

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

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

.

.

.

.

This work is licensed under the Creative Commons Attribution-Noncommercial-Share Alike 4.0 International License. To view a copy of this license, visit
<http://creativecommons.org/licenses/by-nc-sa/4.0/>.

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

Version Date: October 14, 2019

School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

*Department of Mathematics, Uppsala University, Uppsala, Sweden

*Laboratory for Mathematical Statistical Experiments, Uppsala Centre, and

Razeeh Samudin* and Dominic Lee†,

Probability Theory I

Answer (Ex. 5.2) —

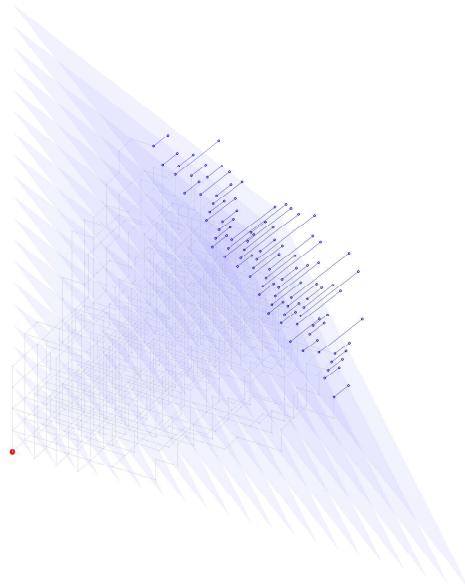
We want $1 - \alpha = 0.95$, and from the standard Normal Table we know that the corresponding $z_{\alpha/2} = 1.96$. Then we can get the right sample size n from the CLT implied Equation (61) in the lecture notes, which is,

$$n = \left(\sqrt{V(\bar{X}_1)} z_{\alpha/2} / \epsilon \right)^2 ,$$

as follows:

$$\begin{aligned} n &= \left(\sqrt{V(\bar{X}_1)} z_{\alpha/2} / \epsilon \right)^2 = \left((\sqrt{1/4} \times 1.96) / (1/10) \right)^2 \\ &= ((1/2) \times 1.96) / (1/10)^2 = (0.98 \times 10)^2 = 9.8^2 = 96.04 \end{aligned}$$

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, $\frac{\sqrt{n}(\bar{X}_n - \mathbf{E}(X_1))}{\sqrt{V(X_1)}} \rightsquigarrow Z \sim \text{Normal}(0, 1)$. 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 $\mathbf{P}(\bar{X}_n < 5.5)$.

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

(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 .

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

ASSESSMENT

Combination. The probability concept. Calculation of probabilities. Stochastic variable. Probability distributions. Independence and conditioned distributions. Expectation and variance. Central limit theorem. The law of large numbers. Practical examples of design of probability models.

CONTENT

- 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 variances 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.

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

SEE <https://www.unise/en/admissions/master/seema/kurssplan/?kp1d=38971&typy=1>.

Official Course Syllabus

Course Code: IMS034	Semester: Autumn 2019	Report Code: 10504	33%, DAG, NALI	Uppsala University	Probability Theory I, 5.0 e	2019-09-02 - 2019-11-03	12 = $\left\lceil \frac{33}{37.5} \right\rceil$ hours / week	work
Courses Syllabus: Friday, 4th of October 2019								
Course Coordinator: Razeezeh Samidunn								

Course Syllabus and Overview

Answer (Ex. 4.1) — This is nothing but the inverse sampling scheme for the standard Cauchy RV X .

Answer (Ex. 3.35) — Apply the formulas for the sample mean, \bar{X}_7 , and sample variance.

Thus, $\phi_{-Z}(t) = \phi_Z(t)$ 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 $z \in \mathbb{R}$.

$$(\tau)Z\phi = \zeta_{/\tau}{}^\partial = \zeta_{/(\tau \times 1^-)}{}^\partial = (\tau-)Z\phi = (\tau \times 1-)Z\phi_{\tau \times 0 \times \tau^\partial} = (\tau)\lambda\phi$$

We can use the facts by noting $Z = -Z = 0 + (-1) \times Z$, with $a = 0$ and $b = -1$ in $X = a + bZ$.

Answer (Ex. 3.54) —

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.

$$(\eta + \gamma)_{-\nu}{}^\partial (\eta + \gamma)^\partial = \eta_{-\nu}{}^\partial \eta + \gamma_{-\nu}{}^\partial \gamma = \eta_{-\nu}{}^\partial \eta^\partial \times \gamma_{-\nu}{}^\partial \gamma^\partial = (\eta) \lambda \phi \times (\gamma) X \phi = (\gamma) \lambda^+ X \phi = (\gamma) M \phi$$

ANSWER (Ex. 3.53) —

$$\cdot \chi = \bar{z}\chi - \chi + \bar{z}\chi = \bar{z}((X)\mathcal{E}) - (\bar{z}X)\mathcal{E} = (X)\mathcal{E}$$

Finally,

$$\begin{aligned} \gamma + \bar{z}\gamma &= ((1+\gamma)z\gamma)\frac{\bar{z}}{1} = ((i+\gamma)\gamma)\frac{\bar{z}}{1} = ((i+{}_0\partial\gamma) {}_{0+\gamma-{}_0\partial\gamma}\partial\gamma)\frac{\bar{z}}{1} = \\ &\stackrel{0=i}{=} \left[(i+0-{}_{n\partial\gamma}) {}_{n+\gamma-{}_n\partial\gamma}\partial\gamma \right] \frac{\bar{z}}{1} = \stackrel{0=i}{=} \left[(i\gamma + \gamma - {}_{n\partial\gamma}) \frac{p\bar{p}}{p} {}_{n+\gamma-{}_n\partial\gamma}\partial\gamma \right] \frac{\bar{z}}{1} = \\ &\stackrel{0=i}{=} \left[(n+\gamma-{}_{n\partial\gamma}) \frac{p\bar{p}}{p} \right] \frac{\bar{z}}{1} = \stackrel{0=i}{=} \left[({}_{n\partial\gamma}\gamma - {}_{n\partial\gamma}\gamma) \frac{p\bar{p}}{p} \right] \frac{\bar{z}}{1} = \stackrel{0=i}{=} \left[(t)X\phi \frac{\bar{z}p}{z\bar{p}} \right] \frac{\bar{z}}{1} = (\bar{z}X)E \end{aligned}$$

$$\begin{aligned} \gamma &= (\gamma\chi) \frac{i}{1} = \left(i\chi_{\gamma-\gamma^2} \right) \frac{i}{1} = {}^{0=i} \left[p^2 i\chi_{\gamma-n^2\gamma^2} \right] \frac{i}{1} = \\ {}^{0=i} \left[(\gamma-n^2\gamma) \frac{ip}{p} \right] \frac{i}{1} &= {}^{0=i} \left[\gamma - n^2\gamma^2 \frac{ip}{p} \right] \frac{i}{1} = {}^{0=i} \left[(i) X \phi \frac{ip}{p} \right] \frac{i}{1} = (X) E \end{aligned}$$

2. To find $V(X)$ using $\phi X(t)$ we need $E(X)$ and $E(X^2)$.

The second-last equality above is using $e_a = \sum_{x=0}^{\infty} \frac{a}{x!}$ with $a = Ae_n$.

$$\gamma_{-n} \varphi^\partial = {}_{n\partial} \chi^\partial \gamma^{-\partial} = \frac{1}{x} \sum_{n=0}^{\infty} \chi_n^\partial \gamma^{-\partial} = \frac{1}{x} \sum_{n=0}^{\infty} x^n \chi_n^\partial \gamma^{-\partial} = \gamma^{-\partial} \frac{1}{x} \sum_{n=0}^{\infty} x^n \chi_n^\partial = (\gamma x^n)^\partial \mathcal{I} = (\gamma) x^\partial \phi$$

1.

— Answer (Ex. 3.52) —

Time Table & Course Overview – In Progress

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

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

Lec.	Lab.	Week	Topics	Comprehension \leftrightarrow 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
*			Preliminaries: 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; EX try 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; EX try 2.5
06			Random variables	RD 3.1; EX 3.1; UD 1,41i, EX 3.2
			Discrete Random variables and IID Bernoulli Trials	RD 3.1, 3.2.1, 3.2.3
			Common Discrete Random variables	RD 3.2.3; UD 45; UD 4 ;EX 3.3;
07		38	Common Discrete Random variables	UD 46, 5, 47, 48, 6, 49, 50, 51
08			Common Discrete Random variables	EX 3.4 & EX try 3.3;
			Continuous RVs	RD 3.4; UD 52, 53, 54, RD 3.4.1
			Common Continuous RVs	RD 3.4.2; UD 55, 56; EX 3.18
			Common Continuous RVs	EX 57, 58; UD 59; EX try 3.5
09			Transformations of RVs – Discrete	RD 3.6; UD 60, 61; RW 3.6.1
10		39	Transformations of RVs – Discrete	RD 3.6.2, UD 62, 63, 64
			Transformations of RVs – Continuous (one-to-one & monotone)	RD 3.6.3; UD 65, 66, 67, 68
11			Transformations of RVs – Continuous (Direct method)	RD 3.6.3; UD 69, 70; EX try 3.7
			Expectations	RD 3.8, 3.8.1; UD 71, 72, 74, 75, 76, UD 77, 78, 79, 80; EX try 3.9
12		40	Multivariate Random Variables	RD 3.10; UD 83, 84, 85, 86, UD 87, 88, 89, 90, 91, 92
13				RD 3.10.2, 3.10.3; UD 95, 96
			Common \mathbb{R}^m -valued RVs	RD 3.10.4; UD 97, 98; EX 3.35, 3.36 try 3.11
14			Characteristic Functions	RD 3.12; UD 101, 102, 103, 104, 105; EX try 3.13
15		41	Statistics & Random Number Generation	RD 3.14, 4.2, 4.3
16			Statistics	EX try 3.15
17			Simulation – Inversion & Rejection Samplers	EX try 4.4
18		42	Convergence of RVs & Limit Laws	RD 5.1, UD 157, 158, 159, 160
19			Basic Inequalities & Law of Large Numbers	UD 161, 163, 164; EX 5.1
20			Law of Large Numbers & Central Limit Theorem via CFs	UD 165, 166, 167, 168; EX try 5.4
			Model exam with live-scribed solutions	

Table 2: Time Table for Inference Theory I

Lec.	Lab.	Week	Topics	Section/Labworks/Simulations

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

$$\frac{d}{dt}\phi_X(t) = \frac{d}{dt}\left(\frac{1}{3}(1 + e^{it} + e^{i2t})\right) = \frac{1}{3}(0 + ie^{it} + 2ie^{i2t}) = \frac{i}{3}(e^{it} + 2e^{i2t})$$

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

$$\begin{aligned} E(X) &= \frac{1}{i}\left[\frac{d}{dt}\phi_X(t)\right]_{t=0} = \frac{1}{i}\left[\frac{i}{3}(e^{it} + 2e^{i2t})\right]_{t=0} = \frac{1}{i}\frac{i}{3}(e^0 + 2e^0) = \frac{1}{3}(1 + 2) = 1 . \\ E(X^2) &= \frac{1}{i^2}\left[\frac{d^2}{dt^2}\phi_X(t)\right]_{t=0} = \frac{1}{i^2}\left[\frac{d}{dt}\frac{i}{3}(e^{it} + 2e^{i2t})\right]_{t=0} = \frac{1}{i^2}\left[\frac{i}{3}(ie^{it} + 4ie^{i2t})\right]_{t=0} \\ &= \frac{1}{3}(e^0 + 4e^0) = \frac{5}{3} . \end{aligned}$$

Finally,

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

Answer (Ex. 3.50) —

1.

$$\phi_X(t) = E(e^{itX}) = \sum_{x=0}^{\infty} e^{itx} \theta(1-\theta)^x = \sum_{x=0}^{\infty} \theta(e^{it}(1-\theta))^x = \frac{\theta}{1-e^{it}(1-\theta)} .$$

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. 3.51) —

$$\begin{aligned} \phi_X(t) &= E(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} f_X(x; a, b) dx = \int_a^b e^{itx} \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b e^{itx} dx = \frac{1}{b-a} \left[\frac{e^{itx}}{it} \right]_{x=a}^b \\ &= \frac{1}{(b-a)it} (e^{itb} - e^{ita}) . \end{aligned}$$

Contents

Time Table
Course Syllabus

1 Preliminaries

1.1 Elementary Set Theory

1.2 Exercises

1.3 Natural Numbers, Integers and Rational Numbers

1.4 Real Numbers

1.5 Introduction to MATLAB

1.6 Elementary Combinatorics

1.7 Array, Sequence, Limit,

1.8 Elementary Real Analysis

1.8.1 Limits of Real Numbers - A Review

1.9 Elementary Number Theory

1.10 Using the Multinomial($n = 10, \theta_1 = 0.6, \theta_2 = 0.3, \theta_3 = 0.08, \theta_4 = 0.02$) RV as our model

2 Probability Model

2.1 Experiments

2.2 Probability

2.2.1 Consequences of our Definition of Probability

2.2.2 Sigma Algebras of Typical Experiments*

2.3 Exercises in Probability

2.4 Conditional Probability

2.4.1 Bayes' Theorem

2.4.2 Independence and Dependence

2.5 Exercises in Conditional Probability

..... 56

..... 54

..... 50

..... 48

..... 47

..... 45

..... 43

..... 41

..... 39

..... 38

..... 36

..... 35

..... 33

..... 32

..... 30

..... 29

..... 28

..... 26

..... 24

..... 23

..... 22

..... 21

..... 20

..... 19

..... 18

..... 17

..... 16

..... 15

..... 14

..... 13

..... 12

..... 11

..... 10

..... 9

..... 8

..... 7

..... 6

..... 5

..... 4

..... 3

..... 2

..... 1

$$\phi(t) = E(e^{tX}) = \sum_{x \in \{0,1,2\}} e^{tx} f_X(x) = e^{0 \times 0} \times \frac{3}{1} + e^{1 \times 1} \times \frac{3}{1} + e^{2 \times 2} \times \frac{3}{1} = \frac{3}{1} (1 + e^{\mu} + e^{2\mu})$$

1. The CF of the discrete RV X is

Answer (Ex. 3.49) —

$$P((X_1, X_2, X_3, X_4) = (6, 2, 0, 0); n = 10, \theta_1 = 0.6, \theta_2 = 0.3, \theta_3 = 0.08, \theta_4 = 0.02)$$

$$= \frac{10!}{6! \times 2! \times 0!} \times 0.6^6 \times 0.3^2 \times 0.08^2 \times 0.02^0$$

$$= \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(6 \times 5 \times 4 \times 3 \times 2 \times 1) \times (2 \times 1) \times 1} \times 0.6^6 \times 0.3^2 \times 0.08^2 \times 1 \approx 0.03386$$

$$= \frac{(6 \times 5 \times 4 \times 3 \times 2 \times 1) \times (2 \times 1) \times 1}{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} \times 0.6^6 \times 0.3^2 \times 0.08^2 \times 1 \approx 0.03386$$

Using the Multinomial($n = 10, \theta_1 = 0.6, \theta_2 = 0.3, \theta_3 = 0.08, \theta_4 = 0.02$) RV as our model

Answer (Ex. 3.48) —

$$\mathbf{P}(X_{10} < 12.01) = P\left(\frac{X_{10} - \mathbb{E}(X_{10})}{\sqrt{0.001}} < \frac{\sqrt{0.001}}{\sqrt{0.001}}\right) = P\left(Z < \frac{0.0316}{\sqrt{0.01}}\right)$$

$$\approx \mathbf{P}(Z < -2.85) = 1 - \mathbf{P}(Z > 2.85) = 1 - \Phi(2.85) = 1 - 0.9978 = 0.0022$$

normal - we know that X_{10} is a Normal(12.1, 0.001) RV. Consequently, the probability of interest is

By the special property of Normal RVs - a linear combination of independent normal RVs is also distributed RVs.

Or by directly using the "formula" $\mathbb{V}(X_{10}) = V(X_1)/10$ for these 10 independently and identically

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

RVs. Similarly,

Or by directly using the "formula" $\mathbb{E}(\underline{X}_{10}) = \mathbb{E}(X_1) = 12.1$ for these 10 identically distributed

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

By property of Expectations and Variances for linear combinations

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

Let X_1, X_2, \dots, X_{10} denote the full volumes of 10 cans. The average full volume is the sample mean

Answer (Ex. 3.47) —

The values for the distribution function $\Phi(z)$ of the Normal(0, 1) RV Z are in the table on page 67. The thickness of the coating layer represented by Y_3 has the least probability (0.788) of meeting specifications. Consequently, a priority should be to reduce variability in this part of the process.

The values for the distribution function $\Phi(z)$ of the Normal(0, 1) RV Z are in the table on page 67.

CHAPFER 3. LIMIT LAWS OF STATISTICS 244

3 Random Variables	59
3.1 Basic Definitions	61
3.2 Discrete Random Variables	64
3.2.1 An Elementary Family of Bernoulli Random Variables	68
3.2.2 Independent Bernoulli Trials	69
3.2.3 Some Common Discrete Random Variables	70
3.3 Exercises in Discrete Random Variables	79
3.4 Continuous Random Variables	81
3.4.1 An Elementary Continuous Random Variable	84
3.4.2 Some Common Continuous Random Variables	85
3.5 Exercises in Continuous Random Variables	91
3.6 Transformations of random variables	91
3.6.1 A Review of Inverse Images	92
3.6.2 Transformations of discrete random variables	94
3.6.3 Transformations of continuous random variables	95
3.7 Exercises in Transformations of Random Variables	102
3.8 Expectations	102
3.8.1 Expectations of functions of random variables	103
3.8.2 Properties of expectations	106
3.8.3 Expectation of Common Random Variables	107
3.9 Exercises in Expectations of Random Variables	112
3.10 Multivariate Random Variables	113
3.10.1 \mathbb{R}^2 -valued Random Variables	114
3.10.2 Conditional Random Variables	124
3.10.3 \mathbb{R}^m -valued Random Variables	126
3.10.4 Some Common \mathbb{R}^m -valued RVs	130
3.10.5 Dependent Random Variables	135
3.11 Exercises in Multivariate Random Variables	136
3.12 Characteristic Functions	139
3.12.1 Obtaining Moments from Characteristic Function	139
3.12.2 Moment Generating Function	144
3.13 Exercises in Characteristic Functions	144
3.14 Statistics	145
3.14.1 Data and Statistics	145
3.14.2 Univariate Data	152
3.14.3 Bivariate Data	154
3.14.4 Trivariate Data	155
3.14.5 Multivariate Data	156
3.14.6 Loading and Exploring Real-world Data	157
3.14.7 Geological Data	157
3.14.8 Metereological Data	160
3.14.9 Textual Data	164
3.14.10 Machine Sensor Data	165
3.15 Exercises in Statistics	166

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. 3.45) —

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

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. 3.46) —

Due to independence of Y_1 , Y_2 and Y_3

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

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

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

218

Answers to Selected Exercises

4.1	Physical Random Number Generators	167
4.2	Pseudo-Random Number Generators	167
4.2.1	Linear Congruential Generators	168
4.2.2	Generalized Feedback Shift Register and the "Mersenne Twister" PRNG	171
4.3	Simulation of non-U uniform (I, J) Random Variables	173
4.3.1	Inversion Sampler for Continuous Random Variables	173
4.3.2	Inversion Sampler for Discrete Random Variables	181
4.3.3	von Neumann Rejection Sampler (RS)	190
4.4	Exercises in Simulation	195
5.1	Covergence of Random Variables	196
5.1.1	Properties of Convergence of RVs**	202
5.2	Law of Large Numbers	202
5.2.1	Point Estimation of $E(X_1)$	205
5.3	Central Limit Theorem	207
5.3.1	Application: Tolerating Errors in our estimate of $E(X_1)$	208
5.3.2	Application: Set Estimation of $E(X_1)$	210
5.4	Exercises in Limit Laws of Statistics	211
5.5	Standard normal distribution function table	212

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

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 $E(X) = \sum_x x \cdot f_X(x)$ and $E(Y) = \sum_y y \cdot f_Y(y)$.

$$= (0, 0) + (0.1, 0.1) + (0.1, 0.2) + (0.2, 0.1) + (0.2, 0.2) + (1.2, 1.2)$$

$$= (0, 0) \times 0.2 + (1, 1) \times 0.1 + (1, 2) \times 0.1 + (2, 1) \times 0.1 + (2, 2) \times 0.1 + (3, 3) \times 0.4$$

$$\lambda^x S \ni (i:x)$$

$$(\hbar' x) \lambda' X f \times (\hbar' x) \quad \swarrow \quad = ((\lambda'$$

$$(\hbar[x])\lambda[X]f \times (\hbar[x]) \quad \swarrow \quad = ((\lambda[X])\hbar) f$$

You can notice 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 and y plane as follows (make a picture!):

$$F_{X,A}(x,y) = \begin{cases} 1 - \frac{1}{e} e^{-\frac{x}{y}} & \text{if } x > 2 \text{ and } y \geq 0 \\ \frac{1}{2} (1 - \frac{1}{e} e^{-\frac{x}{y}} - \frac{1}{2} e^{-\frac{2x}{y}}) & \text{if } 1 \leq x < 2 \text{ and } y \geq 0 \\ 0 & \text{if } x < 1 \text{ or } y < 0 \end{cases}$$

Answer (Ex. 3.43) — 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:

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.

You can notice 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 $3 = 6$ partitions of the $x \times y$ plane as follows (make a picture!):

First obtain marginal PDF of Y . If $y \in [2, 3]$ then

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

$$f_Y(y) = \begin{cases} 0 & \text{otherwise} \\ 1 & \text{if } y \in \mathbb{R} \end{cases}$$

Therefore,

$$x_{-\vartheta} = (\zeta - \varepsilon)x_{-\vartheta} = \frac{\varepsilon}{\zeta}[\hbar]x_{-\vartheta} = \hbar p \Gamma \intop_{\xi} x_{-\vartheta} = \hbar p x_{-\vartheta} \intop_{\xi} = \hbar p(\hbar \cdot x) \lambda' X f \intop_{\infty} = (x) X f$$

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

Finally, verifying that $f_X(x, y) = f_X(x)f_Y(y)$ 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 a picture on the plane to work out the cases from the distinct expressions taken by $f_{X,Y}(x, y)$. Draw a picture on the plane to work out the cases from the distinct expressions taken by $f_{X,Y}(x, y)$.

non-zero val

Answer (Ex. 3.40) —

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. 3.41) —

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. 3.42) —

Find $E(XY)$, $E(X)$ and $E(Y)$ to get $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$ as follows:

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

List of Tables

1	Time Table for Virtual Student of Probability Theory I	4
2	Time Table for Inference Theory I	4
1.1	Symbol Table: Sets and Numbers	22
3.1	The 8 ω 's in the sample space Ω of the experiment \mathcal{E}_θ^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.	125
5.1	Symbol Table: Probability and Statistics	215
5.2	Random Variables with PDF and PMF (using indicator function), Mean and Variance	216
5.3	Symbol Table: Sets and Numbers	216
5.4	Symbol Table: Probability and Statistics	217

- | | |
|-----|---|
| 1.1 | Union and intersection of sets shown by Venn diagrams |
| 1.2 | These Venn diagram illustrate De Morgan's Laws. |
| 1.3 | A function f ("father of") from \mathbb{X} (a set of children) to \mathbb{Y} (their fathers) and its inverse ("children of"). |
| 1.4 | A pictorial depiction of addition and its inverse. The domain is plotted in orthogonal Cartesian coordinates. |
| 1.5 | A depiction of the real line segment $[-10, 10]$. |
| 1.6 | Point plot and stem plot of the finite sequence $\{b_{1:10}\}$ declared as an array. |
| 1.7 | A plot of the sine wave over $[-2\pi, 2\pi]$. |
| 2.1 | A binary tree whose leaves are all possible outcomes. |
| 2.2 | First ball number in 1114 NZ Lotto draws from 1987 to 2008. |
| 2.3 | Preference 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. |
| 3.1 | The indicator function of event $A \in \mathcal{F}$ is a RV I_A with DF F . |
| 3.2 | A RV X from a sample space \mathbb{U} with 8 elements to \mathbb{R} and its DF F . |
| 3.3 | $f(x)$ and $F(x)$ of the fair coin toss random variable X , a discrete uniform RV on $\{0, 1\}$. |
| 3.4 | $f(x)$ and $F(x)$ of the fair die toss random variable X , a discrete uniform RV on $\{1, 2, 3, 4, 5, 6\}$. |
| 3.5 | $f(x)$ and $F(x)$ of sumised astrogath toss random variable X , a discrete (non-uniform) RV on $\{1, 2, 3, 4\}$. |
| 3.6 | PMF $f(x; \theta)$ and DF $F(x; \theta)$ with $\theta = 0.33$. You should see how PMF and DF change as θ goes from 0 to 1. |
| 3.7 | PMF of $X \sim \text{Geometric}(\theta = 0.5)$ and the relative frequency histogram based on 100 samples from X according to Simulation 144 and Labwork 145, you will see in the sequel. |
| 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 according to Simulation 144 and Labwork 145, you will see in the sequel. |
| 3.9 | Figures from Sir Francis Galton, F.R.S., <i>Natural Inheritance</i> , Macmillan, 1889. |
| 4.1 | 3.10 $f(x)$ and $F(x)$ of the Uniform(0, 1) random variable X . |

List of Figures

$$\begin{aligned} \frac{\zeta}{\xi} &= \frac{9}{\xi} \times \frac{\zeta}{9} = \left(0 - 0 - + \frac{\xi}{1} + \frac{9}{1}\right) \frac{\zeta}{9} = \stackrel{0=\tilde{n}}{=} \left[\frac{\xi}{\varepsilon^{\tilde{n}}} + \frac{9}{\varepsilon^{\tilde{n}}} \right] \frac{\zeta}{9} = \tilde{n} p \left(0 - 0 - \varepsilon^{\tilde{n}} + \frac{\xi}{\tilde{n}}\right) \int_1^0 \frac{\zeta}{9} = \\ &\quad \tilde{n} p \stackrel{0=x}{=} \left[x \varepsilon^{\tilde{n}} + \frac{\xi}{\tilde{n} \varepsilon^x} \right] \int_1^0 \frac{\zeta}{9} = \tilde{n} p x p \left(\varepsilon^{\tilde{n}} + \tilde{n} \varepsilon^x\right) \int_1^0 \int_1^0 \frac{\zeta}{9} = \tilde{n} p x p \left(\tilde{n} + \varepsilon^x\right) \frac{\zeta}{9} \int_1^0 \int_1^0 = \\ &\quad \tilde{n} p x p (\tilde{n} \cdot x)^{\mathcal{X}} X f \tilde{n} \int_{\infty}^{\infty} \int_{\infty}^{\infty} = (\mathcal{X}) E \end{aligned}$$

$$\begin{aligned} & \left(0 - 0 - \frac{9}{1} + \frac{8}{1}\right) \frac{5}{9} = \int_1^{0-\delta} \left[\frac{9}{\varepsilon^{\delta}} + \frac{8}{\varepsilon^{\delta}} \right] \frac{5}{9} d\varepsilon = \delta p \left(0 - 0 - \frac{2}{\varepsilon^{\delta}} + \frac{4}{\varepsilon^{\delta}}\right) \int_1^0 \frac{5}{9} d\varepsilon = \\ & \delta p \int_1^{0=x} \left[\frac{2}{\varepsilon^{\delta} x} + \frac{4}{\varepsilon^{\delta} x} \right] \int_1^0 \frac{5}{9} d\varepsilon = \delta p x \int_1^0 \int_1^0 \frac{5}{9} d\varepsilon = \delta p x \int_1^0 \int_1^0 \frac{5}{9} d\varepsilon = \\ & \delta p x p(\delta x) \int_1^{\infty} \int_{-\infty}^{\infty} d\delta x d\delta = (\lambda X)(E) \end{aligned}$$

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

Answer (Ex. 3.39) — Note that $Y = (X - \mu)^2$ is not one-to-one so it is better to use the direct method by differentiating the distribution function of Y , $F_Y(y)$, to obtain $f_Y(y)$.

$$\begin{aligned}
 & ((\underline{\beta}^\wedge - \eta') X f + (\underline{\beta}^\wedge + \eta') X f) \frac{\underline{\beta}^\wedge}{1} = \\
 & \eta' X f \left(\frac{\underline{\beta}}{\frac{\underline{\beta}}{1} - \eta' \frac{\underline{\beta}}{1}} - \right) - (\underline{\beta}^\wedge + \eta') X f \frac{\underline{\beta}}{\frac{\underline{\beta}}{1} - \eta' \frac{\underline{\beta}}{1}} = \\
 & ((\underline{\beta}^\wedge - \eta') X f - (\underline{\beta}^\wedge + \eta') X f) \frac{\eta' d}{p} = (\underline{\beta}^\wedge - \eta') X f \frac{\eta' d}{p} \\
 & \text{pression gives} \\
 & (\underline{\beta}^\wedge + \eta') X f - (\underline{\beta}^\wedge + \eta') X f = \\
 & (\underline{\beta}^\wedge + \eta' \geq X \geq \underline{\beta}^\wedge - \eta') \mathbf{d} = \\
 & (\underline{\beta}^\wedge \geq \eta' - X \geq \underline{\beta}^\wedge - \eta') \mathbf{d} = \\
 & (\eta' \geq \zeta(\eta' - X)) \mathbf{d} = \\
 & (\eta' \geq \lambda) \mathbf{d} = (\eta') \lambda f
 \end{aligned}$$

Differentiating this expression gives

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	85
3.12 $f(x; \lambda)$ and $F(x; \lambda)$ of an exponential random variable where $\lambda = 1$ (dotted) and $\lambda = 2$ (dashed).	87
3.13 Density and distribution functions of $\text{Exponential}(\lambda)$ RVs, for $\lambda = 1, 10, 10^{-1}$, in four different axes scales.	87
3.14 $f(x)$ and $F(x)$ of the Uniform(θ_1, θ_2) random variable X	88
3.15 PDF and DF of a Normal(μ, σ^2) RV for different values of μ and σ^2	99
3.16 Mean ($\mathbf{E}_\theta(X)$), variance ($\mathbf{V}_\theta(X)$) and the rate of change of variance ($\frac{d}{d\theta}\mathbf{V}_\theta(X)$) of a Bernoulli(θ) RV X as a function of the parameter θ	107
3.17 Mean and variance of a Geometric(θ) RV X as a function of the parameter θ	109
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.	110
3.19 Diagrams done on the board!	124
3.20 Visual Cognitive Tool GUI: Quincunx & Septcunx.	132
3.21 Quincunx on the Cartesian plane. Simulations of Binomial($n = 10, \theta = 0.5$) RV as the x-coordinate of the ordered pair resulting from the culmination of sample trajectories formed by the accumulating sum of $n = 10$ IID Bernoulli($\theta = 0.5$) random vectors over $\{(1, 0), (0, 1)\}$ with probabilities $\{\theta, 1 - \theta\}$, respectively. The blue lines and black asterisks perpendicular to and above the diagonal line, i.e. the line connecting $(0, 10)$ and $(10, 0)$, are the density histogram of the samples and the PMF of our Binomial($n = 10, \theta = 0.5$) RV, respectively.	133
3.22 JPDF, Marginal PDFs and Frequency Histogram of Bivariate Standard Normal \vec{R}	135
3.23 JPDF, Marginal PDFs and Frequency Histogram of a Bivariate Normal \vec{R} for lengths of girths of cylindrical shafts in a manufacturing process (in cm).	136
3.24 Sample Space, Random Variable, Realisation, Data, and Data Space.	146
3.25 Data Spaces $\mathbb{X} = \{0, 1\}^2$ and $\mathbb{X} = \{0, 1\}^3$ for two and three Bernoulli trials, respectively.	146
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.	151
3.27 Frequency, Relative Frequency and Density Histograms	153
3.28 Frequency, Relative Frequency and Density Histograms	154
3.29 2D Scatter Plot	155
3.30 3D Scatter Plot	156
3.31 Plot Matrix of uniformly generated data in $[0, 1]^5$	156
3.32 Matrix of Scatter Plots of the latitude, longitude, magnitude and depth of the 22-02-2011 earth quakes in Christchurch, New Zealand.	159
3.33 Google Earth Visualisation of the earth quakes	161
3.34 Daily rainfalls in Christchurch since March 27 2010	162
3.35 Daily temperatures in Christchurch for one year since March 27 2010	163
3.36 Wordle of JOE 2010	164
3.37 Double Pendulum	165

Finally, the marginal PDF of the RV X in the first component of the \vec{R} (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{R} (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.

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

7.

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

^{4.1} The linear congruential sequence of L'Ucanggen (256, 137, 0, 123, 257) with non-maximal hours is:

$$= \int_{-\infty}^{\infty} 6(0.25 - (x - 1.5)^2) dx$$

$$(x_i^i, x_{i+1}^{i+1}, x_{i+2}^{i+2}) \text{ from two different view points.}$$

P12 173 work 130)

⁴⁻⁴ A plot of the PDF, DF or CDF and inverse DF of the uniform(-1,1) RV X. $\int_{-\infty}^x f(x) dx = F(x)$

$$4.6 \quad \text{The PDF of } DE \text{ and inverse DF of the Exponentiaf(A=1.0) RV.} \dots 126$$

4.7 Visual Cognitive Test GUI: Inversion Sampling from $X \sim \text{Exponentia}(0.5)$ 177

Assume that the three left bubbles function independently of each other. The probability that none of the first two bubbles burst is $(1 - p)^2$. The probability that the third bubble bursts is p . Thus, the probability that all three bubbles burst is $(1 - p)^2 \cdot p$.

4.10 The D.F. $F(x; 0.3, 0.7)$ of the de Molivré(0.3, 0.7) RV and its inverse $F^{-1}(u; 0.3, 0.7)$, 183

$P(\{X_1 < 1.2\} \cap \{X_2 < 1.2\} \cap \{X_3 < 1.2\}) = 0.8960 \approx 0.893$

where A_i is the length of time that bond i lasts.

Figure 3.1: Sequence of {Point Mass(1/2)}, {HWS (left parcel)} and {Point Mass(1/2)}, {HWS}

^{5,2} Distribution functions of several Normal(μ_1, σ_2^2) RVs for $\sigma_2^2 = 1, \frac{1}{10}, \frac{1}{100}, \frac{1}{1000}, \dots, 197$

$$I = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} J(X, Y) f(x, y) dx dy = \int_0^{\pi} \int_0^{\pi} J(X, Y) f(\cos(\theta), \sin(\theta)) d\theta d\phi$$

$n = 10$ [blue], and $n = 100$ [green]. One can see clear convergence respectively. One can see clear convergence respectively.

$$\text{PDFs } j_n(x) \text{ keep oscillating wildly with } n \text{ across } [0, 2] \text{ about } \mathbb{E}^{(0,1)}(x), \text{ the PDF of the }$$

¹⁹⁹ in DFS does not imply convergence in PDFs.

³⁻⁴ Sample mean \bar{X}_n as a function of sample size n for 20 replications from independent observations of $E(Y_i) = \mu + \epsilon_i$ (ϵ_i ~ N(0, σ^2)) and $E(\epsilon_i|\mu) = \mu$.

$$\text{Exponentia}(0.1) \text{ RV (red) with population means } (1+2+3+4+5+6)/6 = 21/6 = 3.5,$$

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

means, based on n IID samples from the Uniform(0, 1) RV for each of the p -folds, the $\hat{F}_k(x, n)$ function is given by

simulations (megabinary mes).

$$0 \rightarrow \left(\zeta / z^{\tilde{n}} + z^{x\tilde{n}} \right) \frac{G}{z} = \tilde{h}p(\tilde{n} + z^x) \frac{G}{z} \stackrel{0}{\rightarrow} \tilde{h}p(\tilde{n} + z^x) X' X f \stackrel{\infty}{\rightarrow} f = (x) X f$$

$$\left(\frac{z}{1} + \varepsilon^x\right) \frac{z}{9} = (0 - 1\varepsilon/2 + 1\varepsilon \times 1) \frac{z}{9} =$$

Chapter 1

Preliminaries

1.1 Elementary Set Theory

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

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

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

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

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

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

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

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:

$$b \notin B.$$

Now,

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

and so the variance is

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

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

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

Therefore,

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

and

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

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

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

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

Answer (Exercise 3.35) —

- 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 3.36) — 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 Binomial(n, θ), and this is analogous to the number of steps taken north in the random walk case.

Answer (Ex. 3.37) — The probability that *one* light bulb doesn't need to be replaced in 1200

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,

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

By drawing Venn diagrams, let us check **De Morgan's Laws**: We say two sets C and D are **disjoint** if they have no elements in common, i.e. $C \cap D = \emptyset$.

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

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.:

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

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:

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

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

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

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

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

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

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

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

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:

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

$$C \subset D$$

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

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

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

and so

$$\left. \begin{array}{l} x = 6 \\ x = 5 \\ x = 4 \\ x = 3 \\ x = 2 \\ x = 1 \end{array} \right\} = (x) f$$

Answer (Ex. 3.34) — 1. The probability mass function of X is

$$\begin{aligned} &= d_2(X) + 2abE(X)^2 - 2abE(X)d_2 = d_2(E(X)) - (E(X))^2 \\ &= E((aX+b)^2) - (aE(X)+b)^2 = E((aX^2+2aXb+b^2)) - (a^2E(X^2)+b^2) \end{aligned}$$

square, we get:

Answer (Ex. 3.32) — Using the definition of variance, expectations and by completing the

square, you understand each step.

Answer (Ex. 3.31) — This was already done in Sec. 3.8.2 on Properties of Expectation. Make

sure you understand each step.

The variance is $V(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 0.3 - 0.5^2 = 0.05$

$$\begin{aligned} &= \frac{4}{9}(1^4 - 0) - \frac{5}{6}(1^5 - 0) \\ &= \int_1^0 \left[\frac{5}{2}x^2 - \frac{5}{4}x^3 \right] dx = \end{aligned}$$

$$\begin{aligned} &= \int_1^0 6x^3 - 6x^4 dx \\ &= \int_1^0 6x(x-1) dx = \end{aligned}$$

$$\begin{aligned} &= 0.5 \\ &= 2(1^3 - 0) - \frac{4}{9}(1^4 - 0) \\ &= \int_1^0 \left[2x^3 - \frac{4}{9}x^4 \right] dx = \end{aligned}$$

$$\begin{aligned} &= \int_1^0 6x^2 - 6x^3 dx \\ &= \int_1^0 6x(x-1) dx = \end{aligned}$$

$$\begin{aligned} &= 0.5 \\ &= 2(1^3 - 0) - \frac{4}{9}(1^4 - 0) \\ &= \int_1^0 \left[2x^3 - \frac{4}{9}x^4 \right] dx = \end{aligned}$$

$$\begin{aligned} &= \int_1^0 6x^2 - 6x^3 dx \\ &= \int_1^0 6x(x-1) dx = \end{aligned}$$

Answer (Ex. 3.30) — The expected value $\mathbb{E}(X)$ is

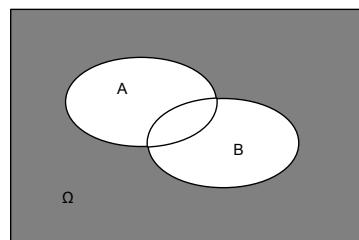
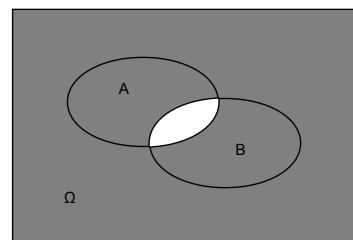
(a) $(A \cup B)^c = A^c \cap B^c$ (b) $(A \cap B)^c = A^c \cup B^c$

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

1. $F \cap C =$

2. $F \cup C =$

3. $F \setminus C =$

4. $C \setminus F =$

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

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

1.2 Exercises

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

Now consider the following subsets:

- The set of 50 students, $S = \{S_1, S_2, S_3, \dots, S_{50}\}$.
- The set of 3 lecturers, $L = \{L_1, L_2, L_3\}$.
- The set of 4 tutors, $T = \{T_1, T_2, T_3, T_4\}$.

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:

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

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 (Ex. 3.29) — 1. The expected number of conditioners that the store sells daily is

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

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.

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

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

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

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

The set of non-negative integers is:

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

number zero may be defined as the size of an empty set:

$$\begin{aligned} & \vdots \\ 2 &= \#\{\ast, \bullet\} = \#\{\bullet, \circ\} = \#\{\circ, \omega\} = \#\{\circ, \circ, \bullet\} = \dots, \\ 1 &= \#\{\ast\} = \#\{\bullet\} = \#\{a\} = \#\{\{\bullet, \circ\}\} = \{\{\bullet, \circ\}, \dots\} \end{aligned}$$

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

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**:

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

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

1.3 Natural Numbers, Integers and Rational Numbers

$\{a_1, a_2, \dots, a_n\}$	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 ,
$A \cap B$	or both.
$\{\} \text{ or } \emptyset$	"intersection", meaning the set of all elements which are in both A and B .
\mathcal{U}	empty set.
A^c	universal set.
A^c	the complement of A , meaning the set of all elements in \mathcal{U} , which are not in A .

SET SUMMARY

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

Ex. 1.3 — Using Venn diagram, sketch and check the rule:

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,

$$f_X(-\log_2(y)) = \theta(1 - \theta)^{-\log_2(y)-1} \quad \text{if } y \in \{2^{-1}, 2^{-2}, 2^{-3}, \dots\} \\ \text{otherwise .} \quad \left\{ \begin{array}{ll} 0 & \text{if } y \in \{2^{-1}, 2^{-2}, 2^{-3}, \dots\} \\ f_X(y) & \text{otherwise .} \end{array} \right\} = (f_Y(y))^{\theta}$$

Therefore, $f_Y(y) = \theta(1 - \theta)^{-\log_2(y)-1}$.

So the probability density function of Y is given by

$$f_Y(y) = \frac{0}{\log_2(y)} \quad \text{if } y \leqslant 1 \\ \frac{\theta}{\log_2(y)} \quad \text{otherwise}$$

Since $e^{-\log_2(y)} = e^{\log_2(y^{-1})} = y^{-1}$.

$$f_Y(y) = f_X(\log_2(y)) \times \frac{y}{1} = \log_2(y) e^{-\log_2(y)} \times \frac{y}{1} = \log_2(y) \frac{y}{1}$$

Therefore

$$\left| \frac{dy}{dx} (g^{-1}(y)) \right| = \left| \frac{dy}{d\log_2(y)} \right| = \frac{y}{1}.$$

Now $x = g^{-1}(y) = \log_2(y)$ is a monotone increasing function for y in $[1, \infty)$

Answer (Ex. 3.26) — Since $y = g(x) = e^x$ is a monotone increasing function for $x \leqslant 0$, we can apply the change of variable formula.

Caution: This is a discrete RV and so don't blindly apply the change of variable formula that only applies to continuous random variables. But Poisson is discrete and "change of variable formula" is hence inapplicable.

Also happens to satisfy these properties. But Poisson is discrete and "change of variable formula" is hence inapplicable.

With a little thought, it's just that in this discrete RV setting the inverse function for x in $(-\infty, 0]$ applies to continuous random variables.

for $y = 1, \frac{1}{2}, \frac{1}{3}, \dots$, and 0 otherwise.

$$f_Y(y) = \mathbf{P}(A = y) = \sum_{x: g(x)=y} \mathbf{P}(X=x) = \sum_{x: g(x)=y} \frac{e^{-\lambda} \lambda^x}{x!} = \frac{e^{-\lambda} \lambda^y}{y!} - 1 \quad \left(\frac{\lambda^y}{y!} - 1 \right)!$$

want to recall quickly https://en.wikipedia.org/wiki/Inverse-function_theorem again, so we get:

and since $y = g(x) = (x + 1)^{-2}$ as it maps or associates each $y \in \{1, \frac{1}{2}, \frac{1}{3}, \dots\}$ to exactly one $x \in \{0, 1, 2, \dots\}$ given by $g^{-1}(y) = y^{-1/2} - 1 = \frac{1}{y} - 1 = x$, it is injective function, this is injective or one-to-one as explained here if you want to recall quickly https://en.wikipedia.org/wiki/Inverse-function_theorem again, so we get:

$$\{ \dots, 2, 1, 0 \} \in x \xrightarrow{x+1=-2} y \in \left\{ \frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \dots \right\}$$

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

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

argument to $f_X(\cdot)$ for notational ease, we get

Answer (Ex. 3.25) — Since X is a Poisson(λ) random variable (by suppressing the $; \lambda$) in the

$$\left\{ \begin{array}{ll} 0 & \text{otherwise .} \\ f_X(-\log_2(y)) = \theta(1 - \theta)^{-\log_2(y)-1} & \text{if } y \in \{2^{-1}, 2^{-2}, 2^{-3}, \dots\} \end{array} \right\} = (f_Y(y))^{\theta}$$

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

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

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.

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

Answer (Ex. 3.22) — Using Equation (3.35), 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 (3.35),

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

whose irreducible unique expression is $1/2$.
For example, $1/2, 2/4, 3/6$, and $100/200$ are different expressions for the same rational number $1/2$, where y is positive and as small as possible. Rational number has a unique irreducible expression p/y , where y is positive and as small as possible. The expressions p/y and $p \cdot q/q$ 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 .

$$\mathbb{Q} := \{d/b : d \in \mathbb{Z}, b \in \mathbb{Z} \setminus \{0\}\}$$

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

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

What is its range?

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

Cartesian product of \mathbb{Z} :

Try to set up the arithmetic operation of addition as a function. The domain for addition is the set of integers $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, \dots\}$.

Classwork 4 (Addition over integers) Consider the set of integers $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, \dots\}$.

Finally, we say that a is greater than b and write $a > b$ if $a < b$. 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 $a + 1$.

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. We say an integer a is **less than** an integer b and write $a < b$ if $b - a$ is positive. We of order. We say an integer a is either positive, negative, or zero. In terms of this we define the notion of integers. Every integer is assumed to be familiar with such arithmetic operations with pairs of integers. The reader is assumed to be familiar with such arithmetic operations with pairs of integers. Some examples of functions you may have encountered are **arithmetic operations** such as **addition** (+), **subtraction** (-), **multiplication** (\cdot) and **division** (/) of ordered pairs are dropped. Some examples of functions you may have encountered are **arithmetic operations** with a **plus or positive sign** (+) before them are called positive integers. Conventionally, + signs we can motivate the set of integers:

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

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:

as the **inverse** of f .

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").

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

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

2. Set

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

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

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

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

Hence

$$F(x) = \int_{-4}^x 0 \, dy + \int_{-4}^4 \frac{8}{1} \, dy + \int_x^4 \frac{8}{1} \, dy = \int_{-4}^4 0 \, dy + \int_x^4 \frac{8}{1} \, dy =$$

If $x \geq 4$, then

$$F(x) = \begin{cases} \frac{8}{1}(x + 4) & x \geq 4 \\ 0 + \int_x^4 \frac{8}{1} \, dy & -4 \leq x \leq 4 \\ 0 & x < -4 \end{cases}$$

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

$$F(x) = \int_x^4 0 \, dy = 0.$$

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

$$y = \frac{8}{1}$$

$$8y = 1$$

$$y(-4) = 1$$

$$y \left[x \right]_4^{-4} = 1$$

That is,

$$\int_4^{-4} y \, dx = 1.$$

grates to one,

Answer (Ex. 3.19) — 1. Since $f(x)$ is a (continuous) probability density function which inter-

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.

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

Answer (Exercise 3.17) — This exercise is optional.

... Try it if you want to and communicate face-to-face your understanding or misunderstanding with Raaz.

Answer (Exercise 3.18) —

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

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

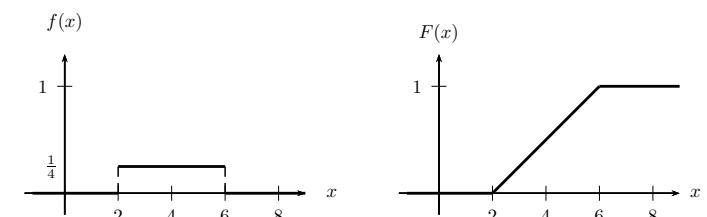
as expected!

(b) Now

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

so the graphs are:

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



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 equinumerous classes of Cauchy sequences 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 construction of real numbers from first principles, we give a more concrete one via decimal expansions.

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 rationals and the sequence $0.d_1d_2d_3\dots$ implies:

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

and so the probability of observing two or more particles during any given second, $\mathbf{P}(X \geq 2) = 1 - \mathbf{P}(X < 2) = 1 - \mathbf{P}(X = 1) - \mathbf{P}(X = 0)$,

Answer (Ex. 3.14) — Since X is Poisson(λ) random variable with $\lambda = 0.5$, $\mathbf{P}(X \geq 2)$ is the probability of observing two or more particles during any given second. Now $\mathbf{P}(X = 0) = e^{-0.5} = 0.6065$ and $\mathbf{P}(X = 1) = \frac{0.5}{1!} \times e^{-0.5} = 0.3033$. and so $\mathbf{P}(X \geq 2) = 1 - 0.9098 = 0.0902$.

Some examples of real numbers that are not rational (**rational numbers**) are:
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} and \underline{y} , \overline{x} and \overline{y} .
 $\sqrt{2} = 1.41421356237309\dots$ the side length of a square with area of 2 units
 $\pi = 3.14159265358979\dots$ the ratio of the circumference to diameter of a circle
 $e = 2.71828182845904\dots$ Euler's constant

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

$$\begin{aligned} & 3 + \frac{1}{10} < \pi < 3 + \frac{1}{10} + \frac{1}{10^2} \\ & 3 + \frac{1}{10} + \frac{1}{10^2} < \pi < 3 + \frac{1}{10} + \frac{1}{10^2} + \frac{1}{10^3} \\ & \vdots \\ & 3 + \frac{1}{10} + \frac{1}{10^2} + \cdots + \frac{1}{10^k} < \pi < 3 + \frac{1}{10} + \frac{1}{10^2} + \cdots + \frac{1}{10^k} + \frac{1}{10^{k+1}} \end{aligned}$$

Think of the real number system as the continuum of points that make up a line, as shown in Figure 1.5.
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 $y \leq x \leq z$.

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

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

Hence

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

Using (b) and solving for λ gives:

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

3. Since $\mathbf{P}(X \geq 1) = 0.1$, and the formula becomes $\mathbf{P}(X = 0) = e^{-\lambda}$.

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

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

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

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

and so

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

Now

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

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

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

probability of observing two or more particles during any given second.

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

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

and so the probability of zero defects is

300 meter tape. If X is the number of defects on a 300 meter tape, then X is Poisson with $\lambda = 6$ defects on a 300 meter tape. If X is the number of defects on every 100 meters, we would expect 6 defects on a 300 meter tape. Therefore, the probability that the target will be hit at least once is

$$1 - 0.3487 \approx 0.6513.$$

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

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

where

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

so

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Answer (Ex. 3.8) — (a)

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

(b) (i) $\mathbf{P}(X \leq 5) = F(5) = 0.17$ (ii) $\mathbf{P}(X < 12) = \mathbf{P}(X \leq 11) = F(11) = 0.84$ (iii) $\mathbf{P}(X > 9) = 1 - \mathbf{P}(X \leq 9) = 1 - F(9) = 1 - 0.79 = 0.21$ (iv) $\mathbf{P}(X \geq 9) = 1 - \mathbf{P}(X < 9) = 1 - \mathbf{P}(X \leq 8) = 1 - 0.59 = 0.41$ (v) $\mathbf{P}(4 < X \leq 9) = F(9) - F(4) = 0.79 - 0.08 = 0.71$ (vi) $\mathbf{P}(4 < X < 11) = \mathbf{P}(4 < X \leq 10) = F(10) - F(4) = 0.82 - 0.08 = 0.74$ **Answer (Ex. 3.9) —** Since we are sampling without replacement,

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

So the probability mass function of X is:

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

The required probabilities are:

1.

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

2.

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

3.

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

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

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

```

<> diary off % turn off the current diary file blah.txt
ans = 59
<> 3+66
<> diary blah.txt % start a diary file named blah.txt

```

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 a diary file by typing **diary filename.txt** 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.

```

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

```

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.

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

```

37
ans =

```

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

```

>> 13+24

```

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

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

We use MATLAB to perform computations and visualizations. MATLAB is a numerical computing environment and programming language that is optimised for vector and matrix processing. MATLAB runs MATLAB. You can remotely connect to these machines from home by following instructions at <http://www.math.centrebury.ac.nz/php/resources/computers/remote>.

1.5 Introduction to MATLAB

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

$$f(x) = \begin{cases} \frac{2}{200} & x = 3 \\ \frac{22}{200} & x = 2 \\ \frac{200}{200} & x = 1 \\ \frac{176}{200} & x = 0 \end{cases}$$

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

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

$$(\text{This answer is easily seen from the distribution function of } X.)$$

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

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

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

so the distribution function is:

$$\mathbf{P}(X = x) \mid \begin{array}{c|ccccc} x & 0 & 1 & 2 & 3 & 4 & 5 \\ \hline 0 & 0.1 & 0.2 & 0.2 & 0.2 & 0.3 & \end{array}$$

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

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

A. If Ω is the sample space, then $\mathbf{P}(\Omega) = 1$ and so the correct probability is

Answer (Ex. 3.5) — $\mathbf{P}(X = 3)$ does not satisfy the condition that $0 \leq \mathbf{P}(A) \leq 1$ for any event

x	observed frequency	observed relative frequency	Prob of x hits
0	229	229/576 = 0.393	$f(0; 0.933) = 0.394$
1	211	211/576 = 0.366	$f(1; 0.933) = 0.367$
2	93	93/576 = 0.161	$f(2; 0.933) = 0.171$
3	35	35/576 = 0.0608	$f(3; 0.933) = 0.0532$
4	7	7/576 = 0.0122	$f(4; 0.933) = 0.0124$
5	1	1/576 = 0.00174	$1 - \sum_{x=0}^{4} f(x; 0.933) = 0.00275$

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

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

```
>> VariableCalledX = 12
```

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

```
VariableCalledX = 12
```

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

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

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

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

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

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

$$\begin{aligned} p &= 45.89 * 1.00009 & d &= 89.0 / 23.3454 \\ m &= 5376.0 - 6.00 & p &= 2^{0.5} \end{aligned}$$

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.

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:

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

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:

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

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 θ in a sequence of $n = 537$ independent Bernoulli(θ) trials. Finally, we can approximate this $\text{Binomial}(n = 537, \theta = \frac{1}{576})$ random variable by Poisson(λ) random variable with $\lambda = n\theta = \frac{537}{576} \cong 0.933$. Using the probability mass function formula for Poisson($\lambda = 0.933$) random variable X we can obtain the probabilities and compare them with the relative frequencies from the data as follows:

and

$$P(F_3|D_e) = \frac{P(D_e)}{P(F_3 \cup D_e)} = \frac{P(D_e)}{P(F_3) - P(D \cap F_3)} = \frac{P(D_e)}{1/12 - 5/72} = \frac{1}{1/12 - 5/72} = \frac{43}{43}.$$

so

$$P(F_1|D_e) = \frac{P(D_e)}{P(F_1 \cup D_e)} = \frac{P(D_e)}{P(F_1) - P(D \cap F_1)} = \frac{P(D_e)}{2/3 - 1/6} = \frac{36}{36}.$$

and so

$$P(F_1 \cup D_e) = P(F_1) - P(F_1 \cap D_e)$$

$$P(F_1) = P(F_1 \cup D_e) + P(F_1 \cap D_e)$$

partitioning idea of the "Total Probability Theorem", and write:

$$P(D_e) = 1 - \frac{29}{43} = \frac{14}{43} = \frac{72}{72}.$$

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

$$\begin{aligned} P(F_3|D) &= \frac{P(D)}{P(F_3 \cup D)} = \frac{P(D)}{5/72} = \frac{29}{5} \\ P(F_2|D) &= \frac{P(D)}{P(F_2 \cup D)} = \frac{P(D)}{1/6} = \frac{29}{12} \\ P(F_1|D) &= \frac{P(D)}{P(F_1 \cup D)} = \frac{P(D)}{1/6} = 12 \end{aligned}$$

(b) Knowing that the gale did cause damage we can calculate the probabilities that it was of the various forms using the probabilities in (a) as follows (Note: $P(D \cap F_1) = P(F_1 \cup D)$ etc.):

$$\begin{aligned} P(D) &= \frac{1}{6} + \frac{1}{6} + \frac{5}{72} = \frac{29}{72} \\ P(D \cup F_3) &= P(D|F_3)P(F_3) = \frac{6}{5} \times \frac{1}{12} = \frac{5}{72}. \end{aligned}$$

and

$$\begin{aligned} P(D \cup F_2) &= P(D|F_2)P(F_2) = \frac{3}{2} \times \frac{1}{1} = \frac{3}{2}, \\ P(D \cup F_1) &= P(D|F_1)P(F_1) = \frac{1}{4} \times \frac{2}{3} = \frac{1}{6}, \end{aligned}$$

where

$$P(D) = P(D \cup F_1) + P(D \cup F_2) + P(D \cup F_3)$$

The probability that a reported gale causes damage is

$$P(D|F_1) = \frac{1}{4}, \quad P(D|F_2) = \frac{3}{2}, \quad P(D|F_3) = \frac{6}{5}.$$

probabilities:

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

$$P(F_1) = \frac{2}{6}, \quad P(F_2) = \frac{1}{4}, \quad P(F_3) = \frac{1}{12}.$$

gale of force 2 occurs and F_3 be the event a gale of force 3 occurs. Now we know that F_2 be the event a gale of force 1 occurs, let F_1 be the event a gale of force 1 occurs. Now we know that

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.

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

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

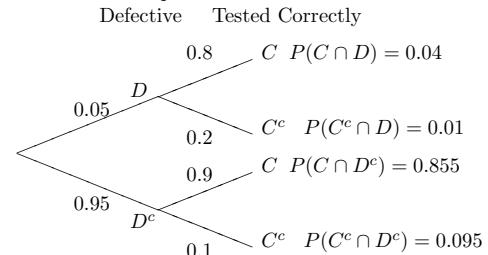
The probability that a micro-chip is defective, and has been declared as effective is

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

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:

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

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:

Example 6 Suppose that a Personal Identification Number (PIN) is a six-symbol code word in

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 total number of possible PINs is:

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

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

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

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

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

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

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 .

Permutations: There are basically two types of permutations:

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

So in mathematics, we use more precise language:

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

"The combination of my PIN is math99".

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

"I have 17 probability texts on my bottom shelf"

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

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

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 total number of possible PINs is:

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

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

First digit: 10 possibilities	Second digit: 10 possibilities

Third letter: 26 possibilities	Fourth letter: 26 possibilities

Second letter: 26 possibilities	Third letter: 26 possibilities

First letter: 26 possibilities	Fourth letter: 26 possibilities

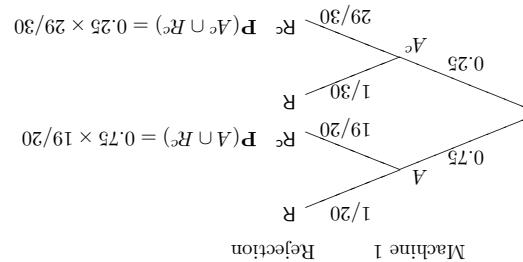
Example 6 Suppose that a Personal Identification Number (PIN) is a six-symbol code word in

2. 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 .

1. Repetition is allowed, as in choosing the letters (unrestricted choice) in the PIN of Example 6.

So the required probability is

$$P(A|R^c) = \frac{P(R^c \cup A)}{\frac{29}{30} \times 0.75 + \frac{30}{30} \times 0.25} \approx 0.747$$



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

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

$$P(H^c \cup A) = P(H^c|A)P(A) = \frac{19}{20} \times 0.75$$

$$P(A|R^c) = \frac{P(H^c)}{P(H^c \cup A)} = \frac{P(H^c \cup A^c)}{P(H^c \cup A) + P(H^c \cup A^c)}$$

We want $P(A|R^c)$ which is give by

$$P(R|A^c) = \frac{1}{10} \quad \text{and} \quad P(H^c|A^c) = \frac{29}{30}$$

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

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

where, H denotes the event that the bottles are accepted. We know the following probabilities: event that bottles are produced by machine 2; R denotes the event that the bottles are rejected; event that bottles are produced by machine 1; and A^c is the event that the bottles are accepted; and R^c denotes the event that the bottles are rejected. We know the following probabilities:

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

3. 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 .

So the required probability is

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.

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

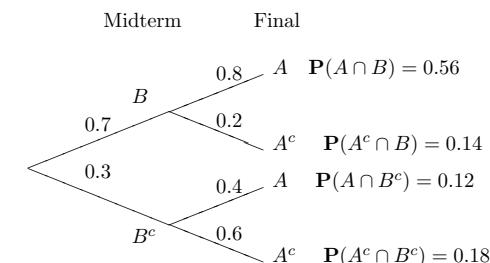
Let A be the event *the sum is 5* and B be the event *the sum is 6*, then A and B are mutually exclusive events with probabilities

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

Therefore,

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

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



Then the probability of passing the final exam is:

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

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

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

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

Some factorials to bear in mind

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

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

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

In general,

Possible ways to choose three distinct objects from a set of n objects and arrange them in a row. 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, made there are $n(n-1)$ ways to select a different object to place next to the left-most one. Thus, given n objects, there are n ways to choose the left-most object, and once this choice has been

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

arrange them in a row in six ways (p3,2):

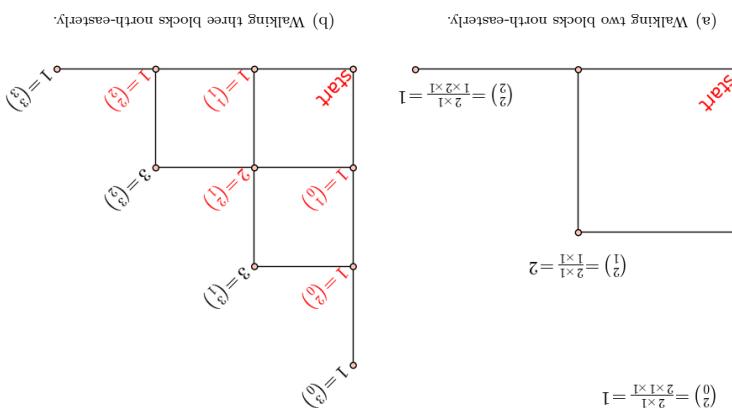
Let the number of ways to choose k objects out of n and to arrange them in a row be denoted by $p_{n,k}$. For example, we can choose two ($k = 2$) objects out of three ($n = 3$) objects, {a,b,c}, and

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

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

$$12, 21, 12, 21,$$

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



CHAPTER 1. PRELIMINARIES

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

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

For case B, the probability of missing both shots is $\frac{3}{5} = \frac{9}{15}$, so the probability hitting some target at least once is $\frac{6}{15} = \frac{2}{5}$. For case A, there is only one shot so the probability of hitting at least one is $\frac{2}{5}$.

Answer (Ex. 2.6) — We can assume that the first shot is independent of the second shot so we can multiply the probabilities here.

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

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

$$P(\{B\} \cup \{l\}) = P(B) + P(l) = \frac{5}{15} + \frac{5}{15} = \frac{2}{3}.$$

4. Since the events {B} and {l} are disjoint,

$$\begin{aligned} P(O) &= P(\{B, l, N, G, O\}) \\ &= P(\{B\} \cup \{l\} \cup \{N\} \cup \{G\} \cup \{O\}) \\ &= P(\{B\} + P(l) + P(N) + P(G) + P(O)) \quad \text{Simplifying notation} \\ &= 1 = \frac{5}{5} + \frac{5}{5} + \frac{1}{5} + \frac{5}{5} + \frac{5}{5} \\ &= 1. \end{aligned}$$

3. Using the addition rule for mutually exclusive events,

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

2. The probabilities of simple events are:

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

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

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

$$P(W) = \frac{1}{11}, P(A) = \frac{3}{11}, P(l) = \frac{1}{11}, P(M) = \frac{1}{11}, P(K) = \frac{1}{11}, P(R) = \frac{2}{11}.$$

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

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, *Piṅgala’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

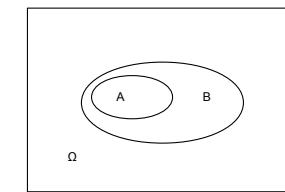
$$\langle x_{m:n} \rangle := x_m, x_{m+1}, \dots, x_{n-1}, x_n$$

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:

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



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) $\mathbf{P}(\{Z\}) = 0.1\% = \frac{0.1}{100} = 0.001$

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

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

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

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

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

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

2. $\{\text{RRRR, RRRR, RRLR, RLRR, LRRR, RLRL, RRLL, LLRR, LRLR, LRRL, RLLL}\}$

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 $\Omega = \{W, A, I, M, K, R\}$.

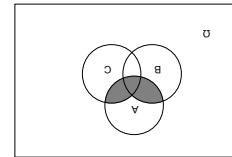
The following Venn diagram illustrates the two implications clearly.

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

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

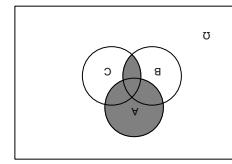
illustrate two implications:

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



We can check $A \cup (B \cup C) = (A \cup B) \cup (A \cup C)$ from the following sketch:

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



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

- (a) $L \cup L = \{L\}$
- (b) $L \cup S = \emptyset$
- (c) $L \cup L = \{L_1, L_2, L_3, L_1, L_2\}$
- (d) $L \cup L = \{L_1, L_2\}$
- (e) $S \cup S = \{y_1, y_2, \dots, y_m\}$
- (f) $S \cup T = \emptyset$
- (g) $S \cup T = \{T_1, T_2, T_3\}$
- (h) $L \cup L = \{L_1, L_2, S_1, S_2, \dots, S_{50}\}$
- (i) $L \cup L = \{L_1, L_2\}$
- (j) $L \cup L = \emptyset$

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

Answers to Selected Exercises

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

The finite sequence $\langle x_{m:n} \rangle$ has $\mathcal{I} = \{m, m+1, m+2, \dots, n\}$ 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 a finite sequence $\langle x_{m':n'} \rangle$ where $m' \leq j \leq k < n$. As a function that maps the index set to the set of real numbers:

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

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

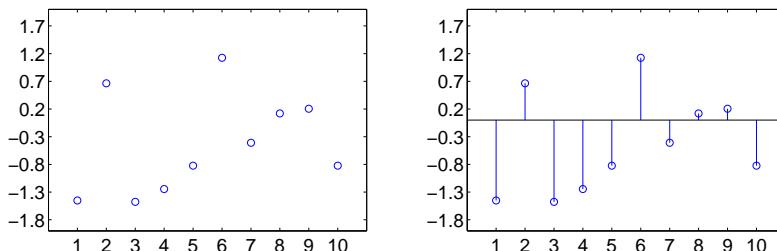
```

>> 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

```

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 $\vec{RV} (X,Y)$
$F_{X,Y}(x)$	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 $\vec{RV} (X,Y)$
$S_{X,Y}$ = $\{(x_i, y_j) : f_{X,Y}(x_i, y_j) > 0\}$	The support set of the discrete $\vec{RV} (X,Y)$
$f_{X,Y}(x)$	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}
$E(g(X,Y)) = \sum_{(x,y) \in S_{X,Y}} g(x,y) f_{X,Y}(x,y)$	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 5.4: Symbol Table: Probability and Statistics

Table 3.2: Random Variables with PDF and PMF (using indicator function), Mean and Variance

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

Table 5.3: Symbol Table: Sets and Numbers

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

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

```

>>> y = %_ the array y is
>>> z = %_ the array z is
>>> x = %_ the sum of vectors y and z (with same size 1 X 3)
% y is updated to 2 * y (each term of y is multiplied by 2)
>>> y = y * 2
% p is the vector obtained by term-by-term product of z and y
>>> p = z / y
>>> d = 1.5000
% d is the vector obtained by term-by-term division of z and y
>>> t = 10.0000
% t has 4 numbers equally-spaced between -10 and 10
>>> t= linspace(-10,10,4)
>>> s = sqrt(t)
% s is a vector obtained from the term-wise square of t
>>> ssq = sin(t) .^ 2
% ssq is an array obtained from term-wise squaring ( .^ 2 ) of the cos(t) array
>>> csg = cos(t) .^ 2
% csg is an array obtained from term-wise squaring ( .^ 2 ) of the cos(t) array
>>> csg = 0.7040 0.9637 0.9637 0.7040

```

```

<>> y % the array y is
<>> z % the array z is
<>> x = 1 1
<>> y = 1 1
<>> z = 1 1
<>> x = y + z
<>> x = 2 3
<>> y = 2 3
<>> z = 2 2
<>> p = 2 4
<>> d = z / y
<>> a = 0.0000 1.5000
<>> t = -10.0000 -3.3333 10.0000
<>> t=Linspace(-10,10,4)
% t has 4 numbers equally-spaced between -10 and 10
<>> s = sin(t)
<>> ssq = sin(t)^2
% ssq is an array obtained from term-wise squaring ( .^ 2 ) of the sin(t) array
<>> csg = cos(t)^2
% csg is an array obtained from term-wise squaring ( .^ 2 ) of the cos(t) array
<>> csg = 0.7040 0.9637 0.9637 0.7040

```

CHAPTER 5. LIMIT LAWS OF STATISTICS

```
>> 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.

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 5.1: Symbol Table: Probability and Statistics

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

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

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

Conditional probabilities obey the 4 axioms of probability.

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

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.
Ω	— universal set.
A^c	— the complement of A , meaning the set of all elements in Ω , the universal set, which are not in A .

EXPERIMENT SUMMARY

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

PROBABILITY SUMMARY

Axioms:

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

Rules:

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

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

5. Standard normal distribution function table

For any given value z , its cumulative probability $\Phi(z)$ was generated by Excel formula NORMSDIST(z).

$$\lim_{\substack{x \rightarrow 1 \\ 0 \leftarrow x}} (x+1)^0 = (x)f^0$$

the Euler's constant e :

$$= \exp\left(\lim_{x \leftarrow 0} \log((x + 1)^{1/x})\right) \quad \text{Transformed using } \exp(\lim_{x \leftarrow 0} \log((x + 1)^{1/x}))$$

Pe 1_∞.

$$= \exp\left(1/\left(\lim_{x \leftarrow 0} (x+1)\right)\right) \quad \text{The limit of } x+1 \text{ as } x \text{ approaches 0 is 1}$$

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:

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

Next we look at some examples of limits at infinity.

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

$\vdash_{\mathcal{M}} \vartheta$ Σ

$$(u \quad) \xrightarrow{\infty \leftarrow u} \quad \quad \quad \xrightarrow{\infty \leftarrow u}$$

proaches ∞ exists and it is 1:

$$\cdot \mathbb{I} = \left(\frac{u}{\sqrt{\cdot}} - \mathbb{I} \right) \stackrel{\infty \leftarrow u}{\lim} = (\mu) f(u) \stackrel{\infty \leftarrow u}{\lim}$$

Definition 5 (Continuity or a function) We say a real-valued function $f(x)$: $D \rightarrow \mathbb{K}$ with the

— 20 —

$$(v)f = (x)f \lim_{x \leftarrow a} \quad \text{or} \quad (v)f = (x)f \lim_{+a \leftarrow x}$$

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

$$\cdot (x)f \underset{-v \leftarrow x}{\lim} = (v)f = (x)f \underset{+v \leftarrow x}{\lim}$$

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

5. Standard normal distribution function table

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)

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

Labwork 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:

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

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 72 The normal-based confidence interval for θ^* (as well as λ^* in the previous example) may not be a valid approximation here with just $n = 7$ samples. After all, the CLT only tells us that the point estimator $\hat{\Theta}_n$ 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 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 $\sum_{i=1}^n X_i/\sqrt{n}$ as $n \rightarrow \infty$ when $X_i \stackrel{\text{IID}}{\sim} \text{Uniform}(-10, 10)$?

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

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

- 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, $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- This time, $\Omega = \{\omega_1, \omega_2\}$ where $\omega_1 = \text{Heads and } \omega_2 = \text{Tails}$.
- $\Omega = \{\text{Heads, Tails}\}$ is our experiment so $\Omega = \{\omega_1, \omega_2\}$ where $\omega_1 = \text{Defective and } \omega_2 = \text{Non-defective}$.
- There are only two outcomes here, so $\Omega = \{\omega_1, \omega_2\}$ where $\omega_1 = \text{Inspect a light bulb}$.
- $\Omega = \{\text{Defective, Non-defective}\}$ is our experiment is to inspect a light bulb.

Example 23 Some standard examples of experiments are the following:

or pair-wise disjoint events. This means that $E_i \cap E_j = \emptyset$ where $i \neq j$. Events, E_1, E_2, \dots, E_n , that cannot occur at the same time are called **mutually exclusive events**, called a **simple event**. The subsets of Ω are called **events**. A single outcome, ω , when seen as a subset of Ω , as in $\{\omega\}$, is called a **sample space**. The set of all outcomes is called the **sample space**, and is denoted by Ω . Definition 6 An experiment is an activity or procedure that produces distinct, well-defined possibilities called **outcomes**. The set of all outcomes is called the **sample space**, and is denoted by Ω .

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.

Probability Model

Chapter 2

2.1 Experiments

company is also within $(6.9016, 13.3841)$ and since the mean waiting time of 10 minutes promised by the Orbiter bus ($6.9016, 13.3841$) is contained in the true population mean $E(X_1) = 1/\alpha^*$ we can be fairly certain that the company sticks to its promise.

respectively. So with 95% probability the true population mean $E(X_1) = 1/\alpha^*$ is contained in the interval $(10.143 \pm 1.96 \times 4.375/\sqrt{7}) = (6.9016, 13.3841)$,

$$x_7 = 10.143 \quad \text{and} \quad (x_7 \pm z_{\alpha/2} s_7 / \sqrt{7}) = (10.143 \pm 1.96 \times 4.375/\sqrt{7}) = (6.9016, 13.3841)$$

respectively. Our point estimate and $1 - \alpha = 95\%$ confidence interval for $E(X_1)$ are:

$$(x_1, x_2, \dots, x_7) = (2, 12, 8, 9, 14, 15, 11), \quad x_7 = 10.143, \quad s_7^2 = 19.143, \quad s_7 = 4.375,$$

and observed the following data, sample mean, sample variance and sample standard deviation:

$$X_1, X_2, \dots, X_n, \quad \text{Exponentia}(x^*)$$

Example 167 We model the waiting times between Orbiter buses with unknown $E(X_1) = 1/\alpha^*$

Let's return to our two examples again.

where, $S_n = \sqrt{S_n^2}$ is the sample standard deviation.

$$(X_n \pm z_{\alpha/2} S_n / \sqrt{n}) = (X_n - z_{\alpha/2} S_n / \sqrt{n}, X_n + z_{\alpha/2} S_n / \sqrt{n}), \quad (5.11)$$

variable $V(X_1)$ the following 1 - α confidence interval for $E(X_1)$ works as that even when we substitute the sample variance $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} (X_i - \bar{X}_n)^2$ for the population variance $V(X_1)$, the following 1 - α confidence interval for $E(X_1)$, Fortunately, a more elaborate form of the CLT tells us that even when we substitute the sample mean to LNN but we won't be able to get a point estimate of $E(X_1)$ from the sample mean with $V(X_1)$. We can still get a point estimate of $E(X_1)$ in an IID experiment with n samples and tried to estimate the population mean $E(X_1)$. But in general, we will not know $V(X_1)$. So far, we have assumed we know the population variance $V(X_1)$ in an IID experiment with n observations will fail to contain $E(X_1)$.

Remark 71 (Heuristic interpretation of the $(1 - \alpha)$ confidence interval) If we repeatedly produced samples of size n to contain $E(X_1)$ within the random interval and 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 $E(X_1)$.

$$\begin{aligned} P(E(X_1) \in (X_n - z_{\alpha/2} \sqrt{V(X_1)/n}, X_n + z_{\alpha/2} \sqrt{V(X_1)/n})) &= 1 - \alpha \\ P(X_n - z_{\alpha/2} \sqrt{V(X_1)/n} < E(X_1) < X_n + z_{\alpha/2} \sqrt{V(X_1)/n}) &= 1 - \alpha \\ P(X_n + z_{\alpha/2} \sqrt{V(X_1)/n} < E(X_1) < X_n - z_{\alpha/2} \sqrt{V(X_1)/n}) &= 1 - \alpha \\ P(-X_n - z_{\alpha/2} \sqrt{V(X_1)/n} < E(X_1) < -X_n + z_{\alpha/2} \sqrt{V(X_1)/n}) &= 1 - \alpha \\ P(-z_{\alpha/2} < Z < z_{\alpha/2}) &= 1 - \alpha \end{aligned}$$

We can easily see how Equation (5.10) is derived from CLT as follows:

$$(X_n \pm z_{\alpha/2} \sqrt{V(X_1)/n}) = (X_n - z_{\alpha/2} \sqrt{V(X_1)/n}, X_n + z_{\alpha/2} \sqrt{V(X_1)/n}). \quad (5.10)$$

A useful byproduct of the CLT is the $(1 - \alpha)$ confidence interval, a random interval (or bivariate RV) that contains $E(X_1)$, the quantity of interest, with probability $1 - \alpha$:

5.3.2 Application: Set Estimation of $E(X_1)$

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 $\omega = HHT$, and another distinct outcome $\omega' = HHH$. An event, say A , that ‘at least two Heads occur’ is the following subset of Ω :

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

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

$$B = \{TTT\}$$

Note that the event B is also an outcome or sample point. Another interesting event is the empty set $\emptyset \subset \Omega$. The event that ‘nothing in the sample space occurs’ is \emptyset .

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,

$$P(\text{error} < 0.1) \cong P\left(-\frac{0.1}{\sqrt{4/80}} < Z < \frac{0.1}{\sqrt{4/80}}\right) = P(-0.447 < Z < 0.447) = 0.345 .$$

Suppose you want the error to be less than tolerance = ϵ 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** = ϵ 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

$$n = \left(\frac{\sqrt{4} \times 1.96}{0.1} \right)^2 \cong 1537$$

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.

$$L = \frac{u}{u} = (u, L \cap H)N$$

1. **Something Happens:** Each time we toss a coin, we are certain to observe Heads or Tails, denoted by H or T. The probability that "something happens" is 1. More formally:

Idea 9 (The long-term relative frequency (LTF) 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 think that it is fair had we observed $N(H, n) \leftarrow 0.5$ as $n \rightarrow \infty$.

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 actions 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.

2.2 Probability

Experiment	- an activity producing distinct outcomes.
Sample Space	- set of all outcomes of the experiment.
Event	- set of outcomes of the experiment.
Simple Event	- an individual outcome in S , called a simple event.
Compound Event	- an event consisting of two or more simple events.
Complement	- the set of all outcomes in S that are not in E .
Intersection	- the set of outcomes common to two or more events.
Union	- the set of outcomes belonging to either one or both of two events.
Disjoint Events	- two events with no outcomes in common.

EXPERIMENT SUMMARY

Classwork 2.1, (A tree-diagramming tree of outcomes) Can you think of a graphic way to enumerate the outcomes of the Experiment 2? 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).

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

$$\begin{aligned} & \left(\frac{u/(^1X)\Lambda^\wedge}{\vartheta} > Z > \frac{u/(^1X)\Lambda^\wedge}{\vartheta} - \right) d = \\ & \left(\frac{u/(^1X)\Lambda^\wedge}{\vartheta} > \frac{u/(^1X)\Lambda^\wedge}{(^1X)\bar{E} - u\bar{X}} > \frac{u/(^1X)\Lambda^\wedge}{\vartheta} - \right) d \approx (\vartheta > (^1X)\bar{E} - u\bar{X} > \vartheta -)d \end{aligned}$$

To be able to do this we needed to know the full distribution of $X_n - E(X_1)$. Instead due to the Central Limit Theorem (CLT) we now know that (assuming n is large)

error < tolerance) = $P(|X^u - E(X^l)| > \epsilon)$ = $P(-\epsilon < X^u - E(X^l) < \epsilon)$.

Recall that we wanted to ensure the error $= |X_n - E(X_1)|$ in our estimate of $E(X_1)$ is within a required tolerance = ϵ and make the following probability statement:

5.3.1 Application: Tolerating Errors in our estimate of $E(X_1)$

For the last limit we have used $(1 + \frac{u}{x})^x \rightarrow e^u$ as $n \rightarrow \infty$. Thus, we have proved Equation (5.8) which is equivalent to Equation (5.7) by a standardization argument that if $W \sim \text{Normal}(u, \sigma^2)$ then $Z = \frac{W - u}{\sigma} \sim \text{Normal}(0, 1)$ through the linear transformation $W = \sigma Z + u$ of Example 6.7.

which implies

$$= 1 + \mu E(Y) + o(\mu)$$

and since we can Taylor expand $\phi y(t)$ as follows:

$$\cdot \quad u\left(\left(\frac{u^\wedge}{\tau}\right)\lambda\phi\right) = (\tau)^u\lambda\phi$$

So, the CE of U_n is

$$\mathbb{E}(Y) = 0, \quad E(Y^2) = 1, \text{ and} \quad V(Y) = 1.$$

then

$$\frac{({}^1X)\Lambda\wedge}{({}^1X)\mathcal{H}-{}^1X} = \Lambda$$

Now, if we let

$$\left(\left(\frac{({}^1X)\Lambda}{({}^1X)\mathcal{E}} \wedge \frac{u\wedge}{\not u} i \right) dx \right) \mathcal{E} \prod_u^{1=\eta} = \left(\left(\frac{({}^1X)\Lambda}{({}^1X)\mathcal{E}} \sum_u^{1=\eta} \frac{u\wedge}{\not u} i \right) dx \right) \mathcal{E} = ((u_{\mathcal{D}} u) dx) \mathcal{E} = (i)^u \Lambda \phi$$

Therefore, the CF of U_n is

$$\left(\frac{({}^1X)\Lambda \wedge}{({}^1X)E - {}^q X} \right) \sum_u \frac{u \wedge}{1} = \frac{({}^1X)\Lambda u \wedge}{({}^1X)Eu - {}^q X} \underset{1=u}{\sum} = \frac{u/({}^1X)\Lambda \wedge}{({}^1X)E - {}^u X} =: {}^u \varPi$$

Second,

This is an intuitively reasonable assumption that simply says that one of the possible outcomes is certain to occur, provided the coin is not so thick that it can land on or even roll along its circumference.

2. **Addition Rule:** Heads and Tails are mutually exclusive events in any given toss of a coin, i.e. they cannot occur simultaneously. The intersection of mutually exclusive events is the empty set and is denoted by $H \cap T = \emptyset$. The event $H \cup T$, namely that the event that “coin lands Heads **or** coin lands Tails” satisfies:

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

3. The coin-tossing experiment is repeatedly performed in an **independent** manner, i.e. the outcome of any individual coin-toss does not affect that of another. This is an intuitively reasonable assumption since the coin has no memory and the coin is tossed identically each time.

We will use the LTRF idea more generally to motivate a mathematical model of probability called probability model. Suppose A is an event associated with some experiment \mathcal{E} , so that A either does or does not occur when the experiment is performed. We want the probability that event A occurs in a specific performance of \mathcal{E} , denoted by $\mathbf{P}(A)$, to intuitively mean the following: if one were to perform a super-experiment \mathcal{E}^∞ by independently repeating the experiment \mathcal{E} and recording $N(A, n)$, the fraction of times A occurs in the first n performances of \mathcal{E} within the super-experiment \mathcal{E}^∞ . Then the LTRF idea suggests:

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

Now, we are finally ready to define probability.

Definition 10 (Probability) Let \mathcal{E} be an experiment with sample space Ω . Let \mathcal{F} denote a suitable collection of events in Ω that satisfy the following conditions:

1. It (the collection) contains the sample space: $[\Omega \in \mathcal{F}]$.
2. It is closed under complementation: $[A \in \mathcal{F} \implies A^c \in \mathcal{F}]$.
3. It is closed under countable unions: $A_1, A_2, \dots \in \mathcal{F} \implies \bigcup_i A_i := A_1 \cup A_2 \cup \dots \in \mathcal{F}$.

Formally, this collection of events is called a **sigma field** or a **sigma algebra**. Our experiment \mathcal{E} has a sample space Ω and a collection of events \mathcal{F} that satisfy the three condition.

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

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

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

trajectories for simulated tosses of a fair coin from IID Bernoulli($\theta^* = 1/2$) RVs and the twenty red sample mean trajectories for simulated waiting times from IID Exponential($\lambda^* = 1/10$) RVs in Figure 5.4 with $n = 7$. Clearly, the point estimates for such a small sample size are fluctuating wildly! However, the fluctuations in the point estimates settles down for larger sample sizes.

The *next natural question is how large should the sample size be in order to have a small interval of width, say 2ϵ , “contain” $E(X_1)$, the quantity of interest, with a high probability, say $1 - \alpha$?* If we can answer this then we can make probability statements like the following:

$$P(\text{error} < \text{tolerance}) = P(|\bar{X}_n - E(X_1)| < \epsilon) = P(-\epsilon < \bar{X}_n - E(X_1) < \epsilon) = 1 - \alpha .$$

In order to ensure the $\text{error} = |\bar{X}_n - E(X_1)|$ in our estimate of $E(X_1)$ is within a required 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 $\text{Normal}(0, 1/3)$ RV, where $X_i \stackrel{\text{IID}}{\sim} \text{Uniform}(-1, 1)$. Hint: Characteristic functions

Proposition 70 (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 $\mathbf{E}(X_1)$ by convention, and variance given by $\frac{1}{n}$ times the variance of any one of the IID RVs, say $V(X_1)$ by convention. More formally, we write:

If $X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1$ and if $\mathbf{E}(X_1) < \infty, V(X_1) < \infty$

$$\text{then } \bar{X}_n \rightsquigarrow \text{Normal}\left(\mathbf{E}(X_1), \frac{V(X_1)}{n}\right) \text{ as } n \rightarrow \infty , \quad (5.7)$$

or equivalently after standardization:

If $X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1$ and if $\mathbf{E}(X_1) < \infty, V(X_1) < \infty$

$$\text{then } \frac{\bar{X}_n - \mathbf{E}(X_1)}{\sqrt{V(X_1)/n}} \rightsquigarrow Z \sim \text{Normal}(0, 1) \text{ as } n \rightarrow \infty . \quad (5.8)$$

Proof: Our proof is based on the convergence of characteristic functions (CFs). We will prove the standardized form of the CLT in Equation (5.8) by showing that the CF of

$$U_n := \frac{\bar{X}_n - \mathbf{E}(X_1)}{\sqrt{V(X_1)/n}}$$

converges to the CF of Z , the $\text{Normal}(0, 1)$ RV. First, note from Equation (3.70) that the CF of $Z \sim \text{Normal}(0, 1)$ is:

$$\phi_Z(t) = E(e^{itZ}) = e^{-t^2/2} .$$

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), are logical consequences of our definition of probability:

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

Proof. One line proof:

$$1. \text{ For any event } A, \boxed{P(A^c) = 1 - P(A)}.$$

2.2.1 Consequences of our Definition of Probability

Remark 69 (Point estimates are realizations of the Point Estimator) We say the statistic X_1, X_2, \dots, X_n , which is a random variable that depends on the data RV , i.e., our observed data estimator of $E(X_1)$. But once we have a realization of the data RV , i.e., our observed data sample (x_1, x_2, \dots, x_n) and its corresponding realization as observed sample mean \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 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 $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$, our point estimator of $E(X_1)$ is still \bar{X}_n but the point estimate \bar{x}_n , may be different from the first point estimate $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n$, that is different from our first data vector $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$, our point estimator of $E(X_1)$ is \bar{X}_n , where \bar{X}_n is the same as the point estimate \bar{x}_n from Figure 5.4.

Typically, we do not know the “true” parameter θ^* $\in [0, 1]$, which is the same as the population mean $E(X_1) = \theta^*$. But by LNN, we know that

Then you can use the observed sample mean $\bar{x}_7 = (2 + 12 + 8 + 9 + 14 + 15 + 11)/7 = 71/7 \approx 10.14$

$$(x_1, x_2, \dots, x_7) = (2, 12, 8, 9, 14, 15, 11).$$

and have the following realization as your observed data:

$$X_1, X_2, \dots, X_7 \stackrel{iid}{\sim} \text{Exponential}(\lambda^*) ,$$

Now, suppose you model seven waiting times in nearest minutes between Orbitr buses at Balgay street as follows:

and therefore, we can use the sample mean \bar{x}_7 as a point estimator of $E(X_1) = 1/\lambda^*$.

events A_1, A_2, \dots, A_n as follows:

5. Once again by the inclusion-exclusion principle, the Boolean inequality generalizes to any n events A_1, A_2, \dots, A_n as follows:

$$\begin{aligned} \Pr(A_1 \cup A_2 \cup A_3) &= \Pr(A_1) + \Pr(A_2) + \Pr(A_3) - \Pr(A_1 \cap A_2) - \Pr(A_1 \cap A_3) - \Pr(A_2 \cap A_3) + \Pr(A_1 \cap A_2 \cap A_3) \\ &= \sum_{i=1}^3 \Pr(A_i) - \sum_{i < j < k} \Pr(A_i \cap A_j \cap A_k) + \sum_{i < j < k < l} \Pr(A_i \cap A_j \cap A_k \cap A_l) - \dots + (-1)^{n-1} \Pr(A_1 \cap A_2 \cap \dots \cap A_n) \end{aligned}$$

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

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

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

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

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

$$\Pr(A) = \Pr(A \setminus B) + \Pr(A \cup B)$$

the addition rule implies that:

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

Proof. Since:

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

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

- If $A = \emptyset$ then $A^c = \emptyset$ and $\Pr(\emptyset) = 1 - P(\emptyset) = 1 - 1 = 0$.

Proof. One line proof:

$$1. \text{ For any event } A, \boxed{P(A^c) = 1 - P(A)}.$$

Now, suppose you model seven waiting times in nearest minutes between Orbitr buses at Balgay street as follows:

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 orbitr 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 which is the same as the probability of Heads is $\bar{x}_7 = 3/7$.

Then you can use the observed sample mean $\bar{x}_7 = (0 + 1 + 0 + 0 + 0 + 1 + 0)/7 = 3/7 \approx 0.4286$ as a **point estimate** of the population mean $E(X_1) = \theta^*$. Thus, our “single best guess” for $E(X_1)$ and have the following realization as your observed data:

$$X_1, X_2, \dots, X_7 \stackrel{iid}{\sim} \text{Ber}(0^*) ,$$

with probability $1 - \theta^*$ as follows:

Now, suppose you model seven coin tosses (including Heads as 1 with probability θ^* and Tails as 0 and therefore, we can use the sample mean \bar{x}_7 as a point estimator of $E(X_1) = \theta^*$.

$$X_1, X_2, \dots, X_7 \stackrel{iid}{\sim} \text{Point Mass}(E(X_1)) ,$$

mean $E(X_1) = \theta^*$. But by LNN, we know that

$$X_1, X_2, \dots, X_7 \stackrel{iid}{\sim} \text{Ber}(\theta^*) .$$

Example 164 Let X_1, X_2, \dots, X_n be an $\text{Ber}(\theta)$ RV, i.e., let

Typically distinctly especially for small n as shown in Figure 5.4.

The sample mean \bar{x}_n from n samples for 20 replications (repetats of the experiment) are x_1, x_2, \dots, x_n . The sample mean \bar{x}_n from $n-1$ samples for 20 replications (repetats of the experiment) are x_1, x_2, \dots, x_{n-1} , but the point estimate \bar{x}_{n-1} may be different from the first point estimate $E(X_1)$ is still \bar{X}_n , but the point estimate from our first data vector $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$, our point estimator of $E(X_1)$ is \bar{X}_n , that is different from our first data vector (x_1, x_2, \dots, x_n) , that is different from our first data vector $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$, our point estimator of $E(X_1)$. In other words, the point estimate \bar{x}_n is a realization of the random variable \bar{X}_n called the point estimator of $E(X_1)$. Therefore, when we observe a new random variable \bar{X}_n estimated the point estimator of $E(X_1)$. In other words, the point estimate \bar{x}_n is a realization of the data RV , i.e., our observed data vector (x_1, x_2, \dots, x_n) and its corresponding realization as observed sample mean \bar{x}_n is a point estimate of $E(X_1)$. But once we have a realization of the data RV , i.e., our observed data estimator of $E(X_1)$. By rearranging $\bar{x}_n = 1/E(X_1)$, we can also obtain a point estimate of the “true” parameter λ^* from $1/\bar{x}_7 = 7/7 \approx 0.0986$.

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

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

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:

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

The n -event case follows by mathematical induction.

We have formally defined the **probability model** specified by the **probability triple** (Ω, \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:

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

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, Ω here is indeed a more complicated set although it only leads to 40 possible outcomes.

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

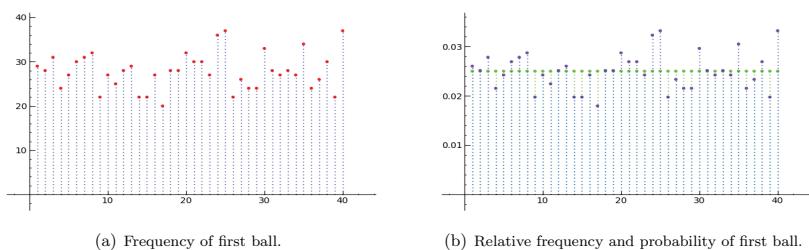


Figure 2.2: First ball number in 1114 NZ Lotto draws from 1987 to 2008.

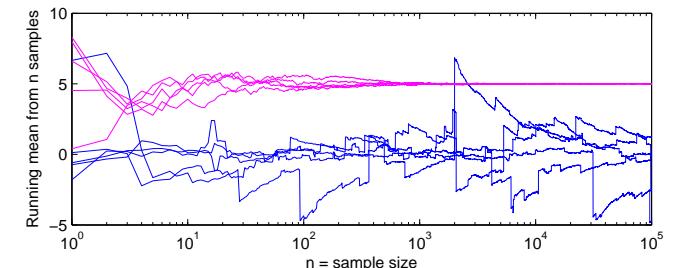
Next, let us take a detour into how one might interpret it in the real world. The following is an adaptation from Williams D, *Weighing the Odds: A Course in Probability and Statistics*, Cambridge University Press, 2001, which henceforth is abbreviated as WD2001.

```

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 $X_1, X_2, \dots, X_n \stackrel{IID}{\sim} X_1$.

Example 163 Let $X_1, X_2, \dots, X_n \stackrel{IID}{\sim} X_1$, where X_1 is an $\text{Exponential}(\lambda^*)$ RV, i.e., let

$$X_1, X_2, \dots, X_n \stackrel{IID}{\sim} \text{Exponential}(\lambda^*) .$$

Typically, we do not know the “true” parameter $\lambda^* \in \Lambda = (0, \infty)$ or the population mean $E(X_1) = 1/\lambda^*$. But by LLN, we know that

$$\bar{X}_n \rightsquigarrow \text{Point Mass}(E(X_1)) ,$$

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

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 Ω (think of intersecting an arbitrary family of sigma algebras)?

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

1. When the sample space $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ is a finite set with k outcomes and $\mathbf{P}(\omega_i)$, the probability for each outcome $\omega_i \in \Omega$ is known, then one typically takes the sigma-algebra \mathcal{F} to be the set of all subsets of Ω called the **power set** and denoted by 2^Ω . The probability of each event $A \in 2^\Omega$ can be obtained by adding the probabilities of the outcomes in A , i.e., $\mathbf{P}(A) = \sum_{\omega_i \in A} \mathbf{P}(\omega_i)$. Clearly, 2^Ω is indeed a sigma-algebra and it contains $2^{\#\Omega}$ events in it.
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 Ω . Note that this is very similar to the case with finite Ω except now $\mathcal{F} = 2^\Omega$ could have uncountably many events in it.
3. If $\Omega = \mathbb{R}^d$ for finite $d \in \{1, 2, 3, \dots\}$ then the **Borel sigma-algebra** is the smallest sigma-algebra containing all **half-spaces**, i.e., sets of the form

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

When $d = 1$ the half-spaces are the half-lines $\{(-\infty, c] : c \in \mathbb{R}\}$ and when $d = 2$ the half-spaces are the south-west quadrants $\{(-\infty, c_1] \times (-\infty, c_2] : (c_1, c_2) \in \mathbb{R}^2\}$, etc. (Equivalently, the Borel sigma-algebra is the smallest sigma-algebra containing all open sets in \mathbb{R}^d).

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

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

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

- **A most primitive sigma-algebra for probability theory:** For example if $\mathbb{S} = \{0, 1\}$, then $\Omega = \{0, 1\}^\infty$ is the set of all infinite sequences made of 0's and 1's. To take advantage of arithmetic and analysis, Ω can be seen as the binary representation of all real numbers in the unit interval $[0, 1]$. We can take advantage of combinatorics and algebra if we further represent the dyadic partition of $[0, 1]$ by a binary tree (as drawn in lectures). Then, a cylinder set such as $1 \times 1 \times 0 \times \{0, 1\} \times \{0, 1\} \times \dots$, 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.

Therefore, for any given $\epsilon > 0$,

$$\begin{aligned} \mathbf{P}(|\bar{X}_n - \mathbf{E}(X_1)| \geq \epsilon) &= \mathbf{P}(|\bar{X}_n - \mathbf{E}(\bar{X}_n)| \geq \epsilon) \quad [\text{by the IID assumption of } X_1, X_2, \dots, \mathbf{E}(\bar{X}_n) = \mathbf{E}(X_1), \text{ as per (3.74)}] \\ &= \frac{\frac{1}{n}\mathbf{V}(X_1)}{\epsilon^2} \rightarrow 0, \quad \text{as } n \rightarrow \infty, \end{aligned}$$

or equivalently, $\lim_{n \rightarrow \infty} \mathbf{P}(|\bar{X}_n - \mathbf{E}(X_1)| \geq \epsilon) = 0$. And the last statement is the definition of the claim made by the law of large numbers (LLN), namely that $\bar{X}_n \xrightarrow{\mathbf{P}} \mathbf{E}(X_1)$.

Proposition 68 (Weak Law of Large Numbers (WLLN): $\bar{X}_n \rightsquigarrow \text{Point Mass}(\mathbf{E}(X_1))$) 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 $\mathbf{E}(X_1)$ exists, i.e. $\mathbf{E}(\text{abs}(X)) < \infty$, then the sample mean \bar{X}_n converges in distribution to the expectation of any one of the IID RVs, say $\text{Point Mass}(\mathbf{E}(X_1))$ by convention. More formally, we write:

If $X_1, X_2, \dots \stackrel{\text{IID}}{\sim} X_1$ and if $\mathbf{E}(X_1)$ exists, then $\bar{X}_n \rightsquigarrow \text{Point Mass}(\mathbf{E}(X_1))$ 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 CFs⁷, the convergence of the corresponding distribution functions (DFs).

First, the CF of $\text{Point Mass}(\mathbf{E}(X_1))$ is

$$\mathbf{E}(e^{it\mathbf{E}(X_1)}) = e^{it\mathbf{E}(X_1)},$$

since $\mathbf{E}(X_1)$ is just a constant, i.e., a Point Mass RV that puts all of its probability mass at $\mathbf{E}(X_1)$.

Second, the CF of \bar{X}_n is

$$\begin{aligned} \mathbf{E}(e^{it\bar{X}_n}) &= \mathbf{E}\left(e^{it\frac{1}{n}\sum_{k=1}^n X_k}\right) = \mathbf{E}\left(\prod_{k=1}^n e^{itX_k/n}\right) = \prod_{k=1}^n \mathbf{E}\left(e^{itX_k/n}\right) = \prod_{k=1}^n \phi_{X_1}(t/n) \\ &= \prod_{k=1}^n \phi_{X_1}(t/n) = (\phi_{X_1}(t/n))^n. \end{aligned}$$

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_ϵ such that for all $x > x_\epsilon$ $|f(x)| < \epsilon |g(x)|$. For example, $\log(x)$ is $o(x)$, x^2 is $o(x^3)$ and x^m is $o(x^{m+1})$ for $m \geq 1$.

Third, we can expand any CF whose expectation exists as a Taylor series with a remainder term that is $o(t)$ as follows:

$$\phi_X(t) = 1 + it\mathbf{E}(X) + o(t).$$

Hence,

$$\phi_{X_1}(t/n) = 1 + it\frac{1}{n}\mathbf{E}(X_1) + o\left(\frac{t}{n}\right)$$

and

$$\mathbf{E}(e^{it\bar{X}_n}) = \left(1 + it\frac{1}{n}\mathbf{E}(X_1) + o\left(\frac{t}{n}\right)\right)^n \rightarrow e^{it\mathbf{E}(X_1)} \text{ as } n \rightarrow \infty.$$

⁷https://en.wikipedia.org/wiki/L%C3%A9vy_continuity_theorem

probability
the derivative case

[https://en.wikipedia.org/wiki/Proofs_of_convergence_in_probability](https://en.wikipedia.org/wiki/Proofs_of_convergence_in_probability#Convergence_in_probability)

[https://en.wikipedia.org/wiki/Proofs_of_convergence_in_probability](https://en.wikipedia.org/wiki/Proofs_of_convergence_in_probability#Convergence_in_probability)

[https://en.wikipedia.org/wiki/Proofs_of_convergence_in_probability](https://en.wikipedia.org/wiki/Proofs_of_convergence_in_probability#Convergence_in_probability)

Proof: Because $V(X_1) < \infty$, we have:

$$\text{If } X_1, X_2, \dots \text{ iid } X_1 \text{ and if } E(X_1) \text{ exists, then } \overline{X}_n \xrightarrow{P} E(X_1).$$

by applying Chebyshev's inequality (5.6) to the RV \overline{X}_n

[by the IID assumption of X_1, X_2, \dots we can apply (3.75)]

$$= \frac{\epsilon^2}{\mathbb{E}(X_1)}$$

$$P(\overline{X}_n - E(\overline{X}_n) | \geq \epsilon) = \frac{\epsilon^2}{\mathbb{E}(X_1)}$$

More formally, we write:
 X_n converges in probability to the expectation of any one of the IID RVs, say $E(X_1)$ by convention.
 (3.43) , i.e., $E(\text{abs}(X_1)) < \infty$, and the variance is finite, i.e., $V(X_1) < \infty$, then the sample mean it independent and identically distributed RVs, $X_1, X_2, \dots, \overline{X}_n$, and if $E(X_1)$ exists, as per

Proposition 67 (Law of Large Numbers (LLN)) If we are given a sequence

5.2 Law of Large Numbers

• In general, convergence in distribution does not imply convergence in probability.

$$X_n \rightsquigarrow \text{Point Mass}(\theta) \iff X_n \xrightarrow{P} \text{Point Mass}(\theta).$$

• Convergence in distribution to a constant θ implies convergence in probability to θ .⁶

$$X_n \xrightarrow{P} X \iff X_n \rightsquigarrow X.$$

• Convergence in probability implies convergence in distribution⁵

• By the Borel-Cantelli Lemma³, convergence in probability does not imply almost sure convergence in the discrete case⁴.

$$X_n \xrightarrow{a.s.} X \iff X_n \xrightarrow{P} X.$$

• Convergence almost surely implies convergence in probability²

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 implications exist between the three notions.

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

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

3. Rolling a die and recording the number on the upturned face until the first 6 appears.

2. Drawing 4 screws from a bucket of left-handed and right-handed screws denoted by L and R, respectively.

1. Tossing 2 coins whose faces are sprayed with black paint denoted by B and white paint denoted by W.

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

(f) P(picking any letter in the word WAZZZZUP or a vowel).

(e) P(picking any letter in the word WAZZZZUP).

(d) P(picking a vowel).

(c) P({E, Z})

(b) P(picking any letter)

(a) P({Z})

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?

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

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

2.3 Exercises in Probability

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

Rules:

$$3. \text{ If } A_1, A_2, \dots \text{ are disjoint then } P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$$

[This is true only when A and B are disjoint.]

$$2. \text{ If } A, B \text{ are disjoint events, then } P(A \cup B) = P(A) + P(B).$$

$$1. \text{ If } A \subseteq \Omega \text{ then } 0 \leq P(A) \leq 1 \text{ and } P(\Omega) = 1.$$

Axioms:

PROBABILITY SUMMARY

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

- (a) Write down the sample space of this experiment.
- (b) Find the probabilities of each simple event.
- (c) Show that $\mathbf{P}(\Omega)$ is indeed 1.
- (d) Check that the addition rule for mutually exclusive events holds for the simple events $\{B\}$ and $\{I\}$.
- (e) Consider the following events: $C = \{B, I, G\}$ and $D = \{G, I, N\}$. Using the addition rule for two arbitrary events, find $\mathbf{P}(C \cup D)$.

2.4 Conditional Probability

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

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

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

or

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

or

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

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

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

Recall that $N(A, n)$ as defined in (2.1) is the fraction of times A occurs out of n independent repetitions of our experiment \mathcal{E} (ie. the experiment \mathcal{E}^n). If $A \cap B$ is the event that ‘ A and B occur simultaneously’, then we intuitively want

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

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

Proposition 65 (Chebychev’s Inequality) For any RV X and any $\epsilon > 0$,

$$\mathbf{P}(|X| > \epsilon) \leq \frac{\mathbf{E}(|X|)}{\epsilon} \quad (5.4)$$

$$\mathbf{P}(|X| > \epsilon) = \mathbf{P}(X^2 \geq \epsilon^2) \leq \frac{\mathbf{E}(X^2)}{\epsilon^2} \quad (5.5)$$

$$\mathbf{P}(|X - \mathbf{E}(X)| \geq \epsilon) = \mathbf{P}((X - \mathbf{E}(X))^2 \geq \epsilon^2) \leq \frac{\mathbf{E}(X - \mathbf{E}(X))^2}{\epsilon^2} = \frac{\mathbf{V}(X)}{\epsilon^2} \quad (5.6)$$

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 $X_n \xrightarrow{\mathbf{P}} X$?

To find out if $X_n \xrightarrow{\mathbf{P}} X$, we need to show that for any $\epsilon > 0$, $\lim_{n \rightarrow \infty} \mathbf{P}(|X_n - X| > \epsilon) = 0$.

Let ϵ be any real number greater than 0, then

$$\begin{aligned} \mathbf{P}(|X_n| > \epsilon) &= \mathbf{P}(|X_n|^2 > \epsilon^2) \\ &\leq \frac{\mathbf{E}(X_n^2)}{\epsilon^2} \quad [\text{by Markov's Inequality (5.2)}] \\ &= \frac{1}{\epsilon^2} \rightarrow 0, \quad \text{as } n \rightarrow \infty \quad [\text{in the sense of Definition 4}]. \end{aligned}$$

Hence, we have shown that for any $\epsilon > 0$, $\lim_{n \rightarrow \infty} \mathbf{P}(|X_n - X| > \epsilon) = 0$ and therefore by Definition 63, $X_n \xrightarrow{\mathbf{P}} X$ or $X_n \xrightarrow{\mathbf{P}} 0$.

Convention: When X has a Point Mass(θ) distribution and $X_n \xrightarrow{\mathbf{P}} X$, we simply write $X_n \xrightarrow{\mathbf{P}} \theta$.

Definition 66 (Convergence Almost Surely (or with Probability 1)) To say that the sequence of RVs $\{X_n\}_{n=1}^\infty$ converges almost surely (or with probability 1 or strongly) towards another RV X on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$, as denoted by

$$X_n \xrightarrow{a.s.} X$$

means that

$$\mathbf{P}\left(\left\{\lim_{n \rightarrow \infty} X_n = X\right\}\right) = 1 \iff \mathbf{P}\left(\left\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

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.

	Have Disease (D)	Don't have disease (D^c)
Test positive (+)	0.009	0.99
Test negative (-)	0.001	0.981

probabilities are:

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

$P(A \cup B) = P(A)P(B A) = P(B)P(A B)$.
If A and B are events, and if $P(A) \neq 0$ and $P(B) \neq 0$, then
Multiplication rule for two likely events:

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

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

Addition rule for two arbitrary events B_1 and B_2 :

$$\text{Complement rule: } P(B|A) = 1 - P(B^c|A).$$

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

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

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

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

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

$$\text{Axiom (2): } P(\Omega | A) = 1 \quad \text{Meaning: Something Happens given the event A happens.}$$

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

satisfied:

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

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

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:

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

conditional probability of B given A ,

Definition 12 (Conditional Probability) Suppose we are given an experiment \mathcal{E} with a triple (Ω, \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

as $n \rightarrow \infty$, the expression below the first overbrace $\rightarrow 1$, while the second overbrace approaches 1. Finally, we get the desired limit:

Let us look at some immediate consequences of Markov's inequality.

$$E(X) \geq E(I_{\{X \geq \epsilon\}}(x)) = P(X \geq \epsilon).$$

get the desired result:

expectation of an indicator function of an event is simply the probability of that event (3.44), we finally, taking expectations on both sides of the above inequality and then using the fact that the

(5.3)

$$\begin{aligned} &\geq P_{\{X \geq \epsilon\}}(x) \\ &\geq X_{\{X \geq \epsilon\}}(x) \\ X &= X_{\{X \geq \epsilon\}}(x) + X_{\{X < \epsilon\}}(x) \end{aligned}$$

Proof:

$$P(X \geq \epsilon) \leq \frac{\epsilon}{E(X)}, \quad \text{for any } \epsilon < 0. \quad (5.2)$$

be a non-negative RV. Then,

Proposition 64 (Markov's Inequality) Let (Ω, \mathcal{F}, P) be a probability triple and let $X = X(\omega)$ need some elementary inequalities in probability to help us answer this question. We visit these measures next.

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 measures next.

$$\lim_{n \rightarrow \infty} P(\{\omega : |X_n(\omega) - X(\omega)| < \epsilon\}) = 0, \quad \text{ie, } P(\{\omega : |X_n(\omega) - X(\omega)| < \epsilon\}) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

Once again, the above limit, by (3.1) in our Definition 18 of a RV, can be equivalently expressed as follows:

$$\lim_{n \rightarrow \infty} P(|X_n - X| < \epsilon) = 0 \quad \text{[in the sense of Definition 4].}$$

If for every real number $\epsilon > 0$,

$$X_n \xrightarrow{P} X$$

converges to X in probability, and write:

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 in probability to X . Definition 63 (Convergence in Probability) Let X_1, X_2, \dots , be a sequence of RVs and let X be a sequence of convergent RVs in probability in the sense of Definition 4.

The second notion of convergence of RVs is convergence in probability.

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

approaches 1. Finally, we get the desired limit:

As $n \rightarrow \infty$, the expression over the first underbrace approaches $e^{-\lambda}$ while that over the second underbrace 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:

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

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

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

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

Thus, the test is quite accurate since sick people test positive 90% of the time and healthy people test negative 90% of the time.

Now, suppose you go for a test and test positive. What is the probability that you have the disease ?

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

Most people who are not used to the definition of conditional probability would intuitively associate a number much bigger than 0.08 for the answer. Interpret conditional probability in terms of the meaning of the numbers that appear in the numerator and denominator of the above calculations.

2.4.1 Bayes' Theorem

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

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

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

Proposition 13 (Bayes' Rule)

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

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?

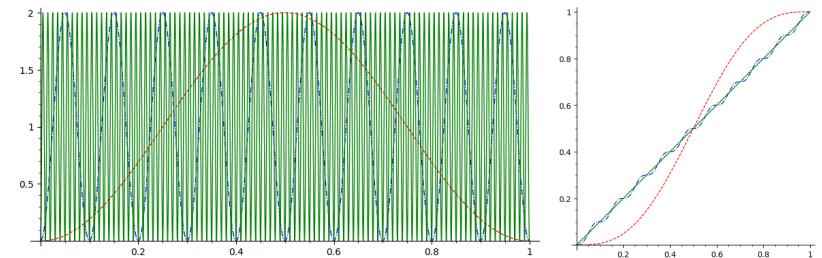


Figure 5.3: $f_{X_n}(x) := \mathbf{1}_{(0,1)}(x)(1 - \cos(2\pi nx))$ of the RV X_n [the left sub-figure] and its DF $F_n(x) := \int_{-\infty}^x \mathbf{1}_{(0,1)}(v)(1 - \cos(2\pi nv))dv$ [the right sub-figure], for $n = 1$ [red '---'], $n = 10$ [blue '-.-'], and $n = 100$ [green '-'], respectively. One can see clear convergence of the DFs F_n to $\mathbf{1}_{(0,1)}(x)x$, the DF of the Uniform(0,1) RV, while the corresponding PDFs $f_n(x)$ keep oscillating wildly with n across $[0, 2]$ about $\mathbf{1}_{(0,1)}(x)$, the PDF of the Uniform(0,1) RV X . Thus giving a counter-example to the claim that convergence in DFs does not imply convergence in PDFs.

Since $F(x) = \mathbf{P}(X \leq x)$, 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 $\lim_{n \rightarrow \infty} \mathbf{P}(X_n = x) = \mathbf{P}(X = x)$ for each $x \in \mathbb{Z}_+$. We will use this fact to prove why we can approximate Binomial RVs by a Poisson under some limiting conditions.

Example 159 ($\text{Binomial}(n, \lambda/n) \rightsquigarrow \text{Poisson}(\lambda)$) 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.

$\boxed{\text{Binomial}(n, \lambda/n) \text{ converges in distribution to } \text{Poisson}(\lambda) \text{ as } n \rightarrow \infty, \theta = \lambda/n \rightarrow 0}$

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 λ/n , approaches the Poisson distribution with expectation λ , as n approaches ∞ (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 $\lim_{n \rightarrow \infty} \mathbf{P}(X_n = x) = \mathbf{P}(Y = x) = e^{-\lambda} \lambda^x / x!$ for any $x \in \{0, 1, 2, 3, \dots, n\}$.

$$\begin{aligned} \mathbf{P}(X = x) &= \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \frac{n(n-1)(n-2)\cdots(n-x+1)}{x(x-1)(x-2)\cdots(2)(1)} \left(\frac{\lambda^x}{n^x}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \\ &= \overbrace{\binom{n}{x} \left(\frac{n-1}{n}\right) \left(\frac{n-2}{n}\right) \cdots \left(\frac{n-x+1}{n}\right)}^{\left(\frac{\lambda^x}{x!}\right)} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\left(\underbrace{\left(1 - \frac{\lambda}{n}\right)^{-x}}_{\left(\frac{\lambda}{n}\right)^x}\right)} \end{aligned} \quad (5.1)$$

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.

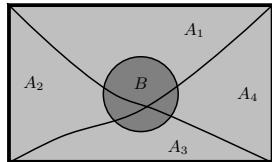


Figure 2.3: Reference to the Venn diagram will help you understand this idea behind the proof of the total probability theorem in Proposition 14 for the four event case.

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

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

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.

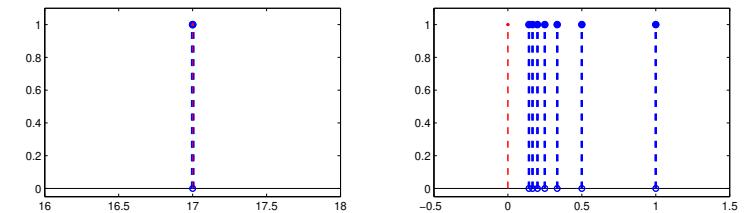
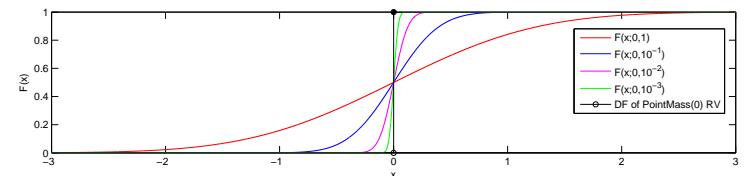


Figure 5.2: Distribution functions of several $\text{Normal}(\mu, \sigma^2)$ RVs for $\sigma^2 = 1, \frac{1}{10}, \frac{1}{100}, \frac{1}{1000}$.



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 61 (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:

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbf{P}(\{\omega : X_n(\omega) \leq t\}) &= \mathbf{P}(\{\omega : X(\omega) \leq t\}), \\ \text{i.e. } \mathbf{P}(\{\omega : X_n(\omega) \leq t\}) &\rightarrow \mathbf{P}(\{\omega : X(\omega) \leq t\}), \quad \text{as } n \rightarrow \infty . \end{aligned}$$

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 .

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

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

$$\begin{aligned} P(A_h|B) &= \frac{P(A_h \cap B)}{P(B)} = \frac{P(B|A_h)P(A_h)}{P(B)} = \frac{P(B|A_h)P(A_h)}{\sum_{j=1}^k P(B|A_j)P(A_j)} \\ &= \frac{P(B|A_h)P(A_h)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_k)P(A_k)} \end{aligned}$$

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

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

Let $B \in \mathcal{F}$ be some event with $P(B) > 0$, then

Thus, precisely one of the A_h 's will occur on any performance of our experiment \mathcal{E} .

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

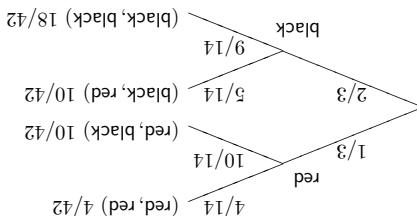
Proposition 15 (Bayes, Theorem, 1763) Suppose the events $A_1, A_2, \dots, A_k \in \mathcal{F}$, with $P(A_h) > 0$ for each $h \in \{1, 2, \dots, k\}$, partition the sample space Ω , i.e. they are mutually exclusive (disjoint) and exhaustive events with positive probability:

$$= (4/14)(1/3) + (5/14)(2/3) = 1/3.$$

$$\begin{aligned} P(R_2) &= P(R_2 \cap R_1) + P(R_2 \cap B_1) \\ &= P(R_2|R_1)P(R_1) + P(R_2|B_1)P(B_1) \end{aligned}$$

Now R_1 and B_1 partition Ω so we can write:

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



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 the number in its support, such as 0.

In other words, a continuous RV, such as X_n , has 0 probability of realizing any single real large. 0, as depicted in Figure 5.2. Based on this observation, can we expect $\lim_{n \rightarrow \infty} X_n = X$, where the masses of X_n increases significantly compared to 0 ? Take a look at Figure 5.2 for instance. The probability of $X_n \sim \text{Normal}(0, 1/n)$ as n approaches ∞ ? The probability of $X_n \sim \text{Normal}(0, 1/n)$ as n approaches $1/n$ approaches 1 . How would you talk about the convergence sequence of RVs $\{X_n\}_{n=1}^{\infty}$, where $X_n \sim \text{Normal}(0, 1/n)$? Suppose you are given an independent sequence of RVs $\{X_i\}_{i=1}^{\infty}$. Suppose you are given the realis! See Figure 5.1.

Yes why not – just move to space of distributions over the reals! See Figure 5.1. Examples 12 and 13?

Can the sequences of $\{\text{Point Mass}(\theta_i = 17)\}_{i=1}^{\infty}$ and $\{\text{Point Mass}(\theta_i = 1/\epsilon)\}_{i=1}^{\infty}$ RVs be the same as the two sequences of real numbers $\{x_i\}_{i=1}^{\infty} = 17, 17, 17, \dots$ and $\{x_i\}_{i=1}^{\infty} = \frac{1}{1}, \frac{2}{1}, \frac{3}{1}, \dots$ we saw in Section 1.8.1 before proceeding further.

Let us first refresh ourselves with notions of convergence, limits and continuity in the real line (See 1.8.1) before proceeding further.

Convergence_of_random_variables.

We need different notions of convergence to characterize such a behavior: two simplest behaviors are that the sequence eventually takes a constant value θ , i.e. X_n approaches $X \sim \text{Point Mass}(\theta)$ are that values in the sequence converge to a constant value θ , i.e. X_n approaches $X \sim F(x)$. We need different notions of convergence to characterize Theory I course, what happens to the limiting RV $X = \lim_{n \rightarrow \infty} X_n$ when given the DFs $F_n(x)$ for each X_n is associated with the amount of information $\leftarrow \infty$. More abstractly, we are interested in what happens to the amount of data or information $\leftarrow \infty$. More abstractly, we are interested in is associated with the amount of data or information, as you will see in Inference Theory I course, $n \leftarrow \infty$.

From a statistical or decision-making viewpoint, as you will see in Inference Theory I course, $n \leftarrow \infty$ is 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 (Ω, \mathcal{F}, P) .

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 (Ω, \mathcal{F}, P) .

5.1 Convergence of Random Variables

Limits Laws of Statistics

Chapter 5

large. In other words, a continuous RV, such as X_n , has 0 probability of realizing any single real large. 0, as depicted in Figure 5.2. Based on this observation, can we expect $\lim_{n \rightarrow \infty} X_n = X$, where the

number in its support, such as 0.

Limiting RV $X \sim \text{Point Mass}(0)$?

0, as depicted in Figure 5.2. Based on this observation, can we expect $\lim_{n \rightarrow \infty} X_n = X$, where the

masses of $X_n \sim \text{Normal}(0, 1/n)$ as n approaches ∞ ? Take a look at Figure 5.2 for instance. The probability

of $X_n \sim \text{Normal}(0, 1/n)$ as n approaches $1/n$ approaches 1 . How would you talk about the convergence

sequence of RVs $\{X_i\}_{i=1}^{\infty}$, where $X_i \sim \text{Normal}(0, 1/i)$? Suppose you are given an independent

sequence of RVs $\{X_i\}_{i=1}^{\infty}$. Suppose you are given the realis! See Figure 5.1.

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

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

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

Solution:

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

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

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

2.4.2 Independence and Dependence

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

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

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

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

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

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

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

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

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

4. A related distribution, denoted by $N_-(\mu, \tau, \sigma^2)$, is the right-truncated normal distribution truncated on the right at τ . Describe how samples from $N_-(\mu, \tau, \sigma^2)$ can be obtained by simulating from an appropriate left-truncated normal distribution.
5. Write a MATLAB function that provides samples from a truncated normal distribution. The function should have the following inputs: number of samples required, left or right truncation, μ , σ^2 and τ .

4.4 Exercises in Simulation

Ex. 4.1 — Suppose the continuous RV X has PDF:

$$f_X(x) = (\pi(1+x^2))^{-1}$$

Devise an algorithm to transform samples from Uniform(0,1) RV to those from X . Present your answer as pseudo-code.

and therefore A and B are not independent. The reason for the events A and B being dependent and therefore A and B are not independent. The chance of getting a total of six depends on the outcome of the first die (not because the chance of getting a total of six depends on the outcome of the first die (not

$$\{B\} = \mathbf{P}\{1, 5\}, \{2, 4\}, \{3, 3\}, \{4, 2\}, \{5, 1\} \} \mathbf{P}\{\{1, 5\}, \{2, 4\}, \{3, 3\}, \{4, 2\}, \{5, 1\}\} = \frac{36}{5} \times \frac{36}{5} = \frac{36}{5} = \frac{216}{5},$$

but

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

for each $w \in U$. Then

Example 36 (independence and independence) Suppose we toss two fair dice. Let A denote the event that the sum of the die 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 $P(\omega) = 1/36$.

in the sample space \mathcal{U} has equal probability of $\frac{1}{2^m}$ due to independence.

Suppose you toss a fair coin independently m times. Then each of the 2^m possible outcomes

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

(a) Suppose you independently toss a fair die three times. Let D , 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

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

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

trials is independent.

Example 33 (Some Standard Examples) A sequence of events in a sequence of independent

$$\mathbf{P}(A^{i_1} \cap A^{i_2} \cdots \cap A^{i_k}) = \mathbf{P}(A^{i_1})\mathbf{P}(A^{i_2}) \cdots \mathbf{P}(A^{i_k})$$

Definition 17 (**Independence of a sequence of events**) We say that a sequence of events A_1, A_2, \dots are independent if whenever i_1, i_2, \dots, i_k are distinct indices, N , such that $A_{i_1}, A_{i_2}, \dots, A_{i_k}$ are defined (elements of F), then

probability of their joint occurrence $P(A \cup B)$ is simply the product of their individual probabilities provided that the two events are mutually exclusive or if one event does not affect the occurrence of the other.

3. Describe how samples from $N^+(u, \tau, o_2)$ can be obtained by simulating from $N^+(u = 0, \tau, o_2)$ and using location-scale transformation.

7. This is the maximum acceptable percentage of anomalies for the following points:
7 = 0, 0.5, 1, 1.5, 2, 2.5 and 3. What can you conclude about efficiency as τ gets further out

$$\alpha = \frac{2}{\tau + \sqrt{\tau^2 + 4}}$$

^{1.} Show that for simulating from $N^+(\mu = 0, \tau, \sigma^2 = 1)$ where $\tau \geq 0$, the best choice of λ that maximizes the expected acceptance probability for the rejection sampler is given by

$$\cdot \omega^{\bar{f}} \mathbb{I}((\omega - f) \vee -) dx \partial \chi = (\omega \vee |f|) t$$

When $\tau < \mu$, the rejection sample can readily be used to simulate from $N^+(\mu, \tau, \sigma^2)$ by simulating from $\text{Normal}(\mu, \sigma^2)$ until a number larger than τ is obtained. When $\tau > \mu$, however, this can be inefficient and increases rapidly so as to get further out into the right tail. In this case, a more efficient approach is to use the rejection sample with the following exponential distribution as the proposal distribution:

$$\cdot x \in \mathbb{I} \frac{[(\sigma/(n-t))\Phi - 1]\sigma \zeta}{(\sigma \zeta / (n-x) -) dx \sigma} = (\bar{\sigma}, t, n | x) f$$

Labwork 16 (**Sampling from truncated normal distributions**) [Linton H.; Raderer, Simulations, Statistics and Computing (1995) 5, 121-125] Let $N_+(\cdot, \tau, \sigma^2)$ denote the left-truncated normal distribution with truncation point τ and density given by

```

>>> random('state', 6/9); % initialize the seed at 6/9 and method as Ziggurat -- type help random
>>> r = randn(2,8) % generate the desired mean parameter mu
r =
    3.1416
    1.4142
    1.1416
    1.4286
    0.8442
    1.3955
    1.7107
    3.9572
    3.2618
    6.1652
    2.6971
    2.4940
    4.5928
    0.7617
    3.5397
    5.0282
    1.6139
    5.0977
    2.0477
    2.3286
ans =
    0.8442
    1.3955
    1.7107
    3.9572
    3.2618
    6.1652
    2.6971
    2.4940
    4.5928
    0.7617
    3.5397
    5.0282
    1.6139
    5.0977
    2.0477
    2.3286

```

Suppose we want samples from $X \sim \text{Normal}(\mu = 1, \sigma^2 = 2)$, then we can do the following:

$$\cdot (1,0)_{\text{POSITION}} \approx z \quad \quad \cdot z \varrho + \eta \Rightarrow x$$

If we want to produce samples from $\alpha \sim \text{Normal}(\mu, \sigma^2)$ with some user-specified μ and σ , then we can use the following relationship between X and $Z \sim \text{Normal}(0, 1)$:

Now, let C be the event that the sum of the two dice equals seven. Then

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

while

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

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

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

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

but,

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

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

CONDITIONAL PROBABILITY SUMMARY

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

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

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

Conditional probabilities obey the axioms and rules of probability.

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.

Proof: For the continuous case:

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

And the number of proposals before acceptance is the Geometric($1/a$) RV with expectation $\frac{1}{1/a} = a$.

The closer $ag(x)$ is to $f(x)$, especially in the tails, the closer a will be to 1, and hence the more efficient the rejection method will be.

The rejection method can still be used only if the un-normalised form of f or g (or both) is known. In other words, if we use:

$$f(x) = \frac{\tilde{f}(x)}{\int \tilde{f}(x)dx} \text{ and } g(x) = \frac{\tilde{g}(x)}{\int \tilde{g}(x)dx}$$

we know only $\tilde{f}(x)$ and/or $\tilde{g}(x)$ in closed-form. Suppose the following are satisfied:

- (a) we can generate random variables from g ;
- (b) the support of g contains the support of f , i.e. $\mathbb{Y} \supset \mathbb{X}$;
- (c) a constant $\tilde{a} > 0$ exists, such that:

$$\tilde{f}(x) \leq \tilde{a}\tilde{g}(x), \tag{4.7}$$

for any $x \in \mathbb{X}$, the support of X . Then x can be generated from Algorithm 8.

Algorithm 8 Rejection Sampler (RS) of von Neumann – target shape

1: *input:*

- (1) shape of a target density $\tilde{f}(x) = \left(\int \tilde{f}(x)dx\right) f(x)$,
- (2) a proposal density $g(x)$ satisfying (a), (b) and (c) above.

2: *output:* a sample x from RV X with density f

3: **repeat**

4: Generate $y \sim g$ and $u \sim \text{Uniform}(0, 1)$

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 .¹

Labwork 155 (Gaussian Sampling with `randn`) We can use MATLAB function `randn` that implements the Ziggurat method to draw samples from an RV $Z \sim \text{Normal}(0, 1)$ as follows:

¹ See http://en.wikipedia.org/wiki/Ziggurat_algorithm for more details.

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

$$\text{false alarm rate} = P(\text{detection declared} \mid \text{no intrusion})$$

$$\text{detection rate} = P(\text{detection declared} \mid \text{intrusion})$$

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

- (a) the probability that a patient who tests negative is free from the disease.
- (b) the probability that a patient who tests positive actually has the disease.

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

$$\text{specificity} = P(\text{test is negative} \mid \text{patient does not have the disease})$$

$$\text{sensitivity} = P(\text{test is positive} \mid \text{patient has the disease})$$

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

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

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

Ex. 2.10 — A process producing micro-chips, produces 5% defective, at random. Each micro-chip is tested the test will correctly detect a defective one $\frac{4}{5}$ of the time, and if a good chip is tested, and the test will randomly, and tested to be defective, what was the probability that it was good anyway?

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

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.

2. What is the probability of obtaining a sum greater than 4 but less than 7?

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?

3. What is the probability that a proposed point is accepted by a binary variable to indicate whether a proposed point is accepted or rejected?

4. Use the following code to convert 1000 zeros into samples from the $\text{Normal}(0,1)$ distribution.

5. Use the following code to convert 1000 samples from the $\text{Normal}(0,1)$ with PDF f based on 100 proposals from $\text{Laplace}(1)$ with PDF g .

Proposition 60 (Acceptance Probability of RS) The expected number of iterations of the rejection algorithm to get a sample x is the constant a .

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 .

The next result tells us how many iterations of the algorithm are lower than the former. distribution using a normal distribution, because the latter has lower tails than the former.

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, because its tails are very long.

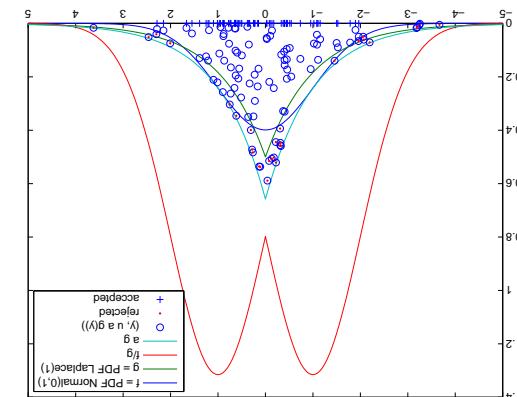


Figure 4.12: Rejection Sampling from $X \sim \text{Normal}(0,1)$ with PDF f based on 100 proposals from $\text{Laplace}(1)$ with PDF g .

```
>>> hist(y, 20) # histogram with 20 bins
>>> y=arrray(zeros(1,1000)) : zeros(1,1000));
>>> % use function to convert 1000 zeros into samples from the Normal(0,1)
>>> % array to accept or reject samples
for u = rand();
    if u <= bound;
        if rrand() <= f(y)/g(y);
            accept = 1;
            x = y;
            y=u*g(y);
            g=g+1;
        else
            reject = 1;
            x = y;
            y=u/g(y);
            g=g+1;
        end
    end
end % while
```

We may obtain a large number of samples and plot them as a histogram using the following command:

```
function x = RejectionNormalLaplace()
    RejectionNormalLaplace()
    Accept = 0; % Reject if random variable to indicate whether a proposed point is accepted
    Reject = 1; % Accept if random variable to indicate whether a proposed point is rejected
    y=LaplacePDF(rand(),1); % Sample Laplace(1) RV
    while ~Accept % Is the LaplacePDF random(1); is accepted
        if u <= bound % If the LaplacePDF random(1); is accepted
            if rrand() <= f(y)/g(y); % If the ratio f(x)/g(x) is less than 1
                Accept = 1;
                x = y;
                y=u*g(y);
                g=g+1;
            else
                Reject = 1;
                x = y;
                y=u/g(y);
                g=g+1;
            end
        end
    end % while
```

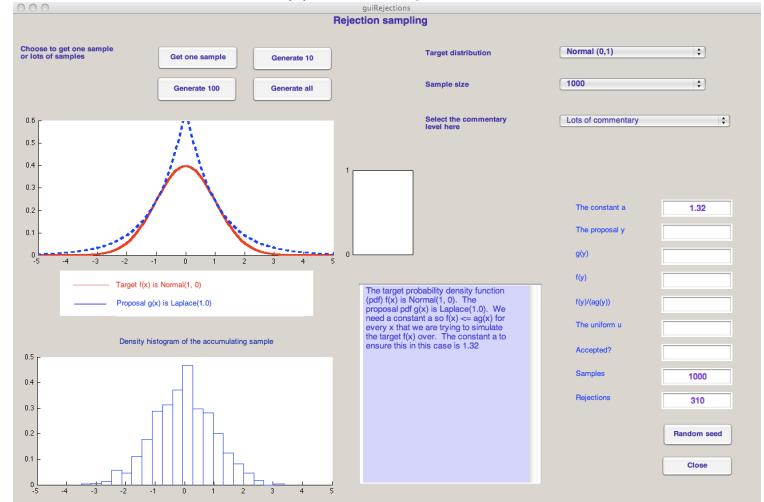
- (a) the probability that there is an intrusion when a detection is declared,
- (b) the probability that there is no intrusion when no detection is declared.

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

```
>> 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`.

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

create a random variable X with this experiment as follows:

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

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

	Experiment	Possible measured outcomes
Counting the number of types up to now		$\mathbb{Z}_+ = \{0, 1, 2, \dots\} \subset \mathbb{R}$
Length in centimetres of some shells on New Brighton beach		$(0, +\infty) \subset \mathbb{R}$
Waiting time in minutes for the next Orbit bus to arrive		$\mathbb{R}_+ = [0, \infty) \subset \mathbb{R}$
Vertical displacement from current position of a pollen on water		\mathbb{R}

including. Crucially, it can become inconvenient to work with a set of outcomes Ω upon which arithmetic is not possible. We are often measuring out outcomes with subsets of real numbers. Some examples 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. We need a different kind of variable to deal with real-world situations where the same variable we can solve for them.

What these **classical variables** have in common is that they take a fixed or deterministic value when

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

Yet another example is the use of variables to represent sequences such as:
over the real line $\mathbb{R} = (-\infty, \infty)$.

where the variable y for the y -axis is determined by the value taken by the variable x , as x varies

$$y = 3x - 2 ,$$

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

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

Random Variables

Chapter 3

Rejection sampling [John von Neumann, 1947, in *Statistical Theory 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 $x \in \mathbb{R}$. Typically, the target density f is only known up to a constant and therefore the (normalized) 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 for which the following are true:
(b) the support of g contains the support of f , i.e. $\mathbb{Y} \subseteq \mathbb{X}$;
(c) a constant $a < 1$ exists, such that:

for any $x \in \mathbb{X}$, the support of X . Then x can be generated from Algorithm 7.

(4.6)
$$f(x) \leq ag(x).$$

the inputs g and a and $u \sim \text{Uniform}(0, 1)$

repeat
3: a proposal density g satisfying (a), (b) and (c) above.
2: output: a sample RV X with density f

3: generate $y \sim g$ and $u \sim \text{Uniform}(0, 1)$

4: if $u \leq \frac{ag(y)}{f(y)}$

5: return: $x \rightarrow y$

6: return: $x \rightarrow y$

Labwork 152 (Rejection Sampler Demo) Let us understand the rejection sampler by calling

$$\begin{aligned} &= \int_{-\infty}^{\infty} f(y) dy \\ &= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-1)^2}{2}} \right) dy \\ &= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-1)^2}{2}} \right) \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-2)^2}{2}} dy \\ &= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-1)^2}{2}} \right) \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-2)^2}{2}} \right) dy \\ &= \frac{\int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-1)^2}{2}} \right) \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-2)^2}{2}} \right) dy}{\int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-1)^2}{2}} \right) dy} \\ &= \frac{\mathbf{P}(Y \leq 2) - \mathbf{P}(Y \leq 1)}{\mathbf{P}(Y \leq 1)} = \frac{\mathbf{P}(X \leq 2) - \mathbf{P}(X \leq 1)}{\mathbf{P}(X \leq 1)} \end{aligned}$$

Proof. We shall prove the result for the continuous case. For any real number t :

player of Algorithm 7 produces a sample x from the random variable X with density $f(x)$. The von Neumann rejection sam-

6: *input:* $x \rightarrow y$
5: *until* $u \leq \frac{ag(y)}{f(y)}$
4: Generate $y \sim g$ and $u \sim \text{Uniform}(0, 1)$
3: repeat
2: output: a sample x from RV X with density f
(2) a proposal density g satisfying (a), (b) and (c) above.
(1) a target density $f(x)$,

Algorithm 7 Rejection Sampler (RS) of von Neumann

for any $x \in \mathbb{X}$, the support of X . Then x can be generated from Algorithm 7.

(4.6)
$$f(x) \leq ag(x).$$

(c) the support of g contains the support of f , i.e. $\mathbb{Y} \subseteq \mathbb{X}$;

(b) we can generate random variables from g for which the following are true:

(a) we can generate random variables from g for which the following are true:

Suppose we have another density or mass function g for which the following are true:

(d) f is itself may be unknown and it is difficult to generate samples directly from X .

(e) Draw 100 samples from the Geometric(Mytetra) RV and report the sample mean. [Note:

the inputs θ_0 and $C(i)$ for the Geometric(θ) RV should be defined and the workings

should].

(b) Set the variable Mytetra=rand.

(c) Draw 100 samples from the Geometric(Mytetra) RV and report the sample mean. [Note:

the inputs θ_0 and $C(i)$ for the Geometric(θ) RV should be defined and the workings

should].

(d) CHAPTE R 4. SIMULATION

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 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,

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

2. $u \sim \text{Uniform}(0, 1)$

2: *output:* a sample from X

3: *initialise:* $p \leftarrow \theta_0$, $q \leftarrow \theta_0$, $i \leftarrow 0$

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

5: $i \leftarrow i + 1$, $p \leftarrow p C(i)$, $q \leftarrow q + p$

6: **end while**

7: *return:* $x = i$

Simulation 150 (Binomial(n, θ)) To simulate from a Binomial(n, θ) 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 θ_0 and $C(x+1)$, we can also simulate from the Geometric(θ) and Poisson(λ) RVs.

Labwork 151 This is a challenging exercise for the student who is finding the other Labworks too easy. So those who are novice to MATLAB may skip this Labwork.

1. Implement Algorithm 6 via a function named `MyGenDiscInvSampler` in MATLAB. Hand in the **M-file** named `MyGenDiscInvSampler.m` giving detailed comments explaining your understanding of each step of the code. [Hint: $C(i)$ should be implemented as a function (use function handles via `@`) that can be passed as a parameter to the function `MyGenDiscInvSampler`].
2. Show that your code works for drawing samples from a Binomial(n, p) RV by doing the following:
 - (a) Seed the fundamental sampler by your Student ID (if your ID is 11424620 then type `rand('twister', 11424620);`)
 - (b) Draw 100 samples from the Binomial($n = 20, p = 0.5$) RV and report the results in an 2×2 table with column headings `x` and No. of observations. [Hint: the inputs θ_0 and $C(i)$ for the Binomial(n, p) RV is given above].
3. Show that your code works for drawing samples from a Geometric(p) RV by doing the following:
 - (a) Seed the fundamental sampler by your Student ID.

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

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

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

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

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

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

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:

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

such that for every $x \in \mathbb{R}$, the inverse image of the half-open interval $(-\infty, x]$ is an element of the collection of events \mathcal{F} , i.e.:

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

Variable (RV), say X , is a function from the sample space Ω to the set of real numbers \mathbb{R} such that for every $x \in \mathbb{R}$, the image of the half-open interval $(-\infty, x]$ is an element of a random variable.

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.

3.1 Basic Definitions

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

“What is $P(X = 1) ?$ ”

or, more simply,

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

in the rain/not experiment of Example 38 becomes

“What is the probability of it raining tomorrow?”

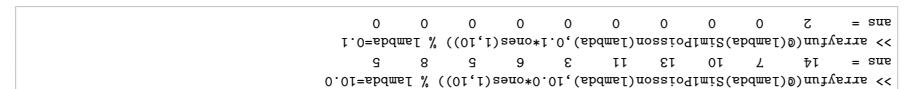
$$f(x; \theta_0, \theta_1, \dots) = \begin{cases} \theta_1, & \text{if } x = 1 \\ \theta_0, & \text{if } x = 0 \\ 0, & \text{if } x \notin \{0, 1, 2, \dots\} \end{cases}$$

is defined as follows:

Model 20 ($G(D(\theta_0, \theta_1, \dots))$) We say X is a General Discrete($\theta_0, \theta_1, \dots$) or $G(D(\theta_0, \theta_1, \dots))$ 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: Simulating from a Poisson(λ) RV is also a special case of simulating from the following more general RV.

Figure 3.18 depicts a comparison of the PDF of Poisson($\lambda = 10$) RV and a relative frequency histogram based on 1000 simulations from it.



Here is a call to simulate 10 samples from Poisson($\lambda = 10.0$) and Poisson($\lambda = 0.1$) RVs:

```
%>>> arrrayfun(@(lambda)SimPoisson(lambda),10,0.*ones(1,10)) % lambda=10.0
ans = 14 7 10 13 11 3 6 5 8 5
%>>> arrrayfun(@(lambda)SimPoisson(lambda),0.1*ones(1,10)) % lambda=0.1
ans = 2 0 0 0 0 0 0 0 0 0
%>>> arrrayfun(@(lambda)SimPoisson(lambda),10,0.*ones(1,10)) % lambda=10.0
ans = 14 7 10 13 11 3 6 5 8 5
%>>> arrrayfun(@(lambda)SimPoisson(lambda),0.1*ones(1,10)) % lambda=0.1
```

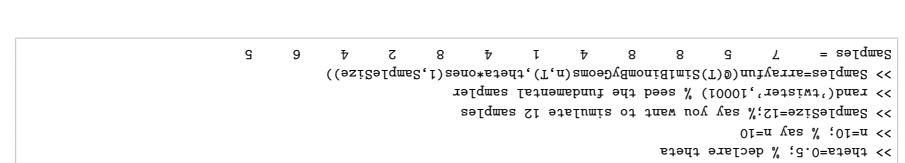
```
function x = SimPoisson(lambda)
    % Simulate one sample from Poisson(lambda) via Exponentials
    YSum=0; k=0; % initialize
    while (YSum < 1), % initialize
        YSum = YSum + -(1/lambda) * Log(rand);
        k=k+1; % return x
    end
    x=k-1; % return x
end
```

We implement the above algorithm via the following M-File:

Simulation 149 (Poisson(λ) from IID Exponential(λ) RVs) By this principle, we can simulate as soon as $\sum_{i=1}^k Y_i \geq 1$ and Step 3: setting $x = k - 1$.
from the Poisson(λ) X by Step 1: generating IID Exponential(λ) RVs Y_1, Y_2, \dots , Step 2: stopping

Let us simulate from the Poisson(λ) RV of Model 6 as shown in Figure 3.18.

Figure 3.8 depicts a comparison of the PDF of Binomial($n = 10, \theta = 0.5$) RV and a relative frequency histogram based on 100,000 simulations from it.



Here is a call to simulate 12 samples from Binomial($n = 10, \theta = 0.5$) RV:

CHAPTER 4. SIMULATION

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

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

```
>> 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 Geometric(θ) RV and the $\text{Binomial}(n, \theta)$ RV.

The $\text{Binomial}(n, \theta)$ RV X is related to the IID Geometric(θ) 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 Geometric(θ) RV.

Simulation 148 ($\text{Binomial}(\theta)$ from IID Geometric(θ) RVs) By this principle, we can simulate from the $\text{Binomial}(\theta)$ X by Step 1: generating IID Geometric(θ) RVs Y_1, Y_2, \dots , Step 2: stopping as soon as $\sum_{i=1}^k (Y_i + 1) > n$ and Step 3: setting $x \leftarrow k - 1$.

We implement the above algorithm via the following M-file:

```
function x = Sim1BinomByGeoms(n,theta)
% Simulate one sample from Binomial(n,theta) via Geometric(theta) RVs
YSum=0; k=0; % initialise
while (YSum <= n),
    TrialsToSuccess=floor(log(rand)/log (1-theta)) + 1; % sample from Geometric(theta)+1
    YSum = YSum + TrialsToSuccess; % total number of trials
    k=k+1; % number of Bernoulli successes
end
x=k-1; % return x
```

X , as follows:

For example, we can compute the desired PDF for an array of samples x from $\text{Binomial}(8, 0.5)$ RV

```
function fx = BinomialPDF(x,n,theta)
    % Binomial probability mass function. Needs BinomialCoefficient(n,x)
    % f is the prob mass function for the Binomial(x;n,theta)
    % and x is an array of samples.
    fx = zeros(size(x));
    for i=1:n+1
        fx(i) = BinomialCoefficient(n,i,theta);
    end
end
```

and call `BinomialCoefficient` in the function `BinomialPDF` to compute the PDF $f(x; n, \theta)$ of the

```
function BC = BinomialCoefficient(n,x)
    % returns the binomial coefficient of n choose x
    % x and n are scalar integers and 0 <= x <= n
    % returns the binomial coefficient of n choose x at a time
    % x and n are scalar integers and 0 <= x <= n
    % numeratorBC = prod(n:-1:max([minusX, x+1]));
    % denominatorPostCancel = prod(2:min([minusX, x]));;
    minusX = n-x;
    NumeratorPostCancel = prod(n:-1:max([minusX, x+1]));
    DenominatorPostCancel = prod(2:min([minusX, x]));
```

with the following M-L file:

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

to compute:

Labwork 146 (Binomial coefficient) The MATLAB function `BinomialCoefficient` can be used

Let us simulate from the $\text{Binomial}(n, \theta)$ RV of Model 5.

```
theta=0.5;
Samplesize=1000;
% simulate 1000 samples from Geometric(theta)
Samplesize=1000;
% Samplesize, % relative frequency histogram
Samplesize=1000;
% relative frequency histogram
Samplesize=1000/Samplesize; % relative frequencies of first 100 samples
Xs = 0:10; % get some values for x
Reffreqs=hist(Samples, Xs); % relative frequencies over Xs
stem(Xs,theta*(1-theta), 'r'); % PDF of Geometric(theta) over Xs
hold on;
Reffreqs=hist(Samples, Xs, 'r'); % relative frequency histogram
plot(Xs,Reffreqs, 'r'); % relative frequency histogram
Reffreqs100=hist(Samples(1:100), Xs)/100; % Relative Frequencies of first 100 samples
plot(Xs,Reffreqs100, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (100 samples)', ...);
% Relative freq. hist. (1000 samples), ...
hold on;
Reffreqs1000=hist(Samples(1:1000), Xs)/1000; % Relative Frequencies of first 1000 samples
plot(Xs,Reffreqs1000, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (1000 samples)', ...);
% Relative freq. hist. (10000 samples), ...
hold on;
Reffreqs10000=hist(Samples(1:10000), Xs)/10000; % Relative Frequencies of first 10000 samples
plot(Xs,Reffreqs10000, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (10000 samples)', ...);
% Relative freq. hist. (100000 samples), ...
hold on;
Reffreqs100000=hist(Samples(1:100000), Xs)/100000; % Relative Frequencies of first 100000 samples
plot(Xs,Reffreqs100000, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (100000 samples)', ...);
% Relative freq. hist. (1000000 samples), ...
hold on;
Reffreqs1000000=hist(Samples(1:1000000), Xs)/1000000; % Relative Frequencies of first 1000000 samples
plot(Xs,Reffreqs1000000, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (1000000 samples)', ...);
% Relative freq. hist. (10000000 samples), ...
hold on;
Reffreqs10000000=hist(Samples(1:10000000), Xs)/10000000; % Relative Frequencies of first 10000000 samples
plot(Xs,Reffreqs10000000, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (10000000 samples)', ...);
% Relative freq. hist. (100000000 samples), ...
hold on;
Reffreqs100000000=hist(Samples(1:100000000), Xs)/100000000; % Relative Frequencies of first 100000000 samples
plot(Xs,Reffreqs100000000, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (100000000 samples)', ...);
% Relative freq. hist. (1000000000 samples), ...
hold on;
```

Figure 3.7:

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

Labwork 145 (PMF versus relative frequency histogram of simulated Geometric(θ) RV)

Some useful properties of the Indicator Function are:

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

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

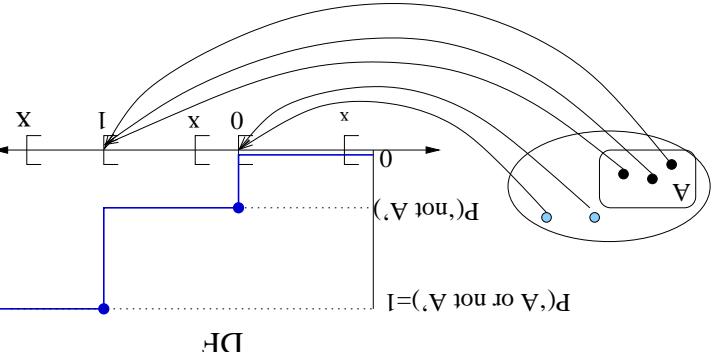


Figure 3.1: The indicator function of event A if F is a RV \mathbb{I}_A with DF F .

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

Some useful properties of the Indicator Function are:

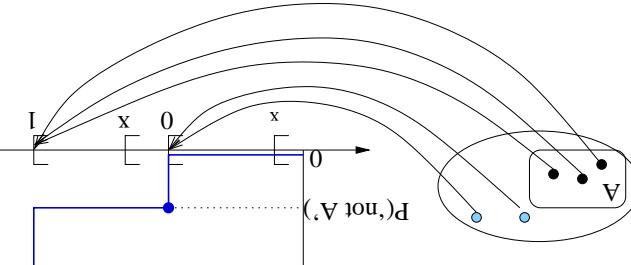


Figure 3.2: A RV X from a sample space Ω with 8 elements to \mathbb{R} and its DF F .

Classwork 41 (A random variable with three values and eight sample points) Consider the RV X of Figure 3.2. First draw this property as done in Ex. 3.1. Let the events $A = \{w_1, w_2\}$, $B = \{w_3, w_4, w_5\}$ and $C = \{w_6, w_7, w_8\}$. Define the RV X formally. What sets should F minimally include? What do you need to do to make sure that F is a sigma algebra?

Exercice 3.1 (Drawing discontinuous functions) Identify the mistakes in how the \mathbb{I}_A is drawn as a discontinuous function in Figure 3.1.

Exercice 3.1 (Drawing discontinuous functions) Identify the mistakes in how the \mathbb{I}_A is drawn as a discontinuous function in Figure 3.1.

Labwork 146 (Binomial coefficient) The MATLAB function `BinomialCoefficient` can be used

Let us simulate from the $\text{Binomial}(n, \theta)$ RV of Model 5.

```
theta=0.5;
Samplesize=1000;
% simulate 1000 samples from Geometric(theta)
Samplesize=1000;
% Samplesize, % relative frequency histogram
Samplesize=1000/Samplesize; % relative frequencies of first 100 samples
Xs = 0:10; % get some values for x
Reffreqs=hist(Samples, Xs); % relative frequencies over Xs
stem(Xs,theta*(1-theta), 'r'); % PDF of Geometric(theta) over Xs
hold on;
Reffreqs=hist(Samples, Xs, 'r'); % relative frequency histogram
plot(Xs,Reffreqs, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (100 samples)', ...);
% Relative freq. hist. (1000 samples), ...
hold on;
Reffreqs100=hist(Samples(1:100), Xs)/100; % Relative Frequencies of first 100 samples
plot(Xs,Reffreqs100, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (100 samples)', ...);
% Relative freq. hist. (1000 samples), ...
hold on;
Reffreqs1000=hist(Samples(1:1000), Xs)/1000; % Relative Frequencies of first 1000 samples
plot(Xs,Reffreqs1000, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (1000 samples)', ...);
% Relative freq. hist. (10000 samples), ...
hold on;
Reffreqs10000=hist(Samples(1:10000), Xs)/10000; % Relative Frequencies of first 10000 samples
plot(Xs,Reffreqs10000, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (10000 samples)', ...);
% Relative freq. hist. (100000 samples), ...
hold on;
Reffreqs100000=hist(Samples(1:100000), Xs)/100000; % Relative Frequencies of first 100000 samples
plot(Xs,Reffreqs100000, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (100000 samples)', ...);
% Relative freq. hist. (1000000 samples), ...
hold on;
Reffreqs1000000=hist(Samples(1:1000000), Xs)/1000000; % Relative Frequencies of first 1000000 samples
plot(Xs,Reffreqs1000000, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (1000000 samples)', ...);
% Relative freq. hist. (10000000 samples), ...
hold on;
Reffreqs10000000=hist(Samples(1:10000000), Xs)/10000000; % Relative Frequencies of first 10000000 samples
plot(Xs,Reffreqs10000000, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (10000000 samples)', ...);
% Relative freq. hist. (100000000 samples), ...
hold on;
Reffreqs100000000=hist(Samples(1:100000000), Xs)/100000000; % Relative Frequencies of first 100000000 samples
plot(Xs,Reffreqs100000000, 'r'); % relative frequency histogram
legend('PDF of Geometric(0.5)', 'Relative freq. hist. (100000000 samples)', ...);
% Relative freq. hist. (1000000000 samples), ...
hold on;
```

Figure 3.7:

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

Labwork 145 (PMF versus relative frequency histogram of simulated Geometric(θ) RV)

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 $(\Omega, \mathcal{F}, \mathbf{P})$. We define the **probability mass function** (PMF) f of X to be the function $f : \mathbb{R} \rightarrow [0, 1]$ defined as follows:

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

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

- For any $x \in \mathbb{R}$,

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

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

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

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

- From the fact that $\mathbf{P}(\Omega) = 1$, we get that the sum of all the probabilities is 1:

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

```
>> 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 θ_i are equal, we generated samples in constant time, which is denoted by $O(1)$.

Simulation 144 (Geometric(θ)) We can simulate a sample x from a Geometric(θ) RV X using the following simple algorithm:

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

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(θ) 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
```

this by producing five samples from a randomly generated PMF f_{θ} .

Clearly, Algorithm 5 may be used to sample any de Moivre($\theta_1, \dots, \theta_k$) RV X . We demonstrate

```

Algorithm 5 Inversion Sampler for de Moivre( $\theta_1, \dots, \theta_k$ ) RV  $X$ 
1: input:  $\theta_1, \dots, \theta_k$ 
2: output: a sample from  $X$ 
3: initialise:  $F \rightarrow 0$ ,  $i \rightarrow 1$ 
4:  $u \sim \text{Uniform}(0, 1)$ 
5:  $i \rightarrow i + 1$ 
6:  $F \rightarrow F + \theta_i$ 
7: end while
8:  $x \rightarrow i$ 
The M-File implementing Algorithm 5 is:
SimdeMoivre.m
```

```

function x = SimdeMoivre(theta)
% Returns a sample from the de Moivre(theta_1, ..., theta_k) RV X
% Call Syntax: x = SimdeMoivre(theta);
% Input: u = a uniform random number e.g. rand
% Output: theta = an array of probabilities theta_1 ... theta_k
% Notes: x = SimdeMoivre(u, theta);
% deMoivre(u, theta);
% SimdeMoivre(theta);
% SimdeMoivre.m
```

```

Let us use the function deMoivreEqn to draw five samples from a fair seven-faced dice.
```

```

Algorithm 6 SimdeMoivreEqn( $\theta$ )
1: input:  $\theta$ 
2: output:  $x = \sum_{i=1}^k \text{rand}(\theta_i)$ 
3: initialise:  $x \rightarrow 0$ 
4:  $u = \text{rand}(1, k)$ 
5:  $i \rightarrow 1$ 
6:  $x = x + \theta_i$ 
7: end for
8:  $x = \lfloor x \rfloor$ 
9: return:  $x$ 
```

```

function x = SimdeMoivreEqn(theta)
% SimdeMoivreEqn( $\theta$ )
% Returns a sample from the de Moivre( $\theta_1, \dots, \theta_k$ ) RV  $X$ 
% Call Syntax: x = SimdeMoivreEqn(theta);
% Input: theta = an array of probabilities theta_1 ... theta_k
% Output: x = sum(theta);
% Notes: x = SimdeMoivreEqn(theta);
% SimdeMoivreEqn.m
```

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

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:

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

Out of the class of discrete random variables we will define specific kinds as they arise often in

what is "under the hood".

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

It is customary to use P , instead of θ , for the probabilities. But we try to avoid it as it will

Probability: $P(X = x_i) = \theta_i$	x_1	x_2	x_3	...
Possible values: x_i				

Probability: $P(X = x_i) = \theta_i$	x_1	x_2	x_3	...
Possible values: x_i				

Hence, we can describe a discrete random variable by the table:
examples it is convenient to associate the possible values x_1, x_2, \dots with the outcomes w_1, w_2, \dots
variables are defined as functions, is much reduced. The reason is that in stratified forward
Notice that in equations (3.5), (3.6) and (3.7) the use of the " $w \in \Omega$ " notation, where random

DISCRETE RANDOM VARIABLES - SIMPLIFIED NOTATION

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

$$\Delta_k := \{(\theta_1, \theta_2, \dots, \theta_k) \in \mathbb{R}_{+}^k : \sum_i \theta_i = 1 \text{ and } \theta_i \geq 0, \text{ for all } i\} \quad (3.10)$$

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:

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
5. Notice that in equations (3.5), (3.6) and (3.7) the use of the " $w \in \Omega$ " notation, where random variables are defined as functions, is much reduced. The reason is that in stratified forward
6. Notice that in equations (3.5), (3.6) and (3.7) the use of the " $w \in \Omega$ " notation, where random variables are defined as functions, is much reduced. The reason is that in stratified forward

RV X when $(\theta_1, \theta_2, \dots, \theta_k)$ are specified as an input vector via the following algorithm.

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 .

RV X which ($\theta_1, \theta_2, \dots, \theta_k$) are specified as an input vector via the following algorithm.

Clearly, Algorithm 5 may be used to sample any de Moivre($\theta_1, \dots, \theta_k$) RV X . We demonstrate

```

Algorithm 5 Inversion Sampler for de Moivre( $\theta_1, \dots, \theta_k$ ) RV  $X$ 
1: input:  $\theta_1, \dots, \theta_k$ 
2: output: a sample from  $X$ 
3: initialise:  $F \rightarrow \theta_1$ ,  $i \rightarrow 1$ 
4:  $u \sim \text{Uniform}(0, 1)$ 
5:  $i \rightarrow i + 1$ 
6:  $F \rightarrow F + \theta_i$ 
7: end while
8:  $x \rightarrow i$ 
The M-File implementing Algorithm 5 is:
SimdeMoivre.m
```

```

Algorithm 6 SimdeMoivreEqn( $\theta$ )
1: input:  $\theta$ 
2: output:  $x = \sum_{i=1}^k \text{rand}(\theta_i)$ 
3: initialise:  $x \rightarrow 0$ 
4:  $u = \text{rand}(1, k)$ 
5:  $i \rightarrow 1$ 
6:  $x = x + \theta_i$ 
7: end for
8:  $x = \lfloor x \rfloor$ 
9: return:  $x$ 
```

```

Algorithm 6 SimdeMoivreEqn( $\theta$ )
1: input:  $\theta$ 
2: output:  $x = \sum_{i=1}^k \text{rand}(\theta_i)$ 
3: initialise:  $x \rightarrow 0$ 
4:  $u = \text{rand}(1, k)$ 
5:  $i \rightarrow 1$ 
6:  $x = x + \theta_i$ 
7: end for
8:  $x = \lfloor x \rfloor$ 
9: return:  $x$ 
```

```

function x = SimdeMoivreEqn(theta)
% SimdeMoivreEqn( $\theta$ )
% Returns a sample from the de Moivre( $\theta_1, \dots, \theta_k$ ) RV  $X$ 
% Call Syntax: x = SimdeMoivreEqn(theta);
% Input: theta = an array of probabilities theta_1 ... theta_k
% Output: x = sum(theta);
% Notes: x = SimdeMoivreEqn(theta);
% SimdeMoivreEqn.m
```

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

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

and its distribution function is given by

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

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

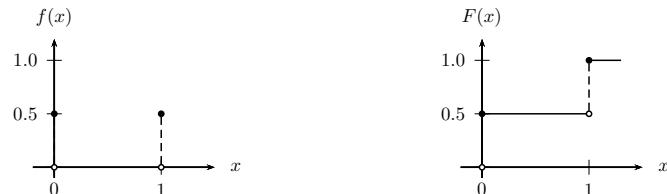
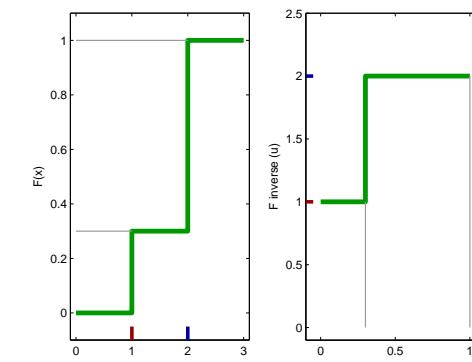


Figure 3.3: $f(x)$ and $F(x)$ of the *fair coin toss* random variable X , a discrete uniform RV on $\{0, 1\}$.

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 Ω 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

Possible values, x_i	1	2	3	4	5	6
Probability, θ_i	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Figure 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($1/k, 1/k, \dots, 1/k$) RV

-
- 1: *input:*
 1. k in de Moivre($1/k, 1/k, \dots, 1/k$) RV X
 2. $u \sim \text{Uniform}(0, 1)$ - 2: *output:* a sample from X
 - 3: *return:* $x \leftarrow \lceil ku \rceil$
-

```
SimdeMoivreEqui.m
function x = SimdeMoivreEqui(u,k);
% return samples from de Moivre(1/k,1/k,...,1/k) RV X
% Call Syntax: x = SimdeMoivreEqui(u,k);
% Input      : u = array of uniform random numbers eg. rand
%               k = number of equi-probabble outcomes of X
% Output     : x = samples from X
x = ceil(k * u); % ceil(y) is the smallest integer larger than y
%x = floor(k * u); if outcomes are in {0,1,...,k-1}
```

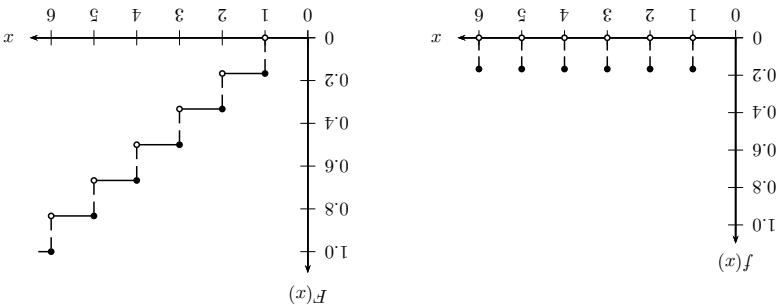
Let us use the function `SimdeMoivreEqui` to draw five samples from a fair seven-faced cylindrical dice.

```
>> k=7; % number of faces of the fair dice
>> n=5; % number of trials
>> rand('twister',78657); % initialise the fundamental sampler
>> u=rand(1,n); % draw n samples from Uniform(0,1)
>> % inverse transform samples from Uniform(0,1) to samples
>> % from de Moivre(1/7,1/7,1/7,1/7,1/7,1/7,1/7)
>> outcomes=SimdeMoivreEqui(u,k); % save the outcomes in an array
>> disp(outcomes);
6 5 5 5 2
```

Now, let us consider the more general problem of implementing a sampler for an arbitrary but specified de Moivre($\theta_1, \theta_2, \dots, \theta_k$) RV. That is, the values of θ_i need not be equal to $1/k$.

to rest on only four sides, the other two sides being rounded. The upper side of the bone, broad games was at first provided by tossing astreaghi, the ankle bones of sheep. These bones could come chance were known in Egypt, 3000 years before Christ. The location of chance needed for those chance were known in Egypt, 3000 years before Christ. The location of chance needed for those **Example 44** (*Astreegah with a Kiwi sheep ankle bone*) *Astreegah*. Board games involving 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 greater than or equal to y , e.g. $\lceil 13.1 \rceil = 14$. Algorithm 4 produces samples from the de Moivre($1/k, 1/k, \dots, 1/k$) RV X .

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



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

and the distribution function is:

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

The probability mass function of this random variable is:

Solution:

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

Simulation 141 (*Point Mass(θ)*) Let us simulate a sample from the Point Mass(θ) RV X . Since this RV produces the same realization θ we can implement it via the following M-file:

The M-File implementing Algorithm 4 is:

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 greater than or equal to y , e.g. $\lceil 13.1 \rceil = 14$. Algorithm 4 produces samples from the de Moivre($1/k, 1/k, \dots, 1/k$) RV, with the ceiling function. First we simulate from an equi-probable special case of the de Moivre($\theta_1, \theta_2, \dots, \theta_k$) RV, with RV X and its inverse $F_{[-1]}$ for a specific $\theta_1 = 0.3$ are depicted in Figure 4.10.

When $k = 2$ in the de Moivre(θ_1, θ_2) model, we have an RV that is similar to the Bernoulli($p = \theta_1$) RV.

First we simulate from an equi-probable special case of the de Moivre($\theta_1, \theta_2, \dots, \theta_k$) RV, with RV X and its inverse $F_{[-1]}$ for a specific $\theta_1 = 0.3$ are depicted in Figure 4.10.

$$F_{[-1]}(u; \theta_1, \theta_2, \dots, \theta_k) = \begin{cases} 1, & \text{if } \theta_1 + \theta_2 + \dots + \theta_{k-1} \leq u < 1 \\ \vdots \\ 3, & \text{if } \theta_1 + \theta_2 \leq u > \theta_1 + \theta_2 + \theta_3 \\ 2, & \text{if } \theta_1 \leq u > \theta_1 + \theta_2 \\ 1, & \text{if } 0 \leq u < \theta_1 \end{cases} \quad (4.5)$$

Given by:

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

Next we simulate from the de Moivre($\theta_1, \theta_2, \dots, \theta_k$) RV X of Model 14 via its inverse DF

```
>>> arrayfun(g(u).SimPointMass,u,17),zeros(2,8)
ans =
    17    17    17    17    17    17    17    17
    17    17    17    17    17    17    17    17
    17    17    17    17    17    17    17    17
    17    17    17    17    17    17    17    17
```

Note that it is not necessary to have input IID samples from Uniform(0,1) RV via rand in order to draw samples from the Point Mass(θ) RV. For instance, an input matrix of zeros can do the job:

```
>>> arrayfun(g(u).SimPointMass,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    17    17    17    17    17    17    17
    17    17    17    17    17    17    17    17    17
```

Here is call to the function.

```
function x = SimPointMass(u,theta)
    % Input : u = one uniform random number e.g. rand()
    % Output : x = SimPointMass(u,theta);
    % Call Syntax: x = SimPointMass(u,theta);
    % Returns one sample from the Point Mass(theta) RV X
    % Simulates to apply SimPointMass to any array of Uniform(0,1) samples
    % we can use arrayfun to apply SimPointMass to any array of Uniform(0,1) samples
    % arrayfun to apply SimPointMass to any array of Uniform(0,1) samples
    % SimPointMass(rand(),2)
    ans = 2
    ans = SimPointMass(rand(),2)
```

This M-File implements the same realization θ we can implement it via the following M-file:

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

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

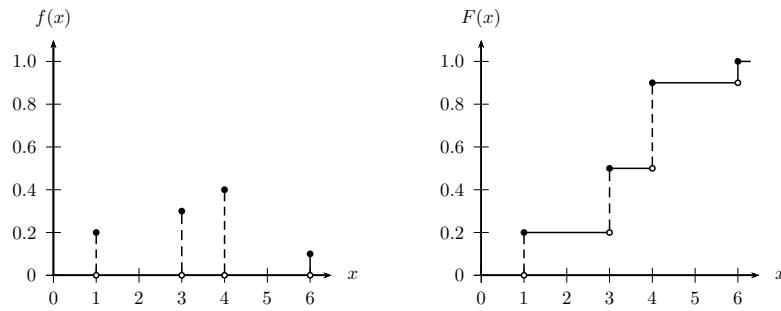


Figure 3.5: $f(x)$ and $F(x)$ of surmised *astragali toss* random variable X , a discrete (non-uniform) RV on $\{1, 2, 3, 4\}$.

3.2.1 An Elementary Family of Bernoulli Random Variables

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

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

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

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

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

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

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 58 (Inversion sampler with compact support) Let the support of the RV X be over some real interval $[a, b]$ and let its inverse DF be defined as follows:

$$F^{[-1]}(u) := \inf\{x \in [a, b] : F(x) \geq u, 0 \leq u \leq 1\}.$$

If $U \sim \text{Uniform}(0, 1)$ then $F^{[-1]}(U)$ has the DF F , i.e. $F^{[-1]}(U) \sim F \sim X$.

Proof: The proof is a consequence of the following equalities:

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

Simulation 140 (Bernoulli(θ)) Consider the problem of simulating from a Bernoulli(θ) RV based on an input from a Uniform(0, 1) RV. Recall that $\lfloor x \rfloor$ (called the ‘floor of x ’) is the largest integer that is smaller than or equal to x , e.g. $\lfloor 3.8 \rfloor = 3$. Using the floor function, we can simulate a Bernoulli(θ) RV X as follows:

```
>> theta = 0.3; % set theta = Prob(X=1)
% return x -- floor(y) is the largest integer less than or equal to y
>> x = floor(rand + theta); % rand is the Fundamental Sampler
>> disp(x) % display the outcome of the simulation
0
>> n=10; % set the number of IID Bernoulli(theta=0.3) trials you want to simulate
>> x = floor(rand(1,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(θ) trials.

```
>> % a demonstration of Central Limit Theorem --
>> % the sample mean of a sequence of n IID Bernoulli(theta) RVs is Gaussian(theta,theta*(1-theta)/n)
>> theta=0.5; Reps=10000; n=10; hist(sum(floor(rand(n,Reps)+theta))/n)
>> theta=0.5; Reps=10000; n=100; hist(sum(floor(rand(n,Reps)+theta))/n,20)
>> theta=0.5; Reps=10000; n=1000; hist(sum(floor(rand(n,Reps)+theta))/n,30)
```

Recall the Point Mass(θ) RV. Formally, we can simulate from it trivially as follows.

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

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

Since the Bernoulli(θ) RV has only two outcomes, i.e., simple events, we know how to obtain the probability of each of the two outcomes in a given Bernoulli trial so we have sequence of Bernoulli(θ_i) trials, say,

probability of flooding is the same each year.

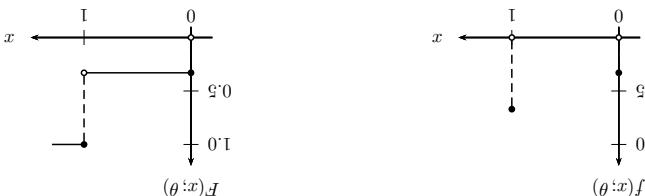
blooded. Note: we assume that flooding is independent from year to year, and that the years; count the number of years, during the 20-year period, during which the property is flooded. Note: we assume that flooding is the same each year.

- Provide a property near a particular bridge in our archipelago with flood insurance for 20 years.
- Roll a die 100 times; count the number of sixes you throw.
- Test 50 randomly selected circuits from an assembly line; count the number of defective circuits.
- by possibility allowing for the coins $P(H)$ to change each time because each of them are manufactured in a terrible mint.
- Flip a coin 10 times; count the number of heads

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

3.2.2 Independent Bernoulli Trials

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



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

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

and its DF is:

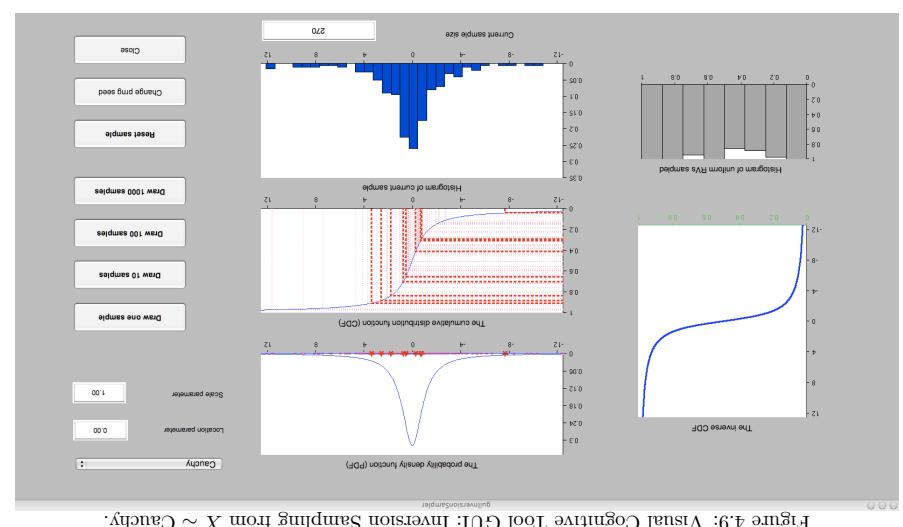


Figure 4.9: Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Cauchy}$.

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 compare 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 and see the location parameter 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 (3.51) with an additional location parameter θ and scale parameter a .

Labwork 138 (Inversion Sampler Demo – Cauchy) Let us comprehend the inversion sampler by calling the interactive visual cognitive tool:

```
>> guiInversionSampler
```

```
0.3530 1.6127 -0.3621 0.3637 0.9953
>> display(x);
% draw 5 samples from Laplace(1) RV using inverse CDF
>> x=laplaceInvCDF(u,lambdab); % draw 5 samples from Laplace(1) RV using inverse CDF
0.6487 0.9003 0.3481 0.6524 0.8152
>> disp(u);
% draw 5 IID samples from Sample
0.6487 0.9003 0.3481 0.6524 0.8152
>> u=rmad(1,5);
% initialize the fundamental sampler
>> lambdab=1.0;
% some value for lambda
>> rand('twister',6567);
% draw 5 IID samples from Uniform(0,1) RV
>> x=rmad(1,5);
% draw 5 samples from Laplace(1) RV using inverse CDF
>> display(x);
% draw 5 samples from Laplace(1) RV using inverse CDF
```

We can simply call the function to draw a sample from, say the `Laplace(A = 1.0)` RV by

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

$$\mathbf{P}(x; (\theta_1, \theta_2, \dots)) = \prod_i f(x_i; \theta_i) = \prod_i \theta_i^{x_i} (1 - \theta_i)^{1-x_i} \mathbf{1}_{\{0,1\}}(x_i), \quad (3.15)$$

where $x := (x_1, x_2, \dots) \in \mathbb{X}_\infty = \{0, 1\}^\infty := \{0, 1\} \times \{0, 1\} \times \dots$

By further assuming that all the θ_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(θ) trials** as follows:

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

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

3.2.3 Some Common Discrete Random Variables

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

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

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

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

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

the probability mass function of X is:

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

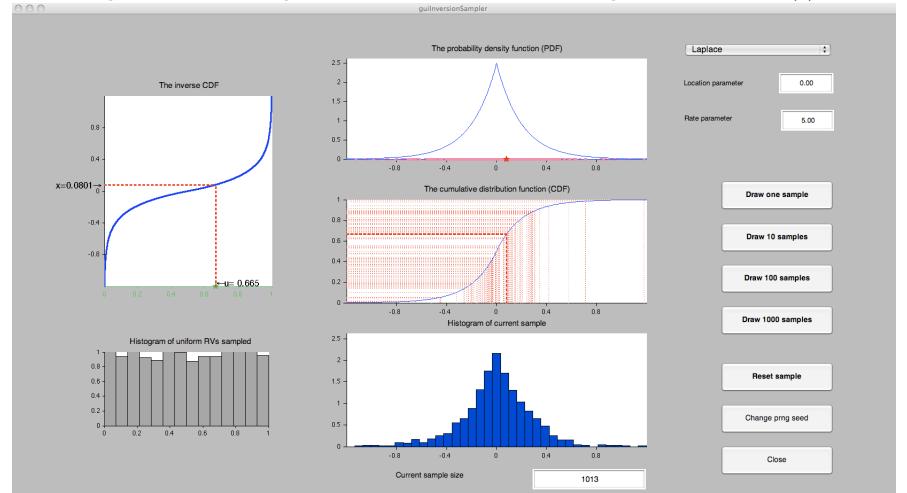
In the previous Example, noting that we have independent trials, we get:

Labwork 136 (Rejection Sampler Demo – Laplace(5)) Let us comprehend the rejection sampler by calling the interactive visual cognitive tool:

```
>> guiInversionSampler
```

The M-file `guiInversionSampler.m` will bring a graphical user interface (GUI) as shown in Figure 4.8. Using the drop-down menu change from the default target distribution Uniform(-5, 5) to Laplace(5). Now repeatedly push the “Draw one sample” button several times and comprehend the simulation process. You can also press “Draw 1000 samples” and see the density histogram of the generated samples. Next try changing the numbers in the “Rate parameter” box from 5.00 to 1.00 in order to alter the parameter λ of Laplace(λ) RV. If you are more adventurous then try to alter the number in the “Location parameter” box from 0.00 to some thing else, say 10.00. Although our formulation of Laplace(λ) implicitly had a location parameter of 0.00, we can easily introduce a location parameter μ into the PDF. With a pencil and paper try to rewrite the PDF in (4.2) with an additional location parameter μ .

Figure 4.8: Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Laplace}(5)$.



Simulation 137 (Laplace(λ)) Here is an implementation of an inversion sampler to draw IID samples from a Laplace(λ) RV X by transforming IID samples from the Uniform(0, 1) RV U :

```
LaplaceInvCDF.m
function x = LaplaceInvCDF(u,lambda);
% Call Syntax: x = LaplaceInvCDF(u,lambda);
% or LaplaceInvCDF(u,lambda);
%
% Input      : lambda = rate parameter > 0,
%               u = an 1 X n array of IID samples from Uniform[0,1] RV
% Output     : an 1Xn array x of IID samples from Laplace(lambda) RV
% or Inverse CDF of Laplace(lambda) RV
x=-(1/lambda)*sign(u-0.5).* log(1-2*abs(u-0.5));
```

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

The outcome of a Geometric(θ) RV can be thought of as "the number of tosses needed before the appearance of the first 'Head' when tossing a coin with probability of 'Heads' equal to θ in a independent and identical manner".

$$\begin{aligned} & \lim_{n \rightarrow \infty} \sum_{x=0}^{n-\infty} a\theta^x = a \left(\frac{1-\theta}{1-\theta^n} \right), \text{ provided, } 0 < \theta < 1 \\ & \left(\frac{\theta}{1-\theta^n} \right) = a \sum_{x=0}^{n-\infty} a\theta^x = a(1-\theta^{n+1}) \\ & a(1-\theta^{n+1}) = a - a\theta^{n+1} = \left(\sum_{x=0}^{n-\infty} a\theta^x \right) - \left(\sum_{x=0}^{n-\infty} a\theta^{x+1} \right) \\ & a + a\theta + a\theta^2 + \dots + a\theta^n = \sum_{x=0}^{n-\infty} a\theta^x - a\theta^{n+1} \end{aligned}$$

Therefore,
Proof:

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

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

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

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

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

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

This is called a **geometric random variable** with "success probability" parameter θ . We can spot a geometric distribution because there will be a sequence of independent trials until 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.

$$P(X = x) = P(F \cdot \cdot \cdot F^{\underline{x-1}}) = (1-\theta)^{x-1}\theta.$$

$P(F) = 1 - \theta$ so that:

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

$$f(x) = P(X = x) = P(\overbrace{TT \cdot \cdot \cdot TH}^x) = P(T \cdot \cdot \cdot T^{\underline{x-1}} H) = \left(\frac{2}{1} \right)^{x-1} \frac{1}{2}.$$

0.4204	0.1219	2.2182	3.6692	0.9654	0.0093	1.1226	5.2216	2.1856	6.6378	7.6490	2.0168	1.3073	7.9990
>> <code>s=sim(4); disp(s);</code> >> %sum up the outcomes of the sequence of 5 draws in each replicate													

Note that the mean is 0 due to the symmetry about 0 and the variance is twice that of the Exponential(λ) RV.

and the variance is

$$\mathbb{E}(X) = \int_{-\infty}^0 xf(x)\lambda e^{-\lambda|x|} dx = \int_{\infty}^0 x\frac{2}{\lambda}e^{-\lambda|x|} dx = 0,$$

Mean and Variance of Laplace(λ) RV X : Show that the mean of a Laplace(λ) RV X is

$$F_{[-1]}(u; \lambda) = -\frac{\lambda}{1} \operatorname{sign}(u - \frac{2}{\lambda}) \log \left(1 - \frac{2}{\lambda} \right), \quad u \in [0, 1], \quad (4.4)$$

and its inverse DF is

$$F(x; \lambda) = \int_x^{\infty} f(y; \lambda) dy = \frac{2}{\lambda} \left(1 - e^{-\lambda|x|} \right), \quad (4.3)$$

Then, the DF of the Laplace(λ) RV X is

$$\operatorname{sign}(x) = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Let us define the sign of a real number x by

$$f(x; \lambda) = \operatorname{sign}(x) = \begin{cases} \frac{2}{\lambda} e^{-\lambda|x|} & \text{if } x < 0 \\ \frac{2}{\lambda} e^{\lambda|x|} & \text{if } x > 0 \end{cases} \quad (4.2)$$

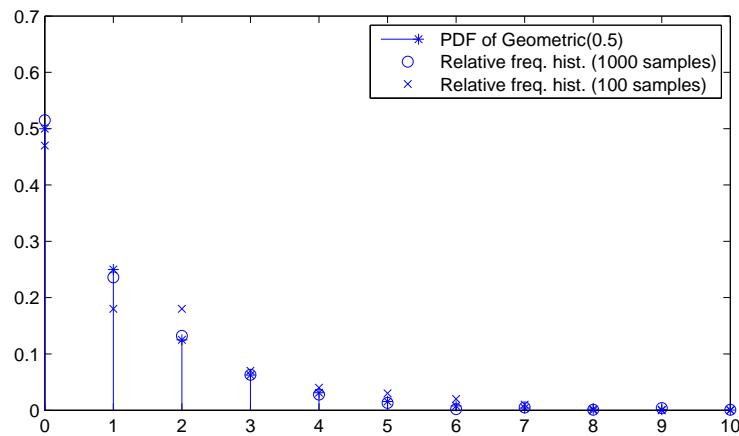
Laplace(λ) RV given by $f(x; \lambda)$ is equal to the entire real line as its support. Model 19 (Laplace(λ) or Double Exponential(λ) RV) If a RV X is equal likely to be either positive or negative with an exponential density, then the Laplace(λ) or Double Exponential(λ) RV, with the rate parameter $\lambda > 0$, $\lambda \in \mathbb{R}$, may be used to model it. The density function for the positive or negative with an exponential density, then the Laplace(λ) or Double Exponential(λ) RV, with the rate parameter $\lambda > 0$, $\lambda \in \mathbb{R}$, may be used to model it. The density function for the

Exponential(λ) RV about the origin with the entire real line as its support.

The support of the Exponential(λ) RV is \mathbb{R}^+ = $[0, \infty)$. Let us consider a RV built by mirroring

times of the next seven buses at your ID-needed bus stop. Speed the fundamental sampler by your Student ID (e.g., if your ID is 11424620 then stop-watch. Simulate the times of arrival for the next seven buses as indicated by your zero to a stop-watch. Suppose you go to your bus stop and arrival of buses at a bus stop. Suppose that the time between arrivals is an Exponential($\lambda = 0.1$) arrival at your bus stop. Let the time between arrivals be T . Consider the problem of modelling the Labwork 135 (Next seven buses at your bus-stop) Consider the problem of modelling the

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

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

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, θ) RV) Let the RV $X = \sum_{i=1}^n X_i$ be the sum of n independent and identically distributed Bernoulli(θ) RVs, i.e.:

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

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

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

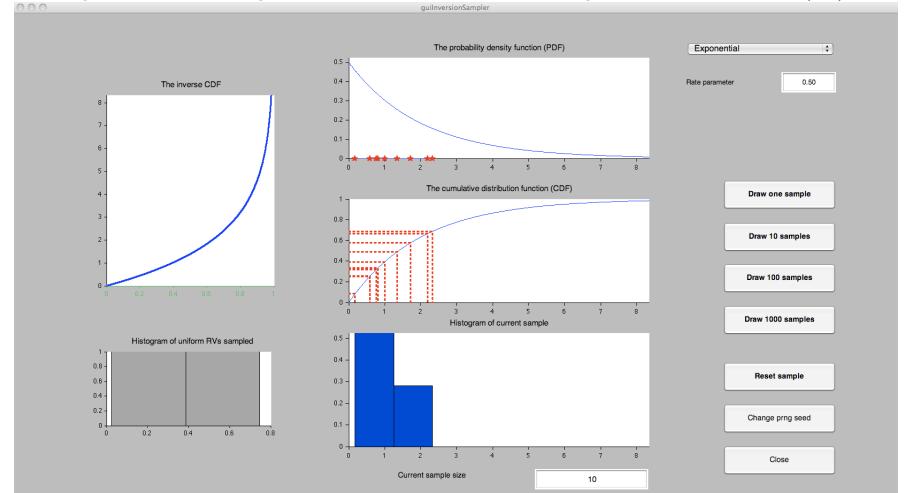
```
>> rand('twister',46678); % initialise the fundamental sampler
>> Lambda=1.0; % declare Lambda=1.0
>> x=ExpInvSam(rand(1,5),Lambda); % pass an array of 5 Uniform(0,1) samples from rand
>> disp(x); % display the Exponential(1.0) distributed samples
0.5945 2.5956 0.9441 1.9015 1.3973
```

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.

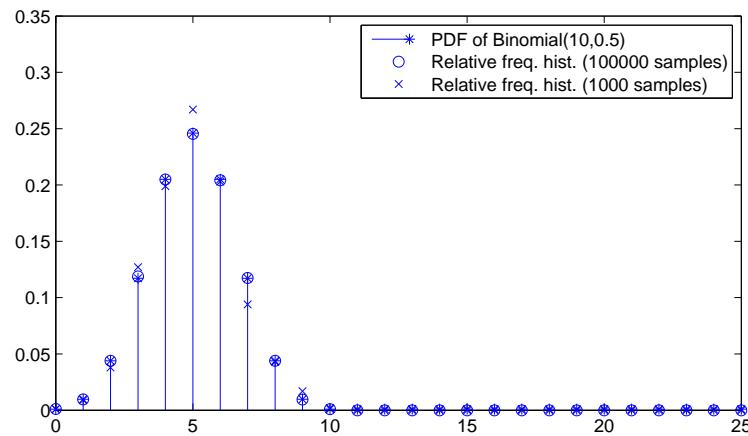
Figure 4.7: Visual Cognitive Tool GUI: Inversion Sampling from $X \sim \text{Exponential}(0.5)$.



It is straightforward to do replicate experiments. Consider the experiment of drawing five independent samples from the Exponential($\lambda = 1.0$) RV. Suppose we want to repeat or replicate this experiment seven times and find the sum of the five outcomes in each of these replicates. Then we may do the following:

```
>> rand('twister',1973); % initialise the fundamental sampler
>> % store 7 replications of 5 IID draws from Exponential(1.0) RV in array a
>> lambda=1.0; a= -1/lambda * log(rand(5,7)); disp(a);
0.7267 0.3226 1.2649 0.4786 0.3774 0.0394 1.8210
1.2698 0.4401 1.6745 1.4571 0.1786 0.4738 3.3690
```


Figure 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

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:

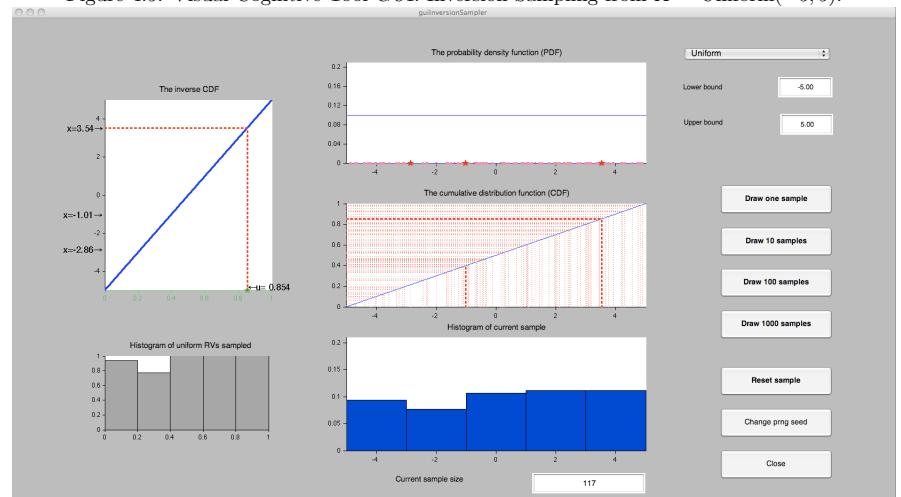
```
>> rand('twister',786543); % initialise the fundamental sampler
>> theta1=-83; theta2=1004; % declare values for parameters a and b
>> u=rand(1,5); % now u is an array of 5 samples from Uniform(0,1)
>> x=theta1+(theta2 - theta1)*u; % x is an array of 5 samples from Uniform(-83,1004) RV
>> disp(x); % 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 – 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 θ_1 and θ_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.

Mechanica Illustration of the Cause of Frequency.—The Curve of Frequency, &c. of Distribution, are convertible : therefore if the genesis of either of them can be made clear, that of Distribution, will be known.

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

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

FIGURE 3: Figures from Sir Francis Galton, F.R.S., *Natural Inheritance*, Macmillan, 1889.

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 definitely handled by the higher methods, and are warily interpreted, their power of detailing with compactness and exactness is extraordinary. They are the path of those which an opening can be cut through the formidable thickness that bars the path of man.

Others, and do not recoil 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 to Avergne, is but a solitary fact, whereas if a single other fact be added to it, an entire Normal distribution is difficult to understand why statistics commonly limit their inquiries.

Charms of Statistics.—It is difficult to understand why statistics commonly limit their inquiries in the sequel as a way of giving precise meanings to Galton's observations with his Quanticum. The

```

>>> n=500;
>>> theta1=1; theta2=1; % initialize the fundamental sampler for Uniform(-1,1) RV X
>>> u=rand(theta1+theta2-1); % read is the fundamental Sampler and u is a sample from it
>>> x=u*(theta2-theta1); % solve for parameters theta1 and theta2
>>> f_x=theta2-u; % derivative of f(x) wrt u
>>> f_x
ans = 0.5000

```

Here is a simple implementation of the Inversion Sampler for the Uniform(θ_1, θ_2) RV in MATLAB:

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

Simulator 131 ($F_{[-1]}(\theta_1, \theta_2)$) To simulate from Uniform(θ_1, θ_2) RV X using the Inversion Sampler, we first need to find $F_{[-1]}(u)$ by solving for x in terms of $u = F(x; \theta_1, \theta_2)$:

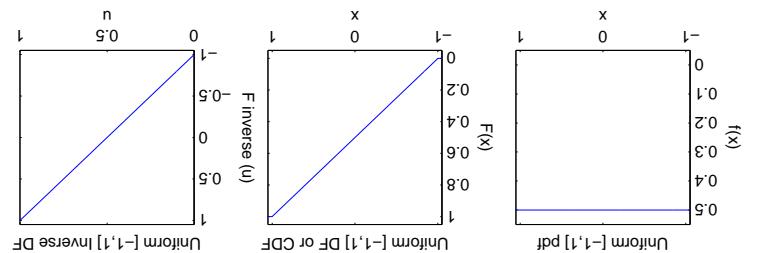


FIGURE 4: A plot of the PDF, DF or CDF and inverse DF of the Uniform(-1,1) RV X .

This algorithm emphasises the fundamental sampler's availability in an *input* step, and its set-up will consider some examples.

Recall the Uniform(θ_1, θ_2) RV of Model 9 with the following PDF, DF and inverse DF. Let us simulate from it using the inversion sampler.

for which the inverse distribution function is explicitly known. The next section will be taken for granted. The direct applicability of Algorithm 3 is limited to univariate densities needs in an *initialise* step. In the following sections, we will not mention these universal steps: they will be considered some examples.

Algorithm 3 Inversion Sampler or inverse (CDF) Sampler

```

1: input: (1)  $F_{[-1]}(x)$ , inverse of the DF of the target RV  $X$ , (2) the fundamental sampler
2: initialise: set the seed, if any, for the fundamental sampler
3: output: a sample from  $X$  distributed according to  $F$ 
4: draw  $u \sim \text{Uniform}(0, 1)$ 
5: return:  $x = F_{[-1]}(u)$ 

```

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

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

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

become more nearly identical with the Normal Curve of Frequency, if the rows of pins were much more numerous, the shot smaller, and the compartments narrower; also if a larger quantity of shot was used.

The principle on which the action of the apparatus depends is, that a number of small and independent accidents befall each shot in its career. In rare cases, a long run of luck continues to favour the course of a particular shot towards either outside place, but in the large majority of instances the number of accidents that cause Deviation to the right, balance in a greater or less degree those that cause Deviation to the left. Therefore most of the shot finds its way into the compartments that are situated near to a perpendicular line drawn from the outlet of the funnel, and the Frequency with which shots stray to different distances to the right or left of that line diminishes in a much faster ratio than those distances increase. This illustrates and explains the reason why mediocrity is so common."

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

We might, for example, ask:

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

A Poisson experiment has the following characteristics:

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

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

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

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

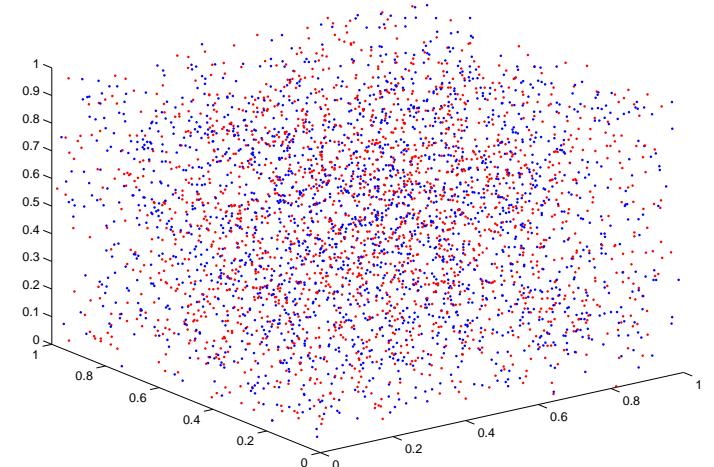
Note that the PDF integrates to 1:

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

where we exploit the Taylor series of e^{λ} to obtain the second-last equality above.

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

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 57 (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 .

the sum of a sequence of n IID Bernoulli(θ) RVs with $\lambda = n\theta$ converges to the Poisson(λ) RV as $n \rightarrow \infty$ and $\theta \rightarrow 0$ in a specific sense.

In the sequel we will see more formally, after understanding notions of convergence of RVs, that

$\lambda = n\theta$. The smaller the value of θ and larger the value of n , the better the approximation.

$$\begin{aligned}
 &= 0.323 \quad (\exists \text{ sig. } \bar{y}_k) \\
 &= 1 - 0.1353 - 0.2707 - 0.2707 \\
 &= 1 - f(0;2) - f(1;2) - f(2;2) \\
 &= 1 - \mathbf{P}(X=0) - \mathbf{P}(X=1) - \mathbf{P}(X=2) \\
 &= 1 - \mathbf{P}(X > X_0) = 1 - \mathbf{P}(\exists \bar{y}_k > X_0)
 \end{aligned}$$

(q)

$$D(X) = 0.135 \quad (3 \text{ sig. figs.})$$

(a)

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

Let the random variable X denote the number of cars arriving in a 15 minute continuum. 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

Solution:

(b) At least three cars arrive?

(a) No cars arrive?

average, 8 cars arrive per hour on Saturday. What is the probability that during a randomly selected one-hour period on Saturday, exactly 6 cars will arrive?

($\cdot \vec{g}_{II} \cdot \vec{g}_{IS}$ c) = 168.0 =

$$= e^{-z} \left(\frac{0!}{2^0} + \frac{1!}{2^1} + \frac{2!}{2^2} + \frac{3!}{2^3} \right) =$$

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

cars enter on average.

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

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

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

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?

CHAPTER 3. RANDOM VARIABLES

```

    >>> plot3(y(1),y(2),:),y(3,:); % plot triplets as red dots
    >>> hold on;
    >>> plot3(x(1),x(2),:);x(3,:); % plot triplets as blue dots
    >>> xrand(5000,2000);% score PMS seeded by 5000 matrix named y
    >>> xrand(5000,2000);% score PMS seeded by 5000 matrix named z
    >>> rand('twister',5489);% the recommended default seed
    >>> xrand(5000,2000);% score PMS in a 5000 matrix named x
    >>> xrand('twister',112347);% same seed as before

```

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?

```

>>> x=rand(3,2000); % store PRNs in a 3x2000 matrix named x
>>> x=rotm(3,2000); % correlate between triplets generated by the "Mersenne Twister" by rotating the 3D plot generated by the following code:
>>> plot(x(1,:);x(2,:);x(3,:));
>>> xlabel('twister',1234)

```

In general, you can use any seed value to initialize your PRNG. You may use the `clock` command to set the seed:

```

>>> rand(1,10) % generate a 1 X 10 array of PRNs
ans =
0.8171 0.9058 0.1270 0.9134 0.6524 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(1,10) % generate a 1 X 10 array of PRNs
ans =
0.8171 0.9058 0.1270 0.9134 0.6524 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

>>> rand(1,10) % reset the PRNs to default state Mersenne Twister with seed=5489
ans =
0.8171 0.9058 0.1270 0.9134 0.6524 0.0975 0.2785 0.5469 0.9575 0.9649

>>> rand(1,10) % reproduce the first array
ans =
0.1576 0.9706 0.9572 0.4854 0.8003 0.1419 0.4218 0.9157 0.7922 0.9595

```

Example 51 (Still-born Babies) About 0.01% of babies are stillborn in a certain hospital. We find the probability that of the next 5000 babies born, there will be no more than 1 stillborn baby. Let the random variable X denote the number of stillborn babies. Then X has a binomial distribution with parameters $n = 5000$ and $\theta = 0.0001$. Since θ is so small and n is large, this binomial distribution may be approximated by a Poisson distribution with parameter

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

Hence

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

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

THINKING POISSON

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

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

DISCRETE RANDOM VARIABLE SUMMARY

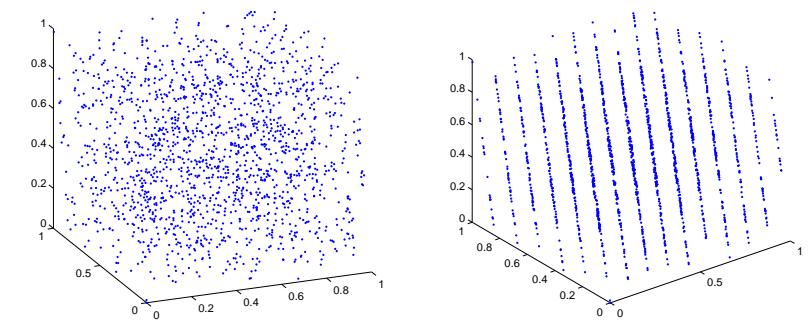
Probability mass function

$$f(x) = \mathbf{P}(X = x_i)$$

Distribution function

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

Figure 4.2: The LCG called `RANDU` with $(m, a, c) = (2147483648, 65539, 0)$ has strong correlation between three consecutive points as: $x_{i+2} = 6x_{k+1} - 9x_k$. The two plots are showing (x_i, x_{i+1}, x_{i+2}) from two different view points. .



The number of random numbers n should at most be about $m/1000$ in order to avoid the future sequence from behaving like the past. Thus, if $m = 2^{32}$ then a new generator, with a new suitable set of (m, a, c, x_0, n) should be adopted after the consumption of every few million pseudo-random numbers.

The LCGs are the least sophisticated type of PRNGs. They are easier to understand but are not recommended for intensive simulation purposes. The next section briefly introduces a more sophisticated PRNG we will be using in this course. Moreover our implementation of LCGs using the variable precision integer package is extremely slow in MATLAB and is only of pedagogical interest.

4.2.2 Generalized Feedback Shift Register and the “Mersenne Twister” PRNG

The following generator termed `twister` in MATLAB is recommended for use in simulation. It has extremely long periods, low correlation and passes most statistical tests (the DIEHARD statistical tests). The `twister` random number generator of Makoto Matsumoto and Takuji Nishimura is a variant of the twisted generalized feedback shift-register algorithm, and is known as the “Mersenne Twister” generator [Makoto Matsumoto and Takuji Nishimura, *Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator*, ACM Transactions on Modeling and Computer Simulation, Vol. 8, No. 1 (Jan. 1998), Pages 3–30]. It has a Mersenne prime period of $2^{19937} - 1$ (about 10^{6000}) and is **equi-distributed** in 623 dimensions. It uses 624 words of state per generator and is comparable in speed to the other generators. The recommended default seed is 5489. See <http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html> and http://en.wikipedia.org/wiki/Mersenne_twister for details.

Let us learn to implement the MATLAB function that generates PRNs. In MATLAB the function `rand` produces a deterministic PRN sequence. First, read `help rand`. We can generate PRNs as follows.

Labwork 129 (Calling PRNG in MATLAB) In MATLAB `rand` is basic PRNG command.

Since f maps a the finite set $\{1, 2, \dots, m-1\}$ into itself, such systems are bound to have a repeating cycle of numbers called the **period**. In Labwork 124, the generator $\text{Lincogenen}(13, 12, 10, 12)$ has period $(10, 1)$ of length 2, the generator $\text{Lincogenen}(13, 10, 9, 8, 12)$ has period $(8, 11, 2, 3, 0, 9)$ of length 6 and the generator $\text{Lincogenen}(256, 137, 0, 123, 257)$ has a period of length 32. All these generators have a non-maximal period 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. a 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 of 256) Consider the linear congruential generator with parameters $(m, a, c, x_0, n) = (256, 137, 123, 256)$. First check that these parameters do indeed satisfy the three conditions above and therefore can produce the maximal period length of 256 as opposed to the generator of length 257. Do you see that the repeat Labwork 125 in order to only $m = 256$. Modify the input parameter to Lincogenen and repeat Labwork 125 Next produce a triplets of maximum length of 256 as opposed to the sequence of first 500! / 3 triplets (x^i, x^{i+1}, x^{i+2}) seeded from $x_0 = 1$ commands help visualize the sequence of first 500! / 3 triplets (x^i, x^{i+1}, x^{i+2}) seeded from $x_0 = 1$ designed. Notably, it fails the **central test** badly for dimensions greater than 2. The following code illustrates the quality of pseudo-random numbers of a LCG is extremely sensitive to the choice of modulus m , and 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 in history. It fails the **central test** badly for dimensions greater than 3. The following code illustrates the quality of pseudo-random numbers of a LCG is extremely sensitive to the choice of 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 parameters a . Let X be the number of years before a particular type of machine will need replacement. Assume that X has the probability function $f(1) = 0.1, f(2) = 0.2, f(3) = 0.2, f(4) = 0.2$, $f(5) = 0.3$.

Labwork 128 (Fishman20 and Leycyer21 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.

```
ans = 0.0006 0.3934 0.4117 0.7893 0.3913 0.3942 0.6790 0.3337 0.2192 0.1883
>>> lincogenen(2147483647,48271,0,0878458,10) . / 2147483647
ans = 0.0041 0.5293 0.0755 0.6242 0.0769 0.9030 0.4259 0.9948 0.8868
>>> lincogenen(2147483647,48271,0,01234567,10) . / 2147483399
ans = 0.0006 0.3934 0.4117 0.7893 0.3913 0.3942 0.6790 0.3337 0.2192 0.1883
```

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

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

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 $f(1) = 0.1, f(2) = 0.2, f(3) = 0.2, f(4) = 0.2$, $f(5) = 0.3$.

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

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?

$P(X=x)$	1	2	3	4	5
	0.1	0.1	0.1	0.1	0.1

3.3 EXERCISES IN DISCRETE RANDOM VARIABLES

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(θ)	$\{0, 1\}$	$P(X=0) = 1 - \theta$ $P(X=1) = \theta$	Situations with only 2 outcomes, coded 1 for success and 0 for failure. Parameter: $\theta = P(\text{success}) \in (0, 1)$.
Geometric(θ)	$\{1, 2, 3, \dots\}$	$P(X=x) = (1 - \theta)^{x-1}\theta$	Situations where you count the number of trials until the first success in a sequence of independent trials with a constant probability of success.
Binomial(n, θ)	$\{0, 1, 2, \dots, n\}$	$P(X=x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$	Situations where you count the number of successes in n trials where each trial is independent and there is a constant probability of success. Parameter: $n \in \{1, 2, \dots\}; \theta = P(\text{success}) \in (0, 1)$.
Poisson(λ)	$\{0, 1, 2, \dots\}$	$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}$	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: $\lambda = \text{rate } \in (0, \infty)$.

(a) Construct a row of cumulative probabilities for this table, that is, find the distribution function of X .

(b) Find the following probabilities.

$$\begin{array}{ll} \text{(i)} \mathbf{P}(X \leq 5) & \text{(iii)} \mathbf{P}(X > 9) \\ \text{(ii)} \mathbf{P}(X < 12) & \text{(iv)} \mathbf{P}(X \geq 9) \end{array}$$

$$\begin{array}{ll} \text{(v)} \mathbf{P}(4 < X \leq 9) & \text{(vi)} \mathbf{P}(4 < X < 11) \end{array}$$

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

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

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

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

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

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

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?

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

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

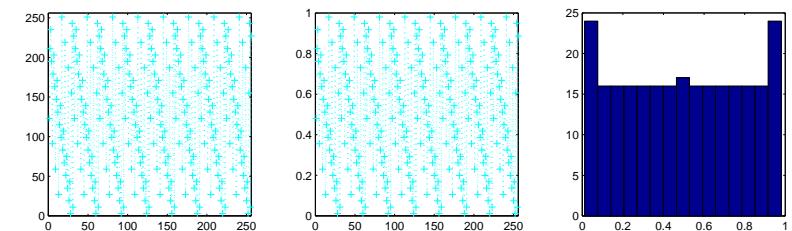
```
>> LinConGen(13,12,11,10,12)
ans = 10 1 10 1 10 1 10 1 10 1 10 1
>> LinConGen(13,10,9,8,12)
ans = 8 11 2 3 0 9 8 11 2 3 0 9
```

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.

```
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*:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Solution

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 forty 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.

$$\left. \begin{array}{ll} \frac{\pi}{2} \leq x & 0 \\ \frac{\pi}{2} > x > 0 & \cos x \\ 0 > x & 0 \end{array} \right\} = (x) f = (x) f$$

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

Solution

$$(b) \text{Find } P(X < \frac{\pi}{2})$$

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

$$\left. \begin{array}{ll} \frac{\pi}{2} \leq x & 1 \\ \frac{\pi}{2} > x > 0 & \sin(x) \\ 0 \geq x & 0 \end{array} \right\} .$$

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

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

(q)

$$\left. \begin{array}{ll} 1 \leq x & 1 \\ 1 > x > 0 & e^x \\ 0 \geq x & 0 \end{array} \right\} = (x) F$$

Hence

$$\begin{aligned} 1 &= \\ 0 + 0[e^a] &+ 0 = \\ ap_0 \int_x^1 + ap_e \int_1^0 + ap_0 \int_0^\infty &= (x) F \end{aligned}$$

If $x \geq 1$, then

$$\begin{aligned} e^x &= \\ 0[e^a] + 0 &= \\ ap_e \int_x^0 + ap_0 \int_0^\infty &= (x) F \end{aligned}$$

If $0 < x < 1$, then

$$F(x) = \int_x^\infty 0 da = 0.$$

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

1, 3, 2, 1, 2, 3

Ex. 3.55 — What is the sample mean and sample variance of the following dataset:

3.15 Exercises in Statistics

(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.

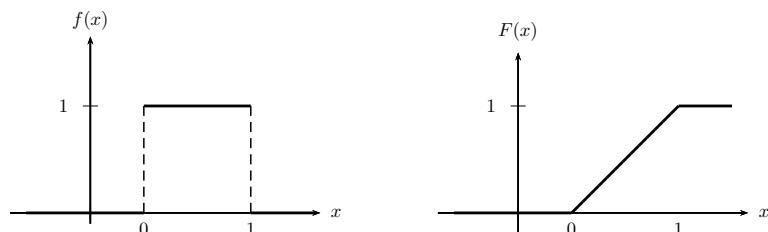


Figure 3.10: $f(x)$ and $F(x)$ of the Uniform(0, 1) random variable X .

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 another 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

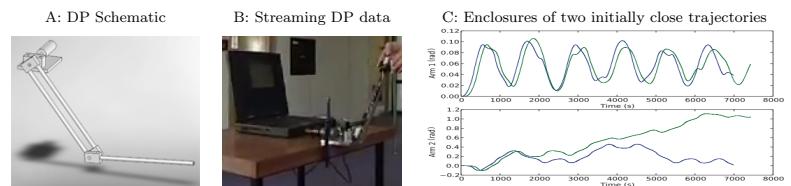


Figure 3.37: Double Pendulum

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.

$$\text{Therefore, } \quad 0 \leq x \leq \infty \quad x^{\alpha} = \frac{1}{\Gamma(\alpha)} \int_x^{\infty} (t-x)^{\alpha-1} e^{-t} dt$$

Solution:

- (a) Find the distribution function.
 (b) Find the probabilities, $P(\frac{1}{4} \leq X \leq 2)$ and $P(-\frac{5}{4} \leq X \leq \frac{5}{4})$.
 (c) Find x such that $P(X \leq x) = 0.95$.

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

B.4.2 Some Common Continuous Random Variables

Simulation in Chapter 4, we can obtain any other random variable from the fundamental mode. Figure 4.1 is its inverse, i.e., $F^{-1}(x) = F_{\text{[---]}}(x)$, as you will see from Neumann's Fundamentals. Theorem 4.1

*Universality of the fundamental model

— The information model has inherently many copies of its entities within it; you can see this in $\text{DF } F$ 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.

If you want a sum suggested in optional Exercise 2.1 on introducing a most primitive algebra (signature algebra) for any $x \in (0, 1)$,

The fundamental model is equivalent to infinite tosses of a fair coin (see using binary expansion tossing a fair coin infinitely often, i.e., the sequence of Bernoulli($1/2$) trials, and the fundamental model).

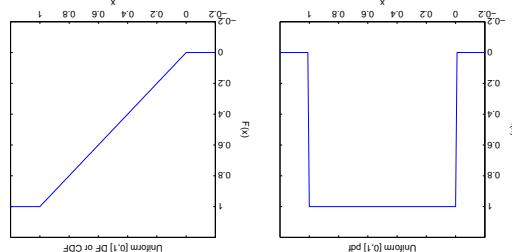


Figure 3.11: A convolutional but mathematically imprecise Matlab plot from a polyline interpolation for the PDF and CDF of the Uniform(0,1) continuous RV X .

68

Worldle is a toy for generating word clouds from plain text that you provide. The clouds give greater prominence to words that appear more frequently in the source text than others. You can hit them out, or save them to the Worldle gallery to share with your friends.

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 closest possible resolutions of the text content in the JOE 2010 Report. The "word cloud" shown in Figure 3-36 was produced by Phillip Wilson using wordle from www.wordle.net/. A description from the wordle URL says:

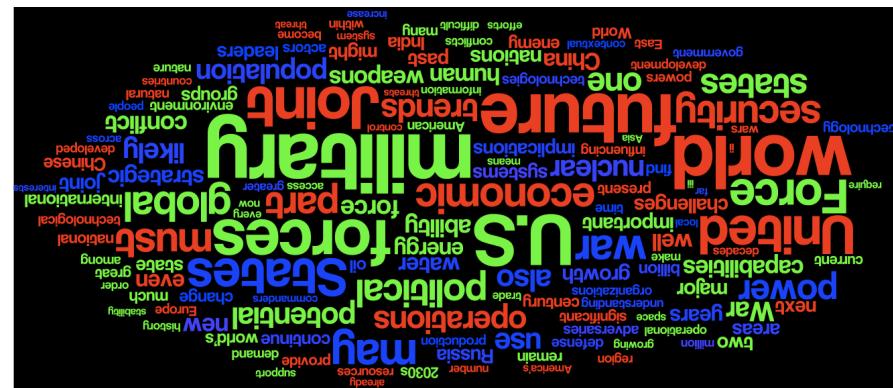


Figure 3.36: Wordle of JOE 2010

APPENDIX I THIS STUDY The joint Operations Environment is intended to monitor joint concepts, agreements, and experiences throughout the Department of Defense. It provides a perspective on future trends, shocks, conflicts, and implications for future joint force commanders and other leaders and professionals in the national security field. This document is specific to the nature and does not propose 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 Miltischer Avenue, Suite 200, Norfolk, VA 23551-2488, (757) 366-9555.

An interesting document with a central thesis project is the joint operating plan outlined 2010 Report by the US Department of Defense. This document was downloaded from http://www.jcs.mil/newslink/storystackive/2010/JOE_2010.pdf. The first paragraph of this 74 page document (JOE 2010 Report) reads:

- twitter messages within an online social network of interest

Processing and analyzing 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

3.14.9 Textual Data

104

(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,

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

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

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

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

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

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

The last two equations are derived from definitions as follows:

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

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

This distribution is unique because of its property of **memorylessness**, i.e., $\mathbf{P}(X > x+y | X > y) = e^{-\lambda x}$, and plays a fundamental role in modeling continuous time processes, such as time between occurrence of events of interest, as we will see in the sequel.**Example 56 (On a dark desert highway)** At a certain location on a dark desert highway, the time in minutes between arrival of cars that exceed the speed limit is an Exponential($\lambda = 1/60$) random variable. If you just saw a car that exceeded the speed limit then what is the probability of waiting less than 5 minutes before seeing another car that will exceed the speed limit?*Solution:*The waiting time in minutes is simply given by the Exponential($\lambda = 1/60$) random variable. Thus, the desired probability is

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

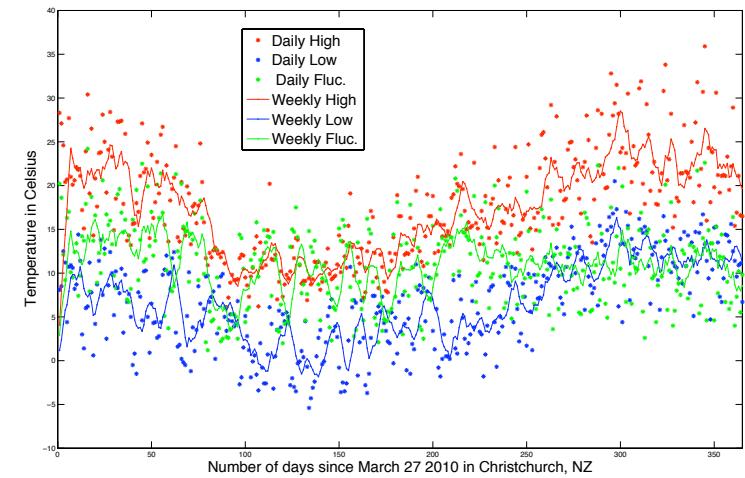
```

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



```

% if the matrix has a different size now than the data-less days now!
size(t) % report the dimensions or size of the matrix T
T = [1,2,4]; % just pull the time
% print size before removing missing data rows are removed
% raw text from NIMA (Copyright NIMA 2011 Subject to NIMA's Terms and
% conditions. See: http://ccl10.nima.co.nz/PDFs/NIPD/Geoms.html)
% raw values - see the file for various manipulations that were done to the
% 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 from NIMA (Copyright NIMA 2011 Subject to NIMA's Terms and
% conditions. See: http://ccl10.nima.co.nz/PDFs/NIPD/Geoms.html)
% any(jmean(t,:)) = [] % Remove any rows containing NaNs from a matrix
T = (jmean(t,:)); % report the dimensions or size of the matrix T
T = (1,2,4); % just pull the time
% keep column 1,2 and 4 named NZST,Tmax(C),Period(Hrs),Min(C),
% i.e. date in YYYYMMDD format, maximum temperature, minimum temperature
T = dimread('NIMAC11P0ChacteristicTemps.txt','[27,1,20904,5]');
% and we read only columns NZST,Tmax(C),Period(Hrs),Min(C),Period(Hrs)
% here we start from line number 27 and end at the last line number 20904
% over the next file 'NIMAC11P0ChacteristicTemps.txt'.
% specify the upper-left and lower-right corners of an imaginary rectangle
% the matrix T is about to be assigned to a matrix; the option [27,1,20904,5] to
% the data as a matrix
% NZST,Max(C),Period(Hrs),Min(C),Period(Hrs),RMean(%),Period(Hrs)
% NZST is daily temperature in Christchurch New Zealand
% Max-Min: Daily Temperature column IDs
% Load the data from the comma delimited text file 'NIMAC11P0ChacteristicTemps.txt'
% with the following command
% ChctempLoad.m

```

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

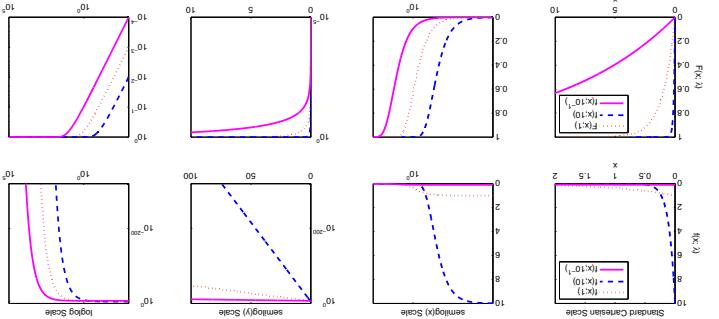
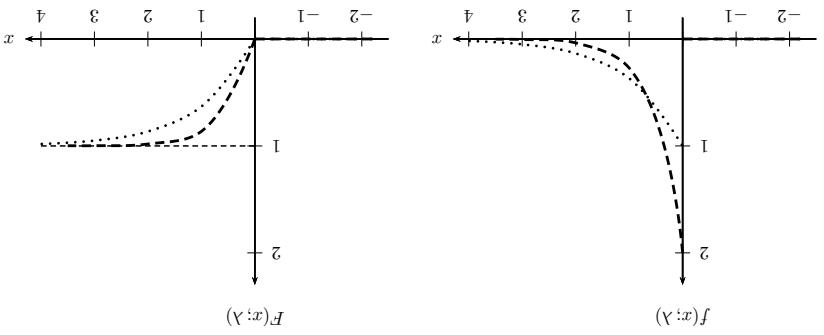


Figure 3.12: $f(x; \lambda)$ and $F(x; \lambda)$ of an exponential random variable where $\lambda = 1$ (dotted) and $\lambda = 2$ (dashed).



script file ChctempLoad.m by typing:
Labwork 121 Understand how Figure 3.35 is being generated by following the comments in the
Daily Temperatures in Christchurch

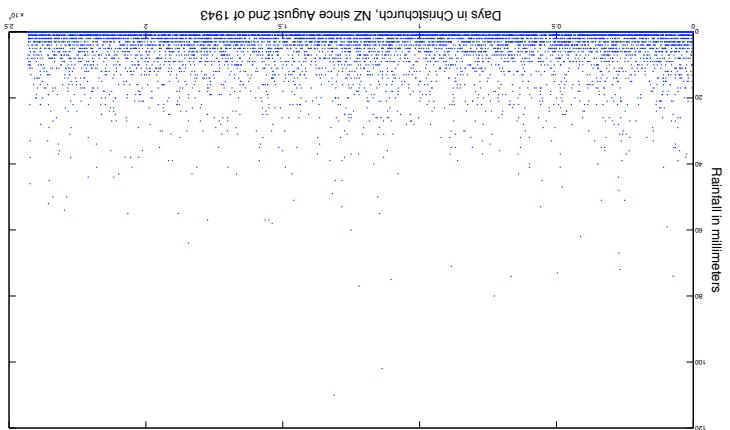


Figure 3.34: Daily rainfall in Christchurch since March 27 2010

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:

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

Exercise 3.17 (Memorylessness of Exponential(λ) RV) [This exercise is optional.]

Try to prove that if $X \sim \text{Exponential}(\lambda)$, then it has the property of **memorylessness**, i.e., prove that $\mathbf{P}(X > x + y | X > y) = e^{-\lambda x}$.

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

Model 9 (Uniform(θ_1, θ_2)) Given two real parameters $\theta_1, \theta_2 \in \mathbb{R}$, such that $\theta_1 < \theta_2$, the PDF of the *Uniform*(θ_1, θ_2) RV X is:

$$f(x; \theta_1, \theta_2) = \begin{cases} \frac{1}{\theta_2 - \theta_1} & \text{if } \theta_1 \leq x \leq \theta_2, \\ 0 & \text{otherwise} \end{cases} \quad (3.30)$$

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

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

Recall that we emphasise the dependence of the probabilities on the two parameters θ_1 and θ_2 by specifying them following the semicolon in the argument for f and F .

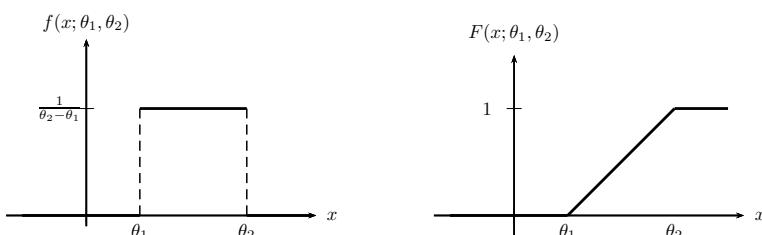
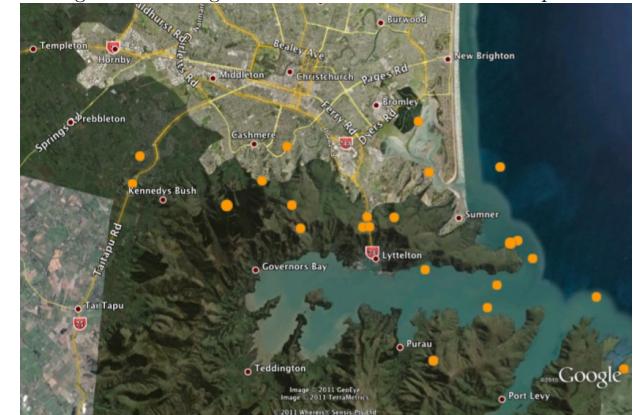


Figure 3.14: $f(x)$ and $F(x)$ of the *Uniform*(θ_1, θ_2) random variable X .

Exercise 3.18 Consider a random variable with a probability density function

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

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
FirstDayOfData = 19430802
LastDayOfData = 20100721
```

```
%% 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{:} RainFallsChch{:}]; % 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)
```

DISTRIBUTIONS FOR ACCESS TO DATA (See <http://cliffo.niwa.co.nz/doc/terms-print.html>.)

New Zealand's meteorological service NIWA provides weather data under its TERMS AND CONDITIONS FOR ACCESS TO DATA (See <http://cliffo.niwa.co.nz/doc/terms-print.html>.)

3.14.8 Meteorological Data

<http://www.googleearth.com/earth/index.html> and the freely available GOOGLE EARTH package Google Earth from <http://www.mathworks.com/products/earth/index.html> wide, the freely available EARTH software bundle from <http://www.google.com/earth/index.html>.

produced 43 earth quakes worldwide, including those in Christchurch as shown in Figure 3.33. One

do a lot more than a mere visualisation with the USGS/NEIC database of earth-quakes world-

wide.

Catalogue: USGS/NEIC (PDE-Q)

Date Range: 2011 2 22 to 2011 2 22

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

Geostatistical exploratory data analysis with Google Earth

```
%>>> % clear any existing figure windows
%>>> % read MagData, LatData, LonData, DepData;
%>>> % make another figure
%>>> % plot(LonData,LatData,MagData,'r.-');
```

```
%>>> % clear any existing figure windows
%>>> % read MagData, LatData;
%>>> % see the depth in Log Scale
%>>> % figure % tell Matlab you are about to make another figure
%>>> % scatter(LonData,LatData,MagData,'r.-');
%>>> % figure % tell Matlab you are about to make another figure
%>>> % scatter(LonData,LatData,MagData,'r.-');
%>>> % figure % tell Matlab you are about to make another figure
%>>> % scatter(LonData,LatData,MagData,'r.-');
```

```
%>>> % tell Matlab you are about to make another figure
%>>> % plot(LonData,LatData,MagData,'r.-');
%>>> % figure % tell Matlab you are about to make another figure
%>>> % scatter(LonData,LatData,MagData,'r.-');
```

```
%>>> % tell Matlab you are about to make another figure
%>>> % plot(LonData,LatData,MagData,'r.-');
%>>> % figure % tell Matlab you are about to make another figure
%>>> % scatter(LonData,LatData,MagData,'r.-');
```

```
%>>> % tell Matlab you are about to make another figure
%>>> % plot(LonData,LatData,MagData,'r.-');
%>>> % figure % tell Matlab you are about to make another figure
%>>> % scatter(LonData,LatData,MagData,'r.-');
```

```
%>>> % tell Matlab you are about to make another figure
%>>> % plot(LonData,LatData,MagData,'r.-');
%>>> % figure % tell Matlab you are about to make another figure
%>>> % scatter(LonData,LatData,MagData,'r.-');
```

```
%>>> % tell Matlab you are about to make another figure
%>>> % plot(LonData,LatData,MagData,'r.-');
%>>> % figure % tell Matlab you are about to make another figure
%>>> % scatter(LonData,LatData,MagData,'r.-');
```

```
%>>> % tell Matlab you are about to make another figure
%>>> % plot(LonData,LatData,MagData,'r.-');
%>>> % figure % tell Matlab you are about to make another figure
%>>> % scatter(LonData,LatData,MagData,'r.-');
```

```
%>>> % tell Matlab you are about to make another figure
%>>> % plot(LonData,LatData,MagData,'r.-');
%>>> % figure % tell Matlab you are about to make another figure
%>>> % scatter(LonData,LatData,MagData,'r.-');
```

CHAPTER 3. RANDOM VARIABLES

hard normal distribution will take on a value:

Example 59 Find the probabilities, using normal tables, that a random variable having the stan-

just do it!

from 0 at $-\infty$ to 1 at ∞ , and intersects the vertical axis at $1/2$. Draw this by hand too.

Classwork 58 Note that the curve of $\phi(z)$ is S-shaped, increasing in a strictly monotone way

Remark 26 This integral 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. 5.5.

The distribution function of Z is given by

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_z^{\infty} e^{-a^2/2} da. \quad (3.33)$$

An exercise in calculus yields the first two derivatives of ϕ as follows:

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

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

Classwork 57 From the above exercise in calculus let us draw the graph of ϕ by hand now! Do it step by step: z^2 , $-z^2$, $-\frac{z^2}{2}$, $\exp(-\frac{z^2}{2})$, $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})$ now!

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

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right). \quad (3.32)$$

This standard normal distribution is the most important continuous probability distribution.

The standard normal variables in a relatively simple fashion. Furthermore, the normal distribution is a useful approximation of more complicated distributions.

normal random variables have a normal distribution, or they are approximately normal, or can be transformed into variables have a normal distribution in year, month, day, hour, minute, sec:

first described by De Moivre in 1733 and subsequently by C. F. Gauss (1777 - 1855). Many random

variables have a normal distribution, or they are approximately normal. Furthermore, the normal distribution is a useful approximation of more complicated distributions.

(b) Sketch the graphs of $f(x)$ and $F(x)$.

(a) Find the value of k .

- (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
- $F(x) = \int_{-\infty}^x f(v)dv$
- $P(a < X \leq b) = F(b) - F(a) = \int_a^b f(v)dv$

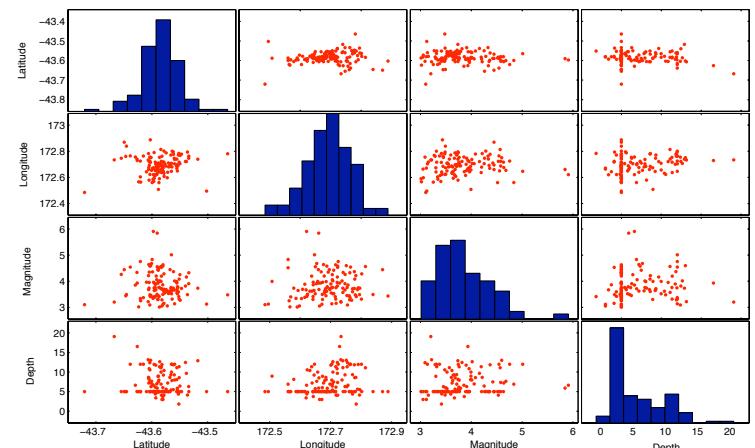
```
>> LatData=EQ(:,2); LonData=EQ(:,3); MagData=EQ(:,12); DepData=EQ(:,13);
>> % finally make a plot matrix of these 124 4-tuples as red points
>> plotmatrix([LatData,LonData,MagData,DepData], 'r.');
```

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`, `Magnitude` and `Depth`, 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
```

```

ans = 124 14
ans = 124
    % now report the size of the earthquakes in Christchurch in variable Eq
    % report the size of Eq and see if it is different from before we removed and NaN containing rows
    % remove locations outside Christchurch and assign it to a new variable called Eq
    % Eq=(Eq11-(43.76*Eq11(:,2))&Eq11(:,3)<12.9&Eq11(:,12)>3,:);
    % remove locations outside of Eq11 and see if it is different from before we removed and NaN containing rows
    % report the size of Eq11 and see if it is different from before we removed and NaN containing rows
    % remove locations outside of Eq11 and assign it to a new variable called Eq
    ans = 145 14
    % move locations outside of Eq11 and see if it is different from before we removed and NaN containing rows
    % Eq11(Any(isnan(eq11),2),:)=[],%move any rows containing NaNs from the matrix Eq11
    % Eq11(any(isnan(eq11),2),:);

```

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.,

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

Example 61 In a cell-phone system a mobile signal may have a signal-to-noise-ratio of X , but the random variable Y which gives the daily profit is of more interest, where profit per item is \$5 and the overhead costs are \$500 per day. The original random variable is X ,

$$Y = 5X - 500.$$

Example 60 Consider a simple financial example where an individual sells X items per day, the 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?

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?

3.6 Transformations of random variables

1. Find k .

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

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

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

Note: X is measured in years.

(a) Assume that $X = 0.01$, and find the probability that the bulb will not burn out before τ hours. This context, λ is often called the failure rate of the bulb.

In this context, λ is often called the failure rate of the bulb.

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

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

(b) For what value of τ is the reliability of the bulb exactly $\frac{1}{2}$?

Hint: Use the distribution function for an Exponential(λ) random variable (recall, $F(\tau; \lambda) = \int_{-\infty}^{\tau} f(t)dt$).

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

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

Note: X is measured in years.

2. Find the distribution function, F .

3. Graph f and F .

Ex. 3.19 — Consider the probability density function

1. Find the value of k .

Ex. 3.19 — Consider the probability density function

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

3. The next set of commands show you how to load, manipulate and visually explore this data.

The thirteen columns correspond to forty 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.

XX Using MATLAB's dimmed command we can assignthe data as a matrix to Eq;

XX the following command XX note that the option 1,0 to dimmed skips first row of column descriptors

XX note that the variable Eq11 is about to be assigned the data as a matrix

XX report the dimensions or size of the matrix Eq11

ans =

4. In order to understand the syntax in detail get help from MATLAB!

```

445 14
ans =
    % Load the data from the comma delimited text file 'NZ2011022earthquakes.csv', with
    % the following column IDs
    % the following command we can assignthe data as a matrix to Eq;
    % note that the option 1,0 to dimmed skips first row of column descriptors
    % note that the variable Eq11 is about to be assigned the data as a matrix
    % report the dimensions or size of the matrix Eq11
    Eq11 = dimread('NZ2011022earthquakes.csv', ',', 1, 0);
    % The thirteen columns correspond to forty 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.
    % 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.
    % the following command we can assignthe data as a matrix to Eq;
    % note that the option 1,0 to dimmed skips first row of column descriptors
    % note that the variable Eq11 is about to be assigned the data as a matrix
    % report the dimensions or size of the matrix Eq11
    ans =

```

CUSP_ID	LAT	LONG	NZMGN	ORT_YEAR	ORT_MONTH	ORT_DAY	ORT_HOUR	ORT_MINUTE	ORT_SECOND	MAG	DEPTH
3418760,-43.58807,172.67126,2433470,5736809,2011,2,22,23,50,1,06171,3,117,3,	348114,-43.5949,-172.70396,2486103,5738805,2011,2,22,23,51,11014,3,117,3,	3469122,-43.5949,172.70621,2486287,5738106,2011,2,22,0,43,70276,3,76,5,4045,	3481760,-43.56579,172.70621,2486287,5738106,2011,2,22,0,43,70276,3,76,5,4045,	3418761,-43.56532,172.69899,2414930,5733975,2011,2,22,0,43,72714,3,79,5,6869,	3419114,-43.56949,172.70396,2486103,5738805,2011,2,22,23,50,1,06171,3,117,3,						

3.5 Exercises in Continuous Random Variables

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; $g^{-1}(y) = \frac{y}{2}$

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:

- if $g(x) = 2x + 1$ then $g : \mathbb{R} \rightarrow \mathbb{R}$ is injective and subjective and therefore its inverse function is:

$$\text{Step 1: } y = 2x + 1, \text{ Step 2: } x = \frac{y - 1}{2}, \text{ Step 3: } g^{-1}(y) = \frac{y - 1}{2}$$

- if $g(x) = x^3$ then $g : \mathbb{R} \rightarrow \mathbb{R}$ is injective and subjective and therefore its inverse function is:

$$\text{Step 1: } y = x^3, \text{ Step 2: } x = \sqrt[3]{y}, \text{ Step 3: } g^{-1}(y) = \sqrt[3]{y}$$

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 $(-\infty, 0]$ 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 $g(x) = x^2$ and domain of g is $[0, \frac{\pi}{2}]$ then its inverse function $g^{-1}(y) = \arcsin(y)$, i.e.,

$g(x) = \sin(x) : [-\frac{\pi}{2}, \frac{\pi}{2}] \rightarrow [-1, 1]$ then the inverse image $g^{-1}(y)$ for $y \in [-1, 1]$ is given by the inverse function $g^{-1}(y) = \arcsin(y) : [0, \frac{\pi}{2}] \rightarrow [-\frac{\pi}{2}, \frac{\pi}{2}]$.

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 $g^{-1}[A]$ satisfies the following properties:

- if $A_1 \cup A_2 \cup \dots = A$ then $g^{-1}(A_1 \cup A_2 \cup \dots) = g^{-1}(A)$
- if A_1, A_2, \dots are disjoint sets then $g^{-1}(A_1), g^{-1}(A_2), \dots$ are disjoint sets
- if $A \subseteq \mathbb{R}$ then $g^{-1}(A) \subseteq \mathbb{R}$
- if A_1, A_2, \dots is a collection of sets A_1, A_2, \dots , then $g^{-1}(A_1, A_2, \dots) = g^{-1}(A_1) \cup g^{-1}(A_2) \cup \dots$

Consequently, 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 $g^{-1}(A)$ and therefore satisfies the axioms of probability and gives the desired probability of the event A from the transformation $Y = g(X)$ we will subscript the probability density of mass function and the distribution function of Y by $F_Y(y)$.

Because we have more than one random variable to consider, namely, X and its transformation $Y = g(X)$ we will subscript the probability density of mass function and the distribution function of Y by the random variable itself. For example we denote the distribution function of X by $F_X(x)$ and $Y = g(X)$ just short-hand for its actual meaning:

$\mathbb{P}(Y \in U) = \mathbb{P}(g(X) \in U) = \mathbb{P}(X \in g^{-1}(U))$ of the underlying experiment in the sense that Equation (3.34) is just short-hand for its sample space underpinned by the random variable X . It is crucial to understand this from the inverse image of A for the probability of the event $g^{-1}(A)$ given by the inverse image $g^{-1}(A)$ satisfies the following properties:

- 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, 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 $g^{-1}(A)$ and therefore satisfies the axioms of probability and gives the desired probability of the event A from the transformation $Y = g(X)$.

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 $\{(i, j) \in \{1, 2, \dots, d\} : i \neq j\}$ and at histograms for every pair of co-ordinates $i \in \{1, 2, \dots, d\}$ 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**.

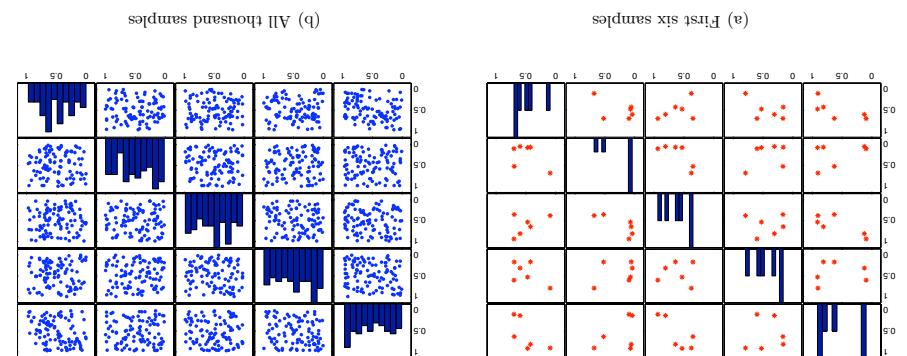


Figure 3.31: Plot Matrix of uniformly generated data in $[0, 1]^5$

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 $\{(i, j) \in \{1, 2, \dots, d\} : i \neq j\}$ and at histograms for every pair of co-ordinates $i \in \{1, 2, \dots, d\}$ 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**.

There are several other techniques for visualizing 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

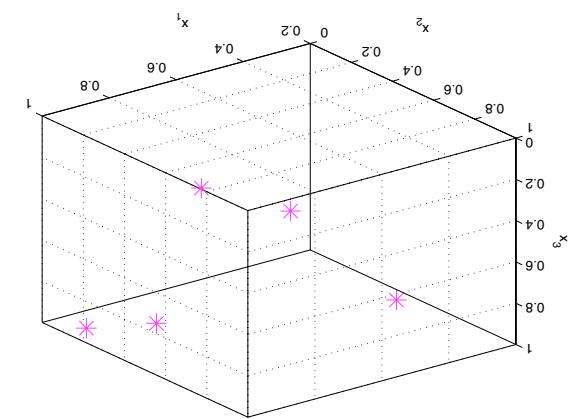


Figure 3.30: 3D Scatter Plot

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 (3.34) as follows:

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

This gives the formula:

$$f_Y(y) = \mathbf{P}(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{x: g(x)=y\}} f_X(x) . \quad (3.35)$$

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

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

If $Y = 2X$ then the transformation $g(X) = 2X$ has inverse image $g^{[-1]}(y) = \{y/2\}$. Then, by Equation (3.35) the probability mass function of Y is expressed in terms of the known probabilities of X as:

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

and tabulated below:

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

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

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

and tabulated below:

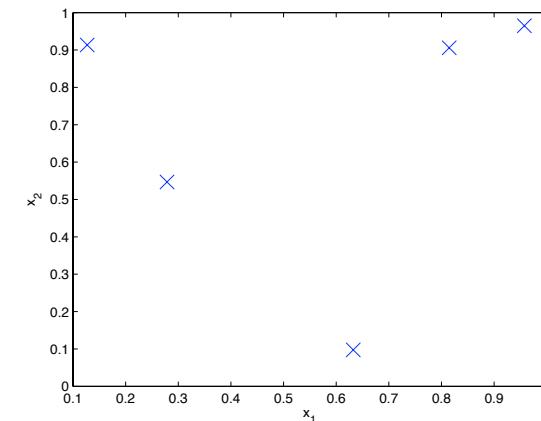
	y	-1	1	3
$f_Y(y)$		$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

```

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

$$f_Y(y) = \frac{dy}{dY} F_Y(y) = \frac{dy}{dY} F_{g^{-1}(y)}(y) = f_X(g^{-1}(y))$$

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

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

distribution function of $Y = g(X)$ in terms of the distribution function of X as random variable X . In this case g^{-1} is also an increasing function and we can obtain the

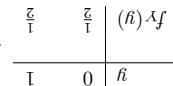
- First, let us consider the case when g is **monotone and increasing** on the range of the random variable X .

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

One-to-one transformations

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

3.6.3 Transformations of continuous random variables



and finally tabulated below:

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

computed for each $y \in [0, 1]$ as follows:

$$f_Y(y) = P(Y = y) = \sum_{x:g(x)=y} f_X(x) = (x) f_X(x).$$

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 one-to-one function. Then, by Equation (3.35) the probability mass function of Y is expressed in terms of the known probabilities of X as:

In fact, obtaining the probability of a one-to-one transformation of a discrete random variable is not one-to-one the number of terms in the sum can be more than one as shown in the next Example.

the transformation is not one-to-one the number of terms in the sum that appears in Equation (3.35). When 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 $R(x)$.

```
>>> rand('twister',5489);
>>> x=rand(2,5) % create a sequence of 5 ordered pairs uniformly from unit square [0,1]x[0,1]
x =
    0.847   0.1270   0.6324   0.2785   0.9575
```

2D scatter plot as shown in Figure 3.29 as follows:

of 5 ordered pairs sampled uniformly at random over the unit square $[0, 1] \times [0, 1]$. We can make

Labwork 115 (Visualising bivariate data) Let us generate a 2×5 array representing samples in orthogonal Cartesian co-ordinates. Such plots are termed 2D **scatter plots** in statistics of points (x_1, x_2) as $i = 1, 2, \dots, n$. The most elementary visualisation of these n ordered pairs is in the Stats Toolbox of MATLAB. Such plots are termed 2D **scatter plots** in statistics.

By bivariate data array x we mean a $2 \times n$ matrix of real numbers or equivalently n ordered pairs

3.14.3 Bivariate Data

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 in the Statistics Toolbox of MATLAB. We can also visually summarise univariate data using the **box-whisker Plot** available

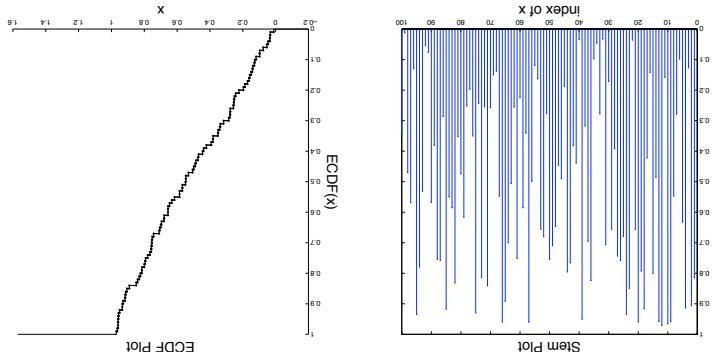


Figure 3.28: Frequency, Relative Frequency and Density Histograms

```
>>> x=rand(1,100); % produce 100 samples with random
>>> stem(x,6); % make a stem plot of the 100 data points in x (the option ' ' gives solid circles for x)
>>> ecdf(x,6); % make a step ECDF plot of the 100 data points in x (the option ' ' gives solid circles for x)
>>> %% ECDF(6,2,6); % second parameter 6 makes the dots in the plot smaller.
>>> %% (second parameter 6 makes the dots in the plot smaller).
>>> %% histogram(6,2,6); % produce 100 samples with random
>>> %% x=rand(1,100); % make a stem plot of the 100 data points in x (the option ' ' gives solid circles for x)
>>> %% stem(x,6); % make a stem plot of the 100 data points in x (the option ' ' gives solid circles for x)
>>> %% ECDF(6,2,6); % second parameter 6 makes the dots in the plot smaller.
>>> %% (second parameter 6 makes the dots in the plot smaller).
```

100 data points in the array x using stem plot and ECDF plot as shown in Figure 3.28 as follows:

We can also visualise the

Labwork 114 (Stem plots and ECDF plots for univariate data) We can make a stem plot and ECDF plot as shown in Figure 3.28 as follows:

Try making a density histogram with 1000 samples from 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

- 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 .

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| . \quad (3.36)$$

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 (3.36) 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 tranformed random variable Y . Let us follow the four steps and use the change of variable formula to obtain f_Y from f_X and g .

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

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

4. we can use Equation (3.36) 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 $0 \leq y \leq 1$

For a given partition of the data range $\mathcal{R}(x)$ or some superset of $\mathcal{R}(x)$, three types of histograms are possible: frequency histogram, relative frequency histogram and density histogram. Typically, the partition b is assumed to be composed of m overlapping intervals of the same width $w = \bar{b}_i - \underline{b}_i$ for all $i = 1, 2, \dots, m$. Thus, a histogram can be obtained by a set of bins along with their corresponding heights:

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

Thus, h_k , the height of the k -th bin, is some function g of the number of data points that fall in the bin b_k . Formally, a histogram is a sequence of ordered pairs:

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

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

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

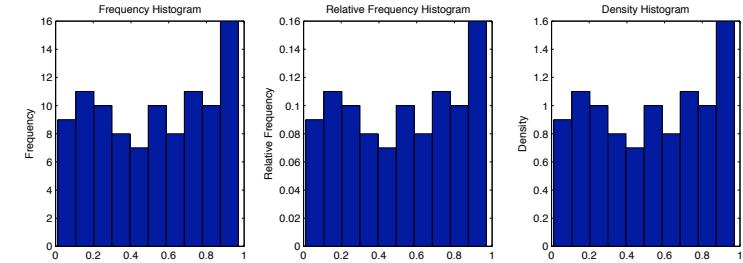
a **relative frequency histogram** is the histogram:

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

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

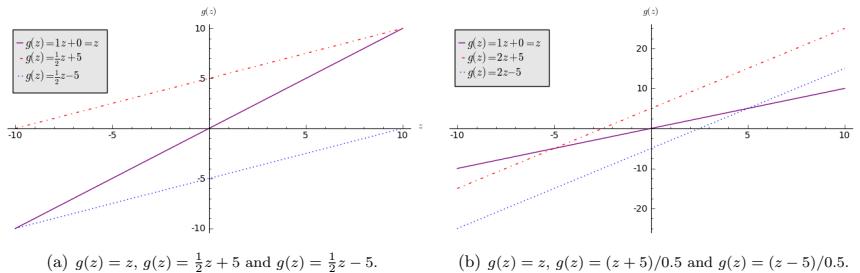
$$((b_1, h_1), (b_2, h_2), \dots, (b_m, h_m)) , \text{ where } h_k := (w_k n)^{-1} \#\{x_i : x_i \in b_k\} , w_k := \bar{b}_k - \underline{b}_k .$$

Figure 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
```

4. we can use Equation (3.36) and Equation (3.32) which gives

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

to find the density of Y as follows:

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

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

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

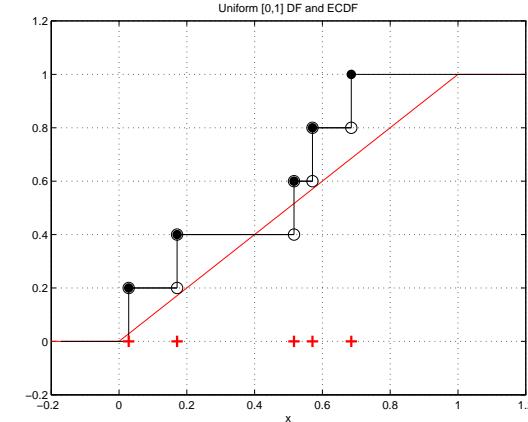
Model 11 (Normal(μ, σ^2) RV) Given a location parameter $\mu \in (-\infty, +\infty)$ and a scale parameter $\sigma^2 > 0$, the Normal(μ, σ^2) or Gaussian(μ, σ^2) random variable X has probability density function:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2\right] \quad (\sigma > 0). \quad (3.37)$$

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

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

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 56 (q^{th} Sample Quantile) For some $q \in [0, 1]$ and n IID RVs $X_1, X_2, \dots, X_n \stackrel{\text{IID}}{\sim} F$, we can obtain the ECDF \hat{F}_n using (3.80). The q^{th} sample quantile is defined as the statistic (statistical functional):

$$T(\hat{F}_n) = \hat{F}_n^{[-1]}(q) := \inf \{x : \hat{F}_n^{[-1]}(x) \geq q\}. \quad (3.81)$$

By replacing q in this definition of the q^{th} sample quantile by 0.25, 0.5 or 0.75, we obtain the first, second (**sample median**) or third **sample quartile**, respectively.

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

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

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

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^{th} sample quantile) Use the implementation of Algorithm 1 as the MATLAB function `qthSampleQuantile` to find the q^{th} sample quantile of two simulated data arrays:

1. `SortedXsFromUni01Twstr101`, the order statistics that was constructed in Labwork 110 and
2. Another sorted array of 7 samples called `SortedXs`

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

$$F(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp \left[-\frac{1}{2} \left(\frac{v-\mu}{\sigma} \right)^2 \right] dv \quad (3.38)$$

The normal distribution has the distribution function

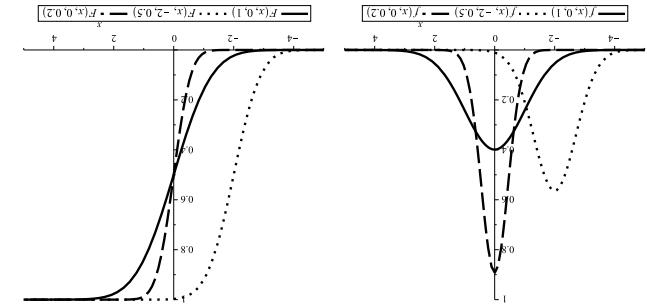


Figure 3.15: PDF and DF of a $\text{Normal}(\mu, \sigma^2)$ RV for different values of μ and σ^2

Proposition 27 (One Table to Rule Them All Gaussians) The distribution function $F(x; \mu, \sigma^2)$ 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:

$$F(x; \mu, \sigma^2) = P(X \leq x) = P(g(Z) \leq x) = P(\sigma Z + \mu \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right).$$

We know that if $X = g(Z) = \sigma Z + \mu$ then X is the $\text{Normal}(\mu, \sigma^2)$ random variable. Therefore,

Proof: Let Z be a $\text{Normal}(0, 1)$ random variable with distribution function $\Phi(z) = P(Z \leq z)$.

$$F(x; \mu, \sigma^2) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

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:

$$F(x; \mu, \sigma^2) = P(X \leq x) = P(g(Z) \leq x) = P(\sigma Z + \mu \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right).$$

Proposition 28 (CDF of Empirical CDF) Let us plot the ECDF for the five samples drawn from the $Unif(0, 1)$ 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:

```
xs = -1:0.01:2; % vector xs from -1 to 2 with increment .05 for x values
cdfzeroes = find(xs>=1); cdf(zeroes) = 1; % set cdf as 1 when xs >= 1
cdf = zeros(1,1); % initialize cdf to zero
for i=1:10 % loop over 10 samples
    plot(xs, cdf, 'r'); % plot the DF or cdf of xs in vector cdf
    hold on; % turn on grid
    plot(x,zeros(1,5), 'x+', 'LineWidth', 2, 'MarkerSize', 10); % plot the data samples
    plot(x,zeros(1,5), 'x+', 'LineWidth', 2, 'MarkerSize', 10); % plot the data samples
    xlabel('x'), ylabel('cdf(z)'), axes([0.2 1.2 -0.2 1.2])
    title('Empirical CDF vs True DF');
end;
```

Labwork 111 (**Plot of empirical CDF**) Let us plot the ECDF for the five samples drawn from the $Unif(0, 1)$ 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:

$$\hat{F}_n(x) = \sum_{i=1}^n \mathbb{I}(X_i \leq x), \quad \text{where } \mathbb{I}(X_i \leq x) := \begin{cases} 0 & \text{if } x_i < x \\ 1 & \text{if } x_i \leq x \end{cases} \quad (3.80)$$

that puts a probability mass of $1/n$ at each sample point x_i . Suppose we have n IID sample empirical distribution function (EDF or ECDF) \hat{F} over the discrete distribution function F_n RVs, X_1, X_2, \dots, X_n . \hat{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 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:

Definition 55 (Empirical Distribution Function (EDF or ECDF)) Suppose we have n IID

$$T(F) = F_{[-1]}(0.75).$$

4. The third quartile or the 0.75^{th} quantile of the RV $X \sim F$:

$$T(F) = F_{[-1]}(0.50).$$

3. The median or the second quartile or the 0.50^{th} quantile of the RV $X \sim F$:

$$T(F) = F_{[-1]}(0.25).$$

2. The first quartile or the 0.25^{th} quantile of the RV $X \sim F$:

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

1. The q^{th} quantile of RV $X \sim F$:

$$T(F) = F_{[-1]}(q).$$

Other functionals of F that depend on the quantile function $F_{[-1]}$ are:

$$T(F) = F(x).$$

3. The value of DF at a given $x \in \mathbb{R}$ of RV $X \sim F$ is also a function of DF F :

$$T(F) = \mathbb{E}(X) - \mathbb{E}(X)^2 = \int (x - \mathbb{E}(X))^2 dF(x).$$

2. The variance of RV $X \sim F$ is a function of the DF F :

$$T(F) = \mathbb{E}(X) = \int x dF(x).$$

1. The mean of RV $X \sim F$ is a function of the DF F :

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 (3.34) as follows:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(Y \in (-\infty, y]) \\ &= P(g(X) \in (-\infty, y]) = P\left(X \in g^{[-1]}((-\infty, y])\right) = P(X \in \{x : g(x) \in (-\infty, y]\}) \end{aligned} \quad (3.39)$$

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:

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 53 (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:

$$X_{([n])}((X_1, X_2, \dots, X_n)) := (X_{(1)}, X_{(2)}, \dots, X_{(n)}), \text{ such that, } X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)} . \quad (3.78)$$

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^{th} quantile, by first defining the function:

Definition 54 (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:

$$F^{[-1]}(q) := \inf\{x : F(x) > q\}, \quad \text{for some } q \in [0, 1] . \quad (3.79)$$

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:

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

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

For brevity, we write $S_n(x_1, x_2, \dots, x_n)$ as S_n and its realization $S_n(x_1, x_2, \dots, x_n)$ as s_n .

$$S_n(x_1, x_2, \dots, x_n) = \sqrt{S_n^2(x_1, x_2, \dots, x_n)} \quad (3.77)$$

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

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

$$T_n(x_1, x_2, \dots, x_n) = S_n^2(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X}_n)^2. \quad (3.76)$$

simply the sample variance :

Definition 52 (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

```
ans = 0.3945
>>> Xsample=matrix1 * ones(5,1) % multiplying an 1 x 5 matrix with a 5 x 1 matrix of ones
ans = 1.9723
>>> Xsample=matrix1 * ones(5,1) % multiplying an 1 x 5 matrix with a 5 x 1 matrix of ones
ans = 1
>>> Xsample=matrix1 * ones(5,1) % here ones(5,1) is an array of 1's with size or dimension 5 x 1
ans = 1
>>> size(Xsample)=gives the size or dimensions of the array Somearray
ans = 1
>>> sum(Xsample)=take the sum of the elements of the Xsample array
ans = 0.3945
>>> ans=(Xsample)/5 % divide the sum by the sample size 5
ans = 0.9723
>>> disp(SampleMean); % The Sample mean is :
0.3945
>>> disp(SampleMean); % The Sample mean is :
0.5707 0.0285 0.1715 0.6853
>>> SampleMean=mean(Xsample); % find sample mean
>>> Xsample=matrix1 * ones(5,1) % simulate n=5 IID samples from uniform(0,1) RV
>>> rand('twister',101); % initialize the fundamental uniform(0,1) sampler
>>> Xsample=matrix1 * ones(5,1) ; simulate n=5 IID samples from uniform(0,1) RV
>>> disp(SampleMean); % The data-points x_1,x_2,x_3,x_4,x_5 are:
0.5164 0.5044 0.4923 0.4802 0.4681
```

We can also obtain the sample mean via matrix product or multiplication as follows:

```
ans = 0.3945
>>> sum(Xsample)=take the sum of the elements of the Xsample array
ans = 1.9723
>>> ans=(Xsample)/5 % divide the sum by the sample size 5
ans = 0.3945
>>> disp(SampleMean); % The Sample mean is :
```

We may also obtain the sample mean using the sum function and a division by sample size:

We can thus use mean to obtain the sample mean \bar{x} of n sample points x_1, x_2, \dots, x_n .

```
0.3945
>>> disp(SampleMean); % The Sample mean is :
0.5707 0.0285 0.1715 0.6853
>>> SampleMean=mean(Xsample); % find sample mean
>>> Xsample=matrix1 * ones(5,1) % simulate n=5 IID samples from uniform(0,1) RV
>>> disp(SampleMean); % The data-points x_1,x_2,x_3,x_4,x_5 are:
0.5164 0.5044 0.4923 0.4802 0.4681
```

the samples stored in the array **Xsample** and then obtain the sample mean using the MATLAB function **mean**. In the following, we will reuse and then obtain the sample mean using the MATLAB function **mean**. After initializing the fundamental sampler, we draw five samples

- If $y \geq 0$ then

$$\bullet \text{ If } y < 0 \text{ then } f_Y(y) = \frac{dy}{p} (F_X(\sqrt{y}) - F_X(-\sqrt{y})) = 0.$$

- If $y \geq 0$ then

$$\bullet \text{ If } y < 0 \text{ then } f_Y(y) = \frac{dy}{p} (F_X(\sqrt{y}) - F_X(-\sqrt{y})) = 0.$$

the beginning of last century to settle “supposedly evidence-based disputes” among scientists using 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.

$$Y \text{ is called the } \text{chi-square} \text{ random variable with one degree of freedom. This distribution plays a fundamental role in hypothesis testing as we will see in Inference Theory and was derived at the beginning of last century to settle “supposedly evidence-based disputes” among scientists using }$$

$$\left. \begin{cases} \frac{z^2}{1} f_X(\sqrt{y}) & \text{if } y \geq 0 \\ 0 & \text{if } y < 0 \end{cases} \right\} = f_Y(y) = \frac{\sqrt{y}}{1} \exp(-\frac{y}{2})$$

then by Equation (3.41) the density of $Y = X^2$ is:

$$f_X(x)\phi = (x) = \frac{\sqrt{x}}{1} \exp(-x/2)$$

Example 70 If X is the standard normal random variable with density

$$\left. \begin{cases} \frac{z^2}{1} f_X(\sqrt{y}) & \text{if } y \geq 0 \\ 0 & \text{if } y < 0 \end{cases} \right\} = f_Y(y) = \frac{\sqrt{y}}{1} \exp(-y/2)$$

and the probability density function of $Y = X^2$ is:

$$\left. \begin{cases} F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \text{if } y \geq 0 \\ 0 & \text{if } y < 0 \end{cases} \right\} = F_Y(y)$$

Therefore, the distribution function of $Y = X^2$ is:

$$\begin{aligned} & \left. \begin{aligned} & F_Y(y) = \frac{2}{1} \exp(-y/2) \\ & = \frac{2}{1} \exp(-y/2) - \frac{2}{1} \exp(-y/2) \end{aligned} \right\} = F_Y(y) = \frac{dy}{p} (F_X(\sqrt{y}) - F_X(-\sqrt{y})) \end{aligned}$$

- If $y \geq 0$ then

$$\bullet \text{ If } y < 0 \text{ then } f_Y(y) = \frac{dy}{p} (F_X(\sqrt{y}) - F_X(-\sqrt{y})) = 0.$$

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(θ) random variable and $Y = (\frac{1}{2})^X$ then find an expression for $f_Y(y)$.

Ex. 3.25 — If X is a Poisson(λ) random variable find the probability mass function, $f_Y(y)$, of

$$Y = \frac{1}{(X+1)^2}.$$

Ex. 3.26 — If X is a continuous random variable with probability density function

$$f_X(x) = \begin{cases} xe^{-x} & x \geq 0 \\ 0 & x < 0, \end{cases}$$

find the probability density function of $Y = e^X$.

Ex. 3.27 — If X , the received power at an antenna is an Exponential(λ) random variable then find the probability density function of the amplitude $Y = \sqrt{X}$.

Ex. 3.28 — If X is a Uniform(a, b) random variable where $0 < a < b$, find the probability density function, $f_Y(y)$, of

$$Y = \log_e(X).$$

3.8 Expectations

Expectation is perhaps the most fundamental concept in probability theory. In fact, probability is itself an expectation as you will soon see!

Expectation is one of the fundamental concepts in probability. The expected value of a real-valued random variable gives the population mean, a measure of the centre of the distribution of the variable in some sense. Its variance measures its spread and so on.

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

Next, we define two important statistics called the **sample mean** and **sample variance**. Since they are obtained from the sample data, they are called **sample moments**, as opposed to the **population moments**. The corresponding population moments are $\mathbf{E}(X_1)$ and $\mathbf{V}(X_1)$, respectively.

Definition 51 (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:

$$T_n((X_1, X_2, \dots, X_n)) = \bar{X}_n((X_1, X_2, \dots, X_n)) := \frac{1}{n} \sum_{i=1}^n X_i. \quad (3.73)$$

For brevity, we write

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

and its realisation

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

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

$$\begin{aligned} \mathbf{E}(\bar{X}_n) &= \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) && [\text{by definition (3.73)}] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{E}(X_i) && [\text{by property (3.48)}] \end{aligned}$$

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

$$\mathbf{E}(\bar{X}_n) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}(X_i) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}(X_1) = \frac{1}{n} n \mathbf{E}(X_1) = \mathbf{E}(X_1). \quad (3.74)$$

Similarly, we can show that:

$$\begin{aligned} \mathbf{V}(\bar{X}_n) &= \mathbf{V}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) && [\text{by definition (3.73)}] \\ &= \left(\frac{1}{n}\right)^2 \mathbf{V}\left(\sum_{i=1}^n X_i\right) && [\text{by property (3.47)}] \end{aligned}$$

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

$$\mathbf{V}(\bar{X}_n) = \left(\frac{1}{n}\right)^2 \mathbf{V}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \mathbf{V}(X_i) \quad [\text{by property (3.49)}]$$

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

$$\mathbf{V}(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^n \mathbf{V}(X_i) = \frac{1}{n^2} \sum_{i=1}^n \mathbf{V}(X_1) = \frac{1}{n^2} n \mathbf{V}(X_1) = \frac{1}{n} \mathbf{V}(X_1). \quad (3.75)$$

Definition 29 (Variance of a RV) Let X be a RV with mean or expectation $\mathbb{E}(X)$. Variance of X denoted by $\mathbb{V}(X)$ or $V(X)$

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

provided this expectation exists. The standard deviation denoted by $\text{sd}(X) =: \sqrt{\mathbb{V}(X)}$. Thus

$$\mathbb{E}^\theta X := \int x dF(x; \theta)$$

with respect to which the integration is understood. We say the expectation exists if

$$\int |x| dF(x) < \infty.$$
(3.43)

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

Sometimes, we denote $\mathbb{E}(X)$ by $\mathbb{E}X$ for brevity. Thus, the expectation is a single-number summary of the RV X and may be thought of as the average. We use subscript θ to specify the parameter $\theta \in \Theta$ of the RV X , for brevity. Thus, the expectation is a single-number summary provided the sum or integral is well-defined. We say the expectation exists if

Definition 28 (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

$$\mathbb{E}(X) := \begin{cases} \int x f(x) dx & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous,} \end{cases}$$
(3.42)

Definition 29 (Expectation of a RV) The expectation, or expected value, or mean, or

provided $\mathbb{E}(g(X))$ exists, i.e., $\int |g(x)| dF(x) < \infty$.

$$\mathbb{E}(g(X)) := \begin{cases} \sum_x g(x) f(x) & \text{if } X \text{ is a discrete RV} \\ \int_x g(x) f(x) dx & \text{if } X \text{ is a continuous RV} \end{cases}$$

Definition 31 (Expectation of a function of a RV) The Expectation of a function $g(X)$ of a random variable X is defined as:

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.

3.8.1 Expectations of functions of random variables

$$\mathbb{E}(X^k) = \int x^k dF(x)$$

as the k -th moment of the RV X and say that the k -th moment exists when $\mathbb{E}(|X|^k) < \infty$. We call the following expectation as the k -th central moment:

$$\mathbb{E}(X^k) = \int (x - \mathbb{E}(X))^k dF(x)$$

Definition 30 (k -th moment of a RV) We call

provided this expectation exists. The standard deviation denoted by $\text{sd}(X) =: \sqrt{\mathbb{V}(X)}$. Thus

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

which takes values in the data space $\mathbb{X} = \{0, 1\}^n$. Then the random vector $X = (X_1, X_2, \dots, X_n)$, upon performance of the n -dimensional hyper-cube, measures the outcomes of this experiment. A particular realization 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 $(x_1, x_2, \dots, x_{n-1}, x_n) = (0, 0, \dots, 0, 1)$.

Example 106 (Tossing a coin n times) For some given parameter $\theta \in \Theta = \{0, 1\}$, consider

Figure 3.25: Data Spaces $\mathbb{X} = \{0, 1\}^2$ and $\mathbb{X} = \{0, 1\}^3$ for two and three Bernoulli trials, respectively.

Classwork 107 (Is data a statistic?) Is the RV X , for which the realization is the observed

we use $T^n(X)$ and \mathbb{T}^n to emphasize that X is an n -dimensional random vector, i.e., $\mathbb{X} \subset \mathbb{R}^n$. Thus, a statistic T is also an RV that takes values in the space \mathbb{T} . When $x \in \mathbb{X}$ is the realization of the statistic T , some times an experiment, we let $T(x) = t$ denote the corresponding realization of the statistic T . Some times a random variable X is also an RV the realization of which is the observed vector x . When $x \in \mathbb{X}$ is the realization of the statistic T , some times an experiment, we let $T(x) : \mathbb{X} \rightarrow \mathbb{T}$.

Definition 50 (Statistic) A statistic T is any function of the data:

Figure 3.24: Sample Space, Random Variable, Realization, Data, and Data Space. For some given parameter $\theta \in \Theta = \{0, 1\}$, consider n IID Bernoulli(θ) trials, i.e., X_1, X_2, \dots, X_n iid Bernoulli(θ). Then the random vector $X = (X_1, X_2, \dots, X_n)$, which takes values in the data space $\mathbb{X} = \{0, 1\}^n$. The realization of X is the random vector $x = (x_1, x_2, \dots, x_n) : x_i \in \{0, 1\}$, $i = 1, 2, \dots, n$, made up of vertices of the n -dimensional hyper-cube, measures the outcomes of this experiment. A particular realization of X , upon performance of this experiment, is the realization, data or data vector (x_1, x_2, \dots, x_n) . For instance, if we observed $n - 1$ tails and 1 head, in that order, then our data vector $(x_1, x_2, \dots, x_{n-1}, x_n) = (0, 0, \dots, 0, 1)$.

Figure 3.25: Data Spaces $\mathbb{X} = \{0, 1\}^2$ and $\mathbb{X} = \{0, 1\}^3$ for two and three Bernoulli trials, respectively.

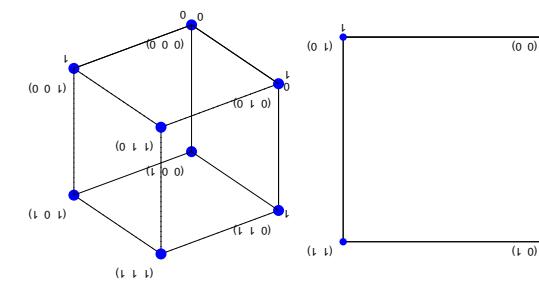


Figure 3.24: Sample Space, Random Variable, Realization, Data, and Data Space.

The **mean** which characterises the central location of the random variable X is merely the expectation of the identity function $g(x) = x$:

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

Often, mean is denoted by μ .

The **variance** which characterises the spread or the variability of the random variable X is also the expectation of the function $g(x) = (x - \mathbf{E}(X))^2$:

$$\mathbf{V}(X) = \mathbf{E}((X - \mathbf{E}(X))^2) = \begin{cases} \sum_x (x - \mathbf{E}(X))^2 f(x) & \text{if } X \text{ is a discrete RV} \\ \int_{-\infty}^{\infty} (x - \mathbf{E}(X))^2 f(x) dx & \text{if } X \text{ is a continuous RV} \end{cases}$$

Often, variance is denoted by σ^2 .

INTUITIVELY, WHAT IS EXPECTATION?

Definition 31 gives expectation as a “weighted average” of the possible values. This is true but some intuitive idea of expectation is also helpful.

- Expectation is what you expect.

Consider tossing a fair coin. If it is heads you lose \$10. If it is tails you win \$10. What do you expect to win? Nothing. If X is the amount you win then

$$\mathbf{E}(X) = -10 \times \frac{1}{2} + 10 \times \frac{1}{2} = 0.$$

So what you expect (nothing) and the weighted average ($\mathbf{E}(X) = 0$) agree.

- Expectation is a long run average.

Suppose you are able to repeat an experiment independently, over and over again. Each experiment produces one value x of a random variable X . If you take the average of the x values for a large number of trials, then this average converges to $\mathbf{E}(X)$ as the number of trials grows. In fact, this is called the **law of large numbers**.

We can concretize the above two intuitive insights by the following two examples.

Example 71 (Winnings on Average) Let $Y = r(X)$. Then

$$\mathbf{E}(Y) = \mathbf{E}(r(X)) = \int r(x) dF(x).$$

Think of playing a game where we draw $x \sim X$ and then I pay you $y = r(x)$. Then your average income is $r(x)$ times the chance that $X = x$, summed (or integrated) over all values of x .

Example 72 (Probability is an Expectation) Let A be an event and let $r(X) = \mathbb{1}_A(x)$. Recall $\mathbb{1}_A(x)$ is 1 if $x \in A$ and $\mathbb{1}_A(x) = 0$ if $x \notin A$. Then

$$\mathbf{E}(\mathbb{1}_A(X)) = \int \mathbb{1}_A(x) dF(x) = \int_A dF(x) = \mathbf{P}(X \in A) = \mathbf{P}(A) \quad (3.44)$$

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

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. 3.50 — Recall that the Geometric(θ) RV X has the following PMF

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

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. 3.51 — 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. 3.52 — Recall that the Poisson(λ) RV has the following PMF

$$f_X(x; \lambda) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{if } x \in \{0, 1, 2, 3, \dots\} \\ 0 & \text{otherwise} \end{cases}$$

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. 3.53 — Let X be a Poisson(λ) RV and Y be another Poisson(μ) RV. Suppose X and Y are independent. Use Eqn. (3.72) to first find the CF of the RV $W = X + Y$. From the CF of W try to identify what RV it is.

Ex. 3.54 — Recall from lecture that if $Y = a + bX$ for some constants a and b with $b \neq 0$ then $\phi_Y(t) = e^{yat} \phi_X(bt)$ and that $\phi_Z(t) = e^{-t^2/2}$ 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 49 (Data) The function X measures the outcome ω of an experiment with sample space Ω [Often, the sample space is also denoted by S]. Formally, X is a random variable [or a random vector $X = (X_1, X_2, \dots, X_n)$, i.e. a vector of random variables] taking values in the **data space** \mathbb{X} :

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

The realisation of the RV X when an experiment is performed is the observation or data $x \in \mathbb{X}$. That is, when the experiment is performed once and it yields a specific $\omega \in \Omega$, the data $X(\omega) = x \in \mathbb{X}$ is the corresponding realisation of the RV X .

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

Example 73 (Mean and variance of Point Mass(θ) RV) Let $X \sim \text{Point Mass}(\theta)$. Then:

Thus, Point Mass(θ) RV X is deterministic in the sense that every realization of X is exactly equal to $\theta \in \mathbb{R}$. We will see that this distribution plays a central limiting role in asymptotic statistics.

Using them here as it is more convenient to work in the complex plane.

$$f(x; \theta) = \begin{cases} 1 & \text{if } x = \theta \\ 0 & \text{if } x \neq \theta \end{cases} \quad (3.46)$$

and the PMF is:

$$P(X = x; \theta) = \begin{cases} 1 & \text{if } x < \theta \\ 0 & \text{if } x \geq \theta \end{cases} \quad (3.45)$$

Model 12 (Point Mass(θ)) Given a specific point $\theta \in \mathbb{R}$, we say an RV X has point mass at θ or is Point Mass(θ) distributed if the DF is:

Consider the class of discrete RVs with distributions that place all probability mass on a single real number. This is the probabilistic model for the deterministic real variable, which is often thought of as an unknown constant $\theta \in \mathbb{R}$.

The same ideas naturally extend, via multiple sums and integrals, to define the expectation of functions of \mathbb{R}^k -valued random variables with $k > 2$.

The same ideas naturally extend, via multiple sums and integrals, to define the expectation of two RVs.

Viewing a deterministic real variable as a random variable

$$\text{Cov}(X, Y) := \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

This allows the definition of covariance of X and Y as $\text{Cov}(X, Y) < \infty$ and $\mathbb{E}(|XY|) < \infty$ and $\mathbb{E}(|(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))|) < \infty$.

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

When $r = s = 1$, we have $\mathbb{E}(XY)$, the expectation of the product of two RVs.

$$\mathbb{E}(XY)$$

1. Joint Moments

Some typical expectations for \mathbb{R}^2 -valued random variables are:

$$\mathbb{E}(g(X, Y)) = \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy & \text{if } (X, Y) \text{ is a continuous RV} \\ \sum_{(x,y)} g(x, y) f_{X,Y}(x, y) & \text{if } (X, Y) \text{ is a discrete RV} \end{cases}$$

Definition 32 The Expectation of a function $g(X, Y)$ of the \mathbb{R}^2 -valued RV (X, Y) is defined as:

In the case of a single random variable we saw that its expectation gives the population mean, the 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.

Example 105 Let Z_1 and Z_2 be independent $\text{Normal}(0, 1)$ RVs.

Expectations of functions of \mathbb{R}^2 -valued random variables

$$f_X(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} & \text{if } x = 0 \\ \frac{3}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}} & \text{if } x = 1 \\ \frac{3}{\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}} & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Ex. 3.49 — Let X be a discrete random variable (RV) with probability mass function (PMF)

3.13 Exercises in Characteristic Functions

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.

3.12.2 Moment Generating Function

In other words $2 \times Z$ is not equal to $Z + Z$ in terms of its probability distribution! Adding the same RV twice does not have the same distribution as that of multiplying it by 2. A smaller variance from adding two independent standard normal RVs. Thus, the result of 5. $Z_1 + Z_2$ has a bigger variance from multiplying the standard normal RV by 2 while $Z_1 + Z_2$ has the Normal($0, 4$) RV with mean parameter $\mu = 0$ and variance parameter $\sigma^2 = 4$.

4. The CF of Z_1 is that of the Normal(μ, σ^2) RV with $\mu = 0$ and $\sigma^2 = 2^2 = 4$. Thus Z_1 is

$$\phi_{Z_1}(t) = \phi_Z(t) = e^{-\frac{t^2}{2}}$$

3. We can again use Eqn. (3.72) to find the CF of Z_1 as follows

2. The CF of $Z_1 + Z_2$ is that of the Normal(μ, σ^2) RV with $\mu = 0$ and variance parameter $\sigma^2 = 2$. Thus $Z_1 + Z_2$ is the Normal($0, 2$) RV with mean parameter $\mu = 0$ and variance parameter $\sigma^2 = 2$.

$$\phi_{Z_1+Z_2}(t) = \phi_{Z_1}(t) \times \phi_{Z_2}(t) = e^{-\frac{t^2}{2}} \times e^{-\frac{t^2}{2}} = e^{-2t^2/2} = e^{-t^2}$$

1. By Eqn. (3.72) we just multiply the characteristic functions of Z_1 and Z_2 , both of which are

Solution:
Hint: from lectures we know that $\phi_X(t) = e^{\mu t - (\sigma^2 t^2)/2}$ for a Normal(μ, σ^2) RV X .

5. Try to understand the difference between the distributions of $Z_1 + Z_2$ and $2Z_1$ despite of Z_1 and Z_2 having the same distribution.

4. From the CF of Z_1 identify what RV it is.

3. Use Eqn. (3.72) to find the CF of Z_2 .

2. From the CF of $Z_1 + Z_2$ identify what RV it is.

1. Use Eqn. (3.72) to find the CF of $Z_1 + Z_2$.

3.8.2 Properties of expectations

The following results, where a is a constant, may easily be proved using the properties of summations and integrals:

$$\boxed{\mathbf{E}(a) = a}$$

$$\boxed{\mathbf{E}(a g(X)) = a \mathbf{E}(g(X))}$$

$$\boxed{\mathbf{E}(g(X) + h(X)) = \mathbf{E}(g(X)) + \mathbf{E}(h(X))}$$

Note that here $g(X)$ and $h(X)$ are functions of the random variable X : e.g. $g(X) = X^2$.

Using these results we can obtain the following useful formula for variance:

$$\begin{aligned} V(X) &= E((X - \mathbf{E}(X))^2) \\ &= E(X^2 - 2X\mathbf{E}(X) + (\mathbf{E}(X))^2) \\ &= E(X^2) - E(2X\mathbf{E}(X)) + E((\mathbf{E}(X))^2) \\ &= E(X^2) - 2\mathbf{E}(X)\mathbf{E}(X) + (\mathbf{E}(X))^2 \\ &= E(X^2) - 2(\mathbf{E}(X))^2 + (\mathbf{E}(X))^2 \\ &= \mathbf{E}(X^2) - (\mathbf{E}(X))^2 . \end{aligned}$$

That is,

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

The above properties of expectations imply that for constants a and b ,

$$\boxed{\mathbf{V}(aX + b) = a^2\mathbf{V}(X)} . \quad (3.47)$$

More generally, for random variables X_1, X_2, \dots, X_n and constants a_1, a_2, \dots, a_n

- $\mathbf{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mathbf{E}(X_i) . \quad (3.48)$

- $\mathbf{V}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \mathbf{V}(X_i), \text{ provided } X_1, X_2, \dots, X_n \text{ are independent} . \quad (3.49)$

- Let X_1, X_2, \dots, X_n be independent RVs, then

$$\mathbf{E}\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n \mathbf{E}(X_i), \text{ provided } X_1, X_2, \dots, X_n \text{ are independent} . \quad (3.50)$$

Thus the CF of the standard normal RV Z is

$$\boxed{\phi_Z(t) = e^{-t^2/2}} \quad (3.70)$$

Let X be a RV with CF $\phi_X(t)$. 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

$$\boxed{\phi_Y(t) = \exp(itb)\phi_X(bt)} \quad (3.71)$$

Proof: This is easy to prove using the definition of CF as follows:

$$\begin{aligned} \phi_Y(t) &= E(\exp(itY)) = E(\exp(it(a + bX))) = E(\exp(itbX + ita)) \\ &= E(\exp(itbX))\exp(itbX) = \exp(itbX)E(\exp(itbX)) = \exp(itbX)\phi_X(bt) \end{aligned}$$

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 (3.70) and (3.71) find the CF of Y .

Solution:

$$\begin{aligned} \phi_Y(t) &= \exp(i\mu t)\phi_Z(\sigma t), \quad \text{since } Y = \mu + \sigma Z \\ &= e^{i\mu t}e^{(-\sigma^2 t^2)/2}, \quad \text{since } \phi_Z(t) = e^{-t^2/2} \\ &= e^{i\mu t - (\sigma^2 t^2)/2} \end{aligned}$$

A generalization of (3.71) 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

$$\boxed{\phi_Y(t) = \phi_{X_1}(a_1 t) \times \phi_{X_2}(a_2 t) \times \cdots \times \phi_{X_n}(a_n t) = \prod_{i=1}^n \phi_{X_i}(a_i t)} . \quad (3.72)$$

Example 104 Using the following three facts:

- Eqn. (3.72)
- the Binomial(n, θ) RV Y is the sum of n independent Bernoulli(θ) RVs (from Probability Course)
- the CF of Bernoulli(θ) RV (from lecture notes for Inference Course)

find the CF of the Binomial(n, θ) RV Y .

Solution:

Let X_1, X_2, \dots, X_n be independent Bernoulli(θ) RVs with CF $(1 - \theta + \theta e^{it})$ then $Y = \sum_{i=1}^n X_i$ is the Binomial(n, θ) RV and by Eqn. (3.72) with $a_1 = a_2 = \dots = 1$, we get

$$\phi_Y(t) = \phi_{X_1}(t) \times \phi_{X_2}(t) \times \cdots \times \phi_{X_n}(t) = \prod_{i=1}^n \phi_{X_i}(t) = \prod_{i=1}^n (1 - \theta + \theta e^{it}) = (1 - \theta + \theta e^{it})^n .$$

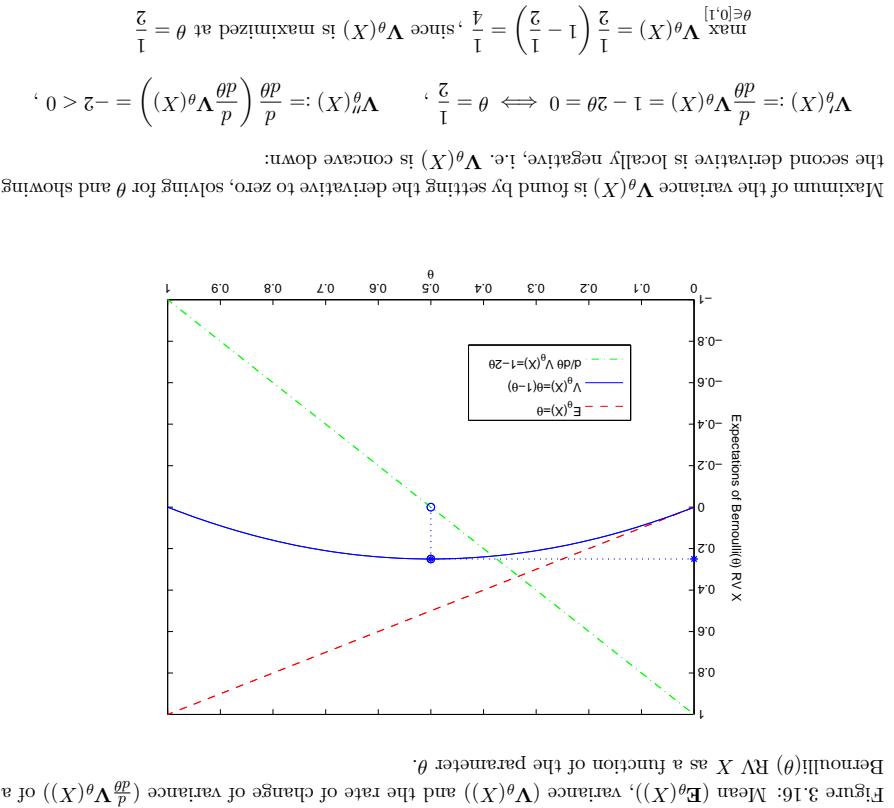


Figure 3.16: Mean ($E^{\theta}(X)$), variance ($V^{\theta}(X)$) and the rate of change of variance ($\frac{d}{d\theta}V^{\theta}(X)$) of a Bernoulli(θ) RV X as a function of the parameter θ .

Parameter specifically,
 $E^{\theta}(X) = \theta$ and $V^{\theta}(X) = \theta(1-\theta)$.

$$V(X) = E(X^2) - (E(X))^2$$

$$E(X^2) = \sum_{x=0}^{\infty} x^2 f_x(x) = (x)(1-\theta) + (1^2)\theta = \theta + \theta^2$$

$$E(X) = \sum_{x=0}^{\infty} x f_x(x) = (0)(1-\theta) + (1)\theta = \theta$$

Example 74 (Mean and variance of Bernoulli(θ) RV) Let $X \sim \text{Bernoulli}(\theta)$. Then,

Let us compute the mean and variance of our familiar RVs.

3.8.3 Expectation of Common Random Variables

$$\max_{\theta \in [0,1]} V^{\theta}(X) = \frac{1}{2} \left(1 - \frac{1}{2} \right)^2 = \frac{1}{4}, \text{ since } V^{\theta}(X) \text{ is maximized at } \theta = \frac{1}{2}$$

$V^{\theta}(X) = \frac{d\theta}{p} V^{\theta}(X) = 1 - 2\theta = 0 \iff \theta = \frac{1}{2}$, $V^{\theta}(X) = \frac{d\theta}{p} V^{\theta}(X) = -2 < 0$,

the second derivative is locally negative, i.e. $V^{\theta}(X)$ is concave down;

Maximum of the variance $V^{\theta}(X)$ is found by setting the derivative to zero, solving for θ and showing

$$\begin{aligned} \phi_Z(t) &= E(e^{tZ}) \\ &= \int_{-\infty}^{\infty} e^{tz} f_Z(z) dz = \int_{-\infty}^{\infty} e^{az - z^2/2} dz = \frac{\sqrt{2\pi}}{1} \int_{-\infty}^{\infty} e^{-z^2/2} dz \\ &= \int_{-\infty}^{\infty} e^{az} dz = \frac{\sqrt{2\pi}}{1} \int_{-\infty}^{\infty} e^{az - z^2/2} dz \\ &= \int_{-\infty}^{\infty} e^{-y^2/2} dy \quad \text{substituting } y = z - it, dy = dz \\ &= e^{-t^2/2} \int_{-\infty}^{\infty} e^{-y^2/2} dy = e^{-t^2/2} \text{ using constant in PDF of Normal}(0,1) \text{ RV} \end{aligned}$$

Let Z be $\text{Normal}(0,1)$, the standard normal RV. We can find the CF for Z using couple of tricks as follows

be much more challenging or impossible to do directly with their DFs. Thus, if we can show that two RVs have the same CF then we know they are the same. This can be done by showing that two DFs have the same characteristic functions, i.e. $F^X(x) = F^Y(x)$ for all $x \in \mathbb{R}$ (for proof see Resnick, S. I. (1999) A Probability Path, Birkhäuser).

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. $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}$. Characteristic functions can be used to characterize the distribution of a random variable.

It and directly computed using integrals for definition of expectation.

Let's check that this is what we had as variance for the Exponential(λ) RV when we first introduced

$$V(X) = E(X^2) - (E(X))^2 = \frac{\lambda^2}{2} - \left(\frac{\lambda}{2} \right)^2 = \frac{\lambda^2}{2} - \frac{\lambda^2}{4} = \frac{\lambda^2}{4}.$$

Finally, from the first and second moments we can get the variance as follows:

$$\begin{aligned} E(X^2) &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx = \frac{1}{2} \lambda^2 \int_0^\infty t^2 e^{-\lambda t} dt = \frac{1}{2} \lambda^2 \left(\frac{\lambda}{2} \right)^2 = \frac{\lambda^4}{4}. \\ E(X) &= \int_0^\infty x \lambda e^{-\lambda x} dx = \frac{1}{2} \lambda^2 \int_0^\infty t e^{-\lambda t} dt = \frac{1}{2} \lambda^2 \left(\frac{\lambda}{2} \right)^2 = \frac{\lambda^3}{4}. \\ \text{Similarly from Equation (3.69) we can get } E(X^2) \text{ as follows:} \end{aligned}$$

imedded $1/\lambda$ (recall from when we introduced this RV).

Let's pause and see if this makes sense.... Yes, because the expected value of Exponential(λ) RV is indeed $1/\lambda$

We get $E(X)$ by evaluating $\frac{d}{dt}\phi_X(t)$ at $t = 0$ and dividing by i according to Equation (3.69) as follows:

$$\begin{aligned} E(X) &= \frac{1}{i} \frac{d}{dt} \phi_X(t) \Big|_{t=0} = \frac{1}{i} \left[\lambda - \lambda e^{-\lambda t} \right]_{t=0} = \frac{1}{i} \left(\lambda - \lambda \right) = 0 \\ \frac{d}{dt} \phi_X(t) &= \lambda \left(1 - e^{-\lambda t} \right) \times \frac{d}{dt} \lambda = \lambda \left(1 - e^{-\lambda t} \right) \times \frac{d\lambda}{dt} = \lambda \left(1 - e^{-\lambda t} \right) \times \frac{d\lambda}{dt} = \lambda \left(1 - e^{-\lambda t} \right) \times \lambda = \lambda^2 \left(1 - e^{-\lambda t} \right) \end{aligned}$$

Let us differentiate the CF to get moments using Equation (3.69) (CF has to be once and twice differentiable at $t = 0$ to get the first and second moments).

The plot depicting these expectations as well as the rate of change of the variance are depicted in Figure 3.16. Note from this Figure that $\mathbf{V}_\theta(X)$ attains its maximum value of $1/4$ at $\theta = 0.5$ where $\frac{d}{d\theta}\mathbf{V}_\theta(X) = 0$. Furthermore, we know that we don't have a minimum at $\theta = 0.5$ since the second derivative $\mathbf{V}_\theta''(X) = -2$ is negative for any $\theta \in [0, 1]$. This confirms that $\mathbf{V}_\theta(X)$ is concave down and therefore we have a maximum of $\mathbf{V}_\theta(X)$ at $\theta = 0.5$. We will revisit this example when we employ a numerical approach called Newton-Raphson method to solve for the maximum of a differentiable function by setting its derivative equal to zero.

Example 75 (Mean and variance of Uniform(0, 1) RV) Let $X \sim \text{Uniform}(0, 1)$. Then,

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

Example 76 (Expected Exponential of the Uniform(0, 1) RV) Let $X \sim \text{Uniform}(0, 1)$ and $Y = r(X) = e^X$. Compute $\mathbf{E}(Y)$.

We can simply apply the definition of $\mathbf{E}(r(X))$, since $Y = r(X)$, is just a function of X , as follows:

$$\mathbf{E}(Y) = \int_0^1 e^x f(x) dx = \int_0^1 e^x \cdot 1 dx = e - 1.$$

Example 77 (Mean and variance of Exponential(λ) RV) Show that the mean of an Exponential(λ) RV X is:

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

and the variance is:

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

Example 78 (Mean and variance of Geometric(θ) RV) Let $X \sim \text{Geometric}(\theta)$ RV. Then,

$$\mathbf{E}(X) = \sum_{x=0}^{\infty} x\theta(1-\theta)^x = \theta \sum_{x=0}^{\infty} x(1-\theta)^x$$

In order to simplify the RHS above, let us employ differentiation with respect to θ :

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

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

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

Similarly, it can be shown that

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

Part 2:

Let's differentiate CF

$$\frac{d}{dt}\phi_X(t) = \frac{d}{dt}(1 - \theta + \theta e^{it}) = \theta i \exp(it)$$

We get $E(X)$ by evaluating $\frac{d}{dt}\phi_X(t)$ at $t = 0$ and dividing by i according to Equation (3.69) as follows:

$$E(X) = \frac{1}{i} \left[\frac{d}{dt}\phi_X(t) \right]_{t=0} = \frac{1}{i} [\theta i \exp(it)]_{t=0} = \frac{1}{i} (\theta i \exp(i0)) = \theta.$$

Similarly from Equation (3.69) we can get $E(X^2)$ as follows:

$$\begin{aligned}E(X^2) &= \frac{1}{i^2} \left[\frac{d^2}{dt^2}\phi_X(t) \right]_{t=0} = \frac{1}{i^2} \left[\frac{d}{dt} \frac{d}{dt}\phi_X(t) \right]_{t=0} = \frac{1}{i^2} \left[\frac{d}{dt} \theta i \exp(it) \right]_{t=0} \\ &= \frac{1}{i^2} [\theta i^2 \exp(it)]_{t=0} = \frac{1}{i^2} (\theta i^2 \exp(i0)) = \theta.\end{aligned}$$

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(θ) RV if we directly computed it using weighted sums in the definition of expectations: $E(X) = 1 \times \theta + 0 \times (1 - \theta) = \theta$, $E(X^2) = 1^2 \times \theta + 0^2 \times (1 - \theta) = \theta$ and thus giving the same $V(X) = E(X^2) - (E(X))^2 = \theta - \theta^2 = \theta(1 - \theta)$.

Example 102 Let X be an Exponential(λ) RV. First show that its CF is $\lambda/(\lambda - it)$. 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(λ) RV for a given parameter $\lambda \in (0, \infty)$ is $\lambda e^{-\lambda x}$ if $x \in [0, \infty)$ and 0 if $x \notin [0, \infty)$.

Part 1: Find the CF.

We will use the fact that

$$\int_0^\infty \alpha e^{-\alpha x} dx = [-e^{-\alpha x}]_0^\infty = 1$$

$$\begin{aligned}\phi_X(t) &= E(\exp(itX)) = E(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} \lambda e^{-\lambda x} dx = \lambda \int_0^\infty e^{-(\lambda-it)x} dx \\ &= \frac{\lambda}{\lambda-it} \int_0^\infty (\lambda-it)e^{-(\lambda-it)x} dx = \frac{\lambda}{\lambda-it} \int_0^\infty \alpha e^{-\alpha x} dx = \frac{\lambda}{\lambda-it},\end{aligned}$$

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:

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} \lambda e^{-\lambda x} dx = \int_{-\infty}^{\infty} \cos(tx) e^{-\lambda x} dx + i \int_{-\infty}^{\infty} \sin(tx) e^{-\lambda x} dx = \frac{\lambda}{\lambda-it}.$$

Part 2:

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:

The RHS evaluated at $t = 0$ is

$$\frac{d}{dt} \phi_X(t) = \frac{d}{dt} \mathbb{E}(\exp(tX)) = \mathbb{E}((X^n) \exp(tX)) = \mathbb{E}((X^n) \mathbb{E}_{\mathbb{P}}(\exp(tX)))$$

This completes the sketch of the proof.

$$\left(\frac{d}{dt} \phi_X(t) \right)_{t=0} = \mathbb{E} \left[\left(\frac{d}{dt} \phi_X(t) \right)_{t=0} \right] = \mathbb{E} \left[\left(\frac{d}{dt} \mathbb{E}(\exp(tX)) \right)_{t=0} \right]$$

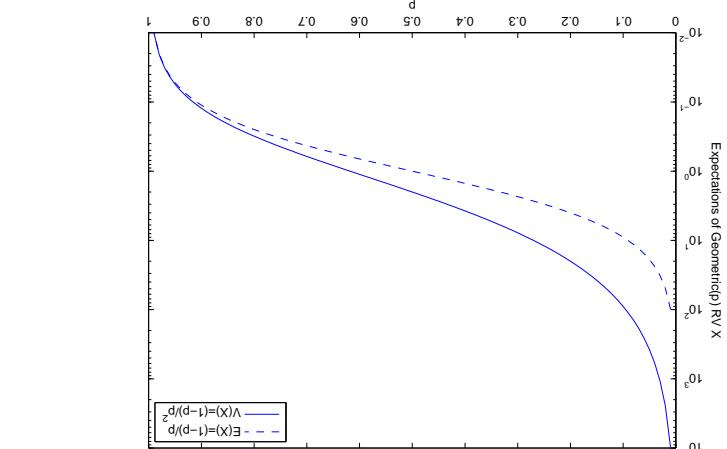


Figure 3.17: Mean and variance of a Geometric(p) RV X as a function of the parameter θ .

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 48 (Moments from CF). Let X be a random variable and $\phi_X(t)$ be its CF. If $\phi_X(t)$ 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,

$$\mathbb{E}(X^k) = \frac{1}{k!} \left[\frac{d^k}{dt^k} \phi_X(t) \right]_{t=0}$$

where $\left[\frac{d^k}{dt^k} \phi_X(t) \right]_{t=0}$ is the k -th derivative of $\phi_X(t)$ with respect to t , evaluated at the point $t = 0$.

Example 79 (Mean and variance of Binomial(n, θ) RV) Let $X \sim \text{Binomial}(n, \theta)$. Based on the definition of expectation:

$$\mathbb{E}(X) = \int x dF(x; n, \theta) = \int x \theta^n (1-\theta)^{n-x} \binom{n}{x} dx$$

However, this is a nontrivial sum to evaluate. Instead, we may use (3.48) and (3.49) by noting that $X = \sum_{i=1}^n X_i$, where the $\{X_1, X_2, \dots, X_n\} \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, $\mathbb{E}(X_i) = \theta$ and $\Delta(X_i) = \theta(1-\theta)$:

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{E}(X_i) = n\theta$$

Example 80 (Mean and variance of Poisson(λ) RV) Let $X \sim \text{Poisson}(\lambda)$. Then:

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

Similarly,

Let's first find the CF of X

$$\phi_X(t) = \mathbb{E}(\exp(itX)) = \sum_x \exp(itx) f(x; \theta) \quad \text{By Defn. in Equation (3.68)}$$

Recall the PMF for this discrete RV with parameter $\theta \in (0, 1)$ is

$$\mathbb{P}(X=x) = \begin{cases} \theta & \text{if } x = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Part 1 Solution:

Example 101 Let X be the Bernoulli(θ) RV. Find the CF of X . Then use CF to find $\mathbb{E}(X)$, $\mathbb{E}(X^2)$ and from this obtain the variance $V(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.

Proof: For proof see e.g., Ushakov, N. G. (1999) Selected topics in characteristic functions, VSF p. 39.

Example 81 Let X be the Bernoulli(θ) RV. Find the CF of X . Then use CF to find $\mathbb{E}(X)$, $\mathbb{E}(X^2)$ and from this obtain the variance $V(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.

For proof see e.g., Ushakov, N. G. (1999) Selected topics in characteristic functions, VSF p. 39.

Example 82 Let X be the Geometric(p) RV. Find the CF of X .

Example 83 Let X be the Geometric(p) RV. Find the mean and variance of X .

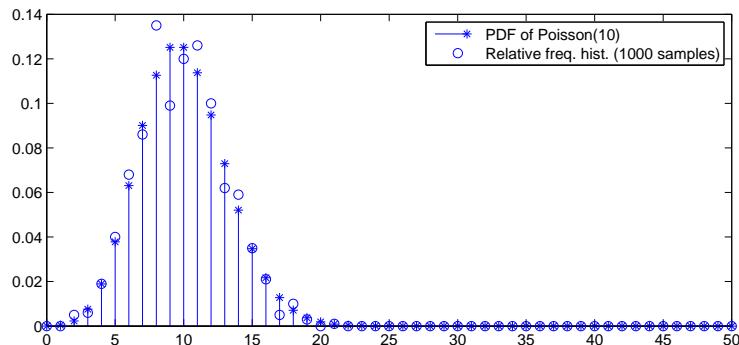
Example 84 (Moments from CF). Let X be a random variable and $\phi_X(t)$ be its CF. If

since

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

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

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 Poisson(λ) RV X is also related to the IID Exponential(λ) RV Y_1, Y_2, \dots : X is the number of occurrences, per unit time, of an instantaneous event whose inter-occurrence time is the IID Exponential(λ) RV. For example, the number of buses arriving at our bus-stop in the next minute, with exponentially distributed inter-arrival times, has a Poisson distribution.

Example 81 (Mean and variance of Normal(μ, σ^2) RV) The location-scale family of RVs is indeed parameterised by its mean and variance, i.e., if $X \sim \text{Normal}(\mu, \sigma^2)$ where $X = g(Z) = \sigma Z + \mu$ and $Z \sim \text{Normal}(0, 1)$ then $\mathbf{E}(X) = \mu$ and $\mathbf{V}(X) = \sigma^2$ follows directly from the properties of Expectations, provided $\mathbf{E}(Z) = 0$ and $\mathbf{V}(Z) = \mathbf{E}(Z^2) - (\mathbf{E}(Z))^2 = \mathbf{E}(Z^2) = 1$.

The mean of a Normal(0, 1) RV Z is:

$$\mathbf{E}(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \exp\left(-\frac{1}{2}z^2\right) dz = \frac{1}{\sqrt{2\pi}} \left[-\exp\left(-\frac{1}{2}z^2\right) \right]_{-\infty}^{\infty} = 0,$$

and the variance is:

$$\mathbf{V}(Z) = \mathbf{E}(Z^2) - (\mathbf{E}(Z))^2 = \mathbf{E}(Z^2) - 0 = \mathbf{E}(Z^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz.$$

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 46 (Characteristic Function (CF)) Let X be a RV and $i = \sqrt{-1}$. The function $\phi_X(t) : \mathbb{R} \rightarrow \mathbb{C}$ defined by

$$\boxed{\phi_X(t) := E(\exp(itX)) = \begin{cases} \sum_x \exp(itx) f_X(x) & \text{if } X \text{ is discrete RV} \\ \int_{-\infty}^{\infty} \exp(itx) f_X(x) dx & \text{if } X \text{ is continuous RV} \end{cases}} \quad (3.68)$$

is called the **characteristic function** of X .

NOTE: $\phi_X(t)$ exists for any $t \in \mathbb{R}$, because

$$\begin{aligned}\phi_X(t) &= E(\exp(itX)) \\ &= E(\cos(tx) + i \sin(tx)) \\ &= E(\cos(tx)) + i E(\sin(tx))\end{aligned}$$

and the last two expected values are well-defined, because the sine and cosine functions are bounded by $[-1, 1]$.

For a continuous RV, $\int_{-\infty}^{\infty} \exp(-itx) f_X(x) dx$ 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

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

The characteristic function can be used to derive the moments of X due to the following nice relationship between the k -th moment of X and the k -th derivative of the CF of X .

Proposition 47 (Moment & CF.) Let X be a random variable and $\phi_X(t)$ be its CF. If $E(X^k)$ exists and is finite, then $\phi_X(t)$ is k times continuously differentiable and

$$E(X^k) = \frac{1}{i^k} \left[\frac{d^k \phi_X(t)}{dt^k} \right]_{t=0}.$$

where $\left[\frac{d^k \phi_X(t)}{dt^k} \right]_{t=0}$ is the k -th derivative of $\phi_X(t)$ with respect to t , evaluated at the point $t = 0$.

$$\left. \begin{array}{l} \exists x \in J \\ \exists y \in I \end{array} \right\} = f(x, y)$$

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

$$\{ \text{I} = \theta \sum_{\theta} \text{I}_{\theta} : (\forall \theta) \theta_1, \theta_2, \dots, \theta_n \in \mathbb{R} \} =: \nabla_{\theta_1, \theta_2, \dots, \theta_n}$$

Model 14 (de Movie($\theta_1, \theta_2, \dots, \theta_k$)) Given a specific point ($\theta_1, \theta_2, \dots, \theta_k$) in the unit $k - 1$ -Simplex:

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

Note that the construction is valid even if we sample X uniformly from $(0, \alpha)$ and take its $\tan(X)$. Example 82 (Mean of Cauchy RV) The expectation of the Cauchy RV X , obtained via infinite parts (set $u = x$ and $v = \tan^{-1}(x)$) does not exist, since:

$$\left| \frac{\tilde{e}^{\tilde{h}} + 1}{1} \right| \frac{u}{1} = \left| (\tilde{h})_{1-u+1} \frac{fp}{p} \right| ((\tilde{h})_{1-u+1}) X f = \left| (\tilde{h})_{1-\tilde{b}} \frac{fp}{p} \right| ((\tilde{h})_{1-\tilde{b}}) X f = (\tilde{h}) X f$$

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 3.36 to obtain the PDF $f_Y(y)$ from the PDF $f_X(x) = \frac{1}{\pi} \mathbb{I}_{(-\pi/2, \pi/2)}(x)$ as follows:

Randomly spinning a LASER emitting improvement of "Darth Mall's double edge 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.

$$(3.51) \quad \frac{u(1+y_2)}{1} = y_1 > \infty > y_2 > -\infty,$$

and its DF is:

Next, let us become familiar with an RV for which the expectation does not exist.

than z grows to $\pm\infty$. The second term equals 1 because it is exactly the total probability integrated of the PDF of the $Normal(0, 1)$ RV.

$$\int_{-\infty}^{\infty} e^{-z^2/2} dz = \sqrt{\frac{2\pi}{1}} = \sqrt{2\pi}$$

Using integration by parts with $u = np$, $v = \int z^{-\theta} dz = z^{-\theta+1}/(-\theta+1)$

CHAPTER 3. RANDOM VARIABLES

111

classification in this digital channel.]

Ex. 3.47 — Soft drink cans are filled by an automated filling machine. Assume the fill volumes of the cans are independent Normal($12.1, 0.01$) RVs. What is the probability that the average volume of ten cans selected from this process is less than 12.01 fluid ounces?

Ex. 3.48 — 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. In a transmission of 10 bits, what is the probability that at least 6 of the bits received are excellent, 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 $0.3, 0.6, 0.3, 0.08$, respectively.

Ex. 3.46 Suppose the RVs Y_1 , Y_2 and Y_3 represent the thicknesses 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 normally distributed. The RVs Y_1 , Y_2 and Y_3 have the following properties:
 $E(Y_1) = 1000$, $\text{Var}(Y_1) = 250^2$, $E(Y_2) = 80$, $\text{Var}(Y_2) = 40^2$, $E(Y_3) = 20$, $\text{Var}(Y_3) = 30^2$. What is the probability that the total thickness of the three layers is less than 1200 micrometers?

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.]

$$f_{X_1^*, X_2^*, X_3^*, X_4^*}(x_1^*, x_2^*, x_3^*, x_4^*) = \begin{cases} 0 & \text{otherwise} \\ 9 \times 10^{-12} e^{-0.001x_1 - 0.002x_2 - 0.0015x_3 - 0.003x_4} & \text{if } x_1^* \geq 0, x_2^* \geq 0, x_3^* \geq 0, x_4^* \geq 0 \end{cases}$$

Ex. 3.45 — 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 PDF of these variables is

Are X and Y independent?

$$\left\{ \begin{array}{ll} 0 & \text{otherwise} \\ e^{-x} & \text{if } x \in [0, \infty) \text{ and } y \in [2, 3] \end{array} \right\} = (f(y, x) X^y X^x f$$

Ex. 3.44 — Let (X, Y) be a continuous RV with joint probability density function (jPDF):

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

$$F(x; \theta_1, \theta_2, \dots, \theta_k) = \begin{cases} 0 & \text{if } -\infty < x < 1 \\ \theta_1 & \text{if } 1 \leq x < 2 \\ \theta_1 + \theta_2 & \text{if } 2 \leq x < 3 \\ \vdots & \\ \theta_1 + \theta_2 + \dots + \theta_{k-1} & \text{if } k-1 \leq x < k \\ \theta_1 + \theta_2 + \dots + \theta_{k-1} + \theta_k = 1 & \text{if } k \leq x < \infty \end{cases} \quad (3.54)$$

The de Moivre($\theta_1, \theta_2, \dots, \theta_k$) RV can be thought of as a probability model for ‘‘the outcome of rolling a polygonal cylindrical die with k rectangular faces that are marked with $1, 2, \dots, k$ ’’. The parameters $\theta_1, \theta_2, \dots, \theta_k$ specify how the die is loaded and may be idealised as specifying the cylinder’s centre of mass with respect to the respective faces. Thus, when $\theta_1 = \theta_2 = \dots = \theta_k = 1/k$, we have a probability model for the outcomes of a fair die.

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

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

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

However, if $X \sim \text{de Moivre}(1/k, 1/k, \dots, 1/k)$, then the mean and variance for the fair k -faced die based on Faulhaber’s formula for $\sum_{i=1}^k i^m$, with $m \in \{1, 2\}$, are,

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

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

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

3.9 Exercises in Expectations of Random Variables

Ex. 3.29 — 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. 3.30 — 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 .

7. $E(X)$, the expectation of X or the first moment of X
8. $E(Y)$, the expectation of Y or the first moment of Y
9. $E(XY)$, the expectation of XY
10. $\mathbf{Cov}(X, Y) = E(XY) - E(X)E(Y)$, the covariance of X and Y .

Ex. 3.39 — Logs are milled to have a width of μ . The actual width of a randomly selected item is X . If X is a Normal(μ, σ^2) random variable then find the probability density function of the squared-error of the milling process,

$$Y = (X - \mu)^2.$$

Ex. 3.40 — 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. 3.41 — 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. 3.42 — 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 $\mathbf{Cov}(X, Y) = E(XY) - E(X)E(Y)$]

Ex. 3.43 — 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 $(x,y) \in \mathbb{R}^2$.]

More generally, we may be interested in heights, weights, blood-sugar levels, family medical history, known alleles, etc. of individuals in the clinical trial and thus need to make m measurements of known variables, we may be interested in the height, weight, blood-sugar levels, family medical history, variables (X_1, X_2, \dots, X_m), ordered triples of random variables (X, Y, Z), or more generally ordered m -tuples of random variables (X, Y), ordered pairs of random variables (X, Z), or more generally ordered m -tuples of random measurements we need the notion of **random vectors** (RVs), i.e. ordered pairs of random variables (X, Y), measurable mappings from $\Omega \rightarrow \mathbb{R}^m$. To deal with such multivariate outcomes in \mathbb{R}^m using a "measurable mapping" from $\Omega \rightarrow \mathbb{R}^m$.

$$\varepsilon \mathbb{M} \leftarrow \mathcal{U} : ((\sigma)Z `(\sigma)A `(\sigma)X) \leftrightarrow \sigma \quad \quad \quad \varepsilon \mathbb{M} \leftarrow \mathcal{U} : ((\sigma)A `(\sigma)X) \leftrightarrow \sigma$$

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 $w \in \mathcal{U}$ 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 $\mathbb{R}^3 = (-\infty, \infty) \times (-\infty, \infty) \times (-\infty, \infty)$:

3.10 Multivariate Random Variables

c. 3.34 — Find the mean and the variance of the following random variables.

1. X is discrete uniform random variable on $\{1, 2, 3, 4, 5, 6\}$, i.e., the number a fair die turns up.
2. X is $\text{Uniform}(0, 8)$ random variable, i.e., a continuous uniform random variable from the interval $[0, 8]$.
3. X has a density function

$$f(x) = \begin{cases} 0 & \text{otherwise} \\ 2e^{-2x} & \text{if } x \geq 0 \end{cases}$$

(a) Find f and
 (i) $P(X = 0)$
 (ii) $P(2.5 < X < 5)$
 (iii) $E(X)$
 (iv) $A(X)$

(b) Write down the DF (or CDF) of X .
 (c) Plot the PMF and CDF of X .

$$f(x) = \begin{cases} 0 & \text{otherwise.} \\ x & \text{if } x \in \{1, 2, 3, 4\}, \end{cases}$$

Ex. 3.33 — Let X be a discrete random variable with PMF given by

Ex. 3.32 — Show that $V(ax + b) = a^2V(x)$ for constants a and b and a random variable X .

$$\cdot \quad \zeta((X)\mathbf{E}) - (\zeta X)\mathbf{E} = (X)\mathbf{A}$$

Ex. 3.31 — Starting from the definition of the variance of a random variable (Definition 29) show that

CHAPTER 3. RANDOM VARIABLES

1. the normalization constant a which will ensure $\mathbf{P}(A) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)$ and the following:
2. $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy$ called the marginal probability density function.
3. $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx$ called the marginal probability density function.
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 said to be independent if $f_X(x)f_Y(y)$.
5. $F_{X,Y}(x,y)$, the joint cumulative distribution function (CDF) of (X,Y) is given by
6. the probability that $X < 0.5$ and $Y > 0.6$, i.e., $\mathbf{P}(X < 0.5, Y > 0.6)$

Ex. 3.38 — Let (X, Y) be a continuous RV with joint probability density function (jPDF)

$$9 \left(0.25 - (x - 1.5)^2 \right) \quad 1 > x > 2 \quad \text{otherwise} \quad 0 \Big\} = (x)f$$

Ex. 3.37 — Find the probability that none of the three bubbles 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

3.11 Exercises in Multivariate Random Variables

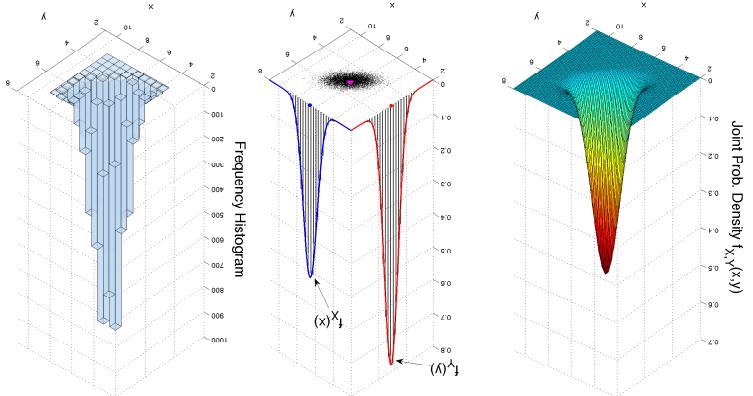


Figure 3.23: JPDF, Marginal PDFs and Frequency Histogram of a Bivariate Normal RV for lengths of grits of cylindrical shafts in a manufacturing process (in cm).

3.10.1 \mathbb{R}^2 -valued Random Variables

We first focus on understanding (X, Y) , a bivariate \vec{RV} 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 33 (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

$$\begin{aligned} F_{X,Y}(x, y) &= \mathbf{P}(X \leq x \cap Y \leq y) = \mathbf{P}(X \leq x, Y \leq y) \\ &= P(\{\omega : X(\omega) \leq x, Y(\omega) \leq y\}), \text{ for any } (x, y) \in \mathbb{R}^2, \end{aligned} \quad (3.55)$$

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 34 (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 $p_{i,j} = \mathbf{P}(X = x_i, Y = y_j) > 0$, then its joint probability mass function (or JPMF) is:

$$f_{X,Y}(x_i, y_j) = \mathbf{P}(X = x_i, Y = y_j) = \begin{cases} p_{i,j} & \text{if } (x_i, y_j) \in \mathcal{S}_{X,Y} \\ 0 & \text{otherwise} \end{cases}. \quad (3.56)$$

Since $\mathbf{P}(\Omega) = 1$, $\sum_{(x_i, y_j) \in \mathcal{S}_{X,Y}} f_{X,Y}(x_i, y_j) = 1$.

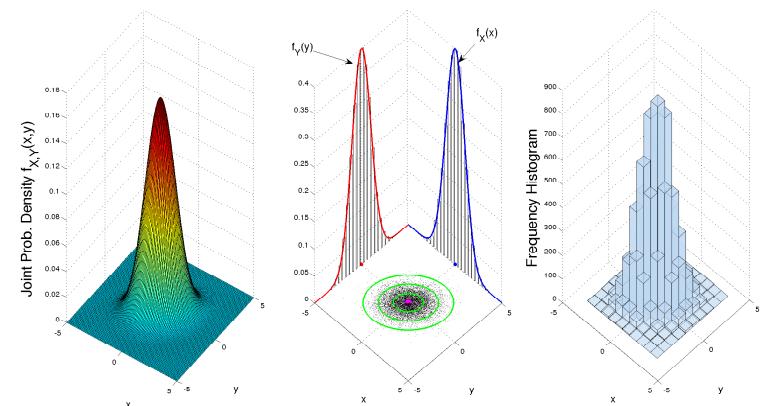
From JPMF $f_{X,Y}$ we can get the values of the JDF $F_{X,Y}(x, y)$ and the probability of any event B by simply taking sums,

$$F_{X,Y}(x, y) = \sum_{x_i \leq x, y_j \leq y} f_{X,Y}(x_i, y_j), \quad \mathbf{P}(B) = \sum_{(x_i, y_j) \in B \cap \mathcal{S}_{X,Y}} f_{X,Y}(x_i, y_j), \quad (3.57)$$

Example 83 Let (X, Y) be a discrete bivariate \vec{RV} with the following joint probability mass function (JPMF):

$$f_{X,Y}(x, y) := P(X = x, Y = y) = \begin{cases} 0.1 & \text{if } (x, y) = (0, 0) \\ 0.3 & \text{if } (x, y) = (0, 1) \\ 0.2 & \text{if } (x, y) = (1, 0) \\ 0.4 & \text{if } (x, y) = (1, 1) \\ 0.0 & \text{otherwise.} \end{cases}$$

Figure 3.22: JPDF, Marginal PDFs and Frequency Histogram of Bivariate Standard Normal \vec{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

$$\Sigma = \begin{pmatrix} \text{Cov}(X, X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Cov}(Y, Y) \end{pmatrix} = \begin{pmatrix} V(X) & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & V(Y) \end{pmatrix} = \begin{pmatrix} 0.59 & 0.24 \\ 0.24 & 0.26 \end{pmatrix}$$

then the Normal(μ, Σ) \vec{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 ± 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**.

$$P(B) = \int \int_B f_{X,Y}(x,y) dx dy \quad (3.59)$$

and

$$F_{X,Y}(x,y) = \int_x^{\infty} \int_y^{\infty} f_{X,Y}(u,v) du dv \quad (3.58)$$

two-dimensional integrals:
From joint PDF $f_{X,Y}$ 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 expectations:

For notational convenience, we sometimes suppress the subscripts 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 $F_{X,Y}(x,y)$.

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

Definition 35 (JPDF) We say (X,Y) is a continuous \mathbb{R}^2 -valued random variable if its JDF $F_{X,Y}(x,y)$ is differentiable and its joint probability density function (JPDF) is given by:

$$5. F_{X,Y}(-4,-1) = \sum_{\{(x,y): x < -4, y < -1\}} f_{X,Y}(x,y) = 1$$

$$4. F_{X,Y}(4,5) = \sum_{\{(x,y): x < 4, y < 5\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) + f_{X,Y}(1,0) + f_{X,Y}(1,1) = 1$$

$$3. F_{X,Y}(3/2,1/2) = \sum_{\{(x,y): x < 3/2, y < 1/2\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) + f_{X,Y}(1,0) = 0.1 + 0.2 = 0.3$$

$$2. F_{X,Y}(1/2,1/2) = \sum_{\{(x,y): x < 1/2, y < 1/2\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) = 0.1$$

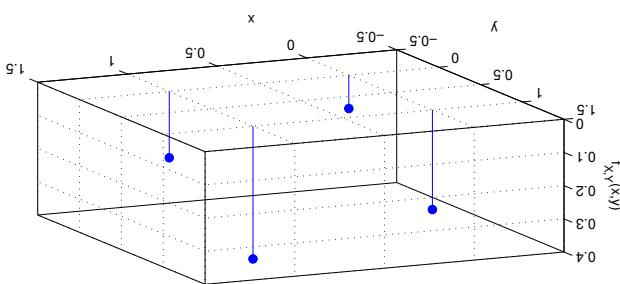
$$1. P(B) = \sum_{\{(x,y): (0,0), (1,1)\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) + f_{X,Y}(1,1) = 0.1 + 0.4$$

Find $P(B)$ for the event $B = \{(0,0), (1,1), F_{X,Y}(1/2,1/2), F_{X,Y}(3/2,1/2), F_{X,Y}(4,5) \text{ and } F_{X,Y}(-4,-1)\}$.

From the above Table we can read for instance that the joint probability $f_{X,Y}(0,0) = 0.1$.

	$X = 0$	$X = 1$
$Y = 0$	0.1	0.2

It is helpful to write down the JPDF $f_{X,Y}(x,y)$ in a tabular form:



positive definite matrix. Setting $\mu = 0$ and $\Sigma = I$ gives back the standard multivariate normal RV.
where $|\Sigma|$ denotes the determinant of Σ , μ is a vector of length m and Σ is a $m \times m$ symmetric matrix.

$$f_X(x; \mu, \Sigma) = f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m; \mu, \Sigma) = \frac{(2\pi)^{m/2}}{\prod_{i=1}^m \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

if it has joint probability density function

More generally, a vector X has a multivariate normal distribution denoted by $X \sim \text{Normal}(\mu, \Sigma)$,

1 along the diagonal entries and 0 on all off-diagonal entries.

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

$$f_Z(z) = f_{Z_1, Z_2, \dots, Z_m}(z_1, z_2, \dots, z_m) = \frac{(2\pi)^{m/2}}{\prod_{i=1}^m \sqrt{\det(z_i)}} \exp\left(-\frac{1}{2} \sum_{i=1}^m z_i^T z_i\right) = \frac{(2\pi)^{m/2}}{\prod_{i=1}^m z_i} \exp\left(-\frac{1}{2} z^T z\right)$$

where, Z_1, Z_2, \dots, Z_m are jointly independent $\text{Normal}(0, 1)$ RVs. Then the JPDF of Z is

$$\begin{pmatrix} Z_m \\ \vdots \\ Z_2 \\ Z_1 \end{pmatrix} = Z$$

matrix Σ . To begin, let
matrix Σ . In the multivariate version $\mu \in \mathbb{R}^{m \times 1}$ is a column vector and σ^2 is replaced by a
 $\sigma^2 \in (0, \infty)$. In the multivariate $\text{Normal}(\mu, \sigma^2)$ RV has two parameters, $\mu \in \mathbb{R}$ and

Model 18 (Normal(μ, Σ) RV) The univariate $\text{Normal}(\mu, \sigma^2)$ RV is a large knowledge-based resources like www, Wikipedia,
Multinomial distributions are at the very foundations of various machine learning algorithms, in-
cluding, fitting junk mail, learning from large knowledge-based resources like www, Wikipedia,
word-net, etc.

Multinomial distributions are at the very foundations of various machine learning algorithms, in-
cluding, fitting junk mail, learning from large knowledge-based resources like www, Wikipedia,
etc.

>>> multinomial

tool as follows:

Let us understand the *Septcunx* construction of the $\text{Molivre}(n, 1/3, 1/3, 1/3)$ RVs by calling the interactive visual cognitive
independent and identical $\text{Molivre}(1/3, 1/3, 1/3)$ RVs by calling the sum of n
Multinomial($n, \theta_1, \theta_2, \theta_3$) RV.

Note again, we can visualize that the sum of n IID $\text{Molivre}(\theta_1, \theta_2, \theta_3)$ RVs constitute the
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
coefficients $\frac{n!}{y_1! y_2! y_3!}$. In the *Septcunx*, balls choose from one of three paths along ϵ_1, ϵ_2 and ϵ_3 with
probabilities θ_1, θ_2 and θ_3 , respectively, in an IID manner at each of the n levels, before they
collect at blocks placed at the integral points in the 3-simplex, $\Sigma^3 = \{(y_1, y_2, y_3) \in \mathbb{Z}_+^3 : \sum_{i=1}^3 y_i = n\}$. Once again, we can visualize that the sum of n IID $\text{Molivre}(\theta_1, \theta_2, \theta_3)$ RVs constitute the
Multinomial($n, \theta_1, \theta_2, \theta_3$) RV.

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

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

where, the multinomial coefficients:

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 $\delta \rightarrow 0$:

$$\mathbf{P}((X, Y) \in \mathbb{B}_\delta(x, y)) \cong \delta f_{X,Y}(x, y). \quad (3.60)$$

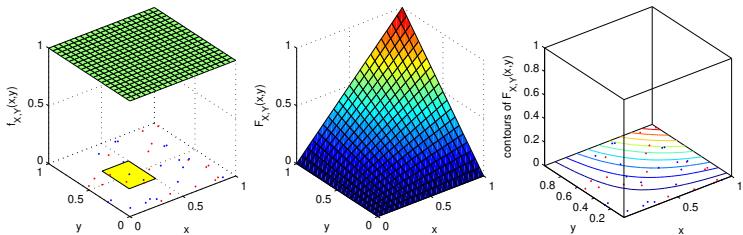
The JPDF satisfies the following two properties:

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 R \vec{V} that is uniformly distributed on the unit square $[0, 1]^2 := [0, 1] \times [0, 1]$ with following JPDF:

$$f(x, y) = \mathbf{1}_{[0,1]^2}(x) \begin{cases} 1 & \text{if } (x, y) \in [0, 1]^2 \\ 0 & \text{otherwise.} \end{cases}$$

Find explicit expressions for the following: (1) DF $F(x, y)$ for any $(x, y) \in [0, 1]^2$, (2) $\mathbf{P}(X \leq 1/3, Y \leq 1/2)$, (3) $P((X, Y) \in [1/4, 1/2] \times [1/3, 2/3])$.



Let us begin to find the needed expressions.

1. Let $(x, y) \in [0, 1]^2$ then by Equation (3.58):

$$F_{X,Y}(x, y) = \int_{-\infty}^y \int_{-\infty}^x f_{X,Y}(u, v) du dv = \int_0^y \int_0^x 1 du dv = \int_0^y [u]_{u=0}^x dv = \int_0^y x dv = [xv]_{v=0}^y = xy$$

2. We can obtain $\mathbf{P}(X \leq 1/3, Y \leq 1/2)$ by evaluating $F_{X,Y}$ at $(1/3, 1/2)$:

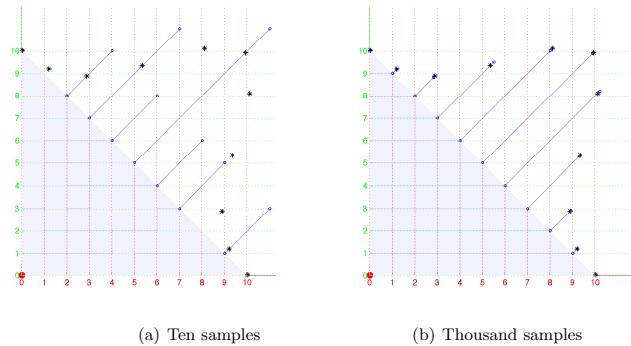
$$\mathbf{P}(X \leq 1/3, Y \leq 1/2) = F_{X,Y}(1/3, 1/2) = \frac{1}{3} \frac{1}{2} = \frac{1}{6}$$

We can also find $\mathbf{P}(X \leq 1/3, Y \leq 1/2)$ by integrating the JPDF over the rectangular event $A = \{X < 1/3, Y < 1/2\} \subset [0, 1]^2$ according to Equation (3.59). This amounts here to finding the area of A , we compute $\mathbf{P}(A) = (1/3)(1/2) = 1/6$.

3. We can find $P((X, Y) \in [1/4, 1/2] \times [1/3, 2/3])$ by integrating the JPDF over the rectangular event $B = [1/4, 1/2] \times [1/3, 2/3]$ according to Equation (3.59):

$$\begin{aligned} P((X, Y) \in [1/4, 1/2] \times [1/3, 2/3]) &= \int \int_B f_{X,Y}(x, y) dx dy = \int_{1/3}^{2/3} \int_{1/4}^{1/2} 1 dx dy \\ &= \int_{1/3}^{2/3} [x]_{1/4}^{1/2} dy = \int_{1/3}^{2/3} \left[\frac{1}{2} - \frac{1}{4} \right] dy = \left(\frac{1}{2} - \frac{1}{4} \right) [y]_{1/3}^{2/3} \\ &= \left(\frac{1}{2} - \frac{1}{4} \right) \left(\frac{2}{3} - \frac{1}{3} \right) = \frac{1}{4} \left(\frac{1}{3} \right) = \frac{1}{12} \end{aligned}$$

Figure 3.21: Quincunx on the Cartesian plane. Simulations of Binomial($n = 10, \theta = 0.5$) RV as the x-coordinate of the ordered pair resulting from the culmination of sample trajectories formed by the accumulating sum of $n = 10$ IID Bernoulli($\theta = 0.5$) random vectors over $\{(1, 0), (0, 1)\}$ with probabilities $\{\theta, 1 - \theta\}$, respectively. The blue lines and black asterisks perpendicular to and above the diagonal line, i.e. the line connecting $(0, 10)$ and $(10, 0)$, are the density histogram of the samples and the PMF of our Binomial($n = 10, \theta = 0.5$) RV, respectively.



Model 16 (de Moivre($\theta_1, \theta_2, \dots, \theta_k$) R \vec{V}) The PMF of the de Moivre($\theta_1, \theta_2, \dots, \theta_k$) R \vec{V} $X := (X_1, X_2, \dots, X_k)$ 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($\theta_1, \theta_2, \dots, \theta_k$) R \vec{V} together, we get the Multinomial($n, \theta_1, \theta_2, \dots, \theta_k$) R \vec{V} as defined below.

Model 17 (Multinomial($n, \theta_1, \theta_2, \dots, \theta_k$) R \vec{V}) We say that a R \vec{V} $Y := (Y_1, Y_2, \dots, Y_k)$ obtained from the sum of n IID de Moivre($\theta_1, \theta_2, \dots, \theta_k$) R \vec{V} 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:

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

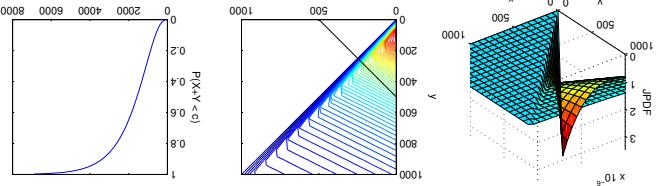
Example 85 Let the RV X denote the time until a web server connects to your computer, and let the RV Y denote the time until the server authorizes you as a valid user. Both of these RVs measure the waiting time from a common starting time (in milliseconds) and $X < Y$. From previous times of the web server we know that a good approximation for the PDF of the RV (X, Y) is the bivariate uniform RV on the unit square $[0, 1] \times [0, 1]$. Thus any two events within 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...).

1. Identify the support of (X, Y) , i.e., the region in the plane where $f_{X,Y}$ takes positive values

2. check that $\int_X f_{X,Y}$ 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 $P(X + Y < 10^2)$?



1. The support is the intersection of the positive quadrant with the $y > x$ half-plane.

Let us answer the questions.

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

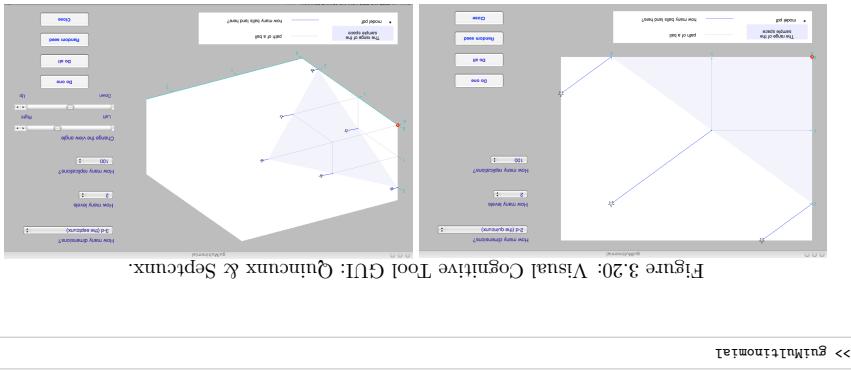


Figure 3.20: Visual Cognitive Tool GUI: Quincunx & Specifix.

Table 99 (Quincunx Sampler Demo – Sum of n IID Bernoulli(1/2) RVs) Let us understand the Quincunx construction of the Binomial($n, 1/2$) RV as the sum of n independent and identically Bernoulli(1/2) RVs by calling the interactive visual cognitive tool as follows:

Exercise 3.36 (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.

2. What is the probability of taking x_1 steps east and x_2 steps north?

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 $\binom{x_1}{n}$?

Exercise 3.35 (Random walk in the first Quadrant) Consider an independent and identical random walk starting from $(0, 0)$ in the first quadrant where you go east, i.e., add $(1, 0)$ to your current position with probability θ , and go north, i.e., add $(0, 1)$ to your current position with probability $1 - \theta$. Suppose you take n such IID steps according to the Bernoulli(θ) RV. Answer the following questions:

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

RVs $X_1 = (X_{1,1}, X_{1,2}), X_2 = (X_{2,1}, X_{2,2}), \dots, X_n = (X_{n,1}, X_{n,2})$. In other words, the Binomial(n, θ) RV $(Y, n - Y)$ is the sum of n IID Bernoulli(θ) drops through n levels of pegs where the probability of a right turn at each peg is independent and identically θ . Thus any two events within the same rectangle 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...).

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 $(X \leq 400, Y \leq 800)$

$$\begin{aligned}
\mathbf{P}(X \leq 400, Y \leq 800) &= \int_{x=0}^{400} \int_{y=x}^{800} f_{X,Y}(x,y) dy dx \\
&= \int_{x=0}^{400} \int_{y=x}^{800} \frac{6}{10^6} \exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right) dy dx \\
&= \frac{6}{10^6} \int_{x=0}^{400} \left[-\frac{1000}{2} \exp\left(-\frac{2}{1000}y\right) \right]_{y=x}^{800} \exp\left(-\frac{1}{1000}x\right) dx \\
&= \frac{6}{10^6} \frac{1000}{2} \int_{x=0}^{400} \left(-\exp\left(-\frac{1600}{1000}x\right) + \exp\left(-\frac{2}{1000}x\right) \right) \exp\left(-\frac{1}{1000}x\right) dx \\
&= \frac{6}{10^6} \frac{1000}{2} \int_{x=0}^{400} \left(\exp\left(-\frac{3}{1000}x\right) - e^{-8/5} \exp\left(-\frac{1}{1000}x\right) \right) dx \\
&= \frac{6}{10^6} \frac{1000}{2} \left(\left(-\frac{1000}{3} \exp\left(-\frac{3}{1000}x\right) \right)_{x=0}^{400} - e^{-8/5} \left(-1000 \exp\left(-\frac{1}{1000}x\right) \right)_{x=0}^{400} \right) \\
&= \frac{6}{10^6} \frac{1000}{2} 1000 \left(\frac{1}{3} \left(1 - e^{-6/5} \right) - e^{-8/5} \left(1 - e^{-2/5} \right) \right) \\
&= 3 \left(\frac{1}{3} \left(1 - e^{-6/5} \right) - e^{-8/5} \left(1 - e^{-2/5} \right) \right) \\
&\approx 0.499 .
\end{aligned}$$

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*Solution:*

1. addition is component-wise

$$(1,0) + (1,0) = (1+1,0+0) = (2,0)$$

$$(1,0) + (0,1) = (1+0,0+1) = (1,1)$$

$$(0,1) + (0,1) = (0+0,1+1) = (0,2)$$

2. $(1,0)$ and $(0,1)$ are vectors for the two sides of unit square and $(1,1)$ is its diagonal.

3. Generally, the diagonal of the parallelogram is the resultant or sum of the vectors representing its two sides

4.

$$(1,0) + (0,1) + (1,0) = (1+0+1,0+1+0) = (2,1)$$

Model 15 (Bernoulli(θ) 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(θ) random vector (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 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(θ) RV in Model 15.

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

Remark 45 We can write the Binomial(n, θ) RV Y as a Binomial(n, θ) RV $X := (Y, n - Y)$. In fact, this is the underlying model and the **bi** in the Binomial(n, θ) does refer to two in Latin. In the coin-tossing context this can be thought of keeping track of the number of Heads and Tails out of an IID sequence of n tosses of a coin with probability θ of observing Heads. In the Quincunx context, this amounts to keeping track of the number of right and left turns made by the ball as it

3.10.4 Some Common \mathbb{R}^m -valued RVs

so far, we have treated our column vectors as random points in \mathbb{R}^n . And how exactly do we whether they are a row or column vectors. We need to be more explicit now in order to perform arithmetic operations and transformations with them.

Let $X = (X_1, X_2, \dots, X_m)$ be a RV in \mathbb{R}^m , i.e., X is a random row vector with 1 row and m columns, with PDF f_{X_1, X_2, \dots, X_m} and CDF F_{X_1, X_2, \dots, X_m} . Similarly, let $Y = (Y_1, Y_2, \dots, Y_m)$ be a RV in \mathbb{R}^m , i.e., Y is a random row vector with 1 row and m columns, with PDF f_{Y_1, Y_2, \dots, Y_m} and CDF F_{Y_1, Y_2, \dots, Y_m} . Let the CDF of the random vectors X and Y together be $F_{X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_m}$ and PDF $f_{X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_m}$. Let the joint CDF of the random vectors X and Y together be $F_{X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_m}$ and joint PDF $f_{X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_m}$.

PDFs) 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

$$\begin{aligned} & \left({}^{A_m} \mu_B^A \right)_{i_1, i_2, \dots, i_m} \times \left({}^{A_m} X^A f \right)_{j_1, j_2, \dots, j_m} = \\ & \quad \left({}^{A_m} \mu_B^A \right)_{i_1, i_2, \dots, i_m} \times \left({}^{A_m} X^A X^{i_1, i_2, \dots, i_m} f \right)_{j_1, j_2, \dots, j_m} = \\ & \quad F^{X^{i_1, i_2, \dots, i_m} X^A} f_{j_1, j_2, \dots, j_m} = \\ & \quad F^{X^{i_1, i_2, \dots, i_m} X^A} f_{j_1, j_2, \dots, j_m} = \end{aligned}$$

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

3. How does the diagonal of the parallelogram relate to its two sides in the geometry of addition in the plane?

2. What is the relationship between $(1, 0)$, $(0, 1)$ and $(1, 1)$ geometrically?

1. What is $(1, 0) + (1, 0)$, $(1, 0) + (0, 1)$, $(0, 1) + (0, 1)$?

Example 9: Let us recall the geometry and arithmetic of vector addition in

$\mathbb{I}_A(x)$ returns 1 if x belongs to A and 0 otherwise.

$$\mathbb{A}^A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

Recall that vector addition and subtraction are done component-wise, i.e. $(x_1, x_2) \pm (y_1, y_2) = (x_1 \mp y_1, x_2 \mp y_2)$. We introduce a useful function called the indicator function of a set, say A ,

```

    <>> c = [100 1000 2000 3000 4000]
    <>> c = 100 1000 2000 3000 4000
    c = 100 1000 2000 3000 4000
    2000 3000 4000
    3000 4000
    4000
    0.9911
    0.9630
    0.8558
    0.5135
    0.0134
    p = 0.0134

```

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.

The JPDF over our triangular event as follows:

$X = 0$	0.1	0.3	0.2	0.4
$X = 1$	0.0	Y	Z	W

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 RV in Example 83. Just sum $f_{X,Y}(x,y)$ over x 's and y 's (reported in a tabular form):

$$\left\{ \int_{-\infty}^x f(X,Y)(x,y) dx \quad \text{if } (X,Y) \text{ is a continuous RV} \right. \\ \left. \int_x^{\infty} f(X,Y)(x,y) dx \quad \text{if } (X,Y) \text{ is a discrete RV} \right\} = f_Y(y)$$

and the marginal PDF or PMF of y is defined by:

$$\left\{ \int_{-\infty}^y f_{X,Y}(x,y) dy \quad \text{if } (X,Y) \text{ is a continuous RV} \right. \\ \left. \sum f_{X,Y}(x,y) \quad \text{if } (X,Y) \text{ is a discrete RV} \right\} = (x)_{X,Y}$$

Joint PMF, then the marginal PDF or PMF of a random vector (X, Y) is defined by :

4. What is $(1, 0) + (0, 1) + (1, 0)$?

in the Plane?

3. How does the diagonal of the parallelogram relate to the two sides in the geometry of addition

From the above Table we can find:

$$\begin{aligned} f_X(x) &= \mathbf{P}(X = x) = \sum_y f_{X,Y}(x,y) \\ &= f_{X,Y}(x,0) + f_{X,Y}(x,1) = \begin{cases} f_{X,Y}(0,0) + f_{X,Y}(0,1) = 0.1 + 0.3 = 0.4 & \text{if } x = 0 \\ f_{X,Y}(1,0) + f_{X,Y}(1,1) = 0.2 + 0.4 = 0.6 & \text{if } x = 1 \end{cases} \end{aligned}$$

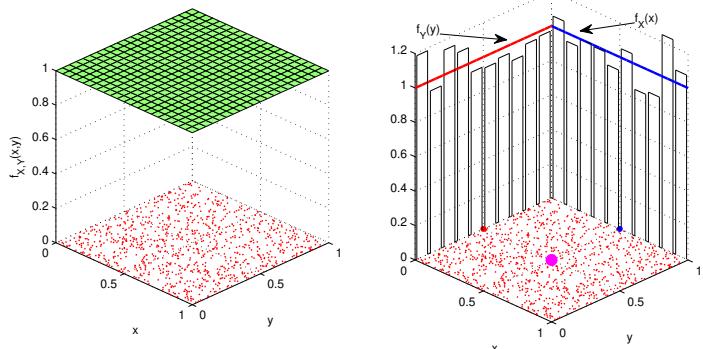
Similarly,

$$\begin{aligned} f_Y(y) &= \mathbf{P}(Y = y) = \sum_x f_{X,Y}(x,y) \\ &= f_{X,Y}(0,y) + f_{X,Y}(1,y) = \begin{cases} f_{X,Y}(0,0) + f_{X,Y}(1,0) = 0.1 + 0.2 = 0.3 & \text{if } y = 0 \\ f_{X,Y}(0,1) + f_{X,Y}(1,1) = 0.3 + 0.4 = 0.7 & \text{if } y = 1 \end{cases} \end{aligned}$$

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.

Linear Combination of Independent Normal RVs is a Normal RV

We can get the following special property of normal RVs using Eqn. (3.72). If X_1, X_2, \dots, X_m be jointly independent RVs, where X_i is Normal(μ_i, σ_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 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 Normal(2, 4), Y be Normal(-1, 2) and Z be Normal(0, 1) RVs that are jointly independent. Obtain the following:

1. $E(3X - 2Y + 4Z)$
2. $V(2Y - 3Z)$
3. the distribution of $6 - 2Z + X - Y$
4. the probability that $6 - 2Z + X - Y > 0$
5. $\mathbf{Cov}(X, W)$, where $W = X - Y$.

Solution

1. $E(3X - 2Y + 4Z) = 3E(X) - 2E(Y) + 4(Z) = (3 \times 2) + (-2 \times (-1)) + 4 \times 0 = 6 + 2 + 0 = 8$
2. $V(2Y - 3Z) = 2^2 V(Y) + (-3)^2 V(Z) = (4 \times 2) + (9 \times 1) = 8 + 9 = 17$
3. From the special property of normal RVs, the distribution of $6 - 2Z + X - Y$ is
Normal $(6 + (-2 \times 0) + (1 \times 2) + (-1 \times -1), ((-2)^2 \times 1) + (1^2 \times 4) + ((-1)^2 \times 2))$
= Normal $(6 + 0 + 2 + 1, 4 + 4 + 2)$
= Normal(9, 10)
4. Let $U = 6 - 2Z + X - Y$ and we know U is Normal(9, 10) RV.
$$\begin{aligned} P(6 - 2Z + X - Y > 0) &= P(U > 0) = P(U - 9 > 0 - 9) = P\left(\frac{U - 9}{\sqrt{10}} > \frac{-9}{\sqrt{10}}\right) \\ &= P\left(Z > \frac{-9}{\sqrt{10}}\right) \\ &= P\left(Z < \frac{9}{\sqrt{10}}\right) \\ &\approx P(Z < 2.85) = 0.9978 \end{aligned}$$
- 5.

$$\begin{aligned} \mathbf{Cov}(X, W) &= E(XW) - E(X)E(W) = E(X(X - Y)) - E(X)E(X - Y) \\ &= E(X^2 - XY) - E(X)(E(X) - E(Y)) = E(X^2) - E(XY) - 2 \times (2 - (-1)) \\ &= E(X^2) - E(X)E(Y) - 6 = E(X^2) - (2 \times (-1)) - 6 \\ &= (V(X) + (E(X))^2) + 2 - 6 = (4 + 2^2) - 4 = 4 \end{aligned}$$

necessarily independent.

Remark 44 The converse is not true: two random variables that have zero covariance are not necessarily independent.

$$\mathbf{P}(Y > 2000) = 0.0475 + 0.0025 = 0.05$$

(try as a tutorial problem)

$$\mathbf{P}(Y > 2000) = \int_{2000}^{\infty} \left(\int_{-\infty}^{y=2000} 6 \times 10^{-6} e^{-0.001x - 0.002y} dy \right) dx +$$

piece $\{(x, y) : y < x, y > 2000, x < 2000\} \dots$ more involved but we get the same answer.
Alternatively, you can obtain $\mathbf{P}(Y > 2000)$ by directly integrating the joint PDF $f_{X,Y}(x, y)$ over the appropriate region (but

$$\begin{aligned} &= \mathbf{E}(X_1)\mathbf{E}(X_2) - \mathbf{E}(X_1)\mathbf{E}(X_2) \\ &= \mathbf{Cov}(X_1, X_2) = \mathbf{E}(X_1X_2) - \mathbf{E}(X_1)\mathbf{E}(X_2) \end{aligned}$$

From the formula for covariance

$$\mathbf{E}(X_1X_2) = \mathbf{E}(X_1)\mathbf{E}(X_2)$$

We know for independent RVs from the properties of expectations that

Solution:

Example 45 If X_1 and X_2 are independent random variables then what is their covariance $\mathbf{Cov}(X_1, X_2)$?

RVs X_i, X_{i-1}, \dots, X_1 is simply determined by the distribution of X_{i+1} .

Equality (3.67) simply says that the conditional distribution of the RV X_{i+1} given all previous

$$\mathbf{P}(X_{i+1} \leq x_{i+1}) =$$

$$\frac{\mathbf{P}(x_{i+1} \leq X_i \leq x_i, \dots, x_1)}{\mathbf{P}(X_i \leq x_i)}$$

$$= \frac{\mathbf{P}(x_{i+1} \leq X_i \leq x_i, \dots, x_1)}{\mathbf{P}(x_{i+1} \leq X_i \leq x_i, \dots, x_1)}$$

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

Proof:

$$\mathbf{P}(X_{i+1} \leq x_{i+1} | X_i \leq x_i, X_{i-1} \leq x_{i-1}, \dots, X_1 \leq x_1) = \mathbf{P}(X_{i+1} \leq x_{i+1}) \quad (3.67)$$

pendent sequence of RVs $\{X_1, X_2, \dots\}$, we have

Proposition 43 (Conditional probability of independent sequence of RVs) For an inde-

$$\begin{aligned} f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) &= f_{X_1}(x_1)f_{X_2}(x_2) \dots f_{X_m}(x_m) \\ f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) &= F_{X_1}(x_1)F_{X_2}(x_2) \dots F_{X_m}(x_m) \end{aligned} \quad (3.66)$$

independent if and only if for every $(x_1, x_2, \dots, x_m) \in \mathbb{R}_m^m$

From Definition 42, we say m random variables X_1, X_2, \dots, X_m are jointly independent or mutually

$$f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) = f_{X_1}(x_1)f_{X_2}(x_2) \dots f_{X_m}(x_m).$$

or equivalently,

$$\mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_m = x_m) = (\mathbf{P}(X_1 = x_1) \dots \mathbf{P}(X_m = x_m)).$$

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:

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

Thus marginal PDF gives us the probability density of a specific RV in a RV, irrespective of the value taken by the other RV in this RV.

By any distinct subset of indices $\{i_1, i_2, \dots, i_m\}$ of $\{1, 2, \dots\}$, the index set of the sequence of RVs

$$f_{X,Y}(x, y) = \begin{cases} 0 & \text{if } \exp(-\frac{1}{100}x - \frac{1}{100}y) \\ \frac{1}{10} \exp(-\frac{1}{100}x - \frac{1}{100}y) & \text{if } x < 0, y < 0, \\ & \text{otherwise.} \end{cases}$$

Example 48 Obtain the marginal PDF $f_Y(y)$ from the joint PDF $f_{X,Y}(x, y)$ of the continuous

RV in Example 45 that gave the response times of a web server.

Thus marginal PDF gives us the probability density of a specific RV in a RV, irrespective of the

value taken by the other RV in this RV.

and any sequence of real numbers $x_{i_1}, x_{i_2}, \dots, x_{i_m}$.

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 37 (Independence of Two RVs) Consider an \mathbb{R}^2 -valued RV $X := (X_1, X_2)$. Then the \mathbb{R} -valued RVs X_1 and X_2 are said to be independent or independently distributed if and only if

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

or equivalently,

$$F_{X_1, X_2}(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2),$$

for any pair of real numbers $(x_1, x_2) \in \mathbb{R}^2$.

By the above definition, for **discrete** RVs X_1, X_2 that are independent, the following equality is satisfied between the joint and marginal PMFs:

$$f_{X_1, X_2}(x_1, x_2) = \mathbf{P}(X_1 = x_1, X_2 = x_2) = \mathbf{P}(X_1 = x_1)\mathbf{P}(X_2 = x_2) = f_{X_1}(x_1)f_{X_2}(x_2) \text{ for any } (x_1, x_2) \in \mathbb{R}^2,$$

and for **continuous** RVs X_1, X_2 that are independent, the following equality is satisfied between the joint and marginal PDFs:

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2) \text{ for any } (x_1, x_2) \in \mathbb{R}^2.$$

In summary, two RVs X and Y are said to be **independent** if and only if for every (x, y)

$$F_{X, Y}(x, y) = F_X(x) \times F_Y(y) \quad \text{or} \quad f_{X, Y}(x, y) = f_X(x) \times f_Y(y)$$

Let us confirm that our familiar experiment of tossing a fair coin twice independently when encoded by a pair of independent Bernoulli(1/2) RVs satisfies the above definition.

Example 89 (Pair of independent Bernoulli(1/2) RVs) Let X_1 and X_2 be a pair of independent Bernoulli(1/2) RVs each taking values in the set $\{0, 1\}$ with the following tabulated probabilities. Verify that the JPMF $f_{X_1, X_2}(x_1, x_2) = 1/4$ for each $(x_1, x_2) \in \{0, 1\}^2$ is indeed given by the marginal PMF $f_{X_i}(x_i) = 1/2$ for each $i \in \{1, 2\}$ and each $x_i \in \{0, 1\}$.

	$X_2 = 0$	$X_2 = 1$	
$X_1 = 0$	1/4	1/4	1/2
$X_1 = 1$	1/4	1/4	1/2
	1/2	1/2	1

From the above Table we can read for instance that the *joint probability* that \mathbb{R}^2 -valued RV (X_1, X_2) takes the value or realization $(0, 0)$ is 1/4 from the first entry of the inner-most tabulated rectangle, i.e., $\mathbf{P}((X_1, X_2) = (0, 0)) = 1/4$, and that the *marginal probability* that the RV X_1 takes the value or realization 0 is 1/2, i.e., $\mathbf{P}(X_1 = 0) = 1/2$. Clearly, $1/4 = 1/2 \times 1/2$, and so our familiar experiment when seen as an \mathbb{R}^2 -valued RV is indeed composed of two independent \mathbb{R} -valued Bernoulli(1/2) RVs.

$$4. F_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) \rightarrow 0 \text{ as } x_1 \rightarrow -\infty, x_2 \rightarrow -\infty, \dots \text{ and } x_m \rightarrow -\infty$$

Definition 40 (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:

$$f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) = \mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_m = x_m). \quad (3.62)$$

Since $\mathbf{P}(\Omega) = 1$, $\sum_{(x_1, x_2, \dots, x_m) \in \mathcal{S}_{X_1, X_2, \dots, X_m}} f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) = 1$.

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 (3.57) but now over all m coordinates.

Definition 41 (Multivariate JPDF) (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 JPDF 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 $m = 3$):

$$F_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) = \int_{-\infty}^{x_m} \dots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) dx_1 dx_2 \dots dx_m, \quad (3.63)$$

and

$$\mathbf{P}(B) = \int \dots \int_B f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) dx_1 dx_2 \dots dx_m. \quad (3.64)$$

The JPDF satisfies the following two properties:

1. integrates to 1, i.e., $\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) dx_1 dx_2 \dots dx_m = 1$
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 JPDF (JPMF) over all other random variables. For example, the marginal PDF of X_1 is

$$f_{X_1}(x_1) = \int_{x_2=-\infty}^{\infty} \dots \int_{x_m=-\infty}^{\infty} f_{X_1, X_2, \dots, X_m}(x_1, x_2, \dots, x_m) dx_2 \dots dx_m$$

Definition 42 (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

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

or equivalently,

$$F_{X_{i_1}, X_{i_2}, \dots, X_{i_m}}(x_{i_1}, x_{i_2}, \dots, x_{i_m}) = F_{X_{i_1}}(x_{i_1})F_{X_{i_2}}(x_{i_2}) \dots F_{X_{i_m}}(x_{i_m}),$$

Example 90 Recall the \mathbb{R}^2 -valued continuous RV (X, Y) of Example 84 that is uniformly distributed on the unit square $[0, 1]^2$. First show that X and Y independent. Then show that both X and Y are identically distributed according to the Uniform(0, 1) RV.

This can be shown by checking that the joint PDF is indeed equal to the product of the marginal PDFs of Uniform(0, 1) RVs as follows:

Solution:

$$\begin{cases} f_{X,Y}(x,y) = f_X(x)f_Y(y) = 1 & \text{if } (x,y) \in [0,1]^2 \\ 0 & \text{otherwise} \end{cases}$$

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 (connected in 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 ?

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 l ?

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 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 RVs X_1 and X_2 are independent, we find that their joint probability density function (JPDF) is:

Solution:

$$f_{X_1, X_2}(x_1, x_2) = \frac{\pi}{2} \cdot \frac{1}{\pi} \cdot \frac{1}{\pi} = \frac{\pi}{4}$$

in the interval $[0, \pi]$. Hence assuming that the RVs X_1 and X_2 are independent, we find that their joint probability density function (JPDF) is:

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:

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 $1, 2, \dots, m$.

Latent 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

Definition 39 (multivariate JDF) The joint distribution function (JDF) or joint cumulative distribution function (CDF) of X_1, X_2, \dots, X_m is

extend the notions of CDF, PDF and JPDF to \mathbb{R}_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 consider the RV X whose components are the RVs X_1, X_2, \dots, X_m , i.e., $X = (X_1, X_2, \dots, X_m)$,

Definition 40 (\mathbb{R}^m -valued Random Variables)

2. What is $P(Y|Y' = 1)$?

$$P(Y \in \{0, 1, 2, 3 | Y' = 0) = \frac{P(Y=0)}{P(Y=0, Y'=0)}$$

$$P(Y = 3 | Y' = 0) = \frac{P(Y=3, Y'=0)}{P(Y'=0)}$$

$$P(Y = 2 | Y' = 0) = \frac{P(Y=2, Y'=0)}{P(Y'=0)}$$

$$P(Y = 1 | Y' = 0) = \frac{P(Y=1, Y'=0)}{P(Y'=0)}$$

$$P(Y = 0 | Y' = 0) = \frac{P(Y=0, Y'=0)}{P(Y'=0)}$$

$$P(Y = y | Y' = 0) = \frac{P(Y=y, Y'=0)}{P(Y'=0)} = \frac{P(\omega: Y'_i = \bar{y} \cup Y_i = y)}{P(\omega: Y'_i = \bar{y})} = \delta$$

1. What is conditional probability $P(Y|Y' = 0)$?

Figure 3.19: Diagrams done on the board!

The event A that the needle intersects one of the parallel ruled lines occurs if and only if

$$X_2 \leq l \sin(X_1) ,$$

i.e., if and only if the corresponding point $X := (X_1, X_2)$ falls in the region B , where B is part of the rectangle $[0, \pi] \times [0, L]$ lying between the x_1 -axis and the curve $x_2 = \sin(x_1)$ (area under the curve in right-subfigure of Figure 3.19). Hence, we can integrate the JPDF to get the probability of the event A of interest:

$$\mathbf{P}(A) = \mathbf{P}((X_1, X_2) \in B) = \int_B \frac{dx_1 dx_2}{\pi L} = \frac{2l}{\pi L}$$

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 $\mathbf{P}(A)$ 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 38 (Conditional PDF or PMF) Let (X_1, X_2) be a discrete bivariate RV. The conditional PMF of $X_1|X_2 = x_2$, where $f_{X_2}(x_2) := \mathbf{P}(X_2 = x_2) > 0$ is:

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

Similarly, if $f_{X_1}(x_1) := \mathbf{P}(X_1 = x_1) > 0$, then the conditional PMF of $X_2|X_1 = x_1$ is:

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

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

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

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

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

Let us consider a few discrete RVs for the simple coin tossing experiment \mathcal{E}_θ^3 that build on the Bernoulli(θ) RV X_i for the i -th toss in an **independent and identically distributed (IID.)** manner.

Table 3.1: The 8 ω 's in the sample space Ω of the experiment \mathcal{E}_θ^3 are given in the first row above. The RV Y is the number of ‘Heads’ in the 3 tosses and the RV Z is the number of ‘Tails’ in the 3 tosses. Finally, the RVs Y' and Z' are the indicator functions of the event that ‘all three tosses were Heads’ and the event that ‘all three tosses were Tails’, respectively.

ω :	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT	RV Definitions / Model
$\mathbf{P}(\omega)$:	$\frac{1}{8}$	$X_i \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\frac{1}{2})$							
$Y(\omega)$:	3	2	2	1	2	1	1	0	$Y := X_1 + X_2 + X_3$
$Z(\omega)$:	0	1	1	2	1	2	2	3	$Z := (1 - X_1) + (1 - X_2) + (1 - X_3)$
$Y'(\omega)$:	1	0	0	0	0	0	0	0	$Y' := X_1 X_2 X_3$
$Z'(\omega)$:	0	0	0	0	0	0	0	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.

$$\mathbf{P}(Y = y) = \left\{ \begin{array}{ll} & \\ & \end{array} \right. \quad \mathbf{P}(Y' = y') = \left\{ \begin{array}{ll} & \\ & \end{array} \right.$$

Classwork 94 (The number of ‘Heads’ given there is at least one ‘Tails’) Consider the following two questions.