the true parameter - formally expressed as (7.19) where  $X$  is random. When  $X = x$  is realized, i.e. no longer random, the confidence statement about  $[-1, 5]$  is derived “in hindsight” from (7.19).

The confidence level  $C = 99.7\%$  is not a commonly used value; these are 99% and 95%. Let us find the corresponding confidence intervals for  $\mu$ , based on  $X \sim \text{Normal}(\mu, 1)$ <sup>2</sup> We have to solve

$$P(|X - \mu| \leq z) = C$$

for  $z > 0$ , upon which  $[X - z, X + z]$  will be a level  $C$  confidence interval. Since  $X - \mu \sim Z$ , we have

$$\begin{aligned} P(|Z| \leq z) &= 1 - 2P(Z > z) = C, \\ P(Z > z) &= (1 - C)/2 =: \alpha. \end{aligned}$$

Thus  $z = z_\alpha$ , the upper  $\alpha$ -quantile of  $Z$  for  $\alpha = (1 - C)/2$ . The commonly used values are easily found from the table, using  $P(Z > z) = P(Z < -z)$ :

$$\begin{aligned} C &= 99\%, \alpha = 0.005, z_\alpha = 2.578 \\ C &= 95\%, \alpha = 0.025, z_\alpha = 1.96. \end{aligned}$$

---

<sup>2</sup>The notation  $X \sim \text{Normal}(\mu, 1)$  is a commonly used equivalent for  $\mathcal{L}(X) = \text{Normal}(\mu, 1)$ . Similarly,  $X \sim Y$  will be used for  $\mathcal{L}(X) = \mathcal{L}(Y)$ . This usage should not be confused with the one for nonrandom sequences  $x_n, y_n$ , where the symbol  $x_n \sim y_n$  means  $x_n/y_n \rightarrow 1$ .

Recall the second part of the 68-95-99.7% rule: there it was claimed that  $P(|Z| \leq 2) \approx 95\%$ ; we just found the corresponding quantile more accurately: it is not 2 but 1.96.

The idea of the confidence interval for the unknown mean  $\mu$  can easily be extended to the case where we observe (one)  $X \sim \text{Normal}(\mu, \sigma^2)$  provided  $\sigma^2$  is known. Let  $C$  be the confidence level and  $\alpha = (1 - C)/2$ ; then  $(X - \mu)/\sigma \sim Z$  and

$$\begin{aligned} C &= P(|Z| \leq z_\alpha) = P\left(\left|\frac{X - \mu}{\sigma}\right| \leq z_\alpha\right) \\ &= P(|X - \mu| \leq \sigma z_\alpha) = P([X - \sigma z_\alpha, X + \sigma z_\alpha] \ni \mu). \end{aligned}$$

**Proposition 107** Suppose a RV  $X \sim \text{Normal}(\mu, \sigma^2)$  is observed where  $\mu$  is unknown and  $\sigma > 0$  is known. Let  $z_\alpha$  be the upper  $\alpha$ -quantile of  $Z$  for some  $0 < \alpha < 1/2$ . Then for  $C = 1 - 2\alpha$ .

$$P([X - \sigma z_\alpha, X + \sigma z_\alpha] \ni \mu) = C,$$

which means that for any observed value  $X = x$ , the interval  $[x - \sigma z_\alpha, x + \sigma z_\alpha]$  is a confidence interval for  $\mu$  of level  $C$ .

In what follows, we will generally abbreviate the statement: “*for any observed value  $X = x$ , the interval  $[x - \sigma z_\alpha, x + \sigma z_\alpha]$  is a confidence interval for..*” by: “*the interval  $[X - \sigma z_\alpha, X + \sigma z_\alpha]$  is a confidence interval for ....*”

**Parameters and statistical inference.** We assumed initially that we observe  $X \sim \text{Normal}(\mu, 1)$  (or equivalently an  $X$  with  $\mathcal{L}(X) = \text{Normal}(\mu, 1)$ ) where the mean  $\mu$  is unknown. In that context  $\mu$  is called a *parameter* of the distribution of  $X$ . Constructing a confidence interval for  $\mu$  is an example of *statistical inference* regarding a parameter. Other examples of inference are hypothesis tests (about a parameter) and estimation of a parameter, also called *point estimation*. In point estimation one just gives a “reasonable guess” of a parameter. Note that if  $X \sim \text{Normal}(\mu, 1)$ , then  $X$  itself is a reasonable guess of  $\mu$ , i.e. a point estimate. In contrast, a confidence interval gives a range and an attached probability statement; a confidence interval is also called an *interval estimate*. In the case  $X \sim \text{Normal}(\mu, \sigma^2)$ , both  $\mu$  and  $\sigma^2$  may be parameters; above we assumed  $\sigma$  “known”, i.e. we constructed a confidence interval which made use of  $\sigma$  (namely  $[X - \sigma z_\alpha, X + \sigma z_\alpha]$ ). When  $\sigma$  is unknown, this interval is not available.

**The case of  $n$  IID normal observations.** The model of *one* normal observation  $X$  appears artificial; it may strike one as a situation with very little data indeed. Consider instead the case of independent observations  $X_1, \dots, X_n$  all with law  $\text{Normal}(\mu, \sigma^2)$ , and as above assume  $\mu$  is unknown while  $\sigma$  is known. To construct a confidence interval for  $\mu$ , one may choose to take the sample mean  $\bar{X}_n$  first and then build an interval estimate using information about the law of  $\bar{X}_n$ . Indeed we have from basic properties of the normal law

$$\bar{X}_n \sim \text{Normal}(\mu, n^{-1}\sigma^2)$$

(this follows from the fact that the sum of independent normals is normal, and a mean and variance computation). As a reminder, let’s compute the variance:

$$\begin{aligned} V(\bar{X}_n) &= V\left(n^{-1} \sum_{i=1}^n X_i\right) = n^{-2} V\left(\sum_{i=1}^n X_i\right) =_{(\text{by independence})} n^{-2} \sum_{i=1}^n V(X_i) \\ &= n^{-2} \sum_{i=1}^n \sigma^2 = n^{-1}\sigma^2. \end{aligned}$$

We immediately obtain a confidence interval for  $\mu$ , by treating  $\bar{X}_n$  as “one normal observation” with mean  $\mu$  and variance  $\tau^2 = n^{-1}\sigma^2$ . We need only apply the previous Proposition 107 setting the standard deviation  $\tau = \sigma/\sqrt{n}$ :

**Proposition 108** Suppose independent observations  $X_1, \dots, X_n$  with  $X_i \sim \text{Normal}(\mu, \sigma^2)$  where  $\mu$  is unknown and  $\sigma > 0$  is known. Let  $\bar{X}_n$  be the sample mean and let  $z_\alpha^*$  be the upper  $\alpha$ -quantile of  $Z$  for some  $0 < \alpha < 1/2$ . Then  $[\bar{X}_n - \sigma z_\alpha^*/\sqrt{n}, \bar{X}_n + \sigma z_\alpha^*/\sqrt{n}]$  is a confidence interval for  $\mu$  of level  $C = 1 - 2\alpha$ .

### Asymptotics for the sample proportion

Let us return to the case of  $n$  IID Bernoulli observations  $X_1, \dots, X_n$ ,  $X_i \sim \text{Bernoulli}(\theta^*)$  where  $\theta^* \in (0, 1)$  is unknown. Our goal is to construct a confidence interval of level  $C$  for the unknown population proportion  $\theta^*$ . The starting point is the normal approximation for the *point estimator*  $\hat{\Theta}_n$  based on the sample proportion:

$$\hat{\Theta}_n = \bar{X}_n = n^{-1} \sum_{i=1}^n X_i \quad \text{and} \quad T_n := \frac{\hat{\Theta}_n - \theta^*}{\sqrt{\theta^*(1-\theta^*)}/\sqrt{n}} \rightsquigarrow Z \quad (7.20)$$

where  $Z \sim \text{Normal}(0, 1)$ .

**Lemma 109** For  $\alpha = (1 - C)/2$  and  $z_\alpha$  such that  $P(Z > z_\alpha) = \alpha$  we have

$$P\left(-z_\alpha \leq \frac{\hat{\Theta}_n - \theta^*}{\sqrt{\theta^*(1-\theta^*)}/\sqrt{n}} \leq z_\alpha\right) \rightarrow P(-z_\alpha^* \leq Z \leq z_\alpha) = C \text{ as } n \rightarrow \infty.$$

**Proof:** Indeed we have

$$P(-z_\alpha \leq T_n \leq z_\alpha) = P(T_n \leq z_\alpha) - P(T_n < -z_\alpha)$$

Now  $P(T_n \leq t)$  is the distribution function of  $T_n$  at  $t$ ; by convergence in distribution  $T_n \rightsquigarrow Z$ ,  $P(T_n \leq t)$  tends to  $P(Z \leq t) = \Phi(t)$  for every  $t$  (since every  $t$  is a continuity point of  $\Phi$ ). We also claim that for every  $t$

$$P(T_n < t) \rightarrow \Phi(t). \quad (7.21)$$

(the distribution function of  $T_n$  may have a jump at  $t$ , but its size tends to 0 as  $n \rightarrow \infty$ ). Indeed for every  $\varepsilon > 0$

$$P(T_n \leq t - \varepsilon) \leq P(T_n < t) \leq P(T_n \leq t + \varepsilon)$$

and  $P(T_n \leq t - \varepsilon) \rightarrow \Phi(t - \varepsilon)$ ,  $P(T_n \leq t + \varepsilon) \rightarrow \Phi(t + \varepsilon)$  and  $|\Phi(t + \varepsilon) - \Phi(t - \varepsilon)|$  can be made arbitrarily small by a choice of  $\varepsilon$ . Hence (7.21) is shown. Setting  $t = z_\alpha$  and  $t = -z_\alpha$  we obtain

$$P(T_n \leq z_\alpha) - P(T_n < -z_\alpha^*) \rightarrow P(Z \leq z_\alpha) - P(Z \leq -z_\alpha) = P(-z_\alpha^* \leq Z \leq z_\alpha).$$

We may write the claim of the lemma as

$$P\left(-z_\alpha \sqrt{\theta^*(1-\theta^*)}/\sqrt{n} \leq \theta^* - \hat{\Theta}_n \leq z_\alpha \sqrt{\theta^*(1-\theta^*)}/\sqrt{n}\right) \quad (7.22)$$

$$= P\left(\hat{\Theta}_n - z_\alpha \sqrt{\theta^*(1-\theta^*)}/\sqrt{n} \leq \theta^* \leq \hat{\Theta}_n + z_\alpha \sqrt{\theta^*(1-\theta^*)}/\sqrt{n}\right) \rightarrow C \quad (7.23)$$

and we are close to an approximate confidence interval for the unknown  $\theta^*$ , except for the fact that  $\theta^*(1 - \theta^*)$  is unknown, hence the upper and lower bounds of the interval cannot be used. There are several methods to overcome this difficulty as shown below.

**a). Standard Confidence Interval of asymptotic level  $C$ :** We can use  $\hat{\theta}_n$  as a point estimate for the unknown  $\theta^*$  in the interval bounds, i.e. we set

$$m = z_\alpha \sqrt{\frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}}$$

use the interval

$$[\hat{\theta}_n - m, \hat{\theta}_n + m] \quad (7.24)$$

as an approximate confidence interval of level  $C$ . Indeed  $\hat{\theta}_n$  is a reasonable estimate of  $\theta^*$  since  $\hat{\Theta}_n \rightarrow_P \theta^*$  by the LLN and also  $E\hat{\Theta}_n = \theta^*$  (i.e.,  $\hat{\Theta}_n$  is an *unbiased* estimator). But then we must establish a convergence result as in (7.23), more precisely

$$\begin{aligned} & P(\hat{\Theta}_n - m \leq \theta^* \leq \hat{\Theta}_n + m) \\ &= P\left(-z_\alpha \leq \frac{\hat{\Theta}_n - \theta^*}{\sqrt{\hat{\Theta}_n(1 - \hat{\Theta}_n)/\sqrt{n}}} \leq z_\alpha\right) \rightarrow C. \end{aligned} \quad (7.25)$$

That should be possible, given that

$$\frac{\hat{\Theta}_n - \theta^*}{\sqrt{\hat{\Theta}_n(1 - \hat{\Theta}_n)/\sqrt{n}}} = T_n \cdot \sqrt{\frac{\theta^*(1 - \theta^*)}{\hat{\Theta}_n(1 - \hat{\Theta}_n)}}$$

and the fact that  $\hat{\Theta}_n \rightarrow_P \theta^*$ , hence  $\theta^*/\hat{\Theta}_n \rightarrow_P 1$ . Below we will establish (7.25), showing that the interval (7.24) is an approximate confidence interval of level  $C$  for  $\theta^*$ . A common synonym for “approximate as  $n \rightarrow \infty$ ” is “asymptotic”; the interval (7.24) is in fact the **standard confidence interval of asymptotic level  $C$**  for the population proportion. The bound  $m$  in the interval  $[\hat{\Theta}_n - m, \hat{\Theta}_n + m]$  is called the *margin of error*; the width of the interval is  $2m$ .

**b) Conservative confidence interval of asymptotic level  $C$ :** We use the inequality

$$\theta^*(1 - \theta^*) \leq \frac{1}{4}$$

(Exercise !) to replace  $\sqrt{\theta^*(1 - \theta^*)}$  by  $1/2$  in (7.23) and thus we work with a wider interval around  $\hat{\theta}_n$  which increases coverage probability of  $\theta^*$ , and thus should also have asymptotic coverage probability *at least*  $C$ . In more detail: set  $m = z_\alpha/2\sqrt{n}$ ; since  $\sqrt{\theta^*(1 - \theta^*)} \leq 1/2$ , we have  $z_\alpha\sqrt{\theta^*(1 - \theta^*)}/\sqrt{n} \leq m$  and hence

$$\begin{aligned} & P(\hat{\Theta}_n - m \leq \theta^* \leq \hat{\Theta}_n + m) \\ &\geq P\left(\hat{\Theta}_n - z_\alpha\sqrt{\theta^*(1 - \theta^*)}/\sqrt{n} \leq \theta^* \leq \hat{\Theta}_n + z_\alpha\sqrt{\theta^*(1 - \theta^*)}/\sqrt{n}\right) \rightarrow C. \end{aligned}$$

hence

$$\liminf_{n \rightarrow \infty} P\left(\hat{\Theta}_n - z_\alpha \frac{1}{2\sqrt{n}} \leq \theta^* \leq \hat{\Theta}_n + z_\alpha \frac{1}{2\sqrt{n}}\right) \geq C. \quad (7.26)$$

So there is also an asymptotic coverage probability of at least  $C$ . This method is known as the **conservative method** for an asymptotic confidence interval of level  $C$ . The conservative margin of error is  $m = z_\alpha/2\sqrt{n}$ .

**c) Exact Chebyshev's confidence interval with coverage probability at least  $C$ :** We use *Chebyshev's inequality* applied to the sample proportion (with  $V(\hat{\Theta}_n) = \theta^*(1 - \theta^*)/n$ )

$$P\left(\left|\hat{\Theta}_n - \theta^*\right| \geq m\right) \leq \frac{\theta^*(1 - \theta^*)}{nm^2} \leq \frac{1}{4nm^2},$$

set  $C = 1 - 1/(4nm^2)$  and solve for  $m$ , which gives  $m = \sqrt{\frac{1}{4n(1-C)}}$  and for the interval  $[\hat{\Theta}_n - m, \hat{\Theta}_n + m]$  a confidence statement

$$P\left([\hat{\Theta}_n - m, \hat{\Theta}_n + m] \ni \theta^*\right) \geq C.$$

This interval has nonasymptotic (or *exact*) coverage probability at least  $C$ , and in that sense it is preferable to an asymptotic confidence interval. But the margin of error is larger: if we compare it with the conservative (asymptotic) margin of error  $z_\alpha \frac{1}{2\sqrt{n}}$ , we notice that both are of order  $1/\sqrt{n}$ , but their ratio

$$\frac{\sqrt{\frac{1}{4n(1-C)}}}{z_\alpha \frac{1}{2\sqrt{n}}} = \frac{1/\sqrt{1-C}}{z_\alpha} = \frac{1/\sqrt{2\alpha}}{z_\alpha} \quad (7.27)$$

is large. This is clear from the fact that the numerator is the solution of  $1/2x^2 = \alpha$  and the denominator is the solution of  $P(Z > x) = \alpha$ , and from what we discussed above about the normal tail. From the table it can be seen that for  $C = 0.95$  we have  $z_\alpha = 1.96$ ; then the above ratio is 2.28 and for  $C = 0.99$  the ratio 3.88.

**Exercise 7.9** Use Mill's inequality to formally show that (7.27) tends to infinity as  $\alpha \rightarrow 0$  (i.e. as  $C \rightarrow 1$ ).

**Exercise 7.10** Show that for any  $\theta$  with  $0 < \theta < 1$

$$\theta(1 - \theta) \leq \frac{1}{4}.$$

*Remark:* this inequality was used to derive the conservative method of building a confidence interval for the population proportion.

**d) Exact Hoeffding's confidence interval with coverage probability at least  $C$ :** For the sample proportion there is a much better inequality available than that of Chebyshev: Hoeffding's inequality of (7.30) in Proposition 111, applied to the sample proportion  $\hat{\Theta}_n$  gives for any  $m > 0$

$$P\left(\left|\hat{\Theta}_n - \theta^*\right| \geq m\right) \leq 2 \exp(-2nm^2) \quad (7.28)$$

so that if we desire a confidence interval with level  $C$ , we set

$$C = 1 - 2 \exp(-2nm^2)$$

and solve for  $m$ , which gives

$$m = \sqrt{\frac{-\log((1-C)/2)}{2n}} = \sqrt{\frac{\log(1/\alpha)}{2n}}.$$

This interval is also nonasymptotic (i.e. has exact coverage probability  $C$  for any  $n$ ) and is narrower than the one derived from the Chebyshev inequality. Nevertheless the interval derived from the normal tail, which is not exact but asymptotic, is more commonly used.

Next we prove Hoeffding's inequality using the following simple idea of Chernoff.

**Idea 110 (Chernoff's bounding method)** By Markov's inequality, if  $s$  is an arbitrary positive real number, then for any RV  $X$ , and any  $t > 0$ :

$$P(X \geq t) = P(e^{sX} \geq e^{st}) \leq \frac{E(e^{sX})}{e^{st}}$$

The idea in Chernoff's bounding method is to find  $s > 0$  that minimises the upper-bound, i.e., the RHS of the above equation, to make it as small as possible. In the case of a sum of independent RVs  $X_1, X_2, \dots, X_n$  given by  $S_n = \sum_{i=1}^n X_i$ :

$$\begin{aligned} P(S_n - E(S_n) \geq t) &\leq e^{-st} E\left(\exp\left(s \sum_{i=1}^n (X_i - E(X_i))\right)\right) \\ &= e^{-st} \prod_{i=1}^n E\left(e^{s(X_i - E(X_i))}\right), \quad \text{by independence.} \end{aligned} \quad (7.29)$$

Thus, the problem of finding better bounds than that given by Chebychev's inequality boils down to finding a good upper-bound for  $E(e^{s(X_i - E(X_i))})$ , i.e., the moment generating function of each of the random variables  $X_i - E(X_i)$ . There are many ways to do this and the most simple approach is due to Hoeffding in 1963 as shown next.

**Proposition 111 (Hoeffding's ≠)** Let  $X_1, X_2, \dots, X_n$  be independent bounded RVs such that  $P(Z_i \in [a_i, b_i]) = 1$  for each  $i \in \{1, 2, \dots, n\}$ . Let  $S_n = \sum_{i=1}^n X_i$ . Then for any  $t > 0$ , we have

$$P(|S_n - E(S_n)| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \quad (7.30)$$

**Proof:** The core of proving Hoeffding's inequality is the following upper bound: if  $X$  is a RV with  $E(X) = 0$  and  $a \leq X \leq b$ , then

$$E(e^{sX}) \leq e^{(s^2(b-a)^2/8)}$$

The above upper-bound is derived from the convexity of the exponential function:

$$e^{sx} \leq \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa}, \quad \text{for } a \leq x \leq b$$

Draw  $e^{sx}$  on y-axis as a function of  $x$  along x-axis and the line from  $e^{sa}$  to  $e^{sb}$  for the upper-bound as  $x$  goes from  $a$  to  $b$ .

Thus, taking expectations on both sides of the above inequality

$$\begin{aligned} E(e^{sX}) &\leq E\left(\frac{X-a}{b-a} e^{sb} + \frac{b-X}{b-a} e^{sa}\right) \\ &= \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb}, \quad \text{because } E(X) = 0 \\ &= \left(1 - c + ce^{s(b-a)}\right) e^{-cs(b-a)}, \quad \text{where, } c = \frac{-a}{b-a}. \end{aligned}$$

Now let

$$u = s(b - a) \quad \text{and define } \phi(u) := -cu + \log(1 - c + ce^u)$$

Then we have

$$E(e^{sX}) \leq (1 - c + ce^{s(b-a)}) e^{-cs(b-a)} = e^{\phi(u)}$$

To minimise the upper-bound let's express  $\phi(u)$  in a Taylor's series with remainder term:

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(v) \quad \text{for some } v \in [0, u]$$

$$\begin{aligned} \phi(0) &= -c \cdot 0 + \log(1 - c + ce^0) = 0 \\ \phi'(u) &= -c + \frac{ce^u}{1 - c + ce^u} \implies \phi'(u) = 0 \\ \phi'' &= \frac{ce^u}{1 - c + ce^u} - \frac{ce^u}{(1 - c + ce^u)^2} \\ &= \frac{ce^u}{1 - c + ce^u} \left(1 - \frac{ce^u}{(1 - c + ce^u)}\right) \\ &= \rho(1 - \rho), \quad \text{where, } \rho := \frac{ce^u}{(1 - c + ce^u)} \end{aligned}$$

Now,  $\phi'' = \rho(1 - \rho)$ , being the familiar quadratic, is maximised by setting

$$\rho = \frac{ce^u}{(1 - c + ce^u)} = \frac{1}{2} \implies \phi'' \leq \frac{1}{4} .$$

Thus we get

$$\phi(u) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8} \implies E(e^{sX}) \leq e^{s^2(b-a)^2/8} .$$

Now, we can just plug-in the above upper-bound  $e^{s^2(b-a)^2/8}$ , specialised to each  $X_i$  that is bounded between  $a_i$  and  $b_i$ , directly into (7.29) of Chernoff's bounding method to derive Hoeffding's inequality:

$$P((S_n - E(S_n)) \geq t) \leq e^{-st} \prod_{i=1}^n e^{s^2(b_i - a_i)^2/8}$$

Similarly, we can also show that

$$P((E(S_n) - S_n) \geq t) \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}$$

Thus we have proved the inequality for the absolute value of  $S_n - E(S_n)$  in (7.30) known as Hoeffding's inequality.

Next we prove (7.28) by specialising (7.30) to the  $\text{Binomial}(n, \theta^*)$  RV  $S_n = \sum_{i=1}^n X_i$ , where  $X_1, X_2, \dots, X_n \stackrel{IID}{\sim} \text{Bernoulli}(\theta^*)$ , in terms of the sample mean as the point estimator  $\widehat{\Theta}_n = \bar{X}_n = S_n/n$ , as originally proved by Chernoff (1952) and Okamoto (1958):

$$\begin{aligned} P(|\bar{\Theta}_n - \theta^*| \geq m) &= P\left(\frac{1}{n}|S_n - E(S_n)| \geq nm\right) = P(|S_n - E(S_n)| \geq nm) \\ &\leq 2e^{-2(nm)^2/\sum_{i=1}^n (b_i - a_i)^2} \\ &= 2e^{-2(nm)^2/\sum_{i=1}^n (1-0)^2} \quad \text{since for each Bernoulli}(\theta) \text{ RV } b_i = 1, a_i = 0 \\ &= 2e^{-2(nm)^2/n} = 2e^{-2nm^2} \end{aligned}$$

**Exercise 7.11** Suppose you toss a possibly biased coin 25 times and observe 18 heads. Assuming IID Bernoulli( $\theta^*$ ) trials where  $\theta^*$  is the probability of coming up heads, obtain exact confidence intervals for the unknown  $\theta^*$  with confidence level  $C$  of at least 95% using exact sample size via: (a) Chebyshev's inequality and (b) Hoeffding's inequality.

### Some technical convergence results

In order to establish the result for the asymptotic coverage probability of the standard confidence interval for  $\theta^*$  (i. e. in order to show (7.25)) some results about convergence in law and in probability are needed. Recall that a continuous random variable is one with a continuous distribution function  $P(X \leq t)$ . The claim that a RV  $X$  has a continuous distribution (or law) means that  $X$  has a continuous distribution function  $P(X \leq t)$ .

**Lemma 112** Suppose  $X_n$  is a sequence of RV which converges in distribution to a continuous RV  $X$ :

$$X_n \rightsquigarrow X$$

and let  $Y_n$  be a sequence of RV which converges in probability to 0:

$$Y_n \rightarrow_P 0.$$

Then

$$X_n + Y_n \rightsquigarrow X.$$

Note that no independence assumptions were made.

**Proof:** Let  $F_n$  be the distribution function of  $X_n$  and  $F$  be the respective d.f. of  $X$ . Convergence in distribution means that

$$P(X_n \leq t) = F_n(t) \rightarrow F(t)$$

for every continuity point of the limit d.f.  $F$ . We assumed that  $F$  is continuous, so it means convergence for every  $t$ . Now for  $\varepsilon > 0$

$$\begin{aligned} P(X_n + Y_n \leq t) &= \\ &= P(\{X_n + Y_n \leq t\} \cap \{|Y_n| \leq \varepsilon\}) + P(\{X_n + Y_n \leq t\} \cap \{|Y_n| > \varepsilon\}). \end{aligned} \tag{7.31}$$

The first term on the right is

$$\begin{aligned} P(\{X_n \leq t - Y_n\} \cap \{|Y_n| \leq \varepsilon\}) &\leq P(\{X_n \leq t + \varepsilon\} \cap \{|Y_n| \leq \varepsilon\}) \\ &\leq P(X_n \leq t + \varepsilon). \end{aligned}$$

For this upper bound we have

$$P(X_n \leq t + \varepsilon) \rightarrow F(t + \varepsilon) \text{ as } n \rightarrow \infty.$$

The second term in (7.31) is

$$P(\{X_n + Y_n \leq t\} \cap \{|Y_n| > \varepsilon\}) \leq P(|Y_n| > \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Hence for every  $\delta > 0$  we can find  $m_1$  such that for all  $n \geq m_1$

$$P(X_n + Y_n \leq t) \leq F(t + \varepsilon) + 2\delta. \tag{7.32}$$

Now take the same  $\varepsilon > 0$ ; we have

$$\begin{aligned} P(X_n \leq t - \varepsilon) &= \\ &\quad P(\{X_n \leq t - \varepsilon\} \cap \{|Y_n| \leq \varepsilon\}) + P(\{X_n \leq t - \varepsilon\} \cap \{|Y_n| > \varepsilon\}) \\ &\leq P(\{X_n + Y_n \leq t\} \cap \{|Y_n| \leq \varepsilon\}) + P(|Y_n| > \varepsilon) \\ &\leq P(X_n + Y_n \leq t) + P(|Y_n| > \varepsilon). \end{aligned}$$

Consequently

$$P(X_n + Y_n \leq t) \geq P(X_n \leq t - \varepsilon) - P(|Y_n| > \varepsilon).$$

Using again the two limits for the probabilities on the right, for every  $\delta > 0$  we can find  $m_2$  such that for all  $n \geq m_2$

$$P(X_n + Y_n \leq t) \geq F(t - \varepsilon) - 2\delta. \quad (7.33)$$

Taking  $m = \max(m_1, m_2)$  and collecting (7.32), (7.33), we obtain for  $n \geq m$

$$F(t - \varepsilon) - 2\delta \leq P(X_n + Y_n \leq t) \leq F(t + \varepsilon) + 2\delta.$$

Since  $F$  is continuous at  $t$ , and  $\varepsilon$  was arbitrary, we can select  $\varepsilon$  such that

$$\begin{aligned} F(t + \varepsilon) &\leq F(t) + \delta, \\ F(t - \varepsilon) &\geq F(t) - \delta \end{aligned}$$

so that for  $n$  large enough

$$F(t) - 3\delta \leq P(X_n + Y_n \leq t) \leq F(t) + 3\delta$$

and since  $\delta$  was also arbitrary, the result follows.

**Lemma 113** Under the assumptions of Lemma 112, we have

$$X_n Y_n \rightarrow_P 0.$$

**Proof:** Let  $\varepsilon > 0$ ; and  $\delta > 0$  be arbitrary and given. Suppose  $|X_n Y_n| \geq \varepsilon$ . Then, for every  $t > 0$ , either  $\{|X_n| > t\}$ , or if that is not the case, then  $|X_n Y_n| \leq t |Y_n|$  and hence  $|Y_n| \geq \varepsilon/t$ . Hence .

$$P(|X_n Y_n| \geq \varepsilon) \leq P(|X_n| > t) + P(|Y_n| \geq \varepsilon/t). \quad (7.34)$$

Let again  $F_n$  be the distribution function of  $X_n$  and  $F$  be the respective d.f. of  $X$ . Now for every  $t > 0$

$$\begin{aligned} P(|X_n| > t) &= 1 - P(X_n \leq t) + P(X_n < -t) \\ &\leq 1 - F_n(t) + F_n(-t). \end{aligned}$$

Since  $F_n$  converges to  $F_0$  at both points  $t, -t$ , we find  $m_1 = m_1(t)$  (depending on  $t$ ) such that for all  $n \geq m_1$

$$P(|X_n| > t) \leq 1 - F_0(t) + F_0(-t) + \delta$$

Select now  $t$  large enough such that

$$1 - F_0(t) \leq \delta, F_0(-t) \leq \delta.$$

Then for all  $n \geq m_1(t)$

$$P(|X_n| \geq t) \leq 3\delta.$$

On the other hand, once  $t$  is fixed, in view of convergence in probability to 0 of  $|Y_n|$ , one can find  $m_2$  such that for all  $n \geq m_2$

$$P(|Y_n| \geq \varepsilon/t) \leq \delta.$$

In view of (7.34) we have for all  $n \geq m = \max(m_1, m_2)$

$$P(|X_n Y_n| \geq \varepsilon) \leq 4\delta.$$

Since  $\delta > 0$  was arbitrary, the result is proved.

We need an auxiliary result which despite its simplicity is still frequently cited as a “Theorem”.

**Proposition 114 (Slutsky’s theorem).** Suppose a sequence of random variables  $X_n$  converges in probability to a value  $x$  ( $X_n \rightarrow_P x$  as  $n \rightarrow \infty$ ). Suppose  $f$  is a real valued function defined in a neighborhood of  $x$  and continuous there. Then

$$f(X_n) \rightarrow_P f(x), \quad n \rightarrow \infty.$$

**Proof:** Consider an arbitrary  $\varepsilon > 0$ . Select  $\delta > 0$  small enough such that  $(x - \delta, x + \delta)$  is contained in the neighborhood of  $x$  where  $f$  is defined and also fulfilling the condition that  $|t - x| \leq \delta$  implies  $|f(t) - f(x)| \leq \varepsilon$  (by continuity of  $f$  such a  $\delta$  can be found). Then the event  $|f(X_n) - f(x)| > \varepsilon$  implies  $|X_n - x| > \delta$  and hence

$$P(|f(X_n) - f(x)| > \varepsilon) \leq P(|X_n - x| > \delta).$$

Since the latter probability tends to 0 as  $n \rightarrow \infty$ , we also have

$$P(|f(X_n) - f(x_0)| > \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty$$

and since  $\varepsilon$  was arbitrary, the result is proved.

### Asymptotic confidence level

With these results we are now able to prove that the standard asymptotic confidence interval for the unknown population proportion  $\theta^*$  has indeed asymptotic level  $C$ , i.e. prove relation (7.25):

$$P\left(-z_\alpha \leq \frac{\hat{\Theta}_n - \theta^*}{\sqrt{\hat{\Theta}_n(1 - \hat{\Theta}_n)/\sqrt{n}}} \leq z_\alpha\right) \rightarrow C \quad (7.35)$$

for  $\alpha = (1 - C)/2$ . As already noted we have

$$\frac{\hat{\Theta}_n - \theta^*}{\sqrt{\hat{\Theta}_n(1 - \hat{\Theta}_n)/\sqrt{n}}} = T_n \cdot \sqrt{\frac{\theta^*(1 - \theta^*)}{\hat{\Theta}_n(1 - \hat{\Theta}_n)}} \quad (7.36)$$

where  $T_n \rightsquigarrow Z$ , and  $T_n$  is the “correctly standardized” sample proportion  $\hat{\Theta}_n$  (as in (7.20)). The function  $f(\theta^*) := \theta^*(1 - \theta^*)$  is continuous in a neighborhood of every  $\theta^* \in (0, 1)$ ; by Slutsky’s theorem we have  $\hat{\Theta}_n(1 - \hat{\Theta}_n) \rightarrow_P \theta^*(1 - \theta^*)$ . A repeated application of Slutsky’s theorem now gives

$$\frac{\theta^*(1 - \theta^*)}{\hat{\Theta}_n(1 - \hat{\Theta}_n)} \rightarrow_P 1, \quad \sqrt{\frac{\theta^*(1 - \theta^*)}{\hat{\Theta}_n(1 - \hat{\Theta}_n)}} \rightarrow_P 1.$$

An application of Lemma (113) to (7.36) gives

$$\hat{T}_n := \frac{\hat{\Theta}_n - \theta^*}{\sqrt{\hat{\Theta}_n(1 - \hat{\Theta}_n)/n}} \rightsquigarrow Z.$$

Analogously to Lemma (109) (replace  $T_n$  there by  $\hat{T}_n$ ) it now follows that

$$P(-z_\alpha \leq \hat{T}_n \leq z_\alpha^*) \rightarrow P(-z_\alpha \leq Z \leq z_\alpha^*) = C$$

i.e. (7.35) is established and thereby proving the following proposition.

**Proposition 115** Suppose  $X_1, \dots, X_n$  are independent Bernoulli observations with  $X_i \sim \text{Bernoulli}(\theta^*)$ . Suppose  $\theta^* \in (0, 1)$  is unknown and we want to find out its value. Let the point estimator  $\hat{\Theta}_n$  of  $\theta^*$  be the sample proportion and let  $z_\alpha$  be the upper  $\alpha$ -quantile of  $Z$  for some  $0 < \alpha < 1/2$ . Then

$$[\hat{\Theta}_n - m, \hat{\Theta}_n + m] \quad \text{where, } m = z_\alpha \sqrt{\frac{\hat{\Theta}_n(1 - \hat{\Theta}_n)}{n}}$$

is a confidence interval for  $\theta^*$  of asymptotic level  $C = 1 - 2\alpha$ .

**Exercise 7.12** Suppose you toss a possibly biased coin 125 times and observe 118 heads. Assuming IID  $\text{Bernoulli}(\theta^*)$  trials where  $\theta^*$  is the probability of coming up heads, obtain a confidence interval based on the CLT with asymptotic level 95% for the unknown  $\theta^*$ .

### Sample size determination

The standard confidence interval has width  $2m$ , where the margin of error is

$$m = z_\alpha \sqrt{\frac{\hat{\Theta}_n(1 - \hat{\Theta}_n)}{n}} \tag{7.37}$$

There are two conflicting aims:

**Aim of precision:** The width should be as small as possible, in order to accurately “pinpoint” the unknown value  $\theta^*$ .

**Aim of certainty:** The confidence level  $C$  should be as high as possible, i.e., close to one, since it can be interpreted as the “reliability” of the interval.

The conflict clearly appears with the quantile  $z_\alpha$ : if  $C \rightarrow 1$  then  $\alpha = (1 - C)/2 \rightarrow 0$  and hence  $z_\alpha \rightarrow \infty$ , and as a consequence the margin of error  $m$  also tends to infinity. For small margin of error at fixed sample size  $n$  we would have to decrease  $z_\alpha$ , thus increase  $\alpha$  and decrease  $C$ . However the sample size  $n$  also influences the margin of error  $m$ ; in fact  $m$  is proportional to  $1/\sqrt{n}$ . Thus both high  $C$  and small  $m$  can easily be achieved if we let  $n \rightarrow \infty$ , but sample size usually is a cost factor.

If we use the conservative margin of error

$$m = z_\alpha \sqrt{\frac{1}{4n}} \tag{7.38}$$

then it is easy to determine a sample size that a given margin of error  $m^*$  is guaranteed:

$$n = \left( \frac{z_\alpha}{2m^*} \right)^2 \quad (7.39)$$

is obtained by solving (7.38) for  $n$ .

If we prefer the standard interval, and try to determine  $n$  for achieving a given margin of error, we are faced with the problem that  $\hat{\Theta}_n$  is not available before we have obtained the sample of that size. Indeed setting  $m = m^*$  in (7.37) and solving for  $n$  gives

$$n = \hat{\Theta}_n(1 - \hat{\Theta}_n) \left( \frac{z_\alpha}{m^*} \right)^2.$$

An obvious idea is to use an initial guess of  $\theta^*$ , or conduct a pilot study with sample size  $n_1$  and obtain an initial estimate  $\hat{\Theta}_{(1)}$  from there. Then determine the final sample size from

$$\hat{n} = \hat{\Theta}_{(1)}(1 - \hat{\Theta}_{(1)}) \left( \frac{z_\alpha}{m^*} \right)^2$$

### Two stage method

To achieve a given margin of error  $m^*$  (for fixed confidence level  $C$ ):

(1) Sample  $n_1$  data  $Y_1, \dots, Y_{n_1}$  (all IID Bernoulli  $Bernoulli(\theta^*)$ ), obtain the corresponding sample proportion  $\hat{\Theta}_{(1)}$  and use it to determine an estimated sample size  $\hat{n}$  by

$$\hat{n} = \hat{\Theta}_{(1)}(1 - \hat{\Theta}_{(1)}) \left( \frac{z_\alpha}{m^*} \right)^2.$$

(2) Take another sample  $X_1, \dots, X_{\hat{n}}$  (all IID Bernoulli  $Bernoulli(\theta^*)$ ), independent of the first one, of size  $\hat{n}$ , obtain the sample proportion  $\hat{\Theta}_{\hat{n}}$  and use it for a confidence interval  $[\hat{\Theta}_{\hat{n}} - m^*, \hat{\Theta}_{\hat{n}} + m^*]$ .

The method is well founded heuristically. Can we show that it works, i.e. that asymptotic confidence level  $C$  is maintained?

Consider an "ideal sample size"  $n_0$  which would guarantee the margin of error  $m^*$  if  $\theta^*$  were known:

$$m^* = z_\alpha \sqrt{\frac{\theta^*(1 - \theta^*)}{n_0}}. \quad (7.40)$$

i.e.

$$n_0 = \theta^*(1 - \theta^*) \left( \frac{z_\alpha}{m^*} \right)^2. \quad (7.41)$$

This is a sample size we cannot use, but if we could use it, then the interval  $[\hat{\Theta}_{n_0} - m^*, \hat{\Theta}_{n_0} + m^*]$  would surely have asymptotic confidence level  $C$ . Of course if we know  $\theta^*$  then there is no need for a confidence interval anymore. But consideration of such "unavailable" methods is frequently useful<sup>3</sup>.

In what follows we will assume firstly, that the sample size for the pilot study  $n_1$  tends to infinity, so that we have a more and more accurate  $\hat{\Theta}_{(1)}$ . On the other hand, we want  $n_1$  to be small compared

<sup>3</sup>Such unavailable choices are sometimes called "of oracle type" since they are only available if an oracle tells us the truth, i.e. gives the true  $\theta^*$  here.

to the size  $\hat{n}$  of our "actual" sample, i.e we want  $n_1/\hat{n} \rightarrow_P 0$  (here  $\hat{n}$  is random). Since there is good reason to believe that  $\hat{n}/n_0 \rightarrow_P 1$ , we should guarantee  $n_1/n_0 \rightarrow 0$ . But  $n_0$  is determined by  $m^*$  via (7.41), so we are led to consider small desired margins of error:  $m^* \rightarrow 0$ . The requirement  $n_1/n_0 \rightarrow 0$  will equivalently be expressed as  $n_1(m^*)^2 \rightarrow 0$  (in view of (7.41)).

**Remark 116** When  $n_1$  is only a fraction of  $n_0$  then the following modification of the Algorithm does not change the essence: let  $Y_1, \dots, Y_{n_1}$  be the first part of the larger sample  $X_1, \dots, X_{\hat{n}}$ . This is how one might proceed in practice, but to show rigorously that this method also works is slightly more involved, though not different in principle from our proof below.

To summarize: we will consider the Algorithm 7.5.4 in a setting where  $m^* \rightarrow 0$ ,  $n_1 \rightarrow \infty$ , and  $n_1(m^*)^2 \rightarrow 0$  (the third requirement means that  $n_1$  tends to infinity slower than  $(m^*)^2$  tends to zero).

**Proposition 117** Suppose independent Bernoulli observations  $Y_1, \dots, Y_{n_1}$  and  $X_1, X_2 \dots$  with law  $\text{Bernoulli}(\theta^*)$  where  $\theta^* \in (0, 1)$  is unknown. Suppose  $n_1 \rightarrow \infty$ ,  $m^* \rightarrow 0$  and  $n_1(m^*)^2 \rightarrow 0$ . Then the confidence interval  $[\hat{\Theta}_{\hat{n}} - m^*, \hat{\Theta}_{\hat{n}} + m^*]$  given by Algorithm 7.5.4 maintains asymptotic level  $C$ .

**Proof:** For the proof, we may take  $n_0$  as our basic index tending to infinity ( $n_0 \rightarrow 0$  is equivalent to  $m^* \rightarrow 0$ ). First we show that

$$\frac{\hat{n}}{n_0} \rightarrow_P 1 \text{ as } n_0 \rightarrow 0. \quad (7.42)$$

Indeed, we have

$$\begin{aligned} \frac{\hat{n}}{n_0} &= \frac{\hat{\Theta}_{(1)}(1 - \hat{\Theta}_{(1)}) \left(\frac{z_\alpha}{m^*}\right)^2}{\theta^*(1 - \theta^*) \left(\frac{z_\alpha}{m^*}\right)^2} \\ &= \frac{\hat{\Theta}_{(1)}(1 - \hat{\Theta}_{(1)})}{\theta^*(1 - \theta^*)}. \end{aligned}$$

By Slutsky's theorem (Theorem 114) it suffices to show that  $\hat{\Theta}_{(1)} \rightarrow_P \theta^*$ . By the LLN, this is implied by our condition  $n_1 \rightarrow \infty$ .

Consider the coverage probability:

$$\begin{aligned} &= P(\hat{\Theta}_{\hat{n}} - m^* \leq \theta^* \leq \hat{\Theta}_{\hat{n}} + m^*) \\ &= 1 - P(\hat{\Theta}_{\hat{n}} - \theta^* \leq -m^*) + P(\hat{\Theta}_{\hat{n}} - \theta^* \geq m^*) \end{aligned} \quad (7.43)$$

Now

$$\begin{aligned} &P(\hat{\Theta}_{\hat{n}} - \theta^* \geq m^*) = \\ &P\left(\frac{\hat{\Theta}_{\hat{n}} - \theta^*}{\sqrt{\theta^*(1 - \theta^*)}/\sqrt{\hat{n}}} \leq \frac{\sqrt{\hat{n}}m^*}{\sqrt{\theta^*(1 - \theta^*)}}\right). \end{aligned} \quad (7.44)$$

Now  $\hat{\Theta}_{\hat{n}}$  is from a sample of random size  $\hat{n}$  but the  $\hat{n}$  is independent of this sample since it is based on  $Y := (Y_1, \dots, Y_{n_1}, \dots)$ . Now by  $P_*$  we will write probabilities for fixed  $Y$  (conditional on  $Y$ );

then we have  $P(\cdot) = EP_*(\cdot)$  where the expected value refers to  $Y$ . Select  $\varepsilon > 0$ ; from (7.42) we have on an event  $A_{n_0}$  (concerning only  $Y$ ) where  $P(Y \notin A_{n_0}) \rightarrow 0$

$$1 - \varepsilon \leq \hat{n}/n_0 \leq 1 + \varepsilon$$

hence for  $Y \in A_{n_0}$

$$\hat{n} \geq n_0(1 - \varepsilon) \text{ and } \hat{n} \leq n_0(1 + \varepsilon).$$

In this case

$$\frac{\sqrt{\hat{n}}m^*}{\sqrt{\theta^*(1 - \theta^*)}} \leq \frac{\sqrt{n_0}m^*}{\sqrt{\theta^*(1 - \theta^*)}} \sqrt{(1 + \varepsilon)} = z_\alpha \sqrt{(1 + \varepsilon)}$$

by (7.40). Let  $B_{n_0}$  be the event in (7.44), then, if  $\mathbf{1}_{\{Y \in A_{n_0}\}}$  denotes the indicator of the event  $\{Y \in A_{n_0}\}$ ,

$$\begin{aligned} P(B_{n_0}) &= EP_*(B_{n_0}) = E \mathbf{1}_{\{Y \in A_{n_0}\}} P_*(B_{n_0}) + E \mathbf{1}_{\{Y \notin A_{n_0}\}} P_*(B_{n_0}) \\ &\leq E \mathbf{1}_{\{Y \in A_{n_0}\}} P_*(B_{n_0}) + E \mathbf{1}_{\{Y \notin A_{n_0}\}} \\ &= E \mathbf{1}_{\{Y \in A_{n_0}\}} P_*(B_{n_0}) + P(Y \notin A_{n_0}) \\ &\leq E \mathbf{1}_{\{Y \in A_{n_0}\}} P_* \left( \frac{\hat{\Theta}_{\hat{n}} - \theta^*}{\sqrt{\theta^*(1 - \theta^*)}/\sqrt{\hat{n}}} \leq z_\alpha \sqrt{(1 + \varepsilon)} \right) + P(Y \notin A_{n_0}). \end{aligned}$$

If  $Y$  is fixed then  $\hat{n}$  is also fixed and on the event  $Y \in A_{n_0}$  we have  $\hat{n} \geq n_0(1 - \varepsilon)$ . Thus  $\hat{n}$  is large and the CLT will hold, so for large  $n_0$  we will have

$$P_* \left( \frac{\hat{\Theta}_{\hat{n}} - \theta^*}{\sqrt{\theta^*(1 - \theta^*)}/\sqrt{\hat{n}}} \leq z_\alpha \sqrt{(1 + \varepsilon)} \right) \leq P(Z \leq z_\alpha \sqrt{(1 + \varepsilon)}) + \varepsilon.$$

Here the right side does not depend on  $Y$ . Collecting these results we obtain

$$\begin{aligned} P(B_{n_0}) &\leq E \mathbf{1}_{\{Y \in A_{n_0}\}} \left( P(Z \leq z_\alpha \sqrt{(1 + \varepsilon)}) + \varepsilon \right) + P(Y \notin A_{n_0}) \\ &\leq P(Z \leq z_\alpha \sqrt{(1 + \varepsilon)}) + \varepsilon + P(Y \notin A_{n_0}) \\ &\leq P(Z \leq z_\alpha \sqrt{(1 + \varepsilon)}) + 2\varepsilon \end{aligned}$$

since  $P(Y \notin A_{n_0}) \leq \varepsilon$  for sufficiently large  $n_0$ . Since  $\varepsilon > 0$  was arbitrary, we obtain

$$\limsup P(\hat{\Theta}_{\hat{n}} - \theta^* \geq m^*) = \limsup P(B_{n_0}) \leq P(Z \leq z_\alpha) = \alpha$$

For the other term in (7.43) we obtain analogously

$$\limsup P(\hat{\Theta}_{\hat{n}} - \theta^* \leq -m^*) \leq \alpha.$$

hence

$$\liminf P(\hat{\Theta}_{\hat{n}} - m^* \leq \theta^* \leq \hat{\Theta}_{\hat{n}} + m^*) \geq 1 - 2\alpha = C.$$

**Remark 118** In the proof we did not use the condition that  $n_1$  is only a fraction of  $n_0$  (i.e.  $n_1(m^*)^2 \rightarrow 0$ ) but this condition would play a role if  $Y_1, \dots, Y_{n_1}$  is the first part of the larger sample  $X_1, \dots, X_{\hat{n}}$ .

**Exercise 7.13** A simple random sample of 200 people aged 18 or over is taken in a large city to see how many of them know the name of the mayor. It turns out that 118 could correctly give her name.

- a Find a 90% (asymptotic) confidence interval for the proportion of people 18 or over who know the mayor's name. In calculating the interval, please also give the margin of error.
- b Suppose the staff said that budget constraints mean that the largest sample they can obtain is 1200 people. They make another survey with that sample size and find  $\hat{p} = 0.6$ . In order to please the mayor they report a margin of error of 0.02. What is the asymptotic confidence level of this interval?

## 7.6 Fundamentals of Estimation

### 7.6.1 Introduction

Now that we have been introduced to two notions of convergence for RV sequences, we can begin to appreciate the basic limit theorems used in statistical inference. The problem of estimation is of fundamental importance in statistical inference and learning. We will formalise the general estimation problem here. There are two basic types of estimation. In point estimation we are interested in estimating a particular point of interest that is supposed to belong to a set of points. In (confidence) set estimation, we are interested in estimating a set with a particular form that has a specified probability of “trapping” the particular point of interest from a set of points. Here, a point should be interpreted as an element of a collection of elements from some space.

### 7.6.2 Point Estimation

**Point estimation** is any statistical methodology that provides one with a “**single best guess**” of some specific quantity of interest. Traditionally, we denote this **quantity of interest as  $\theta^*$**  and its **point estimate as  $\hat{\theta}$  or  $\hat{\theta}_n$** . The subscript  $n$  in the point estimate  $\hat{\theta}_n$  emphasises that our estimate is based on  $n$  observations or data points from a given statistical experiment to estimate  $\theta^*$ . This quantity of interest, which is usually unknown, can be:

- an **integral**  $\vartheta^* := \int_A h(x) dx \in \Theta$ . If  $\vartheta^*$  is finite, then  $\Theta = \mathbb{R}$ , or
- a **parameter**  $\theta^*$  which is an element of the **parameter space**  $\Theta$ , denoted  $\theta^* \in \Theta$ ,
- a **distribution function (DF)**  $F^* \in \mathbb{F} :=$  the set of all DFs
- a **density function (pdf)**  $f \in \{$  “not too wiggly Sobolev functions”  $\}$ , or
- a **regression function**  $g^* \in \mathbb{G}$ , where  $\mathbb{G}$  is a class of regression functions in a regression experiment with model:  $Y = g^*(X) + \epsilon$ , such that  $E(\epsilon) = 0$ , from pairs of observations  $\{(X_i, Y_i)\}_{i=1}^n$ , or
- a **classifier**  $g^* \in \mathbb{G}$ , i.e. a regression experiment with discrete  $Y = g^*(X) + \epsilon$ , or
- a **prediction** in a regression experiment, i.e. when you want to estimate  $Y_i$  given  $X_i$ .

Recall that a statistic is an RV  $T(X)$  that maps every data point  $x$  in the data space  $\mathbb{X}$  with  $T(x) = t$  in its range  $\mathbb{T}$ , i.e.  $T(x) : \mathbb{X} \rightarrow \mathbb{T}$  (Definition 51). Next, we look at a specific class of statistics whose range is the parameter space  $\Theta$ .

**Definition 119 (Point Estimator)** A **point estimator**  $\hat{\Theta}$  of some **fixed and possibly unknown**  $\theta^* \in \Theta$  is a statistic that associates each data point  $x \in \mathbb{X}$  with an estimate  $\hat{\Theta}(x) = \hat{\theta} \in \Theta$ ,

$$\boxed{\hat{\Theta} := \hat{\Theta}(x) = \hat{\theta} : \mathbb{X} \rightarrow \Theta} .$$

If our data point  $x := (x_1, x_2, \dots, x_n)$  is an  $n$ -vector or a point in the  $n$ -dimensional real space, i.e.  $x := (x_1, x_2, \dots, x_n) \in \mathbb{X}_n \subset \mathbb{R}^n$ , then we emphasise the dimension  $n$  in our point estimator  $\hat{\Theta}_n$  of  $\theta^* \in \Theta$ .

$$\boxed{\hat{\Theta}_n := \hat{\Theta}_n(x := (x_1, x_2, \dots, x_n)) = \hat{\theta}_n : \mathbb{X}_n \rightarrow \Theta, \quad \mathbb{X}_n \subset \mathbb{R}^n} .$$

The typical situation for us involves point estimation of  $\theta^* \in \Theta$  on the basis of one realisation  $x \in \mathbb{X}_n \subset \mathbb{R}^n$  of an independent and identically distributed (IID) random vector  $X = (X_1, X_2, \dots, X_n)$ , such that  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} X_1$  and the DF of  $X_1$  is  $F(x_1; \theta^*)$ , i.e. the distribution of the IID RVs,  $X_1, X_2, \dots, X_n$ , is parameterised by  $\theta^* \in \Theta$ .

**Example 195 (Coin Tossing Experiment)** ( $X_1, \dots, X_n \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta^*)$ ) I tossed a coin that has an unknown probability  $\theta^*$  of landing Heads independently and identically 10 times in a row. Four of my outcomes were Heads and the remaining six were Tails, with the actual sequence of Bernoulli outcomes (Heads  $\rightarrow 1$  and Tails  $\rightarrow 0$ ) being  $(1, 0, 0, 0, 1, 1, 0, 0, 1, 0)$ . I would like to estimate the probability  $\theta^* \in \Theta = [0, 1]$  of observing Heads using the natural estimator  $\widehat{\Theta}_n((X_1, X_2, \dots, X_n))$  of  $\theta^*$ :

$$\widehat{\Theta}_n((X_1, X_2, \dots, X_n)) := \widehat{\Theta}_n = \frac{1}{n} \sum_{i=1}^n X_i =: \bar{X}_n$$

For the coin tossing experiment I just performed ( $n = 10$  times), the point estimate of the unknown  $\theta^*$  is:

$$\begin{aligned}\widehat{\theta}_{10} = \widehat{\Theta}_{10}((x_1, x_2, \dots, x_{10})) &= \widehat{\Theta}_{10}((1, 0, 0, 0, 1, 1, 0, 0, 1, 0)) \\ &= \frac{1 + 0 + 0 + 0 + 1 + 1 + 0 + 0 + 1 + 0}{10} = \frac{4}{10} = 0.40.\end{aligned}$$

**Labwork 196 (Bernoulli(38/75) Computer Experiment)** Simulate one thousand IID samples from a  $\text{Bernoulli}(\theta^* = 38/75)$  RV and store this data in an array called **Samples**. Use your student ID to initialise the fundamental sampler. Now, pretend that you don't know the true  $\theta^*$  and estimate  $\theta^*$  using our estimator  $\widehat{\Theta}_n = \bar{X}_n$  from the data array **Samples** for each sample size  $n = 1, 2, \dots, 1000$ . Plot the one thousand estimates  $\widehat{\theta}_1, \widehat{\theta}_2, \dots, \widehat{\theta}_{1000}$  as a function of the corresponding sample size. Report your observations regarding the behaviour of our estimator as the sample size increases.

### 7.6.3 Some Properties of Point Estimators

Given that an estimator is merely a function from the data space to the parameter space, we need choose only the best estimators available. Recall that a point estimator  $\widehat{\Theta}_n$ , being a statistic or an RV of the data has a probability distribution over its range  $\Theta$ . This distribution over  $\Theta$  is called the **sampling distribution** of  $\widehat{\Theta}_n$ . Note that the sampling distribution not only depends on the statistic  $\widehat{\Theta}_n := \widehat{\Theta}_n(X_1, X_2, \dots, X_n)$  but also on  $\theta^*$  which in turn determines the distribution of the IID data vector  $(X_1, X_2, \dots, X_n)$ . The following definitions are useful for selecting better estimators from some lot of them.

**Definition 120 (Bias of a Point Estimator)** The  $\text{bias}_n$  of an estimator  $\widehat{\Theta}_n$  of  $\theta^* \in \Theta$  is:

$$\text{bias}_n = \text{bias}_n(\widehat{\Theta}_n) := E_{\theta^*}(\widehat{\Theta}_n) - \theta^* = \int_{\mathbb{X}_n} \widehat{\Theta}_n(x) dF(x; \theta^*) - \theta^* . \quad (7.45)$$

We say that the estimator  $\widehat{\Theta}_n$  is **unbiased** if  $\text{bias}_n(\widehat{\Theta}_n) = 0$  or if  $E_{\theta^*}(\widehat{\Theta}_n) = \theta^*$  for every  $n$ . If  $\lim_{n \rightarrow \infty} \text{bias}_n(\widehat{\Theta}_n) = 0$ , we say that the estimator is **asymptotically unbiased**.

Since the expectation of the sampling distribution of the point estimator  $\widehat{\Theta}_n$  depends on the unknown  $\theta^*$ , we emphasise the  $\theta^*$ -dependence by  $E_{\theta^*}(\widehat{\Theta}_n)$ .

**Example 197 (Bias of our Estimator of  $\theta^*$ )** Consider the sample mean estimator  $\widehat{\Theta}_n := \overline{X}_n$  of  $\theta^*$ , from  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta^*)$ . That is, we take the sample mean of the  $n$  IID Bernoulli( $\theta^*$ ) trials to be our point estimator of  $\theta^* \in [0, 1]$ . Then, **this estimator is unbiased** since:

$$\mathbb{E}_{\theta^*}(\widehat{\Theta}_n) = \mathbb{E}_{\theta^*}\left(n^{-1} \sum_{i=1}^n X_i\right) = n^{-1} \mathbb{E}_{\theta^*}\left(\sum_{i=1}^n X_i\right) = n^{-1} \sum_{i=1}^n \mathbb{E}_{\theta^*}(X_i) = n^{-1} n \theta^* = \theta^* .$$

**Definition 121 (Standard Error of a Point Estimator)** The standard deviation of the point estimator  $\widehat{\Theta}_n$  of  $\theta^* \in \Theta$  is called the **standard error**:

$$\text{se}_n = \text{se}_n(\widehat{\Theta}_n) = \sqrt{\text{V}_{\theta^*}(\widehat{\Theta}_n)} = \sqrt{\int_{\mathbb{X}_n} \left(\widehat{\Theta}_n(x) - \mathbb{E}_{\theta^*}(\widehat{\Theta}_n)\right)^2 dF(x; \theta^*)} . \quad (7.46)$$

Since the variance of the sampling distribution of the point estimator  $\widehat{\Theta}_n$  depends on the fixed and possibly unknown  $\theta^*$ , as emphasised by  $\text{V}_{\theta^*}$  in (7.46), the  $\text{se}_n$  is also a possibly unknown quantity and may itself be estimated from the data. The estimated standard error, denoted by  $\widehat{\text{se}}_n$ , is calculated by replacing  $\text{V}_{\theta^*}(\widehat{\Theta}_n)$  in (7.46) with its appropriate estimate.

**Example 198 (Standard Error of our Estimator of  $\theta^*$ )** Consider the sample mean estimator  $\widehat{\Theta}_n := \overline{X}_n$  of  $\theta^*$ , from  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta^*)$ . Observe that the statistic:

$$T_n((X_1, X_2, \dots, X_n)) := n \widehat{\Theta}_n((X_1, X_2, \dots, X_n)) = \sum_{i=1}^n X_i$$

is the Binomial( $n, \theta^*$ ) RV. The standard error  $\text{se}_n$  of this estimator is:

$$\text{se}_n = \sqrt{\text{V}_{\theta^*}(\widehat{\Theta}_n)} = \sqrt{\text{V}_{\theta^*}\left(\sum_{i=1}^n \frac{X_i}{n}\right)} = \sqrt{\left(\sum_{i=1}^n \frac{1}{n^2} \text{V}_{\theta^*}(X_i)\right)} = \sqrt{\frac{n}{n^2} \text{V}_{\theta^*}(X_i)} = \sqrt{\theta^*(1 - \theta^*)/n} .$$

Another reasonable property of an estimator is that it converge to the “true” parameter  $\theta^*$  – here “true” means the supposedly fixed and possibly unknown  $\theta^*$ , as we gather more and more IID data from a  $\theta^*$ -specified DF  $F(x; \theta^*)$ . This property is stated precisely next.

**Definition 122 (Asymptotic Consistency of a Point Estimator)** A point estimator  $\widehat{\Theta}_n$  of  $\theta^* \in \Theta$  is said to be **asymptotically consistent** if:

$$\widehat{\Theta}_n \xrightarrow{P} \theta^* \quad \text{i.e., for any real } \epsilon > 0, \quad \lim_{n \rightarrow \infty} P(|\widehat{\Theta}_n - \theta^*| > \epsilon) = 0 .$$

**Definition 123 (Mean Squared Error (MSE) of a Point Estimator)** Often, the quality of a point estimator  $\widehat{\Theta}_n$  of  $\theta^* \in \Theta$  is assessed by the **mean squared error** or  $\text{MSE}_n$  defined by:

$$\text{MSE}_n = \text{MSE}_n(\widehat{\Theta}_n) := \mathbb{E}_{\theta^*} \left( (\widehat{\Theta}_n - \theta^*)^2 \right) = \int_{\mathbb{X}} (\widehat{\Theta}_n(x) - \theta^*)^2 dF(x; \theta^*) . \quad (7.47)$$

The following proposition shows a simple relationship between the mean square error, bias and variance of an estimator  $\widehat{\Theta}_n$  of  $\theta^*$ .

**Proposition 124 (The  $\sqrt{\text{MSE}_n}$  :  $\text{se}_n$  :  $\text{bias}_n$ -Sided Right Triangle of an Estimator)** Let  $\hat{\Theta}_n$  be an estimator of  $\theta^* \in \Theta$ . Then:

$$\boxed{\text{MSE}_n(\hat{\Theta}_n) = (\text{se}_n(\hat{\Theta}_n))^2 + (\text{bias}_n(\hat{\Theta}_n))^2} . \quad (7.48)$$

**Proof:**

$$\begin{aligned}
& LHS \\
&= \text{MSE}_n(\hat{\Theta}_n) \\
&:= E_{\theta^*}((\hat{\Theta}_n - \theta^*)^2), \quad \text{by definition of } \text{MSE}_n \text{ (7.47)} \\
&= E_{\theta^*} \left( \left( \underbrace{\hat{\Theta}_n - E_{\theta^*}(\hat{\Theta}_n)}_A + \underbrace{E_{\theta^*}(\hat{\Theta}_n) - \theta^*}_B \right)^2 \right), \quad \text{by subtracting and adding the constant } E_{\theta^*}(\hat{\Theta}_n) \\
&= E_{\theta^*} \left( \underbrace{(\hat{\Theta}_n - E_{\theta^*}(\hat{\Theta}_n))^2}_{A^2} + 2 \underbrace{(\hat{\Theta}_n - E_{\theta^*}(\hat{\Theta}_n))(E_{\theta^*}(\hat{\Theta}_n) - \theta^*)}_{2AB} + \underbrace{(E_{\theta^*}(\hat{\Theta}_n) - \theta^*)^2}_{B^2} \right), \quad \because (A + B)^2 = A^2 + 2AB + B^2 \\
&= E_{\theta^*} \left( (\hat{\Theta}_n - E_{\theta^*}(\hat{\Theta}_n))^2 \right) + E_{\theta^*} \left( 2(\hat{\Theta}_n - E_{\theta^*}(\hat{\Theta}_n))(E_{\theta^*}(\hat{\Theta}_n) - \theta^*) \right) + E_{\theta^*} \left( (E_{\theta^*}(\hat{\Theta}_n) - \theta^*)^2 \right), \\
&= E_{\theta^*} \left( (\hat{\Theta}_n - E_{\theta^*}(\hat{\Theta}_n))^2 \right) + \underbrace{2(E_{\theta^*}(\hat{\Theta}_n) - \theta^*)}_{C} \underbrace{E_{\theta^*}((\hat{\Theta}_n - E_{\theta^*}(\hat{\Theta}_n)))}_{D} + E_{\theta^*} \left( (E_{\theta^*}(\hat{\Theta}_n) - \theta^*)^2 \right), \quad \because C \text{ is constant} \\
&= E_{\theta^*} \left( (\hat{\Theta}_n - E_{\theta^*}(\hat{\Theta}_n))^2 \right) + 0 + E_{\theta^*} \left( (E_{\theta^*}(\hat{\Theta}_n) - \theta^*)^2 \right), \quad \because D := E_{\theta^*}((\hat{\Theta}_n - E_{\theta^*}(\hat{\Theta}_n))) = E_{\theta^*}(\hat{\Theta}_n) - E_{\theta^*}(\hat{\Theta}_n) = 0 \\
&= V_{\theta^*}(\hat{\Theta}_n) + E_{\theta^*} \left( (E_{\theta^*}(\hat{\Theta}_n) - \theta^*)^2 \right), \quad \because V_{\theta^*}(\hat{\Theta}_n) := E_{\theta^*}((\hat{\Theta}_n - E_{\theta^*}(\hat{\Theta}_n))^2), \text{ by definition of variance} \\
&= \left( \sqrt{V_{\theta^*}(\hat{\Theta}_n)} \right)^2 + E_{\theta^*} \left( (\text{bias}_n(\hat{\Theta}_n))^2 \right), \quad \because \text{bias}_n(\hat{\Theta}_n) = E_{\theta^*}(\hat{\Theta}_n) - \theta^*, \text{ by definition of bias}_n \text{ of an estimator } \hat{\Theta}_n \\
&= (\text{se}_n(\hat{\Theta}_n))^2 + E_{\theta^*} \left( (\text{bias}_n(\hat{\Theta}_n))^2 \right), \quad \because \text{se}_n(\hat{\Theta}_n) := \sqrt{V_{\theta^*}(\hat{\Theta}_n)}, \text{ by definition (7.46)} \\
&= (\text{se}_n(\hat{\Theta}_n))^2 + (\text{bias}_n(\hat{\Theta}_n))^2, \quad \because \text{bias}_n(\hat{\Theta}_n) = E_{\theta^*}(\hat{\Theta}_n) - \theta^* \text{ and } (\text{bias}_n(\hat{\Theta}_n))^2 \text{ are constants.} \\
&= RHS
\end{aligned}$$

**Proposition 125 (Asymptotic consistency of a point estimator)** Let  $\hat{\Theta}_n$  be an estimator of  $\theta^* \in \Theta$ . Then, if  $\text{bias}_n(\hat{\Theta}_n) \rightarrow 0$  and  $\text{se}_n(\hat{\Theta}_n) \rightarrow 0$  as  $n \rightarrow \infty$ , the estimator  $\hat{\Theta}_n$  is asymptotically consistent:

$$\hat{\Theta}_n \xrightarrow{P} \theta^* .$$

**Proof:** If  $\text{bias}_n(\hat{\Theta}_n) \rightarrow 0$  and  $\text{se}_n(\hat{\Theta}_n) \rightarrow 0$ , then by (7.48),  $\text{MSE}_n(\hat{\Theta}_n) \rightarrow 0$ , i.e. that  $E_{\theta^*}((\hat{\Theta}_n - \theta^*)^2) \rightarrow 0$ . This type of convergence of the RV  $\hat{\Theta}_n$  to the *Point Mass( $\theta^*$ )* RV as  $n \rightarrow \infty$  is called convergence in **quadratic mean** or **convergence in  $BL_2$**  and denoted by  $\hat{\Theta}_n \xrightarrow{qm} \theta^*$ . Convergence in quadratic mean is a stronger notion of convergence than convergence in probability, in the sense that

$$E_{\theta^*}((\hat{\Theta}_n - \theta^*)^2) \rightarrow 0 \quad \text{or} \quad \hat{\Theta}_n \xrightarrow{qm} \theta^* \implies \hat{\Theta}_n \xrightarrow{P} \theta^* .$$

Thus, if we prove the above implication we are done with the proof of our proposition. To show that convergence in quadratic mean implies convergence in probability for general sequence of RVs  $X_n$  converging to an RV  $X$ , we first assume that  $X_n \xrightarrow{qm} X$ . Now, fix any  $\epsilon > 0$ . Then by Markov's inequality (5.2),

$$P(|X_n - X| > \epsilon) = P(|X_n - X|^2 > \epsilon^2) \leq \frac{E(|X_n - X|^2)}{\epsilon^2} \rightarrow 0 ,$$

and we have shown that the definition of convergence in probability holds provided convergence in quadratic mean holds.

We want our estimator to be unbiased with small standard errors as the sample size  $n$  gets large. The **point estimator**  $\hat{\Theta}_n$  will then produce a **point estimate**  $\hat{\theta}_n$ :

$$\hat{\Theta}_n((x_1, x_2, \dots, x_n)) = \hat{\theta}_n \in \Theta ,$$

on the basis of the **observed data**  $(x_1, x_2, \dots, x_n)$ , that is close to the **true parameter**  $\theta^* \in \Theta$ .

**Example 199 (Asymptotic consistency of our Estimator of  $\theta^*$ )** Consider the sample mean estimator  $\widehat{\Theta}_n := \bar{X}_n$  of  $\theta^*$ , from  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta^*)$ . Since  $\text{bias}_n(\widehat{\Theta}_n) = 0$  for any  $n$  and  $\text{se}_n = \sqrt{\theta^*(1-\theta^*)/n} \rightarrow 0$ , as  $n \rightarrow \infty$ , by Proposition 125,  $\widehat{\Theta}_n \xrightarrow{P} \theta^*$ . That is  $\widehat{\Theta}_n$  is an **asymptotically consistent estimator** of  $\theta^*$ .

#### 7.6.4 Confidence Set Estimation

As we saw in Section 7.6.2, the point estimate  $\widehat{\theta}_n$  is a “single best guess” of the fixed and possibly unknown parameter  $\theta^* \in \Theta$ . However, if we wanted to make a statement about our confidence in an estimation procedure, then one possibility is to produce subsets from the parameter space  $\Theta$  called **confidence sets** that “engulf”  $\theta^*$  with a probability of at least  $1 - \alpha$ .

Formally, an  $1 - \alpha$  **confidence interval** for the parameter  $\theta^* \in \Theta \subset \mathbb{R}$ , based on  $n$  observations or data points  $X_1, X_2, \dots, X_n$ , is an interval  $C_n$  that is a function of the data:

$$C_n := [\underline{C}_n, \bar{C}_n] = [\underline{C}_n(X_1, X_2, \dots, X_n), \bar{C}_n(X_1, X_2, \dots, X_n)] ,$$

such that:

$$P_{\theta^*} (\theta^* \in C_n := [\underline{C}_n, \bar{C}_n]) \geq 1 - \alpha .$$

Note that the confidence interval  $C_n := [\underline{C}_n, \bar{C}_n]$  is a two-dimensional RV or a random vector in  $\mathbb{R}^2$  that depends on the two statistics  $\underline{C}_n(X_1, X_2, \dots, X_n)$  and  $\bar{C}_n(X_1, X_2, \dots, X_n)$ , as well as  $\theta^*$ , which in turn determines the distribution of the data  $(X_1, X_2, \dots, X_n)$ . In words,  $C_n$  engulfs the true parameter  $\theta^* \in \Theta$  with a probability of at least  $1 - \alpha$ . We call  $1 - \alpha$  as the **coverage** of the confidence interval  $C_n$ .

Formally, a  $1 - \alpha$  **confidence set**  $C_n$  for a vector-valued  $\theta^* \in \Theta \subset \mathbb{R}^k$  is any subset of  $\Theta$  such that  $P_{\theta^*}(\theta^* \in C_n) \geq 1 - \alpha$ . The typical forms taken by  $C_n$  are  $k$ -dimensional boxes or hyper-cuboids, hyper-ellipsoids and subsets defined by inequalities involving level sets of some estimator of  $\theta^*$ .

Typically, we take  $\alpha = 0.05$  because we are interested in the  $1 - \alpha = 0.95$  or 95% confidence interval/set  $C_n \subset \Theta$  of  $\theta^* \in \Theta$  from an estimator  $\widehat{\Theta}_n$  of  $\theta^*$ .

Let us look at an example that makes use of the CLT next (Exercise in Prob. Theor.I).

**Example 200 (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  $X_i$  modelling the number of errors in the  $i^{\text{th}}$  program is the *Poisson*( $\lambda^* = 5$ ) for any  $i = 1, 2, \dots, n$ . Also suppose that they are independently distributed. In short, 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 number of errors per program out of these 125 programs. Since  $E(X_i) = \lambda^* = 5$  and  $V(X_i) = \lambda^* = 5$ , 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  $P(\bar{X}_n < 5.5)$ , for instance, as follows:

$$\begin{aligned}
 P(\bar{X}_n < 5.5) &= P\left(\frac{\sqrt{n}(\bar{X}_n - E(X_1))}{\sqrt{V(X_1)}} < \frac{\sqrt{n}(5.5 - E(X_1))}{\sqrt{V(X_1)}}\right) \\
 &\approx P\left(Z < \frac{\sqrt{n}(5.5 - \lambda^*)}{\sqrt{\lambda^*}}\right) \quad [\text{by CLT, and } E(X_1) = V(X_1) = \lambda^*] \\
 &= P\left(Z < \frac{\sqrt{125}(5.5 - 5)}{\sqrt{5}}\right) \quad [\text{Since, } \lambda^* = 5 \text{ and } n = 125 \text{ in this Example}] \\
 &= P(Z \leq 2.5) = \Phi(2.5) = \int_{-\infty}^{2.5} \left(\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right)\right) dx \approx 0.993790334674224 .
 \end{aligned}$$

To obtain the final number in this approximation, we need the following:

**Labwork 201** The numerical approximation of  $\Phi(2.5)$  was obtained via the call shown below to our erf-based `NormalCdf` function from 234. We could have also found it from a pre-computed Table for  $\Phi(x)$ .

```
>> 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  $Z_n := \sqrt{n}(\bar{X}_n - E(X_1))/\sqrt{V(X_1)}$  is approximately distributed as  $\text{Normal}(0, 1)$ . In Example 200, we knew  $\sqrt{V(X_1)}$  since we assumed knowledge of  $\lambda^* = 5$ . However, in general, we may not know  $\sqrt{V(X_1)}$ . The next proposition says that we may estimate  $\sqrt{V(X_1)}$  using the sample standard deviation  $S_n$  of  $X_1, X_2, \dots, X_n$ , according to (3.79), and still make probability statements about the sample mean  $\bar{X}_n$  using a Normal distribution, **provided n is not too small**, for e.g.  $n \geq 30$ .

**Proposition 126 (CLT based on Sample Variance)** Let  $X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1$  and suppose  $E(X_1)$  and  $V(X_1)$  exists, then:

$$\frac{\sqrt{n}(\bar{X}_n - E(X_1))}{S_n} \rightsquigarrow \text{Normal}(0, 1) . \quad (7.49)$$

The following property of an estimator makes it easy to obtain confidence intervals.

**Definition 127 (Asymptotic Normality of Estimators)** An estimator  $\hat{\Theta}_n$  of a fixed and possibly unknown parameter  $\theta^* \in \Theta$  is **asymptotically normal** if:

$$\frac{\hat{\Theta}_n - \theta^*}{\text{se}_n} \rightsquigarrow \text{Normal}(0, 1) . \quad (7.50)$$

That is,  $\hat{\Theta}_n \rightsquigarrow \text{Normal}(\theta^*, \text{se}_n^2)$ . By a further estimation of  $\text{se}_n := \sqrt{V_{\theta^*}(\hat{\Theta}_n)}$  by  $\widehat{\text{se}}_n$ , we can see that  $\hat{\Theta}_n \rightsquigarrow \text{Normal}(\theta^*, \widehat{\text{se}}_n^2)$  on the basis of (7.49).

**Proposition 128 (Normal-based Asymptotic Confidence Interval)** Suppose an estimator  $\hat{\Theta}_n$  of parameter  $\theta^* \in \Theta \subset \mathbb{R}$  is asymptotically normal:

$$\hat{\Theta}_n \rightsquigarrow \text{Normal}(\theta^*, \widehat{\text{se}}_n^2) .$$

Let the RV  $Z \sim \text{Normal}(0, 1)$  have DF  $\Phi$  and inverse DF  $\Phi^{-1}$ . Let:

$$z_{\alpha/2} = \Phi^{-1}(1 - (\alpha/2)), \quad \text{that is,} \quad P(Z > z_{\alpha/2}) = \alpha/2 \quad \text{and} \quad P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha .$$

Then:

$$P_{\theta^*}(\theta^* \in C_n) = P\left(\theta^* \in [\hat{\Theta}_n - z_{\alpha/2}\hat{s}\hat{e}_n, \hat{\Theta}_n + z_{\alpha/2}\hat{s}\hat{e}_n]\right) \rightarrow 1 - \alpha .$$

Therefore:

$$C_n := [\underline{C}_n, \bar{C}_n] = [\hat{\Theta}_n - z_{\alpha/2}\hat{s}\hat{e}_n, \hat{\Theta}_n + z_{\alpha/2}\hat{s}\hat{e}_n]$$

is the  $1 - \alpha$  Normal-based asymptotic confidence interval that relies on the asymptotic normality of the estimator  $\hat{\Theta}_n$  of  $\theta^* \in \Theta \subset \mathbb{R}$ .

**Proof:** Define the centralised and scaled estimator as  $Z_n := (\hat{\Theta}_n - \theta^*)/\hat{s}\hat{e}_n$ . By assumption,  $Z_n \rightsquigarrow Z \sim \text{Normal}(0, 1)$ . Therefore,

$$\begin{aligned} P_{\theta^*}(\theta^* \in C_n) &= P_{\theta^*}\left(\theta^* \in [\hat{\Theta}_n - z_{\alpha/2}\hat{s}\hat{e}_n, \hat{\Theta}_n + z_{\alpha/2}\hat{s}\hat{e}_n]\right) \\ &= P_{\theta^*}\left(\hat{\Theta}_n - z_{\alpha/2}\hat{s}\hat{e}_n \leq \theta^* \leq \hat{\Theta}_n + z_{\alpha/2}\hat{s}\hat{e}_n\right) \\ &= P_{\theta^*}\left(-z_{\alpha/2}\hat{s}\hat{e}_n \leq \hat{\Theta}_n - \theta^* \leq z_{\alpha/2}\hat{s}\hat{e}_n\right) \\ &= P_{\theta^*}\left(-z_{\alpha/2} \leq \frac{\hat{\Theta}_n - \theta^*}{\hat{s}\hat{e}_n} \leq z_{\alpha/2}\right) \\ &\rightarrow P_{\theta^*}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \\ &= 1 - \alpha \end{aligned}$$

Figure 7.2: Density and Confidence Interval of the Asymptotically Normal Point Estimator

For 95% confidence intervals,  $\alpha = 0.05$  and  $z_{\alpha/2} = z_{0.025} = 1.96 \approx 2$ . This leads to the **approximate 95% confidence interval** of  $\hat{\theta}_n \pm 2\hat{s}\hat{e}_n$ , where  $\hat{\theta}_n = \hat{\Theta}_n(x_1, x_2, \dots, x_n)$  and  $x_1, x_2, \dots, x_n$  are the data or observations of the RVs  $X_1, X_2, \dots, X_n$ .

**Example 202 (Confidence interval for  $\theta^*$  from  $n$  Bernoulli( $\theta^*$ ) trials)** Let  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta^*)$  for some fixed but unknown parameter  $\theta^* \in \Theta = [0, 1]$ . Consider the following point estimator of  $\theta^*$ :

$$\hat{\Theta}_n((X_1, X_2, \dots, X_n)) = n^{-1} \sum_{i=1}^n X_i .$$

That is, we take the sample mean of the  $n$  IID Bernoulli( $\theta^*$ ) trials to be our point estimator of  $\theta^* \in [0, 1]$ . Then, we already saw that **this estimator is unbiased**

We already saw that the standard error  $s\hat{e}_n$  of this estimator is:

$$s\hat{e}_n = \sqrt{\theta^*(1 - \theta^*)/n} .$$

Since  $\theta^*$  is unknown, we obtain the estimated standard error  $\hat{s}_e_n$  from the point estimate  $\hat{\theta}_n$  of  $\theta^*$  on the basis of  $n$  observed data points  $x = (x_1, x_2, \dots, x_n)$  of the experiment:

$$\hat{s}_e_n = \sqrt{\hat{\theta}_n(1 - \hat{\theta}_n)/n}, \quad \text{where,} \quad \hat{\theta}_n = \hat{\Theta}_n((x_1, x_2, \dots, x_n)) = n^{-1} \sum_{i=1}^n x_i.$$

By the central limit theorem,  $\hat{\Theta}_n \rightsquigarrow \text{Normal}(\theta^*, \hat{s}_e_n)$ , i.e.  $\hat{\Theta}_n$  is asymptotically normal. Therefore, an asymptotically (for large sample size  $n$ ) approximate  $1 - \alpha$  normal-based confidence interval is:

$$\hat{\theta}_n \pm z_{\alpha/2} \hat{s}_e_n = \hat{\theta}_n \pm z_{\alpha/2} \sqrt{\frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}} := \left[ \hat{\theta}_n - z_{\alpha/2} \sqrt{\frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}}, \hat{\theta}_n + z_{\alpha/2} \sqrt{\frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}} \right]$$

We also saw that  $\hat{\Theta}_n$  is an **asymptotically consistent estimator** of  $\theta^*$  due to Proposition 125.

The confidence Interval for the coin tossing experiment in Example 195 with the observed sequence of Bernoulli outcomes (Heads  $\rightarrow 1$  and Tails  $\rightarrow 0$ ) being  $(1, 0, 0, 0, 1, 1, 0, 0, 1, 0)$ . We estimated the probability  $\theta^*$  of observing Heads with the **unbiased, asymptotically consistent estimator**  $\hat{\Theta}_n((X_1, X_2, \dots, X_n)) = n^{-1} \sum_{i=1}^n X_i$  of  $\theta^*$ . The point estimate of  $\theta^*$  was:

$$\begin{aligned} \hat{\theta}_{10} &= \hat{\Theta}_{10}((x_1, x_2, \dots, x_{10})) = \hat{\Theta}_{10}((1, 0, 0, 0, 1, 1, 0, 0, 1, 0)) \\ &= \frac{1 + 0 + 0 + 0 + 1 + 1 + 0 + 0 + 1 + 0}{10} = \frac{4}{10} = 0.40. \end{aligned}$$

The normal-based confidence interval for  $\theta^*$  may not be a valid approximation here with just  $n = 10$  samples. Nevertheless, we will compute a 95% normal-based confidence interval:

$$C_{10} = 0.40 \pm 1.96 \sqrt{\frac{0.40(1 - 0.40)}{10}} = 0.40 \pm 0.3036 = [0.0964, 0.7036]$$

with a width of 0.6072. When I increased the sample size  $n$  of the experiment from 10 to 100 by tossing the same coin another 90 times, I discovered that a total of 57 trials landed as Heads. Thus my point estimate and confidence interval for  $\theta^*$  are:

$$\hat{\theta}_{100} = \frac{57}{100} = 0.57 \quad \text{and} \quad C_{100} = 0.57 \pm 1.96 \sqrt{\frac{0.57(1 - 0.57)}{100}} = 0.57 \pm 0.0495 = [0.5205, 0.6195]$$

with a much smaller width of 0.0990. Thus our confidence interval shrank considerably from a width of 0.6072 after an additional 90 Bernoulli trials. Thus, we can make the width of the confidence interval as small as we want by making the number of observations or sample size  $n$  as large as we can.

## 7.7 Fundamentals of Hypothesis Testing

The subset of **all posable hypotheses** that remain **falsifiable** is the space of **scientific hypotheses**. Roughly, a falsifiable hypothesis is one for which a statistical experiment can be designed to produce data that an experimenter can use to falsify or reject it. In the statistical decision problem of hypothesis testing, we are interested in empirically falsifying a scientific hypothesis, i.e. we attempt to reject an hypothesis on the basis of empirical observations or data. Thus, hypothesis testing has its roots in the philosophy of science and is based on Karl Popper's falsifiability criterion for demarcating scientific hypotheses from the set of all posable hypotheses.

### 7.7.1 Introduction

Usually, the hypothesis we attempt to reject or falsify is called the **null hypothesis** or  $H_0$  and its complement is called the **alternative hypothesis** or  $H_1$ . For example, consider the following two hypotheses:

$H_0$ : The average waiting time at an Orbiter bus stop is less than or equal to 10 minutes.

$H_1$ : The average waiting time at an Orbiter bus stop is more than 10 minutes.

If the sample mean  $\bar{x}_n$  is much larger than 10 minutes then we may be inclined to reject the null hypothesis that the average waiting time is less than or equal to 10 minutes. We will learn to formally test hypotheses in the sequel.

Suppose we are interested in the following hypothesis test for the bus-stop problem:

$H_0$ : The average waiting time at an Orbiter bus stop is equal to 10 minutes.

$H_1$ : The average waiting time at an Orbiter bus stop is not 10 minutes.

Once again we can use the sample mean as the test statistic. Our procedure for rejecting this null hypothesis is different and is often called the Wald test.

More generally, suppose  $X_1, X_2, \dots, X_n \stackrel{IID}{\sim} F(x_1; \theta^*)$ , with an unknown and fixed  $\theta^* \in \Theta$ . Let us partition the parameter space  $\Theta$  into  $\Theta_0$ , the null parameter space, and  $\Theta_1$ , the alternative parameter space, ie,

$$\Theta_0 \cup \Theta_1 = \Theta, \quad \text{and} \quad \Theta_0 \cap \Theta_1 = \emptyset.$$

Then, we can formalise testing the null hypothesis versus the alternative as follows:

$$H_0 : \theta^* \in \Theta_0 \quad \text{versus} \quad H_1 : \theta^* \in \Theta_1.$$

The basic idea involves finding an appropriate rejection region  $\mathbb{X}_R$  within the data space  $\mathbb{X}$  and rejecting  $H_0$  if the observed data  $x := (x_1, x_2, \dots, x_n)$  falls inside the rejection region  $\mathbb{X}_R$ ,

If  $x := (x_1, x_2, \dots, x_n) \in \mathbb{X}_R \subset \mathbb{X}$ , then reject  $H_0$ , else do not reject  $H_0$ .

Typically, the rejection region  $\mathbb{X}_R$  is of the form:

$$\mathbb{X}_R := \{x := (x_1, x_2, \dots, x_n) : T(x) > c\}$$

where,  $T$  is the **test statistic** and  $c$  is the **critical value**. Thus, the problem of finding  $\mathbb{X}_R$  boils down to that of finding  $T$  and  $c$  that are appropriate. Once the rejection region is defined, the possible outcomes of a hypothesis test are summarised in the following table.

**Definition 129 (Power, Size and Level of a Test)** The **power function** of a test with rejection region  $\mathbb{X}_R$  is

$$\beta(\theta) := P_\theta(x \in \mathbb{X}_R). \tag{7.51}$$

Table 7.1: Outcomes of an hypothesis test.

	Do not Reject $H_0$	Reject $H_0$
$H_0$ is True	OK	Type I Error
$H_1$ is True	Type II Error	OK

So  $\beta(\theta)$  is the power of the test at the parameter value  $\theta$ , i.e. the probability that the observed data  $x$ , sampled from the distribution specified by  $\theta$ , falls in  $\mathbb{X}_R$  and thereby leads to a rejection of the null hypothesis.

The size of a test with rejection region  $\mathbb{X}_R$  is the supreme power under the null hypothesis, i.e. the supreme probability of rejecting the null hypothesis when the null hypothesis is true:

$$\text{size} := \sup_{\theta \in \Theta_0} \beta(\theta) := \sup_{\theta \in \Theta_0} P_\theta(x \in \mathbb{X}_R) . \quad (7.52)$$

The size of a test is often denoted by  $\alpha$ . A test is said to have level  $\alpha$  if its size is less than or equal to  $\alpha$ .

Let us familiarize ourselves with some terminology in hypothesis testing next.

Table 7.2: Some terminology in hypothesis testing.

$\Theta$	Test: $H_0$ versus $H_1$	Nomenclature
$\Theta \subset \mathbb{R}^m, m \geq 1$	$H_0 : \theta^* = \theta_0$ versus $H_1 : \theta^* \neq \theta_1$	Simple Hypothesis Test
$\Theta \subset \mathbb{R}^m, m \geq 1$	$H_0 : \theta^* \in \Theta_0$ versus $H_1 : \theta^* \in \Theta_1$	Composite Hypothesis Test
$\Theta \subset \mathbb{R}^1$	$H_0 : \theta^* = \theta_0$ versus $H_1 : \theta^* \neq \theta_0$	Two-sided Hypothesis Test
$\Theta \subset \mathbb{R}^1$	$H_0 : \theta^* \geq \theta_0$ versus $H_1 : \theta^* < \theta_0$	One-sided Hypothesis Test
$\Theta \subset \mathbb{R}^1$	$H_0 : \theta^* \leq \theta_0$ versus $H_1 : \theta^* > \theta_0$	One-sided Hypothesis Test

We introduce some widely used tests next.

### 7.7.2 The Wald Test

The Wald test is based on a direct relationship between the  $1 - \alpha$  confidence interval and a size  $\alpha$  test. It can be used for testing simple hypotheses involving a scalar parameter.

**Definition 130 (The Wald Test)** Let  $\widehat{\Theta}_n$  be an asymptotically normal estimator of the fixed and possibly unknown parameter  $\theta^* \in \Theta \subset \mathbb{R}$  in the parametric IID experiment:

$$X_1, X_2, \dots, X_n \stackrel{IID}{\sim} F(x_1; \theta^*) .$$

Consider testing:

$$H_0 : \theta^* = \theta_0 \quad \text{versus} \quad H_1 : \theta^* \neq \theta_0 .$$

Suppose that the null hypothesis is true and the estimator  $\widehat{\Theta}_n$  of  $\theta^* = \theta_0$  is asymptotically normal:

$$\theta^* = \theta_0, \quad \frac{\widehat{\Theta}_n - \theta_0}{\widehat{s}_{\Theta_n}} \rightsquigarrow \text{Normal}(0, 1) .$$

Then, the Wald test based on the test statistic  $W$  is:

$$\text{Reject } H_0 \text{ when } |W| > z_{\alpha/2}, \text{ where } W := W((X_1, \dots, X_n)) = \frac{\widehat{\Theta}_n((X_1, \dots, X_n)) - \theta_0}{\widehat{\text{se}}_n}.$$

The rejection region for the Wald test is:

$$\mathbb{X}_R = \{x := (x_1, \dots, x_n) : |W(x_1, \dots, x_n)| > z_{\alpha/2}\} .$$

**Proposition 131 (Asymptotic size of a Wald test)** As the sample size  $n$  approaches infinity, the size of the Wald test approaches  $\alpha$  :

$$\text{size} = P_{\theta_0} (|W| > z_{\alpha/2}) \rightarrow \alpha .$$

**Proof:** Let  $Z \sim \text{Normal}(0, 1)$ . The size of the Wald test, i.e. the supreme power under  $H_0$  is:

$$\begin{aligned} \text{size} &:= \sup_{\theta \in \Theta_0} \beta(\theta) := \sup_{\theta \in \{\theta_0\}} P_{\theta}(x \in \mathbb{X}_R) = P_{\theta_0}(x \in \mathbb{X}_R) \\ &= P_{\theta_0} (|W| > z_{\alpha/2}) = P_{\theta_0} \left( \frac{|\widehat{\theta}_n - \theta_0|}{\widehat{\text{se}}_n} > z_{\alpha/2} \right) \\ &\rightarrow P (|Z| > z_{\alpha/2}) \\ &= \alpha . \end{aligned}$$

Next, let us look at the power of the Wald test when the null hypothesis is false.

**Proposition 132 (Asymptotic power of a Wald test)** Suppose  $\theta^* \neq \theta_0$ . The power  $\beta(\theta^*)$ , which is the probability of correctly rejecting the null hypothesis, is approximately equal to:

$$\Phi \left( \frac{\theta_0 - \theta^*}{\widehat{\text{se}}_n} - z_{\alpha/2} \right) + \left( 1 - \Phi \left( \frac{\theta_0 - \theta^*}{\widehat{\text{se}}_n} + z_{\alpha/2} \right) \right) ,$$

where,  $\Phi$  is the DF of  $\text{Normal}(0, 1)$  RV. Since  $\widehat{\text{se}}_n \rightarrow 0$  as  $n \rightarrow \infty$  the power increase with sample size  $n$ . Also, the power increases when  $|\theta_0 - \theta^*|$  is large.

Now, let us make the connection between the size  $\alpha$  Wald test and the  $1 - \alpha$  confidence interval explicit.

**Proposition 133 (The size Wald test)** The size  $\alpha$  Wald test rejects:

$$H_0 : \theta^* = \theta_0 \text{ versus } H_1 : \theta^* \neq \theta_0 \text{ if and only if } \theta_0 \notin C_n := (\widehat{\theta}_n - \widehat{\text{se}}_n z_{\alpha/2}, \widehat{\theta}_n + \widehat{\text{se}}_n z_{\alpha/2}).$$

Therefore, testing the hypothesis is equivalent to verifying whether the null value  $\theta_0$  is in the confidence interval.

**Example 203 (Wald test for the mean waiting times at our Orbiter bus-stop)** Let us use the Wald test to attempt to reject the null hypothesis that the mean waiting time at our Orbiter bus-stop is 10 minutes under an IID Exponential( $\lambda^*$ ) model. Let  $\alpha = 0.05$  for this test. We can formulate this test as follows:

$$H_0 : \lambda^* = \lambda_0 = \frac{1}{10} \text{ versus } H_1 : \lambda^* \neq \frac{1}{10}, \text{ where, } X_1, \dots, X_{132} \stackrel{IID}{\sim} \text{Exponential}(\lambda^*) .$$

Based on Example 220 and Labwork 221 we obtained the 95% confidence interval to be [0.0914, 0.1290]. Since our null value  $\lambda_0 = 0.1$  belongs to this confidence interval, we fail to reject the null hypothesis from a size  $\alpha = 0.05$  Wald test.

We can use bootstrap-based confidence interval  $C_n$  in conjunction with Wald test as shown by the next example.

**Example 204 (Wald test of the bootstrapped correlation coefficient)** Recall the problem of estimating the confidence interval for the correlation coefficient between the LSAT scores ( $Y_1, \dots, Y_{15}$ ) and the GPA ( $Z_1, \dots, Z_{15}$ ) in Labwork 233. We assumed that the bivariate data  $(Y_i, Z_i) \stackrel{IID}{\sim} F^*$ , such that  $F^* \in \{\text{all bivariate DFs}\}$ . Suppose we are interested in testing the null hypothesis that the true correlation coefficient  $\theta^*$  is 0:

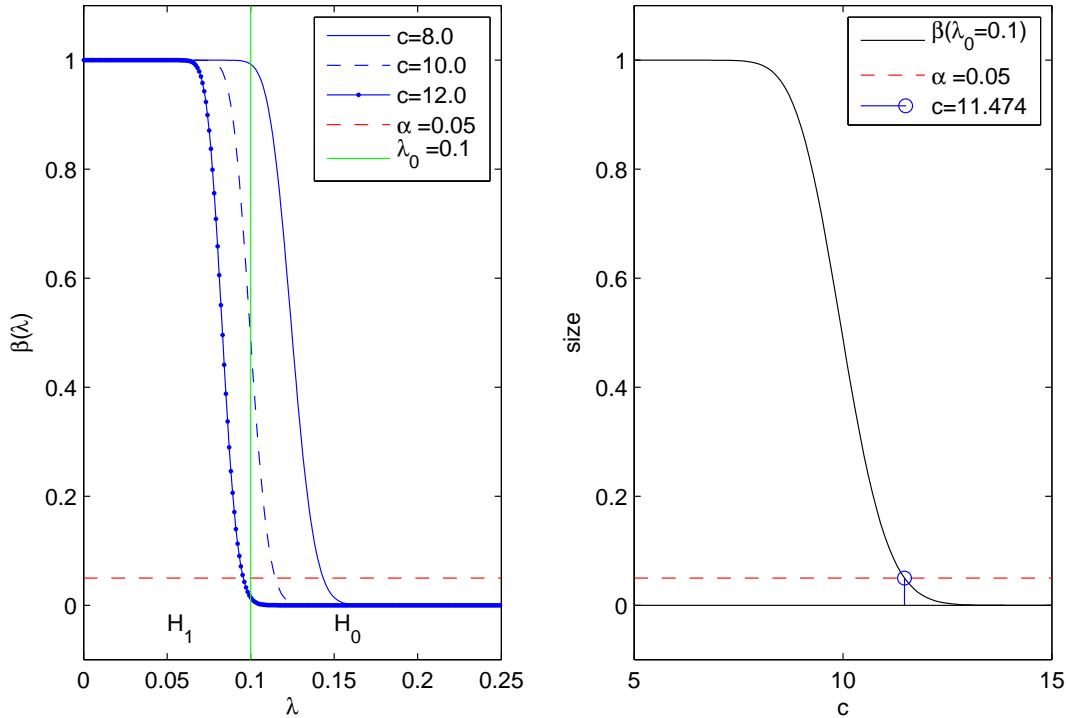
$$H_0 : \theta^* = \theta_0 = 0 \quad \text{versus} \quad H_1 : \theta^* \neq 0, \quad \text{where} \quad \theta^* = \frac{\int \int (y - E(Y))(z - E(Z)) dF(y, z)}{\sqrt{\int (y - E(Y))^2 dF(y) \int (z - E(Z))^2 dF(z)}}.$$

Since the percentile-based 95% bootstrap confidence interval for the plug-in estimate of the correlation coefficient from Labwork 233 was  $[0.2346, 0.9296]$  and this interval does not contain 0, we can reject the null hypothesis that the correlation coefficient is 0 using a size  $\alpha = 0.05$  Wald test.

### 7.7.3 A Composite Hypothesis Test

Often, we are interested in testing a composite hypothesis, i.e. one in which the null hypothesis is not a singleton set. We revisit the Orbiter waiting time problem from this perspective next.

Figure 7.3: Plot of power function  $\beta(\lambda)$  for different values of the critical value  $c$  and the size  $\alpha$  as function of the critical values.



**Example 205 (Testing the Mean Waiting Time at an Orbiter Bus-stop)** Let us test the following null hypothesis  $H_0$ .

$H_0$ : The average waiting time at an Orbiter bus stop is less than or equal to 10 minutes.

$H_1$ : The average waiting time at an Orbiter bus stop is more than 10 minutes.

We have observations of  $n = 132$  waiting times  $x_1, x_2, \dots, x_{132}$  at the Orbiter bus-stop with  $\bar{x}_{132} = 9.0758$ . Let us assume a parametric model, say,

$$X_1, X_2, \dots, X_n \stackrel{IID}{\sim} \text{Exponential}(\lambda^*)$$

with an unknown and fixed  $\lambda^* \in \Lambda = (0, \infty)$ . Since the parameter  $\lambda$  of an  $\text{Exponential}(\lambda)$  RV is the reciprocal of the mean waiting time, we can formalise the above hypothesis testing problem of  $H_0$  versus  $H_1$  as follows:

$$H_0 : \lambda^* \in \Lambda_0 = [1/10, \infty) \quad \text{versus} \quad H_1 : \lambda^* \in \Lambda_1 = (0, 1/10)$$

Consider the test:

Reject  $H_0$  if  $T > c$ .

where the test statistic  $T = \bar{X}_n$  and the rejection region is:

$$\mathbb{X}_R = \{(x_1, x_2, \dots, x_n) : T(x_1, x_2, \dots, x_n) > c\} .$$

Since the sum of  $n$  IID  $\text{Exponential}(\lambda)$  RVs is  $\text{Gamma}(\lambda, n)$  distributed, the power function is:

$$\begin{aligned} \beta(\lambda) &= P_\lambda(\bar{X}_n > c) = P_\lambda\left(\sum_{i=1}^n X_i > nc\right) = 1 - P_\lambda\left(\sum_{i=1}^n X_i \leq nc\right) \\ &= 1 - F(nc; \lambda, n) = 1 - \frac{1}{\Gamma(n)} \int_0^{\lambda nc} y^{n-1} \exp(-y) dy \\ &= 1 - \text{gammainc}(\lambda nc, n) \end{aligned}$$

Clearly,  $\beta(\lambda)$  is a decreasing function of  $\lambda$  as shown in Figure 7.3. Hence the size of the test as a function of the critical region specified by the critical value  $c$  is:

$$\text{size} = \sup_{\lambda \in \Lambda_0} \beta(\lambda) = \sup_{\lambda \geq 1/10} \beta(\lambda) = \beta(1/10) = 1 - \text{gammainc}(132c/10, 132)$$

For a size  $\alpha = 0.05$  test we numerically solve for the critical value  $c$  that satisfies:

$$0.05 = 1 - \text{gammainc}(132c/10, 132)$$

by trial and error as follows:

```
>> lambda0=1/10
lambda0 = 0.1000
>> S=@(C)(1-gammainc(lambda0*n*C,n)); % size as a function of c
>> Cs=[10 11 11.474 12 13] % some critical values c
Cs = 10.0000 11.0000 11.4740 12.0000 13.0000
>> Size=arrayfun(S,Cs) % corresponding size
Size = 0.4884 0.1268 0.0499 0.0143 0.0007
```

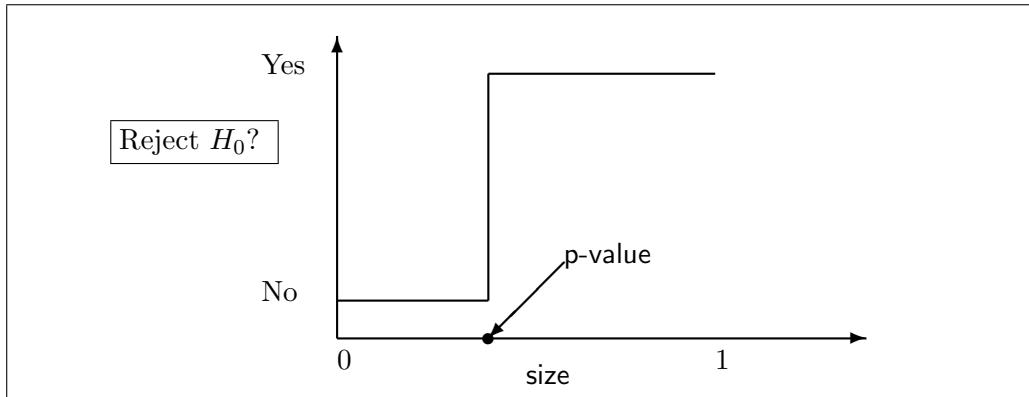
Thus, we reject  $H_0$  when  $\bar{X}_n > 11.4740$  for a level  $\alpha = 0.05$  test. Since our observed test statistic  $\bar{x}_{132} = 9.0758 < 11.4740$  we fail to reject the null hypothesis that the mean waiting time is less than or equal to 10 minutes. Therefore, there is no evidence that the Orbiter bus company is violating its promise of an average waiting time of no more than 10 minutes.

Table 7.3: Evidence scale against the null hypothesis in terms of the range of p-value.

p-value range	Evidence
(0, 0.01]	very strong evidence against $H_0$
(0.01, 0.05]	strong evidence against $H_0$
(0.05, 0.1]	weak evidence against $H_0$
(0.1, 1)	little or no evidence against $H_0$

#### 7.7.4 p-values

It is desirable to have a more informative decision than simply reporting "reject  $H_0$ " or "fail to reject  $H_0$ ." For instance, we could ask whether the test rejects  $H_0$  for each `size =  $\alpha$` . Typically, if the test rejects at `size  $\alpha$`  it will also reject at a larger `size  $\alpha' > \alpha$` . Therefore, there is a smallest `size  $\alpha$`  at which the test rejects  $H_0$  and we call this  $\alpha$  the **p-value** of the test.

Figure 7.4: The smallest  $\alpha$  at which a size  $\alpha$  test rejects the null hypothesis  $H_0$  is the p-value.

**Definition 134 (p-value)** Suppose that for every  $\alpha \in (0, 1)$  we have a size  $\alpha$  test with rejection region  $\mathbb{X}_{R,\alpha}$  and test statistic  $T$ . Then,

$$\text{p-value} := \inf\{\alpha : T(X) \in \mathbb{X}_{R,\alpha}\} .$$

That is, the p-value is the smallest  $\alpha$  at which a size  $\alpha$  test rejects the null hypothesis.

If the evidence against  $H_0$  is strong then the p-value will be small. However, a large p-value is not strong evidence in favour of  $H_0$ . This is because a large p-value can occur for two reasons:

1.  $H_0$  is true.
2.  $H_0$  is false but the test has low power.

Finally, it is important to realise that p-value is not the probability that the null hypothesis is true, i.e.  $\text{p-value} \neq P(H_0|x)$ , where  $x$  is the data. The following tabulation of evidence scale is useful. The next proposition gives a convenient expression for the p-value for certain tests.

**Proposition 135 (The p-value of a hypothesis test)** Suppose that the size  $\alpha$  test based on the test statistic  $T$  and critical value  $c_\alpha$  is of the form:

$$\text{Reject } H_0 \text{ if and only if } T := T((X_1, \dots, X_n)) > c_\alpha,$$

then

$$\text{p-value} = \sup_{\theta \in \Theta_0} P_\theta(T((X_1, \dots, X_n)) \geq t := T((x_1, \dots, x_n))) ,$$

where,  $(x_1, \dots, x_n)$  is the observed data and  $t$  is the observed value of the test statistic  $T$ . In words, the p-value is the supreme probability under  $H_0$  of observing a value of the test statistic the same as or more extreme than what was actually observed.

Let us revisit the Orbiter waiting times example from the p-value perspective.

**Example 206 (p-value for the parametric Orbiter experiment)** Let the waiting times at our bus-stop be  $X_1, X_2, \dots, X_{132} \stackrel{\text{IID}}{\sim} \text{Exponential}(\lambda^*)$ . Consider the following testing problem:

$$H_0 : \lambda^* = \lambda_0 = \frac{1}{10} \quad \text{versus} \quad H_1 : \lambda^* \neq \lambda_0 .$$

We already saw that the Wald test statistic is:

$$W := W(X_1, \dots, X_n) = \frac{\widehat{\Lambda}_n - \lambda_0}{\widehat{s}_{\Lambda}( \widehat{\Lambda}_n )} = \frac{\frac{1}{\bar{X}_n} - \lambda_0}{\frac{1}{\sqrt{n} \bar{X}_n}} .$$

The observed test statistic is:

$$w = W(x_1, \dots, x_{132}) = \frac{\frac{1}{\bar{X}_{132}} - \lambda_0}{\frac{1}{\sqrt{132} \bar{X}_{132}}} = \frac{\frac{1}{9.0758} - \frac{1}{10}}{\frac{1}{\sqrt{132} \times 9.0758}} = 1.0618 .$$

Since,  $W \rightsquigarrow Z \sim \text{Normal}(0, 1)$ , the p-value for this Wald test is:

$$\begin{aligned} \text{p-value} &= \sup_{\lambda \in \Lambda_0} P_\lambda(|W| > |w|) = \sup_{\lambda \in \{\lambda_0\}} P_\lambda(|W| > |w|) = P_{\lambda_0}(|W| > |w|) \\ &\rightarrow P(|Z| > |w|) = 2\Phi(-|w|) = 2\Phi(-|1.0618|) = 2 \times 0.1442 = 0.2884 . \end{aligned}$$

Therefore, there is little or no evidence against  $H_0$  that the mean waiting time under an IID Exponential model of inter-arrival times is exactly ten minutes.

### 7.7.5 Permutation Test for the equality of any two DFs

Permutation test is a non-parametric exact method for testing whether two distributions are the same. It is non-parametric because we do not impose any restrictions on the class of DFs that the unknown DF should belong to. It is exact because we do not have any asymptotic approximations involving sample size approaching infinity. So this test works for any sample size.

Formally, we suppose that:

$$X_1, X_2, \dots, X_m \stackrel{\text{IID}}{\sim} F^* \quad \text{and} \quad X_{m+1}, X_{m+2}, \dots, X_{m+n} \stackrel{\text{IID}}{\sim} G^* ,$$

are two sets of independent samples. The possibly unknown DFs  $F^*, G^* \in \{\text{all DFs}\}$ . Now, consider the following hypothesis test:

$$H_0 : F^* = G^* \quad \text{versus} \quad H_1 : F^* \neq G^* .$$

Let our test statistic  $T(X_1, \dots, X_m, X_{m+1}, \dots, X_{m+n})$  be some sensible one –  $T$  is large when  $F^*$  is too different from  $G^*$ , say:

$$T := T(X_1, \dots, X_m, X_{m+1}, \dots, X_{m+n}) = \text{abs} \left( \frac{1}{m} \sum_{i=1}^m X_i - \frac{1}{n} \sum_{i=m+1}^n X_i \right) .$$

Then the idea of a permutation test is as follows:

1. Let  $N := m + n$  be the pooled sample size and consider all  $N!$  permutations of the observed data  $x_{\text{obs}} := (x_1, x_2, \dots, x_m, x_{m+1}, x_{m+2}, \dots, x_{m+n})$ .
2. For each permutation of the data compute the statistic  $T(\text{permuted data } x)$  and denote these  $N!$  values of  $T$  by  $t_1, t_2, \dots, t_{N!}$ .
3. Under  $H_0 : X_1, \dots, X_m, X_{m+1}, \dots, X_{m+n} \stackrel{\text{IID}}{\sim} F^* = G^*$ , each of the permutations of  $x = (x_1, x_2, \dots, x_m, x_{m+1}, x_{m+2}, \dots, x_{m+n})$  has the same joint probability  $\prod_{i=1}^{m+n} f(x_i)$ , where  $f(x_i) = dF(x_i) = dG(x_i)$ . Therefore, the transformation of the data by our statistic  $T$  also has the same probability over the values of  $T$ , namely  $\{t_1, t_2, \dots, t_{N!}\}$ . Let  $P_0$  be this permutation distribution that is discrete and uniform over  $\{t_1, t_2, \dots, t_{N!}\}$ .
4. Let  $t_{\text{obs}} := T(x_{\text{obs}})$  be the observed value of the statistic.
5. Assuming we reject  $H_0$  when  $T$  is large, the p-value is:

$$\text{p-value} = P_0(T \geq t_{\text{obs}}) = \frac{1}{N!} \left( \sum_{j=1}^{N!} \mathbf{1}(t_j \geq t_{\text{obs}}) \right), \quad \mathbf{1}(t_j \geq t_{\text{obs}}) = \begin{cases} 1 & \text{if } t_j \geq t_{\text{obs}} \\ 0 & \text{otherwise} \end{cases}$$

Let us look at a small example involving the diameters of coarse venus shells (*Dosinia anus*) that Guo Yaozong and Chen Shun found on the left and right sides of the New Brighton pier in Spring 2007. We are interested in testing the hypothesis that the distribution of shell diameters for this bivalve species is the same on both sides of the pier.

**Example 207 (Guo-Chen Experiment with Venus Shell Diameters)** Let us look at the first two samples  $x_1$  and  $x_2$  from the left of pier and the first sample from the right side of pier, namely  $x_3$ . Since the permutation test is exact, we can use this small data set with merely three samples to conduct the following hypothesis test:

$$H_0 : X_1, X_2, X_3 \stackrel{\text{IID}}{\sim} F^* = G^*, \quad H_1 : X_1, X_2 \stackrel{\text{IID}}{\sim} F^*, X_3 \stackrel{\text{IID}}{\sim} G^*, \quad F^* \neq G^* .$$

Let us use the test statistic:

$$T(X_1, X_2, X_3) = \text{abs} \left( \frac{1}{2} \sum_{i=1}^2 X_i - \frac{1}{1} \sum_{i=2+1}^3 X_i \right) = \text{abs} \left( \frac{X_1 + X_2}{2} - \frac{X_3}{1} \right) .$$

The data giving the shell diameters in millimetres and  $t_{\text{obs}}$  are:

$$(x_1, x_2, x_3) = (52, 54, 58) \quad \text{and} \quad t_{\text{obs}} = \text{abs} \left( \frac{52 + 54}{2} - \frac{58}{1} \right) = \text{abs}(53 - 58) = \text{abs}(-5) = 5 .$$

Let us tabulate the  $(2+1)! = 3! = 3 \times 2 \times 1 = 6$  permutations of the data  $(x_1, x_2, x_3) = (52, 54, 58)$ , the corresponding values of  $T$  and their probabilities under the null hypothesis, i.e., the permutation distribution  $P_0(T)$ .

Permutation	$t$	$P_0(T = t)$
(52, 54, 58)	5	$\frac{1}{6}$
(54, 52, 58)	5	$\frac{1}{6}$
(52, 58, 54)	1	$\frac{1}{6}$
(58, 52, 54)	1	$\frac{1}{6}$
(58, 54, 52)	4	$\frac{1}{6}$
(54, 58, 52)	4	$\frac{1}{6}$

From the table, we get:

$$\text{p-value} = P_0(T \geq t_{\text{obs}}) = P_0(T \geq 5) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3} \approx 0.333 .$$

Therefore, there is little to no evidence against  $H_0$ .

When the pooled sample size  $N = m + n$  gets large,  $N!$  would be too numerous to tabulate exhaustively. In this situation, we can use a Monte Carlo approximation of the p-value by generating a large number of random permutations of the data according to the following Steps:

Step 1: Compute the observed statistic  $t_{\text{obs}} := T(x_{\text{obs}})$  of data  $x_{\text{obs}} := (x_1, \dots, x_m, x_{m+1}, \dots, x_{m+n})$ .

Step 2: Randomly permute the data and compute the statistic again from the permuted data.

Step 3: Repeat Step 2  $B$  times and let  $t_1, \dots, t_B$  denote the resulting values ( $B$  is large, say 1000).

Step 4: The (Monte Carlo) approximate p-value is:

$$\frac{1}{B} \sum_{j=1}^B \mathbf{1}(t_j \geq t_{\text{obs}}) .$$

Next we implement the above algorithm on the full data set of Guo and Chen obtained from coarse venus shells sampled from the two sides of the New Brighton pier.

**Labwork 208 (Approximate p-value of a permutation test of shell diameters)** Test the null hypothesis that the distribution of the diameters of coarse venus shells are the same on both sides of the New Brighton pier.

---

```
Shells.m
_____
% this data was collected by Guo Yaozong and Chen Shun as part of their STAT 218 project 2007
% coarse venus shell diameters in mm from left side of New Brighton Pier
left=[52 54 60 60 54 47 57 58 61 57 50 60 60 62 44 55 58 55 60 59 65 59 63 51 61 62 61 60 61 65 ...
43 59 58 67 56 64 47 64 60 55 58 41 53 61 60 49 48 47 42 50 58 48 59 55 59 50 47 47 33 51 61 61 ...
52 62 64 64 47 58 58 61 50 55 47 39 59 64 63 63 62 64 61 50 62 61 65 62 66 60 59 58 58 60 59 61 ...
55 55 62 51 61 49 52 59 60 66 50 59 64 64 62 60 65 44 58 63];
% coarse venus shell diameters in mm from right side of New Brighton Pier
right=[58 54 60 55 56 44 60 52 57 58 61 66 56 59 49 48 69 66 49 72 49 50 59 59 59 66 62 ...
44 49 40 59 55 61 51 62 52 63 39 63 52 62 49 48 65 68 45 63 58 55 56 55 57 34 64 66 ...
54 65 61 56 57 59 58 62 58 40 43 62 59 64 64 65 65 59 64 63 65 62 61 47 59 63 44 43 ...
59 67 64 60 62 64 65 59 55 38 57 61 52 61 61 60 34 62 64 58 39 63 47 55 54 48 60 55 ...
60 65 41 61 59 65 50 54 60 48 51 68 52 51 61 57 49 51 62 63 59 62 54 59 46 64 49 61];
Tobs=abs(mean(left)-mean(right));% observed test statistic
nleft=length(left); % sample size of the left-side data
nright=length(right); % sample size of the right-side data
ntotal=nleft+nright; % sample size of the pooled data
total=[left right]; % observed data -- ordered: left-side data followed by right-side data
B=10000; % number of bootstrap replicates
```

```

TB=zeros(1,B); % initialise a vector of zeros for the bootstrapped test statistics
ApproxPValue=0; % initialise an accumulator for approximate p-value
for b=1:B % enter the bootstrap replication loop
    % use MATLAB's randperm function to get a random permutation of indices{1,2,...,ntotal}
    PermutatedIndices=randperm(ntotal);
    % use the first nleft of the PermutatedIndices to get the bootstrapped left-side data
    Bleft=total(PermutatedIndices(1:nleft));
    % use the last nright of the PermutatedIndices to get the bootstrapped right-side data
    Bright=total(PermutatedIndices(nleft+1:ntotal));
    TB(b) = abs(mean(Bleft)-mean(Bright)); % compute the test statistic for the bootstrapped data
    if(TB(b)>Tobs) % increment the ApproxPValue accumulator by 1/B if bootstrapped value > Tobs
        ApproxPValue=ApproxPValue+(1/B);
    end
end
ApproxPValue % report the Approximate p-value

```

---

When we execute the script to perform a permutation test and approximate the p-value, we obtain:

```

>> Shells
ApproxPValue =      0.8576

```

Therefore, there is little or no evidence against the null hypothesis.

### 7.7.6 Pearson's Chi-Square Test for Multinomial Trials

We derive the Chi-square distribution introduced by Karl Pearson in 1900 [*Philosophical Magazine*, Series 5, **50**, 157-175]. This historical work laid the foundations of modern statistics by showing why an experimenter cannot simply plot experimental data and just assert the correctness of his or her hypothesis. This derivation is adapted from Donald E. Knuth's treatment [*Art of Computer Programming, Vol. II, Seminumerical Algorithms*, 3rd Ed., 1997, pp. 55-56]. We show how de Moivre, Multinomial, Poisson and the Normal random vectors conspire to create the Chi-square random variable.

*Part 1: de Moivre trials*

Consider  $n$  independent and identically distributed de Moivre( $\theta_1, \dots, \theta_k$ ) random vectors ( $\vec{RV}$ s):

$$X_1, X_2, \dots, X_n \stackrel{IID}{\sim} \text{de Moivre}(\theta_1, \dots, \theta_k).$$

Recall from Model 16 that  $X_1 \sim \text{de Moivre}(\theta_1, \dots, \theta_k)$  means  $P(X_1 = e_i) = \theta_i$  for  $i \in \{1, \dots, k\}$ , where  $e_1, \dots, e_k$  are ortho-normal basis vectors in  $\mathbb{R}^k$ . Thus, for each  $i \in \{1, 2, \dots, n\}$ , the corresponding  $X_i$  has  $k$  components, i.e.  $X_i := (X_{i,1}, X_{i,2}, \dots, X_{i,k})$ .

*Part 2: Multinomial trial*

Suppose we are only interested in the experiment induced by their sum:

$$Y := (Y_1, \dots, Y_k) := \sum_{i=1}^n X_i := \sum_{i=1}^n (X_{i,1}, X_{i,2}, \dots, X_{i,k}) = \left( \sum_{i=1}^n X_{i,1}, \sum_{i=1}^n X_{i,2}, \dots, \sum_{i=1}^n X_{i,k} \right).$$

The  $\vec{RV}$   $Y$ , being the sum of  $n$  IID de Moivre( $\theta_1, \dots, \theta_k$ )  $\vec{RV}$ s, is the Multinomial( $n, \theta_1, \dots, \theta_k$ )  $\vec{RV}$  of Model 17 and the probability that  $Y := (Y_1, \dots, Y_k) = y := (y_1, \dots, y_k)$  is:

$$\frac{n!}{y_1!y_2!\cdots y_k!} \prod_{i=1}^k \theta_i^{y_i}.$$

The support of the R $\vec{V} Y$ , i.e. the set of possible realisations of  $y := (y_1, \dots, y_k)$  is:

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

*Part 3: Conditional sum of Poisson trials*

Here we consider an alternative formulation of the Multinomial( $n, \theta_1, \dots, \theta_k$ ) R $\vec{V} Y$ . Suppose,

$$Y_1 \sim \text{Poisson}(n\theta_1), Y_2 \sim \text{Poisson}(n\theta_2), \dots, Y_k \sim \text{Poisson}(n\theta_k) ,$$

and that  $Y_1, \dots, Y_k$  are independent. Recall from Model 6 that  $Y_i \sim \text{Poisson}(n\theta_i)$  means  $P(Y_i = y_i) = e^{-n\theta_i} (n\theta_i)^{y_i} / y_i!$  for  $y_i \in \{0, 1, \dots\}$ . Then, the joint probability probability of the R $\vec{V} (Y_1, \dots, Y_k)$  is the product of the independent Poisson probabilities:

$$\begin{aligned} P((Y_1, \dots, Y_k) = (y_1, \dots, y_k)) &:= P(Y_1 = y_1, \dots, Y_k = y_k) = \prod_{i=1}^k P(Y_i = y_i) = \prod_{i=1}^k \frac{e^{-n\theta_i} (n\theta_i)^{y_i}}{y_i!} \\ &= \frac{\prod_{i=1}^k e^{-n\theta_i} n^{y_i} \theta_i^{y_i}}{\prod_{i=1}^k y_i!} = \left( e^{-n \sum_{i=1}^k \theta_i} n^{\sum_{i=1}^k y_i} \prod_{i=1}^k \theta_i^{y_i} \right) \frac{1}{\prod_{i=1}^k y_i!} \\ &= \frac{e^{-n} n^n \prod_{i=1}^k \theta_i^{y_i}}{\prod_{i=1}^k y_i!} . \end{aligned}$$

Now, the probability that sum  $Y_1 + \dots + Y_k$  will equal  $n$  is obtained by summing over the probabilities of all  $(y_1, \dots, y_k) \in \mathbb{Y}$ :

$$\begin{aligned} P\left(\sum_{i=1}^k Y_i = n\right) &= \sum_{\substack{(y_1, \dots, y_k) \\ \in \mathbb{Y}}} P((Y_1, \dots, Y_k) = (y_1, \dots, y_k)) = \sum_{\substack{(y_1, \dots, y_k) \\ \in \mathbb{Y}}} \frac{e^{-n} n^n \prod_{i=1}^k \theta_i^{y_i}}{\prod_{i=1}^k y_i!} \\ &= e^{-n} n^n \underbrace{\frac{1}{n!} \left( \sum_{\substack{(y_1, \dots, y_k) \\ \in \mathbb{Y}}} \frac{n!}{\prod_{i=1}^k y_i!} \prod_{i=1}^k \theta_i^{y_i} \right)}_{=P(\mathbb{Y})=1} = \frac{e^{-n} n^n}{n!} . \end{aligned}$$

Finally, the conditional probability that  $(Y_1, \dots, Y_k) = (y_1, \dots, y_k)$  given  $\sum_{i=1}^k Y_i = n$  is:

$$\begin{aligned} P\left((Y_1, \dots, Y_k) = (y_1, \dots, y_k) \mid \sum_{i=1}^k Y_i = n\right) &= \frac{P((Y_1, \dots, Y_k) = (y_1, \dots, y_k), \sum_{i=1}^k Y_i = n)}{P(\sum_{i=1}^k Y_i = n)} \\ &= \frac{P((Y_1, \dots, Y_k) = (y_1, \dots, y_k))}{P(\sum_{i=1}^k Y_i = n)} = \frac{e^{-n} n^n \prod_{i=1}^k \theta_i^{y_i}}{\prod_{i=1}^k y_i!} \frac{n!}{e^{-n} n^n} = \frac{n!}{y_1! y_2! \cdots y_k!} \prod_{i=1}^k \theta_i^{y_i} . \end{aligned}$$

Therefore, we may also think of the random vector  $Y := (Y_1, \dots, Y_k) \sim \text{Multinomial}(n, \theta_1, \dots, \theta_k)$  as  $k$  independent Poisson random variables,  $Y_1 \sim \text{Poisson}(n\theta_1), \dots, Y_k \sim \text{Poisson}(n\theta_k)$ , that have been conditioned on their sum  $\sum_{i=1}^k Y_i$  being  $n$ .

*Part 4: The Normal approximation of the centred and scaled Poisson*

Recall from Model 6 that the expectation and variance of a RV  $Y_i \sim \text{Poisson}(n\theta_i)$  are  $E(Y_i) = V(Y_i) = n\theta_i$ . Let  $Z_i$  be  $E(Y_i)$ -centred and  $\sqrt{V(Y_i)}$ -scaled  $Y_i$  and

$$Z_i := \frac{Y_i - E(Y_i)}{\sqrt{V(Y_i)}} = \frac{Y_i - n\theta_i}{\sqrt{n\theta_i}} .$$

The condition that  $Y_1 + \dots + Y_k = n$  is equivalent to requiring that  $\sqrt{\theta_1}Z_1 + \dots + \sqrt{\theta_k}Z_k = 0$ , since:

$$\begin{aligned}\sum_{i=1}^k Y_i = n &\iff \sum_{i=1}^k Y_i - n = 0 \iff \sum_{i=1}^k Y_i - n \sum_{i=1}^k \theta_i = 0 \iff \sum_{i=1}^k Y_i - n\theta_i = 0 \\ &\iff \sum_{i=1}^k \frac{Y_i - n\theta_i}{\sqrt{n}} = 0 \iff \sum_{i=1}^k \sqrt{\theta_i} \frac{Y_i - n\theta_i}{\sqrt{n\theta_i}} = 0 \iff \sum_{i=1}^k \sqrt{\theta_i} Z_i = 0.\end{aligned}$$

Now consider the support of the R.V.  $Z := (Z_1, \dots, Z_k)$  conditioned on  $\sum_{i=1}^k \sqrt{\theta_i} Z_i = 0$ , i.e. the hyper-plane of  $(k-1)$ -dimensional vectors:

$$\mathbb{H} := \{(z_1, \dots, z_k) : \sqrt{\theta_1}z_1 + \dots + \sqrt{\theta_k}z_k = 0\}$$

Each  $Z_i \rightsquigarrow \text{Normal}(0, 1)$  by the central limit theorem. Therefore, for large values of  $n$ , each  $Z_i$  is approximately distributed as the  $\text{Normal}(0, 1)$  RV with PDF  $f(z_i; 0, 1) = (2\pi)^{-1/2} \exp(-z_i^2/2)$ . Since the  $Z_i$ s are independent except for the condition that they lie in  $\mathbb{H}$ , the point in a differential volume  $dz_2 \dots dz_k$  of  $\mathbb{H}$  occur with probability approximately proportional to:

$$\exp(-z_1^2/2) \times \dots \times \exp(-z_k^2/2) = \exp(-(z_1^2 + \dots + z_k^2)/2)$$

*Part 5: Chi-square distribution as the sum of squared Normals*

We are interested in the sum of the area of squares with side-lengths  $Z_1, \dots, Z_k$ . Let  $V$  be the desired sum of squares:

$$V := \sum_{i=1}^k Z_i^2 = \sum_{i=1}^k \frac{(Y_i - n\theta_i)^2}{n\theta_i}, \quad \text{such that } Z_i \rightsquigarrow \text{Normal}(0, 1), \quad \sum_{i=1}^k \sqrt{\theta_i} Z_i = 0.$$

The probability that  $V \leq v$  as  $n \rightarrow \infty$  is:

$$\frac{\int_{(z_1, \dots, z_k) \in \mathbb{H} \text{ and } \sum_{i=1}^k z_i^2 \leq v} \exp(-(z_1^2 + \dots + z_k^2)/2) dz_2 \dots dz_k}{\int_{(z_1, \dots, z_k) \in \mathbb{H}} \exp(-(z_1^2 + \dots + z_k^2)/2) dz_2 \dots dz_k}$$

Since the  $(k-1)$ -dimensional hyper-plane  $\mathbb{H}$  passes through the origin of  $\mathbb{R}^k$ , the domain of integration in the numerator above is the interior of a  $(k-1)$ -dimensional hyper-sphere of radius  $\sqrt{v}$  that is centred at the origin. Using a transformation of the above ratio of integrals into generalised polar co-ordinates with radius  $\chi$  and angles  $\alpha_1, \dots, \alpha_{k-2}$ , we get:

$$\frac{\int_{\chi^2 \leq v} \exp(-\chi^2/2) \chi^{k-2} g(\alpha_1, \dots, \alpha_{k-2}) d\chi d\alpha_1 \dots d\alpha_{k-2}}{\int \exp(-\chi^2/2) \chi^{k-2} g(\alpha_1, \dots, \alpha_{k-2}) d\chi d\alpha_1 \dots d\alpha_{k-2}},$$

for some function  $g$  of the angles [See Problem 15 in *Art of Computer Programming, Vol. II, Seminumerical Algorithms*, 3rd Ed., 1997, pp. 59]. The integration over the  $(k-2)$  angles results in the same factor that cancels between the numerator and the denominator. This yields the formula for the probability that  $V \leq v$  as  $n \rightarrow \infty$ :

$$\lim_{n \rightarrow \infty} P(V \leq v) = \frac{\int_0^{\sqrt{v}} \exp(-\chi^2/2) \chi^{k-2} d\chi}{\int_0^\infty \exp(-\chi^2/2) \chi^{k-2} d\chi}.$$

By substituting  $t = \chi^2/2$ , we can express the integrals in terms of the incomplete Gamma function defined as  $\gamma(a, x) := \int_0^a \exp(-t) t^{a-1} dt$  as follows:

$$P(V \leq v) = \gamma\left(\frac{k-1}{2}, \frac{v}{2}\right) / \Gamma\left(\frac{k-1}{2}\right).$$

This is the DF of the Chi-square distribution with  $k-1$  degrees of freedom.

**Model 29** (Chi-square( $k$ ) RV) Given a parameter  $k \in \mathbb{N}$  called degrees of freedom, we say that  $V$  is a Chi-square( $k$ ) RV if its PDF is:

$$f(v; k) := \frac{v^{(k/2)-1} e^{-v/2}}{2^{k/2} \Gamma(k/2)} \mathbf{1}_{\{v \in \mathbb{R}: v > 0\}}(v)$$

Also,  $E(V) = k$  and  $V(V) = 2k$ .

We can use the test statistic:

$$T := T(Y_1, \dots, Y_k) = \frac{(Y_1 - n\theta_1^*)^2}{n\theta_1^*} + \dots + \frac{(Y_k - n\theta_k^*)^2}{n\theta_k^*}$$

to test the null hypothesis  $H_0$  that may be formalised in three equivalent ways:

$$\begin{aligned} H_0 : X_1, X_2, \dots, X_n &\stackrel{IID}{\sim} \text{de Moivre}(\theta_1^*, \dots, \theta_k^*) \text{ RV} \\ \iff H_0 : Y := (Y_1, \dots, Y_k) &= \sum_{i=1}^n X_i \sim \text{Multinomial}(n, \theta_1^*, \dots, \theta_k^*) \text{ RV} \\ \iff H_0 : Y_1 &\stackrel{IND}{\sim} \text{Poisson}(n\theta_1) \text{ RV}, \dots, Y_k &\stackrel{IND}{\sim} \text{Poisson}(n\theta_k) \text{ RV given that } \sum_{i=1}^k Y_i = n \end{aligned}$$

We have seen that under  $H_0$ , the test statistic  $T \rightsquigarrow V \sim \text{Chi-square}(k-1)$ . Let  $t_{\text{obs}}$  be the observed value of the test statistic and let the upper alpha quantile be  $\chi_{k-1, \alpha}^2 := F^{[-1]}(1 - \alpha)$ , where  $F$  is the CDF of  $V \sim \text{Chi-square}(k-1)$ . Hence the test:

Reject  $H_0$  if  $T > \chi_{k-1, \alpha}^2$  is an asymptotically size  $\alpha$  test and the p-value =  $P(V > t_{\text{obs}})$ .

## 7.8 Parameter Estimation and Likelihood

Now that we have been introduced to point and set estimation of the population mean and the population proportion using the notion of convergence in distribution for sequences of RVs as well as concentration inequalities, we can begin to appreciate the art of estimation in a more general setting. Parameter estimation is the basic problem in statistical inference and machine learning. We will formalize the general estimation problem here.

As we have already seen, when estimating the population mean or population proportion, there are two basic types of estimators. In point estimation, as seen in Definition 119, we are interested in estimating a particular point of interest that is supposed to belong to a set of points. In (confidence) set estimation, we are interested in estimating a set with a particular form that has a specified probability of “trapping” the particular point of interest from a set of points. Here, a point should be interpreted as an element of a collection of elements from some space.

### 7.8.1 Point and Set Estimation – A General Likelihood Approach

**Point estimation** is any statistical methodology that provides one with a “**single best guess**” of some specific quantity of interest. Traditionally, we denote this **quantity of interest as  $\theta^*$**  and its **point estimate as  $\hat{\theta}$  or  $\hat{\theta}_n$** . The subscript  $n$  in the point estimate  $\hat{\theta}_n$  emphasizes that our estimate is based on  $n$  observations or data points from a given statistical experiment to estimate  $\theta^*$ . This quantity of interest, which is usually unknown, can be:

- a **parameter**  $\theta^*$  which is an element of the **parameter space**  $\Theta$ , i.e.  $\theta^* \in \Theta$  such that  $\theta^*$  specifies the “law” of the observations (realizations or samples) of the R $\vec{V}$  ( $X_1, \dots, X_n$ ) modeled by JPDF or JPMF  $f_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta^*)$ , or
- a **regression function**  $\theta^* \in \Theta$ , where  $\Theta$  is a class of regression functions in a regression experiment with model:  $Y = \theta^*(X) + \epsilon$ , such that  $e(\epsilon) = 0$  and  $\theta^*$  specifies the “law” of pairs of observations  $\{(X_i, Y_i)\}_{i=1}^n$ , for e.g., fitting parameters in noisy ODE or PDEs from observed data — one can always do a **prediction** in a regression experiment, i.e. when you want to estimate  $Y_i$  given  $X_i$ , or
- a **classifier**  $\theta^* \in \Theta$ , i.e. a regression experiment with discrete  $Y = \theta^*(X) + \epsilon$ , for e.g. training an scrub-nurse robot to assist a human surgeon, or
- an **integral**  $\theta^* := \int_A h(x) dx \in \Theta$ . If  $\theta^*$  is finite, then  $\Theta = \mathbb{R}$ , for e.g.  $\theta^*$  could be the volume of a high-dimensional irregular polyhedron, a traffic congestion measure on a network of roadways, the expected profit from a new brew of beer, or the probability of an extreme event such as the collapse of a dam in the Southern Alps in the next 150 years.

**Set estimation** is any statistical methodology that provides one with a “**best smallest set**”, such as an interval, rectangle, ellipse, etc. that contains  $\theta^*$  with a high probability  $1 - \alpha$ .

Recall that a statistic is a RV or R $\vec{V}$   $T(X)$  that maps every data point  $x$  in the data space  $\mathbb{X}$  with  $T(x) = t$  in its range  $\mathbb{T}$ , i.e.  $T(x) : \mathbb{X} \rightarrow \mathbb{T}$  (Definition 51). Next, we look at a specific class of estimators based on the likelihood of the data.

### 7.8.2 Likelihood

We take a look at **likelihood** — one of the most fundamental concepts in Statistics.

**Definition 136 (Likelihood Function)** Suppose  $(X_1, X_2, \dots, X_n)$  is a R $\vec{V}$  with JPDF or JPMF  $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta)$  specified by parameter  $\theta \in \Theta$ . Let the observed data be  $(x_1, x_2, \dots, x_n)$ . Then the **likelihood** function given by  $L_n(\theta)$  is merely the joint probability of the data, with the exception that we see it as a function of the parameter:

$$L_n(\theta) := L_n(x_1, x_2, \dots, x_n; \theta) = f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta) . \quad (7.53)$$

The **log-likelihood** function is defined by:

$$\ell_n(\theta) := \log(L_n(\theta)) \quad (7.54)$$

**Example 209 (Likelihood of the IID Bernoulli( $\theta^*$ ) experiment)** Consider our IID Bernoulli experiment:

$X_1, X_2, \dots, X_n \stackrel{IID}{\sim} \text{Bernoulli}(\theta^*)$ , with PDF  $f_{X_i}(x_i; \theta) = \theta^{x_i}(1-\theta)^{1-x_i}\mathbb{1}_{\{0,1\}}(x_i)$ , for  $i \in \{1, 2, \dots, n\}$  .

Let us understand the likelihood function for one observation first. There are two possibilities for the first observation.

If we only have one observation and it happens to be  $x_1 = 1$ , then our likelihood function is:

$$L_1(\theta) = L_1(x_1; \theta) = f_{X_1}(x_1; \theta) = \theta^1(1-\theta)^{1-1}\mathbb{1}_{\{0,1\}}(1) = \theta(1-\theta)^01 = \theta$$

If we only have one observation and it happens to be  $x_1 = 0$ , then our likelihood function is:

$$L_1(\theta) = L_1(x_1; \theta) = f_{X_1}(x_1; \theta) = \theta^0(1-\theta)^{1-0}\mathbb{1}_{\{0,1\}}(0) = 1(1-\theta)^11 = 1 - \theta$$

If we have  $n$  observations  $(x_1, x_2, \dots, x_n)$ , i.e. a vertex point of the unit hyper-cube  $\{0, 1\}^n$  (see top panel of Figure 7.5 when  $n \in \{1, 2, 3\}$ ), then our likelihood function (see bottom panel of Figure 7.5) is obtained by multiplying the densities due to our IID assumption:

$$\begin{aligned} L_n(\theta) &:= L_n(x_1, x_2, \dots, x_n; \theta) = f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta) \\ &= f_{X_1}(x_1; \theta)f_{X_2}(x_2; \theta)\cdots f_{X_n}(x_n; \theta) := \prod_{i=1}^n f_{X_i}(x_i; \theta) \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \end{aligned} \quad (7.55)$$

**Definition 137 (Maximum Likelihood Estimator (MLE))** Let the model for the data be

$$(X_1, \dots, X_n) \sim f_{X_1, X_2, \dots, X_n}(x_1, \dots, x_n; \theta^*) .$$

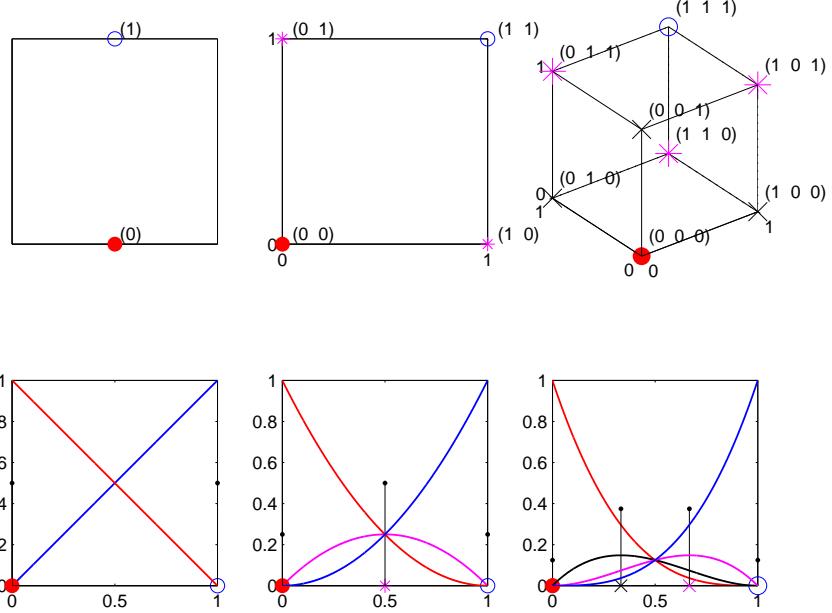
Then the maximum likelihood estimator (MLE)  $\hat{\Theta}_n$  of the fixed and possibly unknown parameter  $\theta^* \in \Theta$  is the value of  $\theta$  that maximizes the likelihood function:

$$\hat{\Theta}_n := \hat{\Theta}_n(X_1, X_2, \dots, X_n) := \underset{\theta \in \Theta}{\operatorname{argmax}} L_n(\theta) ,$$

Equivalently, MLE is the value of  $\theta$  that maximizes the log-likelihood function (since  $\log = \log_e = \ln$  is a monotone increasing function):

$$\hat{\Theta}_n := \underset{\theta \in \Theta}{\operatorname{argmax}} \ell_n(\theta) ,$$

Figure 7.5: Data Spaces  $\mathbb{X}_1 = \{0, 1\}$ ,  $\mathbb{X}_2 = \{0, 1\}^2$  and  $\mathbb{X}_3 = \{0, 1\}^3$  for one, two and three IID Bernoulli trials, respectively and the corresponding likelihood functions.



### Useful Properties of the Maximum Likelihood Estimator

1. The ML Estimator is *asymptotically consistent* (gives the “true”  $\theta^*$  as sample size  $n \rightarrow \infty$ ):

$$\hat{\Theta}_n \rightsquigarrow \text{Point Mass}(\theta^*)$$

2. The ML Estimator is asymptotically normal (has a normal distribution concentrating on  $\theta^*$  as  $n \rightarrow \infty$ ):

$$\hat{\Theta}_n \rightsquigarrow \text{Normal}(\theta^*, (\widehat{s}_{\Theta_n})^2)$$

or equivalently:

$$(\hat{\Theta}_n - \theta^*)/\widehat{s}_{\Theta_n} \rightsquigarrow \text{Normal}(0, 1)$$

where  $\widehat{s}_{\Theta_n}$  is the **estimated standard error**, i.e. the standard deviation of  $\hat{\Theta}_n$ , and it is given by the square-root of the inverse negative curvature of  $\ell_n(\theta)$  at  $\hat{\theta}_n$ :

$$\widehat{s}_{\Theta_n} = \sqrt{\left( \left[ -\frac{d^2 \ell_n(\theta)}{d\theta^2} \right]_{\theta=\hat{\theta}_n} \right)^{-1}}$$

3. Because of the previous two properties, the  $1 - \alpha$  confidence interval is:

$$\hat{\Theta}_n \pm z_{\alpha/2} \widehat{s}_{\Theta_n}$$

MLE is a general methodology for parameter estimation in an essentially arbitrary parameter space  $\Theta$  that is defining or indexing the laws in a parametric family of models, although we are only

seeing it in action when  $\Theta \subset \mathbb{R}$  for simplest parametric family of models involving IID product experiments here. When  $\Theta \subset \mathbb{R}^d$  with  $2 \leq d < \infty$  then MLE  $\hat{\Theta}_n \rightsquigarrow \text{Point Mass}(\theta^*)$ , where  $\theta^* = (\theta_1^*, \theta_2^*, \dots, \theta_d^*)^T$  is a column vector in  $\Theta \subset \mathbb{R}^d$  and  $\hat{\Theta}_n \rightsquigarrow \text{Normal}(\theta^*, \widehat{\Sigma(\text{se})}_n)$ , a multivariate Normal distribution with mean vector  $\theta^*$  and variance-covariance matrix of standard errors given by the *Hessian* (a  $d \times d$  matrix of mixed partial derivatives) of  $\ell_n(\theta_1, \theta_2, \dots, \theta_d)$ . The ideas in the case of dimension  $d = 1$  naturally generalize to an arbitrary, but finite, dimension  $d$ .

**Remark 138** In order to use MLE for parameter estimation we need to ensure that the following two conditions hold:

1. The *support* of the data, i.e. the set of possible values of  $(X_1, X_2, \dots, X_n)$  must not depend on  $\theta$  for every  $\theta \in \Theta$  — of course the probabilities do depend on  $\theta$  in an *identifiable* manner, i.e. for every  $\theta$  and  $\vartheta$  in  $\Theta$ , if  $\theta \neq \vartheta$  then  $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta) \neq f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \vartheta)$  at least for some  $(x_1, x_2, \dots, x_n) \in \mathbb{X}$ .
2. If the parameter space  $\Theta$  is bounded then  $\theta^*$  must not belong to the boundaries of  $\Theta$ .

### Maximum Likelihood Estimation Method in Six Easy Steps

**Background:** We have observed data:

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

which is modeled as a sample or realization from the random vector:

$$(X_1, X_2, \dots, X_n) \sim f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta^*), \quad \theta^* \in \Theta .$$

**Objective:** We want to obtain an estimator  $\hat{\Theta}_n$  that will give:

1. the point estimate  $\hat{\theta}_n$  of the “true” parameter  $\theta^*$  and
2. the  $(1 - \alpha)$  confidence interval for  $\theta^*$ .

#### Steps of MLE:

- Step 1: Find the expression for the log likelihood function:

$$\ell_n(\theta) = \log(L_n(\theta)) = \log(f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta)) .$$

Note that if the model assumes that  $(X_1, X_2, \dots, X_n)$  is jointly independent, i.e. we have an independent and identically distributed (IID) experiment, then  $\ell_n(\theta)$  simplifies further as follows:

$$\ell_n(\theta) = \log(L_n(\theta)) = \log(f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta)) = \log\left(\prod_{i=1}^n f_{X_i}(x_i; \theta)\right) .$$

- Step 2: Obtain the derivative of  $\ell_n(\theta)$  with respect to  $\theta$ :

$$\frac{d}{d\theta}(\ell_n(\theta)) .$$

- Step 3: Set the derivative equal to zero, solve for  $\theta$  and let  $\hat{\theta}_n$  equal to this solution.

- Step 4: Check if this solution is indeed a maximum of  $\ell_n(\theta)$  by checking if:

$$\frac{d^2}{d\theta^2} (\ell_n(\theta)) < 0 .$$

- Step 5: If  $\frac{d^2}{d\theta^2} (\ell_n(\theta)) < 0$  then you have found the maximum likelihood estimate  $\hat{\theta}_n$ .
- Step 6: If you also want the  $(1 - \alpha)$  confidence interval then get it from

$$\hat{\theta}_n \pm z_{\alpha/2} \widehat{s\epsilon}_n , \text{ where } \widehat{s\epsilon}_n = \sqrt{\left( \left[ -\frac{d^2 \ell_n(\theta)}{d\theta^2} \right]_{\theta=\hat{\theta}_n} \right)^{-1}} .$$

Let us apply this method in some examples.

**Example 210 (Maximum likelihood estimation for IID Exponential( $\lambda^*$ ) trials)** Find (or derive) the maximum likelihood estimate  $\hat{\lambda}_n$  and the  $(1 - \alpha)$  confidence interval of the fixed and possibly unknown parameter  $\lambda^*$  for the IID experiment:

$$X_1, \dots, X_n \stackrel{IID}{\sim} \text{Exponential}(\lambda^*), \quad \lambda^* \in \mathbf{A} = (0, \infty) .$$

Note that  $\mathbf{A}$  is the parameter space.

We first obtain the log-likelihood function  $\ell_n(\theta)$  given data  $(x_1, x_2, \dots, x_n)$ .

$$\begin{aligned} \ell_n(\lambda) &:= \log(L(x_1, x_2, \dots, x_n; \lambda)) = \log \left( \prod_{i=1}^n f_{X_i}(x_i; \lambda) \right) = \log \left( \prod_{i=1}^n \lambda e^{-\lambda x_i} \right) \\ &= \log \left( \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \cdots \lambda e^{-\lambda x_n} \right) = \log \left( \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \right) = \log(\lambda^n) + \log \left( e^{-\lambda \sum_{i=1}^n x_i} \right) \\ &= \boxed{\log(\lambda^n) - \lambda \sum_{i=1}^n x_i} \end{aligned}$$

Now, let us take the derivative with respect to  $\lambda$ ,

$$\begin{aligned} \frac{d}{d\lambda} (\ell_n(\lambda)) &:= \frac{d}{d\lambda} \left( \log(\lambda^n) - \lambda \sum_{i=1}^n x_i \right) = \frac{d}{d\lambda} (\log(\lambda^n)) - \frac{d}{d\lambda} \left( \lambda \sum_{i=1}^n x_i \right) = \frac{1}{\lambda^n} \frac{d}{d\lambda} (\lambda^n) - \sum_{i=1}^n x_i \\ &= \frac{1}{\lambda^n} n \lambda^{n-1} - \sum_{i=1}^n x_i = \boxed{\frac{n}{\lambda} - \sum_{i=1}^n x_i} . \end{aligned}$$

Next, we set the derivative to 0, solve for  $\lambda$ , and let the solution equal to the ML estimate  $\hat{\lambda}_n$ .

$$0 = \frac{d}{d\lambda} (\ell_n(\lambda)) \iff 0 = \frac{n}{\lambda} - \sum_{i=1}^n x_i \iff \sum_{i=1}^n x_i = \frac{n}{\lambda} \iff \lambda = \frac{n}{\sum_{i=1}^n x_i} \quad \text{and let } \boxed{\hat{\lambda}_n = \frac{1}{\bar{x}_n}} .$$

Next, we find the second derivative and check if it is negative.

$$\frac{d^2}{d\lambda^2} (\ell_n(\lambda)) = \frac{d}{d\lambda} \left( \frac{d}{d\lambda} (\ell_n(\lambda)) \right) = \frac{d}{d\lambda} \left( \frac{n}{\lambda} - \sum_{i=1}^n x_i \right) = \boxed{-n\lambda^{-2}} .$$

Since  $\lambda > 0$  and  $n \in \mathbb{N}$ ,  $-n\lambda^{-2} = -n/\lambda^2 < 0$ , so we have found the maximum likelihood estimate:

$$\widehat{\lambda}_n = \frac{1}{\bar{x}_n} .$$

Now, let us find the estimated standard error:

$$\begin{aligned}\widehat{s\epsilon}_n &= \sqrt{\left(\left[-\frac{d^2\ell_n(\lambda)}{d\lambda^2}\right]_{\lambda=\widehat{\lambda}_n}\right)^{-1}} = \sqrt{\left(\left[-\left(-\frac{n}{\lambda^2}\right)\right]_{\lambda=\widehat{\lambda}_n}\right)^{-1}} = \sqrt{\left(\frac{n}{\widehat{\lambda}_n^2}\right)^{-1}} = \sqrt{\frac{\widehat{\lambda}_n^2}{n}} = \frac{\widehat{\lambda}_n}{\sqrt{n}} \\ &= \frac{1}{\bar{x}_n\sqrt{n}} .\end{aligned}$$

And finally, the  $(1 - \alpha)$  confidence interval is

$$\widehat{\lambda}_n \pm z_{\alpha/2} \widehat{s\epsilon}_n = \frac{1}{\bar{x}_n} \pm z_{\alpha/2} \frac{1}{\bar{x}_n\sqrt{n}} .$$

Since we have worked “hard” to get the maximum likelihood estimate for a general IID model  $X_1, X_2, \dots, X_n \stackrel{IID}{\sim} \text{Exponential}(\lambda^*)$ . Let us kill two birds with the same stone by applying it to two datasets:

1. Orbiter waiting times and
2. Time between measurable earthquakes in New Zealand over a few months.

Therefore, the ML estimate  $\widehat{\lambda}_n$  of the unknown rate parameter  $\lambda^* \in \mathbb{A}$  on the basis of  $n$  IID observations  $x_1, x_2, \dots, x_n \stackrel{IID}{\sim} \text{Exponential}(\lambda^*)$  is  $1/\bar{x}_n$  and the ML estimator  $\widehat{\Lambda}_n = 1/\bar{X}_n$ .

**Example 211 (Orbiter Waiting Times)** Let us apply this ML estimator of the rate parameter for the supposedly exponentially distributed waiting times at the on-campus Orbiter bus-stop.

Joshua Fenemore and Yiran Wang collected data on waiting times between buses at an Orbiter bus-stop close to campus. They collected a sample of size  $n = 132$  with sample mean  $\bar{x}_{132} = 9.0758$ .

```
% Joshu Fenemore's Data from 2007 on Waiting Times at Orbiter Bust Stop by Balgay Street
%The raw data -- the waiting times to nearest minute between Orbiter buses
>> orbiterTimes=[8 3 7 18 18 3 7 9 9 25 0 0 25 6 10 0 10 8 16 9 1 5 16 6 4 1 3 21 0 28 3 8 ...
6 6 11 8 10 15 0 8 7 11 10 9 12 13 8 10 11 8 7 11 5 9 11 14 13 5 8 9 12 10 13 6 11 13 ...
0 0 11 1 9 5 14 16 2 10 21 1 14 2 10 24 6 1 14 14 0 14 4 11 15 0 10 2 13 2 22 ...
10 5 6 13 1 13 10 11 4 7 9 12 8 16 15 14 5 10 12 9 8 0 5 13 13 6 8 4 13 15 7 11 6 23 1];
>> mean(orbiterTimes)
ans =
9.0758
```

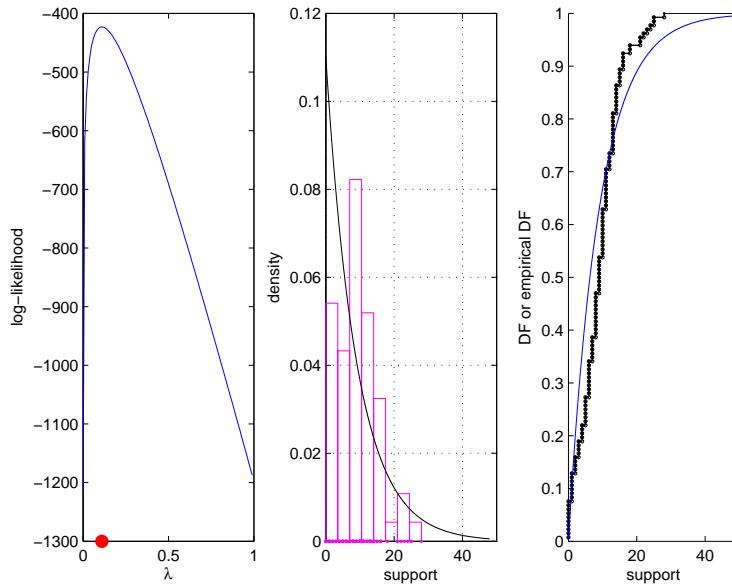
From our work in Example 210 we can now easily obtain the maximum likelihood estimate of  $\lambda^*$  and the 95% confidence interval for it, under the assumption that the waiting times  $X_1, \dots, X_{132}$  are IID  $\text{Exponential}(\lambda^*)$  RVs as follows:

$$\widehat{\lambda}_{132} = 1/\bar{x}_{132} = 1/9.0758 = 0.1102 \quad (0.1102 \pm 1.96 \times 0.1102/\sqrt{132}) = (0.0914, 0.1290) ,$$

and thus the estimated mean waiting time is

$$1/\widehat{\lambda}_{132} = 9.0763 \text{ minutes.}$$

Figure 7.6: Plot of  $\log(L(\lambda))$  as a function of the parameter  $\lambda$  and the MLE  $\hat{\lambda}_{132}$  of 0.1102 for Fenemore-Wang Orbiter Waiting Times Experiment from STAT 218 S2 2007. The density or PDF and the DF at the MLE of 0.1102 are compared with a histogram and the empirical DF.



The estimated mean waiting time for a bus to arrive is well within the 10 minutes promised by the Orbiter bus company. This data and its maximum likelihood analysis is presented visually in Figure 7.6.

The following script was used to generate the Figure 7.6: Notice how the exponential PDF  $f(x; \hat{\lambda}_{132} = 0.1102)$  and the DF  $F(x; \hat{\lambda}_{132} = 0.1102)$  based on the MLE fits with the histogram and the empirical DF, respectively.

**Example 212 (Waiting Times between Earth Quakes in NZ)** Once again from our work in Example 210 we can now easily obtain the maximum likelihood estimate of  $\lambda^*$  and the 95% confidence interval for it, under the assumption that the waiting times (in days) between the 6128 measurable earth-quakes in NZ from 18-Jan-2008 02:23:44 to 18-Aug-2008 19:29:29 are IID Exponential( $\lambda^*$ ) RVs as follows:

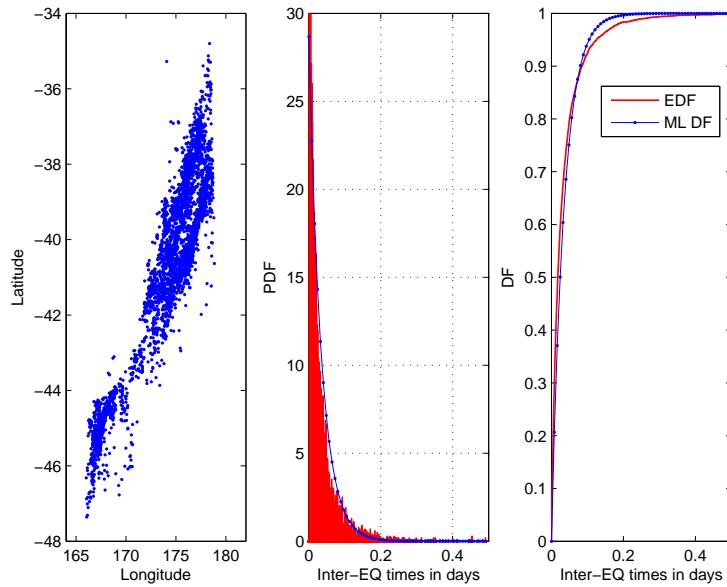
$$\hat{\lambda}_{6128} = 1/\bar{x}_{6128} = 1/0.0349 = 28.6694 \quad (28.6694 \pm 1.96 \times 28.6694/\sqrt{6128}) = (27.95, 29.39) ,$$

and thus the estimated mean time in days and minutes between earth quakes (somewhere in NZ over the first 8 months in 2008), as processed in Labwork 213, is

$$1/\hat{\lambda}_{6128} = \bar{x}_{6128} = 0.0349 \text{ days} \quad = \quad 0.0349 * 24 * 60 = 50.2560 \text{ minutes} .$$

This data and its maximum likelihood analysis is presented visually in Figure 7.7. The PDF and DF corresponding to the  $\hat{\lambda}_{6128}$  (blue curves in Figure 7.7) are the best fitting PDF and DF from the parametric family of PDFs in  $\{\lambda e^{-\lambda x} : \lambda \in (0, \infty)\}$  and DFs in  $\{1 - e^{-\lambda x} : \lambda \in (0, \infty)\}$  to the density histogram and the empirical distribution function given by the data, respectively. Clearly, there is room for improving beyond the model of IID Exponential( $\lambda$ ) RVs, but the fit with just one real-valued parameter is not too bad either. Finally, with the best fitting PDF  $28.6694e^{-28.6694x}$

Figure 7.7: Comparing the Exponential( $\hat{\lambda}_{6128} = 28.6694$ ) PDF and DF with a histogram and empirical DF of the times (in units of days) between earth quakes in NZ. The epicenters of 6128 earth quakes are shown in left panel.



we can get probabilities of events and answer questions like: “what is the probability that there will be three earth quakes somewhere in NZ within the next hour?”, etc.

**Labwork 213 (Inter Earth Quake Time Processing)** To process the data to get the times between earth quakes, we can compute as in the following script:

---

```
%NZSEarthQuakesExponentialMLE.m
%% The columns in earthquakes.csv file have the following headings
%% CUSP_ID, LAT, LONG, NZMGE, NZMGN, ORI_YEAR, ORI_MONTH, ORI_DAY, ORI_HOUR, ORI_MINUTE, ORI_SECOND, MAG, DEPTH
EQ=dlmread('earthquakes.csv',','); % load the data in the matrix EQ
size(EQ) % report thr size of the matrix EQ

% Read help datenum -- converts time stamps into numbers in units of days
MaxD=max(datenum(EQ(:,6:11)));% maximum datenum
MinD=min(datenum(EQ(:,6:11)));% minimum datenum
disp('Earth Quakes in NZ between')
disp(strcat(datestr(MinD), ' and ', datestr(MaxD)))% print MaxD and MinD as a date string

% get an array of sorted time stamps of EQ events starting at 0
Times=sort(datenum(EQ(:,6:11))-MinD);
TimeDiff=diff(Times); % inter-EQ times = times between successtive EQ events
clf % clear any current figures
%figure
%plot(TimeDiff) % plot the inter-EQ times
subplot(1,3,1)
plot(EQ(:,3),EQ(:,2),'.')
axis([164 182 -48 -34])
xlabel('Longitude'); ylabel('Latitude');

subplot(1,3,2) % construct a histogram estimate of inter-EQ times
histogram(TimeDiff',1,[min(TimeDiff),max(TimeDiff)],'r',2);
SampleMean=mean(TimeDiff) % find the sample mean
% the MLE of LambdaStar if inter-EQ times are IID Exponential(LambdaStar)
MLELambdaHat=1/SampleMean
```

```

hold on;
TIMEs=linspace(0,max(TimeDiff),100);
plot(TIMEs,MLELambdaHat*exp(-MLELambdaHat*TIMEs),'b.-')
axis([0 0.5 0 30])
xlabel('Inter-EQ times in days'); ylabel('PDF');

subplot(1,3,3)
[x y]=ECDF(TimeDiff,0,0,0); % get the coordinates for empirical DF
stairs(x,y,'r','linewidth',1) % draw the empirical DF
hold on; plot(TIMEs,ExponentialCdf(TIMEs,MLELambdaHat),'b.-');% plot the DF at MLE
axis([0 0.5 0 1])
xlabel('Inter-EQ times in days'); ylabel('DF'); legend('EDF', 'ML DF')

```

---

We first load the data in the text file `earthquakes.csv` into a matrix `EQ`. Using the `datenum` function in MATLAB we transform the time stamps into a number starting at zero. These transformed time stamps are in units of days. Then we find the times between consecutive events and estimate a histogram. We finally compute the ML estimate of  $\lambda^*$  and super-impose the PDF of the Exponential( $\hat{\lambda}_{6128} = 28.6694$ ) upon the histogram.

```

>> NZSEarthQuakesExponentialMLE
ans =          6128          13

Earth Quakes in NZ between
18-Jan-2008 02:23:44 and 18-Aug-2008 19:29:29

SampleMean =    0.0349
MLELambdaHat =  28.6694

```

Thus, the average time between earth quakes is  $0.0349 * 24 * 60 = 50.2560$  minutes.

**Example 214 (ML Estimation for the IID Bernoulli( $\theta^*$ ) experiment)** Let us do maximum likelihood estimation for the coin-tossing experiment of Example 195 with likelihood derived in Example 209 to obtain the maximum likelihood estimate  $\hat{\theta}_n$  of the unknown parameter  $\theta^* \in \Theta = [0, 1]$  and the  $(1 - \alpha)$  confidence interval for it.

From Equation (7.55) the log likelihood function is

$$\ell_n(\theta) = \log(L_n(\theta)) = \log\left(\theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}\right) = \left[\left(\sum_{i=1}^n x_i\right) \log(\theta) + \left(n - \sum_{i=1}^n x_i\right) \log(1-\theta)\right],$$

Next, we take the derivative with respect to the parameter  $\theta$ :

$$\frac{d}{d\theta} (\ell_n(\theta)) = \frac{d}{d\theta} \left( \left(\sum_{i=1}^n x_i\right) \log(\theta) \right) + \frac{d}{d\theta} \left( \left(n - \sum_{i=1}^n x_i\right) \log(1-\theta) \right) = \boxed{\frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1-\theta}}.$$

Now, set  $\frac{d}{d\theta} \log(L_n(\theta)) = 0$ , solve for  $\theta$  and set the solution equal to  $\hat{\theta}_n$ :

$$\begin{aligned} \frac{d}{d\theta} (\ell_n(\theta)) = 0 &\iff \frac{\sum_{i=1}^n x_i}{\theta} = \frac{n - \sum_{i=1}^n x_i}{1-\theta} \iff \frac{1-\theta}{\theta} = \frac{n - \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i} \\ &\iff \frac{1}{\theta} - 1 = \frac{n}{\sum_{i=1}^n x_i} - 1 \quad \text{let } \boxed{\hat{\theta}_n = \frac{\sum_{i=1}^n x_i}{n}} \end{aligned}$$

Next, we find the second derivative and check if it is negative.

$$\frac{d^2}{d\theta^2}(\ell_n(\theta)) = \frac{d}{d\theta} \left( \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1-\theta} \right) = \boxed{-\frac{\sum_{i=1}^n x_i}{\theta^2} - \frac{n - \sum_{i=1}^n x_i}{(1-\theta)^2}}$$

Since each term in the numerator and the denominator of the two fractions in the above box are non-negative,  $\frac{d^2}{d\theta^2}(\ell_n(\theta)) < 0$  and therefore we have found the maximum likelihood estimate

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}_n$$

We already knew this to be a point estimate for  $E(X_i) = \theta^*$  from LLN and CLT. But now we know that MLE also agrees. Now, let us find the estimated standard error:

$$\begin{aligned} \widehat{\text{se}}_n &= \sqrt{\left( \left[ -\frac{d^2 \ell_n(\theta)}{d\theta^2} \right]_{\theta=\hat{\theta}_n} \right)^{-1}} = \sqrt{\left( \left[ -\left( -\frac{\sum_{i=1}^n x_i}{\theta^2} - \frac{n - \sum_{i=1}^n x_i}{(1-\theta)^2} \right) \right]_{\theta=\hat{\theta}_n} \right)^{-1}} \\ &= \sqrt{\left( \frac{\sum_{i=1}^n x_i}{\hat{\theta}_n^2} + \frac{n - \sum_{i=1}^n x_i}{(1-\hat{\theta}_n)^2} \right)^{-1}} = \sqrt{\left( \frac{n\bar{x}_n}{\bar{x}_n^2} + \frac{n - n\bar{x}_n}{(1-\bar{x}_n)^2} \right)^{-1}} = \sqrt{\left( \frac{n}{\bar{x}_n} + \frac{n}{(1-\bar{x}_n)} \right)^{-1}} \\ &= \sqrt{\left( \frac{n(1-\bar{x}_n) + n\bar{x}_n}{\bar{x}_n(1-\bar{x}_n)} \right)^{-1}} = \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n((1-\bar{x}_n) + \bar{x}_n)}} = \boxed{\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}}. \end{aligned}$$

And finally, the  $(1 - \alpha)$  confidence interval is

$$\hat{\theta}_n \pm z_{\alpha/2} \widehat{\text{se}}_n = \boxed{\bar{x}_n \pm z_{\alpha/2} \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}}.$$

For the coin tossing experiment that was performed ( $n = 10$  times) in Example 195, the maximum likelihood estimate of  $\theta^*$  and the 95% confidence interval for it, under the model that the tosses are IID Bernoulli( $\theta^*$ ) RVs, are as follows:

$$\hat{\theta}_{10} = \bar{x}_{10} = \frac{4}{10} = 0.40 \quad \text{and} \quad \left( 0.4 \pm 1.96 \times \sqrt{\frac{0.4 \times 0.6}{10}} \right) = (0.0964, 0.7036).$$

See Figures 7.8 and 7.9 to completely understand parameter estimation for IID Bernoulli experiments.

Figure 7.8: Plots of the log likelihood  $\ell_n(\theta) = \log(L(1, 0, 0, 0, 1, 1, 0, 0, 1, 0; \theta))$  as a function of the parameter  $\theta$  over the parameter space  $\Theta = [0, 1]$  and the MLE  $\hat{\theta}_{10}$  of 0.4 for the coin-tossing experiment shown in standard scale (left panel) and log scale for  $x$ -axis (right panel).

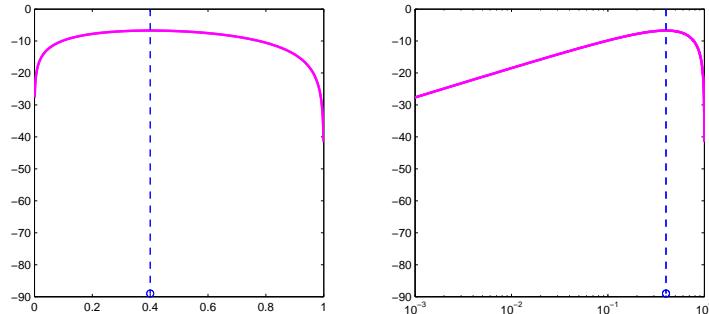
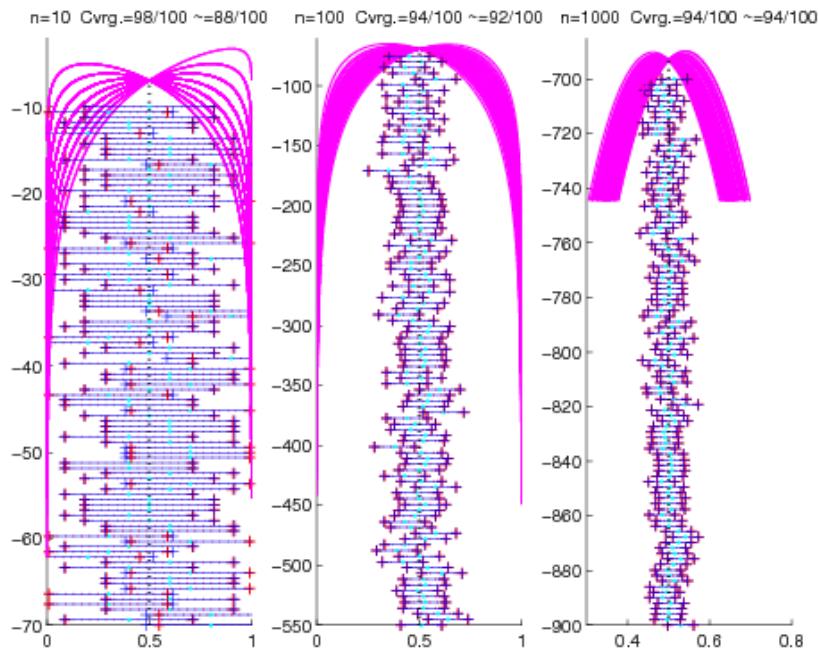


Figure 7.9: 100 realizations of 95% confidence intervals based on samples of size  $n = 10, 100$  and  $1000$  simulated from IID Bernoulli( $\theta^* = 0.5$ ) RVs. The MLE  $\hat{\theta}_n$  (cyan dot) and the log-likelihood function (magenta curve) for each of the 100 replications of the experiment for each sample size  $n$  are depicted. The approximate normal-based 95% confidence intervals with blue boundaries are based on the exact  $\text{se}_n = \sqrt{\theta^*(1 - \theta^*)/n} = \sqrt{1/4}$ , while those with red boundaries are based on the estimated  $\widehat{\text{se}}_n = \sqrt{\hat{\theta}_n(1 - \hat{\theta}_n)/n} = \sqrt{\bar{x}_n(1 - \bar{x}_n)/n}$ . The fraction of times the true parameter  $\theta^* = 0.5$  was contained by the exact and approximate confidence interval (known as *empirical coverage*) over the 100 replications of the simulation experiment for each of the three sample sizes are given by the numbers after `Cvrg.=` and  $\sim=$ , above each sub-plot, respectively.



**Exercise 7.14 (Likelihoods of tiny Bernoulli trials)** Find and plot the likelihood function of the following observations  $(x_1, x_2, \dots, x_n)$  from the following IID sequence of Bernoulli( $\theta$ ) RVs:

1.  $(x_1) = (1)$
2.  $(x_1) = (0)$
3.  $(x_1, x_2) = (0, 0)$
4.  $(x_1, x_2) = (1, 1)$
5.  $(x_1, x_2) = (1, 0)$
6.  $(x_1, x_2) = (0, 1)$
7.  $(x_1, x_2, x_3) = (1, 1, 0)$
8.  $(x_1, x_2, x_3) = (0, 0, 1)$

[Hint: your x-axis is  $\theta$  with values in  $[0, 1]$ , the parameter space, and y-axis is  $L_n(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i; \theta)$ , where  $f_{X_i}(x_i; \theta)$  is the PMF of Bernoulli( $\theta$ ) RV  $X_i$ ]

**Exercise 7.15 (MLE Exercises)** Assume that an independent and identically distributed sample,  $X_1, X_2, \dots, X_n$  is drawn from the distribution of  $X$  with PDF  $f(x; \theta^*)$  for a fixed and unknown parameter  $\theta^*$  and derive the maximum likelihood estimate of  $\theta^*$  (you only need to do Steps 1–5 from **Steps of MLE** in Lecture Notes on pages 126–127). Consider the following PDFs:

1. The parameter  $\theta$  is a real number in  $(0, \infty)$  and the PDF is given by

$$f(x; \theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} . \end{cases}$$

2. The parameter  $\theta$  is a real number in  $(0, \infty)$  and the PDF is given by

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} x^{(1-\theta)/\theta} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} . \end{cases}$$

3. The parameter  $\theta$  is a real number in  $(0, \infty)$  and the PDF is given by

$$f(x; \theta) = \begin{cases} \frac{1}{2\theta^3} x^2 e^{-x/\theta} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise} . \end{cases}$$

4. The parameter  $\theta$  is a real number in  $(0, \infty)$  and the PDF is given by

$$f(x; \theta) = \begin{cases} \frac{x}{\theta^2} e^{-\frac{1}{2}(x/\theta)^2} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise} . \end{cases}$$

### 7.8.3 Moment Estimator (MME)

See notes from class.

## 7.9 Practical Excursion in One-dimensional Optimisation

Numerically maximising a log-likelihood function of one parameter is a useful technique. This can be used for models with no analytically known MLE. A fairly large field of maths, called optimisation, exists for this sole purpose. Conventionally, in optimisation, one is interested in minimisation. Therefore, the basic algorithms are cast in the “find the minimiser and the minimum” of a target function  $f : \mathbb{R} \rightarrow \mathbb{R}$ . Since we are interested in maximising our target, which is the likelihood or log-likelihood function, say  $\log(L(x_1, \dots, x_n; \theta)) : \Theta \rightarrow \mathbb{R}$ , we will simply apply the standard optimisation algorithms directly to  $-\log(L(x_1, \dots, x_n; \theta)) : \Theta \rightarrow \mathbb{R}$ .

The algorithm implemented in `fminbnd` is based on the golden section search and an inverse parabolic interpolation, and attempts to find the minimum of a function of one variable within a given fixed interval. Briefly, the golden section search proceeds by successively **bracketing** the minimum of the target function within an acceptably small interval inside the given starting interval [see Section 8.2 of Forsythe, G. E., M. A. Malcolm, and C. B. Moler, 1977, *Computer Methods for Mathematical Computations*, Prentice-Hall]. MATLAB’s `fminbnd` also relies on Brent’s inverse parabolic interpolation [see Chapter 5 of Brent, Richard. P., 1973, *Algorithms for Minimization without Derivatives*, Prentice-Hall, Englewood Cliffs, New Jersey]. Briefly, additional smoothness conditions are assumed for the target function to aid in a faster bracketing strategy through polynomial interpolations of past function evaluations. MATLAB’s `fminbnd` has several limitations, including:

- The likelihood function must be continuous.
- Only local MLE solutions, i.e. those inside the starting interval, are given.
- One needs to know or carefully guess the starting interval that contains the MLE.
- MATLAB’s `fminbnd` exhibits slow convergence when the solution is on a boundary of the starting interval.

**Labwork 215 (Coin-tossing experiment)** The following script was used to study the coin-tossing experiment in MATLAB. The plot of the log-likelihood function and the numerical optimisation of MLE are carried out using MATLAB’s built-in function `fminbnd` (See Figure 7.8).

---

```
BernoulliMLE.m
%
% To simulate n coin tosses, set theta=probability of heads and n
% Then draw n IID samples from Bernoulli(theta) RV
% theta=0.5; n=20; x=floor(rand(1,n) + theta);
% enter data from a real coin tossing experiment
x=[1 0 0 0 1 1 0 0 1 0]; n=length(x);
t = sum(x); % statistic t is the sum of the x_i values
% display the outcomes and their sum
display(x)
display(t)

% Analytically MLE is t/n
MLE=t/n
%
% l is the log-likelihood of data x as a function of parameter theta
l=@(theta)log(theta ^ t * (1-theta)^(n-t));
ThetaS=[0:0.001:1]; % sample some values for theta

% plot the log-likelihood function and MLE in two scales
subplot(1,2,1);
plot(ThetaS,arrayfun(l,ThetaS),'m','LineWidth',2);
hold on; stem([MLE],[-89],'b--'); % plot MLE as a stem plot
subplot(1,2,2);
```

```

semilogx(ThetaS,arrayfun(l,ThetaS),'m','LineWidth',2);
hold on; stem([MLE],[-89],'b--'); % plot MLE as a stem plot

% Now we will find the MLE by finding the minimiser or argmin of -l
% negative log-likelihood function of parameter theta
negl=@(theta)-(log(theta.^t * (1-theta).^(n-t)));
% read help fminbnd
% you need to supply the function to be minimised and its search interval
% NumericalMLE = fminbnd(negl,0,1)
% to see the iteration in the numerical minimisation
NumericalMLE = fminbnd(negl,0,1,optimset('Display','iter'))

```

```

>> BernoulliMLE
x = 1 0 0 0 1 1 0 0 1 0
t = 4
MLE = 0.4000
Func-count x f(x) Procedure
1 0.381966 6.73697 initial
2 0.618034 7.69939 golden
3 0.236068 7.3902 golden
4 0.408979 6.73179 parabolic
5 0.399339 6.73013 parabolic
6 0.400045 6.73012 parabolic
7 0.400001 6.73012 parabolic
8 0.399968 6.73012 parabolic
Optimisation terminated:
the current x satisfies the termination criteria using OPTIONS.TolX of 1.000000e-04
NumericalMLE = 0.4000

```

**Labwork 216 (Numerical MLE of  $\lambda$  from n IID Exponential( $\lambda$ ) RVs)** Joshua Fenemore and Yiran Wang collected data on waiting times between buses at an Orbiter bus-stop close to campus and modelled the waiting times as IID Exponential( $\lambda^*$ ) RVs (<http://www.math.canterbury.ac.nz/~r.sainudiin/courses/STAT218/projects/Stat218StudentProjects2007.pdf>). We can use their data `sampleTimes` to find the MLE of  $\lambda^*$  under the assumption that the waiting times  $X_1, \dots, X_{132}$  are IID Exponential( $\lambda^*$ ). We find the ML estimate  $\hat{\lambda}_{132} = 0.1102$  and thus the estimated mean waiting time is  $1/\hat{\lambda}_{132} = 9.0763$  minutes. The estimated mean waiting time for a bus to arrive is well within the 10 minutes promised by the Orbiter bus company. The following script was used to generate the Figure 7.6:

---

```

-- ExponentialMLEOrbiter.m --
% Joshu Fenemore's Data from 2007 on Waiting Times at Orbiter Bust Stop
%The raw data -- the waiting times i minutes for each direction
antiTimes=[8 3 7 18 18 3 7 9 9 25 0 0 25 6 10 0 10 8 16 9 1 5 16 6 4 1 3 21 0 28 3 8 ...
6 6 11 8 10 15 0 8 7 11 10 9 12 13 8 10 11 8 7 11 5 9 11 14 13 5 8 9 12 10 13 6 11 13];
clockTimes=[0 0 11 1 9 5 14 16 2 10 21 1 14 2 10 24 6 1 14 14 0 14 4 11 15 0 10 2 13 2 22 ...
10 5 6 13 1 13 10 11 4 7 9 12 8 16 15 14 5 10 12 9 8 0 5 13 13 6 8 4 13 15 7 11 6 23 1];
sampleTimes=[antiTimes clockTimes];% pool all times into 1 array
% L = Log Likelihood of data x as a function of parameter lambda
L=@(lambda)sum(log(lambda*exp(-lambda * sampleTimes)));
LAMBDA=0.0001:0.01:1; % sample some values for lambda
clf;
subplot(1,3,1);
plot(LAMBDA,arrayfun(L,LAMBDA)); % plot the Log Likelihood function
% Now we will find the Maximum Likelihood Estimator by finding the minimizer of -L
MLE = fminbnd(@(lambda)-sum(log(lambda*exp(-lambda * sampleTimes))),0.0001,1)
MeanEstimate=1/MLE
hold on; % plot the MLE
plot([MLE],[-1300],'r.', 'MarkerSize',25); ylabel('log-likelihood'); xlabel('\lambda');
subplot(1,3,2); % plot a histogram estimate

```

```

histogram(sampleTimes,1,[min(sampleTimes),max(sampleTimes)],'m',2);
hold on; TIMES=[0.00001:0.01:max(sampleTimes)+20]; % points on support
plot(TIMES,MLE*exp(-MLE * TIMES ),'k-') % plot PDF at MLE to compare with histogram
% compare the empirical DF to the best fitted DF
subplot(1,3,3)
ECDF(sampleTimes,5,0.0,20); hold on
plot(TIMES,ExponentialCdf(TIMES,MLE),'b-')
ylabel('DF or empirical DF'); xlabel('support');

```

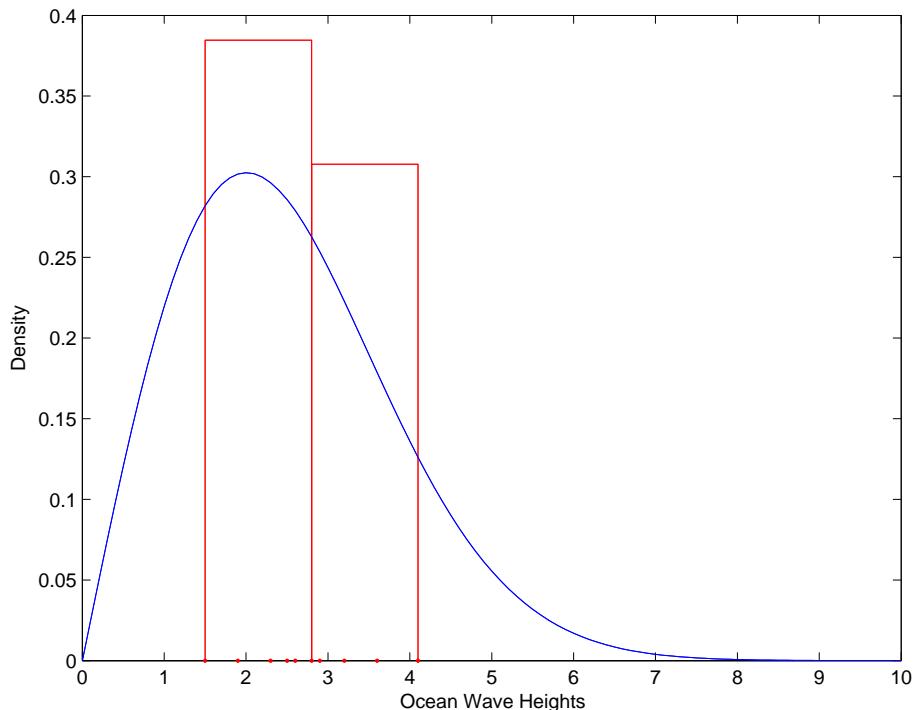
The script output the following in addition to the plot:

```

>> ExponentialMLEOrbiter
MLE =      0.1102
MeanEstimate =    9.0763

```

Figure 7.10: The ML fitted Rayleigh( $\hat{\alpha}_{10} = 2$ ) PDF and a histogram of the ocean wave heights.



**Example 217 (6.7, p. 275 of Ang & Tang)** The distribution of ocean wave heights,  $H$ , may be modeled with the Rayleigh( $\alpha$ ) RV with parameter  $\alpha$  and probability density function,

$$f(h; \alpha) = \frac{h}{\alpha^2} \exp\left(-\frac{1}{2}(h/\alpha)^2\right), \quad h \in \mathbb{H} := [0, \infty).$$

The parameter space for  $\alpha$  is  $\mathbb{A} = (0, \infty)$ . Suppose that the following measurements  $h_1, h_2, \dots, h_{10}$  of wave heights in meters were observed to be

$$1.50, 2.80, 2.50, 3.20, 1.90, 4.10, 3.60, 2.60, 2.90, 2.30,$$

respectively. Under the assumption that the 10 samples are IID realisations from a Rayleigh( $\alpha^*$ ) RV with a fixed and unknown parameter  $\alpha^*$ , find the ML estimate  $\hat{\alpha}_{10}$  of  $\alpha^*$ .

We first obtain the log-likelihood function of  $\alpha$  for the data  $h_1, h_2, \dots, h_n \stackrel{IID}{\sim} \text{Rayleigh}(\alpha)$ .

$$\begin{aligned}\ell(\alpha) &:= \log(L(h_1, h_2, \dots, h_n; \alpha)) = \log \left( \prod_{i=1}^n f(h_i; \alpha) \right) = \sum_{i=1}^n \log(f(h_i; \alpha)) \\ &= \sum_{i=1}^n \log \left( \frac{h_i}{\alpha^2} e^{-\frac{1}{2}(h_i/\alpha)^2} \right) = \sum_{i=1}^n \left( \log(h_i) - 2 \log(\alpha) - \frac{1}{2}(h_i/\alpha)^2 \right) \\ &= \sum_{i=1}^n (\log(h_i)) - 2n \log(\alpha) - \sum_{i=1}^n \left( \frac{1}{2} h_i^2 \alpha^{-2} \right)\end{aligned}$$

Now, let us take the derivative with respect to  $\alpha$ ,

$$\begin{aligned}\frac{\partial}{\partial \alpha} (\ell(\alpha)) &:= \frac{\partial}{\partial \alpha} \left( \sum_{i=1}^n (\log(h_i)) - 2n \log(\alpha) - \sum_{i=1}^n \left( \frac{1}{2} h_i^2 \alpha^{-2} \right) \right) \\ &= \frac{\partial}{\partial \alpha} \left( \sum_{i=1}^n (\log(h_i)) \right) - \frac{\partial}{\partial \alpha} (2n \log(\alpha)) - \frac{\partial}{\partial \alpha} \left( \sum_{i=1}^n \left( \frac{1}{2} h_i^2 \alpha^{-2} \right) \right) \\ &= 0 - 2n \frac{1}{\alpha} - \sum_{i=1}^n \left( \frac{1}{2} h_i^2 (-2\alpha^{-3}) \right) = -2n\alpha^{-1} + \alpha^{-3} \sum_{i=1}^n (h_i^2)\end{aligned}$$

Next, we set the derivative to 0, solve for  $\alpha$ , and set the solution equal to the ML estimate  $\hat{\alpha}_n$ .

$$\begin{aligned}0 = \frac{\partial}{\partial \alpha} (\ell(\alpha)) &\iff 0 = -2n\alpha^{-1} + \alpha^{-3} \sum_{i=1}^n h_i^2 \iff 2n\alpha^{-1} = \alpha^{-3} \sum_{i=1}^n h_i^2 \\ &\iff 2n\alpha^{-1}\alpha^3 = \sum_{i=1}^n h_i^2 \iff \alpha^2 = \frac{1}{2n} \sum_{i=1}^n h_i^2 \iff \hat{\alpha}_n = \sqrt{\frac{1}{2n} \sum_{i=1}^n h_i^2}\end{aligned}$$

Therefore, the ML estimate of the unknown  $\alpha^* \in \mathbb{A}$  on the basis of our 10 observations  $h_1, h_2, \dots, h_{10}$  of wave heights is

$$\begin{aligned}\hat{\alpha}_{10} &= \sqrt{\frac{1}{2 * 10} \sum_{i=1}^{10} h_i^2} \\ &= \sqrt{\frac{1}{20} (1.50^2 + 2.80^2 + 2.50^2 + 3.20^2 + 1.90^2 + 4.10^2 + 3.60^2 + 2.60^2 + 2.90^2 + 2.30^2)} \approx 2\end{aligned}$$

We use the following script file to compute the MLE  $\hat{\alpha}_{10}$  and plot the PDF at  $\hat{\alpha}_{10}$  in Figure 7.10.

---

```
RayleighOceanHeightsMLE.m
OceanHeights=[1.50, 2.80, 2.50, 3.20, 1.90, 4.10, 3.60, 2.60, 2.90, 2.30];% data
histogram(OceanHeights,1,[min(OceanHeights),max(OceanHeights)],'r',2); % make a histogram
Heights=0:0.1:10; % get some heights for plotting
AlphaHat=sqrt(sum(OceanHeights .^ 2)/(2*length(OceanHeights))) % find the MLE
hold on; % superimpose the PDF at the MLE
plot(Heights,(Heights/AlphaHat.^2) .* exp(-((Heights/AlphaHat).^2)/2))
xlabel('Ocean Wave Heights'); ylabel('Density');
```

---

```
>> RayleighOceanHeightsMLE
AlphaHat = 2.0052
```

## 7.10 More Properties of the Maximum Likelihood Estimator

Next, we list some nice properties of the ML Estimator  $\hat{\Theta}_n$  for the fixed and possibly unknown  $\theta^* \in \Theta$ .

1. The ML Estimator is asymptotically consistent, i.e.  $\hat{\Theta}_n \xrightarrow{P} \theta^*$ .
2. The ML Estimator is asymptotically normal, i.e.  $(\hat{\Theta}_n - \theta^*)/\hat{s}\hat{e}_n \rightsquigarrow \text{Normal}(0, 1)$ .
3. The estimated standard error of the ML Estimator,  $\hat{s}\hat{e}_n$ , can usually be computed analytically using the **Fisher Information**.
4. Because of the previous two properties, the  $1 - \alpha$  confidence interval can also be computed analytically as  $\hat{\Theta}_n \pm z_{\alpha/2}\hat{s}\hat{e}_n$ .
5. The ML Estimator is **equivariant**, i.e.  $\hat{\psi}_n = g(\hat{\theta}_n)$  is the ML Estimate of  $\psi^* = g(\theta^*)$ , for some smooth function  $g(\theta) = \psi : \Theta \rightarrow \Psi$ .
6. We can also obtain the estimated standard error of the estimator  $\hat{\Psi}_n$  of  $\psi^* \in \Psi$  via the **Delta Method**.
7. The ML Estimator is **asymptotically optimal** or **efficient**. This means that the MLE has the smallest variance among the well-behaved class of estimators as the sample size gets larger.
8. ML Estimator is close to the Bayes estimator (obtained in the Bayesian inferential paradigm).

## 7.11 Fisher Information

Let  $X_1, X_2, \dots, X_n \stackrel{IID}{\sim} f(X_1; \theta)$ . Here,  $f(X_1; \theta)$  is the probability density function (pdf) or the probability mass function (pmf) of the RV  $X_1$ . Since all RVs are identically distributed, we simply focus on  $X_1$  without loss of generality.

**Definition 139 (Fisher Information)** The **score function** of an RV  $X$  for which the density is parameterised by  $\theta$  is defined as:

$$\mathcal{S}(X; \theta) := \frac{\partial \log f(X; \theta)}{\partial \theta}, \quad \text{and} \quad E_\theta(\mathcal{S}(X; \theta)) = 0.$$

The **Fisher Information** is

$$I_n := V_\theta \left( \sum_{i=1}^n \mathcal{S}(X_i; \theta) \right) = \sum_{i=1}^n V_\theta(\mathcal{S}(X_i; \theta)) = n I_1(\theta), \quad (7.56)$$

where  $I_1$  is the Fisher Information of just one of the RVs  $X_i$ , e.g.  $X$ :

$$\begin{aligned} I_1(\theta) &:= V_\theta(\mathcal{S}(X; \theta)) = E_\theta(\mathcal{S}^2(X; \theta)) \\ &= -E_\theta \left( \frac{\partial^2 \log f(X; \theta)}{\partial \theta^2} \right) = \begin{cases} -\sum_{x \in \mathbb{X}} \left( \frac{\partial^2 \log f(x; \theta)}{\partial \theta^2} \right) f(x; \theta) & \text{for discrete } X \\ -\int_{x \in \mathbb{X}} \left( \frac{\partial^2 \log f(x; \theta)}{\partial \theta^2} \right) f(x; \theta) dx & \text{for continuous } X \end{cases} \end{aligned} \quad (7.57)$$

Next, we give a **general method** for obtaining:

1. The standard error  $\text{se}_n(\widehat{\Theta}_n)$  of **any** maximum likelihood estimator  $\widehat{\Theta}_n$  of the possibly unknown and fixed parameter of interest  $\theta^* \in \Theta$ , and
2. The  $1 - \alpha$  confidence interval for  $\theta^*$ .

**Proposition 140 (Asymptotic Normality of the ML Estimator & Confidence Intervals)**

Let  $\widehat{\Theta}_n$  be the maximum likelihood estimator of  $\theta^* \in \Theta$  with standard error  $\text{se}_n := \sqrt{V_{\theta^*}(\widehat{\Theta}_n)}$ . Under appropriate regularity conditions, the following propositions are true:

1. The standard error  $\text{se}_n$  can be approximated by the side of a square whose area is the inverse Fisher Information at  $\theta^*$ , and the distribution of  $\widehat{\Theta}_n$  approaches that of the  $\text{Normal}(\theta^*, \text{se}_n^2)$  distribution as the samples size  $n$  gets larger. In other terms:

$$\text{se}_n \approx \sqrt{1/I_n(\theta^*)} \quad \text{and} \quad \frac{\widehat{\Theta}_n - \theta^*}{\text{se}_n} \rightsquigarrow \text{Normal}(0, 1)$$

2. The approximation holds even if we substitute the ML Estimate  $\widehat{\theta}_n$  for  $\theta^*$  and use the estimated standard error  $\widehat{\text{se}}_n$  instead of  $\text{se}_n$ . Let  $\widehat{\text{se}}_n = \sqrt{1/I_n(\widehat{\theta}_n)}$ . Then:

$$\frac{\widehat{\Theta}_n - \theta^*}{\widehat{\text{se}}_n} \rightsquigarrow \text{Normal}(0, 1)$$

3. Using the fact that  $\widehat{\Theta}_n \rightsquigarrow \text{Normal}(\theta^*, \widehat{\text{se}}_n^2)$ , we can construct the estimate of an approximate Normal-based  $1 - \alpha$  confidence interval as:

$$C_n = [\underline{C}_n, \bar{C}_n] = [\widehat{\theta}_n - z_{\alpha/2} \widehat{\text{se}}_n, \widehat{\theta}_n + z_{\alpha/2} \widehat{\text{se}}_n] = \widehat{\theta}_n \pm z_{\alpha/2} \widehat{\text{se}}_n$$

Now, let us do an example.

**Example 218 (MLE and Confidence Interval for the IID Poisson( $\lambda$ ) experiment)** Suppose the fixed parameter  $\lambda^* \in \Lambda = (0, \infty)$  is unknown. Let  $X_1, X_2, \dots, X_n \stackrel{IID}{\sim} \text{Poisson}(\lambda^*)$ . We want to find the ML Estimate  $\widehat{\lambda}_n$  of  $\lambda^*$  and produce a  $1 - \alpha$  confidence interval for  $\lambda^*$ .

The MLE can be obtained as follows:

The likelihood function is:

$$L(\lambda) := L(x_1, x_2, \dots, x_n; \lambda) = \prod_{i=1}^n f(x_i; \lambda) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$$

Hence, the log-likelihood function is:

$$\begin{aligned} \ell(\theta) := \log(L(\lambda)) &= \log \left( \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \right) = \sum_{i=1}^n \log \left( e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \right) = \sum_{i=1}^n (\log(e^{-\lambda}) + \log(\lambda^{x_i}) - \log(x_i!)) \\ &= \sum_{i=1}^n (-\lambda + x_i \log(\lambda) - \log(x_i!)) = \sum_{i=1}^n -\lambda + \sum_{i=1}^n x_i \log(\lambda) - \sum_{i=1}^n \log(x_i!) \\ &= n(-\lambda) + \log(\lambda) \left( \sum_{i=1}^n x_i \right) - \sum_{i=1}^n \log(x_i!) \end{aligned}$$

Next, take the derivative of  $\ell(\lambda)$ :

$$\frac{\partial}{\partial \lambda} \ell(\lambda) = \frac{\partial}{\partial \lambda} \left( n(-\lambda) + \log(\lambda) \left( \sum_{i=1}^n x_i \right) - \sum_{i=1}^n \log(x_i!) \right) = n(-1) + \frac{1}{\lambda} \left( \sum_{i=1}^n x_i \right) + 0$$

and set it equal to 0 to solve for  $\lambda$ , as follows:

$$0 = n(-1) + \frac{1}{\lambda} \left( \sum_{i=1}^n x_i \right) + 0 \iff n = \frac{1}{\lambda} \left( \sum_{i=1}^n x_i \right) \iff \lambda = \frac{1}{n} \left( \sum_{i=1}^n x_i \right) = \bar{x}_n$$

Finally, the ML Estimator of  $\lambda^*$  is  $\hat{\Lambda}_n = \bar{X}_n$  and the ML estimate is  $\hat{\lambda}_n = \bar{x}_n$ .

Now, we want an  $1 - \alpha$  confidence interval for  $\lambda^*$  using the  $\hat{s}\epsilon_n \approx \sqrt{1/I_n(\hat{\lambda}_n)}$  that is based on the Fisher Information  $I_n(\lambda) = nI_1(\lambda)$  given in (7.56). We need  $I_1$  given in (7.57). Since  $X_1, X_2, \dots, X_n \sim \text{Poisson}(\lambda)$ , we have discrete RVs:

$$I_1 = - \sum_{x \in \mathbb{X}} \left( \frac{\partial^2 \log(f(x; \lambda))}{\partial^2 \lambda} \right) f(x; \lambda) = - \sum_{x=0}^{\infty} \left( \frac{\partial^2 \log(f(x; \lambda))}{\partial^2 \lambda} \right) f(x; \lambda)$$

First find

$$\begin{aligned} \frac{\partial^2 \log(f(x; \lambda))}{\partial^2 \lambda} &= \frac{\partial}{\partial \lambda} \left( \frac{\partial}{\partial \lambda} \log(f(x; \lambda)) \right) = \frac{\partial}{\partial \lambda} \left( \frac{\partial}{\partial \lambda} \log \left( e^{-\lambda} \frac{\lambda^x}{x!} \right) \right) \\ &= \frac{\partial}{\partial \lambda} \left( \frac{\partial}{\partial \lambda} (-\lambda + x \log(\lambda) - \log(x!)) \right) = \frac{\partial}{\partial \lambda} \left( -1 + \frac{x}{\lambda} - 0 \right) = -\frac{x}{\lambda^2} \end{aligned}$$

Now, substitute the above expression into the right-hand side of  $I_1$  to obtain:

$$I_1 = - \sum_{x=0}^{\infty} \left( -\frac{x}{\lambda^2} \right) f(x; \lambda) = \frac{1}{\lambda^2} \sum_{x=0}^{\infty} (x) f(x; \lambda) = \frac{1}{\lambda^2} \sum_{x=0}^{\infty} (x) e^{-\lambda} \frac{\lambda^x}{x!} = \frac{1}{\lambda^2} \text{E}_{\lambda}(X) = \frac{1}{\lambda^2} \lambda = \frac{1}{\lambda}$$

In the third-to-last step above, we recognise the sum as the expectation of the  $\text{Poisson}(\lambda)$  RV  $X$ , namely  $\text{E}_{\lambda}(X) = \lambda$ . Therefore, the estimated standard error is:

$$\hat{s}\epsilon_n \approx \sqrt{1/I_n(\hat{\lambda}_n)} = \sqrt{1/(nI_1(\hat{\lambda}_n))} = \sqrt{1/(n(1/\hat{\lambda}_n))} = \sqrt{\hat{\lambda}_n/n}$$

and the approximate  $1 - \alpha$  confidence interval is

$$\hat{\lambda}_n \pm z_{\alpha/2} \hat{s}\epsilon_n = \hat{\lambda}_n \pm z_{\alpha/2} \sqrt{\hat{\lambda}_n/n}$$

Thus, using the MLE and the estimated standard error via the Fisher Information, we can carry out point estimation and confidence interval construction in **most** parametric families of RVs encountered in typical engineering applications.

**Example 219 (Fisher Information of the Bernoulli Experiment)** Suppose  $X_1, X_2, \dots, X_n \stackrel{IID}{\sim} \text{Bernoulli}(\theta^*)$ . Also, suppose that  $\theta^* \in \Theta = [0, 1]$  is unknown. We have already shown in Example 214 that the ML estimator of  $\theta^*$  is  $\hat{\theta}_n = \bar{X}_n$ . Using the identity:

$$\hat{s}\epsilon_n = \frac{1}{\sqrt{I_n(\hat{\theta}_n)}}$$

(1) we can compute  $\widehat{\text{se}}_n(\widehat{\theta}_n)$ , the estimated standard error of the unknown parameter  $\theta^*$  as follows:

$$\widehat{\text{se}}_n(\widehat{\theta}_n) = \frac{1}{\sqrt{I_n(\widehat{\theta}_n)}} = \frac{1}{\sqrt{nI_1(\widehat{\theta}_n)}} .$$

So, we need to first compute  $I_1(\theta)$ , the Fisher Information of one sample. Due to (7.57) and the fact that the Bernoulli( $\theta^*$ ) distributed RV  $X$  is discrete with probability mass function  $f(x; \theta) = \theta^x(1 - \theta)^{1-x}$ , for  $x \in \mathbb{X} := \{0, 1\}$ , we have,

$$I_1(\theta) = -\mathbb{E}_\theta \left( \frac{\partial^2 \log f(X; \theta)}{\partial^2 \theta} \right) = -\sum_{x \in \mathbb{X}=\{0,1\}} \left( \frac{\partial^2 \log (\theta^x(1 - \theta)^{1-x})}{\partial^2 \theta} \right) \theta^x(1 - \theta)^{1-x}$$

Next, let us compute,

$$\begin{aligned} \frac{\partial^2 \log (\theta^x(1 - \theta)^{1-x})}{\partial^2 \theta} &:= \frac{\partial}{\partial \theta} \left( \frac{\partial}{\partial \theta} (\log (\theta^x(1 - \theta)^{1-x})) \right) = \frac{\partial}{\partial \theta} \left( \frac{\partial}{\partial \theta} (x \log(\theta) + (1 - x) \log(1 - \theta)) \right) \\ &= \frac{\partial}{\partial \theta} (x\theta^{-1} + (1 - x)(1 - \theta)^{-1}(-1)) = \frac{\partial}{\partial \theta} (x\theta^{-1} - (1 - x)(1 - \theta)^{-1}) \\ &= x(-1)\theta^{-1-1} - (1 - x)(-1)(1 - \theta)^{-1-1}(-1) = -x\theta^{-2} - (1 - x)(1 - \theta)^{-2} \end{aligned}$$

Now, we compute the expectation  $I_1$ , i.e. the sum over the two possible values of  $x \in \{0, 1\}$ ,

$$\begin{aligned} I_1(\theta) &= -\sum_{x \in \mathbb{X}=\{0,1\}} \left( \frac{\partial^2 \log (\theta^x(1 - \theta)^{1-x})}{\partial^2 \theta} \right) \theta^x(1 - \theta)^{1-x} \\ &= -((-0\theta^{-2} - (1 - 0)(1 - \theta)^{-2})\theta^0(1 - \theta)^{1-0} + (-1\theta^{-2} - (1 - 1)(1 - \theta)^{-2})\theta^1(1 - \theta)^{1-1}) \\ &= -((0 - 1(1 - \theta)^{-2})1(1 - \theta)^1 + (-\theta^{-2} - 0)\theta^11) = (1 - \theta)^{-2}(1 - \theta)^1 + \theta^{-2}\theta^1 \\ &= (1 - \theta)^{-1} + \theta^{-1} = \frac{1}{1 - \theta} + \frac{1}{\theta} = \frac{\theta}{\theta(1 - \theta)} + \frac{1 - \theta}{\theta(1 - \theta)} = \frac{\theta + (1 - \theta)}{\theta(1 - \theta)} = \frac{1}{\theta(1 - \theta)} \end{aligned}$$

Therefore, the desired estimated standard error of our estimator, can be obtained by substituting the ML estimate  $\widehat{\theta}_n = \bar{x}_n := n^{-1} \sum_{i=1}^n x_i$  of the unknown  $\theta^*$  as follows:

$$\widehat{\text{se}}_n(\widehat{\theta}_n) = \frac{1}{\sqrt{I_n(\widehat{\theta}_n)}} = \frac{1}{\sqrt{nI_1(\widehat{\theta}_n)}} = \sqrt{\frac{1}{n \frac{1}{\widehat{\theta}_n(1 - \widehat{\theta}_n)}}} = \sqrt{\frac{\widehat{\theta}_n(1 - \widehat{\theta}_n)}{n}} = \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}} .$$

(2) Using  $\widehat{\text{se}}_n(\widehat{\theta}_n)$  we can construct an approximate 95% confidence interval  $C_n$  for  $\theta^*$ , due to the asymptotic normality of the ML estimator of  $\theta^*$ , as follows:

$$C_n = \widehat{\theta}_n \pm 1.96 \sqrt{\frac{\widehat{\theta}_n(1 - \widehat{\theta}_n)}{n}} = \bar{x}_n \pm 1.96 \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}}$$

Recall that  $C_n$  is the realisation of a random set based on your observed samples or data  $x_1, x_2, \dots, x_n$ . Furthermore,  $C_n$ 's construction procedure ensures the engulfing of the unknown  $\theta^*$  with probability approaching 0.95 as the sample size  $n$  gets large.

**Example 220 ([Fisher Information of the Exponential Experiment] )** Let us get our hands dirty with a continuous RV next. Let  $X_1, X_2, \dots, X_n \stackrel{IID}{\sim} \text{Exponential}(\lambda^*)$ . We saw that the ML

estimator of  $\lambda^* \in \Lambda = (0, \infty)$  is  $\widehat{\Lambda}_n = 1/\bar{X}_n$  and its ML estimate is  $\widehat{\lambda}_n = 1/\bar{x}_n$ , where  $x_1, x_2, \dots, x_n$  are our observed data.

(1) Let us obtain the Fisher Information  $I_n$  for this experiment to find the standard error:

$$\widehat{\text{se}}_n(\widehat{\Lambda}_n) = \frac{1}{\sqrt{I_n(\widehat{\lambda}_n)}} = \frac{1}{\sqrt{nI_1(\widehat{\lambda}_n)}}$$

and construct an approximate 95% confidence interval for  $\lambda^*$  using the asymptotic normality of its ML estimator  $\widehat{\Lambda}_n$ .

So, we need to first compute  $I_1(\theta)$ , the Fisher Information of one sample. Due to (7.57) and the fact that the Exponential( $\lambda^*$ ) distributed RV  $X$  is continuous with probability density function  $f(x; \lambda) = \lambda e^{-\lambda x}$ , for  $x \in \mathbb{X} := [0, \infty)$ , we have,

$$I_1(\theta) = -E_\theta \left( \frac{\partial^2 \log f(X; \theta)}{\partial^2 \theta} \right) = - \int_{x \in \mathbb{X} = [0, \infty)} \left( \frac{\partial^2 \log (\lambda e^{-\lambda x})}{\partial^2 \lambda} \right) \lambda e^{-\lambda x} dx$$

Let us compute the above integrand next.

$$\begin{aligned} \frac{\partial^2 \log (\lambda e^{-\lambda x})}{\partial^2 \lambda} &:= \frac{\partial}{\partial \lambda} \left( \frac{\partial}{\partial \lambda} (\log (\lambda e^{-\lambda x})) \right) = \frac{\partial}{\partial \lambda} \left( \frac{\partial}{\partial \lambda} (\log(\lambda) + \log(e^{-\lambda x})) \right) \\ &= \frac{\partial}{\partial \lambda} \left( \frac{\partial}{\partial \lambda} (\log(\lambda) - \lambda x) \right) = \frac{\partial}{\partial \lambda} (\lambda^{-1} - x) = -\lambda^{-2} - 0 = -\frac{1}{\lambda^2} \end{aligned}$$

Now, let us evaluate the integral by recalling that the expectation of the constant 1 is 1 for any RV  $X$  governed by some parameter, say  $\theta$ . For instance when  $X$  is a continuous RV,  $E_\theta(1) = \int_{x \in \mathbb{X}} 1 f(x; \theta) = \int_{x \in \mathbb{X}} f(x; \theta) = 1$ . Therefore, the Fisher Information of one sample is

$$\begin{aligned} I_1(\theta) &= - \int_{x \in \mathbb{X} = [0, \infty)} \left( \frac{\partial^2 \log (\lambda e^{-\lambda x})}{\partial^2 \lambda} \right) \lambda e^{-\lambda x} dx = - \int_0^\infty \left( -\frac{1}{\lambda^2} \right) \lambda e^{-\lambda x} dx \\ &= - \left( -\frac{1}{\lambda^2} \right) \int_0^\infty \lambda e^{-\lambda x} dx = \frac{1}{\lambda^2} 1 = \frac{1}{\lambda^2} \end{aligned}$$

Now, we can compute the desired estimated standard error, by substituting in the ML estimate  $\widehat{\lambda}_n = 1/(\bar{x}_n) := 1/(\sum_{i=1}^n x_i)$  of  $\lambda^*$ , as follows:

$$\widehat{\text{se}}_n(\widehat{\Lambda}_n) = \frac{1}{\sqrt{I_n(\widehat{\lambda}_n)}} = \frac{1}{\sqrt{nI_1(\widehat{\lambda}_n)}} = \frac{1}{\sqrt{n \frac{1}{\widehat{\lambda}_n^2}}} = \frac{\widehat{\lambda}_n}{\sqrt{n}} = \frac{1}{\sqrt{n} \bar{x}_n}$$

Using  $\widehat{\text{se}}_n(\widehat{\lambda}_n)$  we can construct an approximate 95% confidence interval  $C_n$  for  $\lambda^*$ , due to the asymptotic normality of the ML estimator of  $\lambda^*$ , as follows:

$$C_n = \widehat{\lambda}_n \pm 1.96 \frac{\widehat{\lambda}_n}{\sqrt{n}} = \frac{1}{\bar{x}_n} \pm 1.96 \frac{1}{\sqrt{n} \bar{x}_n} .$$

Let us compute the ML estimate and the 95% confidence interval for the rate parameter for the waiting times at the Orbiter bus-stop (see labwork 216). The sample mean  $\bar{x}_{132} = 9.0758$  and the ML estimate is:

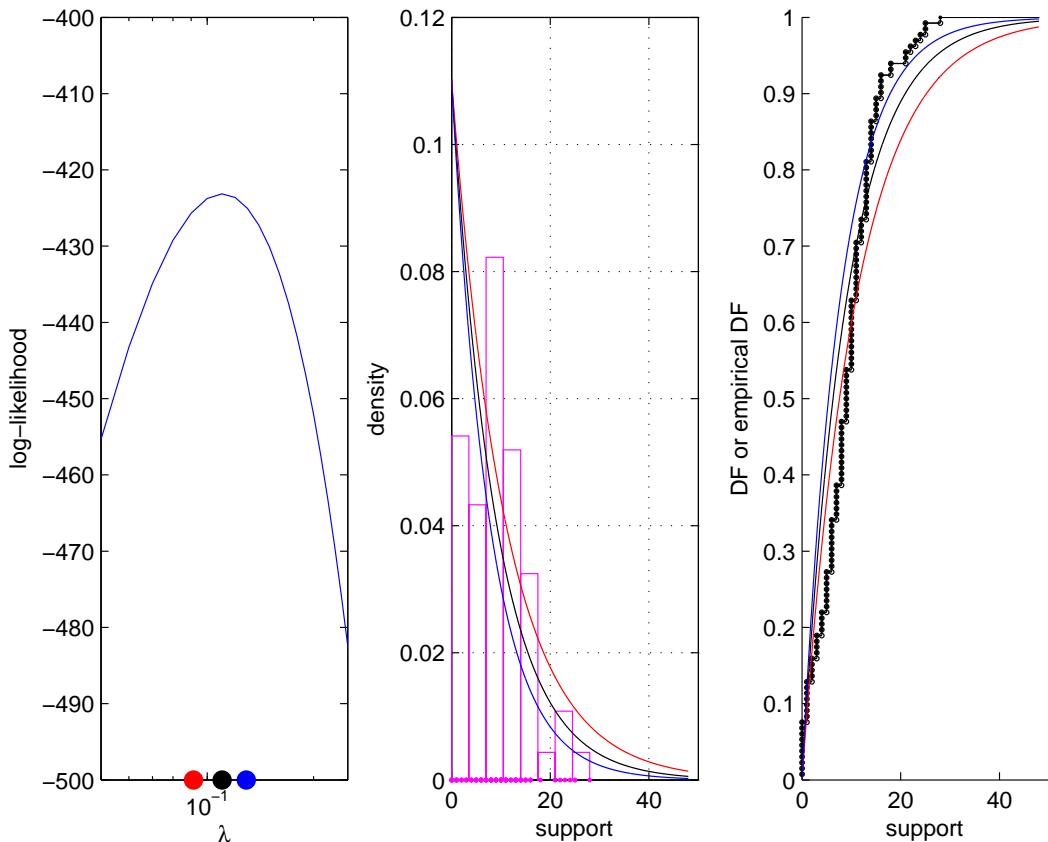
$$\widehat{\lambda}_{132} = 1/\bar{x}_{132} = 1/9.0758 = 0.1102 ,$$

and the 95% confidence interval is:

$$C_n = \hat{\lambda}_{132} \pm 1.96 \frac{\hat{\lambda}_{132}}{\sqrt{132}} = \frac{1}{\bar{x}_{132}} \pm 1.96 \frac{1}{\sqrt{132} \bar{x}_{132}} = 0.1102 \pm 1.96 \cdot 0.0096 = [0.0914, 0.1290] .$$

Notice how poorly the exponential PDF  $f(x; \hat{\lambda}_{132} = 0.1102)$  and the DF  $F(x; \hat{\lambda}_{132} = 0.1102)$  based on the MLE fits with the histogram and the empirical DF, respectively, in Figure 7.11, despite taking the the confidence interval into account. This is a further indication of the inadequacy of our parametric model.

Figure 7.11: Plot of  $\log(L(\lambda))$  as a function of the parameter  $\lambda$ , the MLE  $\hat{\lambda}_{132} = 0.1102$  and 95% confidence interval  $C_n = [0.0914, 0.1290]$  for Fenemore-Wang Orbiter Waiting Times Experiment from STAT 218 S2 2007. The PDF and the DF at (1) the MLE 0.1102 (black), (2) lower 95% confidence bound 0.0914 (red) and (3) upper 95% confidence bound 0.1290 (blue) are compared with a histogram and the empirical DF.



**Labwork 221 (Maximum likelihood estimation for Orbiter bus-stop)** The above analysis was undertaken with the following M-file:

---

```
ExponentialMLECIOrbiter.m
OrbiterData; % load the Orbiter Data sampleTimes
% L = Log Likelihood of data x as a function of parameter lambda
L=@(lambda)sum(log(lambda*exp(-lambda * sampleTimes)));
LAMBDA=[0.01:0.01:1]; % sample some values for lambda
```

---

```

clf;
subplot(1,3,1);
semilogx(LAMBDA, arrayfun(L,LAMBDA)); % plot the Log Likelihood function
axis([0.05 0.25 -500 -400])
SampleMean = mean(sampleTimes) % sample mean
MLE = 1/SampleMean % ML estimate is 1/mean
n=length(sampleTimes) % sample size
StdErr=1/(sqrt(n)*SampleMean) % Standard Error
MLE95CI=[MLE-(1.96*StdErr), MLE+(1.96*StdErr)] % 95 % CI
hold on; % plot the MLE
plot([MLE], [-500], 'k.', 'MarkerSize',25);
plot([MLE95CI(1)], [-500], 'r.', 'MarkerSize',25);
plot([MLE95CI(2)], [-500], 'b.', 'MarkerSize',25);
ylabel('log-likelihood'); xlabel('\lambda');
subplot(1,3,2); % plot a histogram estimate
histogram(sampleTimes,1,[min(sampleTimes),max(sampleTimes)],'m',2);
hold on; TIMES=[0.00001:0.01:max(sampleTimes)+20]; % points on support
% plot PDF at MLE and 95% CI to compare with histogram
plot(TIMES,MLE*exp(-MLE*TIMES), 'k-')
plot(TIMES,MLE*exp(-MLE95CI(1)*TIMES), 'r-'); plot(TIMES,MLE*exp(-MLE95CI(2)*TIMES), 'b-')
% compare the empirical DF to the best fitted DF at MLE and 95% CI
subplot(1,3,3)
ECDF(sampleTimes,5,0.0,20); hold on; plot(TIMES,ExponentialCdf(TIMES,MLE), 'k-');
plot(TIMES,ExponentialCdf(TIMES,MLE95CI(1)), 'r-'); plot(TIMES,ExponentialCdf(TIMES,MLE95CI(2)), 'b-')
ylabel('DF or empirical DF'); xlabel('support');

```

---

A call to the script generates Figure 7.11 and the following output of the sample mean, MLE, sample size, standard error and the 95% confidence interval.

```

>> ExponentialMLECIOrbiter
SampleMean =    9.0758
MLE =      0.1102
n =      132
StdErr =    0.0096
MLE95CI =    0.0914    0.1290

```

**Labwork 222 (Maximum likelihood estimation for your bus-stop)** Recall labwork 135 where you modeled the arrival of buses using  $\text{Exponential}(\lambda^* = 0.1)$  distributed inter-arrival time with a mean of  $1/\lambda^* = 10$  minutes. Using the data of these seven inter-arrival times at your ID-seeded bus stop and pretending that you do not know the true  $\lambda^*$ , report (1) the ML estimate of  $\lambda^*$ , (2) 95% confidence interval for it and (3) whether the true value  $\lambda^* = 1/10$  is engulfed by your confidence interval.

## 7.12 Delta Method

A more general estimation problem of interest concerns some function of the parameter  $\theta \in \Theta$ , say  $g(\theta) = \psi : \Theta \rightarrow \Psi$ . So,  $g(\theta) = \psi$  is a function from the parameter space  $\Theta$  to  $\Psi$ . Thus, we are not only interested in estimating the fixed and possibly unknown  $\theta^* \in \Theta$  using the ML estimator  $\hat{\Theta}_n$  and its ML estimate  $\hat{\theta}_n$ , but also in estimating  $\psi^* = g(\theta^*) \in \Psi$  via an estimator  $\hat{\Psi}_n$  and its estimate  $\hat{\psi}_n$ . We exploit the equivariance property of the ML estimator  $\hat{\Theta}_n$  of  $\theta^*$  and use the Delta method to find the following analytically:

1. The ML estimator of  $\psi^* = g(\theta^*) \in \Psi$  is

$$\hat{\Psi}_n = g(\hat{\Theta}_n)$$

and its point estimate is

$$\widehat{\psi}_n = g(\widehat{\theta}_n)$$

2. Suppose  $g(\theta) = \psi : \Theta \rightarrow \Psi$  is **any** smooth function of  $\theta$ , i.e.  $g$  is differentiable, and  $g'(\theta) := \frac{\partial}{\partial \theta} g(\theta) \neq 0$ . Then, the distribution of the ML estimator  $\widehat{\Psi}_n$  is asymptotically  $\text{Normal}(\psi^*, \widehat{s}\text{e}_n(\widehat{\Psi}_n)^2)$ , i.e.:

$$\frac{\widehat{\Psi}_n - \psi^*}{\widehat{s}\text{e}_n(\widehat{\Psi}_n)} \rightsquigarrow \text{Normal}(0, 1)$$

where the standard error  $\widehat{s}\text{e}_n(\widehat{\Psi}_n)$  of the ML estimator  $\widehat{\Psi}_n$  of the unknown quantity  $\psi^* \in \Psi$  can be obtained from the standard error  $\widehat{s}\text{e}_n(\widehat{\Theta}_n)$  of the ML estimator  $\widehat{\Theta}_n$  of the parameter  $\theta^* \in \Theta$ , as follows:

$$\widehat{s}\text{e}_n(\widehat{\Psi}_n) = |g'(\widehat{\theta}_n)| \widehat{s}\text{e}_n(\widehat{\Theta}_n)$$

3. Using  $\text{Normal}(\psi^*, \widehat{s}\text{e}_n(\widehat{\Psi}_n)^2)$ , we can construct the estimate of an approximate Normal-based  $1 - \alpha$  confidence interval for  $\psi^* \in \Psi$ :

$$C_n = [\underline{C}_n, \bar{C}_n] = \widehat{\psi}_n \pm z_{\alpha/2} \widehat{s}\text{e}_n(\widehat{\psi}_n)$$

Let us do an example next.

**Example 223** Let  $X_1, X_2, \dots, X_n \stackrel{IID}{\sim} \text{Bernoulli}(\theta^*)$ . Let  $\psi = g(\theta) = \log(\theta/(1-\theta))$ . Suppose we are interested in producing a point estimate and confidence interval for  $\psi^* = g(\theta^*)$ . We can use the Delta method as follows:

First, the estimated standard error of the ML estimator of  $\theta^*$ , as shown in Example 219, is

$$\widehat{s}\text{e}_n(\widehat{\Theta}_n) = \sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}.$$

The ML estimator of  $\psi^*$  is:

$$\widehat{\Psi}_n = \log(\widehat{\Theta}_n/(1-\widehat{\Theta}_n))$$

and the ML estimate of  $\psi^*$  is:

$$\widehat{\psi}_n = \log(\widehat{\theta}_n/(1-\widehat{\theta}_n)).$$

Since,  $g'(\theta) = 1/(\theta(1-\theta))$ , by the Delta method, the estimated standard error of the ML estimator of  $\psi^*$  is:

$$\widehat{s}\text{e}_n(\widehat{\Psi}_n) = |g'(\widehat{\theta}_n)|(\widehat{s}\text{e}_n(\widehat{\Theta}_n)) = \frac{1}{\widehat{\theta}_n(1-\widehat{\theta}_n)} \sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} = \frac{1}{\sqrt{n\widehat{\theta}_n(1-\widehat{\theta}_n)}} = \frac{1}{\sqrt{n\bar{x}_n(1-\bar{x}_n)}}.$$

An approximate 95% confidence interval for  $\psi^* = \log(\theta^*/(1-\theta^*))$  is:

$$\widehat{\psi}_n \pm \frac{1.96}{\sqrt{n\widehat{\theta}_n(1-\widehat{\theta}_n)}} = \log(\widehat{\theta}_n/(1-\widehat{\theta}_n)) \pm \frac{1.96}{\sqrt{n\widehat{\theta}_n(1-\widehat{\theta}_n)}} = \log(\bar{x}_n/(1-\bar{x}_n)) \pm \frac{1.96}{\sqrt{n\bar{x}_n(1-\bar{x}_n)}}.$$

**Example 224 (Delta Method for a Normal Experiment)** Let us try the Delta method on a continuous RV. Let  $X_1, X_2, \dots, X_n \stackrel{IID}{\sim} \text{Normal}(\mu^*, \sigma^{*2})$ . Suppose that  $\mu^*$  is known and  $\sigma^*$  is unknown. Let us derive the ML estimate  $\hat{\psi}_n$  of  $\psi^* = \log(\sigma^*)$  and a 95% confidence interval for it in 6 steps.

(1) First let us find the log-likelihood function  $\ell(\sigma)$

$$\begin{aligned}
\ell(\sigma) := \log(L(\sigma)) &:= \log(L(x_1, x_2, \dots, x_n; \sigma)) = \log \left( \prod_{i=1}^n f(x_i; \sigma) \right) = \sum_{i=1}^n \log(f(x_i; \sigma)) \\
&= \sum_{i=1}^n \log \left( \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2\sigma^2}(x_i - \mu)^2 \right) \right) \quad \because f(x_i; \sigma) \text{ in (3.39) is pdf of } \text{Normal}(\mu, \sigma^2) \text{ RV with known } \mu \\
&= \sum_{i=1}^n \left( \log \left( \frac{1}{\sigma \sqrt{2\pi}} \right) + \log \left( \exp \left( -\frac{1}{2\sigma^2}(x_i - \mu)^2 \right) \right) \right) \\
&= \sum_{i=1}^n \log \left( \frac{1}{\sigma \sqrt{2\pi}} \right) + \sum_{i=1}^n \left( -\frac{1}{2\sigma^2}(x_i - \mu)^2 \right) = n \log \left( \frac{1}{\sigma \sqrt{2\pi}} \right) + \left( -\frac{1}{2\sigma^2} \right) \sum_{i=1}^n (x_i - \mu)^2 \\
&= n \left( \log \left( \frac{1}{\sqrt{2\pi}} \right) + \log \left( \frac{1}{\sigma} \right) \right) - \left( \frac{1}{2\sigma^2} \right) \sum_{i=1}^n (x_i - \mu)^2 \\
&= n \log \left( \sqrt{2\pi}^{-1} \right) + n \log(\sigma^{-1}) - \left( \frac{1}{2\sigma^2} \right) \sum_{i=1}^n (x_i - \mu)^2 \\
&= -n \log \left( \sqrt{2\pi} \right) - n \log(\sigma) - \left( \frac{1}{2\sigma^2} \right) \sum_{i=1}^n (x_i - \mu)^2
\end{aligned}$$

(2) Let us find its derivative with respect to the unknown parameter  $\sigma$  next.

$$\begin{aligned}
\frac{\partial}{\partial \sigma} \ell(\sigma) &:= \frac{\partial}{\partial \sigma} \left( -n \log \left( \sqrt{2\pi} \right) - n \log(\sigma) - \left( \frac{1}{2\sigma^2} \right) \sum_{i=1}^n (x_i - \mu)^2 \right) \\
&= \frac{\partial}{\partial \sigma} \left( -n \log \left( \sqrt{2\pi} \right) \right) - \frac{\partial}{\partial \sigma} (n \log(\sigma)) - \frac{\partial}{\partial \sigma} \left( \left( \frac{1}{2\sigma^2} \right) \sum_{i=1}^n (x_i - \mu)^2 \right) \\
&= 0 - n \frac{\partial}{\partial \sigma} (\log(\sigma)) - \left( \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 \right) \frac{\partial}{\partial \sigma} (\sigma^{-2}) \\
&= -n\sigma^{-1} - \left( \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 \right) (-2\sigma^{-3}) = -n\sigma^{-1} + \sigma^{-3} \sum_{i=1}^n (x_i - \mu)^2
\end{aligned}$$

(3) Now, let us set the derivative equal to 0 and solve for  $\sigma$ .

$$\begin{aligned}
0 = -n\sigma^{-1} + \sigma^{-3} \sum_{i=1}^n (x_i - \mu)^2 &\iff n\sigma^{-1} = \sigma^{-3} \sum_{i=1}^n (x_i - \mu)^2 \iff n\sigma^{-1}\sigma^{+3} = \sum_{i=1}^n (x_i - \mu)^2 \\
&\iff n\sigma^{-1+3} = \sum_{i=1}^n (x_i - \mu)^2 \iff n\sigma^2 = \sum_{i=1}^n (x_i - \mu)^2 \\
&\iff \sigma^2 = \left( \sum_{i=1}^n (x_i - \mu)^2 \right) / n \iff \sigma = \sqrt{\sum_{i=1}^n (x_i - \mu)^2 / n}
\end{aligned}$$

Finally, we set the solution, i.e. the maximiser of the concave-down log-likelihood function of  $\sigma$  with a known and fixed  $\mu^*$  as our ML estimate  $\hat{\sigma}_n = \sqrt{\sum_{i=1}^n (x_i - \mu^*)^2 / n}$ . Analogously, the ML estimator

of  $\sigma^*$  is  $\widehat{\Sigma}_n = \sqrt{\sum_{i=1}^n (X_i - \mu^*)^2/n}$ . Don't confuse  $\Sigma$ , the upper-case sigma, with  $\sum_{i=1}^n \bigcirc_i$ , the summation over some  $\bigcirc_i$ 's. This is usually clear from the context.

(4) Next, let us get the estimated standard error  $\widehat{s}\mathbf{e}_n$  for the estimator of  $\sigma^*$  via Fisher Information. The Log-likelihood function of  $\sigma$ , based on one sample from the  $\text{Normal}(\mu, \sigma^2)$  RV with known  $\mu$  is,

$$\log f(x; \sigma) = \log \left( \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2\sigma^2} (x - \mu)^2 \right) \right) = -\log(\sqrt{2\pi}) - \log(\sigma) - \left( \frac{1}{2\sigma^2} \right) (x - \mu)^2$$

Therefore, in much the same way as in part (2) earlier,

$$\begin{aligned} \frac{\partial^2 \log f(x; \sigma)}{\partial^2 \sigma} &:= \frac{\partial}{\partial \sigma} \left( \frac{\partial}{\partial \sigma} \left( -\log(\sqrt{2\pi}) - \log(\sigma) - \left( \frac{1}{2\sigma^2} \right) (x - \mu)^2 \right) \right) \\ &= \frac{\partial}{\partial \sigma} (-\sigma^{-1} + \sigma^{-3}(x - \mu)^2) = \sigma^{-2} - 3\sigma^{-4}(x - \mu)^2 \end{aligned}$$

Now, we compute the Fisher Information of one sample as an expectation of the continuous RV  $X$  over  $\mathbb{X} = (-\infty, \infty)$  with density  $f(x; \sigma)$ ,

$$\begin{aligned} I_1(\sigma) &= - \int_{x \in \mathbb{X} = (-\infty, \infty)} \left( \frac{\partial^2 \log f(x; \sigma)}{\partial^2 \lambda} \right) f(x; \sigma) dx = - \int_{-\infty}^{\infty} (\sigma^{-2} - 3\sigma^{-4}(x - \mu)^2) f(x; \sigma) dx \\ &= \int_{-\infty}^{\infty} -\sigma^{-2} f(x; \sigma) dx + \int_{-\infty}^{\infty} 3\sigma^{-4}(x - \mu)^2 f(x; \sigma) dx \\ &= -\sigma^{-2} \int_{-\infty}^{\infty} f(x; \sigma) dx + 3\sigma^{-4} \int_{-\infty}^{\infty} (x - \mu)^2 f(x; \sigma) dx \\ &= -\sigma^{-2} + 3\sigma^{-4}\sigma^2 \quad \because \sigma^2 = \text{V}(X) = \text{E}(X - \text{E}(X))^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x; \sigma) dx \\ &= -\sigma^{-2} + 3\sigma^{-4+2} = -\sigma^{-2} + 3\sigma^{-2} = 2\sigma^{-2} \end{aligned}$$

Therefore, the estimated standard error of the estimator of the unknown  $\sigma^*$  is

$$\widehat{s}\mathbf{e}_n(\widehat{\Sigma}_n) = \frac{1}{\sqrt{I_n(\widehat{\sigma}_n)}} = \frac{1}{\sqrt{nI_1(\widehat{\sigma}_n)}} = \frac{1}{\sqrt{n2\sigma^{-2}}} = \frac{\sigma}{\sqrt{2n}} .$$

(5) Given that  $\psi = g(\sigma) = \log(\sigma)$ , we derive the estimated standard error of  $\psi^* = \log(\sigma^*)$  via the Delta method as follows:

$$\widehat{s}\mathbf{e}_n(\widehat{\Psi}_n) = |g'(\sigma)| \widehat{s}\mathbf{e}_n(\widehat{\Sigma}_n) = \left| \frac{\partial}{\partial \sigma} \log(\sigma) \right| \frac{\sigma}{\sqrt{2n}} = \frac{1}{\sigma} \frac{\sigma}{\sqrt{2n}} = \frac{1}{\sqrt{2n}} .$$

(6) Finally, the 95% confidence interval for  $\psi^*$  is  $\widehat{\psi}_n \pm 1.96 \widehat{s}\mathbf{e}_n(\widehat{\Psi}_n) = \log(\widehat{\sigma}_n) \pm 1.96 \frac{1}{\sqrt{2n}}$ .

## 7.13 Non-parametric DF Estimation

So far, we have been interested in some estimation problems involved in parametric experiments. In parametric experiments, the parameter space  $\Theta$  can have many dimensions, but these are finite. For example, in the  $n$  IID Bernoulli( $\theta^*$ ) and the  $n$  IID Exponential( $\lambda^*$ ) experiments:

$$\begin{aligned} X_1, \dots, X_n &\stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta^*), & \theta^* \in \Theta = [0, 1] \subset \mathbb{R}^1, \\ X_1, \dots, X_n &\stackrel{\text{IID}}{\sim} \text{Exponential}(\lambda^*), & \lambda^* \in \Lambda = (0, \infty) \subset \mathbb{R}^1, \end{aligned}$$

the parameter spaces  $\Theta$  and  $\Lambda$  are of dimension 1. Similarly, in the  $n$  IID Normal( $\mu, \sigma^2$ ) and the  $n$  IID Lognormal( $\lambda, \zeta$ ), experiments:

$$\begin{aligned} X_1, \dots, X_n &\stackrel{\text{IID}}{\sim} \text{Normal}(\mu, \sigma^2), & (\mu, \sigma^2) \in \Theta = (-\infty, +\infty) \times (0, +\infty) \subset \mathbb{R}^2 \\ X_1, \dots, X_n &\stackrel{\text{IID}}{\sim} \text{Lognormal}(\lambda, \zeta), & (\lambda, \zeta) \in \Theta = (0, +\infty) \times (0, +\infty) \subset \mathbb{R}^2 \end{aligned}$$

the parameter space is of dimension 2.

An experiment with an infinite dimensional parameter space  $\Theta$  is said to be **non-parametric**. Next we consider a non-parametric experiment in which  $n$  IID samples are drawn according to some fixed and possibly unknown DF  $F^*$  from the space of **All Distribution Functions**:

$X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} F^*, \quad F^* \in \Theta = \{\text{All DFs}\} := \{F(x; F) : F \text{ is a DF}\}$

where the DF  $F(x; F)$  is indexed or parameterised by itself. Thus, the parameter space  $\Theta = \{\text{All DFs}\}$  is the **infinite dimensional** space of **All DFs**. In this section, we look at estimation problems in non-parametric experiments with an infinite dimensional parameter space. That is, we want to estimate the DF  $F^*$  from which our IID data are drawn.

The next proposition is often referred to as the **fundamental theorem of statistics** and is at the heart of non-parametric inference, empirical processes, and computationally intensive bootstrap techniques. Recall Definition 56 of the  $n$ -sample empirical distribution function (EDF or ECDF)  $\widehat{F}_n$  that assigns a probability mass of  $1/n$  at each data point  $x_i$ :

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

**Proposition 141 (Gilvenko-Cantelli Theorem)** Let  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} F^*$ . Then:

$$\sup_x |\widehat{F}_n(x) - F^*(x)| \xrightarrow{P} 0.$$

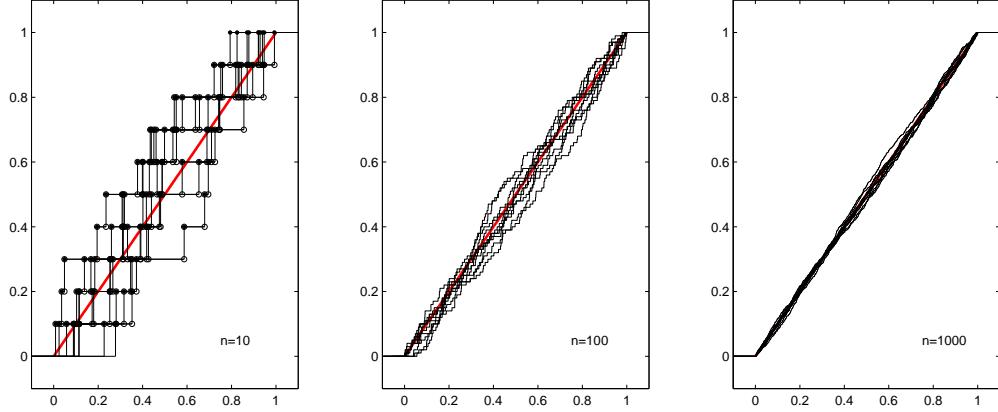
**Heuristic Interpretation of the Gilvenko-Cantelli Theorem:** As the sample size  $n$  increases, the empirical distribution function  $\widehat{F}_n$  converges to the true DF  $F^*$  in probability, as shown in Figure 7.12.

**Proposition 142 (The Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality)** Let  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} F^*$ . Then, for any  $\epsilon > 0$ :

$$P \left( \sup_x |\widehat{F}_n(x) - F^*(x)| > \epsilon \right) \leq 2 \exp(-2n\epsilon^2) \tag{7.58}$$

Recall that  $\sup(A)$  or supremum of a set  $A \subset \mathbb{R}$  is the least upper bound of every element in  $A$ .

Figure 7.12: Plots of ten distinct ECDFs  $\hat{F}_n$  based on 10 sets of  $n$  IID samples from Uniform(0, 1) RV  $X$ , as  $n$  increases from 10 to 100 to 1000. The DF  $F(x) = x$  over  $[0, 1]$  is shown in red. The script of Labwork 240 was used to generate this plot.



### 7.13.1 Estimating DF

Let  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} F^*$ , where  $F^*$  is some particular DF in the space of all possible DFs, i.e. the experiment is non-parametric. Then, based on the data sequence  $X_1, X_2, \dots, X_n$  we want to estimate  $F^*$ .

For any fixed value of  $x$ , the expectation and variance of the empirical DF (3.82) are:

$$E(\hat{F}_n(x)) = F^*(x) \implies \text{bias}_n(\hat{F}_n(x)) = 0 \quad (7.59)$$

$$V(\hat{F}_n(x)) = \frac{F^*(x)(1 - F^*(x))}{n} \implies \lim_{n \rightarrow \infty} \text{se}_n(\hat{F}_n(x)) = 0 \quad (7.60)$$

Therefore, by Proposition 125, the empirical DF evaluated at  $x$ , i.e.  $\hat{F}_n(x)$  is an asymptotically consistent estimator of the DF evaluated at  $x$ , i.e.  $F^*(x)$ . More formally, (7.59) and (7.60), by Proposition 125, imply that for any fixed value of  $x$ :

$$\hat{F}_n(x) \xrightarrow{P} F^*(x).$$

We are interested in a point estimate of the entire DF  $F^*$ , i.e.  $F^*(x)$  over all  $x$ . A point estimator  $T_n = T_n(X_1, X_2, \dots, X_n)$  of a fixed and possibly unknown  $F \in \{\text{All DFs}\}$  is the empirical DF  $\hat{F}_n$ . This estimator has an asymptotically desirable property:

$$\sup_x |\hat{F}_n(x) - F^*(x)| \xrightarrow{P} 0$$

because of the Gilvenko-Cantelli theorem in Proposition 141. Thus, we can simply use  $\hat{F}_n$ , based on the realized data  $(x_1, x_2, \dots, x_n)$ , as a point estimate of  $F^*$ .

On the basis of the DKW inequality (7.58), we can obtain a  $1 - \alpha$  confidence set or **confidence band**  $C_n(x) := [\underline{C}_n(x), \bar{C}_n(x)]$  about our point estimate of  $F^*$ :

$$\begin{aligned} \underline{C}_n(x) &= \max\{\hat{F}_n(x) - \epsilon_n, 0\}, \\ \bar{C}_n(x) &= \min\{\hat{F}_n(x) + \epsilon_n, 1\}, \\ \epsilon_n &= \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}. \end{aligned} \quad (7.61)$$

It follows from (7.58) that for any fixed and possibly unknown  $F^*$ :

$$P(\underline{C}_n(x) \leq F^*(x) \leq \bar{C}_n(x)) \geq 1 - \alpha .$$

Let us look at a simple example next.

**Labwork 225 (Estimating the DF of Uniform(0, 1) RV)** Consider the problem of estimating the DF of Uniform(0, 1) RV  $U$  on the basis of  $n=10$  samples. We use the function `ECDF` of Lab-work 236 and MATLAB's built-in function `stairs` to render the plots. Figure 7.13 was generated by `PlotUniformECDFsConfBands.m` given below.

---

```
% script PlotUniformECDFsConfBands.m to plot the ECDF from 10 and 100 samples
% from Uniform(0,1) RV
rand('twister',76534); % initialize the Uniform(0,1) Sampler
N = 3; % 10^N is the maximum number of samples from Uniform(0,1) RV
u = rand(1,10^N); % generate 1000 samples from Uniform(0,1) RV U

% plot the ECDF from the first 10 samples using the function ECDF
for i=1:N
    SampleSize=10^i;
    subplot(1,N,i)
    % Get the x and y coordinates of SampleSize-based ECDF in x1 and y1 and
    % plot the ECDF using the function ECDF
    if (i==1) [x1 y1] = ECDF(u(1:SampleSize),2,0.2,0.2);
    else
        [x1 y1] = ECDF(u(1:SampleSize),0,0.1,0.1);
        stairs(x1,y1,'k');
    end
    % Note PlotFlag is 1 and the plot range of x-axis is
    % incremented by 0.1 or 0.2 on either side due to last 2 parameters to ECDF
    % being 0.1 or 0.2
    Alpha=0.05; % set alpha to 5% for instance
    Epsn = sqrt((1/(2*SampleSize))*log(2/Alpha)); % epsilon_n for the confidence band
    hold on;
    stairs(x1,max(y1-Epsn,zeros(1,length(y1))), 'g'); % lower band plot
    stairs(x1,min(y1+Epsn,ones(1,length(y1))), 'g'); % upper band plot
    axis([-0.1 1.1 -0.1 1.1]);
    axis square;
    x=[0:0.001:1];
    plot(x,x,'r'); % plot the DF of Uniform(0,1) RV in red
    LabelString=['n=' num2str(SampleSize)];
    text(0.75,0.05,LabelString)
    hold off;
end
```

---

Next we look at a more interesting example involving real-world data.

**Labwork 226 (Non-parametric Estimation of the DF of Times Between Earth Quakes)** Suppose that the 6,128 observed times between Earth quakes in NZ between 18-Jan-2008 02:23:44 and 18-Aug-2008 19:29:29 are:

$$X_1, \dots, X_{6128} \stackrel{IID}{\sim} F^*, \quad F^* \in \{\text{all DFs}\} .$$

Then the non-parametric point estimate of the unknown  $F^*$  is  $\hat{F}_{6128}$ , the ECDF of the inter earth quake times. We plot the non-parametric point estimate as well as the 95% confidence bands for  $F^*$  in Figure 7.14.

Figure 7.13: The empirical DFs  $\hat{F}_n^{(1)}$  from sample size  $n = 10, 100, 1000$  (black), is the point estimate of the fixed and known DF  $F(x) = x, x \in [0, 1]$  of Uniform(0, 1) RV (red). The 95% confidence band for each  $\hat{F}_n$  are depicted by green lines.

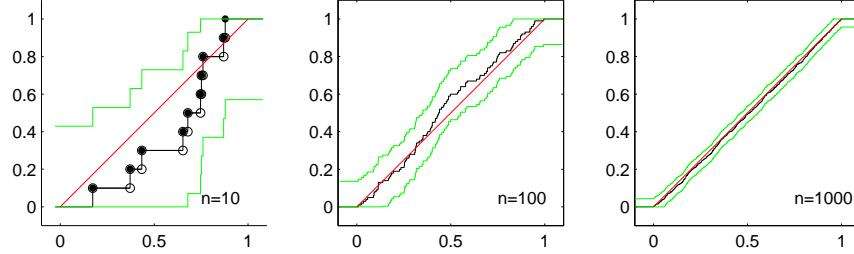
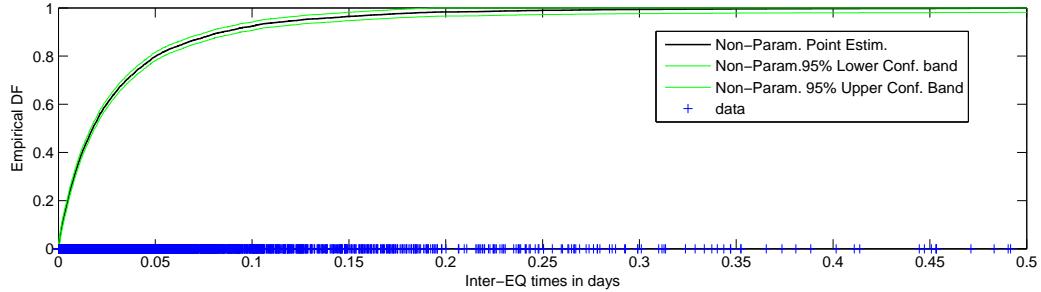


Figure 7.14: The empirical DF  $\hat{F}_{6128}$  for the inter earth quake times and the 95% confidence bands for the non-parametric experiment.




---

```
%>%% The columns in earthquakes.csv file have the following headings
%>%CUSP_ID,LAT,LONG,NZMGE,NZMGN,ORI_YEAR,ORI_MONTH,ORI_DAY,ORI_HOUR,ORI_MINUTE,ORI_SECOND,MAG,DEPTH
EQ=dlmread('earthquakes.csv',','); % load the data in the matrix EQ
size(EQ) % report thr size of the matrix EQ
% Read help datenum -- converts time stamps into numbers in units of days
MaxD=max(datenum(EQ(:,6:11)));% maximum datenum
MinD=min(datenum(EQ(:,6:11)));% minimum datenum
% get an array of sorted time stamps of EQ events starting at 0
Times=sort(datenum(EQ(:,6:11))-MinD);
TimeDiff=diff(Times); % inter-EQ times = times between successtive EQ events
n=length(TimeDiff); %sample size
clf % clear any current figures
%% Non-parametric Estimation X_1,X_2,...,X_132 ~ IID F
[x y] = ECDF(TimeDiff,0,0,0); % get the coordinates for empirical DF
stairs(x,y,'k','linewidth',1) % draw the empirical DF
hold on;
% get the 5% non-parametric confidence bands
Alpha=0.05; % set alpha to 5% for instance
Epsn = sqrt((1/(2*n))*log(2/Alpha)); % epsilon_n for the confidence band
stairs(x,max(y-Epsn,zeros(1,length(y))), 'g'); % non-parametric 95% lower confidence band
stairs(x,min(y+Epsn,ones(1,length(y))), 'g'); % non-parametric 95% upper confidence band
plot(TimeDiff,zeros(1,n),'+')
axis([0 0.5 0 1])
xlabel('Inter-EQ times in days'); ylabel('Empirical DF');
legend('Non-Param. Point Estim.', 'Non-Param. 95% Lower Conf. band', 'Non-Param. 95% Upper Conf. Band', 'data')
```

---

Recall the poor fit of the Exponential PDF at the MLE for the Orbiter waiting time data. We can attribute the poor fit to coarse resolution of the waiting time measurements in minutes and the rigid decaying form of the exponential PDFs. Let us revisit the Orbiter waiting time problem with

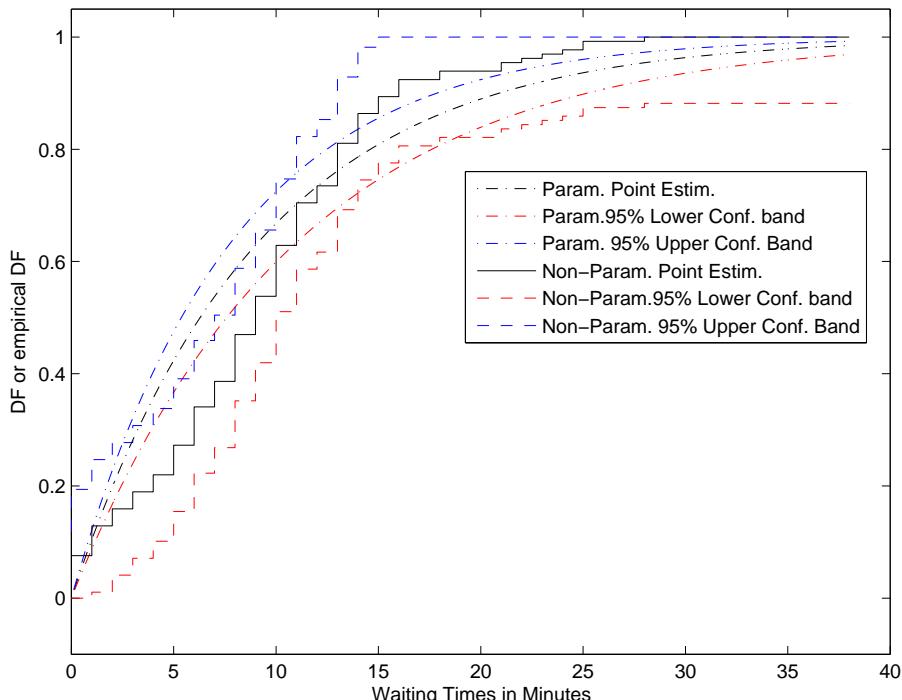
our non-parametric estimator.

**Labwork 227 (Non-parametric Estimation of Orbiter Waiting Times DF)** Suppose that the waiting times at the Orbiter bus stop are:

$$X_1, \dots, X_{132} \stackrel{IID}{\sim} F^*, \quad F^* \in \{\text{all DFs}\} .$$

Then the non-parametric point estimate of  $F^*$  is  $\hat{F}_{132}$ , the ECDF of the 132 Orbiter waiting times. We compute and plot the non-parametric point estimate as well as the 95% confidence bands for

Figure 7.15: The empirical DF  $\hat{F}_{132}$  for the Orbiter waiting times and the 95% confidence bands for the non-parametric experiment.



the unknown DF  $F^*$  beside the parametric estimate and 95% confidence bands from Labwork 221. Clearly, the non-parametric estimate is preferable to the parametric one for this example. Notice how the non-parametric confidence bands do not contain the parametric estimate of the DF.

---

OrbiterECDFsConfBands.m

---

```

OrbiterData; % load the Orbiter Data sampleTimes
clf; % clear any current figures
%% Parametric Estimation X_1,X_2,...,X_132 ~ IID Exponential(lambda)
SampleMean = mean(sampleTimes) % sample mean
MLE = 1/SampleMean % ML estimate is 1/mean
n=length(sampleTimes) % sample size
StdErr=1/(sqrt(n)*SampleMean) % Standard Error
MLE95CI=[MLE-(1.96*StdErr), MLE+(1.96*StdErr)] % 95 % CI
TIMES=[0.00001:0.01:max(sampleTimes)+10]; % points on support
plot(TIMES,ExponentialCdf(TIMES,MLE),'k-'); hold on; % Parametric Point Estimate
plot(TIMES,ExponentialCdf(TIMES,MLE95CI(1)),'r-');% Normal-based Parametric 95% lower C.I.
plot(TIMES,ExponentialCdf(TIMES,MLE95CI(2)),'b-');% Normal-based Parametric 95% upper C.I.
ylabel('DF or empirical DF'); xlabel('Waiting Times in Minutes');

```

---

```

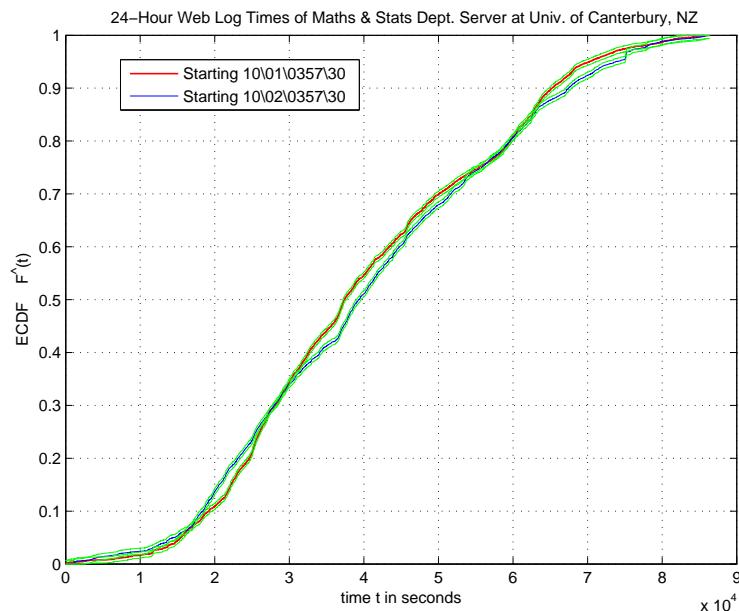
%% Non-parametric Estimation X_1,X_2,...,X_132 ~ IID F
[x1 y1] = ECDF(sampleTimes,0,0.0,10); stairs(x1,y1,'k');% plot the ECDF
% get the 5% non-parametric confidence bands
Alpha=0.05; % set alpha to 5% for instance
Epsn = sqrt((1/(2*n))*log(2/Alpha)); % epsilon_n for the confidence band
stairs(x1,max(y1-Epsn,zeros(1,length(y1))),'r--'); % non-parametric 95% lower confidence band
stairs(x1,min(y1+Epsn,ones(1,length(y1))),'b--'); % non-parametric 95% upper confidence band
axis([0 40 -0.1 1.05]);
legend('Param. Point Estim.', 'Param.95% Lower Conf. band', 'Param. 95% Upper Conf. Band',...
'Non-Param. Point Estim.', 'Non-Param.95% Lower Conf. band', 'Non-Param. 95% Upper Conf. Band')

```

---

**Example 228** First take a look at Data 241 to understand how the web login times to our Maths & Stats Department's web server (or requests to our WWW server) were generated. Figure 7.16 shows the login times in units of seconds over a 24 hour period starting at 0357 hours and 30 seconds (just before 4:00AM) on October 1st, 2007 (red line) and on October 2nd, 2007 (magenta). If we assume

Figure 7.16: The empirical DFs  $\hat{F}_{n_1}^{(1)}$  with  $n_1 = 56485$ , for the web log times starting October 1, and  $\hat{F}_{n_2}^{(2)}$  with  $n_2 = 53966$ , for the web log times starting October 2. Their 95% confidence bands are indicated by the green.



that some fixed and unknown DF  $F^{(1)}$  specifies the distribution of login times for October 1st data and another DF  $F^{(2)}$  for October 2nd data, then the non-parametric point estimates of  $F^{(1)}$  and  $F^{(2)}$  are simply the empirical DFs  $\hat{F}_{n_1}^{(1)}$  with  $n_1 = 56485$  and  $\hat{F}_{n_2}^{(2)}$  with  $n_2 = 53966$ , respectively, as depicted in Figure 7.16. See the script of `WebLogDataProc.m` in Data 241 to appreciate how the ECDF plots in Figure 7.16 were made.

## 7.14 Plug-in Estimators of Statistical Functionals

Recall from Chapter 3.14 that a **statistical functional** is simply any function of the DF  $F$ . For example, the median  $T(F) = F^{[-1]}(1/2)$  is a statistical functional. 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. The idea behind the

plug-in estimator for a statistical functional is simple: just plug-in the point estimate  $\hat{F}_n$  instead of the unknown DF  $F^*$  to estimate the statistical functional of interest.

**Definition 143 (Plug-in estimator)** Suppose,  $X_1, \dots, X_n \stackrel{IID}{\sim} F^*$ . The plug-in estimator of a statistical functional of interest, namely,  $T(F^*)$ , is defined by:

$$\hat{T}_n := \hat{T}_n(X_1, \dots, X_n) = T(\hat{F}_n) .$$

**Definition 144 (Linear functional)** If  $T(F) = \int r(x)dF(x)$  for some function  $r(x) : \mathbb{X} \rightarrow \mathbb{R}$ , then  $T$  is called a **linear functional**. Thus,  $T$  is linear in its arguments:

$$T(aF + a'F') = aT(F) + a'T(F') .$$

**Proposition 145 (Plug-in Estimator of a linear functional)** The plug-in estimator for a linear functional  $T = \int r(x)dF(x)$  is:

$$T(\hat{F}_n) = \int r(x)d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n r(X_i) .$$

Some specific examples of statistical linear functionals we have already seen include:

1. The **mean** of RV  $X \sim F$  is a function of the DF  $F$ :

$$T(F) = \mathbb{E}(X) = \int x dF(x) .$$

2. The **variance** of RV  $X \sim F$  is a function of the DF  $F$ :

$$T(F) = \mathbb{E}(X - \mathbb{E}(X))^2 = \int (x - \mathbb{E}(X))^2 dF(x) .$$

3. The **value of DF at a given**  $x \in \mathbb{R}$  of RV  $X \sim F$  is also a function of DF  $F$ :

$$T(F) = F(x) .$$

4. The  $q^{\text{th}}$  **quantile** of RV  $X \sim F$ :

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

5. The **first quartile** or the  $0.25^{\text{th}}$  **quantile** of the RV  $X \sim F$ :

$$T(F) = F^{[-1]}(0.25) .$$

6. The **median** or the **second quartile** or the  $0.50^{\text{th}}$  **quantile** of the RV  $X \sim F$ :

$$T(F) = F^{[-1]}(0.50) .$$

7. The **third quartile** or the  $0.75^{\text{th}}$  **quantile** of the RV  $X \sim F$ :

$$T(F) = F^{[-1]}(0.75) .$$

**Labwork 229 (Plug-in Estimate for Median of Web Login Data)** Compute the plug-in estimates for the median for each of the data arrays:

WebLogSeconds20071001035730 and WebLogSeconds20071002035730

that can be loaded into memory by following the commands in the first 13 lines of the script file `WebLogDataProc.m` of Data 241.

**Labwork 230 (Plug-in Estimates of Times Between Earth Quakes)** Compute the plug-in estimates for the median and mean time in minutes between earth quakes in NZ using the data in `earthquakes.csv`.

---

```
%>> NZSIEQTimesPlugInEstimates.m
%% The columns in earthquakes.csv file have the following headings
%%CUSP_ID,LAT,LONG,NZMGE,NZMGN,ORI_YEAR,ORI_MONTH,ORI_DAY,ORI_HOUR,ORI_MINUTE,ORI_SECOND,MAG,DEPTH
EQ=dlmread('earthquakes.csv',','); % load the data in the matrix EQ
% Read help datenum -- converts time stamps into numbers in units of days
MaxD=max(datenum(EQ(:,6:11)));% maximum datenum
MinD=min(datenum(EQ(:,6:11)));% minimum datenum
% get an array of sorted time stamps of EQ events starting at 0
Times=sort(datenum(EQ(:,6:11))-MinD);
TimeDiff=diff(Times); % inter-EQ times = times between successive EQ events
n=length(TimeDiff); %sample size
PlugInMedianEstimate=median(TimeDiff) % plug-in estimate of median
PlugInMedianEstimateMinutes=PlugInMedianEstimate*24*60 % median estimate in minutes
PlugInMeanEstimate=mean(TimeDiff) % plug-in estimate of mean
PlugInMeanEstimateMinutes=PlugInMeanEstimate*24*60 % mean estimate in minutes
```

---

```
>> NZSIEQTimesPlugInEstimates
PlugInMedianEstimate =    0.0177
PlugInMedianEstimateMinutes =   25.5092
PlugInMeanEstimate =    0.0349
PlugInMeanEstimateMinutes =   50.2278
```

Note that any statistical functional can be estimated using the plug-in estimator. However, to produce a  $1 - \alpha$  confidence set for the plug-in point estimate, we need bootstrap methods. The subject of next chapter.

## 7.15 Bootstrap

The **bootstrap** is a statistical method for estimating standard errors and confidence sets of statistics, such as estimators.

### 7.15.1 Non-parametric Bootstrap for Confidence Sets

Let  $T_n := T_n((X_1, X_2, \dots, X_n))$  be a statistic, i.e. any function of the data  $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} F^*$ . Suppose we want to know its variance  $V_{F^*}(T_n)$ , which clearly depends on the fixed and possibly unknown DF  $F^*$ .

If our statistic  $T_n$  is one with an analytically unknown variance, then we can use the bootstrap to estimate it. The bootstrap idea has the following two basic steps:

Step 1: Estimate  $V_{F^*}(T_n)$  with  $V_{\hat{F}_n}(T_n)$ .

Step 2: Approximate  $V_{\hat{F}_n}(T_n)$  using simulated data from the “Bootstrap World.”

For example, if  $T_n = \bar{X}_n$ , in Step 1,  $V_{\hat{F}_n}(T_n) = s_n^2/n$ , where  $s_n^2 = n^{-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$  is the sample variance and  $\bar{x}_n$  is the sample mean. In this case, Step 1 is enough. However, when the statistic  $T_n$  is more complicated (e.g.  $T_n = \tilde{X}_n = F^{[-1]}(0.5)$ ), the sample median, then we may not be able to find a simple expression for  $V_{\hat{F}_n}(T_n)$  and may need Step 2 of the bootstrap.

$$\begin{aligned} \text{Real World Data come from } & F^* \implies X_1, X_2, \dots, X_n \implies T_n((X_1, X_2, \dots, X_n)) = t_n \\ \text{Bootstrap World Data come from } & \hat{F}_n \implies X_1^\bullet, X_2^\bullet, \dots, X_n^\bullet \implies T_n((X_1^\bullet, X_2^\bullet, \dots, X_n^\bullet)) = t_n^\bullet \end{aligned}$$

Observe that drawing an observation from the ECDF  $\hat{F}_n$  is equivalent to drawing one point at random from the original data (think of the indices  $[n] := \{1, 2, \dots, n\}$  of the original data  $X_1, X_2, \dots, X_n$  being drawn according to the equi-probable de Moivre( $1/n, 1/n, \dots, 1/n$ ) RV on  $[n]$ ). Thus, to simulate  $X_1^\bullet, X_2^\bullet, \dots, X_n^\bullet$  from  $\hat{F}_n$ , it is enough to draw  $n$  observations with replacement from  $X_1, X_2, \dots, X_n$ .

In summary, the algorithm for Bootstrap Variance Estimation is:

Step 1: Draw  $X_1^\bullet, X_2^\bullet, \dots, X_n^\bullet \sim \hat{F}_n$

Step 2: Compute  $t_n^\bullet = T_n((X_1^\bullet, X_2^\bullet, \dots, X_n^\bullet))$

Step 3: Repeat Step 1 and Step 2  $B$  times, for some large  $B$ , say  $B > 1000$ , to get  $t_{n,1}^\bullet, t_{n,2}^\bullet, \dots, t_{n,B}^\bullet$

Step 4: Several ways of estimating the bootstrap confidence intervals are possible:

(a) The  $1 - \alpha$  Normal-based bootstrap confidence interval is:

$$C_n = [T_n - z_{\alpha/2} \hat{s}e_{\text{boot}}, T_n + z_{\alpha/2} \hat{s}e_{\text{boot}}],$$

where the bootstrap-based standard error estimate is:

$$\hat{s}e_{\text{boot}} = \sqrt{v_{\text{boot}}} = \sqrt{\frac{1}{B} \sum_{b=1}^B \left( t_{n,b}^\bullet - \frac{1}{B} \sum_{r=1}^B t_{n,r}^\bullet \right)^2}$$

- (b) The  $1 - \alpha$  percentile-based bootstrap confidence interval is:

$$C_n = [\widehat{G}^{\bullet}_n^{-1}(\alpha/2), \widehat{G}^{\bullet}_n^{-1}(1 - \alpha/2)],$$

where  $\widehat{G}^{\bullet}_n$  is the empirical DF of the bootstrapped  $t_{n,1}^{\bullet}, t_{n,2}^{\bullet}, \dots, t_{n,B}^{\bullet}$  and  $\widehat{G}^{\bullet}_n^{-1}(q)$  is the  $q^{\text{th}}$  sample quantile (3.83) of  $t_{n,1}^{\bullet}, t_{n,2}^{\bullet}, \dots, t_{n,B}^{\bullet}$ .

**Labwork 231 (Confidence Interval for Median Estimate of Inter Earth Quake Times)**  
 Let us find the 95% Normal-based bootstrap confidence interval as well as the 95% percentile-based bootstrap confidence interval for our plug-in estimate of the median of inter earth quake times from Labwork 230 using the following script:

---

```
NZSIEQTimesMedianBootstrap.m
%% The columns in earthquakes.csv file have the following headings
%%CUSP_ID,LAT,LONG,NZMGE,NZMGN,ORI_YEAR,ORI_MONTH,ORI_DAY,ORI_HOUR,ORI_MINUTE,ORI_SECOND,MAG,DEPTH
EQ=dlmread('earthquakes.csv',','); % load the data in the matrix EQ
% Read help datenum -- converts time stamps into numbers in units of days
MaxD=max(datenum(EQ(:,6:11)));% maximum datenum
MinD=min(datenum(EQ(:,6:11)));% minimum datenum
% get an array of sorted time stamps of EQ events starting at 0
Times=sort(datenum(EQ(:,6:11))-MinD);
TimeDiff=diff(Times); % inter-EQ times = times between successive EQ events
n=length(TimeDiff) %sample size
Medianhat=median(TimeDiff)*24*60 % plug-in estimate of median in minutes
B= 1000 % Number of Bootstrap replications
% REPEAT B times: PROCEDURE of sampling n indices uniformly from 1,...,n with replacement
BootstrappedDataSet = TimeDiff([ceil(n*rand(n,B))]);
size(BootstrappedDataSet) % dimension of the BootstrappedDataSet
BootstrappedMedians=median(BootstrappedDataSet)*24*60; % get the statistic in Bootstrap world
% 95% Normal based Confidence Interval
SehatBoot = std(BootstrappedMedians); % std of BootstrappedMedians
% 95% C.I. for median from Normal approximation
ConfInt95BootNormal = [Medianhat-1.96*SehatBoot, Medianhat+1.96*SehatBoot]
% 95% Percentile based Confidence Interval
ConfInt95BootPercentile = ...
[qthSampleQuantile(0.025,sort(BootstrappedMedians)),...
qthSampleQuantile(0.975,sort(BootstrappedMedians))]
```

---

We get the following output when we call the script file.

```
>> NZSIEQTimesMedianBootstrap
n =       6127
Medianhat =   25.5092
B =       1000
ans =    6127      1000
ConfInt95BootNormal =  24.4383  26.5800
ConfInt95BootPercentile =  24.4057  26.4742
```

**Labwork 232 (Confidence Interval for Median Estimate of Web Login Data)** Find the 95% Normal-based bootstrap confidence interval as well as the 95% percentile-based bootstrap confidence interval for our plug-in estimate of the median for each of the data arrays:

WebLogSeconds20071001035730 and WebLogSeconds20071002035730 .

Once again, the arrays can be loaded into memory by following the commands in the first 13 lines of the script file `WebLogDataProc.m` of Section 241. Produce four intervals (two for each data-set). Do the confidence intervals for the medians for the two days intersect?

```

>> WebLogDataProc % load in the data
>> Medianhat = median(WebLogSeconds20071001035730) % plug-in estimate of median
Medianhat =
      37416
>> % store the length of data array
>> K=length(WebLogSeconds20071001035730)
K =
      56485
>> B= 1000 % Number of Bootstrap replications
B =
      1000
>> BootstrappedDataSet = WebLogSeconds20071001035730([ceil(K*rand(K,B))]);
>> size(BootstrappedDataSet) % dimension of the BootstrappedDataSet
ans =
      56485      1000
>> BootstrappedMedians=median(BootstrappedDataSet); % get the statistic in Bootstrap world
>> % 95% Normal based Confidence Interval
>> SehatBoot = std(BootstrappedMedians); % std of BootstrappedMedians
>> % 95% C.I. for median from Normal approximation
>> ConfInt95BootNormal = [Medianhat-1.96*SehatBoot, Medianhat+1.96*SehatBoot]
ConfInt95BootNormal =
      37242      37590
>> % 95% Percentile based Confidence Interval
ConfInt95BootPercentile = ...
[qthSampleQuantile(0.025,sort(BootstrappedMedians)),...
 qthSampleQuantile(0.975,sort(BootstrappedMedians))]
ConfInt95BootPercentile =
      37239      37554

```

**Labwork 233 (Confidence interval for correlation)** Here is a classical data set used by Bradley Efron (the inventor of bootstrap) to illustrate the method. The data are LSAT (Law School Admission Test in the U.S.A.) scores and GPA of fifteen individuals.

Thus, we have bivariate data of the form  $(Y_i, Z_i)$ , where  $Y_i = \text{LSAT}_i$  and  $Z_i = \text{GPA}_i$ . For example, the first individual had an LSAT score of  $y_1 = 576$  and a GPA of  $z_1 = 3.39$  while the fifteenth individual had an LSAT score of  $y_{15} = 594$  and a GPA of  $z_{15} = 3.96$ . We suppose that the bivariate data  $(Y_i, Z_i) \stackrel{IID}{\sim} F^*$ , such that  $F^* \in \{\text{all bivariate DFs}\}$ . This is a bivariate non-parametric experiment. The bivariate data are plotted in Figure .

The law school is interested in the correlation between the GPA and LSAT scores:

$$\theta^* = \frac{\int \int (y - \mathbb{E}(Y))(z - \mathbb{E}(Z)) dF(y, z)}{\sqrt{\int (y - \mathbb{E}(Y))^2 dF(y) \int (z - \mathbb{E}(Z))^2 dF(z)}}$$

The plug-in estimate of the population correlation  $\theta^*$  is the sample correlation:

$$\widehat{\Theta}_n = \frac{\sum_{i=1}^n (Y_i - \bar{Y}_n)(Z_i - \bar{Z}_n)}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2 \sum_{i=1}^n (Z_i - \bar{Z}_n)^2}}$$

---

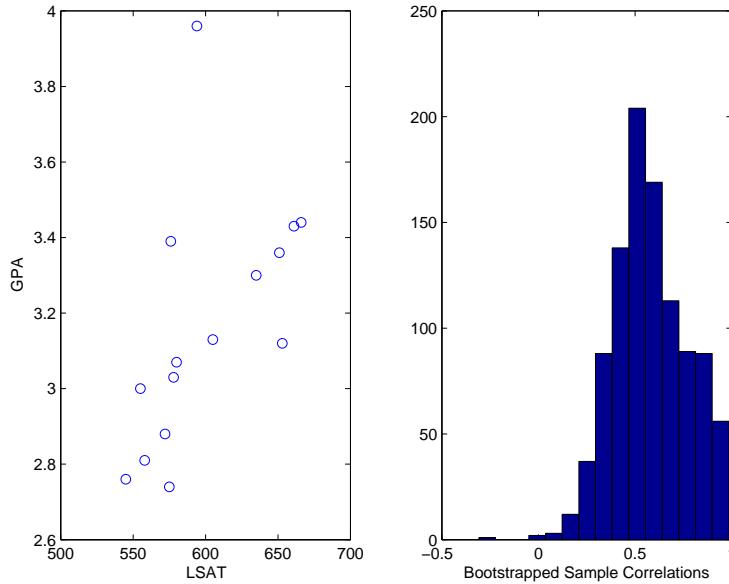
LSATGPACorrBootstrap.m

```

%% Data from Bradley Efron's LSAT,GPA correlation estimation
LSAT=[576 635 558 578 666 580 555 661 651 605 653 575 545 572 594]; % LSAT data
GPA=[3.39 3.30 2.81 3.03 3.44 3.07 3.00 3.43 3.36 3.13 3.12 2.74 2.76 2.88 3.96]; % GPA data
subplot(1,2,1); plot(LSAT,GPA,'o'); xlabel('LSAT'); ylabel('GPA') % make a plot of the data
CC=corrcoef(LSAT,GPA); % use built-in function to compute sample correlation coefficient matrix
SampleCorrelation=CC(1,2) % plug-in estimate of the correlation coefficient
%% Bootstrap
B = 1000; % Number of Bootstrap replications
BootstrappedCCs=zeros(1,B); % initialise a vector of zeros
N = length(LSAT); % sample size
rand('twister',767671); % initialise the fundamental sampler
for b=1:B
    Indices=ceil(N*rand(N,1));% uniformly sample random indices from 1 to 15 with replacement
    BootstrappedLSAT = LSAT([Indices]); % bootstrapped LSAT data
    BootstrappedGPA = GPA([Indices]); % bootstrapped GPA data
end

```

Figure 7.17: Data from Bradley Efron's LSAT and GPA scores for fifteen individuals (left). The confidence interval of the sample correlation, the plug-in estimate of the population correlation, is obtained from the sample correlation of one thousand bootstrapped data sets (right).



```

CCB=corrcoef(BootstrappedLSAT,BootstrappedGPA);
BootstrappedCCs(b)=CCB(1,2); % sample correlation of bootstrapped data
end
%plot the histogram of Bootstrapped Sample Correlations with 15 bins
subplot(1,2,2);hist(BootstrappedCCs,15);xlabel('Bootstrapped Sample Correlations')

% 95% Normal based Confidence Interval
SehatBoot = std(BootstrappedCCs); % std of BootstrappedMedians
% 95% C.I. for median from Normal approximation
ConfInt95BootNormal = [SampleCorrelation-1.96*SehatBoot, SampleCorrelation+1.96*SehatBoot]
% 95% Percentile based Confidence Interval
ConfInt95BootPercentile = ...
    [qthSampleQuantile(0.025,sort(BootstrappedCCs)),...
    qthSampleQuantile(0.975,sort(BootstrappedCCs))]
```

We get the following output when we call the script file.

```

>> LSATGPACorrBootstrap
SampleCorrelation =      0.5459
ConfInt95BootNormal =    0.1770    0.9148
ConfInt95BootPercentile =   0.2346    0.9296
```

### 7.15.2 Parametric Bootstrap for Confidence Sets

The **bootstrap** may also be employed for estimating standard errors and confidence sets of statistics, such as estimators, even in a parametric setting. This is much easier than the the variance calculation based on Fisher Information and/or the Delta method.

The only difference in the **parametric bootstrap** as opposed to the **non-parametric bootstrap** we saw earlier is that our statistic of interest  $T_n := T_n((X_1, X_2, \dots, X_n))$  is a function of the data:

$$X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} F(x; \theta^*) .$$

That is, our data come from a parametric distribution  $F(x; \theta^*)$  and we want to know the variance of our statistic  $T_n$ , i.e.  $V_{\theta^*}(T_n)$ .

The parametric bootstrap concept has the following two basic steps:

**Step 1:** Estimate  $V_{\theta^*}(T_n)$  with  $V_{\hat{\theta}_n}(T_n)$ , where  $\hat{\theta}_n$  is an estimate of  $\theta^*$  based on maximum likelihood or the method of moments.

**Step 2:** Approximate  $V_{\hat{\theta}_n}(T_n)$  using simulated data from the “Bootstrap World.”

For example, if  $T_n = \bar{X}_n$ , the sample mean, then in **Step 1**,  $V_{\hat{\theta}_n}(T_n) = n^{-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$  is the sample variance. Thus, in this case, **Step 1** is enough. However, when the statistic  $T_n$  is more complicated, say  $T_n = \tilde{X}_n = F^{[-1]}(0.5)$ , the sample median, then we may not be able to write down a simple expression for  $V_{\hat{\theta}_n}(T_n)$  and may need **Step 2** of the bootstrap.

$$\begin{array}{lll} \text{Real World Data come from} & F(\theta^*) & \implies X_1, X_2, \dots, X_n \implies T_n((X_1, X_2, \dots, X_n)) = t_n \\ \text{Bootstrap World Data come from} & F(\hat{\theta}_n) & \implies X_1^\bullet, X_2^\bullet, \dots, X_n^\bullet \implies T_n((X_1^\bullet, X_2^\bullet, \dots, X_n^\bullet)) = t_n^\bullet \end{array}$$

To simulate  $X_1^\bullet, X_2^\bullet, \dots, X_n^\bullet$  from  $F(\hat{\theta}_n)$ , we must have a simulation algorithm that allows us to draw IID samples from  $F(\theta)$ , for instance the inversion sampler. In summary, the algorithm for Bootstrap Variance Estimation is:

**Step 1:** Draw  $X_1^\bullet, X_2^\bullet, \dots, X_n^\bullet \sim F(\hat{\theta}_n)$

**Step 2:** Compute  $t_n^\bullet = T_n((X_1^\bullet, X_2^\bullet, \dots, X_n^\bullet))$

**Step 3:** Repeat **Step 1** and **Step 2**  $B$  times, for some large  $B$ , say  $B \geq 1000$ , to get  $t_{n,1}^\bullet, t_{n,2}^\bullet, \dots, t_{n,B}^\bullet$

**Step 4:** We can estimate the bootstrap confidence intervals in several ways:

(a) The  $1 - \alpha$  normal-based bootstrap confidence interval is:

$$C_n = [T_n - z_{\alpha/2} \hat{s}e_{boot}, T_n + z_{\alpha/2} \hat{s}e_{boot}] ,$$

where the bootstrap-based standard error estimate is:

$$\hat{s}e_{boot} = \sqrt{v_{boot}} = \sqrt{\frac{1}{B} \sum_{b=1}^B \left( t_{n,b}^\bullet - \frac{1}{B} \sum_{r=1}^B t_{n,r}^\bullet \right)^2}$$

(b) The  $1 - \alpha$  percentile-based bootstrap confidence interval:

$$C_n = [\widehat{G}_n^{\bullet-1}(\alpha/2), \widehat{G}_n^{\bullet-1}(1 - \alpha/2)],$$

where  $\widehat{G}_n^{\bullet}$  is the empirical DF of the bootstrapped  $t_{n,1}^\bullet, t_{n,2}^\bullet, \dots, t_{n,B}^\bullet$  and  $\widehat{G}_n^{\bullet-1}(q)$  is the  $q^{\text{th}}$  sample quantile (3.83) of  $t_{n,1}^\bullet, t_{n,2}^\bullet, \dots, t_{n,B}^\bullet$ .

Let us apply the bootstrap method to the previous problem of estimating the standard error of the coefficient of variation from  $n = 100$  samples from  $\text{Normal}(100, 10^2)$  RV. The confidence intervals from bootstrap-based methods are similar to those from the Delta method.

---

```
CoeffOfVarNormalBoot.m
```

---

```
n=100; Mustar=100; Sigmastar=10; % sample size, true mean and standard deviation
rand('twister',67345);
x=arrayfun(@(u)(Sample1NormalByNewRap(u,Mustar,Sigmatstar^2)),rand(n,1)); % normal samples
Muhat=mean(x) Sigmahat=std(x) Psihat=Sigmahat/Muhat % MLE of Mustar, Sigmastar and Psistar
Sehat = sqrt((1/Muhat^4)+(Sigmahat^2/(2*Muhat^2))/sqrt(n)) % standard error estimate
% 95% Confidence interval by Delta Method
ConfInt95DeltaMethod=[Psihat-1.96*Sehat, Psihat+1.96*Sehat] % 1.96 since 1-alpha=0.95
B = 1000; % B is number of bootstrap replications
% Step 1: draw n IID samples in Bootstrap World from Normal(Muhat,Sigmahat^2)
xBoot = arrayfun(@(u)(Sample1NormalByNewRap(u,Muhat,Sigmahat^2)),rand(n,B));
% Step 2: % Compute Bootstrapped Statistic Psihat
PsihatBoot = std(xBoot) ./ mean(xBoot);
% 95% Normal based Confidence Interval
SehatBoot = std(PsihatBoot); % std of PsihatBoot
ConfInt95BootNormal = [Psihat-1.96*SehatBoot, Psihat+1.96*SehatBoot] % 1-alpha=0.95
% 95% Percentile based Confidence Interval
ConfInt95BootPercentile = ...
[qthSampleQuantile(0.025,sort(PsihatBoot)),qthSampleQuantile(0.975,sort(PsihatBoot))]
```

---

```
>> CoeffOfVarNormal
Muhat = 100.3117
Sigmahat = 10.9800
Psihat = 0.1095
Sehat = 0.0077
ConfInt95DeltaMethod = 0.0943 0.1246
ConfInt95BootNormal = 0.0943 0.1246
ConfInt95BootPercentile = 0.0946 0.1249
```

## 7.16 Linear Regression

### 7.16.1 Introduction

Regression is a method for studying the relationship between a *response variable*  $Y$  and a *covariate*  $X$ . The covariate is also called a *feature* or a *predictor* variable.

A simple way to summarise the relationship between  $X$  and  $Y$  is through the regression function  $r(x)$ :

$$r(x) = E(Y|X = x) = \int y f(y|x) dy$$

Our objective is to estimate the regression function  $r(x)$  from data of the form:

$$(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n) \stackrel{IID}{\sim} F_{X,Y}$$

We assume that  $F_{X,Y}$ , the joint distribution of  $X$  and  $Y$ , is parametric and  $r$  is linear.

### 7.16.2 Simple Linear Regression

The *simple linear regression model* is when  $X_i$  is real-valued (one-dimensional) and  $r(x)$  is assumed to be linear:

$$r(x) = \beta_0 + \beta_1 x, \quad \text{and} \quad V(Y|X = x) = \sigma^2 \text{ is independent of } x$$

Thus simple linear regression model is the following:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \text{where, } E(\epsilon_i|X_i) = 0 \text{ and } V(\epsilon_i|X_i) = \sigma^2$$

The unknown parameters and their estimates in the model are:

- the intercept  $\beta_0$  and its estimate  $\hat{\beta}_0$ ,
- the slope  $\beta_1$  and its estimate  $\hat{\beta}_1$  and
- the variance  $\sigma^2$  and its estimate  $\hat{\sigma}^2$

The *fitted line* is:

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

The *fitted* or *predicted values* are:

$$\hat{Y}_i = \hat{r}(X_i)$$

The *residuals* are:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

The *residual sum of squares* or *RSS*, that measures how well the line fits the data, is defined by

$$RSS = \sum_{i=1}^n \hat{\epsilon}_i^2$$

The *least squares estimates* are the values  $\hat{\beta}_0$  and  $\hat{\beta}_1$  that minimise *RSS* and they are given by:

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}, \quad \hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n, \quad \hat{\sigma}^2 = \left( \frac{1}{n-2} \right) \sum_{i=1}^n \hat{\epsilon}_i^2} \quad (7.62)$$

Geometric understanding of simple linear regression can be obtained from interactive animations<sup>4</sup>.

### 7.16.3 Least Squares and Maximum Likelihood

Suppose we add the assumption about the model's noise that

$$\boxed{\epsilon_i | X_i \sim \text{Normal}(0, \sigma^2) \quad \text{i.e.,} \quad Y_i | X_i \sim \text{Normal}(\mu_i, \sigma^2), \quad \text{where} \quad \mu_i = \beta_0 + \beta_1 X_i}$$

Then, the likelihood function is:

$$\prod_{i=1}^n f_{X,Y}(X_i, Y_i) = \prod_{i=1}^n f_X(X_i) f_{Y|X}(Y_i | X_i) \quad (7.63)$$

$$= \prod_{i=1}^n f_X(X_i) \prod_{i=1}^n f_{Y|X}(Y_i | X_i) \quad (7.64)$$

$$=: L_{n,X} L_{n,Y|X} \quad (7.65)$$

where,  $L_{n,X} := \prod_{i=1}^n f_X(X_i)$  is the marginal likelihood of  $X_1, \dots, X_n$  that does not depend on the parameters  $(\beta_0, \beta_1, \sigma)$ , and  $L_{n,Y|X} := \prod_{i=1}^n f_{Y|X}(Y_i | X_i)$  is the \*conditional likelihood\* that does depend on the parameters. Therefore the likelihood function is given by the conditional likelihood:

$$L(\beta_0, \beta_1, \sigma) \propto \prod_{i=1}^n f(X_i, Y_i) \quad (7.66)$$

$$\propto L_{n,Y|X} = \prod_{i=1}^n f_{Y|X}(Y_i | X_i) \quad (7.67)$$

$$\propto \sigma^{-n} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \mu_i)^2 \right) \quad (7.68)$$

$$(7.69)$$

and the conditional log-likelihood is:

$$\boxed{l(\beta_0, \beta_1, \sigma) = -n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \mu_i)^2}$$

To find the MLE of  $(\beta_0, \beta_1)$  we need to maximise  $\ell(\beta_0, \beta_1, \sigma)$  for a given  $\sigma$ . From the above expression it is clear that maximising the log-likelihood is equivalent to minimising the *residual sum of squares* or *RSS* given by

---

<sup>4</sup>Explore at <http://setosa.io/ev/ordinary-least-squares-regression/>

$$\boxed{\sum_{i=1}^n (Y_i - \mu_i)^2}$$

Therefore, we have proved the following proposition.

**Proposition 146 (MLE is LSE)** Under the assumption of normally distributed noise in linear regression model, the maximum likelihood estimator (MLE) is the least squares estimator (LSE).

We can maximise  $l(\beta_0, \beta_1, \sigma)$  over  $\sigma$  and obtain the MLE for  $\sigma$  as follows:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^2.$$

But it is more common in practise to use the unbiased estimator, with  $E(\hat{\sigma}^2) = \sigma^2$ , that we saw earlier for sample size  $n > 2$ :

$$\hat{\sigma}^2 = \left( \frac{1}{n-2} \right) \sum_{i=1}^n \hat{\epsilon}_i^2.$$

#### 7.16.4 Properties of the Least Squares Estimator (LSE)

It's finally time to obtain the standard errors and limititng distribution of the least quares estimator (also the MLE).

In regression we are interested in the properties of the estimators conditional on the covariates

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

##### Conditional Mean and Variance of LSE

Let  $\hat{\beta}^T = (\hat{\beta}_0, \hat{\beta}_1)^T$  denote the least squares estimators (which is also the MLE). Then

$$E(\hat{\beta} | X_{1:n}) = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad (7.70)$$

$$V(\hat{\beta} | X_{1:n}) = \frac{\sigma^2}{ns_X^2} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n X_i^2 & -\bar{X}_n \\ -\bar{X}_n & 1 \end{pmatrix} \quad (7.71)$$

where,

$$s_X^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

### Estimated Standard Errors

The estimated standard errors for  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , or more precisely, the estimated standard errors conditional on the covariates, are given by the square-root of the diagonal terms of the variance-covariance matrix  $V(\hat{\beta} | X_{1:n})$  and substituting the estimate  $\hat{\sigma}$  for  $\sigma$ , as follows:

$$\hat{se}(\hat{\beta}_0) := \hat{se}(\hat{\beta}_0 | X_{1:n}) = \frac{\hat{\sigma}}{s_X \sqrt{n}} \sqrt{\frac{\sum_{i=1}^n X_i^2}{n}} \quad (7.72)$$

$$\hat{se}(\hat{\beta}_1) := \hat{se}(\hat{\beta}_1 | X_{1:n}) = \frac{\hat{\sigma}}{s_X \sqrt{n}} \quad (7.73)$$

Thus under appropriate modeling assumptions in simple linear regression we have the following four properties.

### Four Asymptotic Properties of the LSE

1. Asymptotic Consistency: As  $n \rightarrow \infty$ , the LSE, i.e.  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , converges in probability to the parameters, i.e.,  $\beta_0, \beta_1$ , generating the data  $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$  as summarised below.

$$\boxed{\hat{\beta}_0 \xrightarrow{P} \beta_0 \quad \text{and} \quad \hat{\beta}_1 \xrightarrow{P} \beta_1}$$

2. Asymptotic Normality: As  $n \rightarrow \infty$ , the LSE, i.e.  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , converges in distribution to the parameters, i.e.,  $\beta_0, \beta_1$ , generating the data  $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$  as summarised below.

$$\boxed{\frac{\hat{\beta}_0 - \beta_0}{\hat{se}(\hat{\beta}_0)} \xrightarrow{d} \text{Normal}(0, 1) \quad \text{and} \quad \frac{\hat{\beta}_1 - \beta_1}{\hat{se}(\hat{\beta}_1)} \xrightarrow{d} \text{Normal}(0, 1)}$$

3. Approximate  $1 - \alpha$  Confidence Interval: The  $1 - \alpha$  confidence interval for  $\beta_0$  and  $\beta_1$  that is obtained from the approximately normal distribution as  $n$  gets large is:

$$\boxed{\hat{\beta}_0 \pm z_{\alpha/2} \hat{se}(\hat{\beta}_0) \quad \text{and} \quad \hat{\beta}_1 \pm z_{\alpha/2} \hat{se}(\hat{\beta}_1)}$$

4. The Wald Test: Recall Wald test statistic for testing the null hypothesis with the null value  $\beta^{(0)}$ :

$$H_0 : \beta = \beta^{(0)} \quad \text{versus} \quad H_1 : \beta \neq \beta^{(0)} \quad \text{is} \quad W = \frac{(\hat{\beta} - \beta^{(0)})}{\hat{se}(\hat{\beta})}$$

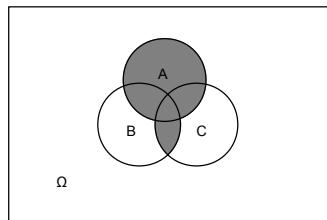
Thus the Wald test for testing  $H_0 : \beta_1 = 0$  versus  $H_1 : \beta_1 \neq 0$  is to reject  $H_0$  if  $|W| > z_{\alpha/2}$  where  $W = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)}$ .

# Answers to Selected Exercises

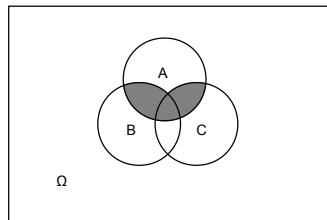
**Answer (Ex. 1.1)** — By operating with  $\Omega$ ,  $T$ ,  $L$  and  $S$  we can obtain the answers as follows:

- |   |  |
|---|--|
| (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:



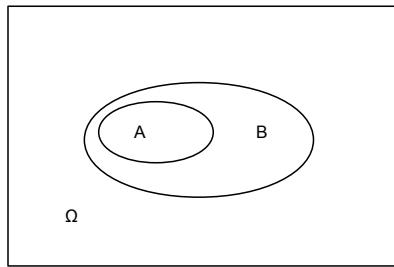
**Answer (Ex. 1.3)** — We can check  $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$  from the following sketch:  
We can check  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$  from the following sketch:



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

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

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

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

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

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

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

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

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

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

(f)  $P(\text{'picking any letter in the word WAZZZUP or a vowel'}) =$

$P(\{W, A, Z, U, P\}) + P(\{A, E, I, O, U\}) - P(\{A, U\}) = 14.4\% + 37.8\% - 10\% = 42.2\%$ , by the addition rule for two arbitrary events, rule (3).

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

2.  $\{\text{RRRR}, \text{RRRL}, \text{RRLR}, \text{RLRR}, \text{LRRR}, \text{RLRL}, \text{RRLR}, \text{LLRR}, \text{LRLR}, \text{LRRL}, \text{RLLL}\}$

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

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

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

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

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

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

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

2. The probabilities of simple events are:

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

3. Using the addition rule for mutually exclusive events,

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

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

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

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

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

**Answer (Ex. 2.6) —** We can assume that the first shot is independent of the second shot so we can multiply the probabilities here.

For case A, there is only one shot so the probability of hitting at least once is  $\frac{1}{2}$ .

For case B, the probability of missing both shots is  $\frac{2}{3} \cdot \frac{2}{3} = \frac{4}{9}$ , so the probability hitting some target at least once is

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

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

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

$$\{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}$$

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

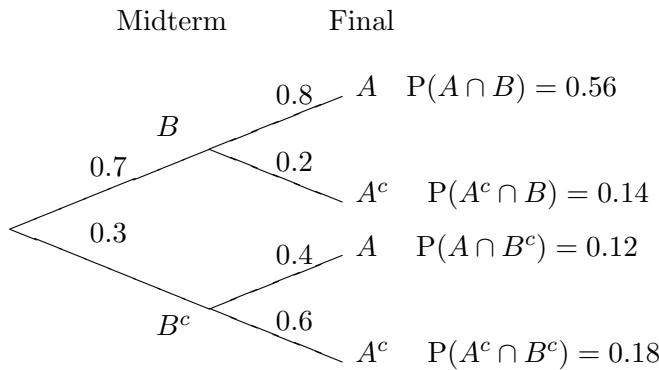
Let  $A$  be the event *the sum is 5* and  $B$  be the event *the sum is 6*, then  $A$  and  $B$  are mutually exclusive events with probabilities

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

Therefore,

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

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



Then the probability of passing the final exam is:

$$P(A) = 0.56 + 0.12 = 0.68.$$

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

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

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

$$\begin{aligned} P(A) &= 0.75 \quad \text{and} \quad P(A^c) = 0.25 \\ P(R|A) &= \frac{1}{20} \quad \text{and} \quad P(R^c|A) = \frac{19}{20} \\ P(R|A^c) &= \frac{1}{30} \quad \text{and} \quad P(R^c|A^c) = \frac{29}{30} \end{aligned}$$

We want  $P(A|R^c)$  which is give by

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

where,

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

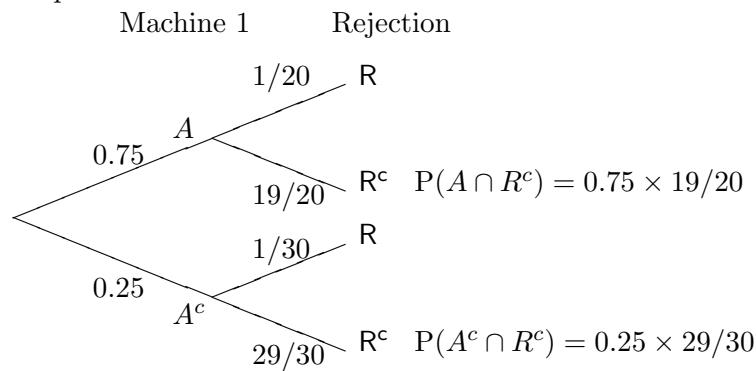
and,

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

Therefore,

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

The tree diagram for this problem is:



So the required probability is

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

**Answer (Ex. 2.10)** — Let the event that a micro-chip is defective be  $D$ , and the event that the test is correct be  $C$ . So the probability that the micro-chip is defective is  $P(D) = 0.05$ , and the probability that it is effective is  $P(D^c) = 0.95$ .

The probability that the test correctly detects a defective micro-chip is the conditional probability  $P(C|D) = 0.8$ , and the probability that if a good micro-chip is tested but the test declares it is defective is the conditional probability  $P(C^c|D^c) = 0.1$ . Therefore, we also have the probabilities  $P(C^c|D) = 0.2$ , and  $P(C|D^c) = 0.9$ .

Moreover, the probability that a micro-chip is defective, and has been declared as defective is

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

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

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

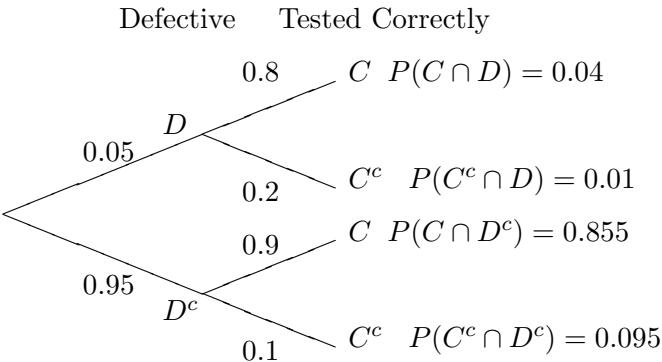
The probability that a micro-chip is defective, and has been declared as effective is

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

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

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

The tree diagram for these events and probabilities is:



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

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

(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. 2.11) —** (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

$$P(D \cap F_1) = P(D|F_1)P(F_1) = \frac{1}{4} \times \frac{2}{3} = \frac{1}{6},$$

$$P(D \cap F_2) = P(D|F_2)P(F_2) = \frac{2}{3} \times \frac{1}{4} = \frac{1}{6},$$

and

$$P(D \cap F_3) = P(D|F_3)P(F_3) = \frac{5}{6} \times \frac{1}{12} = \frac{5}{72}.$$

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:  $P(D \cap F_1) = P(F_1 \cap D)$  etc.):

$$P(F_1|D) = \frac{P(F_1 \cap D)}{P(D)} = \frac{1/6}{29/72} = \frac{12}{29}$$

$$P(F_2|D) = \frac{P(F_2 \cap D)}{P(D)} = \frac{1/6}{29/72} = \frac{12}{29}$$

$$P(F_3|D) = \frac{P(F_3 \cap D)}{P(D)} = \frac{5/72}{29/72} = \frac{5}{29}$$

(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  $P(F_1 \cap D^c)$ . The best way to do this is to use the partitioning idea of the “Total Probability Theorem”, and write:

$$P(F_1) = P(F_1 \cap D^c) + P(F_1 \cap D),$$

Rearranging this gives

$$P(F_1 \cap D^c) = P(F_1) - P(F_1 \cap D)$$

and so

$$P(F_1|D^c) = \frac{P(F_1 \cap D^c)}{P(D^c)} = \frac{P(F_1) - P(F_1 \cap D)}{P(D^c)} = \frac{2/3 - 1/6}{43/72} = \frac{36}{43}.$$

Similarly,

$$P(F_2|D^c) = \frac{P(F_2 \cap D^c)}{P(D^c)} = \frac{P(F_2) - P(F_2 \cap D)}{P(D^c)} = \frac{1/4 - 1/6}{43/72} = \frac{6}{43},$$

and

$$P(F_3|D^c) = \frac{P(F_3 \cap D^c)}{P(D^c)} = \frac{P(F_3) - P(F_3 \cap D)}{P(D^c)} = \frac{1/12 - 5/72}{43/72} = \frac{1}{43}.$$

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

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

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

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

or more simply,

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

The distribution function for  $X$  is:

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

or more simply,

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

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

**Answer (Exercise 3.4)** — We are given that 537 flying bombs hit an area  $A$  of south London made up of  $24 \times 24 = 576$  small equal-sized areas, say  $A_1, A_2, \dots, A_{576}$ . 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  $\theta$  in a sequence of  $n = 537$  independent Bernoulli( $\theta$ ) trials. Finally, we can approximate this  $\text{Binomial}(n = 537, \theta = \frac{1}{576})$  random variable by  $\text{Poisson}(\lambda)$  random variable with  $\lambda = n\theta = \frac{537}{576} \approx 0.933$ . Using the probability mass function formula for  $\text{Poisson}(\lambda = 0.933)$  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	$7/576 = 0.0122$	$f(4; 0.933) = 0.0124$
$\geq 5$	1	$1/576 = 0.00174$	$1 - \sum_{x=0}^4 f(x; 0.933) = 0.00275$

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

$$P(X = 3) = 1 - 0.07 - 0.10 - 0.32 - 0.40 = 0.11 .$$

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

$x$	1	2	3	4	5
$P(X = x)$	0.1	0.2	0.2	0.2	0.3

so the distribution function is:

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

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

$$P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0.1 + 0.2 + 0.2 = 0.5 .$$

(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

$$P(X > 3) = 1 - P(X \leq 3) = 0.5 .$$

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

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

**Answer (Ex. 3.8) — (a)**

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

$$(b) \quad (i) P(X \leq 5) = F(5) = 0.17$$

$$(ii) P(X < 12) = P(X \leq 11) = F(11) = 0.84$$

$$(iii) P(X > 9) = 1 - P(X \leq 9) = 1 - F(9) = 1 - 0.79 = 0.21$$

$$(iv) P(X \geq 9) = 1 - P(X < 9) = 1 - P(X \leq 8) = 1 - 0.59 = 0.41$$

$$(v) P(4 < X \leq 9) = F(9) - F(4) = 0.79 - 0.08 = 0.71$$

$$(vi) P(4 < X < 11) = P(4 < X \leq 10) = F(10) - F(4) = 0.82 - 0.08 = 0.74$$

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

$$P(X = 0) = \frac{4}{10} \cdot \frac{3}{9} = \frac{2}{15} \quad (\text{one way of drawing two right screws}),$$

$$P(X = 1) = \frac{6}{10} \cdot \frac{4}{9} + \frac{4}{10} \cdot \frac{6}{9} = \frac{8}{15} \quad (\text{two ways of drawing one left and one right screw}),$$

$$P(X = 2) = \frac{6}{10} \cdot \frac{5}{9} = \frac{1}{3} \quad (\text{one way of drawing two left screws}).$$

So the probability mass function of  $X$  is:

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

The required probabilities are:

1.

$$P(X \leq 1) = P(X = 0) + P(X = 1) = \frac{2}{15} + \frac{8}{15} = \frac{2}{3}$$

2.

$$P(X \geq 1) = P(X = 1) + P(X = 2) = \frac{8}{15} + \frac{1}{3} = \frac{13}{15}$$

3.

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

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

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

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,

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

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

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

where

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

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

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

1.  $X$  has probability mass function

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

2. The required probabilities are:

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

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

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

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

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

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

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

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

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

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

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

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

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) ,$$

where

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

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

$$1 - 0.3487 \approx 0.6513 .$$

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

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

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

$$P(X \geq 2) = 1 - P(X < 2) = 1 - P(X = 1) - P(X = 0) ,$$

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

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

Now

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

and

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

and so

$$P(X \geq 2) = 1 - 0.9098 = 0.0902 .$$

**Answer (Ex. 3.16)** — 1.The Probability mass function for Poisson( $\lambda$ ) random variable  $X$  is

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

where  $\lambda$  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  $P(X = 0) = e^{-\lambda}$ .

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

$$P(X = 0) = 1 - P(X \geq 1) = 0.9 .$$

Using (b) and solving for  $\lambda$  gives:

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

Hence

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

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

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

**Answer (Exercise 3.18)** —

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

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

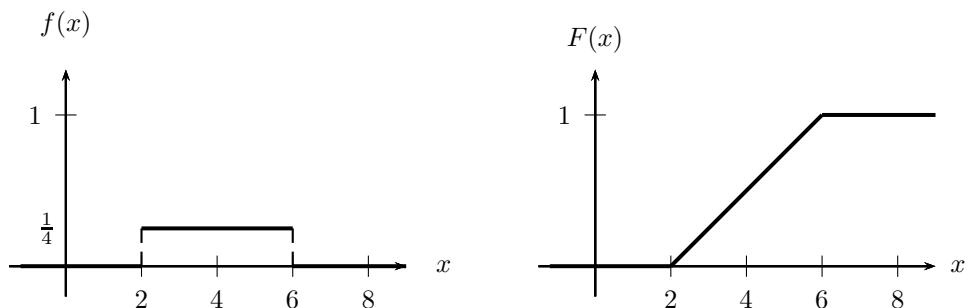
as expected!

(b) Now

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

so the graphs are:

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



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

$$\int_{-4}^4 kdx = 1 .$$

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

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

2. Set

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

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

**Answer (Ex. 3.21)** — 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.

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

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

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

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

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

as follows:

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

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

$$g : \{1, 2, 3, \dots\} \rightarrow \{2^{-1}, 2^{-2}, 2^{-3}, \dots\}$$

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

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

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

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

**Answer (Ex. 3.25)** — Since  $X$  is a  $\text{Poisson}(\lambda)$  random variable (by suppressing the ‘;  $\lambda$ ’ in the argument to  $f_X(\cdot)$  for notational ease), we get

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

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

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

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

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

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

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

**Answer (Ex. 3.26)** — Since  $y = g(x) = e^x$  is a monotone increasing function for  $x \geq 0$ , we can apply the change of variable formula.

Now  $x = g^{-1}(y) = \log_e(y)$  is a monotone increasing function for  $y$  in  $[1, \infty)$

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

Therefore

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

since  $e^{-\log_e(y)} = e^{\log_e(y^{-1})} = y^{-1}$ .

So the probability density function of  $Y$  is given by

$$f_Y(y) = \begin{cases} \frac{\log_e(y)}{y^2} & \text{if } x \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

**Answer (Ex. 3.27)** — Since  $y = g(x) = \sqrt{x}$  is a monotone increasing function for  $x \geq 0$ , we can apply the change of variable formula.

Now  $x = g^{-1}(y) = y^2$  is a monotone increasing function for  $y \geq 0$  so on this interval

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

Therefore

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

So the probability density function of  $Y$  is given by

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

**Answer (Ex. 3.28)** — First note that  $y = g(x) = \log_e(x)$  is a monotone increasing function over  $a \leq x \leq b$ , so we can apply the change of variable formula.

$x = g^{-1}(y) = e^y$  is a monotone increasing function over  $\log_e(a) \leq \log_e(x) \leq \log_e(b)$ , that is, over  $\log_e(a) \leq y \leq \log_e(b)$ .

For  $\log_e(a) \leq y \leq \log_e(b)$ ,

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

Therefore

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

So the probability density function of  $Y$  is given by

$$f_Y(y) = \begin{cases} \frac{e^y}{b-a} & \log_e(a) \leq y \leq \log_e(b) \\ 0 & \text{otherwise} \end{cases}.$$

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

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

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

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

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

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

is \$643.50.

**Answer (Ex. 3.31)** — The expected value  $E(X)$  is

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

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

the variance is  $V(X) = E(X^2) - (E(X))^2 = 0.3 - 0.5^2 = 0.05$

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

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

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

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

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

and so

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

Now,

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

and so the variance is

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

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

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

Therefore,

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

and

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

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

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

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

**Answer (Exercise 3.36) —**

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

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

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

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

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

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

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

**Answer (Ex. 3.39)** —

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.

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

7.

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

8.

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

9.

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

10.

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

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

If  $y \geq 0$ ,

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

Differentiating this expression gives

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

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

**Answer (Ex. 3.41) —**

First derive the marginal PMF of  $X$  and  $Y$  and then check if the JPMF is the product of the marginal PMFs.

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

and

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

Thus,

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

Similarly,

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

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

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

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

**Answer (Ex. 3.42) —**

Let  $X_1, X_2, X_3$  be independent RVs that denote the thickness of the first, second and third layer, respectively. Let  $X$  denote the thickness of the final product. Then

$$X = X_1 + X_2 + X_3$$

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

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

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

**Answer (Ex. 3.43) —**

Find  $E(XY)$ ,  $E(X)$  and  $E(Y)$  to get  $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$  as follows:

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

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

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

Finally,

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

### Answer (Ex. 3.44) —

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

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

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

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

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

### Answer (Ex. 3.45) —

First obtain marginal PDF of  $Y$ . If  $y \in [2, 3]$  then

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

Therefore,

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

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

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

Therefore,

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

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

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

Thus,  $X$  and  $Y$  are independent.

**Answer (Ex. 3.46) —**

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

Since

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

the above quadruply iterated integral becomes

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

**Answer (Ex. 3.47) —**

Due to independence of  $Y_1$ ,  $Y_2$  and  $Y_3$

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

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

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

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

**Answer (Ex. 3.48) —**

Let  $X_1, X_2, \dots, X_{10}$  denote the fill volumes of 10 cans. The average fill volume is the sample mean

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

By property of Expectations and Variances for linear combinations

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

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

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

Or by directly using the “formula”  $V(\bar{X}_{10}) = V(X_1)/10$  for these 10 independently and identically distributed RVs.

By the special property of Normal RVs – a linear combination of independent normal RVs is also normal – we know that  $\bar{X}_{10}$  is a  $\text{Normal}(12.1, 0.001)$  RV. Consequently, the probability of interest is

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

**Answer (Ex. 3.49) —**

Using the Multinomial( $n = 10, \theta_1 = 0.6, \theta_2 = 0.3, \theta_3 = 0.08, \theta_4 = 0.02$ ) RV  $\vec{V}$  as our model

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

**Answer (Ex. 3.50) —**

1. The CF of the discrete RV  $X$  is

$$\begin{aligned} \varphi_X(t) = E(e^{itX}) &= \sum_{x \in \{0, 1, 2\}} e^{itx} f_X(x) = e^{it \times 0} \times \frac{1}{3} + e^{it \times 1} \times \frac{1}{3} + e^{it \times 2} \times \frac{1}{3} \\ &= \frac{1}{3} (1 + e^{it} + e^{i2t}) . \end{aligned}$$

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

$$\frac{d}{dt}\varphi_X(t) = \frac{d}{dt} \left( \frac{1}{3} (1 + e^{it} + e^{i2t}) \right) = \frac{1}{3} (0 + ie^{it} + 2ie^{i2t}) = \frac{i}{3} (e^{it} + 2e^{i2t})$$

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

$$E(X) = \frac{1}{i} \left[ \frac{d}{dt} \varphi_X(t) \right]_{t=0} = \frac{1}{i} \left[ \frac{i}{3} (e^{it} + 2e^{i2t}) \right]_{t=0} = \frac{1}{i} \frac{i}{3} (e^0 + 2e^0) = \frac{1}{3} (1 + 2) = 1 .$$

$$\begin{aligned} E(X^2) &= \frac{1}{i^2} \left[ \frac{d^2}{dt^2} \varphi_X(t) \right]_{t=0} = \frac{1}{i^2} \left[ \frac{d}{dt} \frac{i}{3} (e^{it} + 2e^{i2t}) \right]_{t=0} = \frac{1}{i^2} \left[ \frac{i}{3} (ie^{it} + 4ie^{i2t}) \right]_{t=0} \\ &= \frac{1}{3} (e^0 + 4e^0) = \frac{5}{3} . \end{aligned}$$

Finally,

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

### Answer (Ex. 3.51) —

1.

$$\varphi_X(t) = E(e^{itX}) = \sum_{x=0}^{\infty} e^{itx} \theta(1-\theta)^x = \sum_{x=0}^{\infty} \theta (e^{it}(1-\theta))^x = \frac{\theta}{1 - e^{it}(1-\theta)} .$$

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

2.

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

### Answer (Ex. 3.52) —

$$\begin{aligned} \varphi_X(t) &= E(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} f_X(x; a, b) dx = \int_a^b e^{itx} \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b e^{itx} dx = \frac{1}{b-a} \left[ \frac{e^{itx}}{it} \right]_{x=a}^b \\ &= \frac{1}{(b-a)it} (e^{itb} - e^{ita}) . \end{aligned}$$

**Answer (Ex. 3.53) —**

1.

$$\varphi_X(t) = E(e^{itX}) = \sum_{x=0}^{\infty} e^{itx} \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{e^{itx} \lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^{it})^x}{x!} = e^{-\lambda} e^{\lambda e^{it}} = e^{\lambda e^{it} - \lambda} .$$

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

2. To find  $V(X)$  using  $\varphi_X(t)$  we need  $E(X)$  and  $E(X^2)$ .

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

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

Finally,

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

**Answer (Ex. 3.54) —**

$$\varphi_W(t) = \varphi_{X+Y}(t) = \varphi_X(t) \times \varphi_Y(t) = e^{\lambda e^{it} - \lambda} \times e^{\mu e^{it} - \mu} = e^{\lambda e^{it} - \lambda + \mu e^{it} - \mu} = e^{(\lambda + \mu)e^{it} - (\lambda + \mu)} .$$

So,  $W$  is a Poisson( $\lambda + \mu$ ) RV. Thus the sum of two independent Poisson RVs is also a Poisson RV with parameter given by the sum of the parameters of the two RVs being added. The same idea generalizes to the sum of more than two Poisson RVs.

**Answer (Ex. 3.55) —**

We can use the facts by noting  $Y = -Z = 0 + (-1) \times Z$ , with  $a = 0$  and  $b = -1$  in  $Y = a + bX$  and get

$$\varphi_Y(t) = e^{i \times 0 \times t} \varphi_Z(-1 \times t) = \varphi_Z(-t) = e^{-((-1 \times t))^2 / 2} = e^{-t^2 / 2} = \varphi_Z(t)$$

Thus,  $\varphi_{-Z}(t) = \varphi_Z(t)$  and therefore the distributions of  $Z$  and  $-Z$  are the same. This should make sense because by switching signs of a symmetric (about 0) RV you have not changed its distribution! Note: we are not saying  $Z = -Z$  but just that their distributions are the same, i.e.,  $F_Z(z) = F_{-Z}(z)$  for every  $z \in \mathbb{R}$ .

**Answer (Ex. 3.56) —** Apply the formulas for the sample mean,  $\bar{X}_7$ , and sample variance.

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

**Answer (Ex. 5.2) —**

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

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

as follows:

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

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

**Answer (Ex. 5.3) —**

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

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

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

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

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

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

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

**Answer (Exercise 7.6) —**

(a) The number of questions out of the first five that the student gets correct has a binomial distribution with parameters  $n = 5$  and  $p = 0.2$ . Therefore, the probability that the student gets exactly two of the questions correct is

$$\binom{5}{2} (0.2)^2 (0.8)^{5-2} = \frac{5!}{2! \cdot 3!} (0.2)^2 (0.8)^{5-2} = 0.2048.$$

(b) Let  $X$  be the number of questions that the student answers correctly on the entire test. Then,  $X$  has a binomial distribution with  $n = 100$  and  $p = 0.2$ . Check the success/failure condition :

$$np = 100 \cdot 0.2 = 20 > 10, \quad n(1-p) = 80 > 10.$$

Using the normal approximation to the binomial distribution, we can approximate the distribution of the standardized  $X$  by a standard normal distribution. We have  $\mu = EX = np = 20$  and

$$\sigma = SD(X) = \sqrt{np(1-p)} = \sqrt{100 \cdot (0.2) \cdot (0.8)} = 4.$$

Now we have for  $a = 20$  and  $b = 30$

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\ &\approx P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \end{aligned}$$

where  $Z$  is a standard normal random variable. We find

$$\begin{aligned} \frac{a-\mu}{\sigma} &= \frac{20-20}{\sigma} = 0, \\ \frac{b-\mu}{\sigma} &= \frac{30-20}{4} = 10/4 = 2.5 \end{aligned}$$

The numbers  $\frac{a-\mu}{\sigma}$  and  $\frac{b-\mu}{\sigma}$  are called the  $Z$ -scores of  $a$  and  $b$  respectively; thus from the table of  $Z$

$$\begin{aligned} P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) &= P(0 \leq Z \leq 2.5) \\ &= P(Z \leq 2.5) - P(Z \leq 0) \\ &= 0.9938 - 0.5 = 0.4938. \end{aligned}$$

**Answer (Exercise 7.10)** — HINT: Take the first derivative of  $\theta(1-\theta)$  with respect to  $\theta$ , set it equal to 0 and solve for  $\theta$ . This solution will give the point at which  $\theta(1-\theta)$  has zero slope. Now, find the second derivative of  $\theta(1-\theta)$  with respect to  $\theta$ , and evaluate it at the solution to see if the maximum is achieved with a negative second derivative.

**Answer (Exercise 7.14)** —

Likelihood is just the joint PDF (if continuous) or joint PMF (if discrete) of the data **but** seen as a function of the parameter  $\theta$ :

$$L(\theta) = L(\theta; (x_1, x_2, \dots, x_n)) = f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta)$$

We sometimes write  $L(\theta)$  instead of  $L(\theta; (x_1, x_2, \dots, x_n))$  for notational simplicity. In this case, since we are assuming independent and identically distributed observations, the joint PDF/PMF is simply the product of the marginal PDFs/PMFs:

$$L(\theta) = f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f_{X_i}(x_i; \theta) = \prod_{i=1}^n f_{X_1}(x_i; \theta)$$

Since each  $X_i \stackrel{IID}{\sim} \text{Bernoulli}(\theta)$  RV, the marginal PMF of each  $X_i$  is the same as that of the first RV  $X_1$ , which is:

$$f_{X_1}(x_i; \theta) = \begin{cases} \theta & \text{if } x_i = 1 \\ 0 & \text{if } x_i = 0 \end{cases}.$$

From this we get the eight likelihood functions in the question:

1. When  $(x_1) = (1)$ , there is only one data point so  $n = 1$  and therefore

$$L(\theta) = \prod_{i=1}^1 f_{X_1}(x_1; \theta) = f_{X_1}(x_1 = 1; \theta) = f_{X_1}(1; \theta) = \theta$$

The above step-by-step break down is for understanding only. In the exam, you can just write:

$$L(\theta) = \theta$$

Make a plot of  $L(\theta) = \theta$  as a function of  $\theta$  (with x-axis values taken by  $\theta$  in the unit interval  $[0, 1]$ ).

2. When  $(x_1) = (0)$

$$L(\theta) = \prod_{i=1}^1 f_{X_1}(x_1; \theta) = f_{X_1}(x_1 = 0; \theta) = f_{X_1}(0; \theta) = 1 - \theta$$

In the exam, you can just write:

$$L(\theta) = 1 - \theta$$

Make a plot of  $L(\theta) = 1 - \theta$  as a function of  $\theta$ .

3. When  $(x_1, x_2) = (0, 0)$ , we have  $n = 2$  data points and therefore

$$\begin{aligned} L(\theta) &= \prod_{i=1}^2 f_{X_1}(x_i; \theta) = f_{X_1}(x_1 = 0; \theta) \times f_{X_1}(x_2 = 0; \theta) \\ &= f_{X_1}(0; \theta) \times f_{X_1}(0; \theta) = (1 - \theta) \times (1 - \theta) = (1 - \theta)^2 \end{aligned}$$

Or just

$$L(\theta) = (1 - \theta) \times (1 - \theta) = (1 - \theta)^2 .$$

Make a plot of  $L(\theta) = (1 - \theta)^2$  as a function of  $\theta$ .

4. When  $(x_1, x_2) = (1, 1)$ , we have  $n = 2$  data points and therefore

$$\begin{aligned} L(\theta) &= \prod_{i=1}^2 f_{X_1}(x_i; \theta) = f_{X_1}(x_1 = 1; \theta) \times f_{X_1}(x_2 = 1; \theta) \\ &= f_{X_1}(1; \theta) \times f_{X_1}(1; \theta) = \theta \times \theta = \theta^2 \end{aligned}$$

Or just

$$L(\theta) = \theta \times \theta = \theta^2 .$$

Make a plot of  $L(\theta) = \theta^2$  as a function of  $\theta$ .

5. When  $(x_1, x_2) = (1, 0)$ , we have  $n = 2$  data points and therefore

$$L(\theta) = \theta \times (1 - \theta) = \theta - \theta^2 .$$

Make a plot of  $L(\theta) = \theta - \theta^2$  as a function of  $\theta$ . This plot is easy to draw if you overlay the plot for  $\theta$  and  $\theta^2$  (that you just made separately) and then see where  $\theta > \theta^2$  and by how much.

6. When  $(x_1, x_2) = (0, 1)$ ,

$$L(\theta) = (1 - \theta) \times \theta = \theta - \theta^2 .$$

Notice that the likelihood with data  $(x_1, x_2) = (0, 1)$  is the same as the likelihood with the previous data  $(x_1, x_2) = (1, 0)$ . This is because, the likelihood being a product (from the IID assumption) is invariant to the order of the data, i.e., first observing 0 and then observing 1 has the same likelihood as first observing 1 and then observing 0. This means, in general even when  $n > 2$ , only the number of 1's and 0's in the  $n$  IID Bernoulli( $\theta$ ) experiment affects the likelihood of  $\theta$ . You have already made the plot of  $L(\theta) = \theta - \theta^2$  as a function of  $\theta$  in the previous problem!

7. When  $(x_1, x_2, x_3) = (1, 1, 0)$

$$L(\theta) = \theta \times \theta \times (1 - \theta) = \theta^2(1 - \theta) = \theta^2 - \theta^3 .$$

This plot is easy to draw if you first plot  $\theta^2$  and  $\theta^3$  separately and then see how far apart they are to get a sense for  $\theta^2 - \theta^3$ .

8. When  $(x_1, x_2, x_3) = (0, 0, 1)$

$$L(\theta) = (1 - \theta) \times (1 - \theta) \times \theta = (1 - \theta)^2\theta = 1 - 2\theta + \theta^2 .$$

This is just a polynomial in  $\theta$  and can be plotted. It should be clear from this exercise that the likelihood with observation  $(x_1, x_2, \dots, x_n)$  from  $n$  IID Bernoulli( $\theta$ ) RVs is just

$$L(\theta) = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}$$

where, the number of 1's is  $\sum_{i=1}^n x_i$  and the number of 0's is  $n - \sum_{i=1}^n x_i$ .

See Figure 19 on page 125 of the notes for the plots of these likelihoods as well as those of all possible observations one could have from  $n = 1$ ,  $n = 2$  or  $n = 3$  trials (seen as vertices of the unit interval  $[0, 1]$ , the unit square  $[0, 1]^2$  and the unit cube  $[0, 1]^3$ , respectively).

### Answer (Exercise 7.15) —

Since all five of these problems involve  $n$  IID samples we note that the likelihood function is

$$L(\theta) = f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f_{X_i}(x_i; \theta) = \prod_{i=1}^n f_{X_1}(x_i; \theta)$$

For ease of notation, we just write  $f(x; \theta)$ , instead of the more accurate  $f_{X_1}(x; \theta)$ , for the **common** PDF/PMF of each RV  $X_i$ . Thus, for IID samples we just write the likelihood as

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta) .$$

Recall that the logarithm of a product is the sum of the logarithms of each term in the product, i.e.,  $\log(a \times b) = \log(a) + \log(b)$ . More generally this means:

$$\log \left( \prod_{i=1}^n a_i \right) = \log(a_1 \times a_2 \times \dots \times a_n) = \log(a_1) + \log(a_2) + \dots + \log(a_n) = \sum_{i=1}^n \log(a_i)$$

The above formula won't appear in the formula sheet — you should know this by now. Putting all of the above facts together we can get the log-likelihood as

$$\ell(\theta) = \log(L(\theta)) = \log \left( \prod_{i=1}^n f(x_i; \theta) \right) = \sum_{i=1}^n \log(f(x_i; \theta))$$

Recall the main steps (from Section 7.8.2) to find  $\hat{\theta}_n$ , the maximum likelihood estimate (MLE) of the unknown parameter  $\theta^*$  according to which the data is independently and identically distributed, are as follows:

(Step 1:) find  $\ell(\theta)$ , the log-likelihood as a function of the parameter  $\theta$ , (Step 2:) find  $\frac{d}{d\theta}\ell(\theta)$ , the first derivative of  $\ell(\theta)$  with respect to  $\theta$ , (Step 3:) solve the equation  $\frac{d}{d\theta}\ell(\theta) = 0$  for  $\theta$  and set this solution equal to  $\hat{\theta}_n$ , (Step 4:) find  $\frac{d^2}{d\theta^2}\ell(\theta)$ , the second derivative of  $\ell(\theta)$  with respect to  $\theta$  and finally (Step 5:)  $\hat{\theta}_n$  is the MLE if  $\frac{d^2}{d\theta^2}\ell(\theta) < 0$ .

We are now ready to answer the four questions in this problem.

1.

**Step 1:** If  $x_i \in (0, 1)$  for each  $i \in \{1, 2, \dots, n\}$ , i.e. when each data point lies inside the open interval  $(0, 1)$ , the log-likelihood is

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^n \log(f_X(x_i; \theta)) = \sum_{i=1}^n \left( \log(\theta x_i^{\theta-1}) \right) = \sum_{i=1}^n \left( \log(\theta) + \log(x_i^{\theta-1}) \right) \\ &= \sum_{i=1}^n (\log(\theta) + (\theta - 1)(\log(x_i))) = \sum_{i=1}^n (\log(\theta) + \theta \log(x_i) - \log(x_i)) \\ &= \sum_{i=1}^n \log(\theta) + \sum_{i=1}^n \theta \log(x_i) - \sum_{i=1}^n \log(x_i) = n \log(\theta) + \theta \sum_{i=1}^n \log(x_i) - \sum_{i=1}^n \log(x_i)\end{aligned}$$

**Step 2:**

$$\frac{d}{d\theta}(\ell(\theta)) = \frac{d}{d\theta} \left( n \log(\theta) + \theta \sum_{i=1}^n \log(x_i) - \sum_{i=1}^n \log(x_i) \right) = \frac{n}{\theta} + \sum_{i=1}^n \log(x_i) - 0 = \frac{n}{\theta} + \sum_{i=1}^n \log(x_i)$$

**Step 3:**

$$\frac{d}{d\theta}(\ell(\theta)) = 0 \iff \frac{n}{\theta} + \sum_{i=1}^n \log(x_i) = 0 \iff \frac{n}{\theta} = -\sum_{i=1}^n \log(x_i) \iff \theta = -\frac{n}{\sum_{i=1}^n \log(x_i)}$$

Let

$$\hat{\theta}_n = -\frac{n}{\sum_{i=1}^n \log(x_i)} .$$

**Step 4:**

$$\begin{aligned}\frac{d^2}{d\theta^2}\ell(\theta) &= \frac{d}{d\theta} \left( \frac{d}{d\theta}(\ell(\theta)) \right) = \frac{d}{d\theta} \left( \frac{n}{\theta} + \sum_{i=1}^n \log(x_i) \right) = \frac{d}{d\theta} \left( n\theta^{-1} + \sum_{i=1}^n \log(x_i) \right) \\ &= -n\theta^{-2} + 0 = -\frac{n}{\theta^2}\end{aligned}$$

**Step 5:** The problem states that  $\theta > 0$ . Since  $\theta^2 > 0$  and  $n \geq 1$ , we have indeed checked that

$$\frac{d^2}{d\theta^2}\ell(\theta) = -\frac{n}{\theta^2} < 0$$

and therefore the MLE is indeed

$$\hat{\theta}_n = \frac{-n}{\sum_{i=1}^n \log(x_i)} .$$

2. Step 1: If  $x_i \in (0, 1)$  for each  $i \in \{1, 2, \dots, n\}$ , i.e. when each data point lies inside the open interval  $(0, 1)$ , the log-likelihood is

$$\begin{aligned}
 \ell(\theta) &= \sum_{i=1}^n \log(f_X(x_i; \theta)) = \sum_{i=1}^n \left( \log\left(\frac{1}{\theta} x_i^{(1-\theta)/\theta}\right) \right) = \sum_{i=1}^n \left( \log\left(\frac{1}{\theta}\right) + \log\left(x_i^{(1-\theta)/\theta}\right) \right) \\
 &= \sum_{i=1}^n \left( \log\left(\frac{1}{\theta}\right) + \left(\frac{1-\theta}{\theta}\right) \log(x_i) \right) = \sum_{i=1}^n \left( \log\left(\frac{1}{\theta}\right) + \left(\frac{1}{\theta} - 1\right) \log(x_i) \right) \\
 &= \sum_{i=1}^n \left( \log\left(\frac{1}{\theta}\right) + \frac{1}{\theta} \log(x_i) - \log(x_i) \right) = \sum_{i=1}^n \log\left(\frac{1}{\theta}\right) + \sum_{i=1}^n \frac{1}{\theta} \log(x_i) - \sum_{i=1}^n \log(x_i) \\
 &= n \log\left(\frac{1}{\theta}\right) + \frac{1}{\theta} \sum_{i=1}^n \log(x_i) - \sum_{i=1}^n \log(x_i) = n \log(\theta^{-1}) + \frac{1}{\theta} \sum_{i=1}^n \log(x_i) - \sum_{i=1}^n \log(x_i) \\
 &= -n \log(\theta) + \theta^{-1} \sum_{i=1}^n \log(x_i) - \sum_{i=1}^n \log(x_i)
 \end{aligned}$$

Step 2:

$$\frac{d}{d\theta}(\ell(\theta)) = \frac{d}{d\theta} \left( -n \log(\theta) + \theta^{-1} \sum_{i=1}^n \log(x_i) - \sum_{i=1}^n \log(x_i) \right) = -n\theta^{-1} - \theta^{-2} \sum_{i=1}^n \log(x_i)$$

Step 3:

$$\frac{d}{d\theta}(\ell(\theta)) = 0 \iff -n\theta^{-1} - \theta^{-2} \sum_{i=1}^n \log(x_i) = 0$$

Multiplying both sides of the above equality by  $\theta^2$  we get

$$\begin{aligned}
 \theta^2 \times \left( -n\theta^{-1} - \theta^{-2} \sum_{i=1}^n \log(x_i) \right) &= 0 \times \theta^2 \iff \left( -n\theta - \sum_{i=1}^n \log(x_i) \right) = 0 \\
 \iff n\theta &= -\sum_{i=1}^n \log(x_i) \iff \theta = -\frac{1}{n} \sum_{i=1}^n \log(x_i)
 \end{aligned}$$

Let

$$\hat{\theta}_n = -\frac{1}{n} \sum_{i=1}^n \log(x_i) .$$

Step 4:

$$\frac{d^2}{d\theta^2} \ell(\theta) = \frac{d}{d\theta} \left( \frac{d}{d\theta}(\ell(\theta)) \right) = \frac{d}{d\theta} \left( -n\theta^{-1} - \theta^{-2} \sum_{i=1}^n \log(x_i) \right) = n\theta^{-2} + 2\theta^{-3} \sum_{i=1}^n \log(x_i)$$

Step 5: Since  $\theta > 0$  and  $n \geq 1$ , we know that  $n\theta^{-2} = n/\theta^2 > 0$ ,  $2\theta^{-3} = 2/\theta^3 > 0$ . And since every  $x_i$  only takes values in  $(0, 1)$  we know that  $\log(x_i) < 0$  and therefore  $\sum_{i=1}^n \log(x_i) < 0$ . This problem is more interesting because we have some positive and some negative terms in

$\frac{d^2}{d\theta^2}\ell(\theta)$ . Let us find out when  $\frac{d^2}{d\theta^2}\ell(\theta) < 0$

$$\begin{aligned} \frac{d^2}{d\theta^2}\ell(\theta) &= n\theta^{-2} + 2\theta^{-3} \sum_{i=1}^n \log(x_i) < 0 \\ \iff 2\theta^{-3} \sum_{i=1}^n \log(x_i) &< 0 - n\theta^{-2} && \text{subtracting } n\theta^{-2} \text{ from both sides preserves the inequality} \\ \iff \sum_{i=1}^n \log(x_i) &< -\frac{n\theta^{-2}}{2\theta^{-3}} && \text{dividing by the positive quantity } 2\theta^{-3} \text{ on both sides preserves the inequality} \\ \iff \sum_{i=1}^n \log(x_i) &< -\frac{n\theta^3}{2\theta^2} \iff \sum_{i=1}^n \log(x_i) &< -\frac{n\theta}{2} \end{aligned}$$

and therefore only when the observed data and the parameter jointly satisfy the condition:

$$\sum_{i=1}^n \log(x_i) < -\frac{n\theta}{2}$$

will the MLE be

$$\hat{\theta}_n = -\frac{1}{n} \sum_{i=1}^n \log(x_i) .$$

If the condition is not satisfied then we cannot be sure about the MLE we found by setting the first derivative of the log-likelihood function to 0. This exercise illustrates that just ensuring that the slope or the first derivative of the log-likelihood function is zero at the MLE does not necessarily ensure that the curvature or second derivative of the log-likelihood function will always be negative or concave downward in order to ensure a global maximum at the MLE for every observable data and every possible parameter.

3. Step 1: If  $x_i \in (0, \infty)$  for each  $i \in \{1, 2, \dots, n\}$ , i.e. when each data point is greater than 0, the log-likelihood is

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^n \log(f_X(x_i; \theta)) = \sum_{i=1}^n \left( \log \left( \frac{1}{2\theta^3} x_i^2 e^{-x_i/\theta} \right) \right) \\ &= \sum_{i=1}^n \left( \log \left( \frac{1}{2} \right) + \log \left( \frac{1}{\theta^3} \right) + \log(x_i^2) + \log(e^{-x_i/\theta}) \right) \\ &= \sum_{i=1}^n (\log(2^{-1}) + \log(\theta^{-3}) + 2\log(x_i) + (-x_i/\theta)) \\ &= \sum_{i=1}^n (-\log(2) - 3\log(\theta) + 2\log(x_i) - x_i\theta^{-1}) \\ &= -\sum_{i=1}^n \log(2) - \sum_{i=1}^n 3\log(\theta) + \sum_{i=1}^n 2\log(x_i) - \sum_{i=1}^n x_i\theta^{-1} \\ &= -n\log(2) - 3n\log(\theta) + \sum_{i=1}^n 2\log(x_i) - \sum_{i=1}^n x_i\theta^{-1} \end{aligned}$$

Step 2:

$$\begin{aligned}
 \frac{d}{d\theta}(\ell(\theta)) &= \frac{d}{d\theta} \left( -n \log(2) - 3n \log(\theta) + \sum_{i=1}^n 2 \log(x_i) - \sum_{i=1}^n x_i \theta^{-1} \right) \\
 &= -0 - 3n\theta^{-1} + \frac{d}{d\theta} \left( \sum_{i=1}^n 2 \log(x_i) \right) - \frac{d}{d\theta} \left( \sum_{i=1}^n x_i \theta^{-1} \right) \\
 &= -3n\theta^{-1} + 0 - \sum_{i=1}^n \frac{d}{d\theta} (x_i \theta^{-1}) = -3n\theta^{-1} - \sum_{i=1}^n (-x_i \theta^{-2}) = -3n\theta^{-1} + \sum_{i=1}^n x_i \theta^{-2}
 \end{aligned}$$

Step 3:

$$\begin{aligned}
 \frac{d}{d\theta}(\ell(\theta)) = 0 &\iff -3n\theta^{-1} + \sum_{i=1}^n x_i \theta^{-2} = 0 \iff 3n\theta^{-1} = \sum_{i=1}^n x_i \theta^{-2} \\
 &\iff 3n\theta^{-1} \times \theta^2 = \sum_{i=1}^n x_i \theta^{-2} \times \theta^2 \quad \text{Multiplying both sides of the equality by } \theta^2 \\
 &\iff 3n\theta = \sum_{i=1}^n x_i \iff \theta = \frac{1}{3n} \sum_{i=1}^n x_i
 \end{aligned}$$

Let

$$\hat{\theta}_n = \frac{1}{3n} \sum_{i=1}^n x_i .$$

Step 4:

$$\begin{aligned}
 \frac{d^2}{d\theta^2} \ell(\theta) &= \frac{d}{d\theta} \left( \frac{d}{d\theta} (\ell(\theta)) \right) = \frac{d}{d\theta} \left( -3n\theta^{-1} + \sum_{i=1}^n x_i \theta^{-2} \right) = 3n\theta^{-2} + \frac{d}{d\theta} \left( \sum_{i=1}^n x_i \theta^{-2} \right) \\
 &= 3n\theta^{-2} + \sum_{i=1}^n \frac{d}{d\theta} (x_i \theta^{-2}) = 3n\theta^{-2} + \sum_{i=1}^n (-2x_i \theta^{-3}) = 3n\theta^{-2} - 2\theta^{-3} \sum_{i=1}^n x_i
 \end{aligned}$$

Step 5: The problem states that  $\theta > 0$  and each data point  $x_i > 0$ . Thus  $3n\theta^{-2} = 3n/\theta^2 > 0$ , and more crucially the cubic term  $2\theta^{-3} = 2/\theta^3 > 0$ . Finally with at least one sample  $n \geq 1$  and each data point  $x_i > 0$ , we have the following condition for the negativity of the second derivative

$$\begin{aligned}
 \frac{d^2}{d\theta^2} \ell(\theta) < 0 &\iff 3n\theta^{-2} - 2\theta^{-3} \sum_{i=1}^n x_i < 0 \quad \text{you can stop here for full credit in exam} \\
 &\iff 3n\theta^{-2} < 2\theta^{-3} \sum_{i=1}^n x_i \iff \theta^{-2}\theta^3 < \frac{2}{3n} \sum_{i=1}^n x_i \iff \theta < \frac{2}{3n} \sum_{i=1}^n x_i
 \end{aligned}$$

and therefore when the above condition is satisfied the MLE is indeed

$$\hat{\theta}_n = \frac{\sum_{i=1}^n x_i}{3n} .$$

4. Step 1: If  $x_i \in (0, \infty)$  for each  $i \in \{1, 2, \dots, n\}$ , the log-likelihood is

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^n \log(f_X(x_i; \theta)) = \sum_{i=1}^n \left( \log\left(\frac{x_i}{\theta^2} e^{-\frac{1}{2}(x_i/\theta)^2}\right) \right) = \sum_{i=1}^n \left( \log\left(\frac{x_i}{\theta^2}\right) + \log\left(e^{-\frac{1}{2}(x_i/\theta)^2}\right) \right) \\ &= \sum_{i=1}^n \left( \log(x_i) - \log(\theta^2) - \frac{1}{2}(x_i/\theta)^2 \right) = \sum_{i=1}^n (\log(x_i)) - \sum_{i=1}^n (2 \log(\theta)) - \sum_{i=1}^n \left( \frac{1}{2} x_i^2 \theta^{-2} \right) \\ &= \sum_{i=1}^n \log(x_i) - 2n \log(\theta) - \sum_{i=1}^n \left( \frac{1}{2} x_i^2 \theta^{-2} \right)\end{aligned}$$

Step 2:

$$\begin{aligned}\frac{d}{d\theta}(\ell(\theta)) &= \frac{d}{d\theta} \left( \sum_{i=1}^n (\log(x_i)) - 2n \log(\theta) - \sum_{i=1}^n \left( \frac{1}{2} x_i^2 \theta^{-2} \right) \right) \\ &= \frac{d}{d\theta} \left( \sum_{i=1}^n (\log(x_i)) \right) - \frac{d}{d\theta} (2n \log(\theta)) - \frac{d}{d\theta} \left( \sum_{i=1}^n \left( \frac{1}{2} x_i^2 \theta^{-2} \right) \right) \\ &= 0 - 2n \frac{1}{\theta} - \sum_{i=1}^n \left( \frac{1}{2} x_i^2 (-2\theta^{-3}) \right) = -2n\theta^{-1} + \theta^{-3} \sum_{i=1}^n x_i^2\end{aligned}$$

Step 3:

$$\begin{aligned}\frac{d}{d\theta}(\ell(\theta)) = 0 &\iff -2n\theta^{-1} + \theta^{-3} \sum_{i=1}^n x_i^2 = 0 \iff 2n\theta^{-1} = \theta^{-3} \sum_{i=1}^n x_i^2 \\ &\iff 2n\theta^{-1}\theta^3 = \sum_{i=1}^n x_i^2 \iff \theta^2 = \frac{1}{2n} \sum_{i=1}^n x_i^2 \iff \theta = \sqrt{\frac{1}{2n} \sum_{i=1}^n x_i^2}\end{aligned}$$

Let

$$\hat{\theta}_n = \sqrt{\frac{1}{2n} \sum_{i=1}^n x_i^2}.$$

Step 4:

$$\frac{d^2}{d\theta^2} \ell(\theta) = \frac{d}{d\theta} \left( \frac{d}{d\theta}(\ell(\theta)) \right) = \frac{d}{d\theta} \left( -2n\theta^{-1} + \theta^{-3} \sum_{i=1}^n x_i^2 \right) = 2n\theta^{-2} - 3\theta^{-4} \sum_{i=1}^n x_i^2$$

Step 5: Since  $\theta > 0$ , we have the following condition for the the second derivative to be negative

$$\begin{aligned}\frac{d^2}{d\theta^2} \ell(\theta) = 2n\theta^{-2} - 3\theta^{-4} \sum_{i=1}^n x_i^2 < 0 \quad &\text{you can stop here for full credit in exam} \\ &\iff 2n\theta^{-2} < 3\theta^{-4} \sum_{i=1}^n x_i^2 \iff \theta^2 < \frac{3}{2n} \sum_{i=1}^n x_i^2\end{aligned}$$

and therefore when the inequality above is satisfied the MLE is indeed

$$\hat{\theta}_n = \sqrt{\frac{1}{2n} \sum_{i=1}^n x_i^2}.$$

# Chapter 8

## Appendix

### 8.1 Code

**Labwork 234 (PDF and DF of a Normal( $\mu, \sigma^2$ ) RV)** Here are the functions to evaluate the PDF and DF of a Normal( $\mu, \sigma^2$ ) RV  $X$  at a given  $x$ .

---

```
function fx = NormalPdf(x,Mu,SigmaSq)
% Returns the Pdf of Normal(Mu, SigmaSq), at x,
% where Mu=mean and SigmaSq = Variance
%
% Usage: fx = NormalPdf(x,Mu,SigmaSq)
if SigmaSq <= 0
    error('Variance must be > 0')
    return
end

Den = ((x-Mu).^2)/(2*SigmaSq);
Fac = sqrt(2*pi)*sqrt(SigmaSq);

fx = (1/Fac)*exp(-Den);
```

---

---

```
function Fx = NormalCdf(x,Mu,SigmaSq)
% Returns the Cdf of Normal(Mu, SigmaSq), at x,
% where Mu=mean and SigmaSq = Variance using
% MATLAB's error function erf
%
% Usage: Fx = NormalCdf(x,Mu,SigmaSq)
if SigmaSq <= 0
    error('Variance must be > 0')
    return
end

Arg2Erf = (x-Mu)/sqrt(SigmaSq*2);
Fx = 0.5*erf(Arg2Erf)+0.5;
```

---

Plots of the PDF and DF of several Normally distributed RVs depicted in Figure 3.15 were generated using the following script file:

---

```
% PlotPdfCdfNormal.m script file
% Plot of some pdf's and cdf's of the Normal(mu,SigmaSq) RV X
```

---

```
%  
x=[-6:0.0001:6]; % points from the subset [-5,5] of the support of X  
subplot(1,2,1) % first plot of a 1 by 2 array of plots  
plot(x,NormalPdf(x,0,1),'r') % pdf of RV Z ~ Normal(0,1)  
hold % to superimpose plots  
plot(x,NormalPdf(x,0,1/10),'b') % pdf of RV X ~ Normal(0,1/10)  
plot(x,NormalPdf(x,0,1/100),'m') % pdf of RV X ~ Normal(0,1/100)  
plot(x,NormalPdf(x,-3,1),'r--') % pdf of RV Z ~ Normal(-3,1)  
plot(x,NormalPdf(x,-3,1/10),'b--') % pdf of RV X ~ Normal(-3,1/10)  
plot(x,NormalPdf(x,-3,1/100),'m--') % pdf of RV X ~ Normal(-3,1/100)  
 xlabel('x')  
 ylabel('f(x; \mu, \sigma^2)')  
 legend('f(x;0,1)', 'f(x;0,10^{-1})', 'f(x;0,10^{-2})', 'f(x;-3,1)', 'f(x;-3,10^{-1})', 'f(x;-3,10^{-2})')  
 subplot(1,2,2) % second plot of a 1 by 2 array of plots  
plot(x,NormalCdf(x,0,1),'r') % DF of RV Z ~ Normal(0,1)  
hold % to superimpose plots  
plot(x,NormalCdf(x,0,1/10),'b') % DF of RV X ~ Normal(0,1/10)  
plot(x,NormalCdf(x,0,1/100),'m') % DF of RV X ~ Normal(0,1/100)  
plot(x,NormalCdf(x,-3,1),'r--') % DF of RV Z ~ Normal(-3,1)  
plot(x,NormalCdf(x,-3,1/10),'b--') % DF of RV X ~ Normal(-3,1/10)  
plot(x,NormalCdf(x,-3,1/100),'m--') % DF of RV X ~ Normal(-3,1/100)  
 xlabel('x')  
 ylabel('F(x; \mu, \sigma^2)')  
 legend('F(x;0,1)', 'F(x;0,10^{-1})', 'F(x;0,10^{-2})', 'F(x;-3,1)', 'F(x;-3,10^{-1})', 'F(x;-3,10^{-2})')
```

---

**Labwork 235 (PDF and DF of an Exponential( $\lambda$ ) RV  $X$ )** Here are the functions to evaluate the PDF and DF of an Exponential( $\lambda$ ) RV  $X$  at a given  $x$  (point or a vector).

---

```
function fx = ExponentialPdf(x,Lambda) ExponentialPdf.m  
% Returns the Pdf of Exponential(Lambda) RV at x,  
% where Lambda = rate parameter  
  
% Usage: fx = ExponentialPdf(x,Lambda)  
if Lambda <= 0  
    error('Rate parameter Lambda must be > 0')  
    return  
end  
  
fx = Lambda * exp(-Lambda * x);
```

---



---

```
function Fx = ExponentialCdf(x,Lambda) ExponentialCdf.m  
% Returns the Cdf of Exponential(Lambda) RV at x,  
% where Lambda = rate parameter  
  
% Usage: Fx = ExponentialCdf(x,Lambda)  
if Lambda <= 0  
    error('Rate parameter Lambda must be > 0')  
    return  
end  
  
Fx = 1.0 - exp(-Lambda * x);
```

---

Plots of the PDF and DF of several Exponentially distributed RVs at four axes scales that are depicted in Figure 3.13 were generated using the following script file:

---

```
% PlotPdfCdfExponential.m script file  
% Plot of some pdf's and cdf's of the Exponential(Lambda) RV X  
%  
x=[0:0.0001:100]; % points from the subset [0,100] of the support of X
```

---

```

subplot(2,4,1) % first plot of a 1 by 2 array of plots
plot(x,ExponentialPdf(x,1),'r:','LineWidth',2) % pdf of RV X ~ Exponential(1)
hold on % to superimpose plots
plot(x,ExponentialPdf(x,10),'b--','LineWidth',2) % pdf of RV X ~ Exponential(10)
plot(x,ExponentialPdf(x,1/10),'m','LineWidth',2) % pdf of RV X ~ Exponential(1/10)
xlabel('x')
ylabel('f(x; \lambda)')
legend('f(x;1)', 'f(x;10)', 'f(x;10^{-1})')
axis square
axis([0,2,0,10])
title('Standard Cartesian Scale')
hold off

subplot(2,4,2)
semilogx(x,ExponentialPdf(x,1),'r:','LineWidth',2) % pdf of RV X ~ Exponential(1)
hold on % to superimpose plots
semilogx(x,ExponentialPdf(x,10),'b--','LineWidth',2) % pdf of RV X ~ Exponential(10)
semilogx(x,ExponentialPdf(x,1/10),'m','LineWidth',2) % pdf of RV X ~ Exponential(1/10)
% xlabel('x')
% ylabel('f(x; \lambda)')
% legend('f(x;1)', 'f(x;10)', 'f(x;10^{-1})')
axis square
axis([0,100,0,10])
title('semilog(x) Scale')
hold off

subplot(2,4,3)
semilogy(x,ExponentialPdf(x,1),'r:','LineWidth',2) % pdf of RV X ~ Exponential(1)
hold on % to superimpose plots
semilogy(x,ExponentialPdf(x,10),'b--','LineWidth',2) % pdf of RV X ~ Exponential(10)
semilogy(x,ExponentialPdf(x,1/10),'m','LineWidth',2) % pdf of RV X ~ Exponential(1/10)
% xlabel('x');
% ylabel('f(x; \lambda)');
% legend('f(x;1)', 'f(x;10)', 'f(x;10^{-1})')
axis square
axis([0,100,0,1000000])
title('semilog(y) Scale')
hold off

x=[ 0:0.001:1] [1.001:1:100000]; % points from the subset [0,100] of the support of X
subplot(2,4,4)
loglog(x,ExponentialPdf(x,1),'r:','LineWidth',2) % pdf of RV X ~ Exponential(1)
hold on % to superimpose plots
loglog(x,ExponentialPdf(x,10),'b--','LineWidth',2) % pdf of RV X ~ Exponential(10)
loglog(x,ExponentialPdf(x,1/10),'m','LineWidth',2) % pdf of RV X ~ Exponential(1/10)
% xlabel('x')
% ylabel('f(x; \lambda)')
% legend('f(x;1)', 'f(x;10)', 'f(x;10^{-1})')
axis square
axis([0,100000,0,1000000])
title('loglog Scale')
hold off

x=[0:0.0001:100]; % points from the subset [0,100] of the support of X
subplot(2,4,5) % second plot of a 1 by 2 array of plots
plot(x,ExponentialCdf(x,1),'r:','LineWidth',2) % cdf of RV X ~ Exponential(1)
hold on % to superimpose plots
plot(x,ExponentialCdf(x,10),'b--','LineWidth',2) % cdf of RV X ~ Exponential(10)
plot(x,ExponentialCdf(x,1/10),'m','LineWidth',2) % cdf of RV X ~ Exponential(1/10)
xlabel('x')
ylabel('F(x; \lambda)')
legend('F(x;1)', 'F(x;10)', 'F(x;10^{-1})')
axis square
axis([0,10,0,1])
hold off

```

```

subplot(2,4,6) % second plot of a 1 by 2 array of plots
semilogx(x,ExponentialCdf(x,1),'r:','LineWidth',2) % cdf of RV X ~ Exponential(1)
hold on % to superimpose plots
semilogx(x,ExponentialCdf(x,10),'b--','LineWidth',2) % cdf of RV X ~ Exponential(10)
semilogx(x,ExponentialCdf(x,1/10),'m','LineWidth',2) % cdf of RV X ~ Exponential(1/10)
%xlabel('x')
%ylabel('F(x; \lambda)')
%legend('F(x;1)', 'F(x;10)', 'F(x;10^{-1})')
axis square
axis([0,100,0,1])
%title('semilog(x) Scale')
hold off

subplot(2,4,7)
semilogy(x,ExponentialCdf(x,1),'r:','LineWidth',2) % cdf of RV X ~ Exponential(1)
hold on % to superimpose plots
semilogy(x,ExponentialCdf(x,10),'b--','LineWidth',2) % cdf of RV X ~ Exponential(10)
semilogy(x,ExponentialCdf(x,1/10),'m','LineWidth',2) % cdf of RV X ~ Exponential(1/10)
%xlabel('x');
%ylabel('F(x; \lambda)');
%legend('F(x;1)', 'F(x;10)', 'F(x;10^{-1})')
axis square
axis([0,10,0,1])
%title('semilog(y) Scale')
hold off

x=[ 0:0.001:1] [1.001:1:100000]; % points from the subset of the support of X
subplot(2,4,8)
loglog(x,ExponentialCdf(x,1),'r:','LineWidth',2) % cdf of RV X ~ Exponential(1)
hold on % to superimpose plots
loglog(x,ExponentialCdf(x,10),'b--','LineWidth',2) % cdf of RV X ~ Exponential(10)
loglog(x,ExponentialCdf(x,1/10),'m','LineWidth',2) % cdf of RV X ~ Exponential(1/10)
%xlabel('x')
%ylabel('F(x; \lambda)')
%legend('F(x;1)', 'F(x;10)', 'F(x;10^{-1})')
axis square
axis([0,100000,0,1])
%title('loglog Scale')
hold off

```

---

**Labwork 236 (Plotting the empirical DF)** A MATLAB function to plot the empirical DF (3.82) of  $n$  user-specified samples efficiently for massive number of samples. Read the following M-file for the algorithm:

---

ECDF.m

```

function [x1 y1] = ECDF(x, PlotFlag, LoxD, HixD)
% return the x1 and y1 values of empirical CDF
% based on samples in array x of RV X
% plot empirical CDF if PlotFlag is >= 1
%
% Call Syntax: [x1 y1] = ECDF(x, PlotFlag, LoxD,HixD);
% Input      : x = samples from a RV X (a vector),
%               PlotFlag is a number controlling plot (Y/N, marker-size)
%               LoxD is a number by which the x-axis plot range is extended to the left
%               HixD is a number by which the x-axis plot range is extended to the right
% Output     : [x1 y1] & empirical CDF Plot IF PlotFlag >= 1
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
R=length(x);           % assume x is a vector and R = Number of samples in x
x1=zeros(1,R+2);
y1=zeros(1,R+2);       % initialize y to null vectors
for i=1:1:R            % loop to append to x and y axis values of plot

```

```

y1(i+1)=i/R; % append equi-increments of 1/R to y
end % end of for loop
x1(2:R+1)=sort(x); % sorting the sample values
x1(1)=x1(2)-LoxD; x1(R+2)=x1(R+1)+HixD; % padding x for emp CDF to start at min(x) and end at max(x)
y1(1)=0; y1(R+2)=1; % padding y so emp CDF start at y=0 and end at y=1

% to make a ECDF plot for large number of points set the PlotFlag<1 and use
% MATLAB's plot function on the x and y values returned by ECDF -- stairs(x,y)
if PlotFlag >= 1 % Plot customized empirical CDF if PlotFlag >= 1
    %newplot;
    MSz=10/PlotFlag; % set Markersize MSz for dots and circles in ECDF plot
    % When PlotFlag is large MSz is small and the
    % Markers effectively disappear in the ecdf plot
    R=length(x1); % update R = Number of samples in x
    hold on % hold plot for superimposing plots

    for i=1:1:R-1
        if(i>1 && i ~= R-1)
            plot([x1(i),x1(i+1)], [y1(i),y1(i)], 'k o -', 'MarkerSize',MSz)
        end
        if (i< R-1)
            plot(x1(i+1),y1(i+1), 'k .', 'MarkerSize', 2.5*MSz)
        end
        plot([x1(i),x1(i+1)], [y1(i),y1(i)], 'k -')
        plot([x1(i+1),x1(i+1)], [y1(i),y1(i+1)], 'k -')
    end

    hold off;
end

```

---

Ideally, this function needs to be rewritten using primitives such as MATLAB's `line` commands.

**Labwork 237 (q-th sample quantile)** Let us implement Algorithm 1 as the following MATLAB function:

```

qthSampleQuantile.m
function qthSQ = qthSampleQuantile(q, SortedXs)
%
% return the q-th Sample Quantile from Sorted array of Xs
%
% Call Syntax: qthSQ = qthSampleQuantile(q, SortedXs);
%
% Input      : q = quantile of interest, NOTE: 0 <= q <= 1
%               SortedXs = sorted real data points in ascending order
% Output     : q-th Sample Quantile, ie, inverse ECDF evaluated at q

% store the length of the sorted data array SortedXs in n
N = length(SortedXs);
Nminus1TimesQ = (N-1)*q; % store (N-1)*q in a variable
Index = floor(Nminus1TimesQ); % store its floor in a C-style Index variable
Delta = Nminus1TimesQ - Index;
if Index == N-1
    qthSQ = SortedXs(Index+1);
else
    qthSQ = (1.0-Delta)*SortedXs(Index+1) + Delta*SortedXs(Index+2);
end

```

---

**Labwork 238 (Loading )** Let us save all the steps done in Labwork 118 into the following script M-file:

```

%% 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

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)
% output: ans = 145 14
% remove locations outside Chch and assign it to a new variable called EQ
% only keep earthquake hypocenter locations inside Chch
% only keep earthquakes with magnitude >3
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)
% output: 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.');

```

---

**Labwork 239 (Consistency of MLE in Bernoulli experiment)** Figure 7.9 was made with the following script file.

```

----- BernoulliMLEConsistency.m -----
clf;%clear any figures
rand('twister',736343); % initialize the Uniform(0,1) Sampler
N = 3; % 10^N is the maximum number of samples from RV
J = 100; % number of Replications for each n
u = rand(J,10^N); % generate 10X10^N samples from Uniform(0,1) RV U
p=0.5; % set p for the Bernoulli(p) trials
PS=[0:0.001:1]; % sample some values for p on [0,1] to plot likelihood
for i=1:N
    if(i==1) Pmin=0.; Pmax=1.0; Ymin=-70; Ymax=-10; Y=linspace(Ymin,Ymax,J); end
    if(i==2) Pmin=0.; Pmax=1.0; Ymin=-550; Ymax=-75; Y=linspace(Ymin,Ymax,J); end
    if(i==3) Pmin=0.3; Pmax=0.8; Ymin=-900; Ymax=-700; Y=linspace(Ymin,Ymax,J); end
    n=10^i;% n= sample size, ie, number of Bernoulli trials
    subplot(1,N,i)
    if(i==1) axis([Pmin Pmax Ymin -2]); end
    if(i==2) axis([Pmin Pmax Ymin -60]); end
    if(i==3) axis([Pmin Pmax Ymin -685]); end
    EmpCovSEhat=0; % track empirical coverage for SEhat
    EmpCovSE=0; % track empirical coverage for exact SE
    for j=1:J
        % transform the Uniform(0,1) samples to n Bernoulli(p) samples
        x=floor(u(j,1:n)+p);
        s = sum(x); % statistic s is the sum of x_i's
        % display the outcomes and their sum
        %display(x)
        %display(s)
        MLE=s/n; % Analytically MLE is s/n
        se = sqrt((1-p)*p/n); % standard error from known p
        sehat = sqrt((1-MLE)*MLE/n); % estimated standard error from MLE p
        ZalphaBy2 = 1.96; % for 95% CI
        if(abs(MLE-p)<=2*sehat) EmpCovSEhat=EmpCovSEhat+1; end
    end
end

```

```

line([MLE-2*sehat MLE+2*sehat],[Y(j) Y(j)],'Marker','+', 'LineStyle',':','LineWidth',1,'Color',[1 .0 .0])
if(abs(MLE-p)<=2*se) EmpCovSE=EmpCovSE+1; end
line([MLE-2*se MLE+2*se],[Y(j) Y(j)],'Marker','+', 'LineStyle','-')
% l is the Log Likelihood of data x as a function of parameter p
l=@(p)sum(log(p.^s * (1-p).^(n-s)));
hold on;
% plot the Log Likelihood function and MLE
semilogx(PS,arrayfun(l,PS),'m','LineWidth',1);
hold on; plot([MLE],[Y(j)],'.','Color','c'); % plot MLE
end
hold on;
line([p p], [Ymin, l(p)],'LineStyle',':','Marker','none','Color','k','LineWidth',2)
%axis([-0.1 1.1]);
%axis square;
LabelString=['n=' num2str(n) ' Cvrg.= ' num2str(EmpCovSE) '/ num2str(J) ...
    ' ~= ' num2str(EmpCovSEhat) '/ num2str(J)];
%text(0.75,0.05,LabelString)
title(LabelString)
hold off;
end

```

---

**Labwork 240 (Gilvenko-Cantelli Lemma for Uniform(0, 1))** The following script was used to generate the Figure 7.12.

```

% from Uniform(0,1) RV
rand('twister',76534); % initialize the Uniform(0,1) Sampler
N = 3; % 10^N is the maximum number of samples from Uniform(0,1) RV
u = rand(10,10^N); % generate 10 X 10^N samples from Uniform(0,1) RV U
x=[0:0.001:1];
% plot the ECDF from the first 10 samples using the function ECDF
for i=1:N
    SampleSize=10^i;
    subplot(1,N,i)
    plot(x,x,'r','LineWidth',2); % plot the DF of Uniform(0,1) RV in red
    % Get the x and y coordinates of SampleSize-based ECDF in x1 and y1 and
    % plot the ECDF using the function ECDF
    for j=1:10
        hold on;
        if (i==1) [x1 y1] = ECDF(u(j,1:SampleSize),2.5,0.2,0.2);
        else
            [x1 y1] = ECDF(u(j,1:SampleSize),0,0.1,0.1);
            stairs(x1,y1,'k');
        end
    end
    % Note PlotFlag is 1 and the plot range of x-axis is
    % incremented by 0.1 or 0.2 on either side due to last 2 parameters to ECDF
    % being 0.1 or 0.2
    % Alpha=0.05; % set alpha to 5% for instance
    % Epsn = sqrt((1/(2*SampleSize))*log(2/Alpha)); % epsilon_n for the confidence band
    hold on;
    stairs(x1,max(y1-Epsn,zeros(1,length(y1))), 'g'); % lower band plot
    stairs(x1,min(y1+Epsn,ones(1,length(y1))), 'g'); % upper band plot
    axis([-0.1 1.1 -0.1 1.1]);
    %axis square;
    LabelString=['n=' num2str(SampleSize)];
    text(0.75,0.05,LabelString)
    hold off;
end

```

---

## 8.2 Data

Here we describe some of the data sets we analyze.

**Data 241 (Our Maths & Stats Dept. Web Logs)** We assume access to a **Unix** terminal (**Linux**, **Mac OS X**, **Sun Solaris**, etc). We show how to get your hands dirty with web logs that track among others, every IP address and its time of login to our department web server over the world-wide-web. The raw text files of web logs may be manipulated but they are typically huge files and need some **Unix** command-line utilities.

```
rsa64@mathopt03:~> cd October010203WebLogs/
rsa64@mathopt03:~/October010203WebLogs> ls -al
-rw-r--r--+ 1 rsa64 math 7527169 2007-10-04 09:38 access-07_log.2
-rw-r--r--+ 1 rsa64 math 7727745 2007-10-04 09:38 access-07_log.3
```

The files are quite large over 7.5 MB each. So we need to compress it. We use the **gzip** and **gunzip** utility in any **Unix** environment to compress and decompress these large text files of web logs. After compression the file sizes are more reasonable.

```
rsa64@mathopt03:~/October010203WebLogs> gzip access-07_log.3
rsa64@mathopt03:~/October010203WebLogs> gzip access-07_log.2
rsa64@mathopt03:~/October010203WebLogs> zcat access-07_log.2.gz | grep ' 200 '
| awk '{ print \$4}' | sed -e 's/^\([0-9]\{2\}\)\([a-Z]\{3\}\)\([0-9]\{4\}\)\([0-9]\{2\}\):([0-9]\{2\})\([0-9]\{2\}\)/\3 \1 \4 \5 \6/'
2007 10 02 03 57 48
2007 10 02 03 58 31
.
.
.
2007 10 03 03 56 21
2007 10 03 03 56 52
```

Finally, there are 56485 and 53966 logins for the two 24-hour cycles, starting 01/Oct and 01/Oct, respectively. We can easily get these counts by further piping the previous output into the line counting utility **wc** with the **-l** option. All the **Unix** command-line tools mentioned earlier can be learned by typing **man** followed by the tool-name, for eg. type **man sed** to learn about the usage of **sed** at a **Unix** command shell. We further pipe the output of login times for the two 24-hour cycles starting 01/Oct and 02/Oct in format **YYYY MM DD HH MM SS** to **| sed -e 's/2007 10 //'** > **WebLogTimes20071001035730.dat** and **... > WebLogTimes20071002035730.dat**, respectively to strip away the redundant information on **YYYY MM**, namely **2007 10**, and only save the relevant information of **DD HH MM SS** in files named **WebLogTimes20071001035730.dat** and **WebLogTimes20071002035730.dat**, respectively. These two files have the data of interest to us. Note that the size of these two uncompressed final data files in plain text are smaller than the compressed raw web log files we started out from.

```
rsa64@mathopt03:~/October010203WebLogs> ls -al
-rw-r--r--+ 1 rsa64 math 677820 2007-10-05 15:36 WebLogTimes20071001035730.dat
-rw-r--r--+ 1 rsa64 math 647592 2007-10-05 15:36 WebLogTimes20071002035730.dat
-rw-r--r--+ 1 rsa64 math 657913 2007-10-04 09:38 access-07_log.2.gz
-rw-r--r--+ 1 rsa64 math 700320 2007-10-04 09:38 access-07_log.3.gz
```

Now that we have been familiarized with the data of login times to our web-server over 2 24-hour cycles, let us do some statistics. The log files and basic scripts are courtesy of the Department's computer systems administrators Paul Brouwers and Steve Gourdie. This data processing activity was shared in such detail to show you that statistics is only meaningful when the data and the process that generated it are clear to the experimenter. Let us process the data and visualize the empirical distribution functions using the following script:

---

```
WebLogDataProc.m
load WebLogTimes20071001035730.dat % read data from first file
% multiply day (October 1) by 24*60*60 seconds, hour by 60*60 seconds,
% minute by 60 seconds and seconds by 1, to rescale time in units of seconds
SecondsScale1 = [24*60*60; 60*60; 60; 1];
StartTime1 = [1 3 57 30] * SecondsScale1; % find start time in seconds scale
%now convert time in Day/Hours/Minutes/Seconds format to seconds scale from
%the start time
WebLogSeconds20071001035730 = WebLogTimes20071001035730 * SecondsScale1 - StartTime1;

% repeat the data entry process above on the second file
load WebLogTimes20071002035730.dat %
SecondsScale1 = [24*60*60; 60*60; 60; 1];
StartTime2 = [2 3 57 30] * SecondsScale1;
WebLogSeconds20071002035730 = WebLogTimes20071002035730 * SecondsScale1 - StartTime2;

% calling a more efficient ECDF function for empirical DF's
[x1 y1]=ECDF(WebLogSeconds20071001035730,0,0,0);
[x2 y2]=ECDF(WebLogSeconds20071002035730,0,0,0);
stairs(x1,y1,'r','linewidth',1) % draw the empirical DF for first dataset
hold on;
stairs(x2,y2,'b') % draw empirical cdf for second dataset

% set plot labels and legends and title
xlabel('time t in seconds')
ylabel('ECDF      F^(t)')
grid on
legend('Starting 10\01\0357\30', 'Starting 10\02\0357\30')
title('24-Hour Web Log Times of Maths & Stats Dept. Server at Univ. of Canterbury, NZ')

%To draw the confidence bands
Alpha=0.05; % set alpha
% compute epsilon_n for first dataset of size 56485
Epsn1 = sqrt((1/(2*56485))*log(2/Alpha));
stairs(x1,max(y1-Epsn1,zeros(1,length(y1))), 'g') % lower 1-alpha confidence band
stairs(x1,min(y1+Epsn1,ones(1,length(y1))), 'g') % upper 1-alpha confidence band

% compute epsilon_n for second dataset of size 53966
Epsn2 = sqrt((1/(2*53966))*log(2/Alpha));
stairs(x2,max(y2-Epsn2,zeros(1,length(y2))), 'g') % lower 1-alpha confidence band
stairs(x2,min(y2+Epsn2,ones(1,length(y2))), 'g') % upper 1-alpha confidence band
```

---

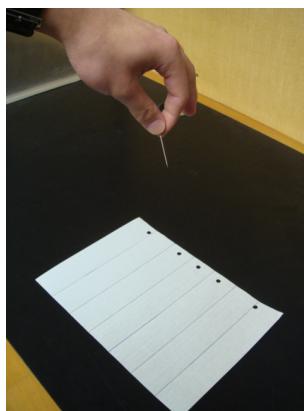
# Chapter 9

## Student Projects

Here are five examples of student projects. Students are strongly encouraged to work in teams of two or three. The projects are meant to encourage team-work. A student who wants to work on his or her own can do so. The following five student projects are model projects work the 5 bonus points. The project is completely optional and it is supposed to be an opportunity for learning certain important skills in mathematical and statistical communication.

### 9.1 Testing the Approximation of $\pi$ by Buffon's Needle Test

Amanda Hughes  
and Sung-Lim Suh



#### Abstract

This project is designed to investigate Buffon's Needle experiment. We replicated the concept of Buffon's Needle and tested the null hypothesis that the outcomes from tossing a needle are directly related to  $\pi$ . The Delta Method and non-parametric methods have been employed in this project.

##### 9.1.1 Introduction & Motivation

The report will firstly cover the background and motivation of this project. Next we will explain the geometric background to why this experiment approximates  $\pi$  and we will look at the statistical methodologies used. Following this we will discuss our results and conclusion. Finally we will discuss potential modifications.

### Background - Comte de Buffon

Comte de Buffon was born September 7 1707 in Montbard, France. He studied law in Dijon and medicine in Angers. After his studies, he had a chance to tour around France, Italy, and England to explore his knowledge in science. When he returned to France, Buffon published translations of one of Isaac Newton's works and his interest in science was now clear.

### Background - Buffon's Needle

Buffon's Needle Problem was first stated by Comte de Buffon in 1733, the solution was published later in 1777. The problem involves finding the probability that a needle of length  $l$  will land on a line, given a floor with equally spaced parallel lines (or floorboards) a distance  $d$  apart.

### Motivation

The motivation behind this project is to reconstruct Buffon's Needle Experiment. We wanted to see if an approximation of  $\pi$  was found by this somewhat simple experiment over 200 years ago.

#### 9.1.2 Materials & Methods

##### Materials

We first constructed a board which had several parallel lines. We did this by ruling lines on a piece of A4 paper. The width between the lines was either the same length as the needle or smaller than the length of the needle or larger than the length of the needle. The needle we used was a sewing needle of length 37mm. Instead of changing the needle to a smaller or larger needle for the three trials we ruled up three different sheets of A4 paper, one for each of the situations. The sheet that had the lines the same distance apart as the length of the needle had lines spaced 37mm apart. The sheet that had the lines closer together than the length of the needle had lines spaced 35mm apart. The sheet that had the lines further apart than the length of the needle had lines spaced 42mm apart. We recorded the data by entering it into Excel as the experiment was taking place. Sung-Lim tossed the needle and Amanda recorded the data.

##### Method

We tossed a needle 200 times for each of the three different distances apart of the lines. Sung-Lim tossed the needle for all trials so that the tossing could be done as identically as possible. Sung-Lim held the needle at the same height and dropped it in exactly the same way for each trial. Sung-Lim called out each toss as either "Yes" for the needle landing on a line or "No" for the needle not landing on a line. Any decisions that had to be made over whether the needle crossed the line or not were made by Sung-Lim. If the needle rolled or bounced off the page we disregarded that trial. A break was taken every 100 trials so that all trials could be done as identically as possible.

##### Geometric Aspect

We need to look at how this experiment approximates  $\pi$ . Let us begin by looking at the general case, where the needle is the same length as the distance between the floorboards.

Imagine a floor on which there are an infinite number of parallel lines (floorboards) spaced two units apart. You toss a needle that is also two units in length onto the floor. We want to find the probability that the needle crosses a line in the floorboards.  $l$ =the length of the needle and  $d$ =the distance between the lines.

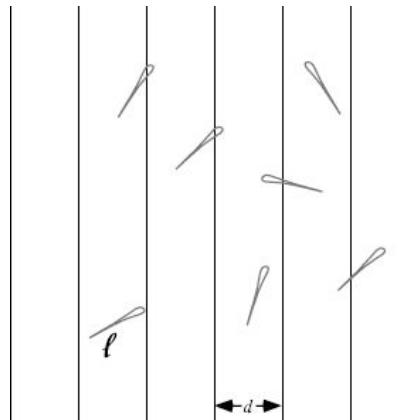


Figure 9.1: Example of Needle Tosses

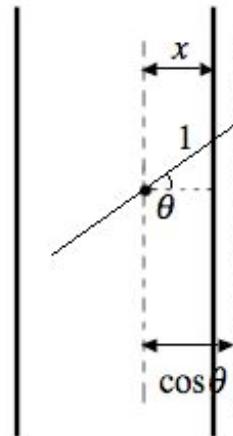


Figure 9.2: Explaining the outcomes of the needle toss

Take one board to examine. Let  $x$  be the shortest distance from the mid point of the needle to the nearest line. Let  $\theta$  be the angle between  $x$  and the needle. Imagine that there is a vertical line down from the end of the needle closest to the line.  $\cos(\theta)$  gives the distance from the midpoint of the line to this imaginary line.

As you can imagine, the needle can fall in an infinite amount of ways and positions relative to the parallel lines. Therefore,  $\theta$  also has an infinite amount of possibilities. We will limit these possibilities in two ways. Firstly only consider the line that is closest to the midpoint of the needle. Secondly we will divide the board into two by drawing an imaginary line down the board halfway between two lines and only consider needle positions on the right side of our halfway line. We will consider the needles whose middle falls to the left of this imaginary line to be a rotation of the right

side of the imaginary line. So we now only have two general situations to look at. One of these situations is shown in the above picture, where  $\theta$  is less than  $\pi/2$  radians. The other situation is where  $\theta$  is larger than  $\pi/2$  radians. In this case we will consider the angle  $\theta$  to actually be the angle below of  $x$  and we will think of this angle as negative. Therefore, the support of  $x$  is 0 to 1 and the support of  $\theta$  is  $-\pi/2$  to  $\pi/2$ . This is shown in figure 9.3.

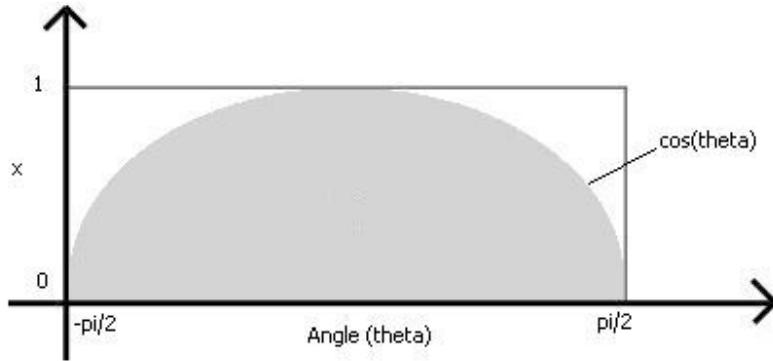


Figure 9.3: Outcome Space of Buffon's Needle Experiment for General Case

The shaded area in figure 9.3, represents when the needle crosses a line. This happens when  $x < \cos(\theta)$ . The total outcome space is  $\pi$  and the area under  $\cos(\theta)$  from  $-\pi/2$  to  $\pi/2$  is 2 (integrating  $\cos(\theta)$  from  $-\pi/2$  to  $\pi/2$ ). Therefore, the chance of a needle falling on a line is  $2/\pi$ .

For the case where the needle is shorter in length than the distance between the floorboards we need to modify the above slightly. We need to account for the exact length difference between the needle and the floorboards. This is added to the above explanation by multiplying  $2/\pi$  by  $y$ , which is the ratio of length of needle and distance between the lines, ie.  $y = l/d$ . Therefore the chance of a shorter needle landing on a line is  $2y/\pi$ .

$$\Pr(\text{needle crosses line}) = \int_0^{2\pi} \left( \frac{l|\cos \theta|}{d} \right) \frac{d\theta}{2\pi}$$

$$\Pr(\text{needle crosses line}) = \frac{2l}{d\pi} \int_0^{\frac{\pi}{2}} \cos \theta d\theta$$

$$\Pr(\text{needle crosses line}) = \frac{2l}{d\pi}$$

$$\Pr(\text{needle crosses line}) = \frac{2y}{\pi}$$

For the case where the needle is longer in length than the distance between the floorboards we get a more complex outcome. We again need to account for the exact length difference between the needle and the floorboards. Therefore, the chance of a longer needle landing on a line is:

$$\Pr(\text{needle crosses line}) = \frac{2}{\pi} (y - \sqrt{y^2 - 1} + \sec^{-1} y)$$

### Statistical Methodology

For this experiment we looked at the three different distances apart of the lines. For each of the three circumstances we used the same hypotheses.

H0(null hypothesis):  $\phi^* = \frac{2}{\pi}$ , the outcomes of tossing a needle are directly related to  $\pi$

H1(alternative hypothesis):  $\phi^* \neq \frac{2}{\pi}$ , the outcomes of tossing a needle are not directly related to  $\pi$   
For the general case:

$$(\Theta, X) \stackrel{IID}{\sim} \text{Uniform}([\frac{\pi}{2}, \frac{\pi}{2}] \times [0, 1])$$

as shown in figure 1.3. We know that for our simplest situation, where the length of the needle is the same as the distance between the lines:

$$\phi^* = \Pr((\Theta, X) \in \text{Shaded Region of figure 9.3}) = \frac{2}{\pi}$$

$$N_1, N_2, \dots, N_{200} \stackrel{IID}{\sim} \text{Bernoulli}(\phi^*)$$

This is the probability of a needle landing on a line. The trials for each of the three different distances apart of the lines are 200 independent and identically distributed Bernoulli trials.

Our maximum likelihood estimator of  $\phi^*$  is:

$$\hat{\phi}_{200} = \frac{n_1 + n_2 + \dots + n_{200}}{200}$$

This is the sample mean. But what we really want is a function of  $\phi^*$ , namely,  $\Psi(\Phi) := 2/\Phi$ . We now need the Delta Method. The Delta Method gives us the needed correction to transform an estimate and its confidence interval. By the Delta Method:

$$\pi \approx \psi^* = g(\phi^*) = \frac{2}{\phi^*}$$

and the maximum likelihood estimate of  $\psi^*$  is:

$$\hat{\psi}_{200} = g(\hat{\phi}_{200}) = \frac{2}{\hat{\phi}_{200}}$$

Next we can calculate the standard error of our estimator of  $\psi^*$ , namely,  $\hat{\Psi}_n = g(\hat{\Phi}_n)$ , and subsequently confidence intervals:

$$se(\hat{\Psi}_n) = |g'(\phi)|se(\hat{\Phi}_n)$$

$$se(\hat{\Psi}_n) = |\frac{d}{d\phi}g(\phi)|se(\hat{\Phi}_n)$$

$$se(\hat{\Psi}_n) = |\frac{d}{d\phi}(2\phi^{-1})|se(\hat{\Phi}_n)$$

$$se(\hat{\Psi}_n) = |-2\phi^{-2}|se(\hat{\Phi}_n)$$

$$se(\hat{\Psi}_n) = (2\phi^{-2})se(\hat{\Phi}_n)$$

where the estimated standard error is:

$$se(\hat{\psi}_{200}) = (2\phi_{200}^{-2})se(\hat{\phi}_{200}) = (2\phi_{200}^{-2})\frac{\hat{\phi}_{200}(1-\hat{\phi}_{200})}{200}$$

Now we can calculate the 95% confidence interval for  $\psi^* \approx \pi$ :

$$\hat{\psi}_{200} \pm 1.96se(\hat{\psi}_{200})$$

Now for the needle shorter in length than the distance between the lines, where  $\phi^* = \frac{2y}{\pi}$  and therefore by the Delta Method:  $\pi \approx \frac{2y}{\phi^*}$ ,  $\hat{\phi}_{200}$  is the sample mean of this set of data.

$$\psi^* = g(\phi^*) = \frac{2y}{\phi^*}$$

$$\hat{\Psi}_{200} = g(\hat{\phi}_{200}) = \frac{2y}{\hat{\phi}_{200}}$$

From this we can calculate the standard error of the estimator of  $\psi^*$ , namely,  $\hat{\Psi}_n$ , and subsequently confidence intervals as follows:

$$se(\hat{\Psi}_n) = |g'(\phi)| * se(\hat{\Phi}_n)$$

$$se(\hat{\Psi}_n) = \left| \frac{d}{d\phi} g(\phi) \right| * se(\hat{\Phi}_n)$$

$$se(\hat{\Psi}_n) = \left| \frac{d}{d\phi} (2y\phi^{-1}) \right| * se(\hat{\Phi}_n)$$

$$se(\hat{\Psi}_n) = | -2y\phi^{-2} | * se(\hat{\Phi}_n)$$

$$se(\hat{\Psi}_n) = (2y\phi^{-2}) * se(\hat{\Phi}_n)$$

where the estimated standard error is

$$se(\hat{\psi}_n) = (2y\hat{\phi}_{200}^{-2}) * se(\hat{\Phi}_n) \quad \text{and} \quad se(\hat{\Phi}_n) = \frac{\hat{\phi}_{200} * (1 - \hat{\phi}_{200})}{200}$$

Now we can calculate the 95% confidence interval for  $\psi^* \approx \pi$ :

$$\hat{\psi}_{200} \pm 1.96se(\hat{\psi}_{200})$$

Now for the needle longer in length than the distance between the lines, where  $\phi^* = \frac{2}{\pi}(y - \sqrt{y^2 - 1} + \sec^{-1}y)$  and therefore by the Delta Method:  $\pi \approx \frac{2}{\phi^*}(y - \sqrt{y^2 - 1} + \sec^{-1}y)$ ,  $\hat{\phi}_{200}$  is the sample mean of this set of data.

$$\psi^* = g(\phi^*) = \frac{2}{\phi^*}(y - \sqrt{y^2 - 1} + \sec^{-1}y)$$

$$\hat{\psi}_{200} = g(\hat{\phi}_{200}) = \frac{2}{\hat{\phi}_{200}}(y - \sqrt{y^2 - 1} + \sec^{-1}y)$$

From this we can calculate the standard error of  $\hat{\Psi}_{200}$ , the estimator of  $\psi^*$ , and subsequently its confidence interval:

$$\begin{aligned}
se(\hat{\Psi}_n) &= |g'(\phi)|se(\hat{\Phi}_n) \\
se(\hat{\Psi}_n) &= \left| \frac{d}{d\phi} g(\phi) \right| se(\hat{\Phi}_n) \\
se(\hat{\Psi}_n) &= \left| \frac{d}{d\phi} \frac{2}{\phi} (y - \sqrt{y^2 - 1} + \sec^{-1} y) \right| se(\hat{\Phi}_n) \\
se(\hat{\Psi}_n) &= |-2\phi^{-2}|(y - \sqrt{y^2 - 1} + \sec^{-1} y) se(\hat{\Phi}_n) \\
se(\hat{\Psi}_n) &= (2\phi^{-2})(y - \sqrt{y^2 - 1} + \sec^{-1} y) se(\hat{\Phi}_n)
\end{aligned}$$

where the estimated standard error is:

$$se(\hat{\Psi}_n) = (2\hat{\phi}_{200}^{-2})(y - \sqrt{y^2 - 1} + \sec^{-1} y) se(\hat{\Phi}_n) \quad \text{and} \quad se(\hat{\Phi}_n) = \frac{\hat{\phi}_{200}(1-\hat{\phi}_{200})}{200}$$

Now we can calculate the 95% confidence interval for  $\psi^* \approx \pi$ :

$$\hat{\psi}_{200} \pm 1.96 se(\hat{\psi}_{200})$$

### 9.1.3 Results

For the needle same in length as the distance between the lines, we got an approximation for  $\pi$  of 3.0769. The 95% Confidence Interval (2.7640, 3.3898), contains  $\pi$ .

For the needle shorter in length than the distance between the lines, we got an approximation for  $\pi$  of 2.9122. The 95% Confidence Interval (2.5861, 3.2384), contains  $\pi$ .

For the needle longer in length than the distance between the lines, we got an approximation for  $\pi$  of 2.8042. However, the 95% Confidence Interval (2.5769, 3.0316), does not contain  $\pi$ .

### 9.1.4 Conclusion

For the first two cases the 95% Confidence intervals both contain  $\pi$  so we cannot reject the null hypothesis that the outcomes from tossing a needle are directly related to  $\pi$ . For the final case the 95% Confidence interval does not contain  $\pi$  so we reject the null hypothesis and conclude that for this case the outcomes from tossing a needle are not directly related to  $\pi$ . The first two outcomes agree with Buffon's Needle while the third one may in fact be false, because our samples were quite small in hindsight.

### Potential Modification

There are several ways we could improve this experiment. Firstly we could increase sample size and see if we can get better approximations of  $\pi$ . Secondly we could investigate the chance of a needle landing on a line on a square tiled floor, this is the Buffon-Laplace Needle. We could also extend the idea to not only cover needles and lines but also shapes and more complicated arrangements of tiles or lines.

## Author Contributions

### Amanda Hughes

Original concept; research; recording of the data collection; MATLAB analysis; writing and construction of the report; LateX writing and compiling; some of the slides for presentation and some of the script for the presentation (mainly sections on the geometric aspect of the problem).

### Sung-Lim Suh

Research; collected data; majority of slides for presentation; majority of script for presentation, some of which was used in the report (Background on Comte de Buffon and Background of Buffon's Needle).

Many thanks to our lecturer Dr. Raazesh Sainudiin who spent much of his time discussing the geometric aspect of this report with us. We really appreciated being able to stop by his office almost whenever and have a discussion.

## References

Texts:

Stigler, Stephen M., *Statistics on the Table: The History of Statistical Concepts and Methods*, Harvard University Press, USA, 2002

Electronic Texts:

Hyna, Irene; Oetiker, Tobias; Partl, Hubert; Schlegl, Elisabeth., *The Not so Short Introduction to LATEX 2e or LATEX 2e in 90 Minutes*, 2000

Websites:

[www.angelfire.com/wa/hurben/buff.html](http://www.angelfire.com/wa/hurben/buff.html)  
[www.mathworld.wolfram.com/BuffonsNeedleProblem.html](http://www.mathworld.wolfram.com/BuffonsNeedleProblem.html)  
[www.mste.uiuc.edu/reese/buffon/buffon.html](http://www.mste.uiuc.edu/reese/buffon/buffon.html)  
[www.nndb.com/people/755/000091482/](http://www.nndb.com/people/755/000091482/)  
[www.ucmp.berkeley.edu/history/buffon2.html](http://www.ucmp.berkeley.edu/history/buffon2.html)  
[www.youtube.com/watch?v=Vws1jvMbs64](https://www.youtube.com/watch?v=Vws1jvMbs64)

## Appendix

Data and MATLAB Code:

---

```
Buffon.m
data% this is the data for the experiment where the needle is the same length as the distance between the lines
datalines=[1 1 1 0 0 1 1 1 1 0 1 0 1 1 1 1 1 0 1 0 0 1 0 0 1 0 1 1 1 ...
1 1 0 0 1 0 1 1 1 0 1 1 1 0 0 0 1 0 1 1 1 1 1 1 1 0 0 1 1 1 1 ...
0 1 0 0 0 1 1 1 0 0 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0 1 1 1 1 0 1 ...
1 0 0 1 1 0 0 0 0 1 1 0 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 ...
1 0 0 1 0 1 0 0 0 1 1 0 0 1 1 1 1 1 0 1 1 1 0 1 1 0 1 0 1 0 1 1 ...
1 0 1 1 0 0 1 1 0 0 1 1 0 1 1 1 1 0 0 1 1 1 1 1 1 0 1 1 1 1 0 1 ...
MLE_theta_star=mean(datalines)% find the maximum likelihood estimator, the sample mean
piapprox=2*(1/MLE_theta_star)% estimate of pi
StdErr=sqrt((MLE_theta_star*(1-MLE_theta_star))/200)% standard error
CI=[piapprox-(1.96*StdErr*2/(MLE_theta_star)^2),piapprox+(1.96*StdErr*2/(MLE_theta_star)^2)]% 95% Confidence Interval ...
for our approximate of pi

short% this is the data for the experiment where the needle is shorter in length than the distance between the lines
datalines=[1 1 0 0 1 1 1 1 1 0 0 0 1 1 0 1 1 1 1 0 1 1 1 1 0 0 0 0 ...]
```

```

1 1 1 0 1 1 0 1 1 0 0 1 0 1 1 0 1 1 1 1 0 0 1 1 1 1 0 1 0 0 1 0 ...
1 1 1 0 0 1 1 0 0 0 1 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 1 1 1 0 1 0 1 ...
1 0 0 0 1 1 1 1 1 0 0 0 1 0 0 1 1 1 1 1 0 0 1 1 0 1 0 1 1 1 1 1 ...
0 1 1 0 1 0 1 0 0 1 1 0 0 1 1 1 1 1 0 1 0 0 0 1 1 1 1 1 0 0 1 0 0 ...
0 1 1 1 1 0 0 1 1 0 0 1 1 1 1 0 1 1 1 0 0 1 1 0 0 0 1 1 1 1 0 1 1]

MLE_theta_star=mean(datalines)% find the maximum likelihood estimator, the sample mean
x=37/42%x=length/distance
piapprox=(2*x)*(1/MLE_theta_star)% estimate of pi
StdErr=sqrt((MLE_theta_star*(1-MLE_theta_star))/200)% standard error
CI=[piapprox-(1.96*StdErr*(2*x)/(MLE_theta_star)^2),piapprox+(1.96*StdErr*(2*x)/(MLE_theta_star)^2)]% 95% Confidence ...
Interval for our approximate of pi

long% this is the data for the experiment where the needle is longer in length than the distance between the lines
datalines=[0 0 1 1 0 0 1 1 1 0 1 0 1 1 1 1 0 1 1 0 1 0 0 1 1 1 ...
0 0 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 ...
1 1 1 1 1 0 1 0 1 1 0 1 1 1 1 1 1 0 0 1 1 1 0 1 1 1 1 1 1 0 1 ...
1 1 1 0 1 1 1 0 1 1 1 1 0 1 1 1 0 0 1 1 1 1 1 0 0 0 1 0 0 1 ...
1 0 0 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 ...
0 1 1 0 1 1 0 1 0 0 0 1 1 1 1 1 1 1 0 1 0 1 0 1 1 1 0 1 1 1]

MLE_theta_star=mean(datalines)% find the maximum likelihood estimator, the sample mean
x=37/35%x=length/distance
piapprox=(2*(1/MLE_theta_star))*(x-sqrt((x^2)-1)+asec(x))% estimate of pi
StdErr=sqrt((MLE_theta_star*(1-MLE_theta_star))/200)% standard error
CI=[piapprox-(1.96*StdErr*2/(MLE_theta_star)^2)*(x-sqrt((x^2)-1)+asec(x)),piapprox+(1.96*StdErr*2/(MLE_theta_star)^2)* ...
(x-sqrt((x^2)-1)+asec(x))]% 95% Confidence Interval for our approximate of pi

```

---

## 9.2 Estimating the Binomial probability $p$ for a Galton's Quincunx

Bry Ashman and Ryan Lawrence

### abstract

Galton's Quincunx is a physical device designed to simulate the discrete binomial distribution. we aim to create a physical model of the quincunx that is characterised by the probability of a ball going left is equal to the probability of it going right. From the conceptual model of the quincunx, we derive the binomial probability mass function. In order to evaluate the parameter of interest  $p$ , we will derive the maximum likelihood estimator and use this to estimate the actual parameter  $p$  of our physical model using 100 samples that are assumed to be independent and identically distributed.



### 9.2.1 Motivation & Introduction

The binomial distribution is a fundamental discrete probability distribution, being the natural extension of the Bernoulli trial to the sum of Bernoulli trials. The distribution describes the number of successes in a sequence of  $n$  binary trials, with a probability  $p$ . Each of these trials is a Bernoulli trial parameterised by  $p$ . A binomial distribution parameterised by  $n = 1$  and  $p$  is simply a *Bernoulli(p)* trial.

The quincunx was invented by Sir Francis Galton originally to demonstrate the normal distribution. The quincunx is simply an array of pegs spaced so that when a ball is dropped into a device, it bounces off the pegs with a probability  $p$  of going right and a probability of  $1-p$  of going left. It bounces off  $n$  pegs before being collected in a bin at the bottom of the device.

To this end, we aim to create a quincunx ideally parameterised by  $p = 0.5$  with  $n = 20$ . To verify this, we will use maximum likelihood estimation to test the null hypothesis that  $p = 0.5$ .

### 9.2.2 Materials and Methods

#### Construction of physical model

The physical model that we created consisted of nails arranged in a pattern on a sheet of wood that we hoped would achieve as close to the ideal probability of  $p=0.5$ .

#### Materials

- Plywood Sheet (1200 x 600 x 20mm)
- Perspex Sheets (1200 x 600 x 2mm and 550 x 800 x 2mm)
- Timber Strips (1200 x 25mm and 600 x 25mm)
- Nails (30 x 2mm)
- Chrome Balls (20mm)

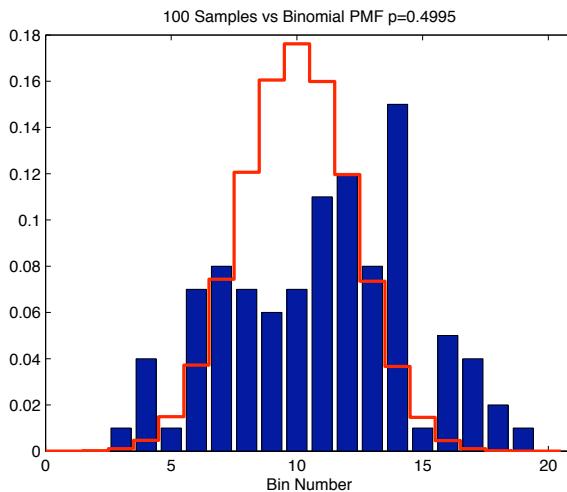
#### Construction Details

1. Mark 20 horizontal lines with 25mm spacings with the board in a portrait orientation.
2. Mark vertical lines at 25mm spacings from the centre of the board.
3. Place a nail at the top centre marking
4. Continue to place nails on the marked grid such that one marked grid point always separates the nails both vertically and horizontally.
5. Create the bins by attaching perspex strips directly below the nails of the last row.
6. Fit the edges to the main sheet.
7. The perspex sheet can now be attached to the edges of the quincunx.

A desirable feature of the quincunx is a release mechanism at the top to release the balls used to simulate a random variable and a release at the bottom to retrieve the balls after the experiment.

#### Sample Collection

To collect samples from the quincunx the balls are dropped into the device as identically as possible with sufficient time between each drop to ensure that the balls do not interfere with each other so as to keep the samples as identical as possible. The balls are collected in a series of bins numbered from 0 to 21, 0 representing the leftmost bin that the sample can be in and 21 being the rightmost bin. Since we assume that each sample is identical and independent, we record the cumulative number of balls in each bin after dropping 100 balls. The data is shown in the blue bars in the next figure.



### 9.2.3 Statistical Methodology

#### Deriving the binomial distribution

The binomial distribution can be thought of as a random walk in one dimension. The parameters map to this model as  $p$  being the probability of taking a step right and  $(1 - p)$  the probability of taking a step left, and  $n$  being the total number of steps taken. From this, it follows that, for a given number of  $n$  steps,  $x$  of which are to the right and  $n - x$  to the left, to find the probability that a combination of those  $n$  steps that will get you to the same point, you have to multiply the probability of the path by how many unique ways you can combine those steps. The number of ways of ordering the  $x$  right steps in a set of  $n$  steps is given by  $\binom{n}{x}$ . Therefore, the probability of ending up at a particular endpoint is as follows:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

A note about the endpoint: I have used the convention that the leftmost bucket is 0. The endpoint numbers also tell you how many right steps you have in the quincunx.

#### Parametric Estimation

In order to estimate the parameter  $p$  for our physical model, we will use a maximum likelihood estimator (MLE) since it is often regarded as asymptotically optimal. However, for the binomial distribution, the MLE is equivalent to the Method of Moments.

#### Deriving the maximum likelihood estimator

$$\begin{aligned} L(p) &= \prod_{i=1}^n \binom{N}{x_i} p^{x_i} (1 - p)^{N-x_i} \\ L(p) &= \prod_{i=1}^n \binom{N}{x_i} p^{\sum x_i} (1 - p)^{nN - \sum x_i} \\ \ln L(p) &= \sum_{i=1}^n \ln \binom{N}{x_i} \sum x_i \ln(p) (nN - \sum x_i) \ln(1 - p) \end{aligned}$$

$$\frac{d}{dp} \ln L(p) = \frac{\sum x_i}{p} - \frac{nN - \sum x_i}{1-p}$$

We can now set  $\frac{d}{dp} \ln L(p) = 0$  to find the maximum:

$$0 = \frac{\sum x_i}{p} - \frac{nN - \sum x_i}{1-p}$$

$$p = \frac{1}{nN} \sum_{i=1}^n x_i$$

Which is equivalent to:

$$p = \frac{1}{N} E(X)$$

#### 9.2.4 Results & Conclusion

##### Maximum Likelihood Estimation

The MLE of the parameter  $p$  of the quincunx is 0.4995, with a 95% normal based confidence interval of [0.4639,0.5351] calculated as derived above.

##### Conclusion

Maximum Likelihood Estimation from the 100 samples from the model of the quincunx has estimated the parameter for the binomial distribution to be in the range [0.4639,0.5351]. This would seem to verify that, in fact, even though the quincunx is a non-linear physical device that, overall, it is remarkably fair with  $p=0.5$  within the 95% normal based confidence interval.

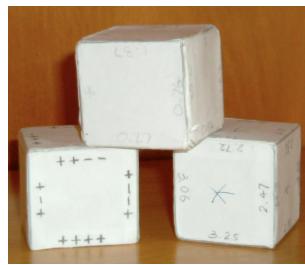
The estimated cumulative distribution function also suggests that the distribution will converge binomially. Thus, we can conclude as  $n \rightarrow \infty$ , it will converge on the standard normal distribution as a consequence of the central limit theorem.

### 9.3 Investigation of a Statistical Simulation from the 19th Century

Brett Versteegh and Zhu Sha

#### Abstract

This project is designed to investigate Sir Francis Galton's statistical dice experiment. We constructed Galton's dice according to his prescriptions and tested the null hypothesis that the outcomes from these dice do indeed follow a discrete approximation to the normal distribution with median error one. The inverse distribution function sampler and Chi Squared test are the statistical methodologies employed in this project.



#### 9.3.1 Introduction and Motivation

The report will firstly cover the background and motivation of this project. Secondly, the methodologies used will be explained before outlining the results and subsequent conclusion found by undertaking this experiment. Finally, a potential modification to Galton's method will be examined as a means of sampling from a standard normal distribution.

#### Francis Galton

Born in 1822, Francis Galton was considered by many, at an early stage, to be a child prodigy. By the age of two, he could read; at five, he already knew some Greek, Latin and long division.

After his cousin, Charles Darwin, published *The Origin of Species* in 1859, Galton became fascinated by it and thus devoted much of his life to exploring and researching aspects of human variation. Galton's studies of heredity lead him to introduce the statistical concepts of regression and correlation. In addition to his statistical research, Galton also pioneered new concepts and ideologies in the fields of meteorology, psychology and genetics.

#### Statistical Dice

This experiment came about from Galton's need, as a statistician, to draw a series of values at random to suit various statistical purposes. Dice were chosen as he viewed them to be superior to any other randomisation device. Cards and marked balls were too tedious to be continually shuffled or mixed following each draw, especially if the required sample size was large.

The dice he created made use of every edge of each face which allowed for 24 equal possibilities as opposed to the six of a normal die.

For further details on Galton's experiment, please refer to his article "Dice for Statistical Experiments"; *Nature* (1890) No 1070, Vol 42 (This article is available free for download. Please refer to the references section for the website.)

## Motivation

The motivation behind this project is to reconstruct Galton's dice using the methods outlined in his 1890 *Nature* article "Dice for Statistical Experiments" and then harness the power of modern computers to determine how effective this technique was for simulating random numbers from the following distribution.

Galton outlines that the samples were taken from a normal distribution with mean zero and median error one. We shall call this distribution Galton's Normal distribution or GN. However, for the experiment to work, we must use a discrete approximation of the normal distribution, which we will define as Galton's Discrete Normal or GDN. Both will be formally explained in the Methodology section.

To determine the success of this experiment, we formulate the following question as a statistical hypothesis test: "Are our sampled values taken independently and identically from an appropriate discrete distribution which approximates Galton's normal distribution?"

## Materials and Methods

### Experiment Process

In order to recreate Galton's Dice Experiment, we have chosen to replicate the design he explains in his *Nature* article.

### Creating the Dice

We chose to use rimu as it was readily available and inexpensive, unlike the mahogany that Galton had access to. As per his specifications, the wood was cut into six cubes of 1.25 inches (3.2 cm) wide, high and deep, before being covered in a paper template that was designed to fit tightly around the wood. The paper was adhered using general PVA glue.

The only change to Galton's original specification was that we chose to write the values to two decimal places on the faces, as opposed to one decimal place. This was to ensure a higher level of precision when plotting the results.

### Collecting the Data

The experiment was carried out by shaking all of the first three dice (dice 1) at once and rolling them across the flat surface of a table top. We interpreted Galton's terminology of the values that "front the eye" to be the results that one can see by looking directly down on top of the dice. The three dice were then lined up into a row and the values called out and entered onto a Notepad document. We used the following formula to calculate the optimal number of trials needed for our investigation:  $f(x)_{min} * sample\ size \approx 5$ , where  $f(x)_{min}$  is the smallest probability for the discrete distribution.

The same rolling process was then performed for dice 2 (two dice at once) and 3 (only one die) with the single exception that we did not need to roll these dice as many times as dice 1.

#### 9.3.2 Statistical Methodology

Firstly, we will define Galton's Normal distribution. As derived from an article published in *Statistical Science*<sup>1</sup>, Galton's Normal Distribution has a mean of zero but the variance is not one. Instead,

---

<sup>1</sup>Stochastic Simulation in the Nineteenth Century. *Statistical Science* (1991) Vol 6, No 1, pg 94.

Galton's sample is taken from a half-normal distribution with a “probable error” (median error) of one. This implies that the probability between zero and one is a quarter, allowing us to solve the following equation to determine the variance:

$$\begin{aligned}\phi(x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \\ \frac{1}{4} &= \int_0^1 \phi(x) dx \\ \frac{1}{4} &= \int_0^1 \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \\ \sigma &= 1.4826\end{aligned}$$

$$\therefore \text{GN} \sim N(0, 1.4826^2)$$

Secondly, we must determine how Galton calculated the values<sup>2</sup> to use on his dice. It was our assumption that he used the midpoints of a set of intervals that partition  $[0, 1]$  and we undertook the following processes to confirm this.

We divided the interval  $[0, 1]$  equally into 24, with the last 3 intervals further divided into 24 subintervals. In total, this gave us 21 + 24 intervals to allocate along the y-axis. The midpoint of each interval was taken in order to compute its corresponding  $x$  value under the inverse CDF map.

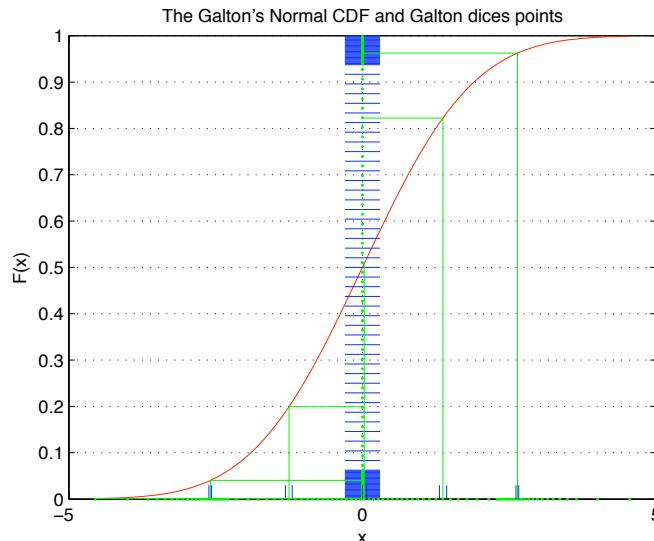


Figure 9.4: Plot showing the midpoints mapping back to specific values on the x axis.

The easiest way to do this would have been to evaluate the inverse CDF function at the midpoints. However, a closed form expression for the inverse CDF does not exist for a Normal distribution. Thus, we applied numerical methods to solve for  $x$  (Newton's method).

We believe the midpoint assumption was correct, as the mapped values are very close to Galton's actual figures and the differences can be attributed to an imprecise value for the standard deviation.

Thirdly, we can now determine Galton's discrete approximation to the Normal. This is necessary as the values drawn from throwing Galton's dice come from a discrete distribution, not the continuous Galton Normal. In doing this, we are also able to define our null hypothesis formally:  $H_0 : x_1, x_2, \dots, x_n \text{ IID } \sim \text{GDN}$  Galton's Discrete Normal (GDN) is an approximation to Galton Normal (GN).

Fourthly, as the distribution is now discrete, we can apply the Chi Squared Test to evaluate our null hypothesis. The test used had the following parameters: Degrees of Freedom:  $90 - 1 = 89$   $\alpha = 0.05$ ; Critical Value = 112.

---

<sup>2</sup>See Appendix B.

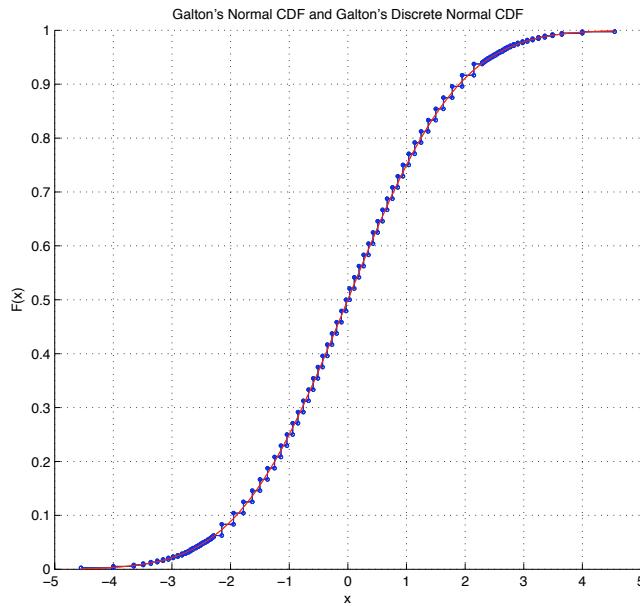


Figure 9.5: Plot showing both the GN and GDN CDFs. They are very similar.

### 9.3.3 Results

Once the experiment was complete and the results collated, they were run through a methodological tester to ensure all values were correct. Testing the data involved running all our sampled values through a Matlab function which checked each number against Galton's 45 possible values. Any values that did not match were outputted as ones and the erroneous data were removed before a graph was plotted to measure how well our experiment sampled from GDN.

#### Chi Squared Test

A Chi Squared test was then performed on the data and the results<sup>1</sup> are summarised below.

	Data Values	Observed Count	Expected Count	$(O - E)^2/E$
	-4.55	5	5.046875	0.000435372
	-4	7	5.046875	0.755853328
	-3.65	5	5.046875	0.000435372
	3.49	6	5.046875	0.180001935
	...	...	...	...
	3.65	8	5.046875	1.727989551
	4	11	5.046875	7.022107198
	4.55	5	5.046875	0.000435372
Total		1938	1938	
Chi Test Result				83.54798762

$$T = \sum_{i=1}^{90} \frac{(Observeverd - Expected)^2}{Expected} = 83.548$$

<sup>1</sup>For the full table, please see Appendix A.

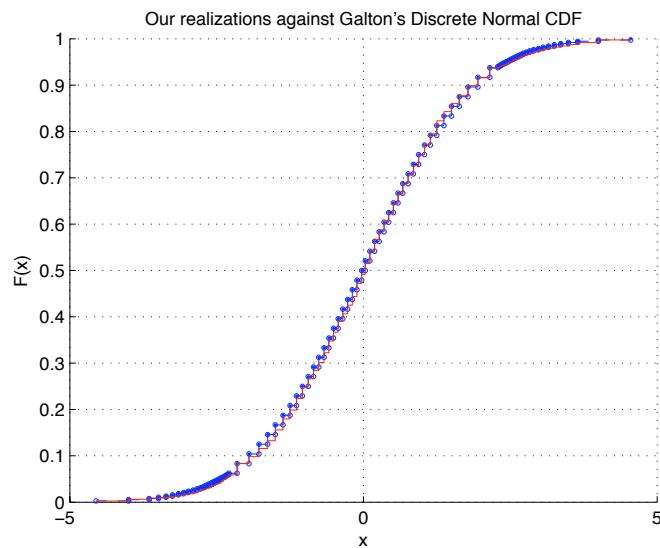
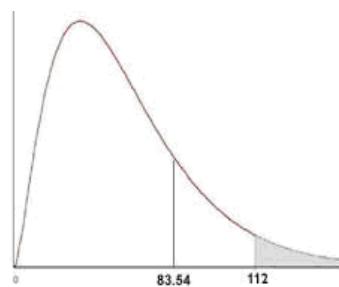


Figure 9.6: Plot showing the empirical DF of our results against GDN. Our values take on a staircase appearance and are very close to GDN. The main deviations occur mostly in the tails.



### 9.3.4 Conclusion

We cannot reject  $H_0$  at  $\alpha = 0.05$  because the observed test statistic is outside the rejection region. In relation to our statistical question, this means that there is insufficient evidence to suggest that our sample is not from GDN.

### Potential Modification

Since the standard normal distribution is more common in all areas, we wanted to convert Galton's Dice into a new set which can be used for simulating the standard normal distribution.

In his experiment, Galton took the mid-point of each probability interval, and then found the corresponding  $x$ -values. Instead of applying a tedious calculation to find the  $x$ -values, we took a  $z$ -value table, and found the corresponding  $z$ -values to the upper bound of those intervals. This enables the creation of two new dice<sup>2</sup>:

Dice (1)	0.05	0.10	0.15	0.21	0.27	0.32
	0.37	0.43	0.49	0.55	0.61	0.67
	0.74	0.81	0.89	0.97	1.05	1.15
	1.26	1.38	1.53	*	*	*
Dice (2)	1.56	1.58	1.60	1.62	1.65	1.68
	1.70	1.73	1.76	1.79	1.83	1.86
	1.90	1.94	1.99	2.04	2.09	2.15
	2.23	2.31	2.42	2.56	2.80	4.00

Through Matlab, we were able to map the data gathered during our original experiment into the values shown in previous table, corresponding to the standard Normal, and develop the following plot:

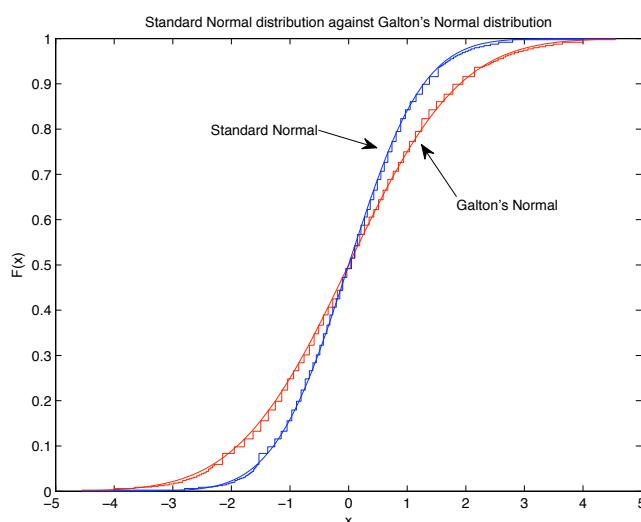


Figure 9.7: Plot showing the Standard Normal Distribution against Galton's Normal Distribution.

---

<sup>2</sup>Tables showing the new values for dice 1 & 2. The third dice can remain the same as Galton's.

## Author Contributions

**Brett** - Constructed dice, gathered majority of the data results, constructed report, conducted spell/grammar check.

**Joe** - Wrote up **Matlab** code to analyse and plot data, entered in data results, constructed presentation and discovered a modification to Galton's experiment.

## References

Dice for Statistical Experiments. *Nature* (1890) Vol 42, No 1070

Stochastic Simulation in the Nineteenth Century. *StatisticalScience* (1991) Vol 6, No 1

<http://www.galton.org>

<http://www.anu.edu.au/nceph/surfstat/surfstat-home/tables/normal.php>

## Appendix A

Data	Count	Expected Count	$(O - E)^2/E$	Data	Count	Expected Count	$(O - E)^2/E$
-4.55	5	5.046875	0.000435372	0.11	53	40.375	3.947755418
-4	7	5.046875	0.755853328	0.19	48	40.375	1.44001548
-3.65	5	5.046875	0.000435372	0.27	40	40.375	0.003482972
-3.49	1	5.046875	3.245017415	0.35	35	40.375	0.715557276
-3.36	3	5.046875	0.830156734	0.43	32	40.375	1.737229102
-3.25	3	5.046875	0.830156734	0.51	42	40.375	0.065402477
-3.15	3	5.046875	0.830156734	0.59	41	40.375	0.009674923
-3.06	4	5.046875	0.217153638	0.67	46	40.375	0.783668731
-2.98	4	5.046875	0.217153638	0.76	35	40.375	0.715557276
-2.9	3	5.046875	0.830156734	0.85	38	40.375	0.139705882
-2.83	8	5.046875	1.727989551	0.94	45	40.375	0.529798762
-2.77	5	5.046875	0.000435372	1.04	44	40.375	0.325464396
-2.72	3	5.046875	0.830156734	1.14	43	40.375	0.170665635
-2.68	3	5.046875	0.830156734	1.25	55	40.375	5.297600619
-2.64	3	5.046875	0.830156734	1.37	38	40.375	0.139705882
-2.59	6	5.046875	0.180001935	1.5	35	40.375	0.715557276
-2.55	4	5.046875	0.217153638	1.63	32	40.375	1.737229102
-2.51	5	5.046875	0.000435372	1.78	42	40.375	0.065402477
-2.47	4	5.046875	0.217153638	1.95	33	40.375	1.347136223
-2.43	6	5.046875	0.180001935	2.15	40	40.375	0.003482972
-2.39	10	5.046875	4.861116486	2.29	3	5.046875	0.830156734
-2.35	6	5.046875	0.180001935	2.32	4	5.046875	0.217153638
-2.32	8	5.046875	1.727989551	2.35	3	5.046875	0.830156734
-2.29	6	5.046875	0.180001935	2.39	4	5.046875	0.217153638
-2.15	47	40.375	1.087074303	2.43	6	5.046875	0.180001935
-1.95	28	40.375	3.792956656	2.47	5	5.046875	0.000435372
-1.78	34	40.375	1.006578947	2.51	3	5.046875	0.830156734
-1.63	33	40.375	1.347136223	2.55	8	5.046875	1.727989551
-1.5	45	40.375	0.529798762	2.59	1	5.046875	3.245017415
-1.37	46	40.375	0.783668731	2.64	7	5.046875	0.755853328
-1.25	37	40.375	0.282120743	2.68	5	5.046875	0.000435372
-1.14	48	40.375	1.44001548	2.72	4	5.046875	0.217153638
-1.04	48	40.375	1.44001548	2.77	4	5.046875	0.217153638
-0.94	35	40.375	0.715557276	2.83	6	5.046875	0.180001935
-0.85	34	40.375	1.006578947	2.9	5	5.046875	0.000435372
-0.76	34	40.375	1.006578947	2.98	5	5.046875	0.000435372
-0.67	41	40.375	0.009674923	3.06	6	5.046875	0.180001935
-0.59	49	40.375	1.84249226	3.15	5	5.046875	0.000435372
-0.51	37	40.375	0.282120743	3.25	4	5.046875	0.217153638
-0.43	44	40.375	0.325464396	3.36	5	5.046875	0.000435372
-0.35	33	40.375	1.347136223	3.49	6	5.046875	0.180001935
-0.27	36	40.375	0.474071207	3.65	8	5.046875	1.727989551
-0.19	36	40.375	0.474071207	4	11	5.046875	7.022107198
-0.11	55	40.375	5.297600619	4.55	5	5.046875	0.000435372
-0.03	38	40.375	0.139705882	Total	1938	1938	
0.03	45	40.375	0.529798762	Chi <sup>2</sup> Result			83.54798762

**Appendix B**

Table 1				Table 2				Table 3			
0.03	0.51	1.04	1.78	2.29	2.51	2.77	3.25	++++	+++	-++	+-+
0.11	0.59	1.14	1.95	2.32	2.55	2.83	3.36	+++-	+--	-+-	+-
0.19	0.67	1.25	2.15	2.35	2.59	2.90	3.49	++-+	-+++	--+	-++
0.27	0.76	1.37	*	2.59	2.64	2.98	3.65	++-	-++-	---	-+-
0.35	0.85	1.50	*	2.43	2.68	3.06	4.00	+--+	-+-+	+++	-+
0.43	0.94	1.63	*	2.47	2.72	3.15	4.55	+---	-+-	++-	---

## 9.4 Testing the average waiting time for the Orbiter Bus Service

J Fenemore and Y Wang

### Abstract

The Metro-owned and operated Orbiter bus service in Christchurch city is a very popular service that links up some of Christchurch's main suburbs, places and attractions. The timetable provided by the Metro bus company claims that on weekdays between 6 a.m. and 7 p.m., a service will arrive at any given stop every ten minutes, regardless of whether that service travels clockwise or anticlockwise. I hypothesise that this is not the case and that arrivals are influenced by many other factors including current traffic volume, traffic accidents, pedestrian volume, traffic light stoppages and passenger boarding times. We tested this hypothesis by sitting at the UCSA bus stops and recording arrival times.



#### 9.4.1 Motivation

The Orbiter is a highly used bus service and I myself often use this service. Many times while waiting for the service, I have noticed that more often than not, two Orbiter buses arrive at the stop at the same time or within a very short time of each other. Because of logistical reasons, I believe the Metro bus company would not run more buses than needed, meaning that if two buses arrived 'back to back' then there would be a twenty minute wait for the next bus (as the waiting time should be only ten minutes, so for two buses, the time is doubled.) This type of scenario significantly affects the times specified by Metro. For this reason, I believe that in reality, the average waiting time/arrival time is not ten minutes. It is important to note that the timetables distributed by Metro give specific times when buses arrive. These times are all ten minutes apart, which I feel can only be interpreted as meaning that a bus will arrive at a stop every ten minutes and the maximum waiting time for a passenger is also ten minutes. So for the two buses arriving in the 'back to back' situation, while the average time of arrival is presented as every ten minutes on paper, in reality, the buses do not arrive specifically ten minutes apart as claimed. This circumstance also gives a variation of ten minutes and decreases the probability of actually waiting only ten minutes. I wish to address this issue of the average waiting times of the buses in relation to the timetables provided and the variation in actual arrivals. Therefore, by examining the arrival times of the buses and recording waiting times, it can be examined just how accurate the timetables given are and whether they are based on average times or specific times. These issues affect Metro's reliability and credibility.

### 9.4.2 Method

The experiment we carried out is relatively simple. We sat at the bus stop outside the UCSA building on Ilam road and recorded the arrival times of each Orbiter bus and then calculated the waiting times between each bus. This was done for both clockwise and anticlockwise directions. The waiting time for the first bus in both directions was taken from the time of our arrival to the stop. After that, the waiting time was calculated as the times between bus arrivals.

A range of times were recorded, which covered an entire working day - 8 a.m. to 5 p.m.. These times were recorded on different days to assess not only the time of day but also different days, so we could see how these differences affect the times. The different times give a fairer assessment of the waiting times. It was assumed that for each day of the week, the waiting times for specific times of the day are relatively the same. A sample taken any day at a specific time would represent all days in the week at that time. The experiment was conducted in this manner because of availability and time restrictions.

While we realise that taking more samples would increase accuracy and reliability while also giving a better description of actual events, we felt it impractical to sit at the stop and record times for an entire day for each day of the week.

### Statistical Methodology

For the experiment, we modelled the distribution of the inter-arrival times or waiting times of the Orbiter bus service using the exponential distribution. The probability distribution function is as shown below. The distribution is continuous.

$$f(x; \lambda) = \lambda * \exp(-\lambda * x)$$

Where:  $x$  is the waiting time, and  $\lambda$  is the rate parameter or  $1/\text{mean}(x)$ .

The mean of this distribution is  $1/\lambda$  and has variance  $1/\lambda^2$ .

The exponential distribution was chosen because of its important memory-less property. Each new waiting time for the next bus is completely independent of the past waiting times. Each bus's arrival is assumed to be independent of the last.

For this experiment, I will be testing whether the average waiting time is ten minutes. More formally:

$H_0$ (null hypothesis):  $\mu = 10$  minutes

$H_A$ (Alternative hypothesis):  $\mu \neq 10$  minutes

To test this hypothesis, we used non-parametric bootstrap methods to estimate  $\lambda$  and obtain a 95% confidence interval for this value. These values will be formed by sampling the data observed with replacement, at equal probabilities, 132 times, of which an average will be taken. The whole process was then repeated 1000 times. An overall average calculated  $\lambda$  will be then transformed into an average waiting time using the formula:

$$\mu = 1/\lambda$$

where  $\mu$  is the average.

This will then be compared and contrasted against the average waiting time found by generating 132 realisations of waiting times then calculating the average of these, then repeating this process 1000 times. This is a parametric bootstrap based technique. For this,  $\lambda = 1/10$  (where  $\mu = 10$  minutes and using the formula above.) An overall average will be found along with a 95% confidence

interval for this value. By comparing these intervals and mean values, an accurate decision will be made as to whether buses do arrive on average every ten minutes or not.

Probabilities of certain arrival times around ten minutes will be evaluated to show the accuracy of the service.

The **Matlab** code for this process is given in Appendix III.

#### 9.4.3 Results

The raw data is given in Appendix IV.

The average waiting time for the anticlockwise direction = 9.19 mins.

The average waiting time for the clockwise direction = 8.95 mins.

The total average = 9.07 mins.

The minimum waiting time = 0 mins.

The maximum waiting time = 28 mins.

There are 66 waiting time samples for each direction.

*Notes for the data:*

- Some buses waited at the stop for random amounts of time in order to space the buses apart (this was never for long: 1 or 2 minutes). This was not taken into account when recording arrival times.
- School rush traffic (heavier volumes) was present from 3 p.m. to 3.30 p.m. approx.
- Evening commuter rush was present from approx 4.30 p.m. onwards.
- Morning commuter rush was from 8 a.m. to 9.30 p.m. approx.

*Observations on the data:* The anticlockwise direction tends to be much more consistent, having a closer average to ten minutes and more observed times close to ten minutes.

#### 9.4.4 Discussion

During less busy hours, buses arrive much more regularly.

The results of the code in (Appendix III) are as follows:

Calculated sample  $\lambda = 0.1105$ . The calculated 95% confidence interval for this value is [0.1078, 0.1129]. The claimed lambda is  $\lambda = 0.1$ . As the calculated sample is within this interval and the claimed  $\lambda$  is below, we can see that the bus arrival is slightly less than claimed (using  $\mu = 1/\lambda$ ).

Using the claimed  $\lambda$ , the randomly obtained mean value for waiting time is  $\mu = 9.9842$ . The calculated 95% confidence interval for the mean waiting time is [9.2882, 10.6237].

I found from the samples that the anticlockwise, clockwise and total mean waiting times are  $\mu = 9.19$ , 8.95 and 9.07, respectively. None of these values is within the interval previously stated.

When the calculated sample  $\lambda$  of 0.1105 was used to produce the mean waiting time and its 95% confidence interval, the following was produced:  $\mu=9.0350$  and [8.4957, 9.6137].

It is important to note that it is seen in the graph, from the empirical CDF, that the probability of having short waiting times is high - the probability of waiting 10 minutes or less according to our observed data is 6288. This value is quite high but the probability of waiting 15 minutes or more is 0.1288 which, in reality, is relatively high also. Practically, this means 1 in every 10 times you wait for an Orbiter to arrive, it will take 15 minutes or more to come. This value may be acceptable by Metro and indeed is good, considering so many unknown factors in traffic, but it would surely frustrate passengers being 5 minutes behind time.

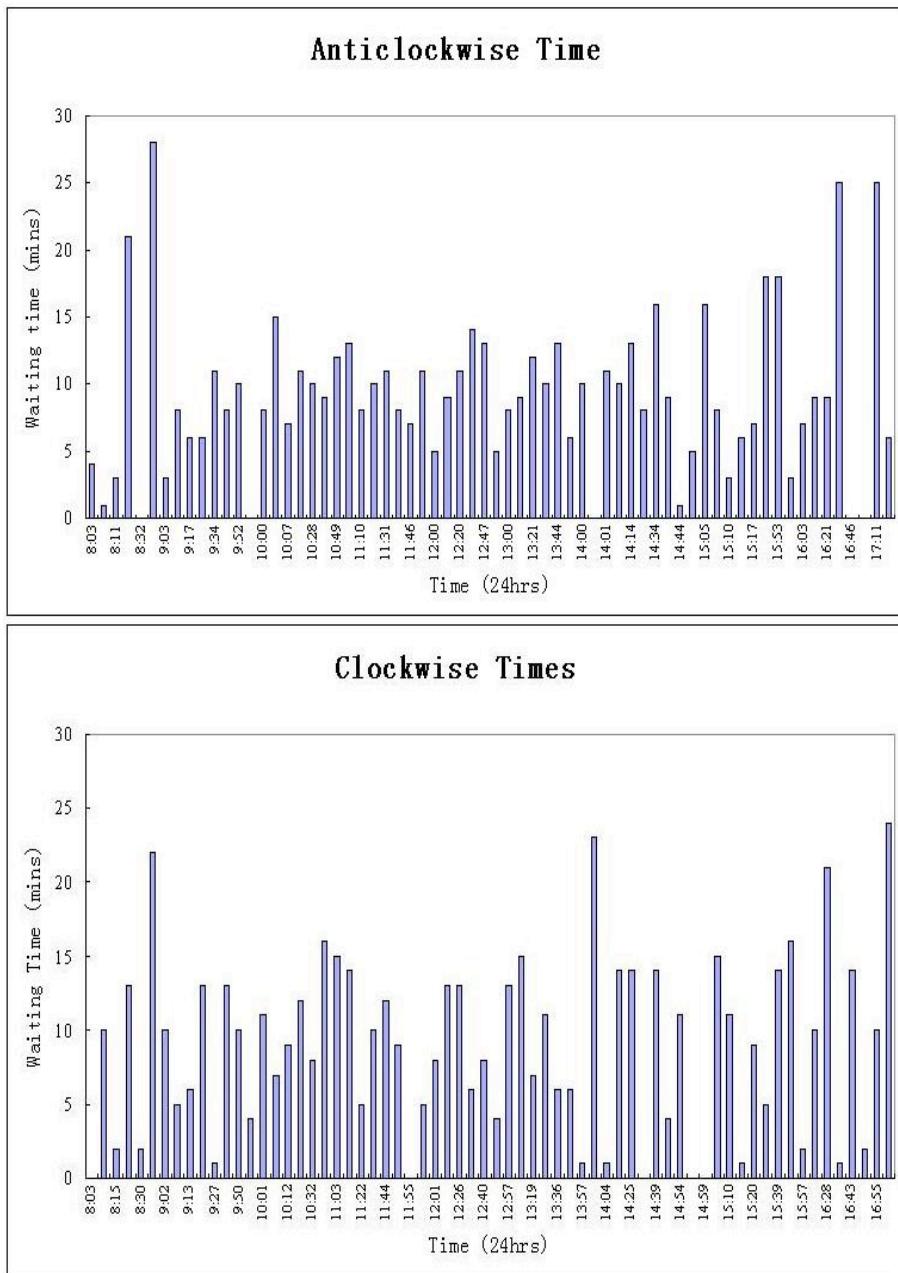


Figure 9.8: From the graphs above, we can see that often, a short wait is followed by a long wait, in both directions. Also, the anticlockwise times are generally much closer to 10 minutes waiting time. It is also seen that around rush hour times (8:30, 15:00, 16:45), a pattern emerged where several buses in quick succession were followed by a long wait for the next bus to arrive. This could be because of the time taken for more passengers than usual to aboard and depart, and areas where traffic volume is greater at these times.

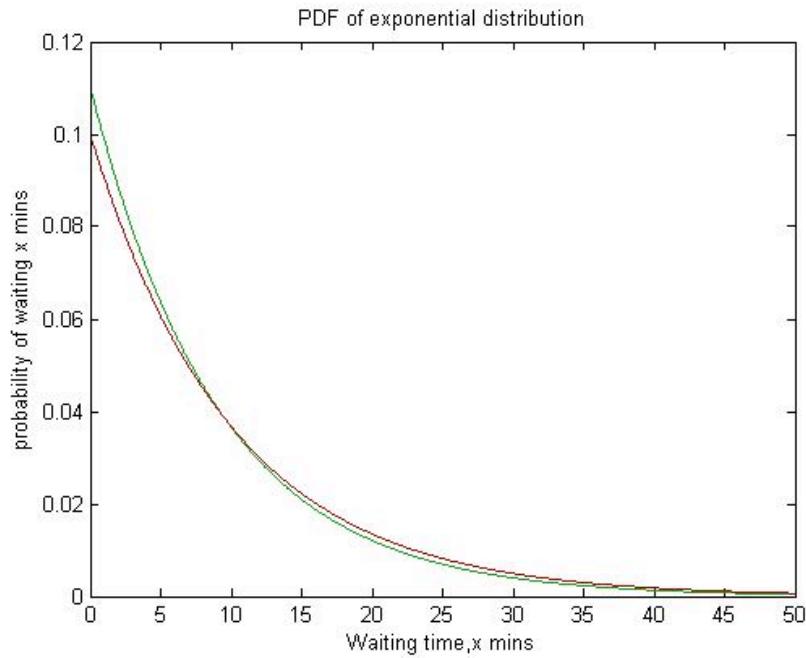


Figure 9.9: This graph shows the probability distribution function for the exponential function with the green line indicating a  $\lambda$  value of 0.1, the claimed  $\lambda$ . The red line indicates the value of  $\lambda$  estimated, 0.1105. From this graph, you can see the probability of getting a short waiting time is high - approximately 0.06, while the probability of a long waiting time is much much lower - approximately 0.01. The **Matlab** code for this graph is shown in Appendix I.

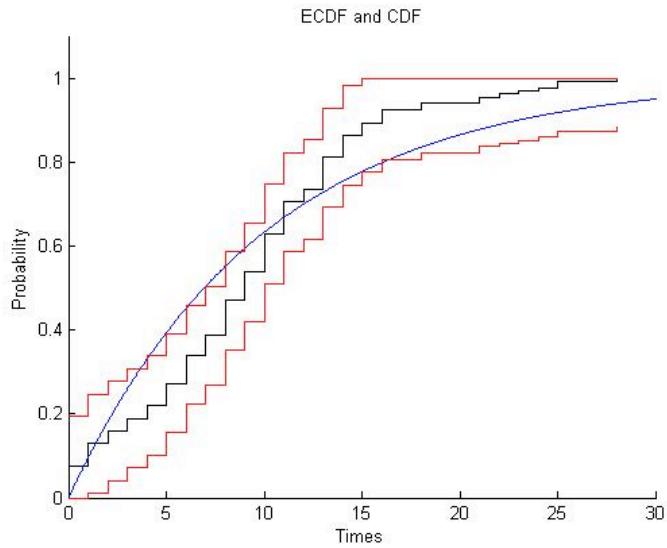


Figure 9.10: This plot is the Empirical CDF plot (black), with a 95% confidence interval (red) and the actual CDF based on claimed  $\lambda = 0.1$  (blue). The **Matlab** code for this graph is given in Appendix II. This graph shows the accuracy of the empirical distribution and hence the accuracy of the data we collected. There are some inconsistencies caused by the randomness of inter-arrival times but our empirical CDF is generally good as the actual CDF lies mostly within the interval lines. With more data points, our accuracy would greatly improve.

### 9.4.5 Conclusion

The calculated value of  $\lambda$  is 0.1105. When this value is used to estimate the mean waiting time, and including the observed waiting times, we can conclude that Metro delivers a service better than claimed. From this  $\lambda$ , we see a mean waiting time of 9.04 minutes - 58 seconds less than the 10 minutes wait claimed.

Furthermore, the mean waiting time estimate calculated and its 95% confidence interval (not including the average waiting times observed) cause us to not accept the null hypothesis,  $H_0$  of  $\mu = 10$  minutes at the 95% confidence level.

From all of this, we can confirm that Metro is quite right in claiming an arrival of an Orbiter bus every ten minutes at any stop. In fact, it appears that they do better than this by a whole minute. However, it is all very well to claim this on paper but it is crucial to note that waiting times of 28 minutes do happen, rarely. This illustrates a very important difference between practical and statistical significance. In this case, it has no major effects, as the observed waiting time is less than claimed.

## Author Contributions

*Josh's Contributions:* The original concept; data recordings for Wednesday, Thursday and Friday (5hrs); data organisation; analysis; methodology and implementation and the preliminary report, final report and presentation notes. *Yirang's Contributions:* Monday's and Tuesday's data recordings (4hrs).

## Appendices

### I

```
x=linspace(0,50,1000);%Array of x points to evaluate
lambda1=0.1105%Estimated lambda
f1=lambda1*exp(-lambda1.*x);%Calculated probabilities
plot(x,f1,'color',[0 0.6 0])%Plot coloured red
hold
lambda2=0.1%Claimed lambda
f2=lambda2*exp(-lambda2.*x);%Calculated probabilities
plot(x,f2,'color',[0.6 0 0])%Plot coloured green
xlabel('Waiting time,x mins')%Graph titles
ylabel('probability of waiting x mins')
title('PDF of exponential distribution')
```

### II

```
antiTimes=[8 3 7 18 18 3 7 9 9 25 0 0 25 6 ...
10 0 10 8 16 9 1 5 16 6 4 1 3 21 0 28 3 8 ...
6 6 11 8 10 15 0 8 7 11 10 9 12 13 8 10 11 8 ...
7 11 5 9 11 14 13 5 8 9 12 10 13 6 11 13];
clockTimes=[0 0 11 1 9 5 14 16 2 10 21 1 14 2 10 ...
24 6 1 14 14 0 14 4 11 15 0 10 2 13 2 22 ...
10 5 6 13 1 13 10 11 4 7 9 12 8 16 15 14 5 ...
10 12 9 8 0 5 13 13 6 8 4 13 15 7 11 6 23 1];
%The raw data-the waiting times for each direction
```

```

sampleTimes=[antiTimes clockTimes];%dd all times into 1 array

x=linspace(0,30,1000);%Create array
lambda1=0.1;%Set claimed lambda
f=1-exp(-lambda1*x);%Create cdf realisations based on claimed lambda

[x1 y1]=ECDF2(sampleTimes,7,0,0);
%Call to class distributed ECDF fuction, save output values in arrays x1
%and y1
hold on%Hold plots for superimposition
plot(x,f)

Alpha=0.05;%set alpha to 5%
SampleSize=132;

Epsn=sqrt((1/(2*SampleSize))*log(2/Alpha));%epsilon_n for the confidence band

stairs(x1,max(y1-Epsn,zeros(1,length(y1))),’r’);%lower band plot
stairs(x1,min(y1+Epsn,ones(1,length(y1))),’r’);%upper band plot
hold off
axis([0,30,0,1])
title(’ECDF and CDF’)
xlabel(’Times’)
ylabel(’Probability’)

```

### III

```

clear
antiTimes=[8 3 7 18 18 3 7 9 9 25 0 0 25 6 ...
10 0 10 8 16 9 1 5 16 6 4 1 3 21 0 28 3 8 ...
6 6 11 8 10 15 0 8 7 11 10 9 12 13 8 10 11 8 ...
7 11 5 9 11 14 13 5 8 9 12 10 13 6 11 13];
clockTimes=[0 0 11 1 9 5 14 16 2 10 21 1 14 2 10 ...
24 6 1 14 14 0 14 4 11 15 0 10 2 13 2 22 ...
10 5 6 13 1 13 10 11 4 7 9 12 8 16 15 14 5 ...
10 12 9 8 0 5 13 13 6 8 4 13 15 7 11 6 23 1];
%The raw data - the waiting times for each direction

rand(’twister’,489110);%set the seed for rand so results can be reproduced
sampleTimes=[antiTimes clockTimes];%All the sample times collected
lambdaTotal=zeros(1000,1);%An empty array

lambdaObs=1/mean(sampleTimes)%The lambda value for observed samples
lambdaClaimed=1/10 %The lambda claimed by Metro

%This is a non-parametric bootstrap
for j=1:1000%Loop to create 1000 lambdas
    for i=1:132 %A loop to sample with replacement 132 times at equal
        %probability
        u1=rand;%Generate a random number
        x1=deMoivreEqui(u1,132);%Select a random number between 1 and 132
        b(i)=sampleTimes(x1);%Array of random sample times, taken from
        %all samples, using random number generated
    end
    lambdaTotal(j)=1/mean(b);%lambda value for each array of random samples
end

sampleLambda=mean(lambdaTotal)
%The mean lambda for all the lambdas calculted
sortedLambdaTotal=sort(lambdaTotal);%Sort lambdas generated
lowerBound=lambdaTotal(25)%Calculate a 95% confidence interval for lambda
upperBound=lambdaTotal(975)

realisationsClaimed=zeros(1000,1);%An empty array
meanClaimed=zeros(1000,1);%An empty array
%This is parametric bootstrap
for x=1:1000%Loop to create 1000 mean waiting times based on claimed lambda

```

```

for z=1:132 %Loop to generate 1000 waiting times based on claimed lambda
    u2=rand;%Generate a random number
    realisationsClaimed(z)=-(1/lambdaClaimed)*log(u2);
    %Create realisation of x, random number u and lambda claimed
end
meanClaimed(x)=mean(realisationsClaimed);
%Find mean of each array of realisations
end
meanOfClaim=mean(meanClaimed)%Overall mean of realisations created
meanClaimed=sort(meanClaimed);%Sort array
lowerBound=meanClaimed(25)
%Create a 95% confidence interval for the mean found
upperBound=meanClaimed(975)

```

The above code was written 15/10/07 by J Fenemore and makes use of the following function:

```

function x = deMoivreEqui(u,k);
%
% return samples from deMoivre(1/k,1/k,...,1/k) RV X
%
% File Dates : Created 08/06/07 Modified 08/06/07
% Author(s) : Raaz
%
% Call Syntax: x = deMoivreEqui(u,k);
%               deMoivreEqui(u,k);
%
% Input      : u = array of uniform random numbers e.g. 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
% floor is useful when the outcomes are {0,1,...,k-1}
%x = floor(k * u);
%%%%%%%%%%%%%%%

```

#### IV. Raw data

<b>Anti-clockwise Route</b>		<b>Clockwise Route</b>	
Bus Arrival Times:	Mins waiting:	Bus Arrival Times:	Mins waiting:
15 : 07	8	14 : 59	0
15 : 10	3	14 : 59	0
15 : 17	7	15 : 10	11
15 : 35	18	15 : 11	1
15 : 53	18	15 : 20	9
15 : 56	3	15 : 25	5
16 : 03	7	15 : 39	14
16 : 12	9	15 : 55	16
16 : 21	9	15 : 57	2
16 : 46	25	16 : 07	10
16 : 46	0	16 : 28	21
16 : 46	0	16 : 29	1
17 : 11	25	16 : 43	14
17 : 17	6	16 : 45	2
		16 : 55	10
		17 : 19	24
14 : 00	10	13 : 56	6
14 : 00	0	13 : 57	1
14 : 10	10	14 : 11	14
14 : 18	8	14 : 25	14
14 : 34	16	14 : 25	0
14 : 43	9	14 : 39	14
14 : 44	1	14 : 43	4
14 : 49	5	14 : 54	11
15 : 05	16	15 : 09	15
15 : 11	6		
8 : 03	4	8 : 03	0
8 : 08	1	8 : 13	10
8 : 11	3	8 : 15	2
8 : 32	21	8 : 28	13
8 : 32	0	8 : 30	2
9 : 00	28	8 : 52	22
9 : 03	3	9 : 02	10
9 : 11	8	9 : 07	5

<b>Anti-clockwise Route</b>		<b>Clockwise Route</b>	
Bus Arrival Times:	Mins waiting:	Bus Arrival Times:	Mins waiting:
9 : 17	6	9 : 13	6
9 : 23	6	9 : 26	13
9 : 34	11	9 : 27	1
9 : 42	8	9 : 40	13
9 : 52	10	9 : 50	10
10 : 07	15	10 : 01	11
9 : 52	0	9 : 56	4
10 : 00	8	10 : 03	7
10 : 07	7	10 : 12	9
10 : 18	11	10 : 24	12
10 : 28	10	10 : 32	8
10 : 37	9	10 : 48	16
10 : 49	12	11 : 03	15
11 : 02	13	11 : 17	14
11 : 10	8	11 : 22	5
11 : 20	10	11 : 32	10
11 : 31	11	11 : 44	12
11 : 39	8	11 : 53	9
11 : 46	7	12 : 01	8
11 : 57	11		
12 : 00	5	11 : 55	0
12 : 09	9	12 : 00	5
12 : 20	11	12 : 13	13
12 : 34	14	12 : 26	13
12 : 47	13	12 : 32	6
12 : 52	5	12 : 40	8
13 : 00	8	12 : 44	4
13 : 09	9	12 : 57	13
13 : 21	12	13 : 12	15
13 : 31	10	13 : 19	7
13 : 44	13	13 : 30	11
13 : 50	6	13 : 36	6
14 : 01	11	13 : 59	23
14 : 14	13	14 : 04	1

## 9.5 Diameter of *Dosinia* Shells

Guo Yaozong and Shen Chun

### 9.5.1 Introduction and Objective

We collected some shells from New Brighton Pier that are commonly called *Dosinia anus* (Coarse Venus Shell). This species is a member of the class Bivalvia. Bivalvia lack a radula, and feed by filtering out fine particles of organic matter either from seawater (suspension feeders) or from surface mud (deposit feeders). In each case, food enters the mantle cavity in a current of water produced by cilia on the gills. Gills have a large surface area in relation to the size of the animal, and secrete copious amounts of slime-like mucus that not only traps the food particles but also acts as a lubricant for the passage of food to the mouth. In addition to having this feeding role, gills are the respiratory structures and are richly supplied with blood dorsally. Sexes are separate, although there is no external dimorphism. Gametes are shed into the seawater, where fertilisation occurs.



Unlike Venus shells from other parts of the world, this species has a flat disc-like shell. Found just below the low-tide mark along Brighton beach, it burrows just below the sand surface and feeds using two short, separate siphons. (*Life in The Estuary*, Malcolm B. Jones & Islay D. Marsden, Canterbury University Press, 2005).

Our objective was to test whether the diameters of *Dosinia anus* shells on the north side of New Brighton Pier are identically distributed to those found on the south side of the pier.

### 9.5.2 Materials and Methods

We collected shells along the New Brighton beach to the left (north) and right (south) of the pier. We walked and picked up all the shells we could see, except broken ones. In about two and a half hours, we collected about two buckets of shells from each side of the pier (i.e. two from the left and two from the right).

After washing, drying and classifying the shells, we found that 254 of them were *Dosinia anus*, 115 collected from north of the pier and 139 from the southern. Then we used mechanical pencils to sketch the outline of each shell onto graph paper and measured the diameter of each in units of millimetres. The way we measured them was from top to bottom (as shown below). After that, we entered the data into a computer, and estimated the empirical CDF as well as confidence bands.

#### Statistical Methodology

In order to test the null hypothesis that the shell diameters of our species are identically distributed on both sides of the pier, we applied the non-parametric permutation test.

By using the permutation test, we tested whether the absolute difference between the two sample means were significantly different from each other.

Step 1: Observe value:  $T = X(\text{left}) - X(\text{right})$

Step 2: Combine [ L1 L2 ..... L115 R1 ..... R139] '254 Data'

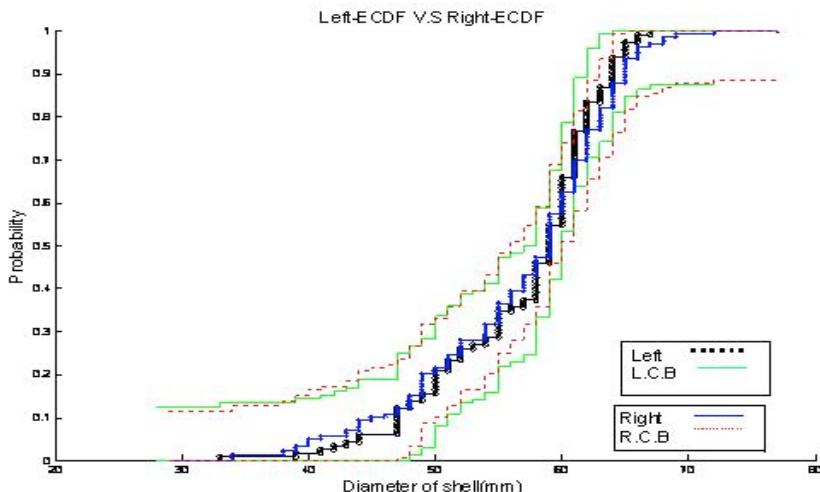
Step 3: Rearrange (MATLAB function: 'randperm'):

$$\begin{array}{c} [\text{R110}, \text{L45}, \text{L78}, \text{R2} \dots | \dots \dots \dots \text{L20}] \quad \text{'254 Data'} \\ \downarrow \qquad \qquad \qquad \downarrow \\ |\text{Mean}(1-115) \quad \quad \quad \text{--} \quad \text{Mean}(116-254)| = D_i(\text{recorded}) \\ (i = 1, 2, 3 \dots 10000) \end{array}$$

Step 4: Repeat Step 3 10000 times

Step 5: Find out how often ' $D_i$ ' is greater than ' $T$ ', then divided this value by **10,000**. This is our **P-value**.

### 9.5.3 Results



### Hypothesis testing

`abs('mean for north'-'mean for south'):  $|56.8173 - 56.6462| = 0.1711$ (observed value)`  $H_0$ : No difference can be observed between north and south  $H_a$ : A difference can be observed Alpha = 0.05

In the test, we found 8470 numbers were greater than 0.1711, so P-value =  $8470 / 10000 = 0.847$

### Conclusion

Since p-value is large, we do not reject the null hypothesis, as we do not have enough evidence to say that there is a difference in the distribution of *Dosinia anus* diameters between the north and south sides of the pier in New Brighton Pier.

### Author contributions

Shen Chun and Yaozong Guo did all the work together.

# **Index**

critical region, 266  
Law of Large Numbers, 268  
null hypothesis, 266  
success / failure rule, 274  
type I error, 266